Computers in Biology and Medicine 164 (2023) 107258

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Topological deep learning based deep mutational scanning

Jiahui Chen [a], Daniel R. Woldring [b], Faqing Huang [c], Xuefei Huang [d,e,f], Guo-Wei Wei [g,h,i,*]

[a] Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA
[b] Department of Chemical Engineering, Michigan State University, East Lansing, MI 48824, USA
[c] Department of Chemistry and Biochemistry, University of Southern Mississippi, Hattiesburg, MS 39406, USA
[d] Department of Chemistry, Michigan State University, MI 48824, USA
[e] Department of Biomedical Engineering, Michigan State University, East Lansing, MI 48824, USA
[f] The Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
[g] Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA
[h] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA
[i] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA

ARTICLE INFO

Keywords:
SARS-coV-2
Infectivity
Antibody-resistance
Deep mutational scanning
Topological deep learning

ABSTRACT

High-throughput deep mutational scanning (DMS) experiments have significantly impacted protein engineering, drug discovery, immunology, cancer biology, and evolutionary biology by enabling the systematic understanding of protein functions. However, the mutational space associated with proteins is astronomically large, making it overwhelming for current experimental capabilities. Therefore, alternative methods for DMS are imperative. We propose a topological deep learning (TDL) paradigm to facilitate *in silico* DMS. We utilize a new topological data analysis (TDA) technique based on the persistent spectral theory, also known as persistent Laplacian, to capture both topological invariants and the homotopic shape evolution of data. To validate our TDL-DMS model, we use SARS-CoV-2 datasets and show excellent accuracy and reliability for binding interface mutations. This finding is significant for SARS-CoV-2 variant forecasting and designing effective antibodies and vaccines. Our proposed model is expected to have a significant impact on drug discovery, vaccine design, precision medicine, and protein engineering.

1. Introduction

Protein mutations refer to changes in the DNA sequence that result in alterations in the amino acid sequence of a protein. These changes can significantly affect the protein's structure, function, and stability, including protein folding stability, protein binding affinity, and protein–protein interactions (PPIs). Protein mutations play a paramount role in evolutionary biology, cancer biology, immunology, directed evolution, and protein engineering.

Accurately analyzing the impact of mutations is crucial in many fields, such as identifying deleterious and benign mutations and developing novel antibody therapies for emerging virus variants. However, experimental evaluation of mutational outcomes can be time-consuming and expensive, as it requires the expression and purification of variant proteins and measurement of their activity over time [1]. Furthermore, measurements of site-directed mutagenesis for a single mutation may vary dramatically across different experimental approaches [2]. Therefore, leveraging accurate and reliable computational methods to predict the impact of mutations could have a profound effect on the throughput and accessibility of protein engineering and drug discovery.

Recent computational predictions of the impact of mutations on protein stability and PPI binding affinity have proven to be an important alternative to experimental mutagenesis analysis for systematically exploring protein structural functions, disease connections, virus infectivity, structural instability, and organism evolution directions [3–5]. Computational approaches offer a rapid, economical, and potentially accurate alternative to site-directed mutational experiments. Many computational methods have been employed for fields as diverse as protein folding energy changes and PPI binding free energy changes upon mutation.

Various computational methods have been developed to predict the impact of mutations on protein stability, each with differing accuracies and computational requirements. Such methods include I-Mutant [6], FoldX [3], SDM [7], DUET [8], PoPMuSiC [9], Rosetta [10], SAAFEC [11], PPSC [12], PROVEAN [13], ELASPIC [14], STRUM [15], EASE-MM [16] etc. DUET, for instance, demonstrates a high correlation in a blind test set and outperforms individual methods like SDM and mCSM [8]. FoldX has the advantage of being easier to run locally,

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

https://doi.org/10.1016/j.compbiomed.2023.107258
Received 9 May 2023; Received in revised form 28 June 2023; Accepted 8 July 2023
Available online 17 July 2023

while PROVEAN offers reasonable results with lower computational costs and without requiring a protein structure [17]. Computational approaches designed to estimate PPI binding free energy changes upon mutation typically rely on physical force fields, electrostatics, conformational sampling, and hydrophobic packing. These methods offer a computationally efficient alternative, including DFIRE [18], FoldX [3], Discovery Studio [19], EGED [20], CC/PBSA [21], Rosetta [22], PoPMuSiC & BeAtMuSic [9,23], and mCSM [24,25]. Several studies have compared the performance of various computational methods in predicting protein stability and binding affinity changes upon mutation. One such study assessed the performance of six methods, including CC/PBSA, EGAD, FoldX, I-Mutant2.0, Rosetta, and Hunter, in predicting protein stability changes [26]. Another investigation evaluated the effectiveness of several methods, including bASA, dDFIRE, DFIRE, STATIUM, Rosetta, FoldX, and Discovery Studio scoring potentials – in predicting antibody binding affinity changes upon mutation. The respective Pearson correlation were 0.22, 0.19, 0.31, 0.32, 0.16, 0.34, and 0.45, respectively [27].

Computational approaches for calculating protein biophysical properties generally fall into three categories: empirical models, physical models, and data-driven machine learning techniques. Empirical models implement potential terms and empirical functions to describe the free energy perturbation under the constraint of the range of conditions for which they are developed [28,29]. Physical modeling makes use of multiscale implicit solvent models and molecular mechanics approaches. On the other hand, these approaches depend on the accurate and self-sufficient predictions derived from the underlying measurements [30].

Alternatively, data-driven approaches employ machine learning (ML) and deep learning (DL) techniques to uncover the mechanism linking protein stability/binding with complex structures or polypeptides. A major advantage of data-driven mutation modeling is its ability to handle high-throughput and diverse mutation datasets. Importantly, the predictive performance of DL approaches heavily relies on the availability and accuracy of training sets. The computational prediction of the impact of mutations on protein stability and protein–protein interactions (PPIs) plays a crucial role in drug repositioning and drug-target interaction. These predictions are essential for identifying deleterious and benign mutations, developing novel antibody therapies for emerging virus variants, and facilitating the throughput and accessibility of protein engineering [31] and drug discovery [32–35].

Deep mutational scanning (DMS) – a high-throughput experimental technique used to study the effects of thousands of mutations on a protein's function, such as fitness, stability, and reactivity [36] – can directly benefit from increasing data availability. This approach combines site-directed mutagenesis with next-generation sequencing to measure the fitness of each mutation in a population based on its enrichment (i.e., change in frequency) during selection or screening. DMS has emerged as a primary approach for protein engineering [36–38] and provides reliable analysis of mutational impacts on protein stability, binding free energy, or evolutionary directions. DMS can measure tens of thousands of variants in a single experiment, providing datasets for machine/deep learning studies. For example, the stochastic gradient boosting model, Envision, uses 21,026 variant effect measurements from nine mutational scan studies to create a unified mutant effect predictor. This predictor outperforms other missense variant effect predictors on both large-scale mutagenesis data and an independent test dataset consisting of 2312 TP53 variants [39]. Sarfati et al. combined deep mutational scanning data and machine learning to predict mutant impacts using sequence and structure features of variants. These were measured in the overall correlation between the predicted and enrichment results [40].

With the advances in experimental techniques and computational approaches, we are now better equipped to study the emergence and evolution of viruses. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has caused a global pandemic since late 2019, evolving into many different variants which have resulted in several waves of Coronavirus Disease 2019 (COVID-19) infections. SARS-CoV-2 exploits mutations to enhance its evolutionary fitness. Two mechanisms of SARS-CoV-2 evolution, namely natural selection via infectivity strengthening and antibody resistance, were identified in early 2020 [4] and late 2021 [41], respectively based on molecular biophysics, topological deep learning, and genotyping of viral genomes isolated from patients. The molecular model underlying the first mechanism is that mutations on the spike protein (S protein) receptor-binding domain (RBD) enhance the virus host cell entry by strengthening the binding of RBD and host angiotensin-converting enzyme 2 (ACE2), giving rise to more infectious variants [1,4,42–46]. The molecular model underlying the second mechanism is that RBD mutations are able to disrupt the RBD and antibody binding, leading to serious vaccine breakthroughs in the populations of Europe and the US that had the earliest access to vaccines [41].

DMS, recognized as a reliable option, is employed to measure the impact of single-amino acid mutations on the RBD-ACE2 binding affinity [47–50] and RBD-antibody binding affinity [48,51,52]. One study reports the stabilization of the original SARS-CoV-2 spike protein RBD through the integration of deep mutational scanning and computational design [53].

Topological Deep Learning (TDL), first introduced in 2017 [54], has emerged as a paradigm that amalgamates topological data analysis (TDA) and deep learning techniques to analyze complex and high-dimensional data. TDA is a branch of mathematics that focuses on understanding the shape and structure of data [55,56]. It is most successful in cases where standard approaches fare poorly, but it can also significantly contribute in the situations where the standard approaches work very well, by contributing novel topological fingerprints. The basic idea behind TDL is to incorporate topological features of the data into deep learning models to improve their performance. This can be done by using topological descriptors to simplify the structural complexity of biomolecules [57–59] and embed physical interactions into topological invariants [54]. TopNetTree [60] model was designed for predicting PPI binding free energy changes upon mutation. These studies have significant implications for the field of computational biology and complex biological systems. Topological deep learning leverages the data analysis of intricate and high-dimensional data. Recently, TopNetmAb model has further been validated with DMS data and has been applied to predict RBD mutation-induced RBD-antibody binding free energy (BFE) changes [5].

Persistent homology, a key method in TDA, was employed in early TopNetTree [60] and TopNetmAb [5] models. We recently developed a persistent Laplacian-based TDL model for predicting PPI binding free energy changes upon mutation [61]. Persistent Laplacians, also referred to as persistent spectral theory [62], are a particular instance within a family of persistent topological Laplacians, including persistent path Laplacians [63], persistent sheaf Laplacians [64], persistent hyperdigraph Laplacians [65], etc. Persistent topological Laplacians are designed to address the limitations of the current TDA methods.

In this work, we propose a TDL-DMS predictor for mutation-induced protein–protein interaction BFE changes. We collect five DMS datasets focusing on SARS-CoV-2 S protein RBD in RBD-ACE2 complexes and RBD-antibody complexes, including deep mutational scanning of the S protein receptor-binding domains (RBD) in the RBD-ACE2 complex [49], in another RBD-ACE2 complex [47,48], in RBD-CTC-445.2 complex [48], and in BA.1 and BA.2 variants [66]. We use an improved TDL model based persistent spectral theory [62] to construct both persistent topological invariants and persistent spectra for predicting single-amino acid mutation impacts on protein–protein interactions using the aforementioned DMS datasets as training sets. The three dimensional (3D) structures of appropriate RBD-ACE2 complexes and RBD-antibody complexes are also utilized in our TDL-DMS models. Our models are validated through leave-one-dataset-out and 10-fold cross – validation. Finally, we demonstrate the performance of TDL-DMS models for *in silico* DMS.

Table 1
The size of each SARS-CoV-2 RBD DMS dataset.

Dataset	RBD-ACE2-1 [47]	RBD-ACE2-2 [48]	RBD-CTC-455.2 [48]	BA.1-RBD-ACE2 [66]	BA.2-RBD-ACE2 [66]
Sample size	3669	1539	2831	3800	3686

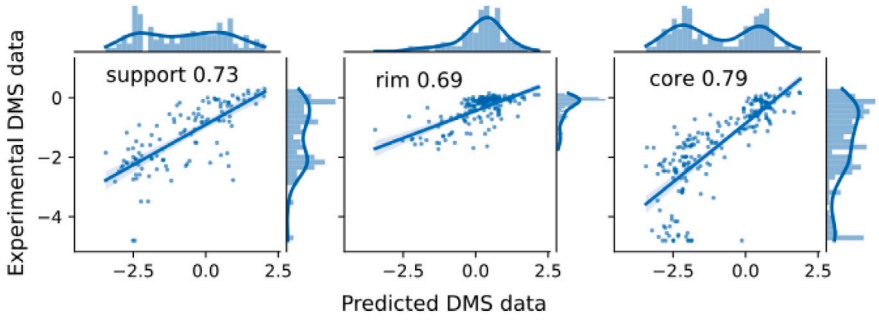

Fig. 1. Leave-one-out cross-validations of mutational scanning on the RBD in the original RBD-ACE2 complex [47] shows an R_p of 0.63. Prediction results for different residue region types according to Fig. 7 with R_ps of 0.73, 0.69, and 0.79 for the support, rim, and core, respectively. Note DMS data does not have a standard unit, and thus the values (functional scores) are compared.

2. Results

There are five SARS-CoV-2 RBD DMS datasets collected as the training set of the TDL model (see Table 1 for more details about datasets). To illustrate the performance of this proposed model, we employ dataset-level leave-one-out validations on these five SARS-CoV-2 RBD DMS datasets (10-fold cross validations are described in the Supplemental Material). Thus, the neural network model consists of six hidden layers with 15,000 neurons in each layer and generates six outputs for each dataset. For the validation process, we use five out of the six DMS datasets as the training data, with the remaining dataset serving as the test set for validation. We provide a comprehensive statistical analysis for each validation, showcasing the results in the form of scatter and histogram plots based on the mutation locations. Additionally, we include five schematic representations, with the definitions of their structural regions derived from the relative accessible surface area (rASA) [67]. Residues with rASA can be classified into structural regions such as interior and surface or interface categories like support, rim, and core, which aids in analyzing TDL-DMS predictions of the SARS-CoV-2 Spike protein RBD while accounting for continuous amino acid exposure.

2.1. DMS of the RBD in the original RBD-ACE2 complex

To guide subsequent experiments and analyses by understanding the mutational impacts on SARS-CoV-2 infectivity and antibody resistance, we initially conducted an *in silico* DMS of the RBD in the original RBD-ACE2 complex, using the dataset with experimental DMS results on SARS-CoV-2 RBD by Starr, et al. [47]. The yeast-surface-display platform was utilized to measure the expression of folded RBD protein and its binding to ACE2. Functional scores for RBD-ACE2 binding affinity were derived from per-barcode counts obtained during the experiments [47]. The dataset was released at the beginning of the pandemic and has been widely used for studying the SARS-CoV-2 RBD-ACE2 interaction and for vaccine design and antibody design. Readers interested in exploring the specifics can refer to the authors' GitHub repository for further information (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS). In our leave-one-out TDL prediction, protein structure 6M0J of RBD-ACE2 complex [68] (see Fig. 1) was used in our TDL model. There are 3669 single mutations on RBD with an overall Pearson correlation of $R_p = 0.63$ between the experimental results and predicted results. It is important to note that experimental DMS enrichment ratios were converted into binding free energies with errors, and some discrepancies were observed in the interior and surface mutations. Despite this, higher correlations were observed in the support, rim, and core regions, indicating that the TDL-DMS model performs well in predicting the binding interface of the RBD-ACE2 complex. There were a significant presence of very negative values (< -4.0) in the DMS data, while the predicted values are in the range from -5 to 1. For example, there are 101 values that were set to be -4.8. On the interior and surface mutations, the correlations are down to 0.52 and 0.57 respectively. Obviously, these mutations with very negative values belong to the interior and surface (see Figure S1). Nevertheless, correlations in the support, rim, and core regions are higher than that of others, with R_p of 0.73, 0.69, and 0.79, respectively. Therefore, the TDL-DMS model performs well on the binding interface of the RBD-ACE2 complex, which is the most relevant and important for understanding mutational impacts to SARS-CoV-2 infectivity and antibody resistance (see Fig. 2).

Our next DMS dataset of the original SARS-CoV-2 RBD was provided by Linsky et al. [48]. The authors studied the *de novo* design of hACE2 decoys to moderate SARS-CoV-2, and provide a monovalent decoy high potently neutralizing SARS-CoV-2. In the experiment, approximately 1700 single mutations were tested, while our analysis considered 1539 single mutations, limited the 6M0J RBD protein structure, which lacks residues at both ends [68]. Here, the proteins use yeast display and the enrichment of DMS data is presented for experiments. Calculation detail can be found at the Supplementary Materials of Ref. [48]. Overall, the predicted values have a correlation of 0.72 (see Figure S2). It is shown that there are two peaks in terms of population ranges for experimental DMS data, while only one peak of that for predicted results. One of the two peaks has the corresponding values around -2.5, which indicates mutations moderating the RBD-ACE2 binding (see Figure S2). Interestingly, when considering the interior mutations, which contribute mostly to the second peak, the correlation is 0.53 (see Figure S2). This difference might be caused either by the TDL-DMS model having lower performance for interior mutations or the dataset has experimental bias in certain regions. For mutations near the binding interface, i.e., support, rim and core, predictions have relatively high correlations with the experimental data. In the interface, residues have differences in their rASAs in monomer and complex, and play key roles in SARS-CoV-2 mutations. High correlations on regions suggest the TDL-DMS model has accurate predictions. The highest correlation between the experimental data and predicted results is observed for core mutations, $R_p = 0.81$ (see Fig. 2).

2.2. DMS of the RBD in a variant RBD-CTC-455.2 complex

In the same work, Linsky, et al. test the RBD binding to their *de novo* design protein CTC-445.2 and scan 1539 mutations on RBD [48]. The experiments are the same as the last one. For this dataset, the overall

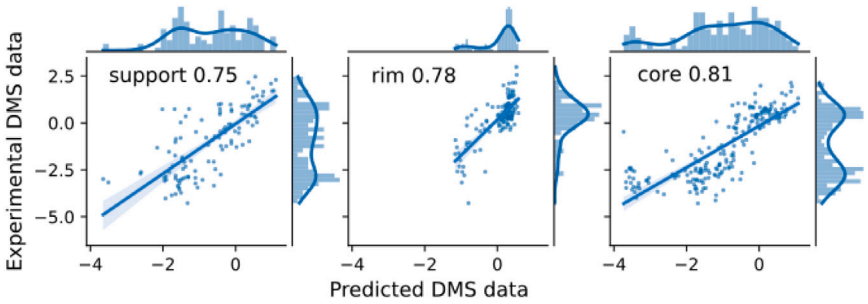

Fig. 2. Leave-one-out cross-validations of mutational scanning on the original RBD binding to ACE2 [48] shows an R_p of 0.72. Prediction results for different residue region types according to Fig. 7 with R_ps of 0.75, 0.78, and 0.81 for the binding interfaces: support, rim, and core, respectively. The average enrichment of the experimental data is compared.

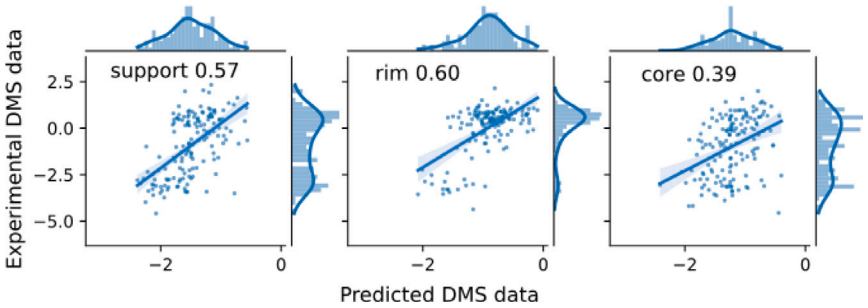

Fig. 3. Leave-one-out cross-validations of DMS on the RBD in the RBD-CTC-445.2 complex [48] shows an R_p of 0.67. Prediction results for different residue region types according to Fig. 7 with R_ps of 0.57, 0.60, and 0.39 for the support, rim, and core, respectively. The average enrichment of the experimental data is compared.

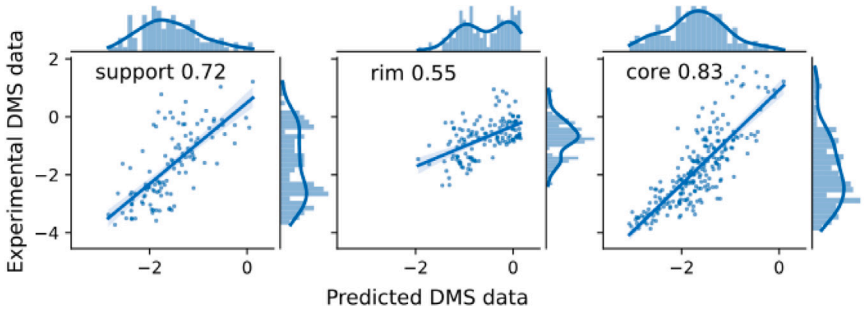

Fig. 4. Leave-one-out cross-validations of mutational scanning on BA.1 RBD binding to ACE2 [66] shows an R_p of 0.84. Prediction results for different residue region types according to Fig. 7 with R_ps of 0.72, 0.55, and 0.83 for the support, rim and core, respectively.

correlation of our TDL-DMS is 0.67. The TDL-DMS model has the worst performance in the interface for this particular dataset among the five DMS datasets, with the correlation of support, rim, and core being 0.57, 0.60, and 0.39, respectively (see Fig. 3). It is observed suboptimal performance of the TDL-DMS model, particularly in the classification of the core region, potentially due to limited data quality and quantity. Similar to the previous dataset, there are also two peaks in terms of population distribution for experimental DMS data, and only one of the peaks was predicted by TDL-DMS.

2.3. DMS of the RBD in variant RBD-ACE2 complexes

Lastly, we examine two datasets featuring distinct mutations on the RBD. Figs. 4 and 5 show the correlations of predictions versus DMS experimental data [66] of the RBD in BA.1 RDB-ACE2 complex (PDB: 7T9L [69]) and BA.2 RBD-ACE2 complex (PDB: 7XB0 [70]), respectively. The converted binding affinity of the experimental DMS data is compared with our prediction values and a yeast-surface display platform was deployed [66]. For the calculation detail of converted binding affinity, please check the author's repository (https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS_Omicron). The correlation analysis for leave-one-out cross-validations on the RBDs in BA.1 and BA.2 variants binding to ACE2 reveals a consistent Pearson correlation coefficient (R_p) of 0.84. For both RBDs, the prediction results show varying R_p values across the support, rim, and core regions (see Figs. 4 and 5). The overall correlations for both datasets are identical, and the correlations for interface mutations are notably high as well. BA.1 and BA.2 exhibit seven unique mutations on the RBD (BA.1: S371L, G446S, G496S; BA.2: S371F, T376A, D405N, R408S). Thus, when performing leave-one-out cross-validation, our proposed TDL-DMS model learns from one dataset and predicts the results for other datasets.

To show the detailed performance, we demonstrate the experimental and predicted DMSs of the RBD in the BA.2 RBD-ACE2 complex in Fig. 6. Overall, our prediction captures the general pattern very well.

3. Discussion

Firstly, residues with their rASA can be considered buried as rASA is less than a certain cutoff, which prompts the definition of two structural regions: the interior and the surface, as shown in Fig. 7. Due to the discreteness caused by the cutoff, a concern might rise as amino

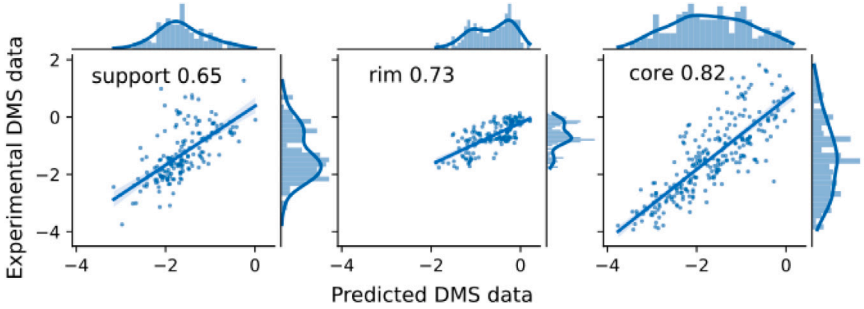

Fig. 5. Leave-one-out cross-validations of the DMS of the RBD in the BA.2 RBD-ACE2 complex [66] shows an R_p of 0.84. Prediction results for different residue region types according to Fig. 7 with R_ps of 0.65, 0.73, and 0.82 for the support, rim, and core, respectively.

Fig. 6. The comparison of experimental [66] and predicted DMSs of the RBD in the BA.2 RBD-ACE2 complex. The top panels show experimental BFE changes (kcal/mol) upon mutation converted from enrichment ratios. The predicted DMS results are given in bottom panels. In the 3D structure of the BA.2 RBD-ACE2 complex (PDB ID: 7XB0 [70]), ACE2 is in purple and RBD is in kelly green. Structures are plotted by the Illustrate [71].

acids' relative exposure is continuous. However, with the studies of *Escherichia coli*, *Saccharomyces cervisiae*, and *Homo sapiens* databases, it was concluded that the rASA cutoff distinguishing the surface and the interior easily is roughly 25% [67]. A similar concept is employed when considering the interface of protein–protein complexes. Within the same work [67], three regions of binding interfaces are defined as *support, rim*, and *core*, which require including rASA on monomer and complex (see Fig. 7). It is noted that interface residues contribute mostly to the binding energy [72]. The classification on the interface is crucial for analyzing TDL-DMS predictions of the RBD of the SARS-CoV-2 Spike protein. In the following discussion, results are analyzed in categories, i.e., interior, surface, support, rim, and core.

We present a comparison of correlations of experimental and predicted DMS values for each mutation in Fig. 8 from all five datasets. There are 361 amino acid mutation types. Among them, 20 mutation types (red color) have negative correlations of experimental and predicted DMS values, while 5 mutation types (white color) have no correlation. The rest mutation types have positive correlations,

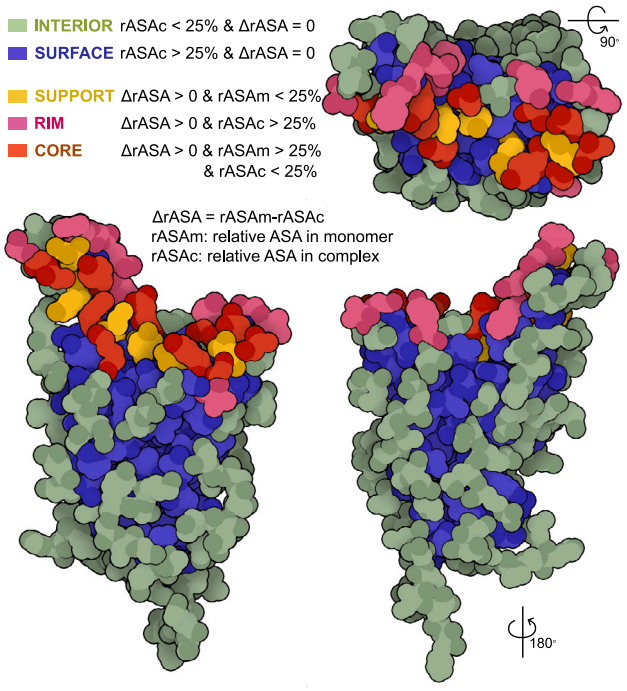

Fig. 7. The definitions of the structural regions on the Spike protein RBD (PDBID: 6M0J [68]). Amino acids are assigned to the surface, interior, support, rim, and core based on the rASA in monomer and complex. Structures are plotted by Illustrate [71].

i.e., more than half mutation types (236) have $R_p > 0.50$ and 76 mutation types have $R_p > 0.70$.

The pattern of DMS results over different mutation types is crucial for protein design, including the design of monoclonal antibodies (mAbs). We evaluate how well our TDL-DMS model predictions resemble the distribution in experimental data by examining the behavior of our model for 20 distinct amino acid types across the five DMS datasets. Remarkably, our predicted patterns align closely with the experimental data in terms of both average DMS results and variance of DMS results (see Fig. 9). The overall predictions exhibit more negative changes, indicated by a higher prevalence of deep red-colored squares. In addition to considering amino acid size, we also classify them into charged, polar, hydrophobic, and special-case groups. Regarding changes in DMS results, we observe that most mutations from charged/polar to other residues yield a positive change (e.g., mutating from K or T to others). This suggests that mutations from charged or polar residues to other types contribute to increased stability within the SARS-CoV-2 PPI system. Although our model exhibits a similar pattern in the variance of value changes as experimental data, the variance of the model predictions is generally lower, as shown in Fig. 9.

Although achieving accurate predictions with a diversity level comparable to experimental data remains a challenging task, future trends are quite clear as shown in Fig. 9. Essentially, residues K, S, and T are relatively stable. In contrast, residues R, C, I, and Y are prone to mutations. Additionally, many mutations will generate D, E, and P.

In this study, we emphasize the leave-one-dataset-out validation approach due to the unique nature of our data and the specific objectives of our research. The test data consists of multiple datasets, each representing a specific SARS-CoV-2 spike RBD associated with ACE2 and antibodies. These datasets are distinct and provide different contexts for the evaluation of the proposed model. The leave-one-dataset-out validation approach allows us to assess the generalizability of the model across these different contexts. In this strategy, it can be evaluated how well the model can adapt to new and unseen data. This validation approach provides a robust estimate of the model's performance. It reduces the risk of overfitting, as the model is tested on data that it has not seen during training. This gives us confidence that our model's good performance is not due to memorizing the training data, but rather its ability to generalize from training.

In light of the findings and challenges encountered in this study, future work will focus on refining the data collection and preprocessing methods to reduce noise and improve data quality. The quality and quantity of experimental data used for training and testing the model significantly impact machine learning performances. It is important to expand experimental datasets, particularly for regions where the model's performance was weak. Additionally, we aim to implement experimental validation as an additional check on the model's predictions. This will provide a more robust evaluation of the model's performance and help identify areas for improvement. We believe that these steps will enhance the model's predictive accuracy and contribute to the development of more effective tools for predicting DMS. Furthermore, we will continue to explore the potential of topological deep learning in the analysis of intricate and high-dimensional data. We are particularly interested in leveraging topological descriptors to simplify the structural complexity of biomolecules and embed physical interactions into topological invariants. This development will have significant implications for computational biology and complex biological systems.

4. Methods

This section provides an overview of spectral graph theory, simplicial complex, and persistent Laplacian methods for feature generation. These mathematical concepts play a crucial role in understanding the topological and spectral properties of protein–protein interactions. Additionally, machine learning and deep learning models are discussed in the context of test datasets and validation settings, highlighting their applications in the analysis and interpretation of these features. This overview aims to equip readers with the essential knowledge required for further exploration and implementation of these techniques.

4.1. Spectral graph theory

Spectral graph theory focuses on the study of graph Laplacian's spectra, connecting the algebraic connectivity and spectral properties of underlying graphs or networks. Mathematically, a graph is an ordered pair $G(V, E)$, where $V = v_i; i = 1, 2, \dots, N$ is the vertex set with size N, and $E = e_{ij} = (v_i, v_j); i \le i < j \le N$ is the edge set. Let $\deg(v)$ denote the degree of each vertex $v_i \in V$, i.e., the number of edges connected to v. A specific Laplacian matrix L^G can be given by

$$L^G = \begin{cases} \deg(v), & \text{if } v_i = v_j, \\ -1, & \text{if } v_i \text{ and } v_j \text{ are adjacent,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where "adjacent" refers to a specific definition or connection rule.

We can order the eigenvalues of the graph Laplacian matrix as

$$\lambda_{\min} = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N = \lambda_{\max}. \tag{2}$$

The kernel dimension of L^G is the multiplicity of 0 eigenvalues, indicating the number of connected components of $G(V, E)$, which is a topological property of the graph. The non-zero eigenvalues of L^G contain information about the graph properties. In particular, λ_2 is called the algebraic connectivity.

4.2. Simplicial complexes

Graph Laplacian allows only pairwise interactions (edges) and excludes high-order many-body interactions. In contrast, simplicial complexes offer a high-order generalization. Simplicial complexes serve as an elegant and robust mathematical framework for capturing the high-order interactions in graphs and networks. At the heart of this

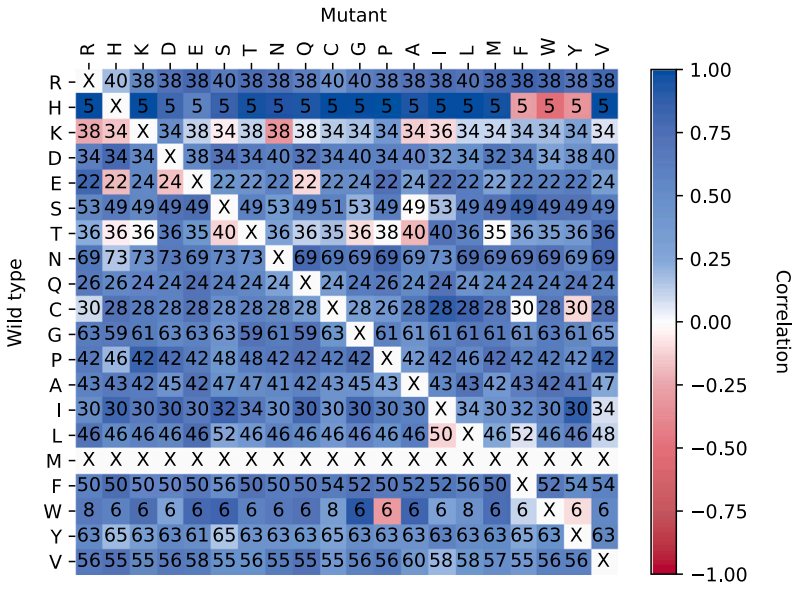

Fig. 8. A comparison of correlations of experimental and predicted DMS values following mutations associated with different amino acid types for all five datasets. Each square shows the numbers of mutations considered. Color indicates the correlation. 'X' indicates no mutation. Note that the SARS-CoV-2 RBD has no amino acid MET (M).

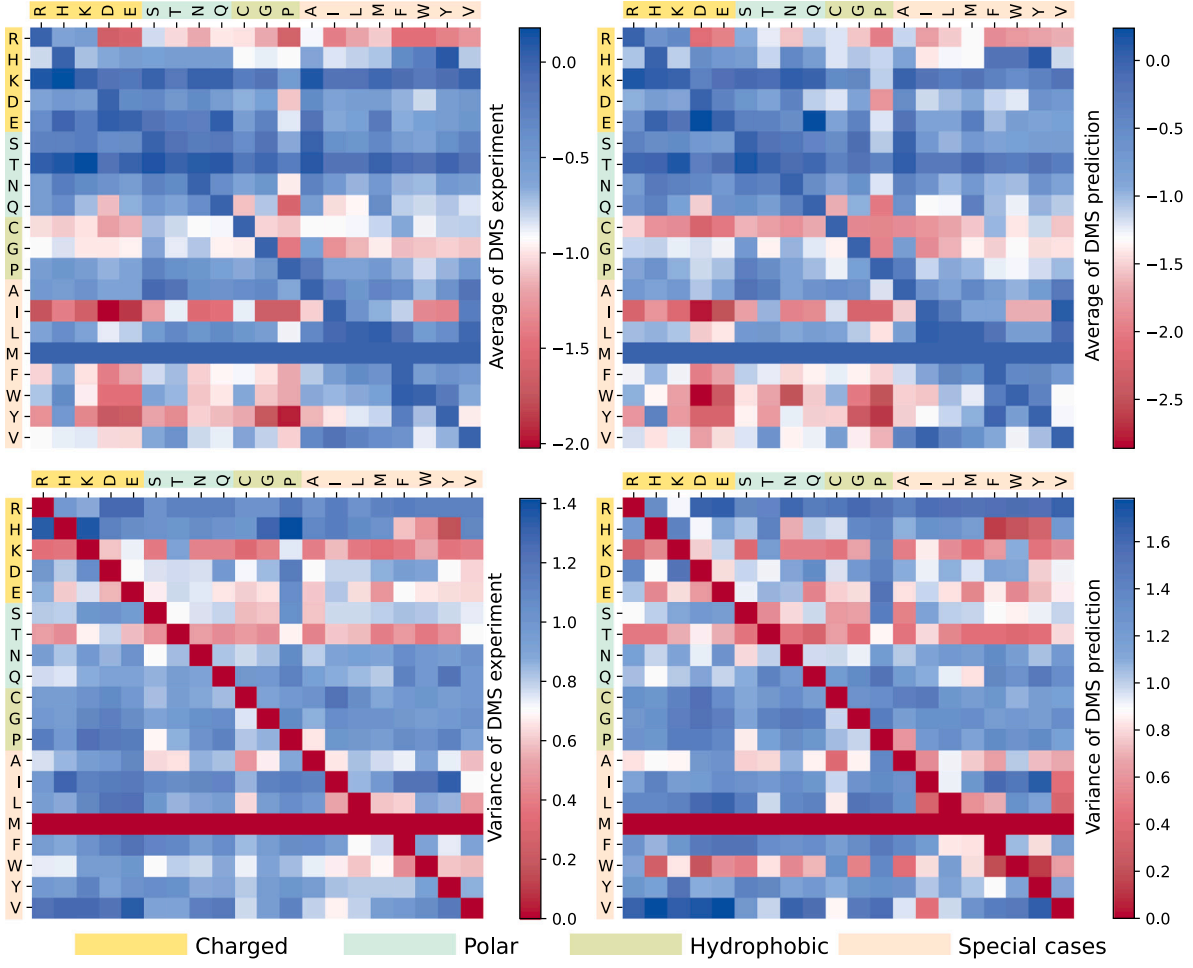

Fig. 9. A comparison of average experimental and predicted DMS values following mutations associated with different amino acid types for all the training sets. The x-axis labels the residue type of the original RBD amino acids, whereas the y-axis labels the residue type of the mutant. Note that there is no amino acid MET (M) on the RBD. **Top**: Average binding affinity changes following mutation. **Bottom**: Variance of binding affinity changes following mutation. **Left**: Experimental values. **Right**: Predicted values.

framework lies the concept of a q-simplex, which is formed from a set of $q+1$ affinely independent points. Examples of simplices encompass various geometric elements such as vertices, edges, triangles, and tetrahedrons. A simplicial complex is an assemblage of simplices that adhere

to specific conditions, and its dimension is established by the maximum dimension of its constituting simplices.

In graph theory, the degree of a vertex encapsulates the number of edges adjacent to it. However, when extending this idea to q-simplices, one must take into account both lower and upper adjacency, as q-simplices can simultaneously have $(q-1)$-simplices and $(q+1)$-simplices adjacent to them. Lower adjacency pertains to the sharing of a common $(q-1)$-face, while upper adjacency entails the sharing of a common $(q+1)$-face.

To delve deeper into the topological properties of simplicial complexes, it is useful to examine the boundary operator and chain complexes. The boundary operator, denoted as ∂_q, maps $C_q(K)$ to $C_{q-1}(K)$:

$$\partial_q \sigma_q = \sum_{i=0}^{q} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k] = \sum_{i=0}^{q} (-1)^i \sigma_{q-1}^i, \tag{3}$$

where $\hat{v}_i$ is the vertex to be excluded.

Chain complexes consist of sequences of chain groups interconnected by boundary operators:

$$\cdots \xrightarrow{\partial_{q+2}} C_{q+1}(K) \xrightarrow{\partial_{q+1}} C_q(K) \xrightarrow{\partial_q} C_{q-1}(K) \xrightarrow{\partial_{q-1}} \cdots \tag{4}$$

In essence, simplicial complexes offer an effective tool for probing the topological properties of graphs and networks. By analyzing the degrees of various simplices, their adjacencies, and the intricacies of boundary operator and chain complexes, we can gain a deeper understanding of the structure and connectivity inherent in complex systems.

4.3. Combinatorial Laplacian

In 1944, Eckmann introduced simplicial complexes into graph Laplacians, which gives rise to combinatorial Laplacian or topological Laplacian [73]. Combinatorial Laplacian was ingeniously devised to enrich the topological and geometric information inherent in simplicial complexes. Foundational concepts revolve around the oriented simplicial complex and the q-combinatorial Laplacian. Comprehensive information on these topics can be explored in the cited literature [74–77]. The subsequent discussion delves into the properties of the q-combinatorial Laplacian matrix and its associated spectra.

The q-combinatorial Laplacian is predicated on oriented simplicial complexes, which harness both lower- and higher-dimensional simplices to investigate a specifically oriented simplicial complex. An oriented simplicial complex, K, is characterized by the orientation of all its constituent simplices. When σ_q^i and σ_q^j are upper adjacent, sharing a common upper $(q+1)$-simplex τ_{q+1}, they are deemed similarly oriented if both exhibit the same sign in $\partial_{q+1}(\tau_{q+1})$, and dissimilarly oriented if the signs are contrary. Moreover, if σ_q^i and σ_q^j are lower adjacent, sharing a common lower $(q-1)$-simplex η_{q-1}, they are similarly oriented if η_{q-1} bears the same sign in both $\partial_q(\sigma_q^i)$ and $\partial_q(\sigma_q^j)$, and dissimilarly oriented if the signs are in opposition. In a similar vein, q-chains can be defined on the oriented simplicial complex K, along with the q-boundary operator.

The q-combinatorial Laplacian is a linear operator $\Delta_q : C_q(K) \longrightarrow C_q(K)$ for integers $q \geq 0$

$$\Delta_q := \partial_{q+1}\partial_{q+1}^* + \partial_q^*\partial_q \tag{5}$$

where ∂_q^* denotes the coboundary operator, mapping $\partial_q^* : C_{q-1}(K) \longrightarrow C_q(K)$. The property $\partial_q \partial_{q+1} = 0$ is preserved, implying that $\mathrm{Im}(\partial_{q+1}) \subset \ker(\partial_q)$. The matrix representation of the q-combinatorial Laplacian operator, denoted by $\mathcal{L}q$, is given by

$$\mathcal{L}q = \mathcal{B}_{q+1}\mathcal{B}_{q+1}^T + \mathcal{B}_q^T\mathcal{B}_q \tag{6}$$

where $\mathcal{B}_q$ and $\mathcal{B}q^T$ represent the matrix representations of the q-boundary operator and q-coboundary operator, respectively, in relation to the standard basis for $C_q(K)$ and $C_{q-1}(K)$ with specific orderings. Consequently, the number of rows in $\mathcal{B}_q$ corresponds to the quantity of $(q-1)$-simplices, while the number of columns reflects the quantity of q-simplices in K. Furthermore, the upper and lower q-combinatorial Laplacian matrices are denoted by $\mathcal{L}_q^U = \mathcal{B}_{q+1}\mathcal{B}_{q+1}^T$ and $\mathcal{L}_q^L = \mathcal{B}_q^T\mathcal{B}_q$, respectively. It is important to note that ∂_0 is the zero map, resulting in $\mathcal{B}_0$ being a zero matrix. Hence, $\mathcal{L}_0(K) = \mathcal{B}_1\mathcal{B}_1^T + \mathcal{B}_0^T\mathcal{B}_0$, with K representing the (oriented) simplicial complex of dimension 1, which is essentially a simple graph. In particular, the 0-combinatorial Laplacian matrix $\mathcal{L}_0(K)$ is actually the Laplacian matrix as defined in the spectral graph theory.

4.4. Persistent Laplacians

Persistent Laplacian, also known as persistent spectral graphs or persistent combinatorial Laplacian [62], has emerged as a popular tool in topological data analysis. It was proposed to overcome the limitation of persistent homology for incapable of capturing the homotopic shape evolution of the data. It is based on a filtration process that converts a data set into a sequence of nested simplicial complexes with increasing levels of complexity. In each level, the Betti numbers are calculated, and the changes in the Betti numbers are tracked as the resolution of the data set increases. These changes in the Betti numbers, called topological persistence, provide a measure of the robustness of the topological features of the data set (see Fig. 10).

In order to study the persistence of the spectral properties of graphs or simplicial complexes, one can use the notion of persistent Laplacians, which are a family of Laplacian matrices that encode the topological and geometric information of the simplicial complexes at different resolutions. The main idea is to construct a sequence of nested simplicial complexes by successively adding simplices to the complex, and associate a Laplacian matrix with each complex. By comparing the spectra of the Laplacian matrices at different resolutions, one can study the persistent spectral properties of simplicial complexes.

There are different ways to construct persistent Laplacians, depending on the filtration process and the type of Laplacian used [62]. One common approach is to use the combinatorial Laplacian matrix $\mathcal{L}_q$ of the simplicial complexes defined in the previous section. Given a sequence of nested simplicial complexes $K_0 \subset K_1 \subset \cdots \subset K_n$ with increasing dimension, one can also define a sequence of combinatorial Laplacian matrices $\mathcal{L}_{q,0}, \mathcal{L}_{q,1}, \ldots, \mathcal{L}_{q,n}$ by setting $\mathcal{L}_{q,i} = \mathcal{L}_q(K_i)$. Then, one can study the persistent spectral properties of the sequence of Laplacian matrices, such as the persistent eigenvalues and eigenvectors.

The harmonic spectra of persistent Laplacians at various scales are the same as the persistent Betti numbers, while the non-harmonic spectra can capture both topological changes and homotopic shape evolution of the data, see Fig. 11. Note that, in the figure, each of the five charts on the top panel is represented by a segment of the non-harmonic spectra, i.e., the first non-zero eigenvalues in red. In contrast, persistent homology (blue bars) does not capture homotopic shape evolution (i.e., the states in the third chart and the fifth chart). As a result, persistent Laplacians offer an enriched representation of data.

4.5. Protein–protein interactions

PPIs are analyzed by topological and shape analysis. We initially partition the atoms in a protein–protein complex into several subsets:

1. $\mathcal{A}_\mathrm{m}$: atoms at the mutation sites.
2. $\mathcal{A}_\mathrm{mn}(r)$: atoms in the vicinity of the mutation site, within a cut-off distance r.
3. $\mathcal{A}_\mathrm{A}(r)$: protein A atoms within r of the binding site.
4. $\mathcal{A}_\mathrm{B}(r)$: protein B atoms within r of the binding site.

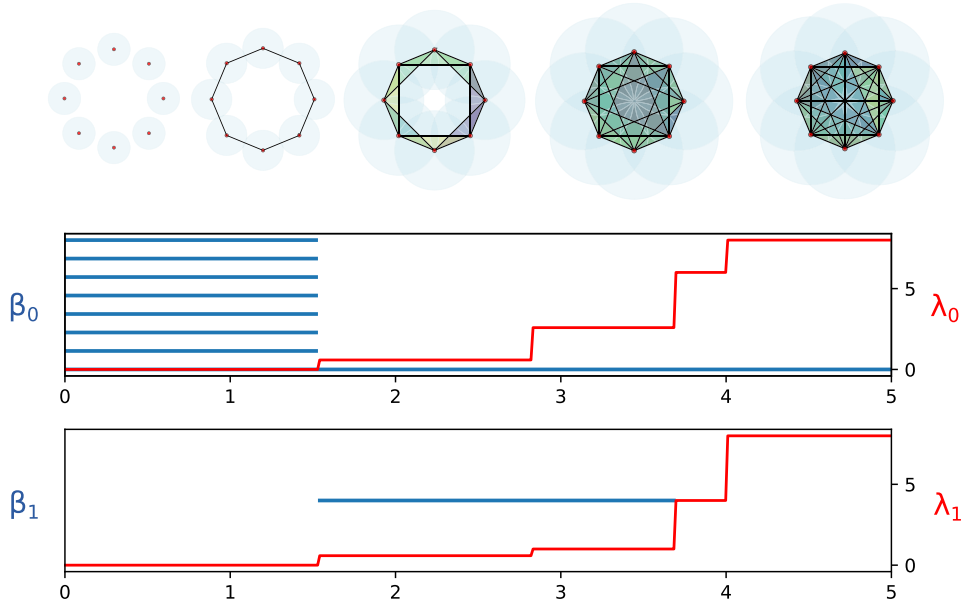

Fig. 10. Comparison of persistent homology (PH) [55,56] and persistent Laplacians (PLs) [62] for eight points. The filtration characterized by the horizontal axis r of eight points is shown in the top panel. The corresponding topological features of dimension 0 and dimension 1 are shown the second and third panels, respectively. PH barcodes ($\beta_0(r)$ and $\beta_1(r)$) are given in blue. 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.

5. $\mathcal{A}_{\text{ele}}$(E): atoms of element type E within the system. We design the distance matrix to exclude interactions between atoms from the same set. For interactions between atoms a_i and a_j in sets $\mathcal{A}$ and/or $\mathcal{B}$, we define the modified distance as follows:

$$D_{\text{mod}}(a_i, a_j) = \begin{cases} \infty, \text{ if } a_i, a_j \in \mathcal{A}, \text{ or } a_i, a_j \in \mathcal{B}, \\ D_e(a_i, a_j), \text{ if } a_i \in \mathcal{A} \text{ and } a_j \in \mathcal{B}, \end{cases} \tag{7}$$

where $D_e(a_i, a_j)$ represents the Euclidean distance between a_i and a_j. Molecular atoms are constructed as points, denoted by v_0, v_1, v_2, ..., v_k, with k!+!1 affinely independent points in a simplicial complex. Persistent spectral graphs are designed to capture multiscale topological and geometrical information across different scales along a filtration [62], yielding essential feature vectors for machine learning methods. Binned barcode vectorization-generated features can represent the strength of atomic bonds and van der Waals interactions, and are readily incorporated into machine learning models that discern and characterize local patterns.

Using atom subsets, such as $\mathcal{A}_{\text{A}}(r)$ and $\mathcal{A}_{\text{B}}(r)$, we create simplicial complexes by considering only the edges from $\mathcal{A}_{\text{A}}(r)$ to $\mathcal{A}_{\text{B}}(r)$ for Vietoris–Rips complexes. Barcodes generated from persistent homology are then enumerated by bar lengths within specific intervals, with numbers 0 or 1, as part of the Vietoris–Rips complex filtration. Concurrently, for each complex in the filtration, we compute eigenvalues using graph Laplacian analysis. We gather statistics of eigenvalues, such as sum, maximum, minimum, mean, and standard deviation, to obtain normalized features for machine learning methods. An alternative vectorization approach involves extracting statistics of bar lengths, birth values, and death values, including sum, maximum, minimum, mean, and standard deviation. This technique is applied to vectorize Betti-1 (H_1) and Betti-2 (H_2) barcodes obtained from alpha complex filtration, based on the observation that higher-dimensional barcodes are sparser than H_0 barcodes.

In summary, this methodology integrates topological representations and persistent Laplacian spectra to analyze protein–protein interactions. By categorizing atoms in a protein–protein complex into subsets, we can construct simplicial complexes and generate feature vectors for machine learning algorithms. This approach effectively captures the essential topological and geometrical information of the underlying molecular structures, facilitating the study of protein–protein interactions and their biological implications.

4.6. Machine learning

The features generated from the persistent spectral graph are tested using the deep neural network (Net) method. Validations are performed on the datasets discussed in the results section. Accurately predicting mutation-induced binding affinity changes in protein–protein complexes is a significant challenge. After generating effective features, machine learning or deep learning models are required for validation and real-world applications. A deep neural network is a network of neurons that maps an input feature layer to an output layer. The neural network mimics the human brain to solve problems with numerous neuron units and employs backpropagation to update weights on each layer. To capture input features at different levels and abstract more properties, one can construct more layers and more neurons in each layer, creating a deep neural network. Optimization methods for feed-forward neural networks and dropout methods are applied to prevent overfitting. The network layers and the number of neurons in each layer are determined by grid searches based on 10-fold cross-validations. Then, the hyperparameters of stochastic gradient descent (SGD) with momentum are set up based on the network structure. The network has 7 layers with 10,000 neurons in each layer. For SGD with momentum, the hyperparameters are `momentum = 0.9` and `weight_decay = 0`. The learning rate is 0.002 and the epoch is 400. The Net is implemented using Pytorch [78].

Fig. 11 provides the workflow of the proposed TDL-DMS methodology. The input is a protein–protein complex, and the output is the predicted DMS (the heatmap on the left). The protein–protein complex is partitioned into subsets, and simplicial complexes are constructed using the Vietoris–Rips complex and filtration. Barcodes are generated from persistent homology, and eigenvalues are computed from persistent graph Laplacians. The barcodes and eigenvalues are used to generate feature vectors. The feature vectors are then used as the inputs for the deep learning network to predict the binding affinity changes of mutations in protein–protein complexes. The model is trained with the experimental DMS data as the ground truth (the heatmap on the right).

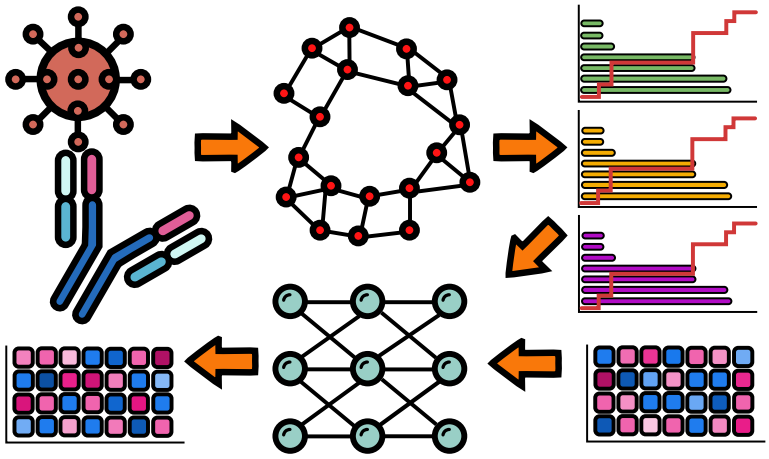

Fig. 11. Illustration of the proposed TDL-DMS methodology. The input is a protein–protein complex, and the output is the predicted DMS (the heatmap on the left). The protein–protein complex is partitioned into subsets, and simplicial complexes are constructed using the Vietoris–Rips complex and filtration. Barcodes are generated from persistent homology and eigenvalues are computed persistent Laplacians. The barcodes and eigenvalues are used to generate feature vectors for deep learning. The heatmap on the right is the training data.

5. Conclusion

Deep mutational scanning (DMS) is a high-throughput experimental technique that enables the systematic analysis of the impact of mutations on protein function, providing insights into the structure-function relationships and evolutionary trends and constraints of proteins. DMS has been successfully applied to a wide range of biological systems, including enzymes, receptors, transcription factors, and viruses. It can be used to design proteins with improved properties, identify drug targets and inhibitors, and understand the mechanisms of protein evolution and adaptation. However, the mutational space of a typical protein is astronomically large and intractable for experimental means.

Computational approaches to DMS offer viable alternatives, although *in silico* DMS has hardly been reported. One challenge is the lack of accurate and reliable biophysical models for dealing with complex protein functions and protein–protein interactions (PPIs). Another challenge is the lack of high-quality DMS data for data-driven machine learning predictions.

Currently, it is well-understood that the SARS-CoV-2 spike protein plays the most important role in viral transmission, and its receptor-binding domain (RBD) binds to human ACE2 to facilitate viral entry into host cells. Emerging SARS-CoV-2 variants are spreading worldwide with increased transmissibility due to the natural selection of RBD mutations with higher infectivity [4] and/or stronger antibody resistance [41]. As a result, researchers have conducted various DMS studies on the original spike RBD and variant RBD in recent years [47,48,66]. This development enables the artificial intelligence (AI)-based prediction of DMS of future SARS-CoV-2 variants.

Topological deep learning (TDL) has led to the discovery of two SARS-CoV-2 evolutionary mechanisms [4,41] and accurate forecasting of future dominant SARS-CoV-2 variants Omicron [79], Omicron BA.2 [80], and Omicron BA.5 [61]. Recently, a new generation of topological data analysis (TDA) techniques was proposed [62] and implemented for SARS-CoV-2 variant prediction [61]. Built on these experimental, mathematical, and computational advances, we develop our TDL-DMS model for SARS-CoV-2 RBDs.

We performed leave-one-dataset-out validation on the proposed TDL-DMS on five datasets involving various SARS-CoV-2 spike RBDs associated with ACE2 and antibodies. We found that our TDL-DMS model works well in general and offers excellent DMS predictions for RBD binding interface mutations, which are particularly important in forecasting future dominant SARS-CoV-2 variants.

We expect the proposed TDL-DMS framework to have potential applications in protein engineering, drug discovery, and directed evolution.

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.

Acknowledgments

This work was supported in part by NIH, United States grants R01GM126189, R01AI164266, and R01AI146210, NSF grants DMS-2052983, DMS-1761320, and IIS-1900473, NASA, United States grant 80NSSC21M0023, MSU Foundation, Bristol-Myers Squibb, United States 65109, and Pfizer, United States. JC thanks Dr. Daniel-Adriano Silva for the assistance in converting the experimental enrichment ratios and BFE changes.

Code availability

The source codes are available at https://github.com/WeilabMSU/TopNetDMS.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.107258.

References

[1] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, et al., SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor, Cell 181 (2) (2020) 271–280.

[2] Ruiqiong Guo, Kristen Gaffney, Zhongyu Yang, Miyeon Kim, Suttipun Sungsuwan, Xuefei Huang, Wayne L Hubbell, Heedeok Hong, Steric trapping reveals a cooperativity network in the intramembrane protease GlpG, Nat. chem. biol. 12 (5) (2016) 353–360.

[3] Raphael Guerois, Jens Erik Nielsen, Luis 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.

[4] Jiahui Chen, Rui Wang, Menglun Wang, Guo-Wei Wei, Mutations strengthened SARS-CoV-2 infectivity, J. Mol. Biol. 432 (19) (2020) 5212–5226.

[5] Jiahui Chen, Kaifu Gao, Rui Wang, Guo-Wei Wei, Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies, Chem. Sci. 12 (20) (2021) 6929–6948.

[6] Emidio Capriotti, Piero Fariselli, Rita Casadio, I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure, Nucleic acids res. 33 (suppl_2) (2005) W306–W310.

[7] Catherine L. Worth, Robert Preissner, Tom L. Blundell, SDM—a server for predicting effects of mutations on protein stability and malfunction, Nucleic acids res. 39 (suppl_2) (2011) W215–W222.

[8] Douglas E.V. Pires, David B. Ascher, Tom L. Blundell, DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach, Nucleic acids res. 42 (W1) (2014) W314–W319.

[9] Yves Dehouck, Aline Grosfils, Benjamin Folch, Dimitri Gilis, Philippe Bogaerts, Marianne Rooman, Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0, Bioinformatics 25 (19) (2009) 2537–2543.

[10] Elizabeth H. Kellogg, Andrew Leaver-Fay, David Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability, Proteins: Struct. Funct. Bioinform. 79 (3) (2011) 830–838.

[11] Ivan Getov, Marharyta Petukh, Emil Alexov, 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 (4) (2016) 512.

[12] Yang Yang, Biao Chen, Ge Tan, Mauno Vihinen, Bairong Shen, Structure-based prediction of the effects of a missense variant on protein stability, Amino Acids 44 (3) (2013) 847–855.

[13] Yongwook Choi, Gregory E Sims, Sean Murphy, Jason R Miller, Agnes P Chan, Predicting the functional effect of amino acid substitutions and indels, Public Library of Science San Francisco, USA, 2012.

[14] Niklas Berliner, Joan Teyra, Recep Colak, Sebastian Garcia Lopez, Philip M Kim, Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation, PLoS One 9 (9) (2014) e107353.

[15] Lijun Quan, Qiang Lv, Yang Zhang, STRUM: structure-based prediction of protein stability changes upon single-point mutation, Bioinformatics 32 (19) (2016) 2936–2946.

[16] Lukas Folkman, Bela Stantic, Abdul Sattar, Yaoqi Zhou, EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models, J. Mol. Biol. 428 (6) (2016) 1394–1405.

[17] Alexey Strokach, Carles Corbi-Verge, Philip M. Kim, Predicting changes in protein stability caused by mutation using sequence-and structure-based methods in a CAGI5 blind challenge, Hum. mutat. 40 (9) (2019) 1414–1423.

[18] C.H.I. Zhang, Song Liu, Yaoqi Zhou, Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential, Prot. sci. 13 (2) (2004) 391–399.

[19] Dassault Systèmes Biovia, et al., Discovery studio modeling environment, 2017.

[20] Navin Pokala, Tracy M. Handel, Energy functions for protein design: adjustment with protein–protein complex affinities, models for the unfolded state, and negative design of solubility and specificity, J. Mol. Biol. 347 (1) (2005) 203–227.

[21] Alexander Benedix, Caroline M Becker, Bert L de Groot, Amedeo Caflisch, Rainer A Böckmann, Predicting free energy changes using structural ensembles, Nat. methods 6 (1) (2009) 3–4.

[22] Kyle A Barlow, Shane O Conchuir, Samuel Thompson, Pooja Suresh, James E Lucas, Markus Heinonen, Tanja Kortemme, Flex ddg: Rosetta ensemble-based estimation of changes in protein–protein binding affinity upon mutation, J. Phys. Chem. B 122 (21) (2018) 5389–5399.

[23] Yves Dehouck, Jean Marc Kwasigroch, Marianne Rooman, Dimitri Gilis, BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations, Nucleic acids res. 41 (W1) (2013) W333–W339.

[24] Douglas E.V. Pires, David B. Ascher, mCSM-AB: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures, Nucleic acids res. 44 (W1) (2016) W469–W473.

[25] Carlos HM Rodrigues, Yoochan Myung, Douglas EV Pires, David B Ascher, mCSM-PPI2: predicting the effects of mutations on protein–protein interactions, Nucleic acids res. 47 (W1) (2019) W338–W344.

[26] Vladimir Potapov, Mati Cohen, Gideon Schreiber, Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details, Protein eng. des. sel. 22 (9) (2009) 553–560.

[27] Sarah Sirin, James R Apgar, Eric M Bennett, Amy E Keating, AB-Bind: antibody binding mutational database for computational affinity predictions, Prot. Sci. 25 (2) (2016) 393–409.

[28] Thomas Steinbrecher, Andreas Labahn, Towards accurate free energy calculations in ligand protein-binding studies, Curr. med. chem. 17 (8) (2010) 767–785.

[29] Gregory King, Arieh Warshel, Investigation of the free energy functions for electron transfer reactions, J. Chem. Phys. 93 (12) (1990) 8682–8692.

[30] Ehecatl Antonio Del Rio-Chanona, Nur Rashid Ahmed, Jonathan Wagner, Yinghua Lu, Dongda Zhang, Keju Jing, Comparison of physics-based and data-driven modelling techniques for dynamic optimisation of fed-batch bioprocesses, Biotechnol. bioeng. 116 (11) (2019) 2971–2982.

[31] Yuchi Qiu, Guo-Wei Wei, Persistent spectral theory-guided protein engineering, Nat. Comput. Sci. 3 (2) (2023) 149–163.

[32] Bo-Wei Zhao, Lei Wang, Peng-Wei Hu, Leon Wong, Xiao-Rui Su, Bao-Quan Wang, Zhu-Hong You, Lun Hu, Fusing higher and lower-order biological information for drug repositioning via graph representation learning, IEEE Trans. Emerg. Top. Comput. (2023).

[33] Xiaorui Su, Pengwei Hu, Haicheng Yi, Zhuhong You, Lun Hu, Predicting drug-target interactions over heterogeneous information network, IEEE J. Biomed. Health Inf. 27 (1) (2022) 562–572.

[34] Hao Wu, Zhongli Chen, Yingfu Wu, Hongming Zhang, Quanzhong Liu, Integrating protein–protein interaction networks and somatic mutation data to detect driver modules in pan-cancer, Interdiscip. Sci.: Comput. Life Sci. (2021) 1–17.

[35] Jinxiang Chen, Miao Wang, Defeng Zhao, Fuyi Li, Hao Wu, Quanzhong Liu, Shuqin Li, MSINGB: A novel computational method based on ngboost for identifying microsatellite instability status from tumor mutation annotation data, Interdiscip. Sci.: Comput. Life Sci. 15 (1) (2023) 100–110.

[36] Douglas M. Fowler, Stanley Fields, Deep mutational scanning: a new style of protein science, Nat. methods 11 (8) (2014) 801–807.

[37] Carlos L. Araya, Douglas M. Fowler, Deep mutational scanning: assessing protein function on a massive scale, Trends Biotechnol. 29 (9) (2011) 435–442.

[38] Molly Gasperini, Lea Starita, Jay Shendure, The power of multiplexed functional analysis of genetic variants, Nat. Protoc. 11 (10) (2016) 1782–1787.

[39] Vanessa E Gray, Ronald J Hause, Jens Luebeck, Jay Shendure, Douglas M Fowler, Quantitative missense variant effect prediction using large-scale mutagenesis data, Cell systems 6 (1) (2018) 116–124.

[40] Hagit Sarfati, Si Naftaly, Niv Papo, Chen Keasar, Predicting mutant outcome by combining deep mutational scanning and machine learning, Proteins: Struct. Funct. Bioinform. 90 (1) (2022) 45–57.

[41] Rui Wang, Jiahui Chen, Guo-Wei Wei, Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America, J. Phys. Chem. Lett. 12 (2021) 11850–11857.

[42] Kaiming Tao, Philip L Tzou, Janin Nouhin, Ravindra K Gupta, Tulio de Oliveira, Sergei L Kosakovsky Pond, Daniela Fera, Robert W Shafer, The biological and clinical significance of emerging SARS-CoV-2 variants, Nature Rev. Genet. 22 (12) (2021) 757–773.

[43] Wendong Li, Zhengli Shi, Meng Yu, Wuze Ren, Craig Smith, Jonathan H Epstein, Hanzhong Wang, Gary Crameri, Zhihong Hu, Huajun Zhang, et al., Bats are natural reservoirs of SARS-like coronaviruses, Science 310 (5748) (2005) 676–679.

[44] Xiu-Xia Qu, Pei Hao, Xi-Jun Song, Si-Ming Jiang, Yan-Xia Liu, Pei-Gang Wang, Xi Rao, Huai-Dong Song, Sheng-Yue Wang, Yu Zuo, et al., Identification of two critical amino acid residues of the severe acute respiratory syndrome coronavirus spike protein for its variation in zoonotic tropism transition via a double substitution strategy, J. Biol. Chem. 280 (33) (2005) 29588–29595.

[45] Huai-Dong Song, Chang-Chun Tu, Guo-Wei Zhang, Sheng-Yue Wang, Kui Zheng, Lian-Cheng Lei, Qiu-Xia Chen, Yu-Wei Gao, Hui-Qiong Zhou, Hua Xiang, et al., Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human, Proc. Natl. Acad. Sci. 102 (7) (2005) 2430–2435.

[46] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, David Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell (2020).

[47] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine HD Crawford, Adam S Dingens, Mary Jane Navarro, John E Bowen, M Alejandra Tortorici, Alexandra C Walls, et al., Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding, Cell 182 (5) (2020) 1295–1310.

[48] Thomas W Linsky, Renan Vergara, Nuria Codina, Jorgen W Nelson, Matthew J Walker, Wen Su, Christopher O Barnes, Tien-Ying Hsiang, Katharina Esser-Nobis, Kevin Yu, et al., De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2, Science 370 (6521) (2020) 1208–1214.

[49] Erik Procko, The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2, BioRxiv (2020).

[50] Tyler N Starr, Allison J Greaney, William W Hannon, Andrea N Loes, Kevin Hauser, Josh R Dillen, Elena Ferri, Ariana Ghez Farrell, Bernadeta Dadonaite, Matthew McCallum, et al., Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution, BioRxiv (2022).

[51] Longxing Cao, Inna Goreshnik, Brian Coventry, James Brett Case, Lauren Miller, Lisa Kozodoy, Rita E Chen, Lauren Carter, Alexandra C Walls, Young-Jun Park, et al., De novo design of picomolar SARS-CoV-2 miniprotein inhibitors, Science 370 (6515) (2020) 426–431.

[52] Allison J Greaney, Tyler N Starr, Pavlo Gilchuk, Seth J Zost, Elad Binshtein, Andrea N Loes, Sarah K Hilton, John Huddleston, Rachel Eguia, Katharine HD Crawford, et al., Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition, Cell host microbe 29 (1) (2021) 44–57.

[53] Alison C Leonard, Jonathan J Weinstein, Paul J Steiner, Annette H Erbse, Sarel J Fleishman, Timothy A Whitehead, Stabilization of the SARS-CoV-2 receptor binding domain by protein core redesign and deep mutational scanning, Protein Eng. Des. Select. 35 (2022).

[54] Zixuan Cang, Guo-Wei Wei, Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, PLoS Comput. Biol. 13 (7) (2017) e1005690.

[55] Herbert Edelsbrunner, John Harer, et al., Persistent homology-a survey, Contemp. math. 453 (26) (2008) 257–282.

[56] Afra Zomorodian, Gunnar Carlsson, Computing persistent homology, in: Proceedings of the Twentieth Annual Symposium On Computational Geometry, 2004, pp. 347–356.

[57] Jacob Townsend, Cassie Putman Micucci, John H Hymel, Vasileios Maroulas, Konstantinos D Vogiatzis, Representation of molecular structures with persistent homology for machine learning applications in chemistry, Nat. commun. 11 (1) (2020) 3230.
[58] Zhenyu Meng, Kelin Xia, Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction, Sci. adv. 7 (19) (2021) eabc5329.
[59] Marcio Gameiro, Yasuaki Hiraoka, Shunsuke Izumi, Miroslav Kramar, Konstantin Mischaikow, Vidit Nanda, A topological measurement of protein compressibility, Japan J. Ind. Appl. Math. 32 (2015) 1–17.
[60] Menglun Wang, Zixuan Cang, Guo-Wei 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.
[61] Jiahui Chen, Yuchi Qiu, Rui Wang, Guo-Wei Wei, Persistent laplacian projected omicron BA. 4 and BA. 5 to become new dominating variants, Comput. Biol. Med. 151 (2022) 106262.
[62] Rui Wang, Duc Duy Nguyen, Guo-Wei Wei, Persistent spectral graph, Int. j. numer. methods biomed. eng. 36 (9) (2020) e3376.
[63] Rui Wang, Guo-Wei Wei, Persistent path laplacian, Found. Data Sci. 5 (2023) 26–55.
[64] Xiaoqi Wei, Guo-Wei Wei, Persistent sheaf laplacians, 2021, arXiv preprint arXiv:2112.10906.
[65] Dong Chen, Jian Liu, Jie Wu, Guo-Wei Wei, Persistent hyperdigraph homology and persistent hyperdigraph laplacians, 2023, arXiv preprint arXiv:2304.00345.
[66] Tyler N Starr, Allison J Greaney, Cameron M Stewart, Alexandra C Walls, William W Hannon, David Veesler, Jesse D Bloom, Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 omicron BA. 1 and BA. 2 receptor-binding domains, PLoS pathog. 18 (11) (2022) e1010951.
[67] Emmanuel 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.
[68] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, et al., Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor, Nature 581 (7807) (2020) 215–220.
[69] Dhiraj Mannar, James W Saville, Xing Zhu, Shanti S Srivastava, Alison M Berezuk, Katharine S Tuttle, Ana Citlali Marquez, Inna Sekirov, Sriram Subramaniam, SARS-CoV-2 omicron variant: Antibody evasion and cryo-EM structure of spike protein–ACE2 complex, Science 375 (6582) (2022) 760–764.
[70] Linjie Li, Hanyi Liao, Yumin Meng, Weiwei Li, Pengcheng Han, Kefang Liu, Qing Wang, Dedong Li, Yanfang Zhang, Liang Wang, et al., Structural basis of human ACE2 higher binding affinity to currently circulating omicron SARS-CoV-2 sub-variants BA. 2 and BA. 1.1, Cell 185 (16) (2022) 2952–2960.
[71] David S. Goodsell, Ludovic Autin, Arthur J. Olson, Illustrate: software for biomolecular illustration, Structure 27 (11) (2019) 1716–1720.
[72] Andrew A. Bogan, Kurt S. Thorn, Anatomy of hot spots in protein interfaces, J. Mol. Biol. 280 (1) (1998) 1–9.
[73] Beno Eckmann, Harmonische funktionen und randwertaufgaben in einem komplex, Comment. Math. Helv. 17 (1) (1944) 240–255.
[74] Daniel Hernández Serrano, Darío Sánchez Gómez, Higher order degree in simplicial complexes, multi combinatorial laplacian and applications of tda to complex networks, 2019, arXiv preprint arXiv:1908.02583.
[75] Slobodan Maletić, Milan Rajković, Consensus formation on a simplicial complex of opinions, Physica A 397 (March) (2014) 111–120.
[76] Timothy E. Goldberg, Combinatorial Laplacians of Simplicial Complexes, Senior Thesis, Bard College, 2002.
[77] Danijela Horak, Jürgen Jost, Spectra of combinatorial laplace operators on simplicial complexes, Adv. Math. 244 (2013) 303–336.
[78] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. neural inf. process. syst. 32 (2019).
[79] Jiahui Chen, Rui Wang, Nancy Benovich Gilby, Guo-Wei Wei, Omicron variant (b. 1.1. 529): Infectivity, vaccine breakthrough, and antibody resistance, J. Chem. Inf. Model. 62 (2) (2022) 412–422.
[80] Jiahui Chen, Guo-Wei Wei, Omicron BA. 2 (b. 1.1. 529.2): High potential for becoming the next dominant variant, J. Phys. Chem. Lett. 13 (2022) 3840–3849.