Computers in Biology and Medicine 151 (2022) 106212

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

D3AI-Spike: A deep learning platform for predicting binding affinity between SARS-CoV-2 spike receptor binding domain with multiple amino acid mutations and human angiotensin-converting enzyme 2

Jiaxin Han [a,b,1], Tingting Liu [c,1], Xinben Zhang [b,1], Yanqing Yang [b,d,1], Yulong Shi [b,d], Jintian Li [b,d], Minfei Ma [b,d], Weiliang Zhu [a,b,d,2,***], Likun Gong [c,2,**], Zhijian Xu [b,d,*,2]

[a] School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing, 210046, China
[b] Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China
[c] Center for Drug Safety Evaluation and Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China
[d] School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China

ARTICLE INFO

Keywords:
D3AI-Spike
ELISA
Protein-protein interaction
COVID-19
Deep learning

ABSTRACT

The number of SARS-CoV-2 spike Receptor Binding Domain (RBD) with multiple amino acid mutations is huge due to random mutations and combinatorial explosions, making it almost impossible to experimentally determine their binding affinities to human angiotensin-converting enzyme 2 (hACE2). Although computational prediction is an alternative way, there is still no online platform to predict the mutation effect of RBD on the hACE2 binding affinity until now. In this study, we developed a free online platform based on deep learning models, namely D3AI-Spike, for quickly predicting binding affinity between spike RBD mutants and hACE2. The models based on CNN and CNN-RNN methods have the concordance index of around 0.8. Overall, the test results of the models are in agreement with the experimental data. To further evaluate the prediction power of D3AI-Spike, we predicted and experimentally determined the binding affinity of a VUM (variants under monitoring) variant IHU (B.1.640.2), which has fourteen amino acid substitutions, including N501Y and E484K, and 9 deletions located in the spike protein. The predicted average affinity score for wild-type RBD and IHU to hACE2 are 0.483 and 0.438, while the determined K_{aff} values are $5.39 \pm 0.38 \times 10^7$ L/mol and $1.02 \pm 0.47 \times 10^7$ L/mol, respectively, demonstrating the strong predictive power of D3AI-Spike. We think D3AI-Spike will be helpful to the viral transmission prediction for the new emerging SARS-CoV-2 variants. D3AI-Spike is now available free of charge at https://www.d3pharma.com/D3Targets-2019-nCoV/D3AI-Spike/index.php.

1. Introduction

As the world has grappled with COVID-19 since December 2019 [1], various variants of SARS-CoV-2 virus were found one after another. Subsequently, potential immune evasions caused by the variants from the existing vaccines and monoclonal antibodies brought new challenges to public health [2,3]. The mutations in virus variants are widespread on various proteins of SARS-CoV-2, which may change different aspects of virus biology, such as pathogenicity, infectivity, transmissibility, and antigenicity, crucial mutations with immune evasion are observed on the receptor-binding domain (RBD) located on the spike protein (S protein) [4]. Human angiotensin-converting enzyme 2 (hACE2) is the receptor of SARS-CoV-2, the S protein binds to hACE2 through its RBD and is then proteolytically activated by human proteases [4–8]. It is also worth noting that a serological analysis of almost 650 SARS-CoV-2-infected individuals indicated that about 90% of the plasma or serum-neutralizing antibody activity target the spike RBD [9]. Mutations, especially multiple mutations bring more uncertainty to

* Corresponding author. Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.
** Corresponding author. Center for Drug Safety Evaluation and Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.
*** Corresponding author. School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing, 210046, China.
E-mail addresses: wlzhu@simm.ac.cn (W. Zhu), lkgong@cdser.simm.ac.cn (L. Gong), zjxu@simm.ac.cn (Z. Xu).
[1] These authors contributed equally: Jiaxin Han, Tingting Liu, Xinben Zhang, Yanqing Yang.
[2] These authors jointly supervised this work: Weiliang Zhu, Likun Gong, Zhijian Xu.

https://doi.org/10.1016/j.compbiomed.2022.106212
Received 20 September 2022; Received in revised form 11 October 2022; Accepted 15 October 2022
Available online 25 October 2022

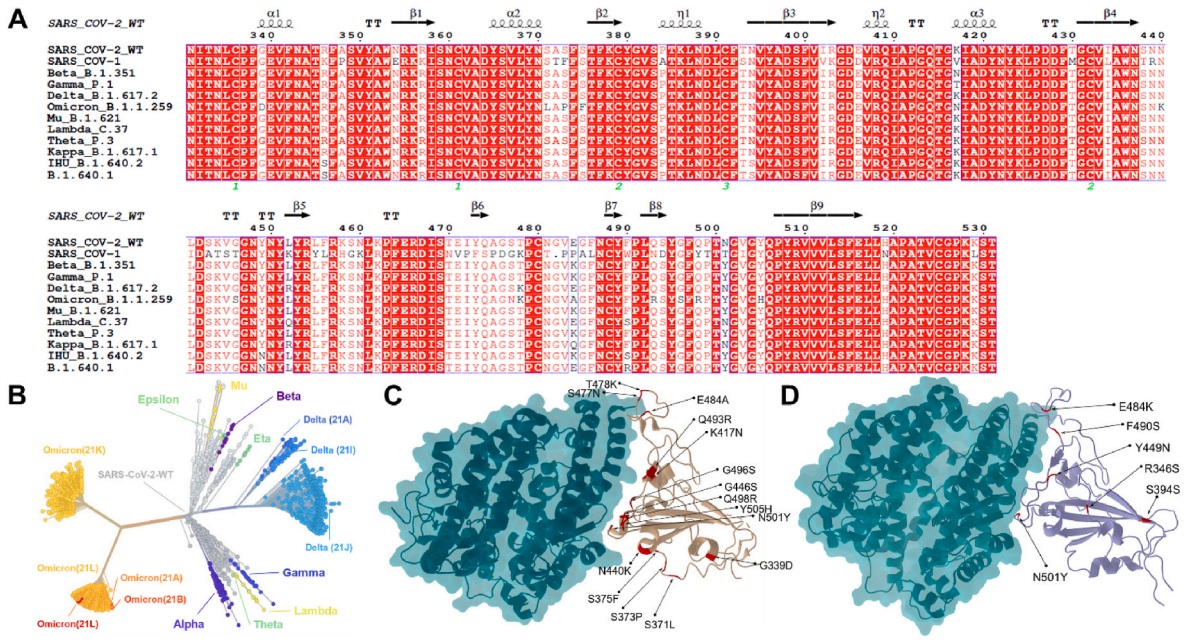

Fig. 1. Sequence alignment of SARS-COV-2 variants, illustration of Omicron and IHU RBDs interacting with hACE2, **(A)** Sequence alignments of VOCs, VOIs, and several related viruses. **(B)** Phylogenetic tree analyses of virus variants. Mutated residues are labeled in the crystal structure of chimeric Omicron RBD (strain BA.1) complexed with human ACE2 [Protein Data Bank (PDB) entry 7U0N] **(C)**. Mutated residues are labeled in the AlphaFold2 [15] predicted structure of IHU RBD (strain B.1.640.2) complexed with human ACE2 3D structures (**D**).

spike-hACE2 binding [10]. A Mutated RBD may strengthen the binding ability of S protein and hACE2, with a higher binding free energy (BFE) which is usually correlated to stronger viral infectivity [11,12]. We aligned the sequences of Variants of Concern (VOC), Variants of Interest (VOI), and the IHU sequence that harbors a very high number of mutations – 46, even higher than Omicron(Fig. 1A). Phylogenetic tree analyses according to GISAID [13] revealed that many variants with different mutations are evolved independently from each other(Fig. 1B). Fig. 1C and D illustrate the binding complexes of hACE2-S protein RBD of Omicron and IHU [14], showing that Omicron has 15 mutations and IHU has 6 mutations on the RBD region.

The affinity changes between RBD with single amino acid mutation and hACE2 have been determined through a deep mutational scanning experiment [16]. However, the number of RBD with multiple amino acid mutations is huge due to random mutations and combinatorial explosions, which make the experimental exploration impossible because of the tremendous human labor and extremely long experimental period. In addition, there is still a possibility that viral evolution will create future variants more infectious than the original SARS-CoV-2 by combining RBD co-mutations [17].

Without a doubt, predicting the possible impact of multiple amino acid mutations on the binding affinity between RBD and hACE2 in a timely manner can help fight against the potential forthcoming risk in advance [18]. Although several artificial intelligence (AI) models have been developed to estimate the mutation effect on protein-protein interactions (PPI), e.g., MuPIPR, MutaBind2, ProAffiMuSeq, and ENDscriptIIWeb server [19–23], which gain unprecedented success in PPI predicting tasks, it's unsatisfactory for these methods in predicting binding affinity between SARS-CoV-2 and hACE2 because these methods are designed for broad-spectrum protein affinity predictions. Hie et al. used a machine learning technique for natural language processing to predict the viral escape [24]. Deep mutational learning (DML) [25], a machine learning-guided protein engineering technology, could accurately predict the impact on ACE2 binding and antibody escape. Based on MD simulations, MM-GBSA, and the neural network, Chen et al. developed the NN_MM-GBSA model [26] to predict binding affinity between SARS-CoV-2 spike RBD and hACE2. Chen et al. developed a comprehensive topology-based AI model TopNetmAb, which shows good predictability in BFE change between Spike RBD and hACE2/-vaccines [27]. However, installing these programs is a difficult task for experimenters with little computer knowledge. Meanwhile, methods such as NN_MM-GBSA require large computing resources which are not available for most experimental scientists. Taken together, an online platform to predict the mutation effect of spike RBD on its hACE2 binding affinity is highly expected. To solve these problems, we developed a free online deep learning platform D3AI-spike to predict the binding affinity changes between RBD with multiple amino acid mutations and hACE2. After submitting the mutated residues on the online platform, users could quickly obtain the predicted binding affinity changes in less than a minute. The website is available at https://www.d3pharma.com/D3Targets-2019-nCoV/D3AI-Spike/index.php. The predicted results are in agreement with the experimental data. In addition, we validated the binding affinity of a VUM variant IHU by bioassay.

2. Materials and methods

2.1. Datasets preparation

The datasets of D3AI-Spike are from two sources, i.e., RBD mutational scanning and SKEMPI2 (Structural database of Kinetics and Energetics of Mutant Protein Interactions) [28]. Firstly, binding affinity change data between RBD and hACE2 was obtained from a single mutational scanning experiment [16]. In order to improve the dataset quality, we retained the RBD mutation sequence data from residues positions 331 to 531 and deleted invalid data with no affinity information in the experiment. Secondly, we collected data from the SKEMPI2, which is widely used as a benchmark set for mutant binding affinity prediction and indispensably used to train deep learning models. The mutated FASTA sequence was not provided directly by SKEMPI2. According to the protein mutation and affinity data given in SKEMPI2, the 3D structures were downloaded from the Protein Data Bank (PDB) [29]. The mutations are mapped to PDB structures, then the mapped protein 3D structures are converted to FASTA through a Python script.

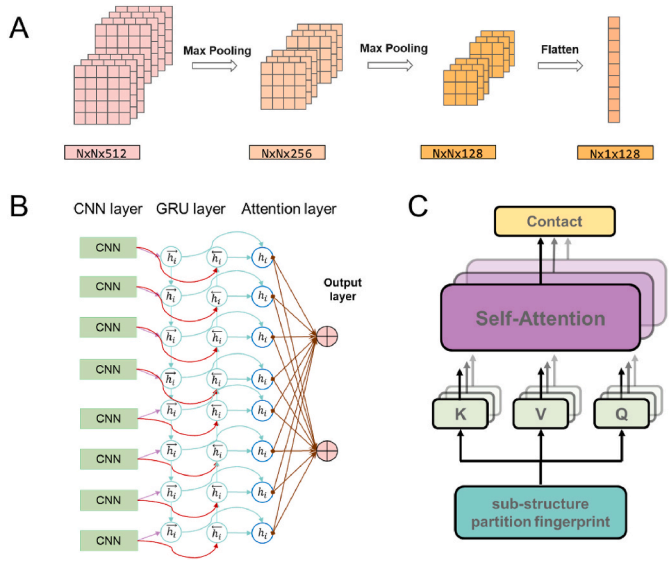

Fig. 2. The schematic illustration of CNN **(A)**, CNN-RNN **(B)**, and Transformer **(C)**.

Finally, 9518 data instances were retained in the dataset, among which 4003 were from the mutational scanning experiment and 5515 from SKEMPI2. The affinity energy values from the mutational scanning experiment and SKEMPI2 were expressed logarithmically to make them comparable. All data were saved in a specific format as fasta_1(RBD), fasta_2(ACE2), affinity change(ΔE).

2.2. Normalization of datasets

The affinity difference between two proteins is described as the affinity change ΔE, which is calculated as the difference between the variant $\log_{10}(K_{D,\ app})_{variant}$ and wild-type $\log_{10}(K_{D,\ app})_{WT}$: $\Delta E = \log_{10}(K_{D,\ app})_{WT} - \log_{10}(K_{D,\ app})_{variant}$. A positive value indicates that the variant has a higher variant hACE2 affinity than the wild type. Based on the ΔE results, four normalization methods (Min-Max, Z-Score, Sigmoid, Tanh) are used to improve the prediction accuracy. The four normalization methods are defined as:

$$\Delta E_{Min-Max} = \frac{\Delta E - \Delta E_{\min}}{\Delta E_{\max} - \Delta E_{\min}} \quad (1)$$

$$\Delta E_{Z-Score} = \frac{\Delta E - \overline{\Delta E}}{\sigma} \quad (2)$$

$$\Delta E_{Sigmoid} = \frac{1}{1 + e^{-\Delta E}} \quad (3)$$

$$\Delta E_{Tanh} = \frac{e^{\Delta E} - e^{-\Delta E}}{e^{\Delta E} + e^{-\Delta E}} \quad (4)$$

where $\Delta E_{Min\text{-}Max}$, $\Delta E_{Z\text{-}Score}$, $\Delta E_{sigmoid}$, and ΔE_{tanh} are the values after normalization. ΔE represents the affinity change relative to wild type. ΔE_{min} and ΔE_{max} are the min and max values from the original dataset. $\overline{\Delta E}$ and σ are the mean value and standard deviation of the original dataset.

Among the four different normalization methods, Min-Max and Z-Score are linear methods, while Sigmoid and Tanh are nonlinear methods. Generally speaking, normalization could improve learning ability. For different deep learning models, different normalization methods show different effects on prediction accuracy. We tried to combine different models and normalization methods to seek the best combinations.

2.3. Regression deep learning models of D3AI-Spike

SBPF (sequence-based protein fold) [30], APAAC (amphiphilic pseudo amino acid composition) [31], Quasi-seq-order (quasi-sequence-order effect) [32], and KFCT (kernel function and a conjoint triad) [33] are four methods focusing on protein sequence information. SBPF is a long-length vector with every position corresponding to an amino acid trimer. APAAC is based on SBPF with additionally extended protein hydrophobicity and hydrophilicity patterns information. KFCT encodes protein by using the continuous three amino acids frequency distribution. Quasi-seq-order follows the principle of the sequence order effect. A one-dimensional (1D) input vector is prepared from the protein FASTA and then encoded with an embedding layer by CNN (convolutional neural networks) [34] (Fig. 2A). With a GRU bidirectional recurrent neural network attached to the CNN output, CNN + RNN (Recurrent Neural Network) [35,36] can get extra sequence order information that CNN doesn't support (Fig. 2B). Transformer [37] uses a self-attention-based transformer encoder and operates on

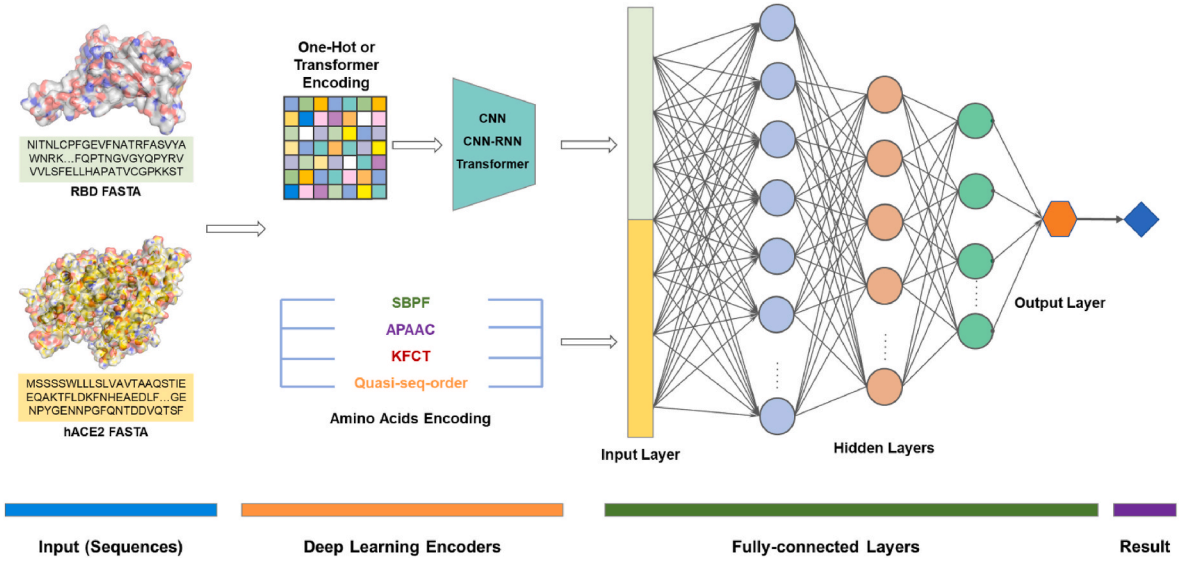

Fig. 3. Flowchart of D3AI-Spike. RBD and hACE2 FASTA are input to different deep learning models, following operations by encoders and fully connected layers, and the predicted result is obtained.

moderate-sized protein substructures, which are fed into the model (Fig. 2C). Many CNN-based 1D sequence deep learning models have achieved outstanding results. Chen et al. [27] indicated that the CNN model has a good effect on the sequence-based RBD-hACE2 affinity prediction. Zhang et al. [38] also proposed that when the training sample size is sufficiently large, LSTM embedding and CNN-based predictive model show superior performance in the RBD-hACE2 affinity prediction. In long sequence processing, RNN is expensive in time. But 1D convolutional neural networks are computationally cheap, so it is a wise idea to use the 1D convolutional neural network as a preprocessing step before RNN, which can make the sequence shorter, and extract useful information to be processed by RNN [34–36]. Dispensing with recurrence and convolutions entirely, Transformer is based totally on the attention mechanisms. Translation tasks experiments showed Transformer models are superior in quality, meanwhile, they are more parallelizable and require significantly less time to train [37].

To explore the effectiveness of each model, we conducted 10-fold cross-validation, where the dataset was split into 10 subsets — 9 for training and 1 for validation. Each run was trained until the loss becomes convergence (about 200 epochs).

2.4. Workflow of D3AI-Spike

To predict the binding affinity changes of RBD mutants and hACE2, we trained and selected several deep learning (DL)-based PPI prediction models. RBD variants' FASTA and hACE2's FASTA are encoded by One-Hot encoding and then the features are extracted through a multi-layer CNN/CNN-RNN. Meanwhile, a self-attention-based Transformer encoder, as well as encoding methods such as SPBF directly utilizing protein FASTA information, are also used to extract features from protein FASTA. Then the FASTA features of RBD and hACE2 are combined and then projected to the multi-layer perceptron. The affinity score between RBD and hACE2 is obtained through the fully connected layer. The prediction results of the top 7 models are normalized by Min-Max, and then the average value is taken as the final predicted score. The overall schematic diagram of D3AI-Spike is shown in Fig. 3.

2.5. Performance test metrics

We tried several deep learning models to predict the affinity change between RBD and hACE2. To evaluate the predictive ability of every model, we used the Pearson correlation coefficient, concordance index, and MSE as performance indicators. These indicators are defined as follows:

$$\text{Pearson correlation} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} \tag{5}$$

where N is the number of all samples. x_i and y_i are the experimental and predicted values of samples, respectively.

$$\text{Concordance index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \tag{6}$$

where η_i represents the risk score of a unit i. $1_{T_j<T_i}$: if $T_j < T_i$, $1_{T_j<T_i} = 1$, else $1_{T_j<T_i} = 0$. $1_{\eta_j>\eta_i}$: if $\eta_j > \eta_i$, $1_{\eta_j>\eta_i} = 1$, else $1_{\eta_j>\eta_i} = 0$. Factor δ_j is multiplied to discards not comparable pairs of observations.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 \tag{7}$$

where n is the number of all samples. x_i and y_i are the label and predicted values of samples, respectively.

2.6. The website of D3AI-Spike

For the user's convenience, we developed D3AI-Spike, a free online platform. Binding affinity can be predicted by inputting the RBD mutations according to the HGVS mutation naming rules [39]. After submitting RBD mutations on the webpage, the FASTA sequence will be automatically generated and fed into the deep learning models to calculate its affinity with hACE2. In less than a minute, the predicted affinity score and the 3D structure mapped with the mutant residues will be displayed on the result page.

The web page of D3AI-Spike is based on Remote Dictionary Server (Redis) and Hypertext Preprocessor (PHP). In the construction of the web page, RBD-hACE2 affinity prediction results calculated by the deep learning models were stored in the Redis database. When a user inputs a sequence, D3AI-Spike will check the database first to look up if there are any corresponding previously obtained prediction results. If the results are already in the database, the results will be displayed on the web page directly, otherwise, the calculation of affinity prediction of the new mutants will be performed and then the predicted results are written into the database. On the backend of the web page, the prediction results from the database will be re-read after the prediction calculation is complete. The usage of the Redis database helps users save time by reducing unnecessary repetitive calculations.

2.7. Binding ELISA

A non-competitive ELISA was performed to measure the affinity constant (K_{aff}) of WT SARS-CoV-2 Spike protein RBD (His Tag) (GenScript), and SARS-CoV-2 B.1.640.2 (IHU) Spike RBD protein (His Tag) (Sino Biological) against the hACE2-Fc protein (GenScript) [40]. Briefly, the 96-well plates were coated with 2, and 4 μg/mL hACE2-Fc Tag protein at 4 °C overnight, then washed with 0.1% PBST. Then the plates were blocked for 1 h at 37 °C using 3% BAS in phosphate-buffered saline (PBS) (Thermo Fisher Scientific, Waltham, MA, USA), and finally incubated with 4- fold serial dilutions of RBDs at 37 °C for 1 h. The tested concentrations were between 10 μg/mL and 9.54 pg/mL. His tag antibody (HRP) (Sino Biological) diluted 1/10,000 was applied, and the plate was incubated at 37 °C for 1 h. After washing the plate thrice with 0.1% PBST, TMB substrate (SeraCare, Milford, MA, USA) was added and the reaction was terminated with 2 M H_2SO_4, and the absorbance was read at 450 nm with an Infinite F50 microplate reader (Tecan Trading AG, Zürich, Switzerland). The following formula for calculation of affinity constant (K_{aff}) in 1/mol (M^{-1}) was used:

$$K_{aff} = \frac{(n-1)}{2(n[Ab]1 - [Ab]2)} \tag{8}$$

where n represents the ratio between the highest and the lowest hACE2 concentration for the three possible comparisons. In a comparison between two hACE2 concentrations, [Ab]1 represents the molar RBD concentration calculated for OD-50 (half of maximum OD450 nm), corresponding to the lower hACE2 concentration. [Ab]2 represents the molar RBD concentration calculated for OD-50 measured at 450 nm, corresponding to the higherhACE2 concentration. The calculation of [Ab]1 and [Ab]2 was carried out by interpolating the value of OD-50 in the curve of OD450 nm vs. RBD concentration, fitting the curve to a four-parameter logistic regression by GraphPad Prism version 9.1.1 (GraphPad Software, San Diego, California, USA). The K_{aff} value for each RBD represents the mean ± the standard deviation (SD) of the three calculated K_{aff} values.

2.8. Statistics analysis

The F test was used to test the Homogeneity of variance. Subsequently, in order to analyze whether there is a significant difference between the binding affinities for either wild type or IHU RBD to hACE2,

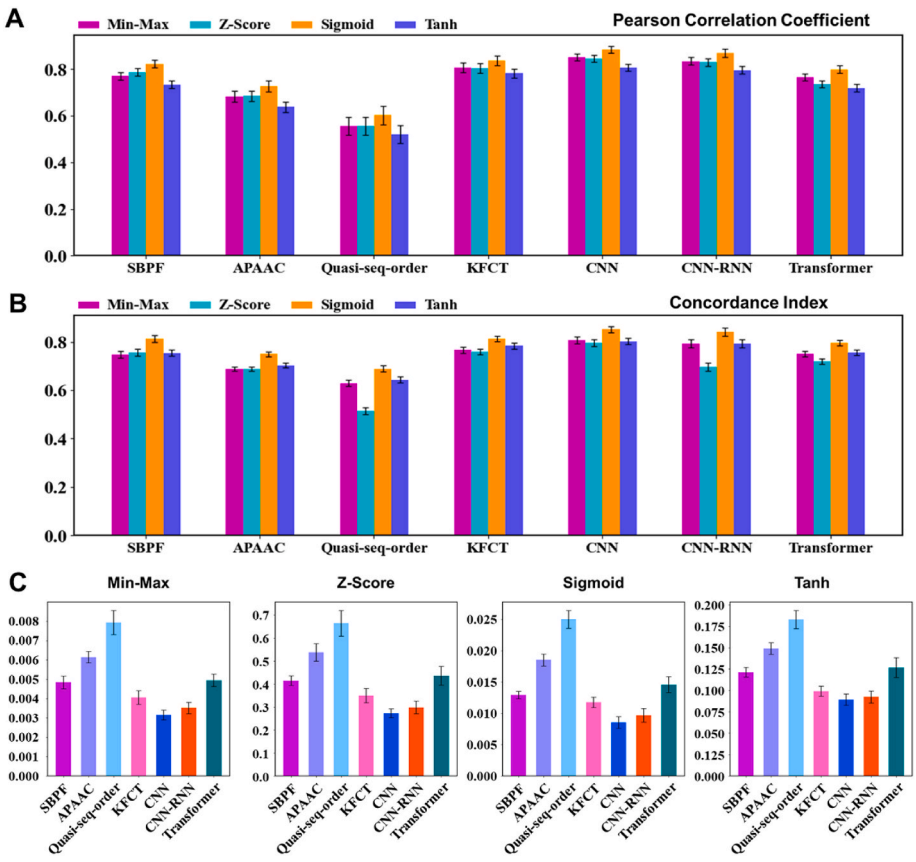

Fig. 4. Model evaluation criteria of all models. **(A)** Pearson Correlation Coefficient **(B)** Concordance Index and **(C)** MSE of each model, each model was trained through the 10-fold cross-validation method. See also Supplement Material 2.

the two-tailed unpaired Student's t-test with equal or unequal variance was used for every two groups. P values of less than 0.05 were considered to be significant.

3. Results and discussion

3.1. Performance evaluation of D3AI-Spike

For different deep learning models and data normalization methods, we use the 10-fold cross-validation, where the dataset was split into 10 subsets — 9 to train each model and 1 to evaluate its performance. Four evaluation metrics: Pearson correlation coefficient, coefficient index, mean squared error (MSE), and loss curve for the validation set were used to assess the quality of prediction and select models.

As shown in Fig. 4A, more than half of the models have a Pearson correlation coefficient >0.8, which stands for an extremely strong correlation. Fig. 4B displays that CNN with sigmoid normalization among all models possesses the best average concordance index = 0.90, while half of all models have a concordance index of around 0.8, indicating good prediction ability. In general, the performance of models normalizing by sigmoid is better than other normalization methods in most encoders, while CNN and CNN-RNN have better overall results than other models. On the other hand, encoders such as Quasi-seq-order which only consider amino acid sequence composition or amino acid triad frequency have relatively low Pearson correlation coefficient and concordance index (Fig. 4A and B), which could be partially attributed to the composition of protein sequence might ignore the interactions between residues. Quasi-seq-order and APAAC have high MSE (Fig. 4C) and high loss values (Fig. S1). As a result, we abandon predicted results by APAAC and Quasi-seq-order.

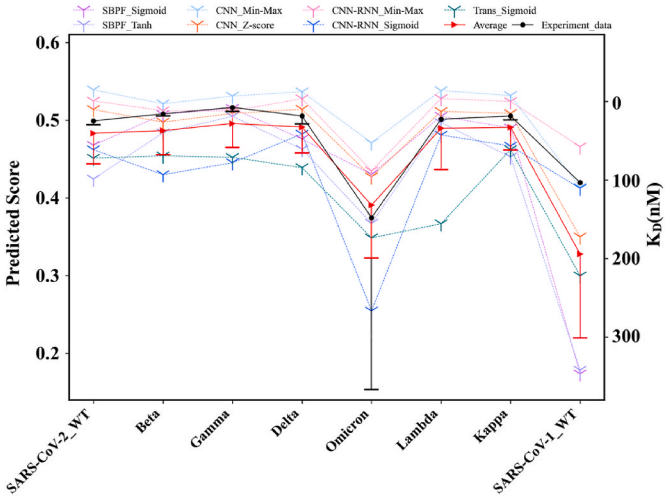

Fig. 5. Predicted results of D3AI-Spike and the experimental results. The predicted affinity score of every single model is shown in points connected by dashed lines, and the average predicted value of the 7 models is shown in the red solid line with the standard deviations shown in red error bars. Black points connected by solid lines represent experimental results with black error bars for the data points with at least two values. A higher predicted score value indicates the variant has a higher hACE2 affinity.

3.2. Test evaluation of D3AI-Spike with experimental data as the external test set

To validate the performance of different deep learning models, we

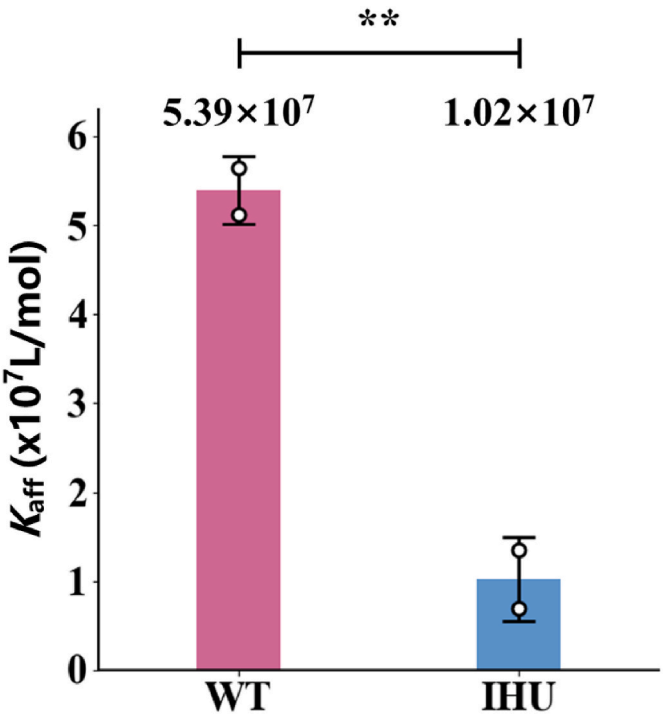

Fig. 6. RBD-hACE2 affinity constants measured by ELISA. The *P* value was determined by the two-tailed unpaired Student's t-test. $^{**}P < 0.01$.

collected affinity experimental data from literature as the external test set. However, experimental results obtained from different laboratories or different experimental methods are of high variability. To integrate these reported affinity data, we selected data from Han et al.'s work [41] as the baseline. The affinities of SARS-CoV-2 variants are converted according to the fold relationship between the same paper's WT data and the baseline, and all integrated data are shown in Table S1. The external test set was used to evaluate all 28 deep learning models (Fig. S2).

Finally, we chose the top 25% of models (top 7 models) as the final models of D3AI-Spike, namely SBPF_Sigmoid, SBPF_Tanh, CNN_Min-Max, CNN_Z-Score, CNN-RNN_Min-Max, CNN-RNN_Sigmoid, and Transformer_Sigmoid (Fig. 5). The final affinity score of D3AI-Spike is the average normalized score of the seven models. Although the predicted trend of an individual model might differ from that of the experimental results, the average predicted value of the 7 models follows almost the same trend as the experimental data (Fig. 5). Therefore, the 7 models as well as their average values were used as the website backend of D3AI-Spike.

3.3. IHU variant RBD shows significantly weaker binding affinity than the wild type (WT) to hACE2 by bioassay

To further evaluate the prediction power of D3AI-Spike, we chose a natural variant, i.e., the IHU variant, which harbors a very high number of mutations – 46, even higher than Omicron [14]. According to D3AI-Spike, the predicted average affinity score for wild-type RBD and IHU to hACE2 are 0.483 and 0.438 respectively, indicating that the affinity of IHU to hACE2 was largely reduced compared to WT. By using the non-competitive ELISA approach, we measured the affinity constant of human ACE2 to RBD_{WT} and RBD_{IHU} respectively. As shown in Fig. 6, the determined K_{aff} values of hACE2-RBD_{WT} and hACE2- RBD_{IHU} are $5.39 \pm 0.38 \times 10^7$ L/mol and $1.02 \pm 0.47 \times 10^7$ L/mol respectively (see also Fig. S3 in supplementary materials). In conclusion, the IHU variant RBD shows significantly weaker binding affinity than the WT to hACE2 with the *P* value of 0.00927 (Fig. 6).

3.4. The web page of D3AI-Spike

D3AI-Spike is a free web server that is user-friendly. Compared with the cumbersome installation of scripts or programs, users can simply

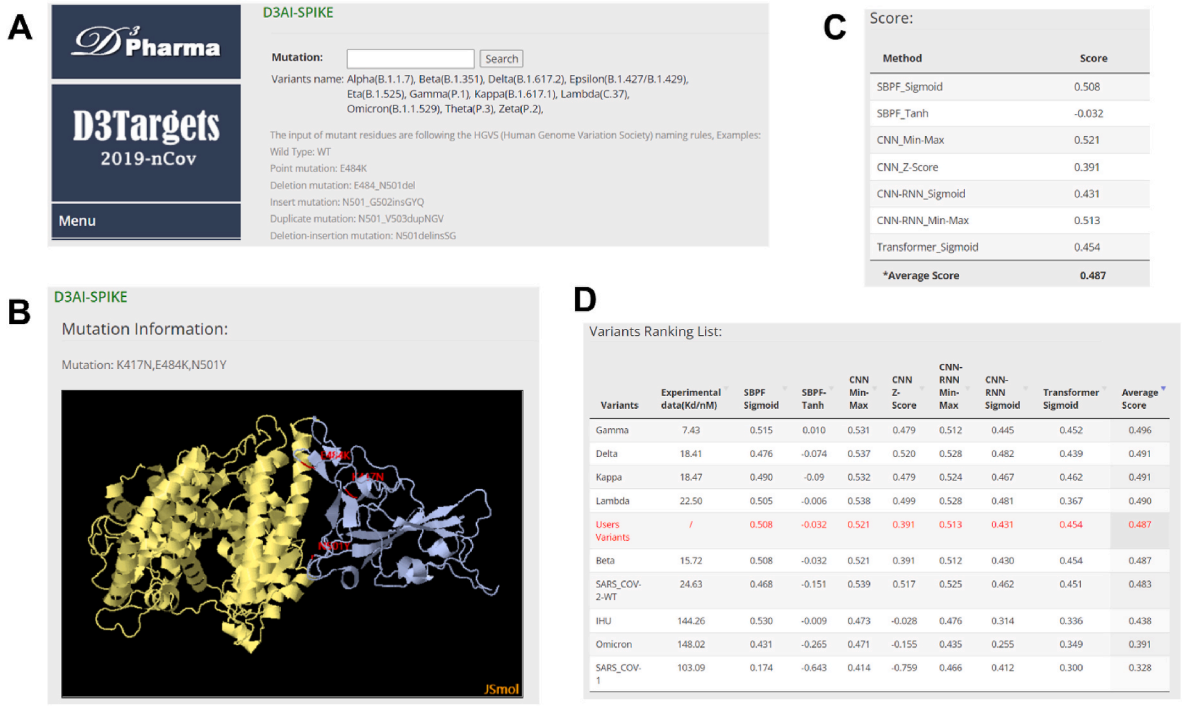

Fig. 7. Graphical interface for input and output of D3AI-Spike. (**A**) Graphical interface for input of D3AI-Spike. (**B**) The mutations (red) were mapped to the three-dimensional (3D) structure of the RBD (light blue) and hACE2 (yellow) complex. (**C**) The output affinity scores of 7 deep learning models for the user's variant. (**D**) The user's variant (red) is compared with the known variants.

submit mutations to the webserver and get results in less than a minute. To predicate the mutant variant's affinity, users only need to input residue mutant type following HGVS (Human Genome Variation Society) rules [39] or select known variants, D3AI-Spike will automatically generate the corresponding RBD sequence and then make the prediction (Fig. 7A). Usually predicting process will last no more than 1 min which is faster than the other approaches partly based on 3D structure or molecular dynamics.

The output of D3AI-Spike contains three parts. To assist intuitive understanding of the mutation information, we mapped the input mutation residues onto the 3D structure of the RBD (light blue) and hACE2 (yellow) complex (Fig. 7B). Prediction results of 7 deep learning models are displayed in Fig. 7C, and the average score is the normalized mean score of the 7 models. The higher the average score, the higher the binding affinity. For the convenience to evaluate the affinity of the user's variant, we put it with the known variants together and rank them from predicted highest binding affinity to lowest binding affinity. (Fig. 7D). D3AI-Spike is accessible free of charge at https://www.d3pharma.com/D3Targets-2019-nCoV/D3AI-Spike/index.php.

4. Conclusion

In this study, we developed D3AI-Spike as a free online platform to facilitate researchers quickly and easily predicting binding affinity between spike RBD mutant and hACE2. D3AI-Spike is a collection of deep learning models which perform the prediction in less than a minute. The average predicted value of the 7 models follows almost the same trend as the experimental data, implying the prediction power of D3AI-Spike. To further evaluate the prediction power of D3AI-Spike, we chose a natural variant named IHU, which harbors a very high number of mutations – 46, even higher than Omicron. The predicted average affinity score for hACE2-RBD_{WT} and hACE2- RBD_{IHU} are 0.483 and 0.438, while the determined K_{aff} values are $5.39 \pm 0.38 \times 10^7$ L/mol and $1.02 \pm 0.47 \times 10^7$ L/mol, demonstrating the prediction power of D3AI-Spike. We hope D3AI-Spike will be helpful to the viral transmission prediction for the new emerging SARS-CoV-2 variants.

Data and software availability

D3AI-Spike is accessible free as a web server at https://www.d3pharma.com/D3Targets-2019-nCoV/D3AI-Spike/index.php. The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

This work was supported by the Key project at central government level: The ability establishment of sustainable use for valuable Chinese medicine resources (2060302), National Key R&D Program of China (2016YFA0502301) and Natural Science Foundation of Shanghai (21ZR1475600, 22S11902100).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106212.

References

[1] P. Zhou, X.L. Yang, X.G. Wang, B. Hu, L. Zhang, W. Zhang, H.R. Si, Y. Zhu, B. Li, C. L. Huang, H.D. Chen, J. Chen, Y. Luo, H. Guo, R.D. Jiang, M.Q. Liu, Y. Chen, X. R. Shen, X. Wang, X.S. Zheng, K. Zhao, Q.J. Chen, F. Deng, L.L. Liu, B. Yan, F. X. Zhan, Y.Y. Wang, G.F. Xiao, Z.L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (2020) 270–273.

[2] S. Molaei, M. Dadkhah, V. Asghariazar, C. Karami, E. Safarzadeh, The immune response and immune evasion characteristics in SARS-CoV, MERS-CoV, and SARS-CoV-2: vaccine design strategies, Int. Immunopharm. 92 (2021), 107051.

[3] W.T. Harvey, A.M. Carabelli, B. Jackson, R.K. Gupta, E.C. Thomson, E.M. Harrison, C. Ludden, R. Reeve, A. Rambaut, C.-G.U. Consortium, S.J. Peacock, D. L. Robertson, SARS-CoV-2 variants, spike mutations and immune escape, Nat. Rev. Microbiol. 19 (2021) 409–424.

[4] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2, Science 367 (2020) 1444–1448.

[5] J. Shang, Y. Wan, C. Luo, G. Ye, Q. Geng, A. Auerbach, F. Li, Cell entry mechanisms of SARS-CoV-2, Proc. Natl. Acad. Sci. U. S. A. 117 (2020) 11727–11734.

[6] K.K. Chan, D. Dorosky, P. Sharma, S.A. Abbasi, J.M. Dye, D.M. Kranz, A.S. Herbert, E. Procko, Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2, Science 369 (2020) 1261–1265.

[7] S. Yu, X. Zheng, B. Zhou, J. Li, M. Chen, R. Deng, G. Wong, D. Lavillette, G. Meng, SARS-CoV-2 spike engagement of ACE2 primes S2' site cleavage and fusion initiation, Proc. Natl. Acad. Sci. U. S. A. 119 (2022), e2111199119.

[8] M. Letko, A. Marzi, V. Munster, Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses, Nat. Microbiol. 5 (2020) 562–569.

[9] L. Piccoli, Y.J. Park, M.A. Tortorici, N. Czudnochowski, A.C. Walls, M. Beltramello, C. Silacci-Fregni, D. Pinto, L.E. Rosen, J.E. Bowen, O.J. Acton, S. Jaconi, B. Guarino, A. Minola, F. Zatta, N. Sprugasci, J. Bassi, A. Peter, A. De Marco, J. C. Nix, F. Mele, S. Jovic, B.F. Rodriguez, S.V. Gupta, F. Jin, G. Piumatti, G. Lo Presti, A.F. Pellanda, M. Biggiogero, M. Tarkowski, M.S. Pizzuto, E. Cameroni, C. Havenar-Daughton, M. Smithey, D. Hong, V. Lepori, E. Albanese, A. Ceschi, E. Bernasconi, L. Elzi, P. Ferrari, C. Garzoni, A. Riva, G. Snell, F. Sallusto, K. Fink, H.W. Virgin, A. Lanzavecchia, D. Corti, D. Veesler, Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology, Cell 183 (2020) 1024–1042 e1021.

[10] O.A. MacLean, R.J. Orton, J.B. Singer, D.L. Robertson, No evidence for distinct types in the evolution of SARS-CoV-2, Virus Evol. 6 (2020), veaa034.

[11] Z. Jahanafrooz, Z. Chen, J. Bao, H. Li, L. Lipworth, X. Guo, An overview of human proteins and genes involved in SARS-CoV-2 infection, Gene 808 (2022), 145963.

[12] R. Mukherjee, R. Satardekar, Why are some coronavirus variants more infectious? J. Biosci. 46 (2021).

[13] S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health, Global Chall. 1 (2017) 33–46.

[14] P. Colson, J. Delerce, E. Burel, J. Dahan, A. Jouffret, F. Fenollar, N. Yahi, J. Fantini, B. La Scola, D. Raoult, Emergence in southern France of a new SARS-CoV-2 variant harbouring both N501Y and E484K substitutions in the spike protein, Arch. Virol. 167 (2022) 1185–1190.

[15] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A. A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589.

[16] T.N. Starr, A.J. Greaney, S.K. Hilton, D. Ellis, K.H.D. Crawford, A.S. Dingens, M. J. Navarro, J.E. Bowen, M.A. Tortorici, A.C. Walls, N.P. King, D. Veesler, J. D. Bloom, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding, Cell 182 (2020) 1295–1310, e1220.

[17] S.B. Kadam, G.S. Sukhramani, P. Bishnoi, A.A. Pable, V.T. Barvkar, SARS-CoV-2, the pandemic coronavirus: molecular and structural insights, J. Basic Microbiol. 61 (2021) 180–202.

[18] J. Zahradnik, S. Marciano, M. Shemesh, E. Zoler, D. Harari, J. Chiaravalli, B. Meyer, Y. Rudich, C. Li, I. Marton, O. Dym, N. Elad, M.G. Lewis, H. Andersen, M. Gagne, R.A. Seder, D.C. Douek, G. Schreiber, SARS-CoV-2 variant prediction and antiviral drug design are enabled by RBD in vitro evolution, Nat. Microbiol. 6 (2021) 1188–1198.

[19] C. Geng, A. Vangone, G.E. Folkers, L.C. Xue, A. Bonvin, iSEE: interface structure, evolution, and energy-based machine learning predictor of binding affinity changes upon mutations, Proteins 87 (2019) 110–119.

[20] N. Zhang, Y. Chen, H. Lu, F. Zhao, R.V. Alvarez, A. Goncearenco, A.R. Panchenko, M. Li, MutaBind2: predicting the impacts of single and multiple mutations on protein-protein interactions, iScience 23 (2020), 100939.

[21] G. Zhou, M. Chen, C.J.T. Ju, Z. Wang, J.Y. Jiang, W. Wang, Mutation effect estimation on protein-protein interactions using deep contextualized representation learning, NAR Genom. Bioinform. 2 (2020), lqaa015.

[22] S. Jemimah, M. Sekijima, M.M. Gromiha, ProAffiMuSeq: sequence-based method to predict the binding free energy change of protein-protein complexes upon mutation using functional classification, Bioinformatics 36 (2020) 1725–1730.

[23] X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server, Nucleic Acids Res. 42 (2014) W320–W324.

[24] B. Hie, E.D. Zhong, B. Berger, B. Bryson, Learning the language of viral evolution and escape, Science 371 (2021) 284–288.

[25] J.M. Taft, C.R. Weber, B. Gao, R.A. Ehling, J. Han, L. Frei, S.W. Metcalfe, M. Overath, A. Yermanos, W. Kelton, S.T. Reddy, Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor binding domain, Cell 185 (2022) 4008–4022.

[26] C. Chen, V.S. Boorla, D. Banerjee, R. Chowdhury, V.S. Cavener, R.H. Nissly, A. Gontu, N.R. Boyle, K. Vandegrift, M.S. Nair, S.V. Kuchipudi, C.D. Maranas,

Computational prediction of the effect of amino acid changes on the binding affinity between SARS-CoV-2 spike RBD and human ACE2, Proc. Natl. Acad. Sci. U. S. A. 118 (2021), e2106480118.
[27] 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 (2022) 412–422.
[28] J. Jankauskaite, B. Jimenez-Garcia, J. Dapkunas, J. Fernandez-Recio, I.H. Moal, Skempi 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation, Bioinformatics 35 (2019) 462–469.
[29] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I. N. Shindyalov, P.E. Bourne, The protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242.
[30] M. Reczko, H. Bohr, The DEF data base of sequence based protein fold class predictions, Nucleic Acids Res. 22 (1994) 3616–3619.
[31] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.
[32] K.C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochem. Biophys. Res. Commun. 278 (2000) 477–483.
[33] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Predicting protein-protein interactions based only on sequences information, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 4337–4341.
[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90.
[35] K. Cho, B.v.M.C. Gulcehre, D. Bahdanau, F.B.H. Schwenk, Y. Bengio, Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation, EMNLP, 2014, pp. 1724–1734.
[36] J. Schmidhuber, S. Hochreiter, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[37] R. Singh, J. Lanchantin, A. Sekhon, Y. Qi, Attend and predict: understanding gene regulation by selective attention on chromatin, Adv. Neural Inf. Process. Syst. 30 (2017) 6785–6795.
[38] R. Zhang, S. Ghosh, R. Pal, Predicting binding affinities of emerging variants of SARS-CoV-2 using spike protein sequencing data: observations, caveats and recommendations, Briefings Bioinf. 23 (2022).
[39] J.T. den Dunnen, R. Dalgleish, D.R. Maglott, R.K. Hart, M.S. Greenblatt, J. McGowan-Jordan, A.F. Roux, T. Smith, S.E. Antonarakis, P.E. Taschner, HGVS recommendations for the description of sequence variants: 2016 update, Hum. Mutat. 37 (2016) 564–569.
[40] J.D. Beatty, B.G. Beatty, W.G. Vlahos, Measurement of monoclonal antibody affinity by non-competitive enzyme immunoassay, J. Immunol. Methods 100 (1987) 173.
[41] P. Han, L. Li, S. Liu, Q. Wang, D. Zhang, Z. Xu, P. Han, X. Li, Q. Peng, C. Su, B. Huang, D. Li, R. Zhang, M. Tian, L. Fu, Y. Gao, X. Zhao, K. Liu, J. Qi, G.F. Gao, P. Wang, Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2, Cell 185 (2022) 630–640 e610.

Jiaxin Han is a postgraduate at Nanjing University of Chinese Medicine. His research interests are deep learning, molecular docking and virtual screening. His affiliation is with School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing, 210046, China; Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.

Tingting Liu is a biological engineer. Her research interest is biological experiments. Her affiliation is with the Center for Drug Safety Evaluation and Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China.

Xinben Zhang got his master's degree at East China University of Science and Technology. His research interest is software development. His affiliation is with Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.

Yanqing Yang is a postgraduate at Shanghai Institute of Materia Medica. His research interests are deep learning, molecular docking and virtual screening. His affiliation is with Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Yulong Shi is a Ph.D. student at Shanghai Institute of Materia Medica. His research interest is molecular docking method development. His affiliation is with Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Jintian Li is a Ph.D. student at Shanghai Institute of Materia Medica. Her research interest is deep learning. Her affiliation is with Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Minfei Ma is a postgraduate at Shanghai Institute of Materia Medica. Her research interests are deep learning, molecular docking and virtual screening. Her affiliation is with Stake Key Laboratory of Drug Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Professor Zhijian Xu got his Ph.D. degree at Shanghai Institute of Materia Medica in 2012. His research interests include computer-aided drug design, computational chemistry, computational biology and artificial intelligence. More information could be found at the website: https://www.researchgate.net/profile/Zhijian_Xu.

Professor Weiliang Zhu received his Ph.D. degree from Shanghai Institute of Materia Medica in 1998. His main research fields are computer-aided drug design, computational biology, computational chemistry and pharmaceutical chemistry, with a special focus on the theoretical research and method development of drug design.

Professor Likun Gong received her Ph.D. degree from Shanghai Institute of Materia Medica in 2005. Her research interests are molecular pharmacology and toxicology, vaccine, medical immunology and drug discovery.