Computational and Structural Biotechnology Journal 21 (2023) 4816–4824

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

Fast end-to-end surface interpretation of SARS-CoV-2 variants by differentiable molecular surface interaction fingerprinting method

Ziyang Zheng [a], Yanqi Jiao [a], Haixin You [a], Junfeng An [b], Yao Sun [a,*]

[a] School of Science, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China
[b] Shen Zhi Xing (Shenzhen) Technology, Shenzhen, Guangdong 518055, China

ARTICLE INFO

Keywords:
SARS-CoV-2
Spike protein
Deep learning
Molecular surface
Interaction site

ABSTRACT

Confronting the challenge of persistent mutations of SARS-CoV-2, researchers have turned to deep learning methods to predict the mutated structures of spike proteins and to hypothesize potential changes in their structures and drug efficacies. However, limited works are focused on the surface learning of spike proteins even though their biological functions are usually defined by the geometric and chemical features of 3D molecular surfaces. In addition, the current used geometric deep learning methods are based on mesh representations of proteins to identify potential binding targets for drugs. However, the use of meshes has limitations and is not applicable for many important tasks in molecular biology. To address these limitations, we adopt the differentiable molecular surface interaction fingerprinting (dMaSIF) method which is based on the 3D point clouds and a novel efficient geometric convolutional layer to fast predict the interaction sites on the protein surface. The different binding site patterns for Delta, Omicron and its subvariants are clearly visualized. We find that Delta and Omicron show the similar surface binding patterns while BA.2, BA.2.13, BA.3 and BA.4 present similar ones. BA.4 possesses higher positive interaction site ratio than the others which may account for its higher transmission and infection among humans. In addition, the positive interaction site ratios of BA.2, BA.2.13, BA.3 are higher than Delta and Omicron, which are accordant with their transmission and infection rates. Hopefully our work offers a new effective route to analyze the protein-protein interaction for the SARS-CoV-2 variants.

1. Introduction

Since the massive outbreak of pneumonia cases in China in December 2019, the epidemic of SARS-CoV-2 virus has turned into a persistent threat to the world [1]. The SARS-CoV-2 has been continuously evolving by acquiring genomic mutations, resulting in the emergence of specific variants of multiple concerns [2]. The mutations in the SARS-CoV-2 spike protein could significantly enhance the binding affinity of receptor-binding domain (RBD) with human Angiotensin-converting enzyme 2 (hACE2), leading to rapid spread in the population. In turn, the increased viral replication can increase the likelihood of mutation formation of SARS-CoV-2 [3]. The available option to possibly terminate the pandemic is the development of effective vaccines and drugs against circulating variants [4]. Therefore, the accurate identification of binding sites of drugs and antibodies on SARS-CoV-2 spike proteins is of great significance towards a better control of the pandemic.

Since its emergence, SARS-CoV-2 has been found to evolve and trigger new variants of concern (VOCs) to avoid host hostility, that is, to evade the host immune response and increase transmission and aggression in the pathogenesis of COVID-19 [5,6]. Among the SARS-CoV-2 variants, the Delta (B.1.617.2) identified in India in December 2020 aroused panic in public which was believed to be 60 % more transmissible than the former Alpha variant [7]. Some researchers have proposed effective antibodies for the termination of the Delta variant, including casirivimab [8], imdevimab [9], celltrion, and regdanvimab [10] *etc*. However, the Delta variant changed from a variable of interest (VOI) to a VOC. In November 2021, the WHO Technical Advisory Group on Virus Evolution (TAG-VE) proposed the identification of the B.1.1.529, which was commonly known as the Omicron variant to be a new VOC. The spike protein of Omicron is determined by 30 mutations, 15 of which occur in the RBD, as well as three deletions and one insertion [11]. The Omicron variant is by far the most highly differentiated strain identified in large numbers during the pandemic, raising concerns associated with greater infectivity, lower vaccine efficiency and greater risk of reinfection [11]. Its subvariant BA.1 started the Omicron wave which has 39 mutations including 30 substitutions,

* Corresponding author.
E-mail address: sunyao0819@hit.edu.cn (Y. Sun).

https://doi.org/10.1016/j.csbj.2023.09.033
Received 11 April 2023; Received in revised form 14 September 2023; Accepted 26 September 2023
Available online 28 September 2023

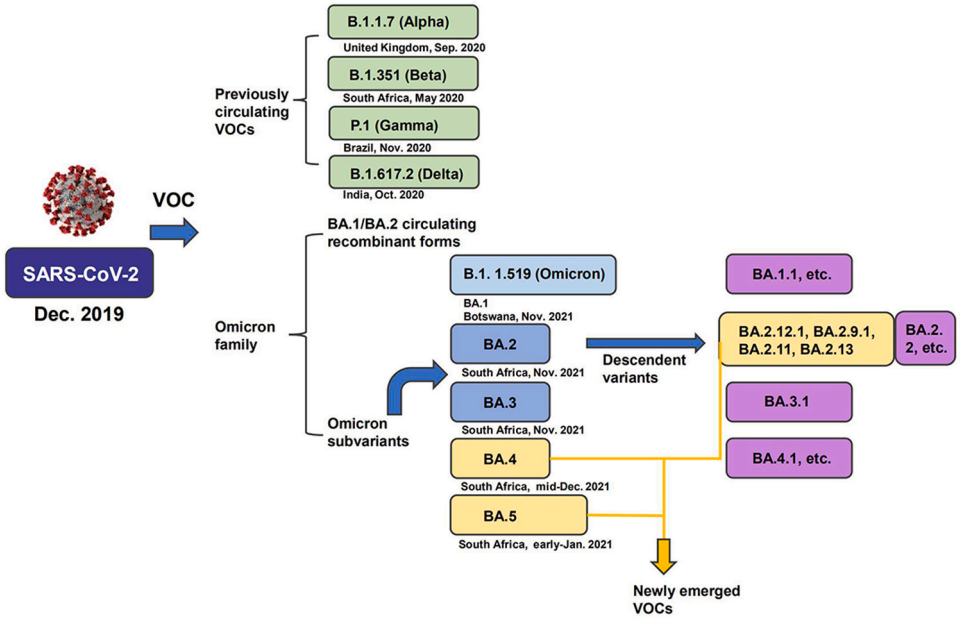

Fig. 1. Some global SARS-CoV-2 VOCs including Delta, Omicron and its subvariants.

six deletions and three insertions comparing to the ancestral strain. The BA.2 and BA.3 appeared at about the same time as BA.1 which are highly related but contain some unique changes in spike protein. The BA.2 and BA.3 have 31 and 34 mutations comparing to the ancestral strain, with 21 shared mutations between them [12]. Followingly, the Omicron subvariants BA.2.12.1, BA.2.13, BA.4, and BA.5 appeared which contain Leu452 substitutions and become more infectious than the BA.2 [13]. Fig. 1 shows some of the global SARS-CoV-2 VOCs including Delta, Omicron and its subvariants [14]. The Omicron subvariants show closely related variations that share a common ancestor. However, their spread rates are quite different which is most certainly due to differences in spike protein.

In recent years, more and more computational studies based on deep learning methods have focused on the structural and binding sites predictions of SARS-CoV-2 mutant spike proteins. For example, the revolutionary AlphaFold2 has taken advantage of the wealth of protein sequences available in the open protein database and shown how the proteins fold barely from the 1D amino acid sequences [15]. This method firstly predicts the distances between amino acids and other geometric relations and then uses them as constraints to refine the 3D structures. There are also some researchers adopting unsupervised learning methods borrowed from the field of Natural Language Processing to predict the biological properties of proteins from sequence information alone [16,17]. In addition, the protein-protein interactions (PPI) can be predicted based on the relations between amino acids of different proteins using deep learning methods [18–25]. However, most of the research papers predict PPI based on the amino acid sequence data and identify nonlinear relationship between the extracted and learned features. Limited works have focused on the surface interpretations of proteins to predict their interactions. In fact, the internal parts of the 3D folded protein do not contribute to protein interactions, but the surface plays the key role instead [26]. Gainza et al. have proposed the Molecular Surface Interaction Fingerprinting (MaSIF) method and pioneered the mesh-based geometric deep learning to predict PPI [27]. This method could classify binding sites for drugs and discriminate surface sites of interaction for protein-protein complexes. But the limitations exist in three aspects. The first one is that the protein surface should be preprocessed as the raw atomic point cloud and then represented by meshes. Secondly, it relies on pre-computed and stored hand-crafted chemical and geometric features. Thirdly, it uses MoNet mesh convolutions [28] on precomputed geodesic patches, which are prohibitively expensive for calculating more than a few thousand proteins. To overcome these limitations, Sverrisson et al. have proposed a method named differentiable molecular surface interaction fingerprinting (dMaSIF) to predict the surface interaction sites for proteins on-the-fly from the underlying atomic point cloud using a novel efficient geometric convolutional layer [26]. This method could end-to-end fast process (tens of milliseconds for pre-processing per protein) large collections of proteins by only taking the raw 3D coordinates and chemical types of their atoms as input without any hand-crafted pre-computed features. The time cost is over 40 times faster than MaSIF and the performance reaches 0.82. It is primarily designed to tackle two important tasks, i.e., binding site identification and interaction prediction. The first task focuses on classifying the surface of a given protein into interaction sites and non-interaction sites. The interaction sites are surface patches that are more likely to mediate interactions with other proteins. Understanding the properties of these interaction sites is of utmost importance for various applications, such as drug design and the study of protein interaction networks. The second task involves taking two surface patches as input with each representing a different protein involved in a complex, and predicting whether these locations tend to come into close contact. This task is particularly critical for tasks like protein docking, which seeks to predict the spatial orientation and arrangement of two proteins when they form a complex. By predicting the likelihood of close contact between the provided surface patches, the method offers valuable insights into the potential interactions between the proteins and assist in predicting the spatial arrangement of the proteins in the complex.

In this work, we adopted the newly emerged dMaSIF method [26] for identifying the interaction sites for SARS-CoV-2 variants. We visualized the interaction sites on 3D surfaces directly and analyzed the similarities of surface binding patterns among different variants. The predicted interaction sites with high positive binding ratios could directly show the surface features of SARS-CoV-2 variants, offering new insights of evolutionary characteristics of the virus.

2. Methods

In this paper, we applied the previously proposed model and modified the input of dMaSIF. For the hyperparameters, we followed the

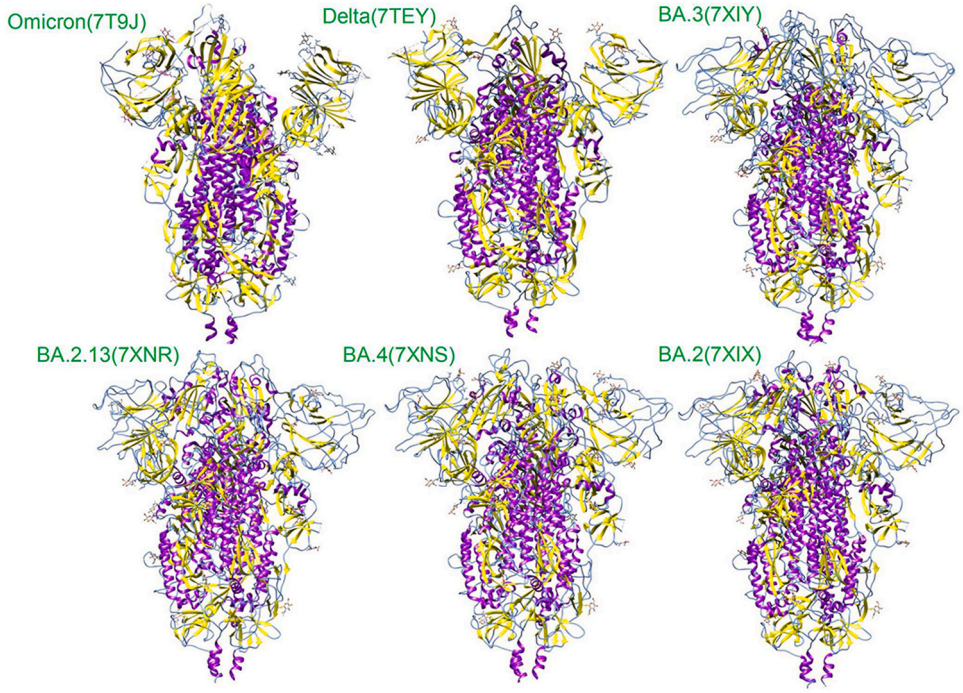

Fig. 2. The structures of SARS-CoV-2 Delta (B.1.617.2) (7TEY.pdb), Omicron (B.1. 1.519) (7T9J.pdb), Omicron BA.2 (7XIX.pdb), BA.2.13 (7XNR.pdb), BA.3 (7XIY.pdb), and BA.4 (7XNS.pdb) spike proteins. The secondary structures of the strand, helix and coil were colored in yellow, purple and blue respectively.

setting in dMaSIF to ensure the performance of neural network. We firstly computed a representation of the protein molecular surface. In our process, we calculated the following results respectively in order: (1) a representation of the protein molecular surface, (2) geometric and chemical features, (3) local coordinate systems (4) the probability of binding site predicted by a geometric convolutional neural network.

2.1. PDB models

The PDB models for SARS-CoV-2 Delta (B.1.617.2) (7TEY.pdb), Omicron (B.1. 1.519) (7T9J.pdb), Omicron BA.2 (7XIX.pdb), BA.2.13 (7XNR.pdb), BA.3 (7XIY.pdb), and BA.4 (7XNS.pdb) spike proteins were adopted in this work. These models are all experimental structures of the corresponding SARS-CoV-2 variants from either cryogenic electron microscopy (Cryo-EM) or electron microscopy (EM). We deleted the small ligands (NAG) using the UCSF Chimera 1.16 software [29] to show the 3 chains clearly. The secondary structures of the strand, helix and coil were colored in yellow, purple and blue respectively, as shown in Fig. 2. The detailed differences of secondary structures among SARS-CoV-2 Delta (B.1.617.2) (7TEY.pdb), Omicron (B.1. 1.519) (7T9J.pdb), Omicron BA.2 (7XIX.pdb), BA.2.13 (7XNR.pdb), BA.3 (7XIY.pdb), and BA.4 (7XNS.pdb) spike proteins were calculated by using VMD 1.9.3 [30], as shown in Fig. S1 in the Supplementary materials.

2.2. Work on protein surface

The dMaSIF model is an end-to-end structure which uses 3D coordinates α_k and chemical properties of all the atom centers of a given protein as inputs and yields results depending on the target task such as a interaction site score (in the range of 0–100, indicating the probability as a interaction site) or a binding/non-binding judgment. The premise of this model is that the molecular surface of the protein carries chemical and geometric information that determines how it binds or interacts with other molecules.

There were six types of atoms (C, H, O, N, S, Se) in our input data, each of which could be encoded as one-hot code $C_k \in \mathbb{R}^6$. The surface of the protein was in the form of oriented point clouds denoted by coordinates and unit normal vectors. The feature vectors associated with these points were updated from 16 dimensions (10 geometric and six chemical) to one dimension by convolution-like steps. Our data scales of the model training were 3–15 K, 30–300 Å, 1 Å and 6–15 K for atoms, molecule size, surface sampling resolution and sampling rate, respectively.

2.3. Surface description

2.3.1. Fast sampling

Fast sampling was divided into three steps, i.e., initial sampling, evolution and screening, prior to which we provided the distance function $SDF(x)$ (Eq. (1)) associated with all atoms and defined squared loss function $E(x_1, x_2, \cdots x_N)$ (Eq. (2)), where σ_k was atomic radius determined by their intrinsic nature and $\sigma(x)$ was an average atomic radius in a neighborhood of point x. The model described the protein surface in terms of the level set of the distance function [31].

$$\begin{cases} SDF(x) = -\sigma(x) \cdot \log \sum_{k=1}^{A} \exp(-||x-\alpha_k||/\sigma_k) \\ \sigma(x) = \dfrac{\sum_{k=1}^{A} \exp(-||x-\alpha_k||)\sigma_k}{\sum_{k=1}^{A} \exp(-||x-\alpha_k||)} \end{cases}, k = 1, 2, \cdots A. \quad (1)$$

$$E\left(x_1, x_2, \cdots x_N\right) = \frac{1}{2}\sum_{i=1}^{N} (SDF(x_i) - r)^2, r = 1.05 \quad (2)$$

Initial sampling: We generated B (B = 20) points from a Gaussian random distribution $\mathcal{N}(\mu = \sigma_k, \sigma^2 = 100)$ in a neighborhood of one atom, and there were total of $A \times B$ sampling points for all.

Evolution: We minimized the loss function (Eq. (2)) by gradient method to get these sampling points evolving towards gradient decent direction. Specific operation was to set r = 1.05 Å, learning rate= 1, and perform gradient descent 4 times.

Screening: Screening consisted of two steps. Step 1: we kept points if

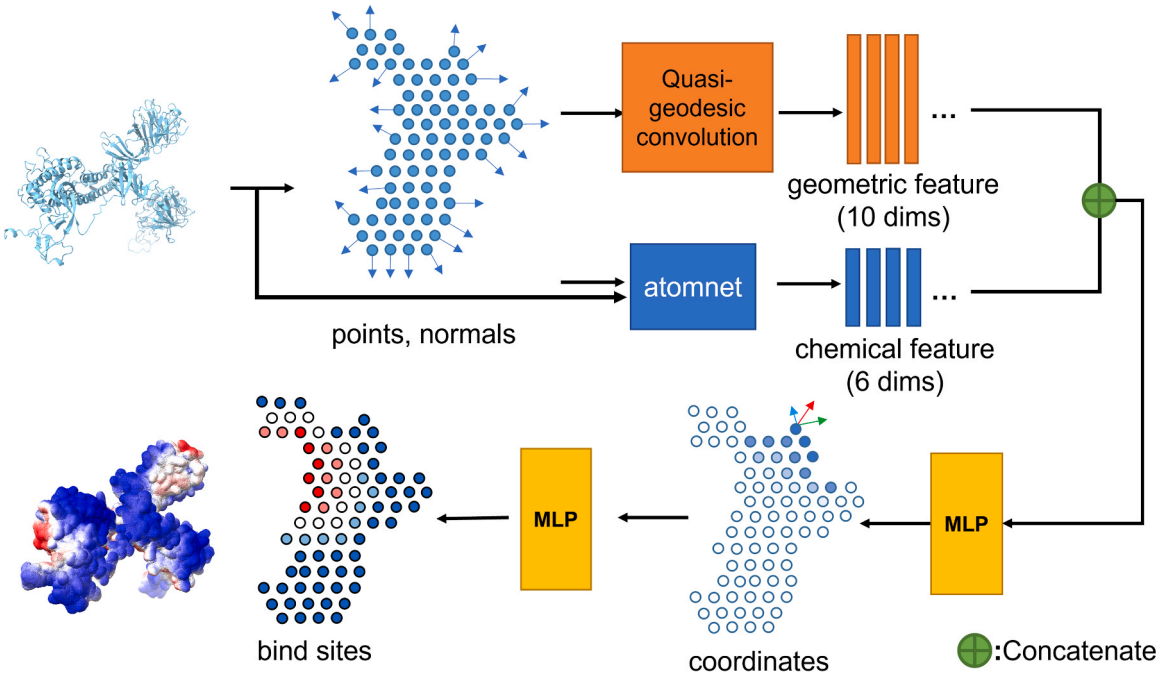

Fig. 3. Illustration of the dMaSIF methodology applied in this work.

Table 1
Software prerequisites of dMaSIF.

Dependency	First Option	Second Option
GCC	7.5.0	8.4.0
CMAKE	3.10.2	3.16.5
CUDA	10.0.130	10.2.89
cuDNN	7.6.4.38	7.6.5.32
Python	3.6.9	3.7.7
PyTorch	1.4.0	1.6.0
PyKeops	1.4	1.4.1
PyTorch Geometric	1.5.0	1.6.1

the distance value was in $(r-0.1, r+0.1)$Å and its increment exceeded 0.5 Å after gradient descending 4 times with a size of 1 Å. That is, the points far away from appointed level set or having persistently negative or relatively flat gradient were excluded. Step 2: we defined cubes with edges 1 Å and kept only one point in each cube to ensure uniform density of sampling point. After the two steps, we obtained N sampling points.

2.3.2. Construction of local coordinate system

We normalized the gradient of the distance function for each sample point $x_i (i = 1, 2, \ldots, N)$ to get their normal vector $\widehat{n_i}$ as initialization. To construct the local coordinate system $(\widehat{n_i}, \widehat{u_i}, \widehat{v_i})$ for subsequent geometric feature analysis, we smoothed the vector field using a Gaussian kernel with $\sigma = \{9, 12\}$, updated normal vector $\widehat{n_i}$ after normalization according to Eq. (3), and calculated the tangential vectors $\widehat{u_i}, \widehat{v_i}$ according to Eq. (4) [32].

$$\widehat{n_j} \leftarrow \text{normalize}\left(\sum_{j=1}^{N} \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right)\widehat{n_j}\right) \tag{3}$$

$$\left\{\widehat{n_i} = \left[x, y, z\right] s = sign(z), a = \frac{-1}{s+z}, b = axy\widehat{u_i} = \left[1 + sax^2, sb, -sx\right], \widehat{v_i} = \left[b, s + ay^2, -y\right]\right. \tag{4}$$

2.3.3. Chemical feature vectors

For each sampling point x_k, we found coordinates of the 16 nearest atomic centers $\alpha_m^i, m = 1, 2, \cdots 16$ with their chemical types $C_m^i, m = 1, 2, \cdots 16$, and the vectors $\left[C_m^i, \frac{1}{||x_i - \alpha_m^i||}\right] \in \mathbb{R}^7$ integrated from the coordinates and chemical types were input into the first Multi-Layer Perceptron (MLP) to generate 16 feature vectors $f_{i,m} \in \mathbb{R}^6$. The second MLP linearly mapped summation of them to chemical feature vectors $f_i \in \mathbb{R}^6$ (each dimension of vectors had a realistic physical meaning, such as Poisson-Boltzmann electrostatic potential, etc.).

2.4. Quasi-geodesic convolution on points clouds

2.4.1. Convolutions on 3D shapes

The geometric convolutional neural network of the model assembled MLP and trainable convolutional operators to simulate quasi-geodesic and predict the molecular surface based on its local chemical and geometric features only, which ensured 3D rotations and translations invariance and protected model from overfitting.

2.4.2. Work on oriented point clouds

The geodesic distance between two points x_i and x_j of a protein surface with weights computed by unit normals $\widehat{n_i}$ and $\widehat{n_j}$ was approximated as Eq. (5).

$$d_{ij} = \| x_i - x_j \| \cdot \left(2 - \langle \widehat{n_i}, \widehat{n_j} \rangle\right) \tag{5}$$

Then we applied a Gaussian window as a filter to d_{ij} (Eq. (6)),

$$w(d_{ij}) = exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right) \tag{6}$$

where the radius $\sigma \in \{9, 12\}$Å. For any point x_i and its neighbor points x_j, we encoded their relative position and orientation in the local coordinate system $(\widehat{n_i}, \widehat{u_i}, \widehat{v_i})$ (Eq. (7)).

$$p_{ij} = \left[\widehat{p}_{ij}^{\widehat{n}}, \widehat{p}_{ij}^{\widehat{u}}, \widehat{p}_{ij}^{\widehat{v}}\right] = \left(x_j - x_i\right)^{\top} \cdot \left[\widehat{n_i} | \widehat{u_i} | \widehat{v_i}\right] \tag{7}$$

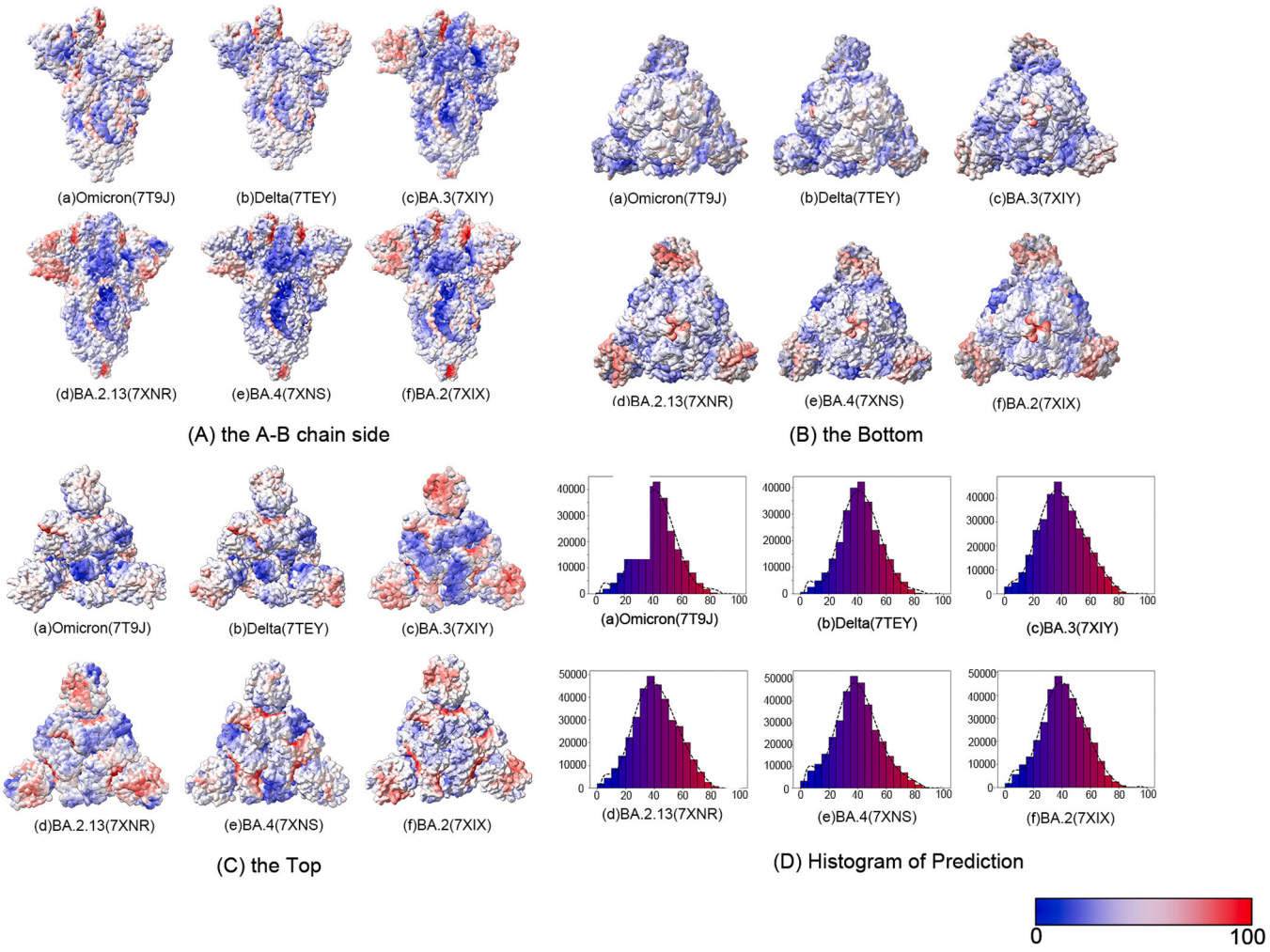

Fig. 4. The visualization results of our predictions. (A) The side views of A-B chains of the SARS-CoV-2 variants. (B) The bottom views and (C) the top views of the SARS-CoV-2 variants. (D) The histogram of our predictions which presents the distribution gap of interaction site score ranging from 0–100. The horizontal axis of the histogram is the interaction site score while the vertical axis is number of the points whose value is in the corresponding range.

$$q_{ij} = \left[\widehat{q}_{ij}^{\hat{n}}, \widehat{q}_{ij}^{\hat{u}}, \widehat{q}_{ij}^{\hat{v}}\right] = \left(n_j - n_i\right)^{\top} \cdot [\widehat{n}_i | \widehat{u}_i | \widehat{v}_i]$$

Here, we applied MLPs as trainable filters.

2.4.3. Local orientation and curvatures

To compute the tangent vectors $(\widehat{u}_i, \widehat{v}_i)$ at a low computation cost, we oriented the first tangent vector $\widehat{u}_i = \widehat{u}(x_i)$ along the geometric gradient $\nabla^{\hat{u},\hat{v}} P(x_i)$ [33], where $P(x_i) = P_i = MLP(f_i)$ and f_i was the input feature. To approximate the gradient, we used a quasi-geodesic convolution (Eq. (8)).

$$\nabla P(x_i) \leftarrow \frac{1}{N} \sum_{j=1}^{N} w(d_{ij}) \left[p_{ij}^{\hat{u}}, p_{ij}^{\hat{v}}\right] P_j \in \mathbb{R}^2 \tag{8}$$

Then we updated the tangent basis $(\widehat{u}_i, \widehat{v}_i)$ through the standard trigonometric formulae. To estimate the local curvatures for oriented point clouds efficiently, we used quasi-geodesic convolutions with Gaussian windows of radii $\sigma \in [1, 2, 3, 5, 10]$Åand quadratic filter functions to estimate the local covariances$Cov_{\sigma,i}^{\hat{u},\hat{v}}(p,p)$ and $Cov_{\sigma,i}^{\hat{u},\hat{v}}(p,q)$ [34], where

$$p = \left[\widehat{p}_{ij}^{\hat{u}}, \widehat{p}_{ij}^{\hat{v}}\right] = \left(x_j - x_i\right)^{\top} \cdot \left[\widehat{u}_i | \widehat{v}_i\right] \tag{9}$$

$$q = \left[\widehat{q}_{ij}^{\hat{u}}, \widehat{q}_{ij}^{\hat{v}}\right] = \left(n_j - n_i\right)^{\top} \cdot \left[\widehat{u}_i | \widehat{v}_i\right].$$

We approximated the 2×2 shape operator at point x_i and scaled σ with a small regularization parameter below,

$$S_{\sigma,i} = \left(\lambda^2 I_{2\times 2} + Cov_{\sigma,i}^{\hat{u},\hat{v}}(p,p)\right)^{-1} Cov_{\sigma,i}^{\hat{u},\hat{v}}\left(p,q\right) \tag{10}$$

where $\lambda = 0.1$Å in our work. We defined the Gaussian curvatures $K_{\sigma,i} = det(S_{\sigma,i})$ and mean curvatures $H_{\sigma,i} = trace(S_{\sigma,i})$ at scaleσ.

2.4.4. Trainable convolutions

Finally, we turned the input feature $f_i \in \mathbb{R}^F$ into the output feature $f_i^{'} \in \mathbb{R}^F$ based on a quasi-geodesic convolution that relied on a trainable MLP (Eq. (11)),

$$f_i^{'} \leftarrow \sum_{j=1}^{N} w(d_{ij}) MLP(p_{ij}) f_j \tag{11}$$

where the MLP was a neural network consisting of three input units, H= 8 hidden units, ReLU non-linearity and F= 16 outputs.

2.5. Pipeline of our method

We combined the steps in the previous sections and made a pipeline for our methods, as illustrated in Fig. 3. The method could be summarized as below:

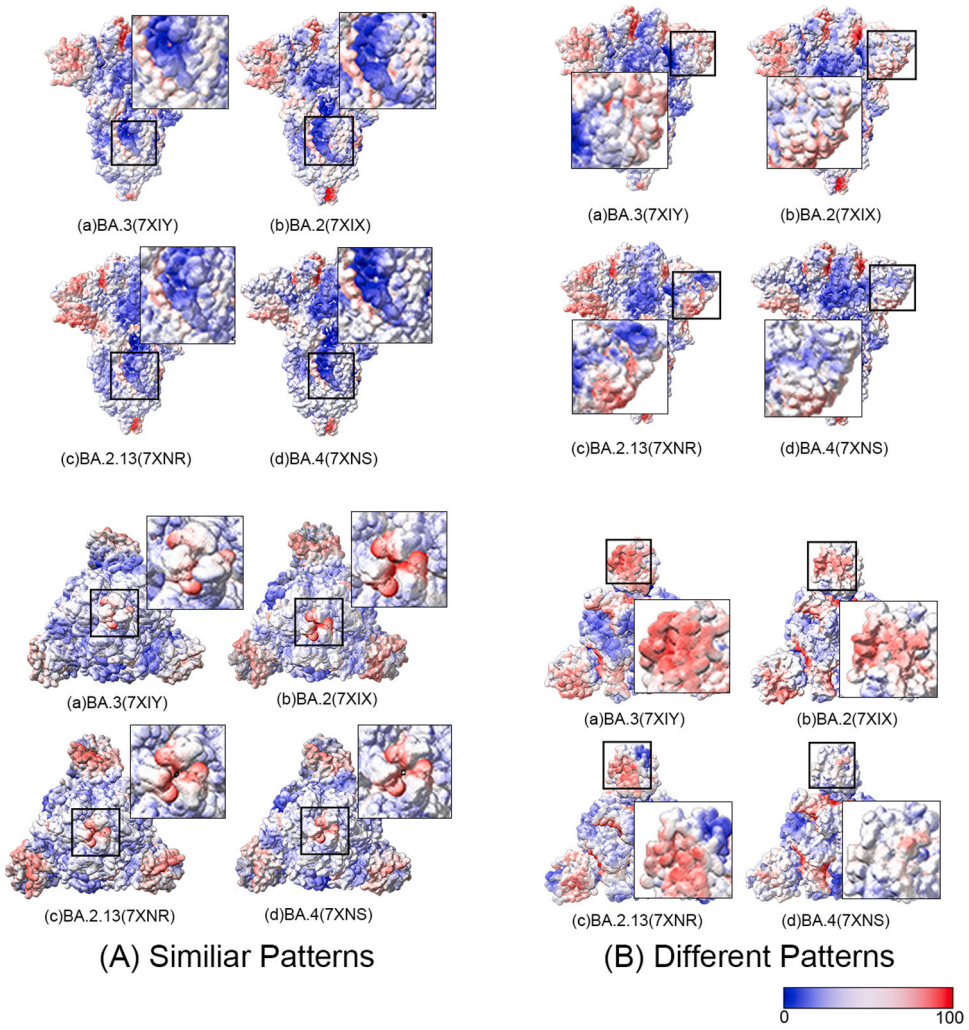

Fig. 5. The enlarged views of specific regions in Fig. 4. (A) The similar patterns and (B) the different patterns of the four Omicron subvariants.

Table 2
The correlation between histograms H_1 and H_2 in the Fig. 4 (D).

Cor	7T9J	7TEY	7XIX	7XIY	7XNR	7XNS
Omicron(7T9J)	1.0000	0.9989	0.9823	0.9611	0.9782	0.9770
Delta(7TEY)	0.9989	1.0000	0.9873	0.9659	0.9833	0.9802
BA.2(7XIX)	0.9823	0.9873	1.0000	0.9897	0.9984	0.9893
BA.3(7XIY)	0.9611	0.9659	0.9897	1.0000	0.9937	0.9902
BA.2.13(7XNR)	0.9782	0.9833	0.9984	0.9937	1.0000	0.9892
BA.4(7XNS)	0.9770	0.9802	0.9893	0.9902	0.9892	1.0000

1. We sampled surface points of SARS-CoV-2 mutant spike proteins and computed their normals using the gradient of the distance function;
2. We used the normal $\hat{n}_i$ to compute the geometric feature, i.e., mean and Gaussian curvatures at five scalesσ ranging from 1 Å to 10 Å;
3. We computed chemical features on the spike protein surface by using atom types and inverse distances to surface points, which further flowed through an MLP. Then, the contributions from the 16 nearest atoms to a surface point$x_i$were summed together and flowed through a linear transformation to obtain the chemical feature;
4. We concatenated these geometric features and chemical features to get a full feature vector of size 16;
5. The full feature vector flowed through a small MLP to predict the orientation score $P_i$for each surface point. Then we used a quasi-geodesic convolution to orient the local coordinates $(\hat{n}_i, \hat{u}_i, \hat{v}_i)$;

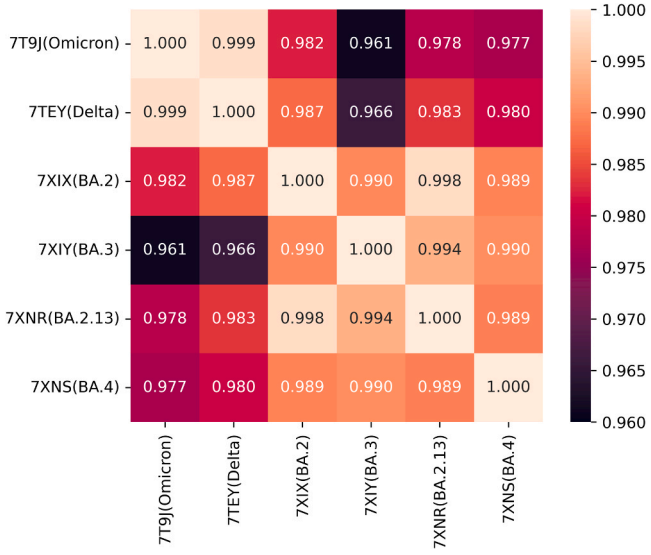

Fig. 6. The heatmap of the correlation between histograms H_1 and H_2 in Table 2.

Table 3
The positive binding site ratios of the SARS-CoV-2 variants under different threshold values.

PDB	positive ratio			
	>50	>60	>70	>80
BA.2.13(7XNR)	0.2770	0.1187	0.0284	0.0020
BA.2(7XIX)	0.2743	0.1127	0.0246	0.0029
BA.3(7XIY)	0.2627	0.1153	0.0325	0.0027
Delta(7TEY)	0.2615	0.1015	0.0238	0.0013
Omicron(7T9J)	0.2550	0.1013	0.0234	0.0014
BA.4(7XNS)	0.2367	0.1033	0.0345	0.0050

6. We applied successive trainable convolutions, MLPs and batch normalizations on the feature vector f_i. At the final step, we applied an MLP to the output of the convolutions to predict the interaction site scores.

2.6. Hardware requirements and software prerequisites

2.6.1. Hardware requirements

Models were trained on either a single NVIDIA RTX 2080 Ti or a single Tesla V100 GPU. Time and memory benchmarks were performed on a single Tesla V100.

2.6.2. Software prerequisites

Scripts were tested using the following sets of core dependencies as shown in Table 1.

2.7. Molecular docking of spike RBD and hACE2

In order to choose the appropriate threshold value for positive binding site ratios in dMaSIF prediction, molecular docking based on grid strategy was adopted to calculate the binding interface area between SARS-CoV-2 spike RBD and hACE2. The crystal structure of spike RBD-hACE2 complex (PDB ID: 6M0J) was downloaded from the Protein Data Bank. The RBD was obtained by removing hACE2 from 6M0J. Similarly, hACE2 was obtained by removing RBD from 6M0J and added with polar hydrogen atoms in the Discovery Studio Visualizer. We used pyDock (https://life.bsc.es/pid/pydockweb) for the molecular docking. Table S1 in the Supplementary materials showed the docking score, RMSD from the overall lowest energy, and buried surface area in our molecular docking. In addition, we calculated the surface areas of spike protein using VMD 1.9.3. For instance, the surface areas of BA.2 (7XIX) were calculated and used to determine the threshold value together with binding interface area in molecular docking in this study.

3. Results and discussion

After the prediction by dMaSIF method, all the PDB models were represented by different degrees of color scale, with the blue and red colors indicating low and high predicted probabilities to be interaction sites respectively. The PDB models of the SARS-CoV-2 variants consist of three chains, for better visualization, we showed the side, top and bottom views in Fig. 4(A–C). It can be seen that the red color is mainly concentrated around the bottoms of the spike proteins (Fig. 4(A)) where the RBDs are located. The Omicron (7T9J) has a similar predicted surface binding pattern with Delta (7TEY) from all the side, bottom and top views (Fig. 4(A–C)). But they show much less red color than the Omicron subvariants (7XIY, 7XNR, 7XNS and 7XIX), demonstrating the lower binding possibilities of Delta and Omicron comparing to the other Omicron subvariants. To some extent, the BA.2 (7XIX), BA.2.13 (7XNR), BA.3 (7XIY) and BA.4 (7XNS) show some similar predicted patterns but also some differences from all the views (Fig. 4(A–C)). The enlarged views of specific regions with similar and different patterns of the four Omicron subvariants in Fig. 4 were further plotted in Fig. 5. Fig. 4(D) shows the histogram of our predictions which presents the distribution gap of interaction site score ranging from 0 to 100. The interaction site score indicates the percentage of the probability to be a binding site. For instance, the interaction site score 80 means there is an 80 % probability that the site is a binding site.

Table 2 summarized the similarity of the histogram in Fig. 4 (D) while Fig. 6 showed the heatmap of the similarity matrix. To get the correlation between histograms H_1 and H_2, we used

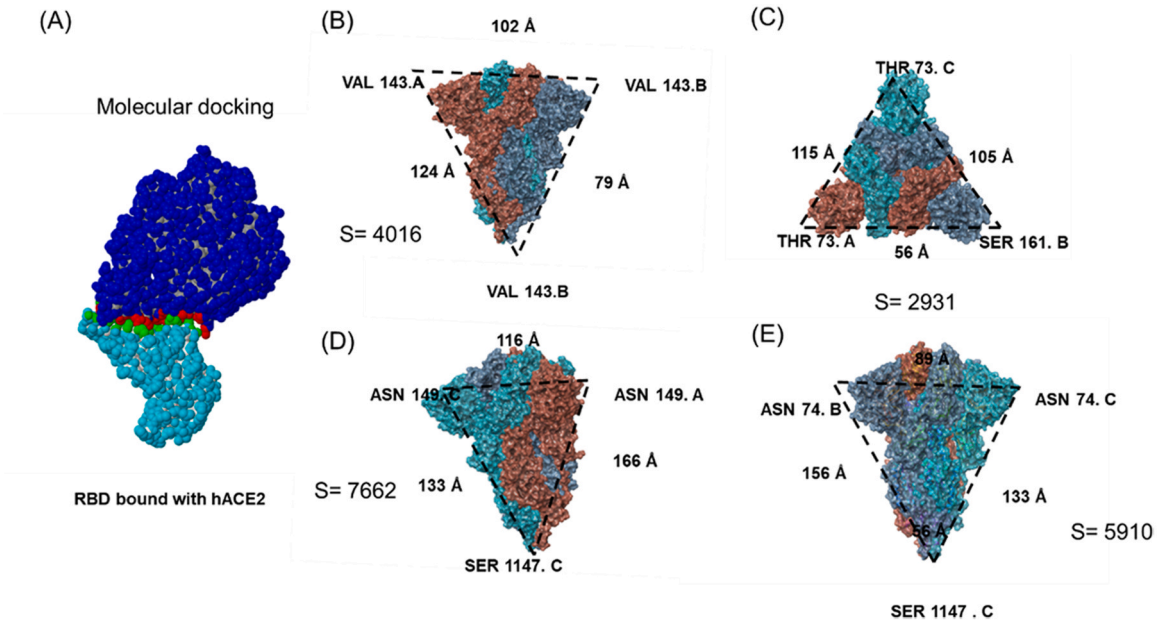

Fig. 7. (A) The binding structure of SARS-CoV-2 spike RBD (cyan) and hACE2 (blue) with interfaces colored in green and red respectively. (B) The side view of the surface area of spike protein of BA.2 (7XIX). (C) The top view of the surface area of BA.2 (7XIX). (D) The side view of the surface area of spike protein of BA.2 (7XIX). (E) The side view of the surface area of spike protein of BA.2 (7XIX). Please note the spike protein presents three different side views as shown in (B), (D) and (E).

$$d\left(H_1, H_2\right) = \frac{\sum_I (H_1(I) - \overline{H}_1)(H_2(I) - \overline{H}_2)}{\sqrt{\sum_I (H_1(I) - \overline{H}_1)^2 \sum_I (H_2(I) - \overline{H}_2)^2}} \quad (12)$$

where I denotes the histogram bin. The value of correlation ranges from –1–1, while the larger of the value indicates the higher similarity. It can be seen that 7T9J and 7TEY, 7XIX and 7XNR have the highest correlation values larger than 0.998, suggesting that the binding possibilities of Omicron and Delta, BA.2 and BA.2.13 are quite similar respectively. The BA.3 (7XIY) and BA.2.13 (7XNR), BA.4 (7XNS) and BA.3 (7XIY), BA.2 (7XIX) and BA.3 (7XIY) subvariants show the decreasing trend of correlation values of 0.9937, 0.9902 and 0.9897 respectively, demonstrating the high similarity of binding possibilities of Omicron subvariants. In addition, Delta (7TEY) and Omicron (7T9J) possess the relative lower correlation values than the Omicron subvariants since the highest value lies below 0.988 (7TEY and 7XIX) while the lowest falls below 0.962 (7T9J and 7XIY).

Table 3 showed the positive interaction site ratios (positive ratios) of SARS-CoV-2 variants under the different threshold values, from which we could compare the potential binding capacities of different variants. The positive ratio indicates the proportion of interaction points to all surface points. It could be noticed that a certain variant represents different positive ratios at the different threshold values, demonstrating that the prediction error can be influenced by the threshold. For instance, when we use 50 as the threshold, 7XNR has the largest positive ratio. However, the positive ratio of 7XNS is greatly larger than the others when 80 is chosen as the threshold. Therefore, it is essential to select an appropriate threshold to analyze the results. Herein, we adopted the molecular docking method to calculate the binding interface area between the SARS-CoV-2 spike RBD and hACE2 (Fig. 7A)). In addition, we calculated the surface areas of spike protein of BA.2 (7XIX), as shown in Fig. 7 (B-E). It is worth noting that the surface areas of the spike protein of BA.2 can be considered as triangles, and their areas can be solved by the side lengths measured by using VMD 1.9.3. The binding interface area and the total surface area of spike protein were calculated approximately to be 932.8 Å and 20519 Å respectively. According to the ratio of the two values (~0.045), we selected 70 as the appropriate threshold value to determine the positive ratios for all the SARS-CoV-2 variants. According to Table 3, BA.4 (7XNS) shows larger positive ratio than the others at threshold value 70, suggesting that BA.4 possesses more active binding behavior than the other variants. In addition, BA.2.13 (7XNR), BA.2 (7XIX) and BA.3 (7XIY) present similar and higher positive ratios than Delta (7TEY) and Omicron (7T9J) which also show similar positive ratios at thresholds 70.

Based on the results of correlation and positive ratio analyses, the surface binding patterns of Omicron subvariants are found to be similar, while the surface binding patterns of Delta and Omicron variants are similar but quite different to the Omicron subvariants. We demonstrate that BA.4 is the most infectious among the Omicron subvariants for its highest positive binding site ratio, which can be supported by the results in the literature [22,35,36]. Moreover, all of the Omicron subvariants are more spreadable than the Delta and Omicron ancestral strain, which can also be demonstrated by the published paper [37].

4. Conclusion

In this paper, we adopt the newly emerged dMaSIF method to visualize and analyze the interaction sites on the surfaces of spike proteins for SARS-CoV-2 variants. We find that the Delta and Omicron show the similar surface binding patterns while the BA.2, BA.2.13, BA.3 and BA.4 present the similar ones. The BA.4 possesses higher positive ratio than the others which may be associated with its higher transmission and infection among humans. In addition, the BA.2, BA.2.13, BA.3 have higher positive ratios than the Delta and Omicron variants which are also accordant with their transmission and infection rates. Our study offers a new deep learning method for fast end-to-end learning on spike proteins of SARS-CoV-2 variants, hopefully could advance the field of function prediction and help guide the design for new SARS-CoV-2 vaccines.

Author contributions

Yao Sun conceptualized the research design. Yao Sun, Ziyang Zheng, Yanqi Jiao, Haixin You and Junfeng An carried out the advanced computation assessments. Yao Sun and Ziyang Zheng drafted and edited the manuscript. All authors read and approved the final manuscript.

CRediT authorship contribution statement

Yao Sun: Conceptualization, Methodology, Investigation, Supervision, Writing-original draft, Writing-review & editing. **Ziyang Zheng:** Investigation, Writing-Original Draft, Visualization, Formal analysis. **Yanqi Jiao:** Investigation, Data Curation. **Haixin You:** Investigation, Formal analysis. **Junfeng An:** Investigation, Data Curation.

Conflict of interest

The authors declare no competing interests.

Data Availability

The protein structures for SARS-CoV-2 Delta (B.1.617.2) (7TEY.pdb), Omicron (B.1. 1.519) (7T9J.pdb), Omicron BA.2 (7XIX.pdb), BA.2.13 (7XNR.pdb), BA.3 (7XIY.pdb), and BA.4 (7XNS.pdb) spike proteins were obtained free of charge from the Protein Data Bank (https://www.rcsb.org/).

Acknowledgements

This work is acknowledged to the National Natural Science Foundation of China (Ref: 12102113) and the Major program of the National Natural Science Foundation of China (T2293720/T2293722).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.09.033.

References

[1] Hu B, Guo H, Zhou P, Shi Z-L. Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol 2021;19(3):141–54.
[2] Otto SP, Day T, Arino J, Colijn C, Dushoff J, Li M, et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Curr Biol 2021;31(14):R918–29.
[3] Zinatizadeh MR, Zarandi PK, Zinatizadeh M, Yousefi MH, Amani J, Rezaei N. Efficacy of mRNA, adenoviral vector, and perfusion protein COVID-19 vaccines. Biomed Pharm 2022;146:112527.
[4] Fontanet A, Autran B, Lina B, Kieny MP, Karim SSA, Sridhar D. SARS-CoV-2 variants and ending the COVID-19 pandemic. Lancet (Lond, Engl) 2021;397 (10278):952–4.
[5] Baral PK, Nuruzzaman M, Uddin MS, Ferdous M, Chowdhury IH, Smrity SZ. Severe acute respiratory syndrome coronavirus 2 invasion in the central nervous system: a host-virus deadlock. Acta Virol 2021;65(2):115–26.
[6] Nemunaitis J, Stanbery L, Senzer N. Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) infection: let the virus be its own demise. Future Virol 2020. https://doi.org/10.2217/fvl-2020-0068. Epub 2020 May 26.
[7] Shiehzadegan S, Alaghemand N, Fox M, Venketaraman V. Analysis of the delta variant B.1.617.2 COVID-19. Clin Pr 2021;11(4):778–84.
[8] Baum A, Ajithdoss D, Copin R, Zhou A, Lanza K, Negron N, et al. REGN-COV2 antibodies prevent and treat SARS-CoV-2 infection in rhesus macaques and hamsters. Science 2020;370(6520):1110–5.
[9] Cicchitto G, Cardillo L, de Martinis C, Sabatini P, Marchitiello R, Abate G, et al. Effects of casirivimab/imdevimab monoclonal antibody treatment among vaccinated patients infected by SARS-CoV-2 Delta variant. Viruses 2022;14(3):650.

[10] Kim C, Ryu D-K, Lee J, Kim Y-I, Seo J-M, Kim Y-G, et al. A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein. Nat Commun 2021;12(1):288.
[11] Kumar S, Thambiraja TS, Karuppanan K, Subramaniam G. Omicron and Delta variant of SARS-CoV-2: a comparative computational study of spike protein. Comp Stud 2022;94(4):1641–9.
[12] Kumar S, Karuppanan K, Subramaniam G. Omicron (BA.1) and sub-variants (BA.1.1, BA.2, and BA.3) of SARS-CoV-2 spike infectivity and pathogenicity: a comparative sequence and structural-based computational assessment. J Med Virol 2022;94(10):4780–91.
[13] Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinform 2021;38(6):1735–7.
[14] Wu Y, Long Y, Wang F, Liu W, Wang Y. Emergence of SARS-CoV-2 Omicron variant and strategies for tackling the infection. Immun Inflamm Dis 2022;10(12):e733.
[15] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577 (7792):706–10.
[16] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 2021;118(15):e2016239118.
[17] Biswas S, Khimulya G, Alley EC, Esvelt KM, Church GM. Low-N protein engineering with data-efficient deep learning. Nat Methods 2021;18(4):389–96.
[18] Fukuda H, Tomii K. DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. BMC Bioinform 2020;21 (1):10.
[19] Townshend R.J.L., Bedi R., Suriana P., Dror R.O. (2019) End-to-end learning on 3D protein structure for interface prediction. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.
[20] Jamasb AR, Day B, Cangea C, Liò P, Blundell TL. Deep learning for protein-protein interaction site prediction. Methods Mol Biol 2021;2361:263–88.
[21] Wang P, Zhang G, Yu Z-G, Huang G. A deep learning and XGBoost-based method for predicting protein-protein interaction sites. Front Genet 2021;12:752732.
[22] Hachmann NP, Miller J, Collier AY, Ventura JD, Yu J, Rowe M, et al. Neutralization escape by SARS-CoV-2 Omicron subvariants BA.2.12.1, BA.4, and BA.5. N Engl J Med 2022;387(1):86–8.
[23] Zhou H, Wang W, Jin J, Zheng Z, Zhou B. Graph neural network for protein-protein interaction prediction: a comparative study. Molecules 2022;27(18):6135.
[24] Hashemifar S, Neyshabur B, Khan AA, Xu J. Predicting protein-protein interactions through sequence-based deep learning. Bioinform 2018;34(17):i802–10.
[25] Wu J, Wang W, Zhang J, Zhou B, Zhao W, Su Z, et al. DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. Front Immunol 2019;10:2559.
[26] Sverrisson F., Feydy J., Correia B.E., Bronstein M.M. (2021) Fast end-to-end learning on protein surfaces. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA.
[27] Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods 2020;17(2):184–92.
[28] Monti F., Boscaini D., Masci J., Rodolà E., Svoboda J., Bronstein M. (2017) Geometric deep learning on graphs and manifolds using mixture model CNNs. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA.
[29] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera-a visualization system for exploratory research and analysis. J Comput Chem 2004;25(13):1605–12.
[30] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph 1996;14(1):33–8.
[31] Blinn JF. A generalization of algebraic surface drawing. ACM SIGGRAPH Comput Graph 1982;16(3):273.
[32] Duff T, Burgess JP, Christensen PH, Hery C, Kensler AE, Liani M, et al. Building an orthonormal basis, revisited. J Comput Graph Tech 2017;6:1.
[33] Melzi S., Spezialetti R., Tombari F., Bronstein M.M., Stefano L.D., Rodolà E. (2019) GFrames: gradient-based local reference frame for 3D shape matching. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA.
[34] Cao Y., Li D., Sun H., Assadi A.H., Zhang S.J.A. (2019) Efficient curvature estimation for oriented point clouds. ArXiv, abs/1905.10725.
[35] Yao L, Zhu KL, Jiang XL, Wang XJ, Zhan BD, Gao HX, et al. Omicron subvariants escape antibodies elicited by vaccination and BA.2.2 infection. Lancet Infect Dis 2022;22(8):1116–7.
[36] Cao Y, Yisimayi A, Jian F, Song W, Xiao T, Wang L, et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. Nature 2022;608(7923):593–602.
[37] Powell AA, Kirsebom F, Stowe J, Ramsay ME, Lopez-Bernal J, Andrews N, et al. Protection against symptomatic infection with delta (B.1.617.2) and omicron (B.1.1.529) BA.1 and BA.2 SARS-CoV-2 variants after previous infection and vaccination in adolescents in England, August, 2021-March, 2022: a national, observational, test-negative, case-control study. Lancet Infect Dis 2023;23(4): 435–44.