*Bioinformatics*, 33(22), 2017, 3549–3557
doi: 10.1093/bioinformatics/btx460
Advance Access Publication Date: 14 July 2017
Original Paper

OXFORD

Structural bioinformatics

# Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology

**Zixuan Cang[1] and Guo-Wei Wei[1,2,3,*]**

[1]Department of Mathematics, [2]Department of Biochemistry and Molecular Biology and [3]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48823, USA

*To whom correspondence should be addressed.
Associate Editor: Burkhard Rost

Received on March 8, 2017; revised on May 8, 2017; editorial decision on June 21, 2017; accepted on July 13, 2017

## Abstract

**Motivation:** Site directed mutagenesis is widely used to understand the structure and function of biomolecules. Computational prediction of mutation impacts on protein stability offers a fast, economical and potentially accurate alternative to laboratory mutagenesis. Most existing methods rely on geometric descriptions, this work introduces a topology based approach to provide an entirely new representation of mutation induced protein stability changes that could not be obtained from conventional techniques.

**Results:** Topology based mutation predictor (T-MP) is introduced to dramatically reduce the geometric complexity and number of degrees of freedom of proteins, while element specific persistent homology is proposed to retain essential biological information. The present approach is found to outperform other existing methods in the predictions of globular protein stability changes upon mutation. A Pearson correlation coefficient of 0.82 with an RMSE of 0.92 kcal/mol is obtained on a test set of 350 mutation samples. For the prediction of membrane protein stability changes upon mutation, the proposed topological approach has a 84% higher Pearson correlation coefficient than the current state-of-the-art empirical methods, achieving a Pearson correlation of 0.57 and an RMSE of 1.09 kcal/mol in a 5-fold cross validation on a set of 223 membrane protein mutation samples.

**Availability and implementation:** http://weilab.math.msu.edu/TML/TML-MP/

**Contact:** wei@math.msu.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Mutagenesis, as a basic biological process that changes the genetic information of organisms, serves as a primary source for many kinds of cancer and heritable diseases, as well as a driving force for natural evolution (Kucukkal *et al.*, 2015; Yue *et al.*, 2005; Zhang *et al.*, 2012). For example, more than 60 human hereditary diseases are directly related to mutagenesis in proteases and their natural inhibitors (Puente *et al.*, 2003). Additionally, mutagenesis often leads to drug resistance (Martinez and Baquero, 2000). Mutation, as a result of mutagenesis, can either occur spontaneously in nature or be caused by the exposure to a large dose of mutagens in living organisms. In laboratories, site directed mutagenesis analysis is a vital experimental procedure for exploring protein functional changes in enzymatic catalysis, structural supporting, ligand binding and signaling (Fersht, 1978). Nonetheless, site directed mutagenesis analysis is both time-consuming and expensive. Additionally, site directed mutagenesis measurements for one specific mutation obtained from different experimental approaches may vary dramatically for membrane protein mutations (Guo *et al.*, 2016). Computational prediction of mutation impacts on protein stability

is an important alternative to experimental mutagenesis analysis for the systematic exploration of protein structural instabilities, functions, disease connections and organism evolution pathways (Guerois *et al.*, 2002). A major advantage of these approaches is that they provide an economical, fast and potentially accurate alternative to site directed mutagenesis experiments. Many state-of-the-art methods have been developed in the past decade, including I-Mutant (Capriotti *et al.*, 2005), PoPMuSiC (Dehouck *et al.*, 2009), knowledge-modified MM/PBSA approach (Getov *et al.*, 2016), Rosetta (high) protocols (Kellogg *et al.*, 2011), FoldX (3.0, beta 6.1) (Guerois *et al.*, 2002), SDM (Worth *et al.*, 2011), DUET (Pires *et al.*, 2014), PPSC (Prediction of Protein Stability, version 1.0) with the 8 (M8) and 47 (M47) feature sets (Yang *et al.*, 2013), PROVEAN (Choi *et al.*, 2014), ELASPIC (Berliner *et al.*, 2014), STRUM (Quan *et al.*, 2016) and EASE-MM (Folkman *et al.*, 2016). In general, computational approaches can be classified into three major classes. Among them, physics based methods typically make use of molecular mechanics (MM) or multiscale implicit solvent models approaches. These approaches might offer physical insights to mutagenesis. Empirical models are another class of methods that utilize empirical functions and potential terms to describe protein mutations. Model parameters are fit with a given set of experimental data and the resulting model is used to predict new mutation induced folding free energy changes. The last class of approaches is knowledge based methods that invoke modern machine learning techniques to uncover hidden relationships between protein stability and protein structure as well as sequence. A major advantage of knowledge based mutation predictors is their ability to handle increasingly large and diverse mutation datasets. However, the performance of these approaches highly depends on the training sets and their results usually cannot be easily interpreted in physical terms.

A common challenge for all existing mutation impact prediction models is in achieving accurate and reliable predictions of membrane protein stability changes upon mutation. As recently noted by Kroncke *et al*, currently there is no reliable method for the prediction of membrane protein mutation impacts on stability (Kroncke *et al.*, 2016). The membrane protein mutation dataset studied by these authors has fewer than 250 data entries, which is too few for most knowledge based methods, and involves 7 membrane protein families, which might be too many for typical physics based methods tuned to work with a specific membrane protein family.

A key feature of all existing structure based mutation impacts on protein stability predictors is that they either fully or partially rely on direct geometric descriptions which rest in excessively high dimensional spaces resulting in large number of degrees of freedom. In practice, the geometry can easily be over simplified. Mathematically, topology, in contrast to geometry, concerns the connectivity of different components in space (Kaczynski *et al.*, 2004), and offers the ultimate level of abstraction of data. However, conventional topology incurs too much reduction of geometric information to be practically useful in biomolecular analysis. Persistent homology, a new branch of algebraic topology, retains partial geometric information in topological description, and thus bridges the gap between geometry and topology (Edelsbrunner *et al.*, 2002; Zomorodian and Carlsson, 2005). It has been applied to biomolecular characterization, identification and analysis via topological fingerprints (Mate *et al.*, 2014; Wang and Wei, 2016; Xia and Wei, 2014, 2015a,b; Xia *et al.*, 2015; Yao *et al.*, 2009). However, conventional persistent homology makes no distinction of different atoms in a biomolecule, which results in a heavy loss of biological information and limits its performance in protein classification (Cang *et al.*, 2015).

In the present work, we introduce element specific persistent homology (ESPH), interactive persistent homology and binned barcode representation to retain essential biological information in the topological simplification of biological complexity. We further integrate ESPH and machine learning to analyze and predict mutation induced protein stability changes. The essential idea of our topological mutation predictor (T-MP) is to use ESPH to transform the biomolecular data in the high-dimensional space with full biological complexity to a space of fewer dimensions and simplified biological complexity, and to use machine learning to deal with massive and diverse datasets. A distinct feature of the present T-MP is that the prediction results can be analyzed and interpreted in physical terms to shed light on the molecular mechanism of protein folding energy changes upon mutation. Additionally, machine learning models might be adaptively optimized according to the performance analysis of ESPH features for different types of mutations. We demonstrate that the performance of proposed T-MP matches or excesses that of other existing methods.

## 2 Materials and methods

### 2.1 Persistent homology characterization of proteins

Unlike physics based models which describe protein structures in terms of covalent bonds, hydrogen bonds, electrostatic and van der Waals interactions, persistent homology, on the other hand characterizes the geometric space of protein structures in terms of a sequence of topological spaces each corresponding to a spatial scale. The topological features at different spatial scales implicitly describe atomic interactions, such as strong or weak hydrogen bonds, and van der Waals interactions. Additionally, topological characterization shows not only pairwise interactions, but also higher order patterns, such as hydrophobic networks. Such properties of persistent homology methods motivate us to develop a topology based representation parallel to the geometric one in physical models.

The natural language of persistent homology is topological invariants, i.e. the intrinsic features of the underlying topological space. More specifically, independent components, rings and cavities are topological invariants in a given dataset and their numbers are called Betti-0, Betti-1 and Betti-2, respectively, as shown in the left column of Figure 1. Loosely speaking, simplicial complexes are generated from discrete data points according to a specific rule such as Vietoris-Rips complex, Cêch complex, or alpha complex. Specifically, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle and a 3-simplex represents a tetrahedron, see the middle column of Figure 1. Algebraic groups built on these simplicial complexes are used in simplicial homology to practically compute Betti numbers of various (topological) dimensions. Furthermore, persistent homology creates a series of homologies through a filtration process, in which the connectivity of a given dataset is systematically reset according to a scale parameter, such as an ever-increasing radius of every atom in a protein, see the right column of Figure 1. As a result, the birth, death and persistence of topological invariants over the filtration give rise to the barcode representation of a given dataset (Ghrist, 2008). When persistent homology is used to analyze three dimensional (3D) protein structures, one-dimensional (1D) persistent homology barcodes are obtained as topological fingerprints (TFs) (Cang *et al.*, 2015; Mate *et al.*, 2014; Xia and Wei, 2014; Yao *et al.*, 2009).

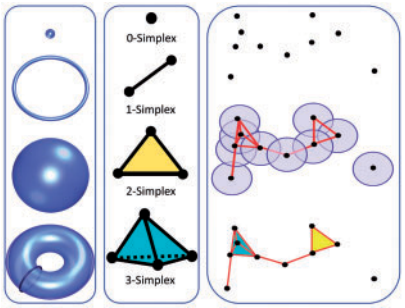


**Fig. 1.** An illustration of topological invariants (Left column), basic simplexes (Middle column) and simplicial complex construction in a given radius of filtration (Right column). Left column: a point, a circle, an empty sphere and a torus are displayed from left to right. Betti-0, Betti-1 and Betti-2 numbers for point are, respectively, 1,0 and 0, for the circle 1, 1, and 0, for the empty sphere 1, 0, and 1, and for the torus 1,2 and 1. Two auxiliary rings are added to the torus explain Betti-1 = 2. Middle column: Four typical simplexes are illustrated. Right column: Illustration of a set of ten points (top chart) at a given filtration radius (middle chart) and the corresponding simplicial complexes (bottom chart), where there are one 0-simplex, three 1-simplexes, one 2-simplex and one 3-simplex

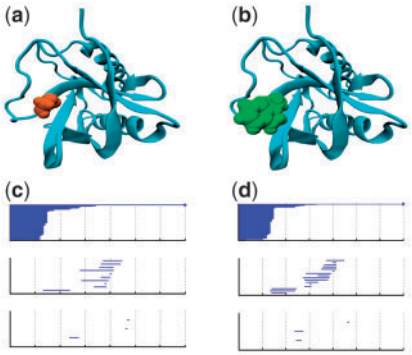


**Fig. 2.** An illustration of persistent homology barcode changes from wild type to mutant proteins. (**a**) The wild type protein (PDB:1ey0) with residue 88 as Gly. (**b**) The mutant with residue 88 as Trp. (**c**) Wild type protein barcodes for heavy atoms within 6 Å of the mutation site. Three panels from top to bottom are Betti-0, Betti-1 and Betti-2 barcodes, respectively. The horizontal axis is the filtration radius (Å). (**d**) Mutant protein barcodes obtained similarly as those for the wild type

As an illustration, we consider the persistent homology analysis of a wild type protein (PDB:1ey0) and its mutant. The mutation (G88W) occurred at residue 88 from Gly to Trp is shown at Figure 2a and b. In this case, a small residue (Gly) is replaced by a large one (Trp). We carry out persistent homology analysis of a set of heavy atoms within 6 Å from the mutation site. Persistent homology barcodes of the wild type and the mutant are respectively given in Figure 2c and d, where the three panels from top to bottom are for Betti-0, Betti-1 and Betti-2, respectively. Since the set of atoms included in the wild type and the mutant is the same except for that in the mutation site, the obvious difference in persistent homology barcodes is induced by the mutation. The increase of residue size results in tighter pattern of Betti-0 bars where there are fewer relatively long bars and more Betti-1 and Betti-2 bars in a shorter distance scale are observed.

Nonetheless, the above topological representation of proteins does not contain sufficient biological information, such as bond length distribution of a given type of atoms, hydrogen bonds, hydrophobic and hydrophilic effects, to offer an accurate model for mutation induced protein stability change predictions. To characterize chemical and biological properties of biomolecules, we introduce *element specific persistent homology* (ESPH). Instead of labeling every atom as in many physics based methods, we distinguish different element types of biomolecules in constructing persistent homology barcodes. For proteins, commonly occurring element types include C, N, O, S and H. Among them, hydrogen atoms are often absent from PDB data and sulfur atoms are too few to be statistically significant in most proteins. Therefore, we focus on the ESPH of C, N and O elements in protein characterization.

## 2.2 Topological descriptors

The most important issue in mutation induced protein stability change analysis is the interactions between the mutation site and the rest of the protein. To describe these interactions, we propose *interactive persistent homology* adopting the distance function $DI(A_i, A_j)$ describing the distance between two atoms $A_i$ and $A_j$ defined as

$$DI(A_i, A_j) = \begin{cases} \infty, & \text{if } \mathrm{Loc}(A_i) = \mathrm{Loc}(A_j), \\ DE(A_i, A_j), & \text{otherwise}, \end{cases} \quad (1)$$

where $DE(\cdot,\cdot)$ is the Euclidean distance between the two atoms and $\mathrm{Loc}(\cdot)$ denotes the location of an atom which is either in a mutation site or in the rest of the protein. In the persistent homology computation, Vietoris-Rips complex (VC) and alpha complex (AC) are used for characterizing first order interactions and higher order patterns respectively. To characterize interactions of different kinds, we construct persistent homology barcodes on the atom sets by selecting one certain type of atoms in mutation site and one other certain type of atoms in the rest of the protein. We denote the set of barcodes from one persistent homology computation as $V^{p,d,b}_{\gamma,\alpha,\beta}$ where $p \in \{\mathrm{VC}, \mathrm{AC}\}$ is the complex used, $d \in \{DI, DE\}$ is the distance function, $b \in \{0, 1, 2\}$ represents topological dimensions, $\alpha \in \{\mathrm{C}, \mathrm{N}, \mathrm{O}\}$ is the element type selected in the rest of the protein, and $\beta \in \{\mathrm{C}, \mathrm{N}, \mathrm{O}\}$ is the element type selected in the mutation site. $\gamma \in \{\mathrm{M}, \mathrm{W}\}$ denotes whether the mutant protein or the wild type protein is used for the calculation. The proposed approach ends up with a total of 54 sets of persistent homology bar codes $\left(V^{\mathrm{VC},DI,0}_{\gamma,\alpha,\beta}\right.$, where $\alpha, \beta = \mathrm{C}, \mathrm{N}, \mathrm{O}; \gamma = \mathrm{M}, \mathrm{W}$ and $V^{\mathrm{AC},DE,b}_{\gamma,\alpha,\beta}$, where $\alpha, \beta = \mathrm{C}, \mathrm{N}, \mathrm{O}$; $\gamma = \mathrm{M}, \mathrm{W}$; $b = 1, 2$). These barcodes are capable of revealing the molecular mechanism of protein stability. For example, interactive ESPH barcodes generated from carbon atoms are associated with hydrophobic interaction networks in proteins. Similarly, interactive ESPH barcodes between nitrogen and oxygen atoms correlate to hydrophilic interactions and/or hydrogen bonds as shown in Figure 3. Interactive ESPH barcodes are also able to reveal other bond information; notwithstanding, they cannot always be interpreted as covalent bond, hydrogen bonds, or van der Waals bonds in general. In fact, interactive ESPH barcodes provide an entirely new representation of molecular interactions.

Features are extracted from the groups of persistent homology barcodes. For the 18 groups of Betti-0 ESPH barcodes, though they cannot be literally interpreted as bond lengths, they can be used to effectively characterize biomolecular interactions. Interatomic distance is a crucial parameter for interaction strength. One can classify hydrogen bonds with donor-acceptor distances of 2.2–2.5 Å as strong and mostly covalent, 2.5-3.2 Å as moderate and mostly electrostatic, and 3.2–4.0 Å as weak and electrostatic (Jeffrey, 1997). Their corresponding energies are about 40-14, 15-4 and less than 4 kcal/mol, respectively (Jeffrey, 1997). To differentiate the interaction distances between various element types, we propose *binned*

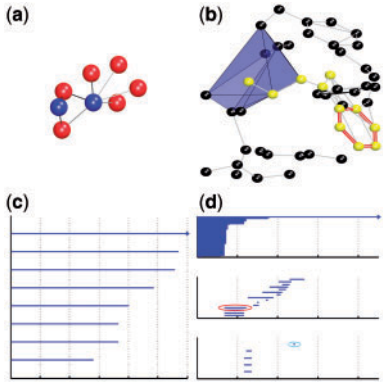


**Fig. 3.** An illustration of element specific persistent homology (ESPH) indicating the hydrophilic network (Left) and hydrophobic network (Right) at a mutation site. **a**: Hydrophilic network showing the connectivity between nitrogen atoms of the mutation residue (blue) and oxygen atoms of the rest of the protein (red). **b**: Hydrophobic network showing the connectivity between carbon atoms of the rest of the protein (black) and of the mutation residue (yellow). The red circle labels a hexagon ring and the blue filling indicates a cavity. **c**: The ESPH Betti-0 barcodes of the hydrophilic network in (a). Betti-0 barcodes show not only the number and strength of hydrogen bonds, but also the hydrophilic environment. Specifically, the shortest four bars can be directly interpreted as conventional hydrogen bonds, while other bars contributing the degree of hydrophilicity at the mutation site. **d**: The ESPH Betti-0, Betti-1 and Betti-2 barcodes of the hydrophobic network in (b). The bar in the red circle is due to the hexagon ring in (b) and the bar in the blue circle is due to the cavity in (b)

*barcode representation* by dividing interactive ESPH barcodes (the Betti-0 barcodes obtained with Rips complex with interactive distance *DI*) into a number of equally spaced bins, namely $[0, 0.5], (0.5, 1], \ldots, (5.5, 6]$Å. The death value of bars is counted in each bin resulting in 12*18 features. Such representation enables us to precisely characterize hydrogen bond, van der Waals, electrostatic, hydrophilic and hydrophobic interactions. For the higher order Betti numbers, the emphasis is given on patterns of both short and long distance scales. Seven features are computed for each group of barcodes for Betti-1 or Betti-2 (the barcodes obtained with alpha complex using the Euclidean distance) which are summation, maximum and average bar length as well as maximum and minimum birth and death values resulting in 7*36 features. To contrast the interactive ESPH barcodes of wild type protein and mutant, we also take the differences between the features described above, which gives rise to a total of 702 features.

## 2.3 Auxiliary descriptors

While the topological descriptors give a thorough examination of the atomic arrangements and connectivities, some other crucial properties are not explicitly characterized. Additionally, due to the diverse quality of the structures examined, some higher level descriptors such as residue level descriptors can enhance the robustness of the model. Therefore, we include some auxiliary descriptors from the aspect of geometry, electrostatics, amino acid types composition and amino acid sequence. The distance from an atom to the mutation site is defined as the shortest distance between this atom to any atom of the mutation site. The distance from a residue to the mutation site is the shortest distance between any pair of atoms containing one atom from this residue and one atom from the mutation site. In the following descriptions of features, an atom or a residue is near the mutation site if they are within a distance of 10 Å from the mutation site.

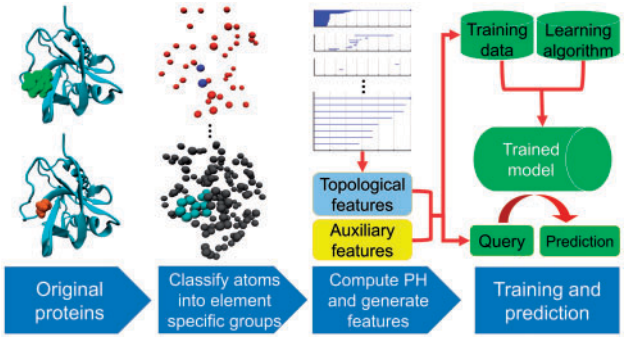


**Fig. 4.** Flowchart for topology based predictions of protein folding stability changes upon mutation

The geometric descriptors contain surface area and van der Waals interaction. The area of solvent excluded surface is assigned to each atom and are summed over various selections of atoms based on their locations and element types. The van der Waals interaction is quantified with Lenard-Johns potential and is computed on each atom by contrasting against all other heavy atoms with a cutoff of 40 Å. The features are generated similarly as surface area.

The electrostatics descriptors are consisted of atomic partial charge, Coulomb interaction and atomic electrostatic solvation energy. Coulomb interactions are computed on each atom by contrasting against all other atoms with a cutoff of 40 Å. The features are generated similarly as surface area with additional summation of absolute values for partial charge and Coulomb interaction.

The high level descriptors include neighborhood amino acid composition and predicted pKa shifts. Amino acid types are divided into 5 non-overlapping groups, i.e. hydrophobic, polar, positively charged, negatively charged and special case groups. The count and percentage of amino acids of each group near the mutation site are collected. The sum, average and variance of surface area, weight, volume and hydropathy scores of the nearby amino acids are also computed. The pKa values for the ionizable amino acids are calculated and characterized for each amino acid type and for each location including the mutation site, C-terminal, N-terminal.

The sequence descriptors describe the secondary structure and residue conservation score collected from Position-specific scoring matrix (PSSM). The secondary structure of the mutation site is categorized from the structure. Since the global minimization of the mutant structures is not performed, the preference score of the mutation site to be in each type of secondary structure is quantified based only on sequence to further describe the favorability of each type of secondary structure at the mutation site for the mutant. The conservation scores derived from PSSM describes the favorability of the mutation from an evolutionary point of view.

Details of auxiliary features can be found in Section 2 of Supplementary Material and a detailed list of all the auxiliary features can be found in Supplementary Table S5.

## 2.4 Gradient boosting trees regressor

The topological features and the auxiliary features are ideally suited for being used as machine learning features to predict protein stability changes upon mutation. Figure 4 shows a schematic illustration of our T-MP. We have examined a number of machine learning algorithms, including decision tree learning, random forest and gradient boosted regression trees (GBRTs) (Friedman, 2001), and found very similar results from these algorithms for the above binned interactive ESPH barcodes. For example, GBRTs are able to integrate

weak learners to form a strong predictor. GBRTs uncover the coupling or nonlinear dependence among highly interactive topological features by choosing an appropriate maximum tree depth. Additionally, GBRTs bypass the normalization of the topological feature vectors, and thus allow mixed attributes of different topological measures and physical units. Finally, GBRTs can effectively avoid overfitting by lowering the learning rate and carrying out subsampling, which is important in dealing with small training datasets, such as datasets for membrane protein mutations.

### 2.5 Datasets and preprocessing

Four globular protein datasets and one membrane protein dataset are used to validate the proposed method. A dataset of 2648 mutation instances in 131 proteins, called *S2648* set collected by (Dehouck *et al.*, 2009) and its subset containing 350 mutation instances involving 67 proteins, named *S350* set are used to compare the performance with other methods. A larger dataset containing 3421 mutation instances in 150 proteins called *Q3421* set collected by (Quan *et al.*, 2016) is used to further validate the method on a more diverse situation. To test the ability of the method on predicting new data, we constructed a testing set containing 293 mutation instances named *T293* set by manually collecting the data with references published on or after year 2009 when S2648 set was constructed to mimic the blind test situation. The detailed list of T293 set can be found in Supplementary Material. A collection of 223 mutation instances in 7 membrane proteins called *M223* set (Kroncke *et al.*, 2016) is used to benchmark the method on membrane protein mutations. Mutation induced protein stability changes are originally obtained from the ProTherm database (Bava *et al.*, 2004). When multiple records for the same mutation are found, a scheme same as Quan *et al.* (2016) and Dehouck *et al.* (2009) is applied. Each dataset is associated to one machine learning task. S2648 set, Q3421 set and M223 set are used for 5-fold cross validation. S350 set is tested with models trained on S2468 set excluding S350 set. T293 set is tested with models trained on S2648 set. There is no overlap between T293 set and S2648 set.

The PDB files of wild type proteins are downloaded from Protein Data Bank (PDB) (Berman *et al.*, 2000). For the mutation instances in S2648 set, multimetric state is considered according to (Dehouck *et al.*, 2009) and instances related to biological assemblies are obtained from either PDB or PISA server (Krissinel and Henrick, 2007). In the productive models, we only consider the chain where mutation is made because that inter-domain interactions may contribute differently to folding energy and this different contribution cannot be sufficiently learned, whilst only the minority of the data involves multimetric states. The protein part is extracted and saved using the VMD software package (Humphrey *et al.*, 1996). The missing heavy atoms and hydrogen atoms are added to the structure using the Profix utility of Jackal software package (Xiang and Honig, 2001). The mutant protein structure is obtained using Scap utility from Jackal software package (Xiang and Honig, 2001) by replacing the side chain of the mutation site with *min* option being set to 4, in which case additional conformers obtained by perturbing the conformers in the rotamer library are explored.

## 3 Software and resources

The machine learning models are built using scikit-learn software package (Pedregosa *et al.*, 2011). Specifically, the gradient boosting trees regressor is used with the parameters: $n\_estimators = 20000$, $max\_depth = 6$, $min\_samples\_split = 3$, $learning\_rate = 0.005$, $subsample = 0.4$, and $max\_features = sqrt$. All results shown in this work if not specified are obtained with this configuration. The pqr files containing the partial charge information of the proteins are obtained from PDB2PQR software package (Dolinsky *et al.*, 2007) using the CHARMM27 force field. The surface area and solvation energy are computed using our in-house online software packages ESES (Liu *et al.*, 2017) and MIBPB (Chen *et al.*, 2011; Yu *et al.*, 2007; Zhou *et al.*, 2006). The pKa values are predicted using PROPKA software package (Li *et al.*, 2005). The position-specific scoring matrices are generated by the BLAST+ software package (Johnson *et al.*, 2008) by searching the nr database. The secondary structure and torsion angle prediction from sequences are obtained using the SPIDER software package (Heffernan *et al.*, 2015). The persistent homology computation with VR complexes are carried out using Javaplex software package (Adams *et al.*, 2014). The persistent homology computation with alpha complexes are done using Dionysus software package (Morozov, 2012) which uses CGAL library (Tran *et al.*, 2017) for alpha shapes. Computational work in support of this research was performed at Michigan State University's High Performance Computing Facility.

## 4 Results

### 4.1 General performance

For various tests shown in this section, 50 repeated runs are conducted separately and the median values of the results are reported to reasonably assess the performance of models with randomness. A comparison of the performances of various methods is summarized in Table 1. Pearson correlations coefficient ($R_P$) and RMSE for test

**Table 1.** Comparison of Pearson correlation coefficients ($R_P$) and RMSEs (kcal/mol) of various methods on the prediction of mutation induced protein stability changes of the S350 set and 5-fold cross validation of mutation induced protein stability changes of the S2648

| Method | S350 | | | S2648 | | |
|---|---|---|---|---|---|---|
| | $n^d$ | $R_P$ | RMSE | $n^d$ | $R_P^e$ | RMSE[f] |
| T-MP-2 | 350 | 0.82 | 0.92 | 2648 | 0.79 | 0.91 |
| STRUM [b] | 350 | 0.79 | 0.98 | 2647 | 0.77 | 0.94 |
| T-MP-1 | 350 | 0.76 | 1.02 | 2648 | 0.75 | 0.98 |
| mCSM[b,c] | 350 | 0.73 | 1.08 | 2643 | 0.69 | 1.07 |
| INPS[b,c] | 350 | 0.68 | 1.25 | 2648 | 0.56 | 1.26 |
| PoPMuSiC 2.0[b] | 350 | 0.67 | 1.16 | 2647 | 0.61 | 1.17 |
| PoPMuSiC 1.0[a] | 350 | 0.62 | 1.23 | – | – | – |
| I-Mutant 3.0[b] | 338 | 0.53 | 1.35 | 2636 | 0.60 | 1.19 |
| Dmutant[a] | 350 | 0.48 | 1.38 | – | – | – |
| Automute[a] | 315 | 0.46 | 1.42 | – | – | – |
| CUPSAT[a] | 346 | 0.37 | 1.46 | – | – | – |
| Eris[a] | 334 | 0.35 | 1.49 | – | – | – |
| I-Mutant 2.0[a] | 346 | 0.29 | 1.50 | – | – | – |

*Note*: T-MP-1 is our topological based mutation predictor that solely utilizes structural information. T-MP-2 is our model that complements T-MP-1 with additional electrostatic, evolutionary and sequence information. The T-MP methods are tested with 50 repeated experiments and the medians are reported.

[a]Data directly obtained from Worth *et al.* (2011).

[b]Data obtained from Quan *et al.* (2016).

[c]The results reported in the publications are listed in the table, however, according to Quan *et al.* (2016), the data from the online server has $R_p$(RMSE) of 0.59(1.28) and 0.70(1.13) for INPS and mCSM respectively in the task of S350 set.

[d]Number of samples successfully processed.

**Table 2.** Pearson correlation coefficients and RMSEs in the unit of kcal/mol of auxiliary features for four tasks in the prediction of mutation impacts on protein stability

| Features | S350 | | S2648 | | Q3421 | | M223 | |
|---|---|---|---|---|---|---|---|---|
| | $R_P$ | RMSE | $R_P$ | RMSE | $R_P$ | RMSE | $R_P$ | RMSE |
| T-MP-2 | 0.817(0.002) | 0.92(0.004) | 0.789(0.005) | 0.91(0.009) | 0.803(0.008) | 1.18(0.020) | 0.575(0.019) | 1.08(0.018) |
| T-MP-1 | 0.765(0.003) | 1.02(0.006) | 0.746(0.006) | 0.98(0.009) | 0.767(0.006) | 1.27(0.014) | 0.543(0.022) | 1.11(0.020) |
| E-MP | 0.760(0.003) | 1.02(0.005) | 0.721(0.005) | 1.02(0.008) | 0.733(0.009) | 1.34(0.018) | 0.525(0.026) | 1.14(0.026) |
| G-MP | 0.759(0.004) | 1.03(0.006) | 0.716(0.004) | 1.03(0.007) | 0.724(0.008) | 1.37(0.015) | 0.474(0.033) | 1.17(0.027) |
| S-MP | 0.609(0.005) | 1.26(0.006) | 0.616(0.006) | 1.16(0.007) | 0.581(0.006) | 1.61(0.008) | 0.379(0.029) | 1.27(0.025) |
| H-MP | 0.686(0.004) | 1.14(0.006) | 0.662(0.009) | 1.11(0.013) | 0.654(0.009) | 1.50(0.016) | 0.231(0.048) | 1.41(0.043) |

*Note*: The medians of 50 repeated runs are reported with the standard deviation across the repeated runs in the parenthesis. Here S350 is a test and its predictions are generated with a model trained with the training set S2648 excluding set S350. Results for S2648 are obtained from 5-fold cross validation. Similarly results for Q3421 and M223 are obtained from 5-fold validation. Here G-MP, E-MP, H-MP and S-MP denote mutation predictors derived respectively from geometric features, electrostatic features, high level features and sequence features described in Section 2.3 and Supplementary Material.

set S350, and 5-fold cross validations for training set S2648, are given for various methods, including ours. The proposed topology based mutation predictor, labeled as T-MP-1, significantly outperforms other existing methods, except for STRUM (Quan *et al.*, 2016). STRUM is constructed by using various descriptors including geometric, evolutionary and sequence information and its $R_P$ and RMSE are 0.79 and 0.98 kcal/mol, respectively for test set S350, and 0.77 and 0.94 kcal/mol, respectively for cross validation of S2648 set (Quan *et al.*, 2016). STRUM's excellent performance motivates us to consider auxiliary features. To this end, we add features generated from geometric, electrostatic and sequence information (see Supplementary Material) to our T-MP-1 to construct T-MP-2. As shown in Table 1, T-MP-2 has the best performance among all methods with $R_P$ and RMSE being 0.82 and 0.92 kcal/mol, respectively for test set S350, and 0.79 and 0.91 kcal/mol, respectively for cross validation of S2648 set. A comparison between T-MP-1 and T-MP-2 indicates that geometric, electrostatic and sequence features give rise to approximately 5% improvement over the original topological prediction, indicating the importance of geometric, electrostatic and sequence information to mutation predictions. However, as shown in Table 2, none of these features has more predictive power than the present topological descriptor.

It is important to understand whether the performance of T-MP-2 was due to overfitting as more features had been used. To this end, we carried out a feature importance analysis as shown in Supplementary Table S6. Based on this analysis, we constructed two new models with respectively 1000 and 800 top features for S2648 training set, excluding S350 test set, under the same parameter setting. We found that for test set S350, Pearson correlations coefficient ($R_P$) and RMSE of results predicted by two new models were unchanged. Similar behavior was found for other datasets used in this work. Therefore, we conclude that the performance of our method was not due to overfitting and the present method is robust with respect to additional non-essential features.

Q3421 set and T293 set are used to further validate the proposed method. In 5-fold cross validation of Q3421 set, $R_P$/RMSE (kcal/mol) of 0.803/1.18 is obtained compared to 0.79/1.2 reported for STRUM method (Quan *et al.*, 2016). For the prediction of T293 set, the model is trained on S2648 set to mimic blind test. T-MP-2 achieves $R_P$/RMSE (kcal/mol)=0.721/1.94. The inferior performance compared to cross validation results is probably due to low overlapping of proteins for the mutation samples in training and testing sets. This observation suggests that, like all machine learning methods, the present model will potentially generate relatively less accurate predictions, when there is no mutation instance of similar proteins in the training set.

**Table 3.** The performance of T-MP-2 model on different datasets and validation methods

| S350 | S2648 | | | Q3421 | M223 |
|---|---|---|---|---|---|
| | M 5-fold | P 5-fold | P LOO | | |
| 0.82/0.92 | 0.79/0.91 | 0.57/1.21 | 0.59/1.19 | 0.80/1.18 | 0.58/1.08 |

*Note*: The results on S350, Q3421 and M223 are the same as those in Table 2. 'M 5-fold', 'P 5-fold' and 'P LOO' are respectively, mutation level 5-fold, protein level 5-fold and protein level leave one out. The first number is PCC and the second number is RMSE in kcal/mol.

Protein level 5-fold cross validation and leave-one-out cross validation on S2648 set are conducted to test the performance of T-MP-2 when testing and training set have no overlapping proteins. In 5-fold cross validation, the 131 proteins are divided into 5 groups. Each time, mutation samples related to one group of proteins are set as testing set with the rest of the samples being the training set. In leave-one-out cross validation, mutation samples related to one protein is set as a testing set each time. The protein level 5-fold and leave-one-out cross validations yield T-MP-2 performance of $R_P$/RMSE (kcal/mol) of 0.57/1.21 and 0.59/1.19. This suggests that, in the future, more data should be used once available, to cover as various as possible proteins for the training set. A comparison among various methods in Supplementary Table S2 shows that T-MP-2 still outperforms other methods in this case. A collection of performance of the present model, T-MP-2 is listed in Table 3. The importance score for each feature and the linear dependency of the target value on each feature are listed in Supplementary Table S6.

## 4.2 Performance in various mutation situations

Figure 5 depicts detailed correlations between experimental mutation impacts on protein stability and T-MP-2 predictions for 25 subsets of 2648 mutations from the cross validation process on the S2648 set. To this end, we adopt a standard classification that categorizes amino acid residues into hydrophobic (HYD), polar (POL), positively charged (POS), negatively charged (NEG) and special case (SPC) types. First, the majority of mutations lead to more unstable structures (i.e. negative free energy changes), as they should be. However, two mutations from POS to HYD and one mutation from POS to POL lead to unusual stabilizing effects. Moreover, the most

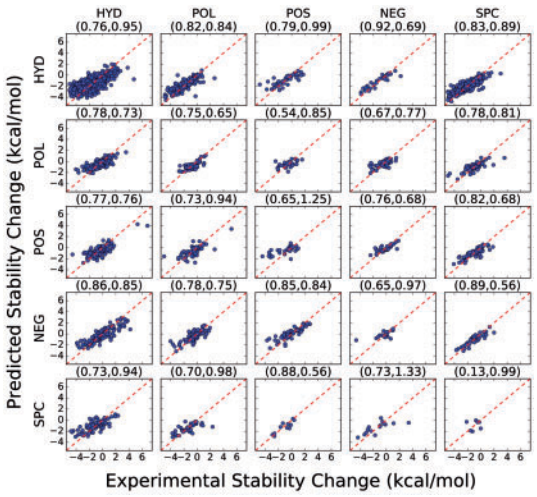


**Fig. 5.** Correlations between experimental stability changes and predicted stability changes (kcal/mol) upon mutation for 25 subsets of the S2648 dataset. All predictions are obtained from 5-fold cross validations. For each subfigure, two numbers in brackets are Pearson correlation coefficients and RMSEs (kcal/mol), respectively. The vertical residue label and horizontal residue label are respectively for the wild type and the mutant such that the second subfigure in the first row denotes a group of mutations from hydrophobic residues to polar residues. The median is taken among 50 repeated experiments. The amino acids are divided into 5 groups, HYD (hydrophobic), POL (polar), POS (positively charged), NEG (negatively charged) and SPC (special case)

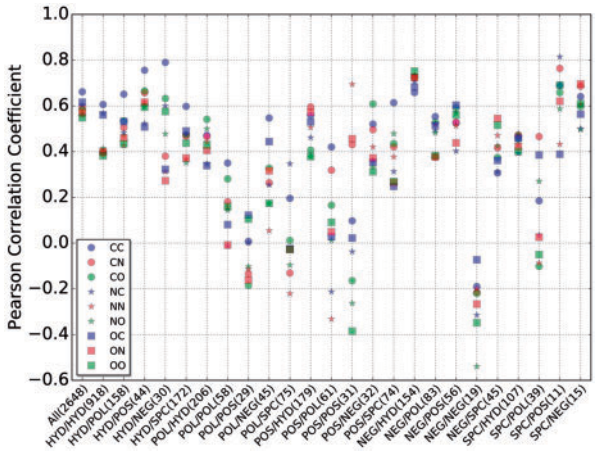


**Fig. 6.** Comparisons of the Pearson correlation coefficients obtained with 9 sets of ESPH features for the S2648 dataset and its 24 subsets. The performance are the medians of 50 repeated runs

accurate prediction in terms of RMSE was for a set of negatively charged residues being mutated to special case ones. Not surprisingly, the worst performance is observed for mutations where little geometric rearrangements happen such as when a negatively charged residue is mutated to another residue of the same type. This performance analysis provides a guidance of how confident the prediction model is in different mutation situations.

### 4.3 Performance of features of various element combinations

To facilitate the discussion of different features, we denote topological features extracted from the atom set containing atoms of element type $\alpha$ in the rest of the protein and atoms of element type $\beta$ in the mutation site by $F_{\alpha\beta}$. Typically, a more important feature has a higher predictive power. Therefore, it is interesting to analyze the predictive powers of individual interactive ESPH features (i.e. $F_{CC}$, $F_{CN}$, $F_{CO}$, $F_{NC}$, $F_{NN}$, $F_{NO}$, $F_{OC}$, $F_{ON}$, $F_{OO}$). In this analysis, we consider 10-fold cross validations for the S2648 set and its subsets due to the small size of some subsets. Random forest regression with 3000 trees is used to reduce computation time. In each analysis, we use only one set of interactive ESPH features, such as hydrophobic effects related feature $F_{CC}$. The left column in Figure 6 depicts our findings based on the whole dataset of 2648 entries. It is found that features associated to carbon atoms in the mutation site, i.e. $F_{\alpha C}$, give rise to some of the best predictions with Pearson correlation coefficients being higher than 0.65 (blue color ones). In fact, $F_{CC}$ gives the best prediction, indicating the key importance of hydrophobic interactions to mutations. Other features have similar performances with Pearson correlation coefficients ranging from 0.55 to 0.58.

We further analyze individual interactive ESPH feature performance with respect to different types of mutations. The same classification of residue types is used as discussed in Section 4.2. We use HYD/POL to represent the situation in which a hydrophobic residue is mutated to a polar residue and similar notations are used for other situations. Our results are also presented in Figure 6. Firstly, for 9 sets of mutation data that involve hydrophobic residues, features that involve carbon atoms in the mutation site (i.e. $F_{\alpha C}$) have a relatively high predictive power. Note that carbon atoms play a major role in hydrophobic interactions and changes in hydrophobic residues can be captured by the changes in Betti-0, Betti-1 and Betti-2 barcodes involving carbon atoms. In fact, other topological features do a good job in predicting hydrophobic residue involved mutations because this set of mutations leads to significant changes in topological invariants. Secondly, all features that involve nitrogen atoms in the mutation site (i.e. $F_{\alpha N}$) have a better predictive power for all positively charged residues (POS/POS). This occurs because three positively charged residues can be distinguished by their numbers of nitrogen atoms, which in turn, can be captured by Betti-0 barcodes (i.e. $V^{VC,DI,0}_{\gamma,\alpha,N}$). Features constructed from oxygen atoms in the mutation site (i.e. $F_{\alpha O}$) have the least prediction power for this dataset. Thirdly, for mutations from one negatively charged residue to another negatively charged residue (i.e. NEG/NEG), features constructed from nitrogen atoms in protein and oxygen atoms in the mutation site (i.e. $F_{NO}$) have the worst predictive power. In fact, none of other topological features does a good job either. This poor performance might be due to negligible mutation induced changes in geometry, topology and structural stability. In this case, small changes in free energies are most likely caused by electrostatic redistribution, which is relatively insensitive to the present topological description. Fourthly, all of the 9 types of features have a similar predictive power for the NEG/HYD dataset. Finally, small data size is hardly a pivotal factor in 10-fold cross validations, although all of the 7 lowest prediction datasets have relatively small data sizes. Note that data sizes in all of the three best predictions (HYD/POS, HYD/NEG, SPC/POS) are fewer than 45 instances.

### 4.4 Performance on membrane proteins

We also examine performance of the proposed topological methods on a challenge problem identified by Kroncke et al. (2016). The proposed method is tested with 5-fold cross validations of a set of 223 mutation instances of membrane proteins in 7 protein families named M223 dataset (Kroncke *et al.*, 2016). A comparison of Pearson correlation coefficients and RMSEs over a number of

**Table 4.** Comparison of Pearson correlation coefficients ($R_P$) and RMSEs (kcal/mol) of various methods for the M223 dataset obtained from 5-fold cross validation in the prediction of mutation impacts on protein stability

| Method | $R_P$ | RMSE | Method | $R_P$ | RMSE |
|---|---|---|---|---|---|
| T-MP-2 | 0.57 | 1.09 | PROVEAN | 0.26 | 4.23 |
| T-MP-1 | 0.54 | 1.12 | Rosetta-MPddG | 0.19 | – |
| Rosetta-MP | 0.31 | – | Rosetta (low)[b] | 0.18 | – |
| Rosetta (High)[a] | 0.28 | – | SDM | 0.09 | 2.40 |
| FoldX | 0.26 | 2.56 | | | |

*Note*: Except for the present results for T-MP-1 and T-MP-2, all other results are adopted from Kroncke *et al.* (2016). The results of Rosetta methods are obtained from Supplementary Figure S1 of Kroncke *et al.* (2016) where RMSE is not given. The results of other methods are obtained from Supplementary Table S1 of Kroncke *et al.* (2016). The results of the machine learning based methods are not listed since those servers are not trained on membrane protein datasets. Among the methods listed, only Rosetta methods have terms describing the membrane protein system. The results reported for T-MP methods are the median values of 50 repeated experiments.

[a]High resolution.

[b]Low resolution.

methods is shown in Table 4. The machine learning based methods are not listed as they are trained on soluble protein datasets. As noted by Kroncke *et al*, there is no reliable method for the prediction of membrane protein mutation impacts on stability at present (Kroncke *et al.*, 2016). Nevertheless, our topology based approaches significantly outperform other existing physical or empirical methods. When auxiliary features are used together with topological features, a 5% improvement in Pearson correlation coefficient is found. Compared with Rosetta-MP, which achieves the best performance with terms designed for membrane proteins (Kroncke *et al.*, 2016), the present T-MP-2 has a 84% higher Pearson correlation coefficient. Nonetheless, Kroncke *et al*'s statement about membrane protein mutation impact on stability predictions still holds as the best Pearson correlation coefficient is only 0.57 and the best RMSE is over 1 kcal/mol. We therefore call for further methodology developments to improve the predictions of membrane protein stability changes upon mutation.

## 5 Conclusion

Contrary to geometry that dominates most biomolecular descriptions, topology is rarely implemented in quantitative analysis of biomolecular science, due to its high level of abstraction and dramatic reduction of biologic information. This article introduces element specific persistent homology to appropriately simplify biomolecular complexity while effectively retain essential biological information in the predictions of mutation impacts on protein stability. Extensive numerical experiments indicate that element specific persistent homology offers some of the most efficient descriptions of protein mutations that cannot be obtained by other conventional techniques.

The advantages of the proposed element specific persistent homology are mainly twofold. Firstly, element specific persistent homology is able to effectively extract unique features, such as loops and cavities associated with set of elements, from complex geometric space. The spatial scale and significance of each identified feature can also be reflected by feature's birth and death over the persistent homology filtration. Secondly, in addition to the faithful characterization of localized pairwise interactions, element specific persistent homology is able to offer a unique description of non-local many-body interactions, such as hydrophobic networks, in terms of high dimensional topological invariants. Such information can be easily picked up by machine learning methods and further boost the performance of the present model.

While persistent homology is good at direct description of the biomolecular structures, some additional assistance may be needed for specific applications. For the mutation problem studied in this work, the electrostatics is characterized with continuum solvent model and Coulomb equation, whereas the amino acid sequences are handled by traditional bioinformatics tools. Therefore, it would be interesting to see the development of persistent homology based tools for the analysis of electrostatics and amino acid sequences in the future work. Having been demonstrated to be a powerful tool for the prediction of mutation impact on protein stability, persistent homology is expected to bring potential improvements when extended to the applications in other complex biological problems.

## Funding

Z.X.C. and G.W.W. were partially supported by NSF grants DMS-1721024 and IIS-1302285, and MSU Center for Mathematical Molecular Biosciences Initiative.

*Conflict of Interest*: none declared.

## References

Adams,H. *et al.* (2014) Javaplex: a research software package for persistent (co)homology. *Int. Congr. Math. Softw.*, 129–136.

Bava,K.A. *et al.* (2004) Protherm, version 4.0: thermodynamic database for proteins and mutants. **32**, D120–D121. *Nucleic Acids Res.*

Berliner,N. *et al.* (2014) Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLoS One*, **9**, e107353.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 35–242.

Cang,Z. *et al.* (2015) A topological approach to protein classification. *Mol. Based Math. Biol.*, **3**, 140–162.

Capriotti,E. *et al.* (2005) I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.

Chen,D. *et al.* (2011) MIBPB: a software package for electrostatic analysis. *J. Comput. Chem.*, **32**, 657–670.

Choi,Y. *et al.* (2014) Predicting the functional effect of amino acid substitutions and indels. *PLoS One*, **7**, e46688.

Dehouck,Y. *et al.* (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: popmusic-2.0. *Struct. Bioinf.*, **25**, 2537–2543.

Dolinsky,T.J. *et al.* (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.*, **35**, W522–W525.

Edelsbrunner,H. *et al.* (2002) Topological persistence and simplification. *Discret. Comput. Geom.*, **28**, 511–533.

Fersht,A.R. (1978) Dissection of the structure and activity of the tyrosyl-trna synthetase by site-directed mutagenesis. *Biochemistry*, **26**, 8031–8037.

Folkman,L. *et al.* (2016) EASEMM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189–1232.

Getov,I. *et al.* (2016) Saafec: predicting the effect of single point mutations on protein folding free energy using a knowledge-modified mm/pbsa approach. *Int. J. Mol. Sci.*, **17**, 512.

Ghrist,R. (2008) Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc.*, **45**, 61–75.

Guerois,R. *et al*. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol*., **320**, 369–387.

Guo,R. *et al*. (2016) Steric trapping reveals a cooperativity network in the intramembrane protease glpg. *Nat. Chem. Biol*., **12**, 353–360.

Heffernan,R. *et al*. (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep*., **5**, 11476.

Humphrey,W. *et al*. (1996) VMD – visual molecular dynamics. *J. Mol. Graph*., **14**, 33–38.

Jeffrey,G.A. (1997) *An Introduction to Hydrogen Bonding*. Oxford University Press, New York.

Johnson,M. *et al*. (2008) Ncbi blast: a better web interface. *Nucleic Acids Res*., **36**, W5–W9.

Kaczynski,T. *et al*. (2004) *Computational Homology. Applied Mathematical Sciences*, vol. **157**. Springer-Verlag, New York.

Kellogg,E.H. *et al*. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Genet*., **79**, 830–838.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol*., **372**, 774–797.

Kroncke,B.M. *et al*. (2016) Documentation of an imperative to improve methods for predicting membrane protein stability. *Biochemistry*, **55**, 5002–5009.

Kucukkal,T.G. *et al*. (2015) Structural and physico-chemical effects of disease and non-disease nssnps on proteins. *Curr. Opin. Struct. Biol*., **32**, 18–24.

Li,H. *et al*. (2005) Very fast empirical prediction and rationalization of protein pka values. *Proteins*, **61**, 704–721.

Liu,B. *et al*. (2017) ESES: software for Eulerian solvent excluded surface. *J. Comput. Chem*., **38**, 446–466.

Martinez,J.L. and Baquero,F. (2000) Mutation frequencies and antibiotic resistance. *Antimicrob. Agents Chemother*., **44**, 1771–1777.

Mate,G. *et al*. (2014) A topological similarity measure for proteins. *Biochim. Biophys. Acta BBA Biomembranes*, **1838**, 1180–1190.

Morozov,D. (2012) Dionysus library for computing persistent homology, vol. 2. Software available at http://www.mrzv.org/software/dionysus

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*., **12**, 2825–2830.

Pires,D.E.V. *et al*. (2014) DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*., **42**, W314–W319.

Puente,X.S. *et al*. (2003) Human and mouse proteases: a comparative genomic approach. *Nat. Rev. Genet*., **4**, 544–558.

Quan,L. *et al*. (2016) Strum: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, **32**, 2936–2946.

Tran,K.F.D. *et al*. (2017) 3D Alpha Shapes. In: *CGAL User and Reference Manual*. CGAL Editorial Board 4.10 edition.

Wang,B. and Wei,G.W. (2016) Object-oriented persistent homology. *J. Comput. Phys*., **305**, 276–299.

Worth,C.L. *et al*. (2011) SDM-a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res*., **39**, W215–W222.

Xia,K.L. and Wei,G.W. (2014) Persistent homology analysis of protein structure, flexibility and folding. *Int. J. Numer. Methods Biomed. Eng*., **30**, 814–844.

Xia,K.L. and Wei,G.W. (2015a) Persistent topology for cryo-EM data analysis. *Int. J. Numer. Methods Biomed. Eng*., **31**, e02719.

Xia,K.L. and Wei,G.W. (2015b) Multidimensional persistence in biomolecular data. *J. Comput. Chem*., **36**, 1502–1520.

Xia,K.L. *et al*. (2015) Multiresolution persistent homology for excessively large biomolecular datasets. *J. Chem. Phys*., **143**, 134103.

Xiang,Z. and Honig,B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*., **311**, 421–430. 405.

Yang,Y. *et al*. (2013) Structure-based prediction of the effects of a missense variant on protein stability. *Amino Acids*, **44**, 847–855.

Yao,Y. *et al*. (2009) Topological methods for exploring low-density states in biomolecular folding pathways. *J. Chem. Phys*., **130**, 144115.

Yu,S.N. *et al*. (2007) Treatment of geometric singularities in implicit solvent models. *J. Chem. Phys*., **126**, 244108.

Yue,P. *et al*. (2005) Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol*., **353**, 459–473.

Zhang,Z. *et al*. (2012) Analyzing effects of naturally occurring missense mutations. *Comput. Math. Methods Med*., **2012**, 805827.

Zhou,Y.C. *et al*. (2006) High order matched interface and boundary method for elliptic equations with discontinuous coefficients and singular sources. *J. Comput. Phys*., **213**, 1–30.

Zomorodian,A. and Carlsson,G. (2005) Computing persistent homology. *Discrete Comput. Geom*., **33**, 249–274.