Persistent Laplacian projected Omicron BA.4 and BA.5 to become new dominating variants

Jiahui Chen [a], Yuchi Qiu [a], Rui Wang [a], Guo-Wei Wei [a,b,c,*]

[a] *Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA*
[b] *Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824, USA*
[c] *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA*

ARTICLE INFO

ABSTRACT

Due to its high transmissibility, Omicron BA.1 ousted the Delta variant to become a dominating variant in late 2021 and was replaced by more transmissible Omicron BA.2 in March 2022. An important question is which new variants will dominate in the future. Topology-based deep learning models have had tremendous success in forecasting emerging variants in the past. However, topology is insensitive to homotopic shape evolution in virus–human protein–protein binding, which is crucial to viral evolution and transmission. This challenge is tackled with persistent Laplacian, which is able to capture both the topological change and homotopic shape evolution of data. Persistent Laplacian-based deep learning models are developed to systematically evaluate variant infectivity. Our comparative analysis of Alpha, Beta, Gamma, Delta, Lambda, Mu, and Omicron BA.1, BA.1.1, BA.2, BA.2.11, BA.2.12.1, BA.3, BA.4, and BA.5 unveils that Omicron BA.2.11, BA.2.12.1, BA.3, BA.4, and BA.5 are more contagious than BA.2. In particular, BA.4 and BA.5 are about 36% more infectious than BA.2 and are projected to become new dominant variants by natural selection. Moreover, the proposed models outperform the state-of-the-art methods on three major benchmark datasets for mutation-induced protein–protein binding free energy changes. Our key projection about BA4 and BA.5's dominance made on May 1, 2022 (see arXiv:2205.00532) became a reality in late June 2022.

1. Introduction

The coronavirus disease, 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has lasted for more than years. The development of effective vaccines, monoclonal antibodies (mABs), and antiviral drugs have significantly improved our ability to bring COVID-19 pandemic under control. Nonetheless, the emerging SARS-CoV-2 variants have become a major threat to existing vaccines, monoclonal antibodies (mABs), and antiviral drugs.

The Omicron variant has mutations on various SARS-CoV-2 proteins, such as non-structure protein 3 (NSP3), NSP4, NSP5, NSP6, NSP12, NSP14, spike (S) protein, envelope protein, membrane protein, and nucleocapsid protein. Specifically, Omicron has three main lineages, BA.1 (B.1.1.529.1), BA.2 (B.1.1.529.2), and BA.3 (B.1.1.529.3), and many sub-lineages. Many new recombinants occurred, including XD, XE, and XF. XD and XE are recombination of Delta and BA.1, while XE is basically a BA.2 Omicron lineage carrying a piece of BA.1 at the front end of its genome. The S protein of XE is still BA.2.

The research community focuses its attention on the mutations at the S protein receptor-binding domain (RBD) due to the fact that the RBD facilitates the binding between the S protein and the host angiotensin-converting enzyme 2 (ACE2), which initiates the viral entry of a host cell and infection. It turns out that the binding strength between the S protein RBD and the ACE2 is proportional to the viral infectivity [1–5]. An artificial intelligence (AI) study revealed that natural selection is the governing mechanism for SARS-CoV-2 evolution [6]. Specifically, viral evolution selects those mutations that are able to strengthen the RBD-ACE2 binding. This mechanism led to the occurrence of many variants, such as Alpha, Beta, Gamma, Delta, Mu, etc. Natural selection in SARS-CoV-2 mutations was confirmed beyond doubt in April 2021 by the genotyping of over half a million viral genomes isolated from patients [7].

Additionally, antibodies are generated by the human immune response to infection or vaccination. A strong RBD-antibody binding would lock off RBD-ACE2 binding and directly neutralize the virus [8–10]. As such, mABs targeting the S protein, particularly the RBD, which are designed to treat viral infection. It was unveiled that viral evolution also selects those mutations that are able to weaken RBD-antibody binding, leading to vaccine breakthrough infections [11,12].

Therefore, a new virus with RBD mutations that make the virus more infectious and more capable of evading antibody protection would become the next dominant variant, which is the underlying principle for the successful forecasting of Omicron BA.2's dominance [13].

In biophysics, the strength of protein–protein complex is measured by binding free energy (BFE). Mutation-induced BFE change $\Delta\Delta G$ is calculated by

$$\Delta\Delta G = \Delta G_{WT} - \Delta G_{MT} \tag{1}$$

where ΔG_{WT} and ΔG_{MT} are the BFE of wild type and mutant. A positive (negative) BFE change indicates the strengthening (weakening) of the protein–protein binding. Protein–protein BFE changes can be carried out in a variety of ways as shown in software packages FOLDX [14], SAAMBE [15], mCSM-AB [16], mCSM-PPI2 [17], BindProfX [18], etc. AI approaches take the advantage of existing data and often outperform other methods when experimental data become available. Due to the structural complexity and high dimensionality of protein–protein interactions (PPIs), methods that are able to effectively reduce the PPI structural complexity and dimensionality have demonstrated great advantages in predicting PPI BFE changes [19]. Advanced mathematics, particularly, persistent homology [20–25], offers tremendous abstraction of PPIs. Persistent homology is the main workhorse in popular topological data analysis (TDA) [26–29]. Element-specific persistent homology (EPH) has had tremendous success in computational biology [30,31] and worldwide competitions in computer-aided drug design [32].

Based on FPH, a topology-based network tree (TopNetTree) model was constructed from conventional neural network and decision trees for predicting PPI BFE changes [19]. In the past two years, this approach has been extended with SARS-CoV-2 related deep mutational data to predict the BFE changes RBD-ACE2 and RBD-antibody complexes up on RBD mutations [33,34]. Initially, in early 2020, TopNetTree model was applied to successfully predict that RBD residues 452 and 501 "have high chances to mutate into significantly more infectious COVID-19 strains" [6]. These RBD mutations later appeared in all major variants, Alpha, Beta, Delta, Gamma, Delta, Epsilon, Theta, Kappa, Lambada, Mu, and Omicron L452R/Q and N501Y mutations. In April 2021, the TopNetTree model predicted a list of 31 RBD antibody-escape mutations, including W353R, I401N, Y449D, Y449S, P491R, P491L, Q493P, etc. [7]. Notably, experimental results confirmed that mutations at RBD residues Y449, E484, Q493, S494, and Y505 enable the virus to escape antibodies [35]. It was revealed that variants found in the United Kingdom and South Africa in late 2020 would strengthen virus infectivity, which is consistent with the experimental results [36]. In summer 2021, a topology-based deep neural network trained with mAbs (TopNetmAb) was developed to forecast a list of most likely vaccine-escape RBD mutations, such as S494P, Q493L, K417N, F490S, F486L, R403K, E484K, L452R, K417T, F490L, E484Q, and A475S [34], and mutations S494P, K417N, E484K/Q, and L452R were designated as the variants of concern or variants of interest denounced by the Worldwide Health Organization (WHO). The correlation between the experimental deep mutational data [37] and AI-predicted RBD-mutation-induced BFE changes for all possible 3686 RBD mutations on the RBD-ACE2 complex is 0.7 [34]. In comparison, experimental deep mutational results for the same set of RBD mutations from 2 different labs only have a correlation of 0.67 [37,38]. TopNetmAb predictions of Omicron [39] and Omicron BA.2 [13] infectivity, vaccine breakthrough, and antibody resistance were nearly perfectly confirmed by experiments and pandemic evolution in the world. These mechanistic discovery and successful predictions may not be achievable via purely experimental means, indicate the indispensable role of AI for scientific discovery.

However, persistent homology and TDA provide only topological invariants, which may not be sufficient for representing PPI data. In particular, the shape of data arising from a family of homotopy geometries cannot be captured by persistent homology. For example,

the geometry of each drum in an acoustic drum set is designed to offer a specific sound or frequency, but persistent homology is insensitive to the change in the sizes (or shapes) in the drum set. This challenge in TDA was addressed by the introduction of persistent Laplacian, or persistent spectral graph [40]. Persistent Laplacian manifests the full set of topological invariants and the homotopic shape evolution of data in its harmonic and non-harmonic spectra, respectively. Additional mathematical analysis [41] and a software package, i.e., HERMES [42], for persistent Laplacian have been reported in the literature. This method has been successfully applied to biological studies, including protein thermal stability [40], protein–ligand binding [43], and protein–protein binding problems [44].

In the present work, we introduce element-specific and site-specific persistent Laplacians to forecast emerging SARS-CoV-2 variants. We hypothesize that persistent Laplacians generate intrinsically low-dimensional representations of PPIs and dramatically reduce the dimensionality of PPI data, leading to a reliable high-throughput screening of emerging SARS-CoV-2 variants. To quantitatively validate this hypothesis, we integrate the harmonic and non-harmonic spectra of persistent Laplacians with efficient machine learning algorithms, i.e., gradient boosting tree (GBT) and deep neural network (Net), to predict PPI $\Delta\Delta G$ following mutations. The resulting topological and spectral-based machine learning models are validated on three major benchmark datasets, the AB-Bind database [45], SKEMPI dataset [46] and SKEMPI v2.0 dataset [47], giving rise to the state-of-the-art performance. Meanwhile, with additional training on SARS-CoV-2 related datasets, our models forecast emerging SARS-CoV-2 variants and recommend four Omicron subvariants, i.e., BA.2.11, BA.2.12.1, BA.4, and BA.5 for active surveillance. Our key projection of BA.4 and BA.5's incoming dominance made in May 1, 2022 [48] had become reality in late June 2022.

2. Results

In this section, we first carry out the infectivity predictions on emerging SARS-CoV-2 variants. Next, three benchmark PPI datasets, i.e., the AB-Bind [45], SKEMPI [46], and SKEMPI 2.0 datasets [47] are employed to demonstrate the proposed persistent Laplacian-based AI models with ten-fold cross validations. Two evaluation metrics, Pearson correlation R_p and the root-mean-square error (RMSE), are used to assess the quality of the present models. Lastly, we present the validation of our models on SARS-CoV-2-related datasets.

2.1. Emerging SARS-CoV-2 variants: Infectivity

Fig. 1 shows the RBD mutations of Omicron subvariants and their BFE changes of SARS-CoV-2 variants. A comparison is also given to other main SARS-CoV-2 variants Alpha, Beta, Gamma, Delta, Theta, Kappa, Lambda, and Mu variants. The Delta variant had the highest BFE change among the earlier variants and was the most infectious variant before the occurrence of the Omicron variant, which explains its dominance in 2021. Omicron BA.1, BA.2, and BA.3 have the common RBD mutations G339D, S373P, S375F, K417N, N440K, S477N, T478K, E484A, Q493R, Q498R, N501Y, and Y505H. Omicron BA.1 has three distinct RBD mutations S371L, G446S, and G496S. Four distinct mutations, S371F, T376A, D405N, and R408S, were found for Omicron BA.2. Omicron BA.3 shares three mutations either with BA.1 or BA.2: S371F, D405N, and G446S. The AI-predicted BFE changes of BA.1, BA.2, and BA.3 are 2.60, 2.98, and 2.88 kcal/mol, respectively [13]. These values are significantly higher than those of other major SARS-CoV-2 variants as shown in Fig. 1. Note that Omicron BA.2 is the most infectious variant. It is about 20 and 4.2 times as infectious as the original SARS-CoV-2 and the Delta variant, respectively. The machine learning model also predicts that BA.2 is about 1.5 times as contagious as BA.1, which is highly consistent with experimental studies [50,51]. BA.2 has been the dominating variant since late March 2022 [13].
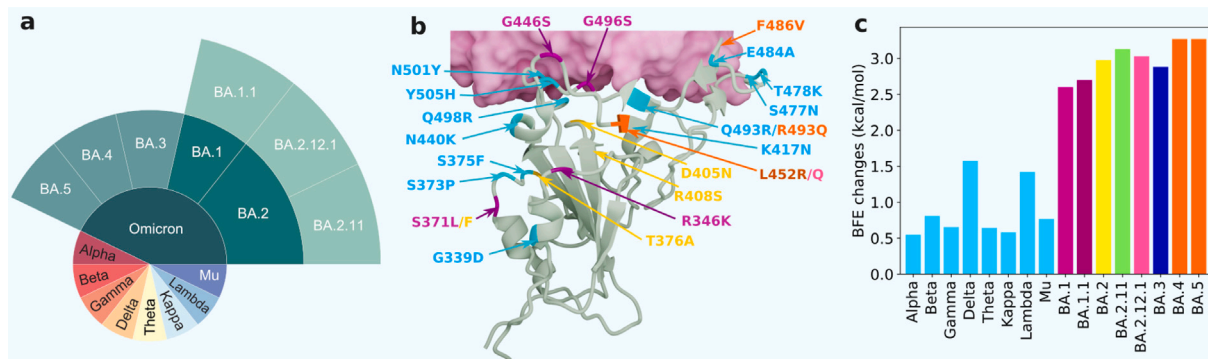
Fig. 1. Illustration of major SARS-CoV-2 (sub)variants, the RBD mutations of Omicron subvariants at the RBD-ACE2 interface, and their mutation-induced BFE changes. **a** Prevailing SARS-CoV-2 variants and Omicron subvariants. **b** RBD mutations of Omicron subvariants at the RBD-ACE2 interface (PDB: 7T9L [49]). The shared 12 mutations are shown in cyan. BA.1 mutations are plotted with magenta. BA.2 mutations are marked in yellow. BA.4 and BA.5 mutations are labeled in orange. The rest colors can be matched from the right chart. **c** A comparison of predicted mutation-induced BFE changes for various SARS-CoV-2 variants and subvariants.

We have also examined other Omicron subvariants, namely, BA.1.1, BA.2.11, BA.2.12.1, BA.4, and BA.5. Compared with BA.1, BA.1.1 has one additional RBD mutation, i.e., R346K. BA.2.11 has one more RBD mutation, L452R, than BA.2 does. BA.2.12.1 has an extra RBD mutation, L452Q, compared with BA.2. BA.4 and BA.5 share the same set of RBD mutations but differ in ORF7b, nucleocapsid (N), and membrane (M) proteins. They have three additional RBD mutations, L452R, F486V, and R493Q compared with BA.2. Note that R493Q is a reversion to the wide type, Q493. It is interesting that L452R is one of Delta's two RBD mutations. Additionally, mutations simultaneously occurred on two RBD residues, L452 and N501, which were singled out by our AI model in early 2020 [6].

Our AI-predicted BFE changes for BA.1.1, BA.2.11, BA.2.12.1, BA.4, and BA.5 are 2.70, 3.13, 3.03, 3.27, and 3.27 kcal/mol, respectively. It is noticed that BA.4 and BA.5 are predicted to be 1.36 times as infectious as BA.2 and have high potential to become new dominating SARS-CoV-2 variants.

2.2. The performance on the AB-Bind dataset

The AB-Bind dataset, including 1101 mutational data entries for experimentally determined BFE changes [45] is considered in the validation of the proposed models. Its 645 single mutations involving 29 antibody–antigen complexes are denoted as the AB-Bind S645 set. In the AB-Bind S645 set, about one-fifth of mutations strengthen the binding, while the rest are destabilizing mutations. In particular, 27 non-binders, which are mutants determined not to bind within the experimental sensitivity of the assay, are in the dataset. The mutation-induced binding free energy changes for these non-binders were set to −8 kcal/mol. For machine learning models, non-binders are outliers and can cause a very negative impact on model accuracy.

As shown in Table 1, our TopLapNetGBT and TopGBT models achieved the R_p of 0.61 and 0.56 for the AB-Bind S645 set. In comparison, TopNet outperforms LapNet because TopNet includes auxiliary features, while LapNet has only Laplacian features. Note that our worst model (LapNet) still outperforms the other best model in the literature, while our best model is about 15% better than the other best model in the literature, indicating the predictive power of our topology and Laplacian-based machine learning models. Both GBTs and Nets models are quite sensitive to system errors as the model training is based on optimizing the mean-square error of the loss function. The BFE changes of 27 non-binders were defined to be −8 kcal/mol in the original dataset and did not follow the distribution of the whole dataset as the values is defined by For the TopLapGBT model, the RMSE of AB-Bind S645set is 1.13 kcal/mol and reduces to 0.82 kcal/mol when 27 non-binder samples are excluded. In this case, the R_p of the TopLapNetGBT model is increased from 0.61 to 0.76 by excluding non-binder samples.

Table 1

Comparison of the Pearson correlation coefficients (R_p) of various methods for the AB-bind S645 set.
Source: The results of methods except for present TopLapGBT and TopLapNet are adopted from Ref. [16].

Method	R_p	Method	R_p
TopLapNetGBT	0.61/0.76[a]	mCSM-AB	0.53/0.56[a]
LapGBT	0.60/0.71[a]	Discovery Studio	0.45
TopNetGBT	0.59/0.76[a]	mCSM-PPI	0.31
LapNetGBT	0.59/0.70[a]	FoldX	0.34
TopLapNet	0.58/0.77[a]	STATIUM	0.32
TopLapGBT	0.58/0.74[a]	DFIRE	0.31
LapNet	0.58/0.70[a]	bASA	0.22
TopNet	0.57/0.76[a]	dDFIRE	0.19
TopGBT	0.56/0.73[a]	Rosetta	0.16

[a]Results exclude 27 non-binders (their ΔΔGs were set to −8 kcal/mol [45]).

The consensus results of GBT and Net have correlations of 0.58–0.59, which are lower than that of GBT but higher than that of Net. GBT models outperform Net models in the validation, showing that GBT performs better than Net on a small dataset.

2.3. The performance on the SKEMPI dataset

The SKEMPI dataset [46] has 3047 entries of BFE changes induced by mutations. This dataset is collected from the literature for protein–protein heterodimeric complexes with experimentally determined structures. It consists of single- and multi-mutations. Among them, 2317 single mutations out of 3047 entries are called the S2317 dataset. Recently, a subset of 1131 non-redundant interface single-mutations is selected and denoted as the S1131 set [18]. Table 2 shows the Pearson correlation coefficients on tenfold cross-validations of various models, including topology- and Laplacian-based models. The proposed topology- and Laplacian-based models are found to be more accurate than other existing methods. One may notice that for a larger training set, the consensus predictions of GBT and Net outperform GBT methods. Additionally, topology-based models contain topology features and auxiliary features, which include more biomolecular information than Laplacian-based models.

2.4. The performance on the SKEMPI 2.0 dataset

The SKEMPI 2.0 [47] database is an updated version of the original SKEMPI database with new mutations from three other databases: AB-bind [45], PROXiMATE [52], and dbMPIKT [53]. This dataset has 7085 entries, including single-mutations and multi-mutations. To validate mCSM-PPI2, David et al. filtered only single-point mutations, selected

Table 2

Comparison of the Pearson correlation coefficients (R_p) of various methods for the S1131 set in the SKEMPI dataset.

Source: The results of other methods are adopted from Ref. [18].

Method	R_p	Method	R_p
TopLapNetGBT	0.87	BindProfX	0.738
TopNetGBT	0.87	Profile-score+FoldX	0.738
TopLapNet	0.86	Profile-score	0.675
TopNet	0.86	SAAMBE	0.624
TopLapGBT	0.86	FoldX	0.457
TopGBT	0.86	BeAtMuSic	0.272
LapNetGBT	0.81	Dcomplex	0.056
LapNet	0.81		
LapGBT	0.78		

Table 3

Comparison of the Pearson correlation coefficients (R_p) of various methods for S4169 set and S8338 set in SKEMPI 2.0.

Source: Results of mCSM-PPI2 are from Ref. [17].

S4169		S8338	
Method	R_p	Method	R_p
TopLapNetGBT	0.82	TopLapNetGBT	0.87
TopNetGBT	0.82	TopLapNet	0.87
TopLapNet	0.81	TopNetGBT	0.87
TopLapGBT	0.81	TopNet	0.86
TopNet	0.81	TopLapGBT	0.85
TopGBT	0.80	TopGBT	0.85
LapNetGBT	0.77	LapNetGBT	0.83
mCSM-PPI2	0.76	mCSM-PPI2	0.82
LapNet	0.76	LapNet	0.81
LapGBT	0.76	LapGBT	0.80

4169 variants in 319 different complexes, and denoted them as the S4169 set [17]. Additionally, set S8338 was derived from set S4169 by setting the BFE changes of the reverse mutations as the negative values of the original BFE changes induced by mutations. We present our tenfold cross-validation results on sets S4169 and S8338 in Table 3. For S4169, TopLapNetGBT has the most accurate result with R_p of 0.82 and RMSE of 1.06 kcal/mol. Topology-based models, aided by auxiliary features, have correlations greater than 0.80 and RMSE from 1.04 kcal/mol to 1.10 kcal/mol. Purely Laplacian-based models also performed quite well, with the Pearson correlation of 0.76, which is the same as that of the mCSM-PPI2.

For the S8338 set, TopLapNetGBT has the highest Pearson correlation R_p of 0.8702 and RMSE of 1.01 kcal/mol as shown in Table 3. TopLapNet has the most accurate results with R_p of 0.8688 and RMSE of 0.984 kcal/mol. Topology models, aided by auxiliary features, have the R_p in the range of (0.848, 0.870) and RMSE in the range of (1.070 kcal/mol, 0.984 kcal/mol). LapNet and LapGBT models have their R_p values slightly lower than that of mCSM-PPI2, but the R_p of their consensus (LapNetGBT) is higher than that of the mCSM-PPI2.

2.5. The performance on SARS-CoV-2 datasets

Training datasets have the utmost importance in implementing our machine learning model for SARS-CoV-2 applications. First, all the datasets mentioned above, including AB-bind,[45] PROXiMATE [52], dbMPIKT [53], SKEMPI [46], and SKEMPI 2.0 [47], are used in our model training. Additionally, SARS-CoV-2-related datasets are also employed to improve the prediction accuracy after a label transformation. These are deep mutational enrichment ratio data, including mutational scNeting data of ACE2 binding to the receptor-binding domain (RBD) of the S protein [54], mutational scNeting data of RBD binding to ACE2 [37,38], and mutational scNeting data of RBD binding to CTC-445.2 and of CTC-445.2 binding to the RBD [37]. Note that in our validation, our training datasets exclude the test dataset, which is a mutational scNeting data of RBD binding to ACE2. Here, these datasets

provide more information on SARS-CoV-2 and can be used to calibrate the models to predict the real experimental results.

Here, we present a validation of our model BFE change prediction for mutations on S protein RBD compared to the experimental deep mutational enrichment data [37]. We compare between experimental deep mutational enrichment data and BFE change predictions on SARS-CoV-2 RBD binding to ACE2 in Fig. 2. Both BFE changes (Fig. 2 top) and enrichment ratios (Fig. 2 bottom) describe the binding affinity changes of the S protein RBD-ACE2 complex induced by mutations. It can be found that the predicted BFE changes are highly correlated to the enrichment ratio data. Pearson correlation is 0.69.

3. Theories and methods

This section presents brief reviews of spectral graph theory, simplicial complex, and persistent Laplacian are presented. Machine learning and deep learning models are discussed in test datasets and validation settings.

3.1. Persistent Laplacians

3.1.1. Spectral graphs

Spectral graph theory studies the spectra of graph Laplacian matrices. It gives rise to the topological and spectral properties of underlying graphs or networks. Mathematically, a graph is an ordered pair $G(V, E)$, where $V = \{v_i; i = 1, 2, \ldots, N\}$ is the vertex set with size N and $E = \{e_{ij} = (v_i, v_j); i \leq i < j \leq N\}$ is the edge set. Denote $\deg(v)$ the degree of each vertex $v_i \in V$, i.e., the number of edges that connects to v. A specific Laplacian matrix L^G can be given by

$$L^G = \begin{cases} \deg(v), & \text{if } v_i = v_j, \\ -1, & \text{if } v_i \text{ and } v_j \text{ are adjacent}, \\ 0, & \text{otherwise}, \end{cases} \quad (2)$$

where "adjacent" is subject to a specific definition or connection rule.

Let order the eigenvalues of the graph Laplacian matrix as

$$\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N = \lambda_{\max}. \quad (3)$$

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

3.1.2. Simplicial complex

To construct a topological description of a graph, simplicial complex is used. For a set of $q+1$ points, $\{v_0, v_1, \ldots, v_q\}$, a q-plane is well defined if the $q + 1$ points are affinely independent, i.e., $v_1 - v_0$, $v_2 - v_0$, ..., $v_q - v_0$ are linearly independent. Thus, one can have at most n linearly independent vectors with at most $n + 1$ affinely independent points in $\mathbb{R}^n$. An affine hull is the set of affine combinations, $v = \sum_{i=0}^{q} c_i v_i$, $c_i \in \mathbb{R}$, and $\sum_{i=0}^{q} c_i = 1$. Such an affine combination is a convex combination if all c_i are non-negative. The convex hull is the set of convex combinations. A q-simplex denoted as σ_q is the convex hull of $q+1$ affinely independent points. For example, 0-, 1-, 2-, and 3-simplex are vertexes, edges, triangles, and tetrahedrons. A simplicial complex K is a collection of simplices in $\mathbb{R}^n$ satisfying the following conditions such as the Cech complexes, Vietoris–Rips complexes, and alpha shapes. For example, the Vietoris–Rips complex of K with radius r consists of all subsets of radius $\mathrm{R}(\sigma)$ at most r as

$$\mathrm{VR}(r) = \{\sigma \subseteq K | \mathrm{R}(\sigma) \leq r\}. \quad (4)$$

For $\sigma_q \in K$, its face σ_{q-1} is also in K. The non-empty intersection of any two simplices σ_q, $\sigma_p \in K$ is a face of them. The dimension of simplicial complex is defined as the maximum dimension of its simplex.
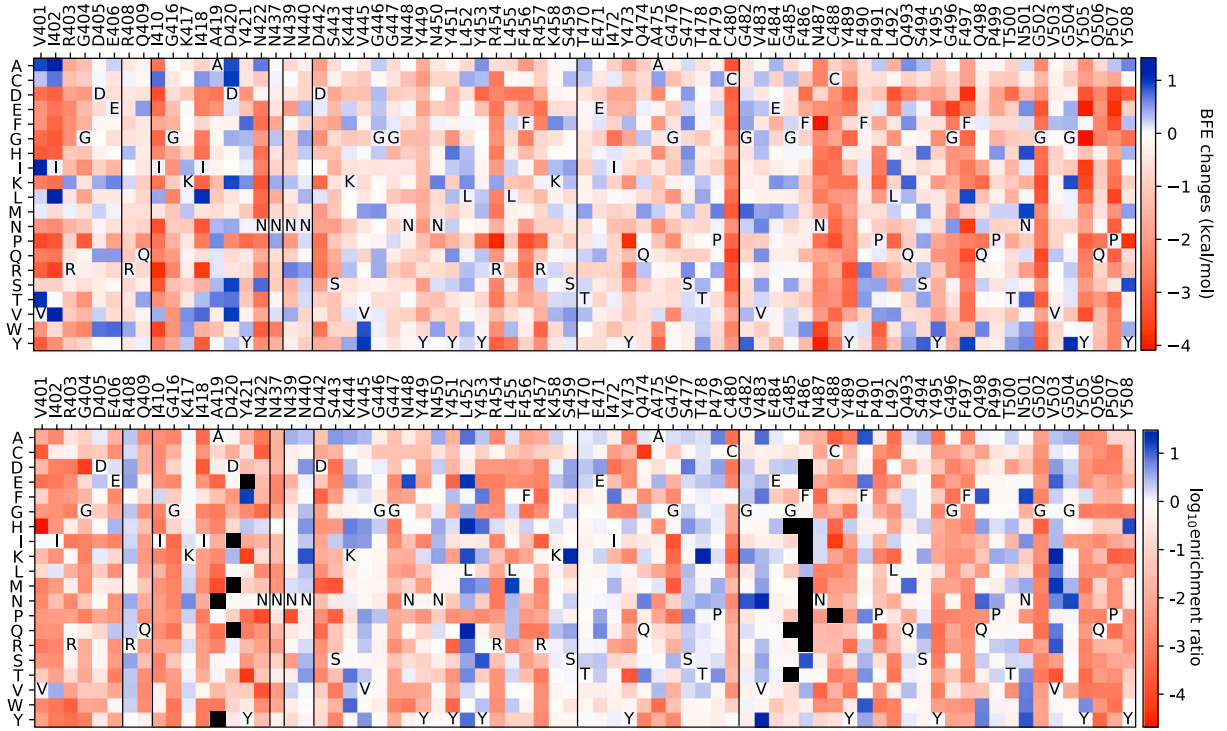
Fig. 2. A comparison between experimental RBD deep mutation enrichment data and predicted BFE changes for SARS-CoV-2 RBD binding to ACE2 (6M0J) [37]. **Top**: machine learning predicted BFE changes for single-site mutants of the S protein RBD. **Bottom**: deep mutational scanning heatmap showing the average effect on the enrichment for single-site mutants of RBD when assayed by yeast display for binding to the S protein RBD [37].

A q-chain is a finite sum of simplices as $\Sigma_i c_i \sigma_i^k$ with $\mathbb{Z}_2$ field of the coefficients c_i for the sum, and the set of all chains in a group $C_q(K)$. The boundary operator ∂_k maps $C_q(K) \rightarrow C_{q-1}(K)$ defined as

$$\partial_q \sigma_q = \sum_{i=0}^{q} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k] = \sum_{i=0}^{q} (-1)^i \sigma_{q-1}^i, \tag{5}$$

where $\sigma_q = [v_0, v_1, \ldots, v_k]$ and $\hat{v}_i$ stands for v_i being omitted. A q-chain is called q-cycle if its boundary is zero. A chain complex is the sequence of chain groups connected by boundary operators

$$\cdots \xrightarrow{\partial_{q+2}} C_{q+1}(K) \xrightarrow{\partial_{q+1}} C_q(K) \xrightarrow{\partial_q} C_{q-1}(K) \xrightarrow{\partial_{q-1}} \cdots \tag{6}$$

and the kth homology group H_k is defined by $H_k = Z_k / B_k$ where $Z_k = \ker \partial_k = \{c \in C_k \mid \partial_k c = 0\}$ and $B_k = \mathrm{im} \ \partial_{k+1} = \{\partial_{k+1} c \mid c \in C_{k+1}\}$. The Betti numbers are defined by the ranks of kth homology group H_k. This, in practice, is counting holes in k-dimension, such as β_0 reflects the number of connected components, β_1 gives the number of loops, and β_2 is the number of cavities. In a nutshell, the Betti number sequence $\{\beta_0, \beta_1, \beta_2, \ldots\}$ reveals the intrinsic topological property of the system.

Recall that in graph theory, the degree of a vertex (0-simplex) v is the number of edges that are adjacent to the vertex, denoted as $\deg(v)$. However, once we generalize this notion to q-simplex, problem aroused since q-simplex can have $(q-1)$-simplices and $(q+1)$-simplices adjacent to it at the same time. Therefore, the upper adjacency and lower adjacency are required to define the degree of a q-simplex for $q > 0$ [55,56].

Definition 3.1. *Given two q-simplices σ_q^i and σ_q^j of a simplicial complex K. We say they are lower adjacent if they share a common $(q-1)$-face, denoted as $\sigma_q^i \overset{L}{\sim} \sigma_q^j$. The lower degree of q-simplex is the number of nonempty $(q-1)$-simplices in K that are faces of σ_q, which is denoted as $\deg_L(\sigma_q)$ and is always $q + 1$.*

Definition 3.2. *Given two q-simplices σ_q^i and σ_q^j of a simplicial complex K. We say they are upper adjacent if they share a common $(q+1)$-face, denoted*

as $\sigma_q^i \overset{U}{\sim} \sigma_q^j$. *The upper degree of q-simplex is the number of $(q+1)$-simplices in K of which σ_q is a face, which is denoted $\deg_U(\sigma_q)$.*

Then, the degree of a q-simplex $(q > 0)$ is defined as:

$$\deg(\sigma_q) = \deg_L(\sigma_q) + \deg_U(\sigma_q) = \deg_U(\sigma_q) + q + 1. \tag{7}$$

3.1.3. Graph Laplacian

The graph Laplacian was introduced to enrich topological and geometric information of simplicial complexes via a filtration process. The preliminary concepts are about the oriented simplicial complex and q-combinatorial Laplacian. More detail information can be found elsewhere [56–59]. The properties of the q-combinatorial Laplacian matrix with its spectra are discussed in the following.

A q-combinatorial Laplacian is defined based on oriented simplicial complexes, and its lower- and higher-dimensional simplexes can be employed to study a specifically oriented simplicial complex. An oriented simplicial complex K is defined if all of its simplices are oriented. If σ_q^i and σ_q^j are upper adjacent with a common upper $(q+1)$-simplex τ_{q+1}, they are similarly oriented if both have the same sign in $\partial_{q+1}(\tau_{q+1})$ and dissimilarly oriented if the signs are opposite. Additionally, if σ_q^i and σ_q^j are lower adjacent with a common lower $(q-1)$-simplex η_{q-1}, they are similarly oriented if η_{q-1} has the same sign in $\partial_q(\sigma_q^i)$ and $\partial_q(\sigma_q^j)$, and dissimilarly oriented if the signs are opposite. Similarly, q-chains can be defined on the oriented simplicial complex K, as well as q-boundary operator.

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

$$\Delta_q := \partial_{q+1} \partial_{q+1}^* + \partial_q^* \partial_q \tag{8}$$

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

$$\mathcal{L}_q = B_{q+1} B_{q+1}^T + B_q^T B_q \tag{9}$$

of operator Δ_q, where $\mathcal{B}_q$ and $\mathcal{B}_q^T$ be the matrix representation of a q-boundary operator and q-coboundary operator, respectively, with respect to the standard basis for $C_q(K)$ and $C_{q-1}(K)$ with some assigned orderings. Then, the number of rows in $\mathcal{B}_q$ corresponds to the number of $(q-1)$-simplices and the number of columns shows the number of q-simplices in K, respectively. In addition, 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. Note that ∂_0 is the zero map which leads to $\mathcal{B}_0$ being a zero matrix. Therefore, $\mathcal{L}_0(K) = \mathcal{B}_1\mathcal{B}_1^T + \mathcal{B}_0^T\mathcal{B}_0$, with K the (oriented) simplicial complex of dimension 1, which is actually a simple graph. Especially, 0-combinatorial Laplacian matrix $\mathcal{L}_0(K)$ is actually the Laplacian matrix defined in the spectral graph theory.

Given an oriented simplicial complex K with $0 \leq q \leq \dim(K)$, the entries of q-combinatorial Laplacian matrices are given by [58]

$$
q > 0, \ (\mathcal{L}_q)_{ij} = \begin{cases} \deg(\sigma_q^i), & \text{if } i = j. \\ 1, & \text{if } i \neq j, \ \sigma_q^i \overset{U}{\nsim} \sigma_q^j \text{ and } \sigma_q^i \overset{L}{\sim} \sigma_q^j \text{ with} \\ & \text{similar orientation.} \\ -1, & \text{if } i \neq j, \ \sigma_q^i \overset{U}{\nsim} \sigma_q^j \text{ and } \sigma_q^i \overset{L}{\sim} \sigma_q^j \text{ with} \\ & \text{dissimilar orientation.} \\ 0, & \text{if } i \neq j \text{ and either }, \ \sigma_q^i \overset{U}{\sim} \sigma_q^j \text{ or } \sigma_q^i \overset{L}{\nsim} \sigma_q^j. \end{cases} \tag{10}
$$

$$
q = 0, \ (\mathcal{L}_q)_{ij} = \begin{cases} \deg(\sigma_0^i), & \text{if } i = j. \\ -1, & \text{if } \sigma_0^i \overset{U}{\sim} \sigma_0^j. \\ 0, & \text{otherwise.} \end{cases} \tag{11}
$$

3.1.4. Persistent spectral graphs

Persistent spectral graphs were introduced by integrating graph Laplacian and multiscale filtration [40]. Both topological and geometric information (i.e. connectivity and robustness of simple graphs) can be derived from analyzing the spectra of q-combinatorial Laplacian matrix. 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, persistent spectral graphs (PSG) is proposed to create a sequence of simplicial complexes induced by varying a filtration parameter, which is inspired by the idea of persistent homology and our earlier work in multiscale graphs. This section mainly introduce the construction of persistent spectral graphs.

First, a q-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. More specifically, the multiplicity of zero spectra of $\mathcal{L}_q(K)$ is denoted by β_q which is actually the qth Betti number defined in the homology:

$$
\beta_q = \dim(\mathcal{L}_q(K)) - \text{rank}(\mathcal{L}_q(K)) = \text{nullity}(\mathcal{L}_q(K))
$$
$$
= \# \text{ of zero eigenvalues of } \mathcal{L}_q(K). \tag{12}
$$

Naturally, persistent spectral theory creates a sequence of simplicial complexes induced by varying a filtration parameter [40]. A filtration of an oriented simplicial complex K is a sequence of sub-complexes $(K_t)_{t=0}^m$ of K

$$
\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m = K. \tag{13}
$$

It induces a sequence of chain complexes Eq. (14) is given in Eq. (14).

$$
\tag{14}
$$

For each sub-complexes K_t, we define its corresponding chain group to be $C_q(K_t)$, and the q-boundary operator will be denoted by $\partial_q^t : C_q(K_t) \to C_{q-1}(K_t)$. We say that if $q < 0$. then $C_q(K_t)$ is an empty set and ∂_q^t is a zero map. If $0 < q \leq \dim(K_t)$, then

$$
\partial_q^t(\sigma_q) = \sum_i^q (-1)^i \sigma_{q-1}^i, \sigma_q \in K_t, \tag{15}
$$

with $\sigma_q = [v_0, \ldots, v_q]$ being the q-simplex, and $\sigma_{q-1}^i = [v_0, \ldots, \hat{v}_i, \ldots, v_q]$ being the $(q-1)$-simplex for which its vertex v_i is removed. Additionally, the adjoint operator is $\partial_q^{t^*} : C_{q-1}(K_t) \to C_q(K_t)$. The topological and spectral information of K_t can be analyzed from $\mathcal{L}_q(K_t)$ along with the filtration parameter by diagonalizing the q-combinatorial Laplacian matrix. We call the multiplicity of zero spectra of $\mathcal{L}_q^t$ as its persistent Betti number β_q^t, which counts the number of q-dimensional holes in K_t:

$$
\beta_q^t = \dim(\mathcal{L}_q^t) - \text{rank}(\mathcal{L}_q^t) = \text{nullity}(\mathcal{L}_q^t) = \# \text{of harmonic spectra of } \mathcal{L}_q^t. \tag{16}
$$

Specifically, β_0^t represents the number of connected components in K_t, β_1^t reveals the number of one-dimensional loops or circles in K_t, and β_2^t shows the number of two-dimensional voids or cavities in K_t. Moreover, the set of spectra of $\mathcal{L}_q^t$ is given by:

$$
\text{Spectra}(\mathcal{L}_q^t) = \{(\lambda_1)_q^t, (\lambda_2)_q^t, \ldots, (\lambda_N)_q^t\}, \tag{17}
$$

where $\mathcal{L}_q^t$ has dimension $N \times N$ and spectra are arranged in ascending order. The smallest non-zero eigenvalue of $\mathcal{L}_q^t$ is defined as $(\tilde{\lambda}_2)_q^t$. The p-persistent q-combinatorial Laplacian operator is defined by extending the boundary operator. Detailed descriptions can be found in Ref. [40].

3.2. Predictive models for mutation-induced protein–protein binding free energy changes

Since the harmonic spectra produced by the kernel of a persistent Laplacian contain exact topological information as that of persistent homology, we utilize a persistent homology software, GUDHI, to generate purely topological representations of PPIs in dimensions 0, 1, and 2. Additionally, persistent Laplacian spectra, including both harmonic and non-harmonic parts, are coded in Python. Machine learning and deep learning algorithms are implemented in Pytorch [60].

3.2.1. Persistent Laplacian representation of PPIs
To facilitate topological and shape analysis of PPIs via persistent Laplacians, we first composite the atoms in a protein–protein complex into various subsets.

1. $\mathcal{A}_m$: atoms of the mutation sites.
2. $\mathcal{A}_{mn}(r)$: atoms in the neighborhood of the mutation site within a cut-off distance r.

3. $\mathcal{A}_A(r)$: protein A atoms within r of the binding site.
4. $\mathcal{A}_B(r)$: protein B atoms within r of the binding site.
5. $\mathcal{A}_{ele}(E)$: atoms in the system that has atoms of element type E.

The distance matrix is specially designed such that it excludes the interactions between the atoms form the same set. For interactions between atoms a_i and a_j in set $\mathcal{A}$ and/or set $\mathcal{B}$, the modified distance is defined as

$$D_{mod}(a_i, a_j) = \begin{cases} \infty, & \text{if } a_i, a_j \in \mathcal{A}, \text{ or } a_i, a_j \in \mathcal{B}, \\ D_e(a_i, a_j), & \text{if } a_i \in \mathcal{A} \text{ and } a_j \in \mathcal{B}, \end{cases} \quad (18)$$

where $D_e(a_i, a_j)$ is the Euclidian distance between a_i and a_j.

Molecular atoms of different can be constructed as points presented by v_0, v_1, v_2, ..., v_k as $k + 1$ affinely independent points in simplicial complex. Persistent spectral graph is devised to track the multiscale topological and geometrical information over different scales along a filtration [40], resulting in significant important feature vectors for the machine learning method. Features generated by binned barcode vectorization can reflect the strength of atom bonds, van der Waals interactions, and can be easily incorporated into a machine learning model, which captures and discriminates local patterns. Using the atom subsets, for example $\mathcal{A}_A(r)$ and $\mathcal{A}_B(r)$, simplicial complexes are constructed by only considering the edges from $\mathcal{A}_A(r)$ to $\mathcal{A}_B(r)$ for Vietoris–Rips complexes. Then from the Vietoris–Rips complex filtration, barcodes generated from persistent homology are enumerated by bar lengths in certain intervals with number 0 or 1. Meanwhile, for each complexes in the filtration, eigenvalues are calculated according to the graph Laplacian analysis. The statistics of eigenvalues such as sum, maximum, minimum, mean, and standard deviation are collected to have a normalized features for machine learning methods. Another method of vectorization is to get the statistics of bar lengths, birth values, and death values, such as sum, maximum, minimum, mean, and standard deviation. This method is applied to vectorize Betti-1 (H_1) and Betti-2 (H_2) barcodes obtained from alpha complex filtration based on the facts that higher-dimensional barcodes are sparser than H_0 barcodes.

3.2.2. Machine learning and deep learning algorithms

The features generated from the persistent spectral graph are tested by the gradient boosting tree (GBT) method and the deep neural network (Net) method. The validations are performed on the datasets discussed in the results section. The accurate prediction of the mutation-induced binding affinity changes of protein–protein complexes is very challenging. After effective feature-generations, a machine learning or deep learning model is also required for validations and real applications. The gradient boosting tree is a popular ensemble method for regression and classification problems. It builds a sequence of weak learners to correct training errors. By the assumption that the individual learners are likely to make different mistakes, the method combines weak learners to eliminate the overall error. Furthermore, a decision tree is added to the ensemble depending on the present prediction error on the training dataset. Therefore, this method is relatively robust against hyperparameter tuning and overfitting, especially for a moderate number of features. The GBT is shown for its robustness against overfitting, good performance for moderately small data sizes, and model interpretability. The present work uses the package provided by scikit-learn (v 0.23.0) [61]. The number of estimators and the learning rate are optimized for ensemble methods as 20 000 and 0.01, respectively. For each set, ten runs (with different random seeds) were done, and the average result is reported in this work. Considering a large number of features, the maximum number of features to consider is set to the square root of the given descriptor length for GBT methods to accelerate the training process. The parameter setting shows that the performance of the average of sufficient runs is decent.

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 with backpropagation to update weights on each layer. To reveal the facts of input features at different levels and abstract more properties, one can construct more layers and more neurons in each layer, known as a deep neural network. Optimization methods for feedforward neural networks and dropout methods are applied to prevent overfitting. The network layers and the number of neurons in each layer are determined by gird 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 on Pytorch [60].

3.2.3. Predictive models

In our previous work, topology-based deep neural network trained with mAbs (TopNetmAb) was introduced to predict mAb binding free energy changes [34]. Persistent homology is the main workhorse for TopNetmAb, but auxiliary features inherited from our earlier TopNet-Tree [19] are utilized.

In this work, we construct a TopNet model from TopNetmAb by excluding mAb training data. A topology-based GBT model (TopBGT) is also developed in the present work by replacing Net in the TopNet model with GBT. Both TopNet and TopGBT include a set of auxiliary features inherited from our earlier TopNetTree [19] and TopNetmAb [34] to enhance their performance.

Additionally, to evaluate the performance of persistent Laplacian (Lap) for PPIs, we construct persistent Laplacian-based GBT (LapGBT) and persistent Laplacian-based deep neural network (LapNet). Note that, unlike TopNet and TopGBT, LapGBT and LapNet employ only persistent Laplacian features extracted from protein structures. Therefore, their performance depends purely on persistent Laplacian.

Moreover, TopLapGBT and TopLapNet are constructed by adding persistent Laplacian features to TopGBT and TopNet, respectively. Furthermore, the consensus of GBT and Net predictions are also used for validations, denoted as TopNetGBT and LapNetGBT, respectively. Finally, the consensus of TopLapNet and TopLapGBT is called TopLapNetGBT.

4. Conclusion

Due to natural selection, emerging SARS-CoV-2 variants are spreading worldwide with their increased transmissibility as a result of higher infectivity and/or stronger antibody resistance. The increase in antibody resistance also leads to vaccine breakthrough infections and jeopardizes the existing monoclonal antibody drugs. The spike protein plays the most important role in viral transmission because its receptor binding domain (RBD) binds to human ACE2 to facilitate the viral entry of host cells. Topological data analysis (TDA) of RBD-ACE2 binding free energy changes induced by RBD mutations enables the accurate forecasting of emerging SARS-CoV-2 variants [6,13,39,62].

However, the earlier TDA method is not sensitive to homotopic shape evolution during filtration, which is important for protein–protein interactions (PPIs). To overcome this obstacle, persistent Laplacian, which characterizes the topology and shape of data, is introduced in this work for analyzing PPIs. Paired with advanced machine learning and deep learning algorithms, the proposed persistent Laplacian method outperforms the state-of-art approaches in validation with mutation-induced binding free energy changes of PPIs using major benchmark datasets. An important forecasting from the present work is that Omicron subvariants BA.2.11, BA.212.1, BA.4, and BA.5 have a high potential to become new dominating variants in the world.

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.

References

[1] W. Li, Z. Shi, M. Yu, W. Ren, C. Smith, J.H. Epstein, H. Wang, G. Crameri, Z. Hu, H. Zhang, et al., Bats are natural reservoirs of SARS-like coronaviruses, Science 310 (5748) (2005) 676–679.

[2] X.-X. Qu, P. Hao, X.-J. Song, S.-M. Jiang, Y.-X. Liu, P.-G. Wang, X. Rao, H.-D. Song, S.-Y. Wang, Y. 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.

[3] H.-D. Song, C.-C. Tu, G.-W. Zhang, S.-Y. Wang, K. Zheng, L.-C. Lei, Q.-X. Chen, Y.-W. Gao, H.-Q. Zhou, H. 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.

[4] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T.S. Schiergens, G. Herrler, N.-H. Wu, A. 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.

[5] A.C. Walls, Y.-J. Park, M.A. Tortorici, A. Wall, A.T. McGuire, D. Veesler, Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, Cell (2020).

[6] J. Chen, R. Wang, M. Wang, G.-W. Wei, Mutations strengthened SARS-CoV-2 infectivity, J. Mol. Biol. 432 (19) (2020) 5212–5226.

[7] R. Wang, J. Chen, K. Gao, G.-W. Wei, Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries, Genomics 113 (4) (2021) 2158–2170.

[8] C. Wang, W. Li, D. Drabek, N.M. Okba, R. van Haperen, A.D. Osterhaus, F.J. van Kuppeveld, B.L. Haagmans, F. Grosveld, B.-J. Bosch, A human monoclonal antibody blocking SARS-CoV-2 infection, Nature Commun. 11 (1) (2020) 1–6.

[9] F. Yu, R. Xiang, X. Deng, L. Wang, Z. Yu, S. Tian, R. Liang, Y. Li, T. Ying, S. Jiang, Receptor-binding domain-specific human neutralizing monoclonal antibodies against SARS-CoV and SARS-CoV-2, Signal Transduct. Target. Ther. 5 (1) (2020) 1–12.

[10] C. Li, X. Tian, X. Jia, J. Wan, L. Lu, S. Jiang, F. Lan, Y. Lu, Y. Wu, T. Ying, The impact of receptor-binding domain natural mutations on antibody recognition of SARS-CoV-2, Signal Transduct. Target. Ther. 6 (1) (2021) 1–3.

[11] R. Wang, J. Chen, G.-W. Wei, Mechanisms of sars-cov-2 evolution revealing vaccine-resistant mutations in europe and america, J. Phys. Chem. Lett. 12 (2021) 11850–11857.

[12] R. Wang, J. Chen, Y. Hozumi, C. Yin, G.-W. Wei, Emerging vaccine-breakthrough SARS-CoV-2 variants, ACS Infect. Dis. 8 (3) (2022) 546–556.

[13] J. Chen, G.-W. 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.

[14] 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.

[15] M. Petukh, L. Dai, E. Alexov, Saambe: webserver to predict the charge of binding free energy caused by amino acids mutations, Int. J. Mol. Sci. 17 (4) (2016) 547.

[16] D.E. Pires, D.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.

[17] C.H. Rodrigues, Y. Myung, D.E. Pires, D.B. Ascher, Mcsm-ppi2: predicting the effects of mutations on protein–protein interactions, Nucleic Acids Res. 47 (W1) (2019) W338–W344.

[18] P. Xiong, C. Zhang, W. Zheng, Y. Zhang, Bindprofx: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts, J. Mol. Biol. 429 (3) (2017) 426–434.

[19] 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.

[20] P. Frosini, Measuring shapes by size functions, in: Intelligent Robots and Computer Vision X: Algorithms and Techniques, Vol. 1607, International Society for Optics and Photonics, 1992, pp. 122–133.

[21] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, in: Proceedings 41st Annual Symposium on Foundations of Computer Science, IEEE, 2000, pp. 454–463.

[22] A. Zomorodian, G. Carlsson, Computing persistent homology, Discrete Comput. Geom. 33 (2) (2005) 249–274.

[23] G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2) (2009) 255–308.

[24] K. Mischaikow, V. Nanda, Morse theory for filtrations and efficient computation of persistent homology, Discrete Comput. Geom. 50 (2) (2013) 330–353.

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

[26] V. De Silva, R. Ghrist, Coverage in sensor networks via persistent homology, Algebr. Geom. Topol. 7 (1) (2007) 339–358.

[27] Y. Yao, J. Sun, X.H. Huang, G.R. Bowman, G. Singh, M. Lesnick, L.J. Guibas, V.S. Pande, G. Carlsson, Topological methods for exploring low-density states in biomolecular folding pathways, J. Chem. Phys. 130 (2009) 144115.

[28] P. Bubenik, J.A. Scott, Categorification of persistent homology, Discrete Comput. Geom. 51 (3) (2014) 600–627.

[29] T.K. Dey, F. Fan, Y. Wang, Computing topological persistence for simplicial maps, in: Proceedings of the Thirtieth Annual Symposium on Computational Geometry, ACM, 2014, p. 345.

[30] 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.

[31] Z. Cang, L. Mu, G.-W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening, PLoS Comput. Biol. 14 (1) (2018) e1005929.

[32] 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 (1) (2019) 71–82.

[33] J. Chen, K. Gao, R. Wang, G.-W. Wei, Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies, Chem. Sci. 12 (20) (2021) 6929–6948.

[34] J. Chen, K. Gao, R. Wang, G.-W. Wei, Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies, J. Mol. Biol. 433 (7744) (2021).

[35] M. Alenquer, F. Ferreira, D. Lousa, M. Valério, M. Medina-Lopes, M.-L. Bergman, J. Gonçalves, J. Demengeot, R.B. Leite, J. Lilue, et al., Signatures in sars-cov-2 spike protein conferring escape to neutralizing antibodies, PLoS Pathog. 17 (8) (2021) e1009772.

[36] L. Dupont, L.B. Snell, C. Graham, J. Seow, B. Merrick, T. Lechmere, T.J. Maguire, S.R. Hallett, S. Pickering, T. Charalampous, et al., Neutralizing antibody activity in convalescent sera from infection in humans with sars-cov-2 and variants of concern, Nat. Microbiol. (2021) 1–10.

[37] T.W. Linsky, R. Vergara, N. Codina, J.W. Nelson, M.J. Walker, W. Su, C.O. Barnes, T.-Y. Hsiang, K. Esser-Nobis, K. Yu, et al., De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2, Science 370 (6521) (2020) 1208–1214.

[38] T.N. Starr, A.J. Greaney, S.K. Hilton, D. Ellis, K.H. Crawford, A.S. Dingens, M.J. Navarro, J.E. Bowen, M.A. Tortorici, A.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.

[39] J. Chen, R. Wang, N.B. Gilby, G.-W. Wei, Omicron variant (b, 1.1. 529) infectivity, vaccine breakthrough, and antibody resistance, J. Chem. Inf. Model. 62 (2) (2022) 412–422.

[40] R. Wang, D.D. Nguyen, G.-W. Wei, Persistent spectral graph, Int. J. Numer. Methods Biomed. Eng. 36 (9) (2020) e3376.

[41] F. Mémoli, Z. Wan, Y. Wang, Persistent laplacians: Properties, algorithms and implications, 2020, arXiv preprint arXiv:2012.02808.

[42] R. Wang, R. Zhao, E. Ribando-Gros, J. Chen, Y. Tong, G.-W. Wei, Hermes: Persistent spectral graph software, Found. Data Sci. (Springfield, Mo.) 3 (1) (2021) 67.

[43] Z. Meng, K. Xia, Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction, Sci. Adv. 7 (19) (2021) eabc5329.

[44] J. Wee, K. Xia, Persistent spectral based ensemble learning (perspect-el) for protein–protein binding affinity prediction, Brief. Bioinform. 23 (2) (2022).

[45] S. Sirin, J.R. Apgar, E.M. Bennett, A.E. Keating, AB-Bind: antibody binding mutational database for computational affinity predictions, Prot. Sci. 25 (2) (2016) 393–409.

[46] I.H. Moal, J. Fernández-Recio, SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models, Bioinformatics 28 (20) (2012) 2600–2607.

[47] J. Jankauskaitė, B. Jiménez-García, J. Dapkūnas, J. Fernández-Recio, I.H. Moal, SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation, Bioinformatics 35 (3) (2019) 462–469.

[48] J. Chen, Y. Qiu, R. Wang, G.-W. Wei, Persistent laplacian projected omicron ba. 4 and ba. 5 to become new dominating variants, 2022, arXiv preprint arXiv:2205.00532.

[49] D. Mannar, J.W. Saville, X. Zhu, S.S. Srivastava, A.M. Berezuk, K. Tuttle, C. Marquez, I. Sekirov, S. Subramaniam, Sars-cov-2 omicron variant: Ace2 binding, cryo-em structure of spike protein-ace2 complex and antibody evasion, BioRxiv (2021).

[50] BA2 reinfection, 0000. https://www.timesofisrael.com/several-cases-of-omicron-reinfection-said-detected-in-israel-with-new-ba2-strain/.

[51] F.P. Lyngse, C.T. Kirkeby, M. Denwood, L.E. Christiansen, K. Mølbak, C.H. Møller, R.L. Skov, T.G. Krause, M. Rasmussen, R.N. Sieber, et al., Transmission of sars-cov-2 omicron voc subvariants ba. 1 and ba. 2: Evidence from danish households, MedRxiv (2022).

[52] S. Jemimah, K. Yugandhar, M. Michael Gromiha, Proximate: a database of mutant protein–protein complex thermodynamics and kinetics, Bioinformatics 33 (17) (2017) 2787–2788.

[53] Q. Liu, P. Chen, B. Wang, J. Zhang, J. Li, Dbmpikt: a database of kinetic and thermodynamic mutant protein interactions, BMC Bioinformatics 19 (1) (2018) 1–7.

[54] E. Procko, The sequence of human ace2 is suboptimal for binding the s spike protein of sars coronavirus 2, BioRxiv (2020).

[55] D.H. Serrano, D.S. Gómez, Centrality measures in simplicial complexes: applications of tda to network science, 2019, arXiv preprint arXiv:1908.02967.

[56] S. Maletić, M. Rajković, Consensus formation on a simplicial complex of opinions, Phys. A 397 (March) (2014) 111–120.

[57] D. Hernández Serrano, D. Sánchez Gómez, Higher order degree in simplicial complexes, multi combinatorial laplacian and applications of tda to complex networks, 2021, arXiv preprint arXiv:1908.02583.

[58] T.E. Goldberg, Combinatorial Laplacians of Simplicial Complexes Senior Thesis, Bard College, 2002.

[59] D. Horak, J. Jost, Spectra of combinatorial laplace operators on simplicial complexes, Adv. Math. 244 (2013) 303–336.

[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[62] R. Wang, J. Chen, Y. Hozumi, C. Yin, G.-W. Wei, Emerging vaccine-breakthrough SARS-CoV-2 variants, ACS Infect. Dis. 8 (3) (2021) 546–556.