Mashaghi Alireza (Orcid ID: 0000-0002-2157-1211)

Circuit topology predicts pathogenicity of missense mutations

Running title: *Topology-guided mutation analysis*

Jaie Woodard[1,2], Sumaiya Iqbal[3,4,5,6], Alireza Mashaghi[1,7]*

[1]Medical Systems Biophysics and Bioengineering, Leiden Academic Centre for Drug Research, Faculty of Science, Leiden University, 2333CC Leiden, Netherlands

[2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA

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

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

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

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

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/prot.26342

[7]Centre for Interdisciplinary Genome Research, Faculty of Science, Leiden University, 2333CC Leiden, Netherlands

ABSTRACT

The contact topology of a protein determines important aspects of the folding process. The topological measure of contact order has been shown to be predictive of the rate of folding. Circuit topology is emerging as another fundamental descriptor of biomolecular structure, with predicted effects on the folding rate. We analyze the residue-based circuit topological environments of 21K mutations labeled as pathogenic or benign. Multiple statistical lines of reasoning support the conclusion that the number of contacts in two specific circuit topological arrangements, namely inverse parallel and cross relations, with contacts involving the mutated residue have discriminatory value in determining the pathogenicity of human variants. We investigate how results vary with residue type and according to whether the gene is essential. We further explore the relationship to a number of structural features and find that circuit topology provides non-redundant information on protein structures and pathogenicity of mutations. Results may have implications for the polymer physics of protein folding and suggest that "local" topological information, including residue-based circuit topology and residue contact order, could be useful in improving state-of-the-art machine learning algorithms for pathogenicity prediction.

INTRODUCTION

One mechanism by which missense mutations can cause disease is by altering the folding properties of the protein in which the amino acid change is located. Databases of pathogenic and benign mutations have been established, where the domain of the protein containing the mutation can often be linked to a crystal structure or NMR structure in the Protein Data Bank (*1-5*). However, it is still a challenge to predict whether a given mutation will lead to disease, suggesting that common measures such as $\Delta\Delta G$, the change in free energy of folding upon mutation (*6*), may not account for all the relevant information contained in the structure (*7, 8*). This "missing information" may include kinetic rates and misfolding propensity (*9-13*). While energetics is important to determining a protein's folding and unfolding rates and folding robustness, the positioning of contacts along the protein chain also seems to play an important role. A quantity known as contact order, defined either as the average distance along the chain between contacting residues or that quantity divided by the chain length, has been shown to anticorrelate significantly with the folding rate (*14-16*). Topology may have general importance in defining pathogenicity of mutations, and as such, recent advances in molecular topology may have relevance to medicine.

There is a substantial history of the application of topology more generally to biomolecular structure and mutational analysis. Much early work investigated knot theoretical aspects of DNA (*17, 18*). Recently, algebraic topology, and persistent homology in particular (*19-22*), as well as differential geometry (*23*), have shown much success in protein-based analyses and predictions, where the combination of topology and machine learning has demonstrated much predictive power (*23*). However, none of these methods examine contact topology as it occurs along the protein chain. Therefore, such methods may lack predictability and proper interpretation where folding

kinetics is an important factor. In addition, although persistent homology has been applied to predict ΔΔG of a mutation within a protein (*21*), such methods have not been applied directly to prediction of pathogenicity. Contact order is a contact-based topological method that has shown some success in predicting protein folding rates (*10, 14*). However, it reduces protein structure to a single value, so it may be less informative than a method with multiple descriptors.

Circuit topology is a theoretical method for describing relations between pairs of contacts, as positioned on the protein backbone (*24-27*). Two contacts may be in parallel, in series, or in cross relation, as illustrated in Figure 1A. Briefly, we define the interval as the span of sequence between the contacting residues. Two contacts in series have non-overlapping intervals, contacts in cross have partially overlapping intervals, and in the case of parallel contacts, one interval is contained within the other. A distinction is made between whether a contact is in strict parallel with another contact (its interval is contained within the contact) or in inverse parallel with another contact (its interval contains the contact). Note that a first contact necessarily shortens the distance along the chain for all contacts in inverse parallel with this contact. This may also be the case for contacts in cross, depending on the nature of overlap. Considering the process of intramolecular diffusion, early folding can then facilitate later folding through establishment of parallel and cross relations. Following this logic, a theoretical dependence of folding rate on the types of relations between contacts has been proposed, for proteins and other linear heteropolymers (*28*) and was demonstrated for proteins with known experimental folding rates (*29*). In general, circuit topology may be related in integral ways to the folding and unfolding processes.

The residue-based circuit topology of a biological protein can be obtained from a crystal structure or NMR structure, following a published method (*24, 29*), while treating individual

residues as interacting segments. An example of the case of myoglobin is shown in Figure 1B-C. In Figure 1B, an arc is drawn between each pair of contacting sequence positions. Each row of the matrix in Figure 1C represents the relations of all other contacts with respect to a single contact drawn in Figure 1B. The use of graphs akin to those utilized in this study to visualize protein topology has precedent in the literature (*30*). In the present work, we omit secondary structure and direction information and denote each single amino acid as a separate node.

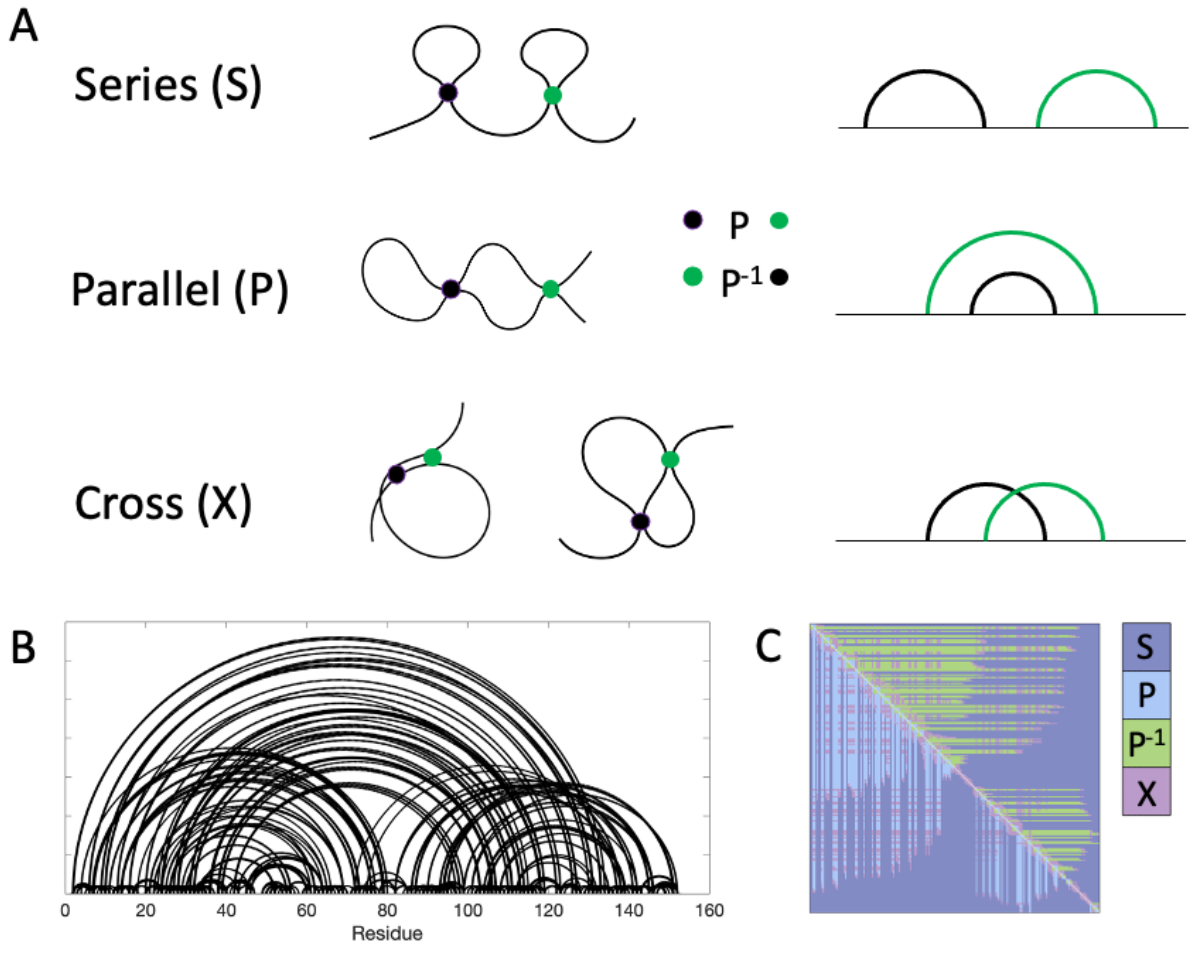

Figure 1. Molecular circuit topology. A) Definitions of contact relations: series (top), parallel (middle) and cross (bottom). Black P green (read black is in parallel with green) indicates that the interval of the black contact is contained in the interval of the green contact. Green P^{-1} black (green is in inverse parallel with black) indicates that the interval of the green contact is contained in the interval of the black contact. B) Circuit topology arc diagram for myoglobin (PDB ID: 1MBN). C) Circuit topology relations diagram for myoglobin. Each row corresponds to a single arc from (B).

Recently, it was shown that circuit topology figures in decision tree analysis in a way that is intuitive given folding kinetics, for lysosomal-storage-disorder-associated proteins exported from the endoplasmic reticulum (*31*). A large number of inverse parallel relations was predictive of non-responsiveness to pharmacological chaperone treatment. For a TIM barrel protein, the structure of which is contained in three lysosomal storage disorder proteins studied, residues with many inverse parallel relations were in fact early to fold (*32*), according to hydrogen-deuterium exchange experiments, consistent with a model where late to fold residues are more likely to be rescued. We speculate, however, that this may be a special case, dependent on the kinetics of the system.

Here, we utilize the newly developed ADDRESS database mapping human variants to structures available in the Protein Data Bank (*1*) to explore the relationship between circuit topology and pathogenicity. We seek to establish statistical differences between pathogenic and benign variants in the numbers of parallel, series, and/or cross relations relative to contacts involving the mutated residue. Using comparison of distributions, logistic regression, single decision tree analysis, and Random Forests, we find that the number of inverse parallel relations and the number of cross relations particularly inform on pathogenicity, where a greater number of such

relations indicates greater likelihood that the mutation will cause disease. We speculate that differences in the local topology of pathogenic and benign mutants reflect the nature of polymer folding in the biological environment.

METHODS

Database

We utilized an existing database containing structural and pathogenicity information, in order to determine the dependence of pathogenicity on various circuit topological features. We reference the ADDRESS database of pathogenic and benign mutations, which contains entries from the UniProt Humsavar database mapped to protein structures from the Protein Data Bank: https://zhanglab.ccmb.med.umich.edu/ADDRESS/download.html. Considering all entries that contain information on numbers of contacts, ΔΔG predicted by EvoEF (*33, 34*), and ΔΔG predicted by FoldX (*35*), our dataset contains 13,624 pathogenic mutations and 7,627 benign mutations. Consistent with the convention in EvoEF and FoldX, positive values of ΔΔG indicate destabilization of the protein, while negative values indication stabilization. Here, "pathogenicity" is curated based on literature references, according to the referenced version of the Humsavar database, and "the number of contacts" describes the number of residues contacting the WT residue at the mutated position, with a cutoff of 6 or more heavy atom contacts within 5 Angstroms of any type. Further, the MISCAST database (*5*) was referenced for the set of mutations that are mappable to protein structures and are common between ADDRESS and MISCAST to explore the relationship of circuit topological information to additional structural features.

Relations calculations

Dependence of pathogenicity on circuit topology relations was considered in this study. Matrices of parallel, series, and cross relations were calculated from contact maps identical to those used in the construction of the ADDRESS database, using methods described previously (*24*). Here, we simplified the matrices: concerted parallel relations were counted as parallel relations of the corresponding type, and concerted series relations were counted as series relations. Identity relations on the diagonal (parallel by definition) were not included in the relations count. For the calculation of numbers of relations associated with a mutation, we considered rows of the relations matrix corresponding to contacts involving the mutated residue, and for each column, we recorded whether any of the rows contained the specified topology relation. The sum of columns with this property corresponded to the number of relations.

Secondary structural element-based graphs

Secondary structural element based graphs were used to visualize protein structures at a resolution more easily visualized than amino acid based graphs such as that shown in Figure 1B Arc plots were generated as previously described (*24*), with a cutoff of 7 or more contacts within 3.8 Angstroms.

Logistic Regression

A logistic regression model was used as one method to determine the relative importance of features used in pathogenicity prediction. Logistic regression was carried out in Matlab 2020b. Coefficients and p-values were calculated using the mnrfit function. Relations values were set to zero when no contacts were present. Features considered were the number of contacts with other residues, number of cross relations, parallel relations, inverse parallel relations, and series relations with contacts involving the mutated residue, and protein length. The signs indicate the signs of coefficients from mnrfit.

Essential genes

Whether or not a given gene is essential was considered as a feature in machine learning methods. A list of essential genes was downloaded from the Database of Essential Genes (DEG) (*36*), which collects genes determined to be essential primarily referencing CRISPR/Cas9 editing on human cell types and high depletion of protein-truncated variants identified by whole-exome sequencing. The database contains 43,294 genes, 13,449 of which were in common with the genes of ADDRESS proteins. Genes that the ADDRESS and DEG had in common were considered to be essential. Proteins were marked as essential if they were contained within a DEG-annotated essential gene.

Decision trees

Single decision trees and random forests were used as additional models to determine feature importance and which features produce the best predictor, in addition to providing a classifier

based solely on structural data and information about mutated residue type. A single decision tree with default parameters was constructed on the entire dataset using the tree function in R. Additionally, the random forest analysis was carried out in R, dividing the data randomly into training (80% of the data) and test (20% of the data) sets. We used five-fold cross validation on the data from the training set to optimize the number of trees (Figure S1). Increasing ntree by 100 at a time, we considered the data to plateau at the point where less than 0.05% performance improvement (AUC) was achieved, at 500 trees. The parameter "ntree" was set to 500, "mtry" was set to 2, and "replacement" was set to TRUE. Relations and contact order values were set to zero when no contacts were present. Features considered for both the single tree and random forests approach were the residue type of the mutated residue, the residue type mutated to, the predicted $\Delta\Delta G$ from EvoEF and FoldX, whether or not the gene is essential, the number of parallel, inverse parallel, and cross relations, the number of concerted parallel and concerted series relations, and the local contact order, defined as the average sequence distance to residues in contact with the mutated residue. This analysis was intended to complement the logistic regression model, which additionally considered information on the number of series relations and protein length (but not contact order or free energy change). The single decision tree provides insight not seen in the logistic regression model, including how dependence on one feature, such as change in stability upon mutation, depends on another, such as whether the gene is essential.

Auto-ML

We used the Auto-ML program H2O (*37*) to train and test pathogenicity data from our dataset. We again trained on 4/5 of the data and tested on 1/5. On the training set, we performed 10-fold cross

validation and considered models with high validation AUC. For the leading model, a stacked ensemble, we calculated MCC on the test set.

RESULTS

Properties of database

Our dataset contains 13,624 pathogenic and 7,627 benign human variants mapped to protein structures, out of 30,255 pathogenic and 39,465 benign human variants total in the Uniprot Humsavar database. Pathogenic mutations stem from 1,192 different genes, while benign are from 2,464 genes. Most genes contain one to a few different mutations (Figure S2). Pathogenic genes are more likely than benign to contain many mutations, with 64 genes containing more than 50 mutations (vs. 9 such genes for benign mutations). Number of mutations from and to each amino acid type are shown in Figure S3. Some expected trends are seen: for instance, mutations from tryptophan or cysteine are particularly likely to be pathogenic as opposed to benign among our data, reflective of the roles of these residues in formation of the hydrophobic core and disulfide bonding. Interestingly, mutations to alanine are especially likely to be benign *vs.* pathogenic. Examination of the ADDRESS database statistics indicates that such mutations are dominated by mutations from threonine and valine, two amino acids that are small and similar to alanine. 77% of mutations are in the SNPdb database, as of the release of ADDRESS. Numbers of genes and mutations with various GO annotations in the database are shown in Table S1. For instance, 2,530 mutations are in DNA-binding proteins, while 2,183 mutations are in proteins with kinase activity.

Contact and relations distributions

A previous publication reported a moderate but highly significant difference in the number of residues in contact with the mutated residue, comparing pathogenic and benign variants in the ADDRESS database, based on UniProt Humsavar (*1*). Here, we compared mutations to a background distribution of all residue positions for all proteins in the database. We found that pathogenic mutations are shifted towards greater numbers of contacts with respect to the background, while benign mutations are shifted towards smaller numbers of contacts (Figure S4A, Table S2). The cutoff values (6 or more contacts within 5 Angstroms) for contacts were chosen based on visual inspection of circuit topology diagrams of simple helices and sheets. An alternate cutoff scheme of 5 or more contacts within 4.5 Angstroms also shows a clear difference in pathogenic vs. benign variants, but with somewhat lower significance (Table S2).

Next, we asked whether a difference exists for the number of circuit topology relations of a particular type associated with the mutated residue, or for the local contact order, defined as the mean intrachain distance to contacting residues. In these comparisons, we excluded cases in which the residue formed zero contacts. The two relations types with the greatest significance of the difference in means were cross and inverse parallel relations (Table S2, Figure S4B-C). While for the number of contacts and cross relations histograms, benign mutations appear more similar to the background than pathogenic, the opposite is true in the case of inverse parallel relations. In both cross and inverse parallel cases, pathogenic mutations are shifted towards greater numbers of relations, while benign mutations are shifted towards fewer.

Other topological measures show less (but still substantial) statistical significance in the difference between distributions (Table S2, Figure S5). Pathogenic mutations are shifted towards larger local

contact order, indicating that disruption of far-reaching contacts is more likely to result in pathogenicity. Pathogenic mutations are shifted towards greater numbers of parallel relations. Finally, there is a difference between pathogenic and benign mutations in the number of series relations. However, this final observation likely reflects, in part, biases in the database, where benign mutations may especially be catalogued for small proteins, given the high correlation of 0.88 between protein length and number of series relations. Preliminary work showed that absolute numbers, not divided by chain length, were far more informative in distinguishing between pathogenic and benign variants, so we do not divide by chain length for these features. Unnormalized local long range order (based on long range order, defined in (*38*)), defined as the number of contacts with amino acids separated by more than 12 residues in sequence distance, shows a highly significant difference that is still somewhat less significant than the total number of contacts. Due to this and due to the low importance of this feature calculated for the Random Forests model, we excluded long range order from other analysis.

Logistic Regression model

To further investigate which measures are most important in determining pathogenicity, we performed logistic regression on the dataset, considering circuit topological features along with protein length and number of contacts (Table 1). Here, the objective was to obtain the relative feature importance of the features. Protein length showed a significant p-value of 0.01, contributing positively towards pathogenicity, while the number of series relations was not a significant feature. It has been shown that longer proteins have slower folding rates, which may lead to higher pathogenicity of mutations (*39*).While the number of contacts showed the most

significant p-value, the number of cross relations and inverse parallel relations were also highly significant, both contributing positively towards pathogenicity. This suggests that the number of cross relations and inverse parallel relations are the two important topological measures in determining the pathogenicity of mutations.

Table 1. Results of logistic regression for pathogenicity prediction

	Sign	p-value
Number of contacts	+	$2x10^{-75}$
Cross relations	+	$5x10^{-22}$
Parallel relations	-	0.07
Inverse parallel relations	+	$2x10^{-8}$
Series relations	-	0.8
Protein length	+	0.01

We find that residues with a large number of cross relations relative to protein length and a residue with a large number of inverse parallel relations likely play important roles within protein structures, suggesting that their mutations would be likely detrimental. Furthermore, it is informative to consider the structural environments surrounding residues with near equal numbers of a given relations type, considering both pathogenic and benign variants.

We examine select proteins with large numbers of a given relation type with respect to a protein residue. First, we note two examples of cross-relation-rich two-domain proteins (Figure 2). Because residue-based diagrams can be difficult to view and interpret visually, we here show diagrams where secondary structural elements are the basic unit. Among proteins with a large number of residue-based cross relations per length is the serine protease neutrophil elastase (1B0F, ranked 9^{th} of all mutations). Neutrophil elastase's active site residues are contributed by both domains; therefore, it is crucial that the two domains bind to each other via a specific interaction.

Residue 141 is at the interaction interface and is part of a loop in the C-terminal domain. Mutation of this tryptophan residue to either cysteine or arginine is pathogenic. The benign-case residue with the largest number of cross relations is beta-B3 crystallin (3QK3) residue 105. The arginine in the N-terminal domain interacts with glutamate in the C-terminal domain. The evolutionarily related protein Beta-B2 crystallin has domains separated, suggesting that interdomain interaction may be less vital to function. The circuit topology diagrams of the two proteins are shown in Figure 2, revealing similarities in topology. Interestingly, the circuit topology diagram of neutrophil elastase exhibits an almost fractal-like structure, with each colored domain in Figure 2C showing a similar structure to the highlighted purple topology in Figure 2D (see comparison in 2E). It is possible that the same types of interactions that stabilizing local regions of the protein are also suitable in the case of longer-range interactions. Serine proteases of this structural type are known to have high kinetic stability, which may be facilitated in part by large numbers of cross relations and the presence of the Greek key topological motif. In general, the degree of conservation of domain-domain interaction varies greatly among proteins (*40*), and it will be interesting to investigate in further detail the relationship between this conservation and circuit topology.

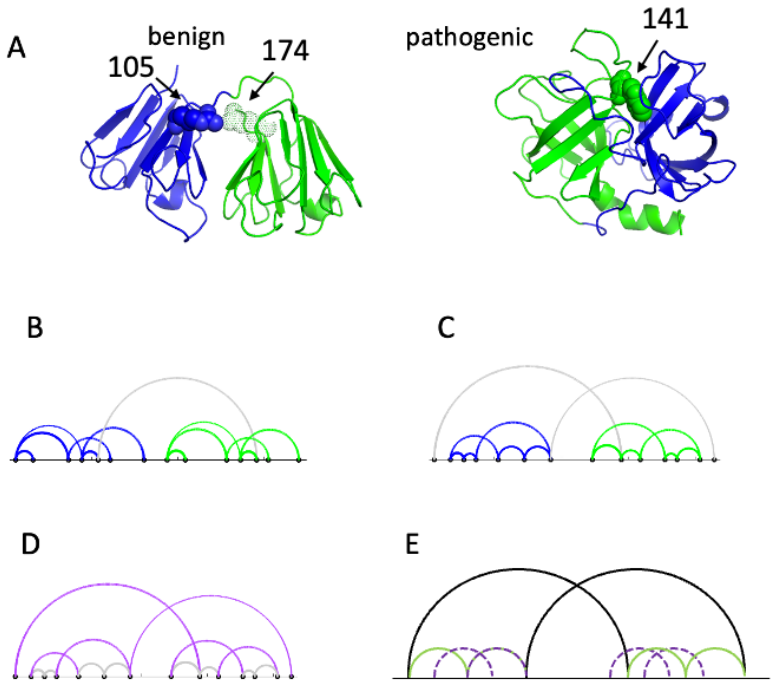

Figure 2. Topologies of beta-B3 crystallin and neutrophil elastase. A) left: beta-B3 crystallin. Residue. 105 shown in sphere representation and interacting residue 174 shown as dots. Right: neutrophil elastase, with residue 141 shown in sphere representation. B) secondary structural element-based topology diagram for beta-B3 crystallin. C) topology diagram for neutrophil elastase. D) topology diagram for neutrophil elastase, with color depicting similarity of long-range structure to shorter range structure in (C). E) superimposed diagrams highlighting similarity of (C), green and black, and (D), purple and black.

The residue with the largest number of cross relations per protein length is residue 651 (tryptophan) of the NC1 domain of collagen X (1GR3). Mutation to arginine causes heritable disease. Figure 3A shows the topology diagram, with contacts with the gray point indicated by colored lines. The contacting secondary structural elements are displayed in like colors in Figure 3B. It is evident that the cross relations have a role in bringing together sequence-distant parts of the chain. The inverse

parallel relation is also important to this structure, as seen by considering the shortest range contact with endpoint P^{-1}, which is in parallel with 7 out of 9 other secondary structural element based contacts. The residue with the most inverse parallel interactions for this protein ranks 4th of all residues, indicating that large numbers of local cross and inverse parallel relations can exist in the same protein. Finally, a residue pair with a large number of parallel relations (502 and 246 of 4WXQ) is shown in Figure 3C-D. This is a beta propeller protein, where a repeated beta structure loops back, such that the terminal beta strands form a strong interaction. We note that while structural element based graphs are shown in Figures 1 and 2 for ease of viewing, thee statistics and rankings were obtained using residue-based contacts.

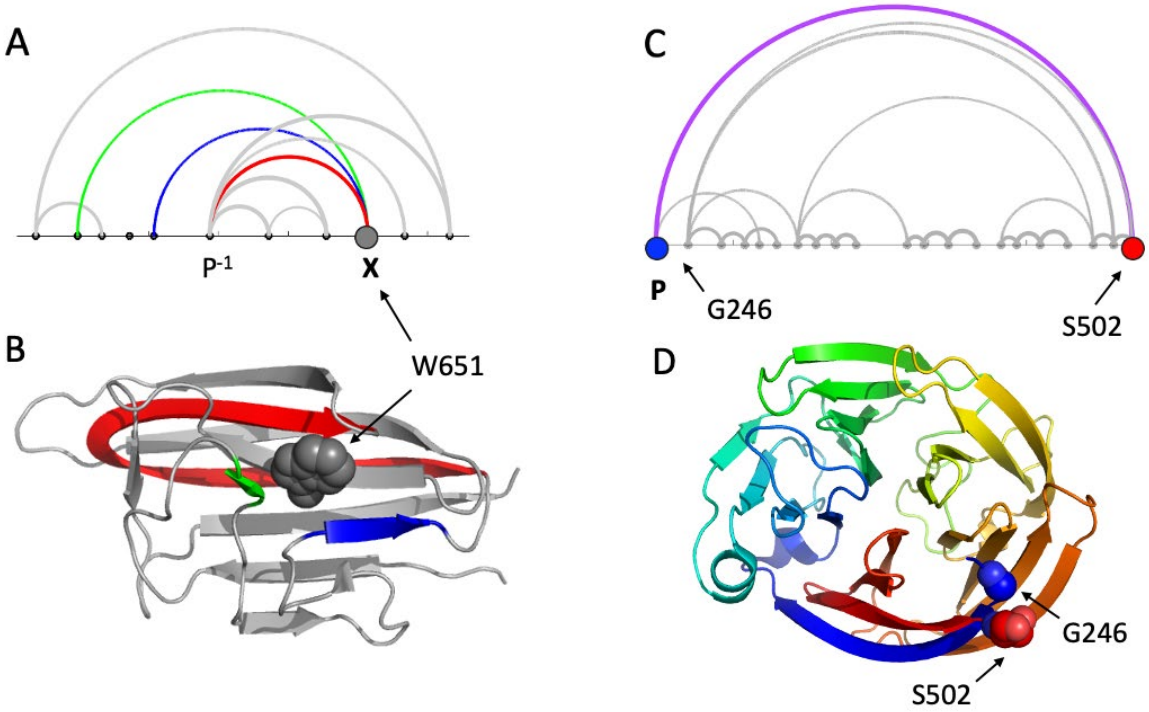

Figure 3. Topology of proteins with large numbers of residue-based topology relations. Mutated residues are labeled. A-B) PDB ID 1GR3. Colors indicate contacts with mutated residue (spheres).

C-D) Propeller protein 4WXQ. Contact formed by residues in sphere representation in D is colored purple in C.

Proteins encoded by essential genes

We considered whether the difference between pathogenic and benign mutations could be affected by whether or not a gene is essential. The difference in the means of number of contacts distributions is slightly greater for proteins encoded by essential genes vs. proteins encoded by non-essential genes (Table 2, Figure S6). However, for cross and especially for inverse parallel relations, which we saw to also be important in determining pathogenicity, non-essential genes actually show a greater difference in means than essential ones. Figure S6C shows that non-essential-gene proteins with relatively few numbers of inverse parallel relations are especially likely to be benign. Other topological measures show less substantial differences (Figure S7).

Table 2. Difference in means of topological features for essential and non-essential genes

	Essential Δmean	**Non-essential Δmean**	**p-value Essential**	**p-value non-essential**
Number of contacts	1.24	1.16	$4x10^{-162}$	$7x10^{-120}$
Cross relations	32.8	39.4	$3x10^{-83}$	$2x10^{-98}$
Parallel relations	50.1	37.9	$2x10^{-30}$	$4x10^{-40}$
Inverse parallel relations	14.3	28.5	$2x10^{-29}$	$1x10^{-83}$
Local contact order	11.7	15.2	$3x10^{-20}$	$3x10^{-24}$

Topology by residue type

Different residue types may be expected to promote different circuit topology relations. We calculated the ratio of the average numbers of inverse parallel and parallel relations and the average number of cross relations, for mutations from each residue type (Figure S8-A). For the inverse

parallel to parallel ratio, polar residues tended to have higher values for benign mutations than for pathogenic mutations, while the two values were generally more similar for hydrophobic and charged residues. This is potentially because mutations in polar residues tend to be less disruptive of the brining together of disparate residues through formation of inner contacts. Cysteine exhibits a different trend, with a substantially greater value for pathogenic mutations vs. benign ones. We considered disulfides alone, estimated by considering contacts between cysteine sulfur atoms with distance of less than 2.3 Angstroms. Of 446 identified mutated disulfides, 93% of mutations were pathogenic. In fact, the inverse parallel to parallel ratio for pathogenic disulfides was higher than this ratio for pathogenic mutations of all residue types. This may indicate an important role of disulfides in bringing together residues far apart in sequence.

The number of cross relations also differs by mutated residue type (Figure S8-B). Pathogenic mutations involving large hydrophobic and aromatic residues have the largest number of cross relations. This is likely because such residues play important structural roles in holding together the protein. Note that values reflect in part the available residue types to which a residue may mutate. For instance, Isoleucine can mutate to valine, which would be expected to be less disruptive of the cross relation than mutations available to tryptophan (cysteine or arginine), perhaps influencing the ratios of benign to pathogenic mutations for these residue types. The disulfide has smaller overall values due to the fact that only one contact is considered, while residues in general may form more than one contact.

Comparative importance of features

We further investigated relationships among factors contributing to pathogenicity. A map of correlations is shown in Figure 4. As previously noted, a high correlation is seen between protein

length and number of series relations. This is intuitive, because given a contact, a longer protein is more likely to have a large number of contacts in tandem. The correlation of length with number of cross and inverse parallel relations is much smaller but still greater than 0.3. High correlations are also seen between contact order and the number of cross and especially the number of parallel relations, since contacts that span a greater sequence length are able to contain a greater number of contacts; a high correlation is likewise seen between the number of parallel and cross relations. The parallel relation likewise has a high correlation with contact order, since contact order is higher for contacts that span a long sequence distance, and these are expected to have more contacts in parallel with (within) the contact span. As noted previously (*1*), there is a correlation of 0.67 between the two computational methods of predicting ΔΔG of the mutation. There is a weak to moderate correlation between predicted ΔΔG predicted by each method and the number of contacts. The number of contacts correlates moderately with the number of cross, parallel, and inverse parallel relations. However, the logistic regression described in a previous section of this paper indicates that the numbers of cross and inverse parallel relations are themselves important to pathogenicity prediction even in the presence of number of contacts information. Particularly low correlations are seen between predicted ΔΔG and protein length, local contact order, or number of series relations. There is a somewhat higher correlation between ΔΔG and number of inverse parallel or cross relations, perhaps indicating that contacts promoting these relations are important in nucleating folding and so are somewhat stronger energetically.

	EvoEF	FoldX	NC	X	P	IP	S	CO	length
EvoEF		0.67	0.37	0.24	0.13	0.16	0.09	0.10	0.03
FoldX	0.67		0.28	0.22	0.13	0.15	0.09	0.10	0.04
NC	0.37	0.28		0.58	0.36	0.38	0.28	0.27	0.12
X	0.24	0.22	0.58		0.70	0.48	0.41	0.64	0.32
P	0.13	0.13	0.36	0.70		0.19	0.29	0.87	0.26
IP	0.16	0.15	0.38	0.48	0.19		0.47	0.12	0.39
S	0.09	0.09	0.28	0.41	0.29	0.47		0.23	0.88
CO	0.10	0.10	0.27	0.64	0.87	0.12	0.23		0.25
length	0.03	0.04	0.12	0.32	0.26	0.39	0.88	0.25	

Figure 4. Statistical relationships among energetic and topological factors. Correlations, with high correlations colored blue and low correlations colored red.

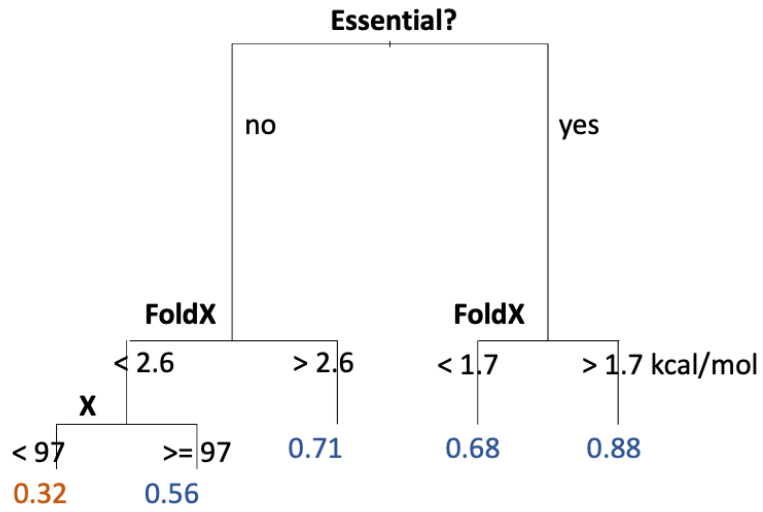

Figure 5. Single decision tree showing probability of pathogenicity, generated in R. The decision tree has an MCC of 0.35. FoldX free energy change predictions are in units of kcal/mol.

Using R, we constructed an optimal single decision tree (Figure 5) incorporating features other than length, which may be problematic due to redundancies and other biases in the database, and

number of series relations, which correlates strongly with length. The decision tree first branches at the distinction between essential and non-essential genes, where mutations in essential genes are predicted to be pathogenic. Both essential and non-essential genes then branch at the FoldX predicted ΔΔG, with larger values of ΔΔG promoting pathogenicity. The boundary value is greater for non-essential genes, indicating that essential genes are more sensitive to small destabilizations. Non-essential genes with low ΔΔG then split according to the number of cross relations, where high numbers of cross relations predict pathogenic mutations. It is possible that mutations of residues with large numbers of cross relations may promote misfolding and aggregation, a hypothesis which can be further investigated in future studies. This analysis illustrates that while whether the gene is essential and the free energy change are most important to determining pathogenicity, relations information plays a non-trivial role.

We next applied Random Forests in R to determine whether inclusion of topological features may improve machine learning performance. In preliminary analysis, we found that performance metrics may be inflated by including global factors of the protein, since mutations in the database are often contained within the same protein, and some proteins contain completely or predominantly pathogenic or benign mutations due to aspects of the Humsavar database on which our database ADDRESS is based. We therefore only considered local features in feature set construction, also excluding the number of series relations, which correlates strongly with length. We considered all possible combinations of features for 2, 3, 4, and 5 total features and identified the set with the highest AUC in each case (Table 3). For two features, the FoldX ΔΔG and whether the gene is essential gave the highest AUC, consistent with the simple decision tree. For greater numbers of features, however, FoldX results are not included and instead the amino acid type

mutated from and the number of cross relations appears important. For five variables, the number of inverse parallel relations appears, along with the position of the mutated residue along the chain. Table S3 shows the results as in Table 3 for a range of ntree values, showing that MCC values tend to level off for each number of features after about 200 trees. The full model has an AUC of 0.81 and an MCC of 0.43 on the test set (20% of the data). We show also the complete confusion matrix in Table 4, taking into consideration that AUC would not be expected to be the best metric of performance, since our dataset is imbalanced. From this can be obtained the recall or sensitivity, 90%, the precision, 76%, and the specificity, 49%, of a mutation being pathogenic, and the accuracy of classification, 75%. Calculated feature importance (Table S4) for the entire dataset shows again that among topological features, the numbers of inverse parallel and cross relations are most important in predicting pathogenicity.

Additionally, we ran H2O AutoML (*37*) on training data to identify models with high performance. The best method was a stacked ensemble with a 10-fold cross validated AUC of 0.81 and MCC on the test set of 0.44. The best single (non-ensemble) method was a Gradient Boosting Machine method with an AUC of 0.80.

Table 3. Random Forests performance for pathogenicity prediction, ntree = 500

number of features	2	3	4	5	complete (14)
AUC	0.693	0.741	0.770	0.794	0.807
features	FoldX ΔΔG essential?	aa1 X essential?	aa1 X position essential?	aa1 X P^{-1} position essential?	

Essential: whether or not the gene is essential
aa1: amino acid type mutated from
X: number of cross relations
position: the position of the residue along the chain divided by chain length
P^{-1}: number of inverse parallel relations

Table 4. Confusion matrix for random forests model based on full feature set. MCC = 0.434.

		actual	
		Benign	Pathogenic
prediction	benign	743	276
	pathogenic	780	2452

Relationship of circuit topology to known structural features

Finally, to test for redundancy to known structural features, we compared the circuit topology information to computed features from the MISCAST database (*5*), for a set of 5,005 mutations (3,225 pathogenic and 1,780 benign) mapped to structures and are in common between ADDRESS (13,634 pathogenic mutations and 7,627 benign mutations) and MISCAST (32, 923 pathogenic mutations and 164, 915 general population variants). Notably, a relationship was seen between circuit topology measures and both secondary structure and measures of solvent exposure. The example of the cross relation, for which the most substantial correlations and differences were seen, is shown in Figure S9. Interestingly, given that in MISCAST beta sheet structures have an odds ratio for pathogenicity greater than 1, larger numbers of cross relations contribute towards pathogenicity (Table 1, Figure S4), and beta sheet residues have a larger number of cross relations on average than helix or coil residues (Figure S9-A). We carried out logistic regression with and without circuit topological features, based on the Humsavar annotations of pathogenicity, considering also the number of contacts from the ADDRESS database and the following features from MISCAST: residues' exposure to solvent; coil, helix, or sheet secondary structure; location of the active site; metal binding; binding site; DNA binding site; nucleotide phosphate binding

region; calcium binding region; and disulfide bond. Inclusion of circuit topological features decreased the deviance of the fit from 5.811×10^3 to 5.776×10^3, demonstrating that circuit topological features are non-redundant with other commonly referenced structural features.

DISCUSSION

Recent monumental advances in protein structure prediction have shown that we can predict, with a high degree of accuracy, the structure of a protein, given its amino acid sequence (*41, 42*). However, a wealth of experimental data has shown that fold switches upon point mutation (the type of sequence change most relevant to our understanding of human disease) are rare (*43*). Instead, missense mutation seems to alter the folding stability of a protein, in addition to other properties such as aggregation propensity, binding affinity to other proteins and/or ligands, catalysis, and dynamic aspects relevant to protein function. Furthermore, there is still much to learn about the details of the folding process and how this process is derailed in cases of misfolding and aggregation (*44-46*). One approach is to study the impact of features that are meaningful in terms of the polymer physics of protein folding. While future efforts will likely incorporate sophisticated applications of machine learning methods, including neural networks, here we have taken a first step towards understanding the importance of yet uninterrogated chain properties, using concepts from circuit topology. Our results suggest that approaches based on deep learning of contact maps may have strong predictive value; however, such approaches may lack interpretability in terms of the details of the chain-folding process.

Although protein stability depends on the ratio of folding and unfolding rates, it is important to remember that proteins are produced and degraded, and interact, at rates comparable to folding. It

will therefore be important to understand the effects of mutations in terms of kinetic models of folding and interaction in the relevant biological environment(s), *e.g.* (*47*). Part of the inability of folding stability predictors to fully capture predictability in mutation pathogenicity may be due to an important role of kinetics (*9, 13*), which topology may in part capture.

The large number of available examples in our dataset of 21K pathogenic and benign mutations allows us to draw statistically significant conclusions based on relatively small trends. We present multiple statistical assessments as multiple lenses for viewing a rich, though somewhat statistically biased, dataset. An important overall conclusion is the discriminatory value of the cross-relation. Further work will be needed to show the reason for this trend and whether it can be explained by considering other aspects of folding and topology. It is possible that mutation of contacts with large numbers of cross relations disrupts important folding nucleation sites and/or especially promotes misfolding and aggregation, which is supported by the observation that residues in beta sheet secondary structures contain greater numbers of cross relations on average than other secondary structure types. The inverse parallel relation also shows substantial contribution to pathogenicity. According to models of folding (*28, 31*), residues with many inverse parallel relations are likely to be early to fold, indicating that they would be expected to disrupt rate of folding. The hypothesis that the relative importance of the inverse parallel and cross relations indicate an importance of folding kinetics and non-native interactions in determining pathogenicity can be validated in the future using newly developed all-atom models of (un)folding and misfolding (*48*).

While number of contacts has the highest feature importance in the logistic regression model, it is relatively unimportant according to random forests. This suggests, perhaps, a more nuanced connection between circuit topology and pathogenicity. The logistic regression model also does not include stability change information, which number of contacts information may report on.

This work relates to a recent study evaluating the importance of structural features in predicting and explaining drug responsiveness in lysosomal storage disorders (*31*). In this study, it was found that the dependence on the inverse parallel relation according to a simple decision tree is consistent with a simple kinetic model of folding, binding, and export. In the present study, we also see a dependence on the inverse parallel relation, according to histograms, logistic regression, and random forests approaches, but we also see a dependence on the cross relation, which may be important for many proteins more generally. The simple decision tree suggests that this dependence on the cross relation is particularly important for non-essential genes. Consistent with the lysosomal storage disorders study, destabilizing mutations were seen as mediating pathogenicity (to a greater extent than stabilizing mutations, for instance).

Pathogenicity prediction is important in prioritizing variants for experimental study and ultimately for genetics-based decision making within personalized medicine and genetic counseling and for drug discovery. Early methods relied primarily on evolutionary conservation, where mutation of a well conserved residue is more likely to be pathogenic (*4, 49*). Recently developed, advanced methods also incorporate structural information and predictions (*4, 8, 50*). In the case that it is non-redundant with other features, circuit topology information has the potential to improve state-of-the-art predictors, as well as providing mechanistic insight. It will be elucidating to combine the

analysis presented here with binding information and predictions, *e.g.,* (*51*). It will also be interesting to compare to other established databases, such as COSMIC (*52, 53*), which catalogs cancer-associated mutations, for which we may see important similarities and also differences.

Performance metrics for our method alone are somewhat less than those for SIFT (MCC of 0.54 on complete HumVar, (*54*)) and especially advanced methods such as DAMPred (MCC of 0.601 on a similar dataset, (*8*)), which utilize multiple sequence alignments of homologous proteins. In fact, sequence-based methods were found to be the best techniques for distinguishing between pathogenic and benign mutants, with structural information thus far improving methods by a relatively small amount (*8, 55*). However, we stress that such methods do not provide information on why, structurally a mutation is pathogenic or benign. Therefore, we believe the appropriate comparison is to other structure-based methods, such as MISCAST, next to which we find our metrics provide non-redundant information on pathogenicity. We also wish to stress that the relation between topological information and pathogenicity of mutations has not been studied extensively, and thus new insights are expected to emerge from future studies. While the present work indicates that circuit topology relations are important in pathogenicity determination, the full circuit topology matrix contains much information beyond the numbers of parallel, series, and cross relations. Treatment of the full matrix, along with machine learning and/or other advanced analysis techniques is likely to yield additional insights and predictive value, perhaps providing further improvement to sequence and structure based methods and structure-only analyses. How well structure-based methods can perform without the aid of sequence alignment, is an interesting question in itself which reflects our current understanding of the biophysics of mutations. An exciting endeavor for future studies, in addition, is the integration of structure-based methods with

information on the systems biology of the relevant molecular networks, beyond information on whether or not a gene is essential.

Much past research has focused on specific folding topologies and aspects unique to particular types of proteins: for instance, specific folds and relationships between them; the mechanism by which loop conformational dynamics promotes catalysis; and the unique features of proteins that are very thermostable or adapted to a particular environment. Other studies focus on commonalities between proteins, such as the two-state folding of many small proteins or the dependence of aggregation potential on properties such as charge and hydrophobicity. Circuit topology strikes a middle ground between these two types of perspectives and may ultimately help to bridge them. While all proteins with substantial order have relations between contacts of parallel, series, and/or cross types, the details of the relations map are specific to each protein. We can therefore use circuit topology to ask questions encompassing a range of levels of detail, using a variety of methods, data, and perspectives. Circuit topology is not limited to proteins, nor to biological matter, and it may have utility in the construction of new molecules and other materials. Biological proteins, however, provide a rich example of possibilities within heteropolymer chains, which are present in large, medically relevant datasets. While physics-based studies of proteins have the potential to spur medical advances, this study has provided an example of how, in addition, medically relevant databases may help us understand the physical polymer properties of proteins, here facilitated by a simple mathematical approach.

DATA AND SOFTWARE SHARING

Data and software are available in the Supplementary Information and via the following Github repository: https://github.com/circuittopology/local_circuit_topology

ACKNOLEDGEMENTS

We thank Dr. Gil Omen and Eric Bell for comments on drafts of the manuscript and Dr. Chengxin Zhang for helpful discussions. We thank Elnaz Banijamali for communication regarding preliminary tests of parameter values.

1. J. Woodard, C. Zhang, Y. Zhang, ADDRESS: A Database of Disease-associated Human Variants Incorporating Protein Structure and Folding Stabilities. *J Mol Biol*, 166840 (2021).
2. T. D. Luu *et al.*, MSV3d: database of human MisSense Variants mapped to 3D protein structure. *Database (Oxford)* **2012**, bas018 (2012).
3. R. Karchin *et al.*, LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* **21**, 2814-2820 (2005).
4. N. M. Ioannidis *et al.*, REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016).
5. S. Iqbal *et al.*, MISCAST: MIssense variant to protein StruCture Analysis web SuiTe. *Nucleic Acids Res* **48**, W132-W139 (2020).
6. E. H. Kellogg, A. Leaver-Fay, D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830-838 (2011).
7. L. Gerasimavicius, X. Liu, J. A. Marsh, Identification of pathogenic missense mutations using protein stability predictors. *Sci Rep* **10**, 15387 (2020).
8. L. Quan, H. Wu, Q. Lyu, Y. Zhang, DAMpred: Recognizing Disease-Associated nsSNPs through Bayes-Guided Neural-Network Model Built on Low-Resolution Structure Prediction of Proteins and Protein-Protein Interactions. *J Mol Biol* **431**, 2449-2459 (2019).
9. U. Bastolla, P. Bruscolini, J. L. Velasco, Sequence determinants of protein folding rates: positive correlation between contact energy and contact range indicates selection for fast folding. *Proteins* **80**, 2287-2304 (2012).
10. A. V. Glyakina, O. V. Galzitskaya, How Quickly Do Proteins Fold and Unfold, and What Structural Parameters Correlate with These Values? *Biomolecules* **10**, (2020).
11. A. M. Fernandez-Escamilla, F. Rousseau, J. Schymkowitz, L. Serrano, Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**, 1302-1306 (2004).
12. P. J. Waters, Degradation of mutant proteins, underlying "loss of function" phenotypes, plays a major role in genetic disease. *Curr Issues Mol Biol* **3**, 57-65 (2001).
13. R. Godoy-Ruiz *et al.*, Natural selection for kinetic stability is a likely origin of correlations between mutational effects on protein energetics and frequencies of amino acid occurrences in sequence alignments. *J Mol Biol* **362**, 966-978 (2006).
14. K. W. Plaxco, K. T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* **277**, 985-994 (1998).
15. M. O. Lindberg *et al.*, Folding of circular permutants with decreased contact order: general trend balanced by protein stability. *J Mol Biol* **314**, 891-900 (2001).
16. A. Broom, S. Gosavi, E. M. Meiering, Protein unfolding rates correlate as strongly as folding rates with native structure. *Protein Sci* **24**, 580-587 (2015).
17. D. Sumners, in *Proceedings of Symposia in Applied Mathematics*. (1992).
18. D. Sumners, The **knot theory** of **molecules**. *Journal of mathematical chemistry* **1**, 1-14 (1987).
19. K. Xia, G. W. Wei, Persistent homology analysis of protein structure, flexibility, and folding. *Int J Numer Method Biomed Eng* **30**, 814-844 (2014).
20. Z. Cang, G. W. Wei, Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int J Numer Method Biomed Eng* **34**, (2018).

21. Z. Cang, G. W. Wei, Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics* **33**, 3549-3557 (2017).
22. M. Wang, Z. Cang, G. W. Wei, A topology-based network tree for the prediction of protein-protein binding affinity changes following mutation. *Nat Mach Intell* **2**, 116-123 (2020).
23. D. D. Nguyen, Z. Cang, G. W. Wei, A review of mathematical representations of biomolecular data. *Phys Chem Chem Phys* **22**, 4343-4367 (2020).
24. O. Schullian, J. Woodard, A. Tirandaz, A. Mashaghi, A circuit topology approach to categorizing changes in biomolecular structure. *Frontiers in Physics* **8**, (2020).
25. N. Nikoofard, A. Mashaghi, Implications of Molecular Topology for Nanoscale Mechanical Unfolding. *J Phys Chem B* **122**, 9703-9712 (2018).
26. A. Mashaghi, R. J. van Wijk, S. J. Tans, Circuit topology of proteins and nucleic acids. *Structure* **22**, 1227-1237 (2014).
27. B. Scalvini *et al.*, Topology of Folded Molecular Chains: From Single Biomolecules to Engineered Origami. *Trends in Chemistry* **2**, 609-622 (2020).
28. A. Mugler, S. J. Tans, A. Mashaghi, Circuit topology of self-interacting chains: implications for folding and unfolding dynamics. *Phys Chem Chem Phys* **16**, 22537-22544 (2014).
29. B. Scalvini, V. Sheikhhassani, A. Mashaghi, Topological principles of protein folding. *Phys Chem Chem Phys* **23**, 21316-21328 (2021).
30. I. Koch, T. Schäfer, Protein super-secondary structure and quaternary structure topology: theoretical description and application. *Curr Opin Struct Biol* **50**, 134-143 (2018).
31. J. Woodard, W. Zheng, Y. Zhang, Protein structural features predict responsiveness to pharmacological chaperone treatment for three lysosomal storage disorders. *PLoS Comput Biol* **17**, e1009370 (2021).
32. R. Jain *et al.*, A conserved folding nucleus sculpts the free energy landscape of bacterial and archaeal orthologs from a divergent TIM barrel family. *Proc Natl Acad Sci U S A* **118**, (2021).
33. R. Pearce, X. Huang, D. Setiawan, Y. Zhang, EvoDesign: Designing Protein-Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. *J Mol Biol* **431**, 2467-2476 (2019).
34. X. Huang, R. Pearce, Y. Zhang, EvoEF2: accurate and fast energy function for computational protein design. *Bioinformatics* **36**, 1135-1142 (2020).
35. J. Schymkowitz *et al.*, The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-388 (2005).
36. H. Luo *et al.*, DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. *Nucleic Acids Res* **49**, D677-D686 (2021).
37. E. LeDell, S. Poirier, H2O AutoML : Scalable Automatic Machine Learning. *7 th ICML Workshop on Automated Machine Learning (AutoML)*, (2020).
38. M. M. Gromiha, S. Selvaraj, Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol* **310**, 27-32 (2001).
39. M. Corrales *et al.*, Machine Learning: How Much Does It Tell about Protein Folding Rates? *PLoS One* **10**, e0143166 (2015).
40. J. H. Han, N. Kerrison, C. Chothia, S. A. Teichmann, Divergence of interdomain geometry in two-domain proteins. *Structure* **14**, 935-945 (2006).

41. A. W. Senior *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710 (2020).
42. J. Yang *et al.*, The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7-8 (2015).
43. R. A. Laskowski, J. M. Thornton, Understanding the molecular machinery of genetics through 3D structures. *Nat Rev Genet* **9**, 141-151 (2008).
44. F. Chiti, C. M. Dobson, Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu Rev Biochem* **86**, 27-68 (2017).
45. I. Pallarés, S. Ventura, Advances in the Prediction of Protein Aggregation Propensity. *Curr Med Chem* **26**, 3911-3920 (2019).
46. J. Ferina, V. Daggett, Visualizing Protein Folding and Unfolding. *J Mol Biol* **431**, 1540-1564 (2019).
47. S. Bershtein, W. Mu, A. W. Serohijos, J. Zhou, E. I. Shakhnovich, Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Mol Cell* **49**, 133-144 (2013).
48. A. Bitran, W. M. Jacobs, E. Shakhnovich, Validation of DBFOLD: An efficient algorithm for computing folding pathways of complex proteins. *PLoS Comput Biol* **16**, e1008323 (2020).
49. P. Kumar, S. Henikoff, P. C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009).
50. L. Ponzoni, I. Bahar, Structural dynamics is a determinant of the functional significance of missense variants. *Proc Natl Acad Sci U S A* **115**, 4164-4169 (2018).
51. Q. Wu, Z. Peng, Y. Zhang, J. Yang, COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res* **46**, W438-W442 (2018).
52. H. C. Jubb, H. K. Saini, M. L. Verdonk, S. A. Forbes, COSMIC-3D provides structural perspectives on cancer genetics for drug discovery. *Nat Genet* **50**, 1200-1202 (2018).
53. J. G. Tate *et al.*, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res* **47**, D941-D947 (2019).
54. R. Vaser, S. Adusumalli, S. N. Leng, M. Sikic, P. C. Ng, SIFT missense predictions for genomes. *Nat Protoc* **11**, 1-9 (2016).
55. J. Zhang *et al.*, Assessing predictions on fitness effects of missense variants in calmodulin. *Hum Mutat* **40**, 1463-1473 (2019).