Computers in Biology and Medicine 169 (2024) 107918

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Integration of persistent Laplacian and pre-trained transformer for protein solubility changes upon mutation

JunJie Wee [a], Jiahui Chen [b], Kelin Xia [c], Guo-Wei Wei [a,d,e,*]

[a] Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA
[b] Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA
[c] Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore
[d] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA
[e] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA

ARTICLE INFO

Dataset link: https://github.com/ExpectozJJ/TopLapGBT, https://github.com/hozumiyu/RSI

Keywords:
Mutation
Protein solubility
Persistent Laplacian
Transformer
AlphaFold2

ABSTRACT

Protein mutations can significantly influence protein solubility, which results in altered protein functions and leads to various diseases. Despite tremendous effort, machine learning prediction of protein solubility changes upon mutation remains a challenging task as indicated by the poor scores of normalized Correct Prediction Ratio (CPR). Part of the challenge stems from the fact that there is no three-dimensional (3D) structures for the wild-type and mutant proteins. This work integrates persistent Laplacians and pre-trained Transformer for the task. The Transformer, pretrained with hundreds of millions of protein sequences, embeds wild-type and mutant sequences, while persistent Laplacians track the topological invariant change and homotopic shape evolution induced by mutations in 3D protein structures, which are rendered from AlphaFold2. The resulting machine learning model was trained on an extensive data set labeled with three solubility types. Our model outperforms all existing predictive methods and improves the state-of-the-art up to 15%.

1. Introduction

Genetic mutations alter the genome sequence, leading to changes in the corresponding amino acid sequence of a protein. These alterations have far-reaching implications on the protein's structure, function, and stability, affecting attributes such as folding stability, binding affinity, and solubility. The consequences of protein mutations have been extensively studied in diverse fields such as evolutionary biology, cancer biology, immunology, directed evolution, and protein engineering [1]. Understanding the impact of genetic mutations on protein solubility is crucial in various fields, including protein engineering, drug discovery, and biotechnology. Accurately analyzing and predicting the impact of mutations on protein solubility is therefore crucial in many fields, facilitating the engineering of proteins with desirable functions. There are numerous intricately interconnected factors impacting protein solubility, ranging from amino acid sequence arrangement, post-translational modifications, protein–protein interactions, to environmental conditions, such as solvent type, ion type and concentration, the presence of small molecules, temperature, etc. Unfortunately, the existing data set does not contain sufficient information. This complexity poses significant challenges for the accurate prediction and modeling of protein solubility, often requiring multifaceted computational approaches for reliable outcomes.

Computational predictions serve as a valuable complement to experimental mutagenesis analysis of protein stability changes upon mutation. Such computational approaches offer several advantages, including being economical, efficient, and provide a viable alternative to labor-intensive site-directed mutagenesis experiments [2]. As a result, the development of accurate and reliable computational techniques for mutational impact prediction could substantially enhance the throughput and accessibility of research in protein engineering and drug discovery.

Over the years, a variety of computational methods have been developed to explore the effects of mutations on protein solubility, including but not limited to CamSol [3], OptSolMut [4], PON-Sol [5], SODA [6], Solubis [7], and others as summarized in a recent review [8]. CamSol employs an algorithm to construct a residue-specific solubility profile, although no explicit method has been made publicly available. OptSolMut is trained on 137 samples, each featuring single or multiple mutations affecting solubility or aggregation. PON-Sol utilizes a random forest model trained on a dataset of 406 single amino acid substitutions labeled as solubility-increasing, solubility-decreasing, or exhibiting no change in solubility. SODA, which is based on the PON-Sol data, specifically targets samples with decreasing solubility [6].

* Corresponding author at: Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA.
E-mail addresses: xiakelin@ntu.edu.sg (K. Xia), weig@msu.edu (G.-W. Wei).

https://doi.org/10.1016/j.compbiomed.2024.107918
Received 28 October 2023; Received in revised form 21 December 2023; Accepted 1 January 2024
Available online 3 January 2024

Solubis is an optimization tool that increases protein solubility and integrates interaction analysis from FoldX [2], aggregation prediction from TANGO [9], and structural analysis from YASARA [10]. Recently, PON-Sol2 [11] extended the original PON-Sol dataset and employed a gradient boosting algorithm for sequence-based predictions. Despite intensive effort, the current prediction accuracy in terms of normalized Correct Prediction Ratio (CPR) remains very low, calling for innovative strategies.

Topological data analysis (TDA) is a relatively new approach for data science. Its main technique is persistent homology [12,13]. The essential idea of persistent homology is to construct a multiscale analysis of data in terms of topological invariants. The resulting changes of topological invariants over scales can be used to characterize the intricate structures of data, leading to an unusually powerful approach in describing protein structure, flexibility, and folding [14]. Persistent homology was integrated with machine learning for the classification of proteins in 2015 [15], which was one the first integrations of TDA and machine learning, and the predictions of mutation-induced protein stability changes [16,17] and protein–protein binding free energy changes [18,19]. One of the major achievements of TDA is its winning of D3R Grand Challenges, an annual worldwide competition series in computer-aided drug design [20,21]. A nearly comprehensive summary of the early success of TDA in biological science was given in a review [22].

However, persistent homology only tracks the changes in topological invariants and cannot capture homotopic shape evolution of data over scales or induced by mutations. To overcome this limitation, Wei and coworkers introduced persistent combinatorial Laplacians, also called persistent spectral graphs, on point clouds [23] and evolutionary de Rham–Hodge method on manifolds [24] in 2019. The essence of these methods is the persistent topological Laplacians (PTLs) either on point clouds or on manifolds. PTLs not only fully capture the topological invariants in its harmonic spectra as those given by persistent homology, but also capture the homotopic shape evolution of data during the multiscale analysis or a mutation process. PTLs were applied to the predictions of protein flexibility [23] and protein–ligand binding free energies [25], protein–protein interactions [26,27], and protein engineering [1]. The most remarkable accomplishment by persistent Laplacian is its accurate forecasting of emerging dominant SARS-CoV-2 variants BA.4 and BA.5 about two months in advance [28].

However, TDA approaches depend on the biomolecular structures, which may not be available. In fact, many proteins involved in the present study do not have 3D structures. In recent years, advanced natural language processing (NLP) models, including Transformers and long short-term memory (LSTM), have been widely implemented across various domains to extract protein sequence information. For example, Tasks Assessing Protein Embedding (TAPE) introduced three different architectures, namely transformer, dilated residual network (ResNet), and LSTM [29]. Additionally, LSTM-based models like Bepler [30] and UniRep [31] have been developed. Additionally, large-scale protein transformer models trained on extensive datasets comprising hundreds of millions of sequences have made significant advancements in the field. These models, including Evolutionary Scale Modeling (ESM) [32] and ProtTrans [33,34], have exhibited exceptional performance in capturing a variety of protein properties. ESM, for instance, allows for fine-tuning based on either downstream task data or local multiple sequence alignments [35]. More recently, a Transformer-based ensemble framework was developed to predict protein–protein interaction sites by integrating transformers and gated convolutional neural networks [36]. In the present work, we leverage the pre-trained ESM transformer model to extract crucial information from protein sequences.

In this work, we will integrate transformer-based sequence embedding and persistent topological Laplacians to predict protein solubility changes upon mutation. While sequence-based models can be applied without 3D structural information, the PTL-based features require high-quality structures. We generate 3D structures of wild type proteins from AlphaFold2 [37] to facilitate topological embedding. By combining both embedding approaches, they naturally complement each other in classifying protein solubility changes upon mutation. These embeddings are fed into an ensemble classifier, gradient boosted trees (GBT), to build a machine learning model, called TopLapGBT. We validate TopLapGBT on the classification of protein solubility changes upon mutation. We demonstrate that this integrated machine learning model gives rise to a substantial improvement as compared to existing state-of-the-art models. Residue-Similarity plots are also applied to assess how well the TopLapGBT model classify three solubility labels.

2. Results

2.1. Overview of TopLapGBT

TopLapGBT integrates both structure-based and sequence-based features, derived from protein structures and sequences respectively, into a unified model. Our architecture comprises three distinct embedding modules: persistent Laplacian-based embeddings, sequence-based embeddings, and auxiliary feature embeddings, all of which feed into an ensemble classifier as depicted in Fig. 1.

In the persistent Laplacian-based feature embedding module, we employ persistent Laplacian techniques to generate features that encapsulate the structural attributes of proteins both pre- and post-mutation. This approach is particularly effective in capturing the structural alterations induced by mutations within the localized neighborhoods of the mutation sites. Mathematically, the persistent Laplacian builds a sequence of simplicial complexes through a filtration process, thereby characterizing atom–atom interactions across multiple scales (details in the Methods section). In the sequence-based feature embedding module, a pre-trained transformer model generates latent feature vectors extracted from protein sequences. Specifically, the transformer model used here is a 650M-parameter protein language model, trained on a corpus of 250M protein sequences spanning multiple organisms [39]. Finally, the auxiliary feature embedding module incorporates a variety of attributes such as surface area, partial charge, pK_a shifts, solvation free energy, and secondary structural information, synthesized from both protein sequences and structures. These three distinct sets of feature embeddings are subsequently concatenated to produce a comprehensive feature vector. This vector is then fed into a gradient-boosting tree classifier to categorize the mutation-induced samples.

2.2. Performance of TopLapGBT on PON-Sol2 dataset

In our study, we utilize the dataset employed by PON-Sol2 as detailed in [11]. The dataset is comprised of 6328 mutation samples, originating from 77 distinct proteins. These samples are categorized into three labels: decrease in solubility, increase in solubility, and no change in solubility. Specifically, the dataset contains 3136 samples demonstrating a decrease in solubility, 1026 samples showing an increase, and 2166 samples with no change. Notably, the dataset exhibits a class imbalance, with a ratio of $1 : 0.69 : 0.34$, indicating a bias towards samples that exhibit a decrease in solubility. To assess the performance of our model, we initially carry out a random 10-fold cross-validation on the dataset. Subsequently, an independent blind test prediction is executed to provide further validation of the model's efficacy.

In Table 1, we present a comparative analysis of the performance of both single layer and double layer classifiers of PON-Sol2 [11] against our proposed model, TopLapGBT, using 10-fold cross-validation. It should be noted that PON-Sol2 incorporates feature selection techniques such as recursive feature elimination (RFE). To provide a robust

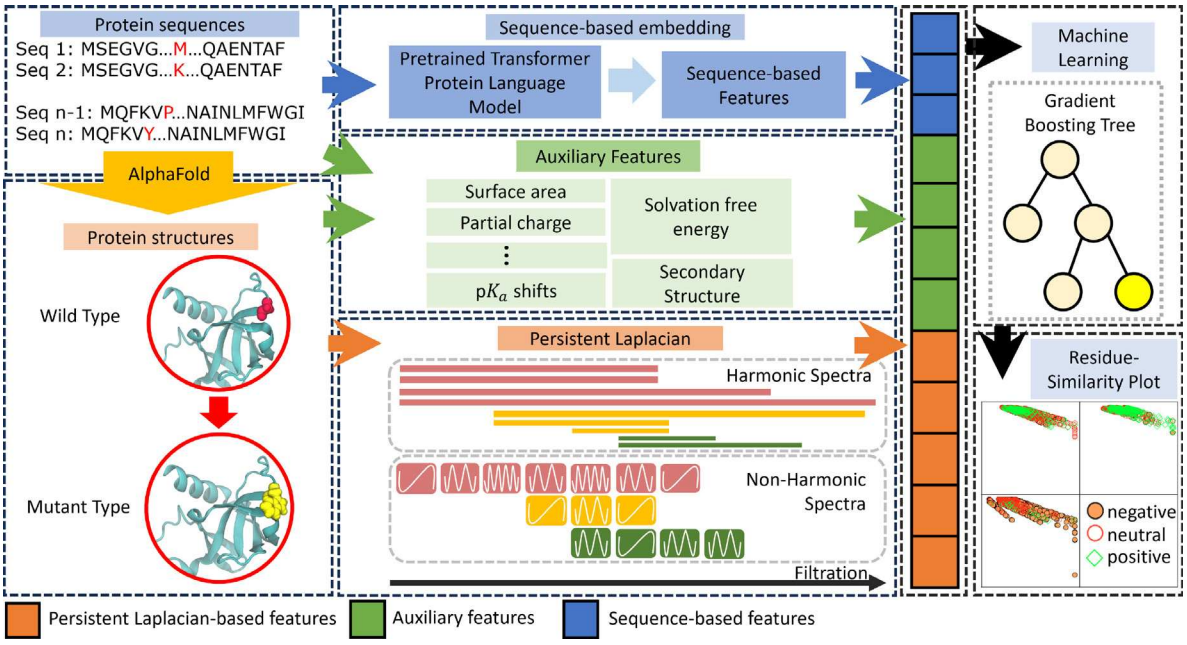

Fig. 1. The illustration of the workflow for TopLapGBT. Protein sequences are first preprocessed by AlphaFold 2 to generate wild type protein structures. Mutant proteins are generated from the Jackal software [38]. The structure-based features from persistent Laplacian, auxiliary and sequence-based features are then concatenated to form a long feature input for gradient boosting tree to classify the protein solubility changes upon mutation. The predicted labels are also analyzed on Residue-Similarity (R-S) plots.

Table 1
Comparison of performance metrics between TopLapGBT and both single layer and double layer classifiers of PON-Sol2 in the 10-fold cross validation. The negative solubility samples are denoted as "–" whereas the positive solubility change samples are denoted as "+". The samples with no solubility change are denoted as "N". Performance metrics include the positive predicted values (PPV), negative predicted values (NPV), sensitivity, specificity, correct prediction ratio (CPR) and generalized correlation (GC^2). PPV refers to the proportions of positive predictions for each solubility class while NPV refers to the proportions of negative predictions for each solubility class. CPR calculates the percentage of correctly classified samples while GC^2 measures the correlation coefficient of the classification. All normalized metrics are also reported. For each metric, the first value is without normalization while the second one is with normalization.

Performance metric		Model					
		PON-Sol2 [11]				TopGBT	TopLapGBT
		Single three-class classifier		Two-layer three-class classifier		–	–
		All Features	30 features selected by RFE	All features	34 features selected by RFE	–	–
PPV	–	0.842/0.742	0.835/0.729	0.875/0.793	0.869/0.781	0.868/0.785	0.873/0.797
	N	0.657/0.536	0.658/0.543	0.635/0.521	0.647/0.534	0.686/0.554	0.681/0.557
	+	0.563/0.730	0.586/0.752	0.520/0.696	0.538/0.714	0.646/0.797	0.627/0.779
NPV	–	0.913/0.954	0.901/0.947	0.893/0.942	0.891/0.941	0.932/0.965	0.931/0.964
	N	0.841/0.824	0.847/0.832	0.847/0.829	0.855/0.838	0.864/0.849	0.858/0.842
	+	0.877/0.737	0.877/0.738	0.877/0.736	0.878/0.739	0.886/0.749	0.888/0.757
Sensitivity	–	0.919/0.919	0.906/0.906	0.892/0.892	0.891/0.891	0.937/0.937	0.934/0.934
	N	0.701/0.701	0.717/0.717	0.724/0.724	0.738/0.738	0.752/0.752	0.735/0.735
	+	0.329/0.329	0.326/0.326	0.336/0.336	0.340/0.340	0.359/0.359	0.395/0.395
Specificity	–	0.831/0.839	0.825/0.831	0.875/0.883	0.868/0.874	0.860/0.872	0.867/0.881
	N	0.812/0.697	0.807/0.697	0.785/0.667	0.792/0.678	0.821/0.697	0.823/0.707
	+	0.948/0.938	0.954/0.947	0.938/0.927	0.941/0.932	0.962/0.954	0.953/0.944
CPR		0.747/0.650	0.746/0.650	0.743/0.651	0.747/0.656	0.780/0.682	**0.792/0.688**
GC^2		0.317/0.298	0.309/0.289	0.322/0.313	0.323/0.312	0.371/0.354	**0.376/0.361**

assessment of TopLapGBT's performance, we conduct 10 repeated runs, and the mean values of these runs are reported to account for any randomness in the model's output.

Performance evaluation of our model, TopLapGBT, is conducted using a range of metrics, including Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, Specificity, Correct Prediction Ratio (CPR), and Generalized Squared Correlation (GC^2). PPV and NPV quantify the proportions of correct positive and negative predictions for each solubility class, respectively. Given that we are dealing with a K-class problem with three distinct solubility classes, CPR and GC^2 are particularly relevant for providing a holistic view of the model's performance [40]. Specifically, CPR measures the overall accuracy of the model, while GC^2 quantifies the correlation coefficient of the classification, ranging from 0 to 1. Larger values for these metrics denote better performance. Importantly, due to the class imbalance in the number of mutation samples across the categories, all performance metrics are normalized to ensure a robust and reliable evaluation of the model's efficacy (further details are elaborated in the Methods section).

The proposed model, TopLapGBT, demonstrates significant performance gains over existing featurization methods in PON-Sol2 across all evaluation metrics [11]. Specifically, normalized CPR and GC^2 scores of TopLapGBT stand at 0.688 and 0.361, marking improvements of 4.88% and 15.71% over PON-Sol2, respectively. These gains underscore the

Table 2
Performance of TopLapGBT with existing state-of-the-art models on the independent blind test classification. The negative solubility samples are denoted as "-" whereas the positive solubility change samples are denoted as "+". The samples with no solubility change are denoted as "N". Performance metrics include the positive predicted values (PPV), negative predicted values (NPV), sensitivity, specificity, correct prediction ratio (CPR) and generalized correlation (GC^2). PPV refers to the proportions of positive predictions for each solubility class while NPV refers to the proportions of negative predictions for each solubility class. CPR calculates the percentage of correctly classified samples while GC^2 measures the correlation coefficient of the classification. All normalized metrics are also reported. For each metric, the first value is without normalization while the second one is with normalization.

Performance metric		Independent Test							
		PON-Sol [5]	SODA	SODA(5 as Threshold)	SODA(10 as Threshold)	SODA(17 as Threshold)	PON-Sol2 [11]	TopGBT	TopLapGBT
PPV	–	0.593/0.428	0.427/0.258	0.606/0.428	0.673/0.468	0.742/0.585	0.804/0.643	0.781/0.649	0.789/0.645
	N	0.427/0.385	NaN/NaN	0.425/0.365	0.397/0.357	0.383/0.350	0.600/0.475	0.617/0.462	0.624/0.475
	+	0.151/0.373	0.080/0.229	0.047/0.149	0.060/0.184	0.098/0.284	0.233/0.472	0.524/0.761	0.476/0.718
NPV	–	0.514/0.691	0.373/0.537	0.508/0.684	0.502/0.677	0.501/0.677	0.794/0.887	0.843/0.920	0.842/0.918
	N	0.685/0.700	0.642/0.667	0.761/0.739	0.797/0.782	0.797/0.782	0.804/0.793	0.816/0.795	0.826/0.809
	+	0.881/0.693	0.832/0.605	0.848/0.633	0.858/0.649	0.858/0.649	0.879/0.684	0.881/0.692	0.880/0.688
Sensitivity	–	0.263/0.263	0.488/0.488	0.195/0.195	0.098/0.098	0.068/0.068	0.802/0.802	0.867/0.867	0.864/0.864
	N	0.456/0.456	0.000/0.000	0.759/0.759	0.886/0.886	0.954/0.954	0.671/0.671	0.692/0.692	0.713/0.713
	+	0.448/0.448	0.253/0.253	0.069/0.069	0.057/0.057	0.046/0.046	0.161/0.161	0.126/0.126	0.115/0.115
Specificity	–	0.812/0.824	0.318/0.297	0.867/0.869	0.951/0.944	0.975/0.976	0.796/0.777	0.747/0.765	0.759/0.763
	N	0.659/0.636	1.000/1.000	0.426/0.340	0.249/0.204	0.144/0.116	0.751/0.630	0.760/0.597	0.760/0.606
	+	0.617/0.623	0.558/0.573	0.786/0.802	0.863/0.872	0.936/0.942	0.920/0.910	0.983/0.980	0.981/0.977
CPR		0.356/0.389	0.282/0.247	0.381/0.341	0.375/0.347	0.382/0.356	0.671/0.545	0.707/0.562	**0.711/0.564**
GC^2		0.010/0.011	NaN/NaN	0.041/0.045	0.022/0.022	0.016/0.016	0.181/0.157	0.205/0.184	**0.206/0.185**

merit of incorporating both structure-based and sequence-based features into the model. To elucidate the contribution of Persistent Laplacian (PL)-based features, we also present a comparative analysis with our TopGBT model in Table 1. The TopGBT model utilizes persistent homology-based embeddings alongside auxiliary and pre-trained transformer features. While TopGBT still outperforms all existing PON-Sol2 models, the incorporation of PL-based features in TopLapGBT leads to an incremental improvement of 1% and 2% in CPR and GC^2 metrics, respectively. This validates our approach of leveraging Persistent Laplacian to comprehensively capture both the topological and homotopic nuances in the evolution of protein structures.

2.3. Performance of TopLapGBT on independent test set

To robustly assess the performance of TopLapGBT, we subjected it to an independent test using the same dataset employed by PON-Sol2 [11]. In this validation, TopLapGBT consistently outperformed all five existing models, as evidenced in Table 2. Specifically, TopLapGBT registers a normalized CPR of 0.564 and a normalized GC^2 of 0.185, surpassing PON-Sol2 by 3.49% and 17.83%, respectively. Relative to TopGBT, the inclusion of PL-based features in TopLapGBT yielded incremental gains in both CPR and GC^2 metrics, thereby further substantiating the utility of Persistent Laplacian in capturing the homotopic shape evolution within protein structures.

3. Discussion

The performance of machine learning models generally relies on the nature of the input features. In our model, the PL-based features depend on one main element which is the quality of the protein structures from AlphaFold 2 (AF2). The quality of AF2 structures are crucial in determining the performance of TopLapGBT. Recently, AF2 structures have been reported to achieve comparable performance to nuclear magnetic resonance (NMR) structures while ensemble methods can be used to enhance the performance by combining multiple NMR structures [1]. This allows AF2 structures to serve as a practical substitute for experimental structural data. Although AF2 structures are not as reliable as X-ray structures, the fusion of sequence-based pre-trained transformer features and PL-based features provides robust featurization even for low quality AF2 structural data. PL elucidates the precise mutation geometry and topology, while sequence-based pre-trained transformer features capture evolutionary patterns from an extensive sequence library. This synergy holds significance and can be applied to a diverse range of other challenges in the field of biomolecular research. For the rest of this section, we analyze the model's performance based on the region of the mutations and the type of mutations. We also discuss the performance of different feature types using the Residue-Similarity plots.

3.1. Performance analysis based on different mutation regions

To delve deeper into the model's performance, we categorize mutation samples based on their structural regions: interior and surface, as depicted in Fig. 2 pre- and post-mutations. These regions are defined by their relative accessible solvent area (rASA), using a cutoff value c. A residue at the mutation site is classified as buried or interior if its rASA falls below this cutoff. While the discrete nature of c initially raised concerns, given that amino acids have a continuous exposure profile, empirical analyses on databases from organisms like Escherichia coli, Saccharomyces cerevisiae, and Homo sapiens have shown that an optimal rASA cutoff of approximately 25% is effective for distinguishing between surface and interior residues [41]. In our analysis, we apply this framework to identify surface and interior residues in the solubility dataset. We observe that some mutation sites undergo a regional transition, moving from one structural domain to another, consequent to the mutation.

To gain nuanced insights into TopLapGBT's performance, we segment the results according to the mutation's structural location within the protein. We present these segmentations as heatmap plots that delineate both mutation regions and amino acid types. Structural regions are defined based on relative accessible surface area (rASA) [41]. By categorizing residues as either interior or surface, we can examine the influence of continuous amino acid exposure on solubility change classification post-mutation. Fig. 2(b) displays accuracy scores for four types of mutations: interior–interior, interior–surface, surface–interior, and surface–surface. TopLapGBT attains an average accuracy score of 0.770 across these categories. Extended data in Figure S1 further breaks down accuracy scores for all 20 distinct amino acids within each region-pair, revealing variations in residue-residue pair performance.

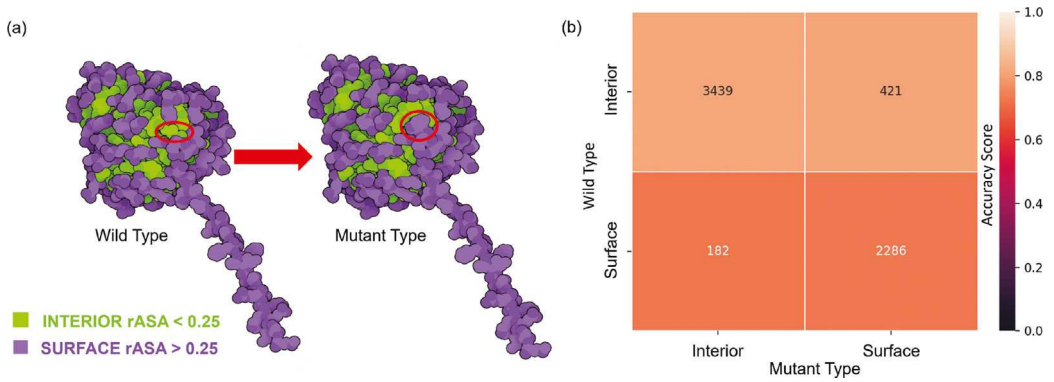

Fig. 2. (a) The definitions of the structural regions on the protein label 213133708 with mutation ID: I283 W. For both wild type and mutant type, amino acids in the proteins are classified under surface or interior regions based on the rASA of the residue. The residue ID 283 of protein label 213133708 was mutated from isoleucine (interior region) to tryptophan (surface region). Structures are plotted with the software Illustrate [42]. (b) A comparison of performance of TopLapGBT among different mutation region types. The *y*-axis represents the region type for the original residue and the *x*-axis represents the region type for the mutated residue. The numbers indicated in each cell corresponds to the number of mutation samples in each region–region mutation pair. The accuracy scores (CPR) for both interior–interior and interior–surface are 0.813 and 0.812 while the accuracy score for both surface–interior and surface–surface are 0.725 and 0.730.

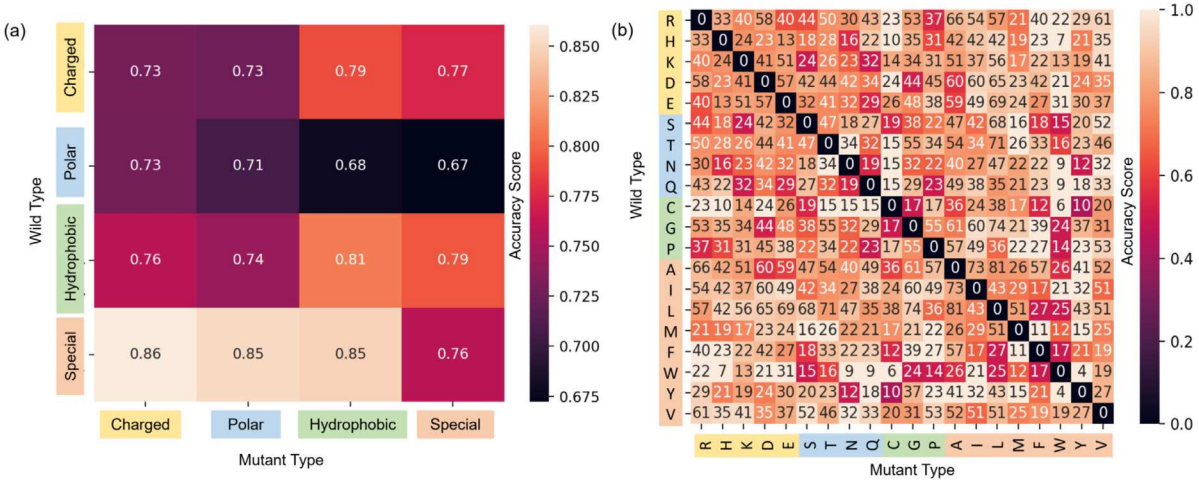

Fig. 3. A comparison of 10-fold cross validation accuracy scores (CPR) for (a) different mutation groups and (b) its associated amino acid types. The *y*-axis labels the residue type of the original protein, whereas the *x*-axis labels the residue type of the mutant. The squares colored in black in (b) have zero mutation samples. For a reverse mutation, the labels are taken with reverse solubility change unless the change is zero.

3.2. Performance analysis based on different mutation types

Switching focus to mutation types, our model's capability in classifying solubility changes also merits exploration across the 20 distinct amino acid types in the dataset. In addition to this, we subgroup amino acids as charged, polar, hydrophobic, or special case. Table S1 enumerates the sample counts for each mutation group pair. Fig. 3(a) displays accuracy scores for each mutation group pair, while Fig. 3(b) shows scores for each amino acid pair. Notably, the special-charged and special-polar groups register the highest accuracy, whereas the polar-hydrophobic and polar-special groups underperform. One plausible reason could be the inherent complexity in accurately classifying mutations with non-negative solubility changes. It is worth noting that PON-Sol2 employed a two-layer classifier to improve classification [11]. Our results indicate that TopLapGBT surpasses the performance of this two-layer system.

3.3. Feature analysis based on residue-similarity plots

The Residue Similarity Index (RSI) serves as a potent metric for evaluating the efficacy of dimensionality reduction in both clustering and classification contexts [43]. RSI has proven its value in generating classification accuracy scores that align well with supervised methods in single-cell typing. When applied to our solubility change dataset, Residue-Similarity (R-S) plots can be constructed to scrutinize how the Residue Index (RI) and Similarity Index (SI) may indicate the quality of cluster separation.

Fig. 4 juxtaposes the R-S plots derived from TopLapGBT against those from various feature sets utilized in model training. Across all visualizations, samples manifest a range of classification outcomes – both correct and incorrect – for each true label. However, a noteworthy observation is that Transformer-pretrain and persistent Laplacian-based features demonstrate superior clustering attributes compared to auxiliary features. The high RI and SI scores for auxiliary features cause these data points to cluster near the upper regions of their respective sections. Despite this, the integrative use of all three feature types in TopLapGBT results in appreciable clustering performance, corroborated by the CPR metrics obtained in 10-fold cross-validation. To solidify the rationale behind adopting robust supervised classifiers like TopLapGBT, we contrast the R-S plots with UMAP visualizations (shown in Figure S2). It becomes evident that UMAP plots fail to form clusters that are as distinct as those observed in R-S plots, thereby reinforcing the need for a specialized approach to classify mutation samples effectively.

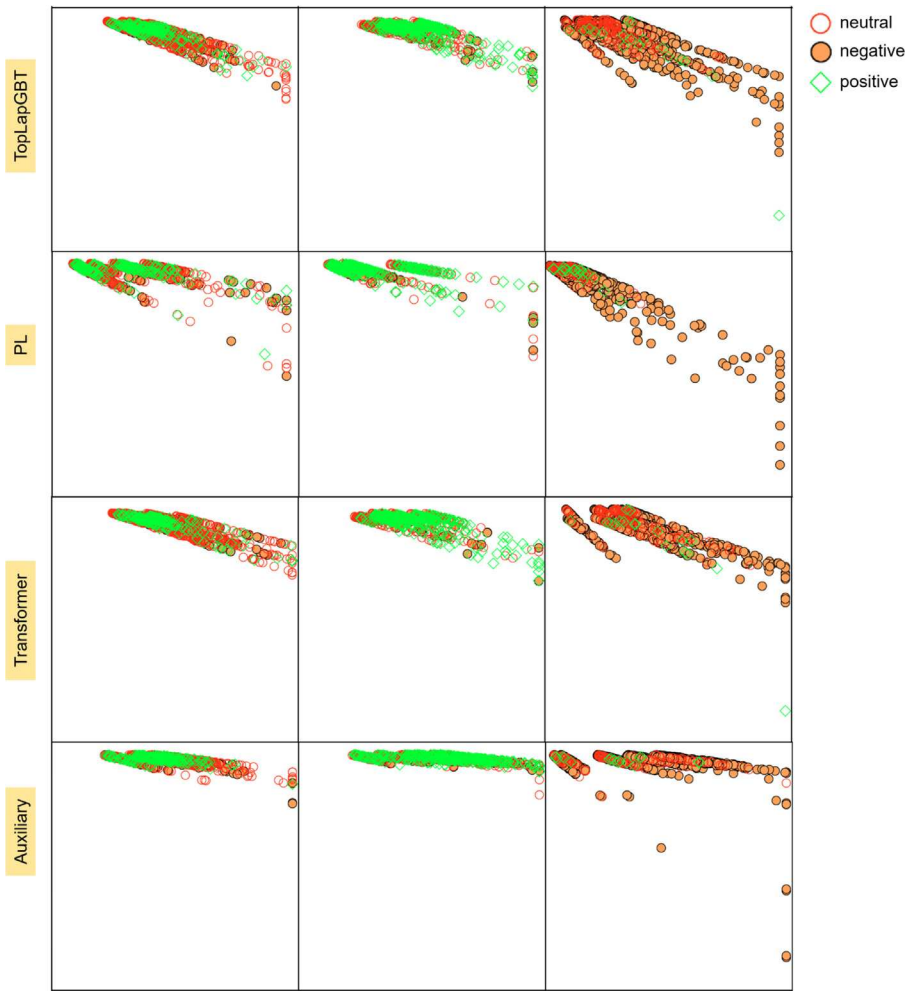

Fig. 4. The comparison of R-S plots between the different types of features used in TopLapGBT model. The *y*-axis represents the residue score, whereas the *x*-axis represents the similarity score. RS scores were computed for the testing set, and all 10-folds were visualized. Each section corresponds to one of the 3 true solubility labels, and the sample's color and marker correspond to the predicted label from TopLapGBT.

The impetus for utilizing structure-based features stems from the multifaceted relationship that exists among protein sequence, structure, and solubility. Factors such as hydrophobicity, charge distribution, and intermolecular interactions contribute to the complexity of protein solubility. Traditional prediction methods, which often rely on empirical rules or rudimentary descriptors, fall short in capturing this intricate molecular interplay. By employing advanced mathematical techniques like persistent Laplacian (PL) coupled with machine learning algorithms, we can decipher the complex patterns and relationships embedded within protein sequences and structures. Persistent Laplacian, in particular, provides a robust mathematical representation that captures both the topological and homotopic evolution of protein structures. Furthermore, machine learning models rooted in advanced mathematics offer several advantages for classifying changes in protein solubility. These models are well-suited for handling high-dimensional and complex data sets, such as those involving protein sequences and structures. They are also capable of learning non-linear relationships and capturing nuanced dependencies that are often overlooked by traditional linear models. Importantly, these advanced models can adeptly manage class-imbalanced datasets, which are commonly encountered in protein solubility studies.

4. Conclusion

In the multifaceted quest to understand mutation-induced solubility changes, various scientific domains including quantum mechanics, molecular mechanics, biochemistry, biophysics, and molecular biology have made significant contributions. Despite these concerted efforts, state-of-art models have limitations, as evidenced by their normalized CPR value of 0.656 even after employing feature selection methods. Persistent homology (PH) has emerged as a powerful tool for capturing the complexity of biomolecular structures and has achieved noteworthy success in drug discovery applications. However, its inability to capture the nuances of homotopic shape evolution, crucial for delineating molecular interactions in proteins, presents a critical shortcoming.

Our study introduces TopLapGBT, a novel model that integrates persistent Laplacian (PL) features with pretrained transformer features, thereby bridging the gap in capturing both topology and homotopic shape evolution. This innovative fusion leads to significant advancements in classification performance. Specifically, TopLapGBT achieves normalized CPR and GC^2 scores of 0.688 and 0.361, respectively, marking improvements of 4.88% and 15.71% over the state-of-the-art

PON-Sol2. These findings are further corroborated by an independent blind test, where TopLapGBT continues to outperform existing models.

In summary, our proposed TopLapGBT model not only achieves superior performance over existing state-of-the-art methods but also introduces a more nuanced approach for the classification of protein solubility changes upon mutation. These results underscore the transformative potential of integrating geometric and topological features with machine learning in advancing the field of molecular biology.

5. Materials and methods

In this section, we endeavor to elucidate key mathematical and computational foundations that are instrumental for the work presented in this study. Specifically, we delve into spectral graph theory, simplicial complex, and persistent Laplacian methods, highlighting their significance in capturing topological and spectral properties essential for the characterization of proteins. Additionally, we discuss machine learning and deep learning paradigms, focusing on their role in processing, analyzing, and interpreting these complex features, especially within the confines of test datasets and validation settings.

5.1. Persistent Laplacian characterization of proteins

Simplicial complex. A simplicial complex is made up of a set of simplices and generalizes beyond graph networks at higher dimensions [44–47]. Every simplex is a finite set of vertices which can be interpreted as the atoms in a protein structure. Essentially, simplices can be a point (0-simplex), an edge (1-simplex), a triangle (2-simplex), a tetrahedron (3-simplex), or in higher dimensions, a p-simplex. In other words, a k-simplex $\sigma^k = \{v_0, v_1, \dots, v_k\}$ is the convex hull formed by $k+1$ affinely independent points $v_0, v_1, \dots, v_k$ as follows,

$$\sigma^k = \left\{ \lambda_0 v_0 + \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_k v_k | \sum_{i=0}^{k} \lambda_i = 1; \forall i, 0 \le \lambda_i \le 1 \right\}$$

A geometric simplicial complex K is a finite set of geometric simplexes that satisfy two essential conditions. First, any face of a simplex from K is also in K. Second, the intersection of any two simplexes in K is either empty or shares faces. Commonly used methods to construct simplicial complexes are Čech complex, Vietoris-Rips complex, Alpha complex, Clique complex, Cubic complex, and Morse complex [44–47].

Chain group. A kth chain group C_k is a free Abelian group generated by oriented k-simplices σ^k. A boundary operator $\partial_k : C_k \to C_{k-1}$ defined on an oriented k-simplex σ^k can be written as

$$\partial_k \sigma^k = \sum_{i=0}^{k} (-1)^i [v_0, v_1, v_2, \dots, \hat{v_i}, \dots, v_k],$$

where $[v_0, v_1, v_2, \dots, \hat{v_i}, \dots, v_k]$ is an oriented $(k-1)$-simplex, which is constructed by the all the vertices except v_i, i.e., removing v_i from the simplex. The boundary operator satisfies $\partial_{k-1}\partial_k = 0$.

The adjoint of ∂_k, which is

$$\partial_k^* : C_{k-1} \to C_k,$$

satisfies the inner product relation $\langle \partial_k(f), g \rangle = \langle f, \partial_k^*(g) \rangle$, for every $f \in C_k$, $g \in C_{k-1}$. This will be used in the combinatorial Laplacian.

Combinatorial Laplacian. For the k-boundary operator $\partial_k : C_k \to C_{k-1}$ in K, define $\mathbf{B}_k$ to be an $m \times n$ matrix representation of the boundary operator under the standard bases $\{\sigma_i^k\}_{i=1}^n$ and $\{\sigma_j^{k-1}\}_{j=1}^m$ of C_k and C_{k-1}. Similarly, the matrix representation of ∂_k^* is the transpose matrix $\mathbf{B}_k^\top$, with respect to the same ordered bases of the boundary operator ∂_k.

More specifically, let m and n be the number of $(k-1)$-simplices and p-simplices respectively in a simplicial complex K. The $m \times n$ boundary matrix $\mathbf{B}_k$ has entries defined as follows:

$$\mathbf{B}_k(i,j) = \begin{cases} 1, & \text{if } \sigma_i^{k-1} < \sigma_j^k, \sigma_i^{k-1} \sim \sigma_j^k. \\ -1, & \text{if } \sigma_i^{k-1} < \sigma_j^k, \sigma_i^{k-1} \nsim \sigma_j^k. \\ 0, & \text{if } \sigma_i^{k-1} \not< \sigma_j^k. \end{cases}$$

where $1 \le i \le m$ and $1 \le j \le n$. Here, $\sigma_i^{k-1} < \sigma_j^k$ represents the ith $(k-1)$-simplex σ_i^{k-1} is a face of jth k-simplex σ_j^k and $\sigma_i^{k-1} \sim \sigma_j^k$ indicates the coefficient of σ_i^{k-1} in $\partial_k(\sigma_j^k)$ is 1. Likewise, $\sigma_i^{k-1} \not< \sigma_j^k$ means that σ_i^{k-1} is not a face of σ_j^k and $\sigma_i^{k-1} \nsim \sigma_j^k$ indicates that the coefficient of σ_i^{k-1} in $\partial_k(\sigma_j^k)$ is -1.

Then the k-combinatorial Laplacian or the topological Laplacian is a linear operator $\Delta_k : C_k(K) \to C_k(K)$

$$\Delta_k := \partial_{k+1}\partial_{k+1}^* + \partial_k^*\partial_k. \tag{1}$$

The k-combinatorial Laplacian exhibits an $n \times n$ matrix representation L_k and is given by

$$\mathbf{L}_k = \mathbf{B}_{k+1}\mathbf{B}_{k+1}^\top + \mathbf{B}_k^\top \mathbf{B}_k. \tag{2}$$

In the case $k = 0$, then $\mathbf{L}_0 = \mathbf{B}_1\mathbf{B}_1^\top$ since ∂_0 is a zero map.

The number of rows in $\mathbf{B}_k$ represents the number of $(k-1)$-simplices in K and the number of columns refers to the number of k-simplices in K. Furthermore, the upper k-combinatorial Laplacian matrix is $\mathbf{L}_k^U = \mathbf{B}_{k+1}\mathbf{B}_{k+1}^\top$ and the lower k-combinatorial Laplacian matrix is $\mathbf{L}_k^L = \mathbf{B}_k^\top\mathbf{B}_k$. Recall that since ∂_0 is a zero map, hence $\mathbf{L}_0(K) = \mathbf{B}_1\mathbf{B}_1^\top$ with $\mathbf{B}_0$ being a zero matrix and K being an oriented simplicial complex of dimension 1. In fact, the 0-combinatorial Laplacian matrix $\mathbf{L}_0(K)$ is actually the graph Laplacian in spectral graph theory.

The above graph Laplacian matrices can be explicitly described in terms of the simplex relations. More precisely, $\mathbf{L}_0$ can be described as

$$\mathbf{L}_0(i,j) = \begin{cases} d(\sigma_i^0), & \text{if } i = j \\ -1, & \text{if } i \ne j \text{ and } \sigma_i^0 \frown \sigma_j^0 \\ 0, & \text{if } i \ne j \text{ and } \sigma_i^0 \not\frown \sigma_j^0, \end{cases}$$

which is equivalent to the graph Laplacian. Furthermore, when $k > 0$, $\mathbf{L}_k$ can be expressed as

$$\mathbf{L}_k(i,j) = \begin{cases} d(\sigma_i^k) + k + 1, & \text{if } i = j \\ 1, & \text{if } i \ne j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \smile \sigma_j^k \text{ and } \sigma_i^k \sim \sigma_j^k \\ -1, & \text{if } i \ne j, \sigma_i^k \not\frown \sigma_j^k, \sigma_i^k \smile \sigma_j^k \text{ and } \sigma_i^k \nsim \sigma_j^k \\ 0, & \text{if } i \ne j, \text{ and either } \sigma_i^k \frown \sigma_j^k \text{ or } \sigma_i^k \not\smile \sigma_j^k. \end{cases}$$

Here, we denote $\sigma_j^{k-1} \sim \sigma_i^k$ if they have the same orientation, i.e. similarly oriented. Furthermore, we say that two k-simplices σ_i^k and σ_j^k are upper adjacent (resp. lower adjacent) neighbors, denoted as $\sigma_i^k \frown \sigma_j^k$ (resp. $\sigma_i^k \smile \sigma_j^k$), if they are both faces of a common $(k+1)$-simplex (resp. they both share a common $(k-1)$-simplex as their face). In addition, if the orientations of their common lower simplex are the same, it is called similar common lower simplex ($\sigma_i^k \smile \sigma_j^k$ and $\sigma_i^k \sim \sigma_j^k$). On the other hand, if the orientations are different, it is called dissimilar common lower simplex ($\sigma_i^k \smile \sigma_j^k$ and $\sigma_i^k \nsim \sigma_j^k$). The (upper) degree of a k-simplex σ_i^k, denoted as $d(\sigma_i^k)$, is the number of $(k+1)$-simplices, of which σ_i^k is a face.

The eigenvalues of combinatorial Laplacian matrices are independent of the choice of the orientation [48]. Furthermore, the multiplicity of zero eigenvalues, i.e. the total number of zero eigenvalues, of $\mathbf{L}_k$ corresponds to the kth Betti number β_k, according to the combinatorial Hodge theorem [49]. The kth Betti numbers are topological invariants that describe the k-dimensional holes in a simplicial complex. In particular, β_0, β_1 and β_2 represents the numbers of independent components, rings and cavities, respectively.

Persistent Laplacian. Persistent Laplacian (PL) were first introduced by integrating graph Laplacian and multiscale filtration [23]. Analyzing the spectra of k-combinatorial Laplacian matrix allows both topological and geometric information (i.e. connectivity and robustness of simple graphs) to be obtained. However, this method is genuinely free of metrics or coordinates, which induced too little topological and geometric information that can be used to describe a single configuration.

Therefore, PL was extended to simplicial complexes. This allows a sequence of simplicial complexes from a filtration process to generate

persistent Laplacian which is largely inspired by persistent homology and in earlier works in multiscale graphs. For the rest of this section, we introduce mainly on the construction of PL. First, a k-combinatorial Laplacian matrix is symmetric and positive semi-definite. Therefore, its eigenvalues are all real and non-negative. The multiplicity of zero spectra (also called harmonic spectra) reveals the topological information, and the geometric information will be preserved in the non-harmonic spectra.

A key concept of PL is the filtration process. Essentially, an ever-increasing filtration value f is used to generate a series of topological spaces, which are represented by a nested sequence of multiscale simplicial complexes. Naturally, PL generates a sequence of simplicial complexes induced by varying a filtration parameter [23]. For an oriented simplicial complex K, its filtration is a nested sequence of simplicial complexes $(K_t)_{t=0}^{m}$ of K

$$\varnothing = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_m = K.$$

This nested sequence of simplicial complexes induces a family of chain complexes

$$\left\{ \cdots \underset{\partial_{k+2}^{t*}}{\overset{\partial_{k+2}^{t}}{\leftrightharpoons}} C_{k+1}^{t} \underset{\partial_{k+1}^{t*}}{\overset{\partial_{k+1}^{t}}{\leftrightharpoons}} C_{k}^{t} \underset{\partial_{k}^{t*}}{\overset{\partial_{k}^{t}}{\leftrightharpoons}} \cdots \underset{\partial_{1}^{t*}}{\overset{\partial_{1}^{t}}{\leftrightharpoons}} C_{0}^{t} \underset{\partial_{0}^{t*}}{\overset{\partial_{0}^{t}}{\leftrightharpoons}} \varnothing \right\}_{t\in\mathbb{R}^{+}}. \tag{3}$$

where $C_k^t = C_k(K_t)$ is the chain group for the subcomplex K_t, and its k-boundary operator is $\partial_k^t : C_k(K_t) \to C_{k-1}(K_t)$. In the case $k < 0$, then $C_k(K_t)$ is an empty set and ∂_k^t is a zero map. For $0 < k < \dim(K_t)$, the boundary operator

$$\partial_k^t(\sigma_k) = \sum_{i=0}^{k} (-1)^i \sigma_i^{k-1}, \quad \sigma^k \in K_t, \tag{4}$$

with $\sigma^k = [v_0, \ldots, v_k]$ being the k-simplex, and $\sigma_i^{k-1} = [v_0, \ldots, \hat{v_i}, \ldots, v_k]$ being the $(k-1)$-simplex for which its vertex v_i is removed. Similarly, the adjoint of ∂_k^t is the operator $\partial_k^{t*} : C_{k-1}(K_t) \to C_k(K_t)$. The topological and spectral characteristics can then be studied from $\mathbf{L}_k(K_t)$ by varying the filtration parameter and diagonalizing the k-combinatorial Laplacian matrix. The multiplicity of the zero spectra of $\mathbf{L}_k^t$ is the persistent Betti number β_k^t, which represents the number of k-dimensional holes in K_t. In other words,

$$\beta_k^t = \dim(\mathbf{L}_k^t) - \operatorname{rank}(\mathbf{L}_k^t) = \operatorname{nullity}(\mathbf{L}_k^t) = \#\text{ of harmonic spectra of } \mathbf{L}_k^t. \tag{5}$$

In particular, β_0^t represents the number of connected components in K_t, β_1^t counts the number of one-dimensional cycles in K_t and β_2^t reveals the number of two-dimensional voids in K_t. In addition, the spectra of $\mathbf{L}_k^t$ can be written in the following ascending order

$$\text{Spectra}(\mathbf{L}_k^t) = \{(\lambda_1)_k^t, (\lambda_2)_k^t, \ldots, (\lambda_n)_k^t\}, \tag{6}$$

where $\mathbf{L}_k^t$ here is an $n \times n$ matrix. The p-persistent k-combinatorial Laplacian can be extended based on the boundary operator as well. Further details can be found in [23].

In order to illustrate the difference between PL and PH, Fig. 5 describes a point cloud, basic simplices, a filtration process and the comparison between persistent Laplacian and persistent homology barcodes of 13 points. The filtration process in Fig. 5(c) shows the different stages of a Rips filtration process for the 13 points. Fig. 5(d) shows the persistent homology barcodes (in blue) and persistent non-harmonic spectra (in red). It can be seen that the non-harmonic spectra provides the additional homotopic shape evolution that is missing in persistent homology in the later part of the filtration process.

5.2. Persistent Laplacian descriptors

In order to capture the mutation-induced solubility change, we apply the persistent Laplacian (PL) to characterize the interactions between the mutation site and the rest of the protein. To describe these interactions, we first propose the interactive PL with the distance function $\mathrm{DI}(A_i, A_j)$ describing the distance between two atoms A_i and A_j defined as

$$\mathrm{DI}(A_i, A_j) = \begin{cases} \infty, & \text{if } \mathrm{Loc}(A_i) = \mathrm{Loc}(A_j), \\ \mathrm{DE}(A_i, A_j), & \text{otherwise.} \end{cases} \tag{7}$$

where $\mathrm{DE}(\cdot,\cdot)$ is the Euclidean distance between the two atoms and $\mathrm{Loc}(\cdot)$ refers to the atom's location which is either in the mutation site or in the rest of the protein. Here, we construct two types of simplicial complexes in our PL computation, such as Vietoris-Rips complex (VC) and Alpha complex (AC). Both complexes are used to characterize the first order interactions and higher order patterns respectively. To capture and characterize different types of atom–atom interactions, we generate the PL based on different atom subsets by selecting one type of atom in the mutation site and one other atom type in the rest of the protein. Different types of atom–atom interactions characterize the different interactions in proteins. For example, interactions generated from carbon atoms are associated with hydrophobic interactions. Similarly, interactions between nitrogen and oxygen atoms correlate to hydrophilic interactions and/or hydrogen bonds. Both types of interactions are illustrated in Fig. 6. Interactive PLs have the capability to unveil additional details about bonding interactions and offer a fresh and distinct representation of molecular interactions in proteins.

The set of persistent spectra from each persistent Laplacian computation consists of $V_{\gamma,\alpha,\beta}^{DI}$ and $V_{\gamma,\alpha,\beta}^{DE}$ where $\gamma \in \{M, W\}$ refers to the mutant protein or the wild type protein, $\alpha \in \{C, N, O\}$ is the atom type chosen in the rest of the protein and $\beta \in \{C, N, O\}$ is the atom type chosen in the mutation site. $V_{\gamma,\alpha,\beta}^{DI}$ applies the distance DI-based filtration to generate 0-dimensional Laplacian using the Vietoris-Rips complex and $V_{\gamma,\alpha,\beta}^{DE}$ applies the Euclidean distance DE-based filtration to generate 1 and 2-dimensional Laplacian using the alpha complex. In total, there are 54 sets of persistent spectra. The persistent spectra from PL contains both harmonic and non-harmonic spectra that are capable of revealing the molecular mechanism of protein solubility.

For zero dimensions, we consider both the harmonic spectra and non-harmonic spectra information for each persistent Laplacian. Filtration using Rips complex with DI distance is used. The 0-dimensional PL features are generated from 0 Å to 6 Å with 0.5 Å gridsize. For the non-harmonic spectra information, we count the number of non-harmonic spectra and calculate seven statistical values of non-harmonic spectra such as sum, minimum, maximum, mean, standard deviation, variance and the sum of eigenvalues squared. This generates eight statistical values for each of the nine atomic pairs. Therefore, the dimension of 0-dimensional PL features for a protein is 72. In total, the 0-dimensional PL-based feature size after concatenating features at different dimensions for wild type and mutant is 1872.

For one or two dimensions, we perform the filtration using Alpha complex with the DE distance. The limited number of atoms in the local protein structure can create only a few high-dimensional simplexes, resulting in minimal alterations in shape. As a result, it suffice to consider features from only harmonic spectra of persistent Laplacians by coding the topological invariants for the high-dimensional interactions. Using GUDHI [51], the persistence of the harmonic spectra can be represented by persistent barcodes. The topological feature vectors are generated by computing the statistics of bar lengths, births and deaths. Bars shorter than 0.1 Å are excluded as they do not exhibit any clear physical meaning. The remaining bars are then used for computing the statistics: (1) sum, maximum and mean for lengths of bars; (2) minimum and maximum for the birth values of bars; (3) minimum and maximum for the death values of bars. Each set of point clouds leads to a seven-dimensional vector. These features are calculated on nine single atomic pairs and one heavy atom pair. The dimension of one- and two-dimensional PL feature vectors for a protein is 140. In total, the higher-dimensional PL-based feature size after concatenating features at different dimensions for wild type, mutant and their difference is 420.

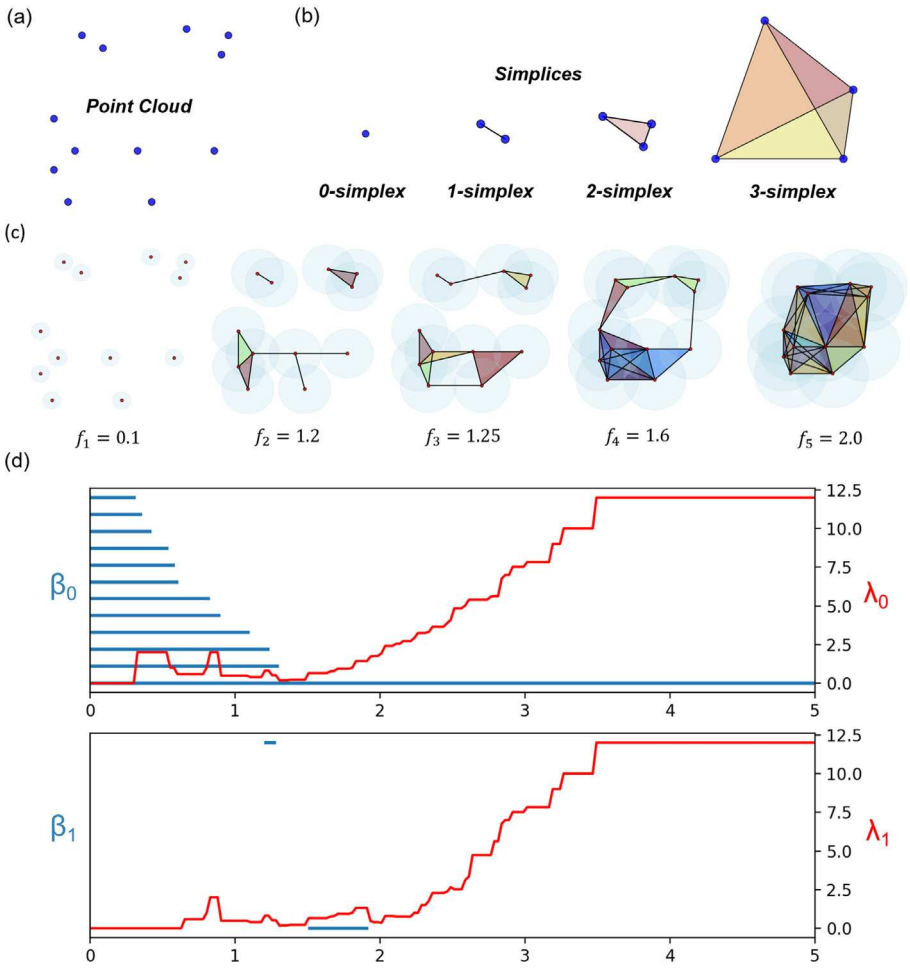

Fig. 5. The illustration of (a) point cloud, (b) basic simplices and (c) the filtration process and (d) Comparison between PH barcodes [12,50] and the non-harmonic spectra of persistent Laplacians (PLs) [23] from the filtration process in (c). The x-axis represents the filtration parameter f. By discretizing the filtration region into equal-sized bins and adding all the Betti bars together, the topological invariants are summarized into persistent Betti numbers that acts a topological descriptor extracted from protein structures. Persistent Laplacians (PLs) [23] for thirteen points. The first non-zero eigenvalues of dimension 0, $\lambda_0(r)$, and dimension 1, $\lambda_1(r)$, of PLs are depicted in red. The harmonic spectra of PLs return all the topological invariants of PH, whereas the non-harmonic spectra of PLs capture the additional homotopic shape evolution of PLs during the filtration that are neglected by PH.

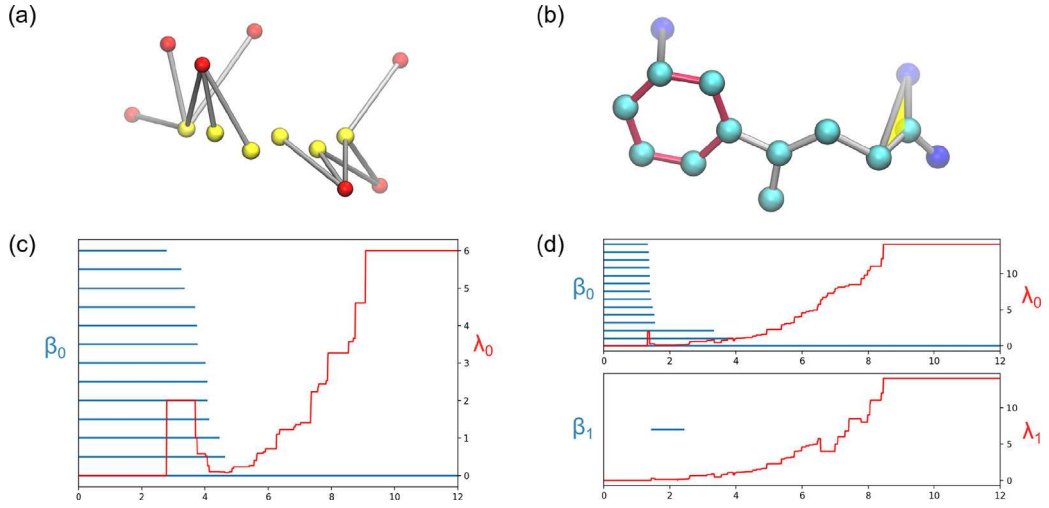

Fig. 6. An illustration of interactive PL showing hydrophilic interactions based on DI-based filtration (left) and hydrophobic interactions based on DE-based filtration (right) at a mutation site. (a) Hydrophilic interactions between the nitrogen atoms (red) and oxygen atoms (yellow). (b) Hydrophobic interactions between the carbon atoms (cyan) and oxygen atoms (dark blue). The hexagon ring is colored in red and the triangle is colored in yellow. (c) The PH barcodes and PL for dimension 0 of the hydrophilic interactions in (a). (d) The PH barcodes and PL for dimension 0 and dimension 1 of the hydrophobic interactions in (b). The Betti-1 bar is due to the red hexagon ring in (b).

5.3. Persistent homology

Persistent homology is part of the harmonic spectra of PL. The homology groups in PH illustrate the persistence of topological invariants, hence providing the harmonic spectral information in PL. The site- and element-specific PH features are generated in a similar way as compared to PL. Similar to PL, filtration construction is also employed to PH. For the zero dimension, the filtration parameter can be discretized into several equally spaced bins, namely [0, 0.5], (0.5, 1], ⋯, (5.5, 6] Å. The death value of the bars are summed in each bin resulting in 12×18 features.

For each bin, we count the numbers of persistent bars, resulting in a nine-dimensional vector for each point cloud. Similarly, this is performed for each of the nine single atomic pairs. Hence, the dimension of PH features for a protein is 216. For one or two dimensions, the identical featurization from the statistics of persistent bars in PH is used. The PH embedding combines features at different dimensions as described above and concatenated for wild type, mutant and their difference, resulting in a 648-dimensional vector.

5.4. Transformer features

Recently, we have seen significant advancements in modeling protein properties using large-scale protein transformer models trained on hundreds of millions of sequences. These models, like ESM [32] (evolutionary scale modeling) and ProtTrans [33,34], have demonstrated impressive performance. Moreover, hybrid fine-tuning approaches that leverage both local and global evolutionary data have proven to enhance these models even further. For instance, eUniRep is an improved LSTM-based UniRep model achieved through fine-tuning with knowledge extracted from local multiple sequence alignments (MSAs). Additionally, the ESM model can be fine-tuned using either downstream task data or local MSAs. In our research, we employed the ESM-1b transformer, a model that falls under the transformer architecture. This particular variant was trained on a dataset of 250 million sequences using a masked filling procedure and boasts an architecture comprising 34 layers with a whopping 650 million parameters. The ESM transformer's primary role in our work was to generate sequence embeddings. At each layer of the ESM model, it encoded a sequence of length L into a matrix sized at 1280×L, excluding the start and terminal tokens. For our study, we utilized the sequence representation derived from the final (34th) layer and computed the average along the sequence length axis, resulting in a 1280-component vector.

5.5. Performance metrics

PPV and NPV assesses the true positive and true negative proportion of the predicted results for each solubility class. PPV and NPV are computed based on TP, TN, FP and FN which represents the true positive, true negative, false positive and false negative values for each solubility class. For each solubility class, PPV and NPV can be computed by:

$$\mathrm{PPV} = \frac{TP}{TP + FP}. \tag{8}$$

$$\mathrm{NPV} = \frac{TN}{TN + FN}. \tag{9}$$

Furthermore, specificity and sensitivity can be computed by the following:

$$\mathrm{Specificity} = \frac{TN}{TN + FP}. \tag{10}$$

$$\mathrm{Sensitivity} = \frac{TP}{TP + FN}. \tag{11}$$

The correct prediction ratio (CPR) and generalized squared correlation (GC^2) are used to evaluate the overall performance of TopLapGBT. CPR and GC^2 can be computed as

$$\mathrm{CPR} = \frac{1}{N}\sum_i z_{ii}, \text{ and} \tag{12}$$

$$GC^2 = \frac{1}{N(K-1)}\sum_{ij}\frac{(z_{ij} - e_{ij})^2}{e_{ij}}, \tag{13}$$

where K is the number of classes and N is the number of samples. Here, z_{ij} represents the number of samples of class i to class j. Let $x_i = \sum_j z_{ij}$ be the number of inputs from class i, and $y_j = \sum_i z_{ij}$ be the number of inputs predicted to class j. Then the expected number of samples in (i, j)-th entry of the multiclass confusion matrix is

$$e_{ij} = \frac{x_i y_j}{N}.$$

Since the mutational samples across the three solubility classes are imbalanced, we normalized the values to provide more reliable calculation of performance metrics.

6. Software and resources

Protein sequences are first preprocessed by AlphaFold 2 to generate wild type protein structures. In particular, 3D protein structures are generated from protein sequences using ColabFold [52]. Mutant proteins are generated from the Jackal software [38]. All TopLapGBT models are built using the sklearn machine learning library [53]. The hyperparameters for all the TopLapGBT are: n_estimators = 20,000, learning_rate = 0.05, max_depth = 7, subsample = 0.4, min_sample_split = 3 and max_features = sqrt. The PQR files, which contains the partial charge information of the proteins, are generated from the PDB2PQR software [54]. The PQR files for both the wild type proteins are generated with AMBER force field. The solvation energy and surface area information are calculated from the in-house online software package ESES [55] and MIBPB [56]. The pKa values are computed from the PROPKA software package [57]. The position-specific-scoring matrices (PSSM) are computed from the BLAST+ software [58] using the nr database. The secondary structure features and torsion angle sequence-based information are calculated from SPIDER [59]. The persistent Laplacian descriptors for both VR complexes and alpha complexes are calculated using the GUDHI software library [60]. All computational work in support of this research was performed using the resources from the National Super Computing Centre of Singapore (NSCC).

Code and data availability

The 3D protein structures and the TopLapGBT code can be found in https://github.com/ExpectozJJ/TopLapGBT. The source code for the R-S plot can be found at https://github.com/hozumiyu/RSI.

CRediT authorship contribution statement

JunJie Wee: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Jiahui Chen:** Conceptualization, Formal analysis, Writing – review & editing. **Kelin Xia:** Funding acquisition, Writing – review & editing. **Guo-Wei Wei:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Code and data availability

The 3D protein structures and the TopLapGBT code can be found in https://github.com/ExpectozJJ/TopLapGBT. The source code for the R-S plot can be found at https://github.com/hozumiyu/RSI.

Acknowledgments

This work was supported in part by NIH grants R01GM126189, R01AI164266, and R35GM148196, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA grant 80NSSC21M0023, MSU Foundation, Bristol-Myers Squibb 65109, and Pfizer. It was supported in part by Nanyang Technological University Startup Grant M4081842.110, Singapore Ministry of Education Academic Research fund Tier 1 RG109/19 and Tier 2 MOE-T2EP20120-0013, MOE-T2EP20220-0010, and MOE-T2EP20221-0003.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.107918.

References

[1] Y. Qiu, G.-W. Wei, Persistent spectral theory-guided protein engineering, Nat. Comput. Sci. 3 (2) (2023) 149–163.
[2] R. Guerois, J.E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, J. Mol. Biol. 320 (2) (2002) 369–387.
[3] P. Sormanni, F.A. Aprile, M. Vendruscolo, The CamSol method of rational design of protein mutants with enhanced solubility, J. Mol. Biol. 427 (2) (2015) 478–490.
[4] Y. Tian, C. Deutsch, B. Krishnamoorthy, Scoring function to predict solubility mutagenesis, Algorithms Mol. Biol. 5 (1) (2010) 1–11.
[5] Y. Yang, A. Niroula, B. Shen, M. Vihinen, PON-Sol: Prediction of effects of amino acid substitutions on protein solubility, Bioinformatics 32 (13) (2016) 2032–2034.
[6] L. Paladin, D. Piovesan, S.C. Tosatto, SODA: Prediction of protein solubility from disorder and aggregation propensity, Nucleic Acids Res. 45 (W1) (2017) W236–W240.
[7] J. Van Durme, G. De Baets, R. Van Der Kant, M. Ramakers, A. Ganesan, H. Wilkinson, R. Gallardo, F. Rousseau, J. Schymkowitz, Solubis: a webserver to reduce protein aggregation through mutation, Protein Eng., Des. Select. 29 (8) (2016) 285–289.
[8] M. Vihinen, Solubility of proteins, ADMET and DMPK 8 (4) (2020) 391–399.
[9] 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 (10) (2004) 1302–1306.
[10] H. Land, M.S. Humble, YASARA: A tool to obtain structural guidance in biocatalytic investigations, Protein Eng.: Methods Protocols (2018) 43–67.
[11] Y. Yang, L. Zeng, M. Vihinen, PON-Sol2: Prediction of effects of variants on protein solubility, Int. J. Mol. Sci. 22 (15) (2021) 8027.
[12] H. Edelsbrunner, J.L. Harer, Computational Topology: An Introduction, American Mathematical Society, 2022.
[13] A. Zomorodian, G. Carlsson, Computing persistent homology, in: Proceedings of the Twentieth Annual Symposium on Computational Geometry, 2004, pp. 347–356.
[14] K. Xia, G.-W. Wei, Persistent homology analysis of protein structure, flexibility, and folding, Int. J. Numer. Methods Biomed. Eng. 30 (8) (2014) 814–844.
[15] Z. Cang, L. Mu, K. Wu, K. Opron, K. Xia, G.-W. Wei, A topological approach for protein classification, Comput. Math. Biophys. 3 (1) (2015).
[16] Z. Cang, G.-W. Wei, TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, PLoS Comput. Biol. 13 (7) (2017) e1005690.
[17] Z.X. Cang, G.W. Wei, Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction, Int. J. Numer. Methods Biomed. Eng. 34 (2) (2018) e2914.
[18] 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 (2) (2020) 116–123.
[19] J. Chen, R. Wang, M. Wang, G.-W. Wei, Mutations strengthened SARS-CoV-2 infectivity, J. Mol. Biol. 432 (19) (2020) 5212–5226.
[20] D.D. Nguyen, Z. Cang, K. Wu, M. Wang, Y. Cao, G.-W. Wei, Mathematical deep learning for pose and binding affinity prediction and ranking in D3R grand challenges, J. Comput. Aided Mol. Des. 33 (2019) 71–82.
[21] D.D. Nguyen, K. Gao, M. Wang, G.-W. Wei, MathDL: mathematical deep learning for D3R grand challenge 4, J. Comput. Aided Mol. Des. 34 (2020) 131–147.
[22] D.D. Nguyen, Z. Cang, G.-W. Wei, A review of mathematical representations of biomolecular data, Phys. Chem. Chem. Phys. 22 (8) (2020) 4343–4367.
[23] R. Wang, D.D. Nguyen, G.-W. Wei, Persistent spectral graph, Int. J. Numer. Methods Biomed. Eng. (2020) e3376.
[24] J. Chen, R. Zhao, Y. Tong, G.-W. Wei, Evolutionary de Rham-Hodge method, Discrete and Continuous Dynamical Systems - B 26 (7) (2021) 3785–3821.
[25] Z. Meng, K. Xia, Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction, Sci. Adv. 7 (19) (2021) eabc5329.
[26] J. Wee, K. Xia, Persistent spectral based ensemble learning (PerSpect-EL) for protein–protein binding affinity prediction, Brief. Bioinform. (2022) bbac024.
[27] J. Bi, J. Wee, X. Liu, C. Qu, G. Wang, K. Xia, Multiscale topological indices for the quantitative prediction of SARS CoV-2 binding affinity change upon mutations, J. Chem. Inf. Model. 63 (13) (2023) 4216–4227.
[28] J. Chen, Y. Qiu, R. Wang, G.-W. Wei, Persistent Laplacian projected Omicron BA.4 and BA.5 to become new dominating variants, Comput. Biol. Med. 151 (2022) 106262.
[29] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, Y. Song, Evaluating protein transfer learning with TAPE, Adv. Neural Inf. Process. Syst. 32 (2019).
[30] T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure, in: International Conference on Learning Representations, 2018.
[31] E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G.M. Church, Unified rational protein engineering with sequence-based deep representation learning, Nat. Methods 16 (12) (2019) 1315–1322.
[32] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C.L. Zitnick, J. Ma, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci. 118 (15) (2021) e2016239118.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, North American Chapter of the Association for Computational Linguistics, 2019.
[35] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Vol. 34, Curran Associates, Inc., 2021, pp. 29287–29303.
[36] M. Mou, Z. Pan, Z. Zhou, L. Zheng, H. Zhang, S. Shi, F. Li, X. Sun, F. Zhu, A transformer-based ensemble framework for the prediction of protein–protein interaction sites, Research 6 (2023) 0240.
[37] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.
[38] J.Z. Xiang, B. Honig, Jackal: A Protein Structure Modeling Package, Columbia University and Howard Hughes Medical Institute, New York, 2002.
[39] R.M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, A. Rives, MSA transformer, in: International Conference on Machine Learning, PMLR, 2021, pp. 8844–8856.
[40] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: An overview, Bioinformatics 16 (5) (2000) 412–424.
[41] E.D. Levy, A simple definition of structural regions in proteins and its use in analyzing interface evolution, J. Mol. Biol. 403 (4) (2010) 660–670.
[42] D.S. Goodsell, L. Autin, A.J. Olson, Illustrate: Software for biomolecular illustration, Structure 27 (11) (2019) 1716–1720.
[43] Y. Hozumi, K.A. Tanemura, G.-W. Wei, Preprocessing of single cell RNA sequencing data using correlated clustering and projection, J. Chem. Inf. Model. (2023).
[44] J.R. Munkres, Elements of Algebraic Topology, CRC Press, 2018.
[45] A.J. Zomorodian, Topology for Computing, Vol. 16, Cambridge University Press, 2005.
[46] H. Edelsbrunner, J. Harer, Computational Topology: An Introduction, American Mathematical Soc., 2010.
[47] K. Mischaikow, V. Nanda, Morse theory for filtrations and efficient computation of persistent homology, Discrete Comput. Geom. 50 (2) (2013) 330–353.
[48] D. Horak, J. Jost, Spectra of combinatorial Laplace operators on simplicial complexes, Adv. Math. 244 (2013) 303–336.
[49] B. Eckmann, Harmonische funktionen und randwertaufgaben in einem komplex, Comment. Math. Helv. 17 (1) (1944) 240–255.
[50] A. Zomorodian, Topological data analysis, Adv. Appl. Comput. Topol. 70 (2012) 1–39.
[51] The GUDHI Project, GUDHI User and Reference Manual, GUDHI Editorial Board, 2015.
[52] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: Making protein folding accessible to all, Nat. Methods 19 (6) (2022) 679–682.
[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[54] T.J. Dolinsky, J.E. Nielsen, J.A. McCammon, N.A. Baker, PDB2PQR: An automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations, Nucleic Acids Res. 32 (suppl_2) (2004) W665–W667.
[55] B. Liu, B. Wang, R. Zhao, Y. Tong, G.-W. Wei, ESES: Software for Eulerian solvent excluded surface, J. Comput. Chem. 38 (7) (2017) 446–466.

[56] D. Chen, Z. Chen, C. Chen, W. Geng, G.-W. Wei, MIBPB: A software package for electrostatic analysis, J. Comput. Chem. 32 (4) (2011) 756–770.
[57] H. Li, A.D. Robertson, J.H. Jensen, Very fast empirical prediction and rationalization of protein pKa values, Proteins: Struct. Funct. Bioinform. 61 (4) (2005) 704–721.
[58] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, T.L. Madden, NCBI BLAST: A better web interface, Nucleic Acids Res. 36 (suppl_2) (2008) W5–W9.
[59] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning, Sci. Rep. 5 (1) (2015) 11476.
[60] C. Maria, J.-D. Boissonnat, M. Glisse, M. Yvinec, The GUDHI library: Simplicial complexes and persistent homology, in: Mathematical Software–ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings 4, Springer, 2014, pp. 167–174.