Journal of Molecular Graphics and Modelling 118 (2023) 108364

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Binding affinity prediction for antibody–protein antigen complexes: A machine learning analysis based on interface and surface areas

Yong Xiao Yang [a], Pan Wang [a,b], Bao Ting Zhu [a,b,*]

[a] Shenzhen Key Laboratory of Steroid Drug Discovery and Development, School of Medicine, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China
[b] Shenzhen Bay Laboratory, Shenzhen, 518055, China

A B S T R A C T

Specific antibodies can bind to protein antigens with high affinity and specificity, and this property makes them one of the best protein-based therapeutics. Accurate prediction of antibody–protein antigen binding affinity is crucial for designing effective antibodies. The current predictive methods for protein–protein binding affinity usually fail to predict the binding affinity of an antibody–protein antigen complex with a comparable level of accuracy. Here, new models specific for antibody–antigen binding affinity prediction are developed according to the different types of interface and surface areas present in antibody–antigen complex. The contacts-based descriptors are also employed to construct or train different models specific for antibody–protein antigen binding affinity prediction. The results of this study show that (***i***) the area-based descriptors are slightly better than the contacts-based descriptors in terms of the predictive power; (***ii***) the new models specific for antibody–protein antigen binding affinity prediction are superior to the previously-used general models for predicting the protein–protein binding affinities; (***iii***) the performances of the best area-based and contacts-based models developed in this work are better than the performances of a recently-developed graph-based model (*i.e.*, CSM-AB) specific for antibody–protein antigen binding affinity prediction. The new models developed in this work would not only help understand the mechanisms underlying antibody–protein antigen interactions, but would also be of some applicable utility in the design and virtual screening of antibody-based therapeutics.

1. Introduction

Antibodies are the central players of the humoral immune response, which can recognize different antigens [1,2]. Rational design of specific antibody molecules has become an important part of modern biopharmaceutical industries [3–6]. Accurate prediction of the antibody–antigen binding affinity is crucial for the virtual screening of the antibody candidates during the rational design of therapeutic antibodies [7–11].

In the past several years, computational methods have been developed to predict different antibody properties [3,12], such as antibody solubility and aggregation [13–15], immunogenicity or degree of humanness [16], and structure [17–19]. When the antigens are protein or peptide molecules, the antibody–antigen interaction is a type of protein–protein interactions, and the predictive models for protein–protein binding affinity theoretically can also be used to predict the antibody–protein antigen binding affinity. However, most of the presently-available models for protein–protein binding affinity prediction generally fail to obtain the same predictive power for antibody–protein antigen binding affinity [20]. Therefore, it is necessary to develop effective models that can be used for prediction of the antibody–protein antigen binding affinities.

Currently, several methods are available for antibody–antigen binding affinity prediction, and these methods can be generally classified into two main categories, *i.e.*, the sequence-based methods [21–24] and the structure-based methods [25,26]. Recently, a new structure-based antibody–antigen binding affinity predictive model (CSM-AB) was developed by modelling the binding interfaces using graph-based descriptors [27]. The contacts-based and area-based descriptors have been proven effective in protein–protein binding affinity prediction in a few earlier studies [28–30]. It is possible that these two classes of effective descriptors can be further expanded to enhance the accuracy for predicting the antibody-protein antigen binding affinities.

In this work, the effectiveness of the area-based and contacts-based descriptors on antibody-protein antigen binding affinity prediction is investigated by constructing and training different predictive models. By evaluating the performances of these models on antibody–protein antigen binding affinity prediction, we found that the best area-based and contacts-based models developed in this study are superior to the previous area-based and contacts-based models constructed for general protein–protein binding affinity prediction [29–31] and the

* Corresponding author. Shenzhen Key Laboratory of Steroid Drug Discovery and Development, School of Medicine, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China.
E-mail address: BTZhu@CUHK.edu.cn (B.T. Zhu).

https://doi.org/10.1016/j.jmgm.2022.108364
Received 6 July 2022; Received in revised form 8 October 2022; Accepted 11 October 2022
Available online 29 October 2022

graph-based CSM-AB model specifically designed for antibody–antigen binding affinity prediction [27].

2. Materials and methods

2.1. Datasets and descriptors

The antibody–antigen binding affinity data used in this work were obtained from SAbDab (http://opig.stats.ox.ac.uk/webapps/newsabdab/sabdab/) [32] and an expanded benchmark for antibody–antigen docking and affinity prediction [20]. In the SAbDab, there were 746 antibody–antigen complexes (accessed on January 17, 2022) [32]. In the expanded benchmark for antibody–antigen docking and affinity prediction (https://github.com/piercelab/antibody_benchmark) (accessed on November 17, 2020), there were 51 antibody–antigen complexes [20].

Similar to our previous work on the protein–protein binding affinity prediction [29], several criteria were adopted in this study to select the data with high quality from the original sources: (1) The resolution of the structure of the antibody–antigen complex is $\leq$ 3.25 Å. (2) Both the heavy and light chains are present in the antibody structure. (3) The antigen type is a protein or a peptide. (4) The number of residues contained in the antibody structure is $\geq$ 400. (5) The number of residues contained in the antigen structure is greater than 25 but smaller than (or equal to) the number of residues in the antibody structure. (6) One antibody only binds to one antigen, and the opposite is also true.

After filtering, 290 antibody–antigen complexes from SAbDab [32] were placed into the training and validation sets (pre-*Set1*), and 33 complexes from the expanded benchmark [20] formed the test set (*Set2*). Since there were 28 shared antibody–antigen complexes in both pre-*Set1* and *Set2*, they were removed from pre-*Set1* to avoid the cross interference between different sets, which forms the *Set1* (containing 262 complexes).

The area-based descriptors were calculated using the same methods reported in the previous work [29]. The residue–residue contact areas across protein–protein binding interface were calculated using $Q_{contact}$ [33], and the solvent accessible surface area of atoms were computed using dr_sasa [34]. The 20 amino acid types were divided into four groups: basic *AA*s (basic amino acids: H, R and K), nonpolar *AA*s (nonpolar (hydrophobic) amino acids: I, F, L, W, A, M, P and V), polar *AA*s (polar but uncharged amino acids: C, N, G, S, Q, Y and T), and acidic *AA*s (acidic amino acids: D and E).

The surface areas were classified into 8 types, which included 4 types of *RSA* (Receptor Surface Area) and 4 types of *LSA* (Ligand Surface Area): *RSA* of basic *AA*s (A_1), *RSA* of nonpolar *AA*s (A_2), *RSA* of polar *AA*s (A_3), and *RSA* of acidic *AA*s (A_4); *LSA* of basic *AA*s (A_5), *LSA* of nonpolar *AA*s (A_6), *LSA* of polar *AA*s (A_7), and *LSA* of acidic *AA*s (A_8). In this study, an antibody was defined as a receptor, and a protein antigen as a ligand. The interface areas were categorized into 10 types: basic *AA*s ~ basic *AA*s (A_9), nonpolar *AA*s ~ nonpolar *AA*s (A_{10}), polar *AA*s ~ polar *AA*s (A_{11}), acidic *AA*s ~ acidic *AA*s (A_{12}), basic *AA*s ~ nonpolar *AA*s (A_{13}), basic *AA*s ~ polar *AA*s (A_{14}), basic *AA*s ~ acidic *AA*s (A_{15}), nonpolar *AA*s ~ polar *AA*s (A_{16}), nonpolar *AA*s ~ acidic *AA*s (A_{17}), and polar *AA*s ~ acidic *AA*s (A_{18}).

The contacts-based descriptors at the amino acid level were calculated using PRODIGY [30,31]. The residues were categorized into three groups: polar residues (C, H, N, Q, S, T, W), nonpolar residues (A, F, G, I, L, V, M, P, Y), and charged residues (E, D, K, R) [30]. There were eight contacts-based descriptors: number of interface contacts of charged residues ~ charged residues, number of interface contacts of charged residues ~ polar residues, number of interface contacts of charged residues ~ nonpolar residues, number of interface contacts of polar residues ~ polar residues, number of interface contacts of nonpolar residues ~ polar residues, number of interface contacts of nonpolar residues ~ nonpolar residues, percentage of nonpolar NIS (Non-Interacting Surface) residues, and percentage of charged NIS residues.

Table 1
The adopted descriptors and formations of the four groups of models for affinity prediction.

Model Type	Adopted descriptors	Formation	Equation No.
Linear model	Different combinations of the 18 area-based or 8 contacts-based descriptors	$\log(K) = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$	(1)
Constructed nonlinear model	different combinations based on no more than 5 variables from the 21 area-based descriptors or different combinations of the 10 area-based descriptors	$\log(K) \propto -x_1^{k1} x_2^{k2} \ldots x_n^{kn}$	(2)
		$\log(K) = ax + b$	(3)
Neural network or random forest model	18 area-based or 8 contacts-based descriptors	No explicit formation	–
Mixed model	Different combinations of the 14 representative nonlinear (neural network or random forest) models	$\log(K) = a_1m_1 + a_2m_2 + \ldots + a_nm_n + b$	(4)

Note: K is the experimentally-determined equilibrium dissociation constant or inhibition constant; x_1, x_2, …, x_n are the descriptors; m_1, m_2, …, m_n are the neural network or random forest models; k_1, k_2, …, k_n are the power exponents taken from four values: −1, −0.5, 0.5, 1; x is the constructed nonlinear term; a_1, a_2, …, a_n, a are the coefficients, and b is the constant term.

The PDB codes, numbers of residues in the antibody and antigen, experimental binding affinities ($\log(K)$, where K is the equilibrium dissociation constant or the equilibrium inhibition constant), values of the 18 area-based descriptors and 8 contacts-based descriptors are summarized in **Supplementary File 1**.

2.2. Constructing or training the predictive models

In our recent work, the effective models for protein–protein binding affinity prediction were constructed or trained manually, by linear regression and neural network [29]. The same methods were used in this study to construct or train models specific for predicting antibody–antigen binding affinities. Four groups of models were trained or constructed to find the models with the best predictive ability for the antibody–antigen binding affinity prediction: the linear models trained using linear regression; the nonlinear models manually constructed; the nonlinear models trained using neural network or random forest; and lastly, the mixed models trained using linear regression of the neural network or random forest models. The adopted descriptors and formations of the models are shown in Table 1.

2.2.1. Adopted descriptors

The adopted descriptors in the linear model included different combinations based on the 18 area-based or 8 contacts-based descriptors; the descriptors in the constructed nonlinear model included different combinations based on no more than 5 variables from the 21 area-based descriptors (18 original descriptors, one receptor surface area (RSA) (A_{19}), one ligand surface area (LSA) (A_{20}), and one total interface area (A_{21})) or different combinations base on the 10 contacts-based descriptors (8 original descriptors, one total number of intermolecular contacts, and one total percentage of nonpolar and charged NIS residues); the descriptors adopted in the neural network and random forest models included all 18 area-based or 8 contacts-based descriptors; the descriptors in the mixed model included different combinations based on 14 representative area-based neural network (or random forest) models or 14 representative contacts-based neural network (or random forest) models.

2.2.2. Formations of the four groups of models

The formations of the four groups of models are listed in Table 1. The

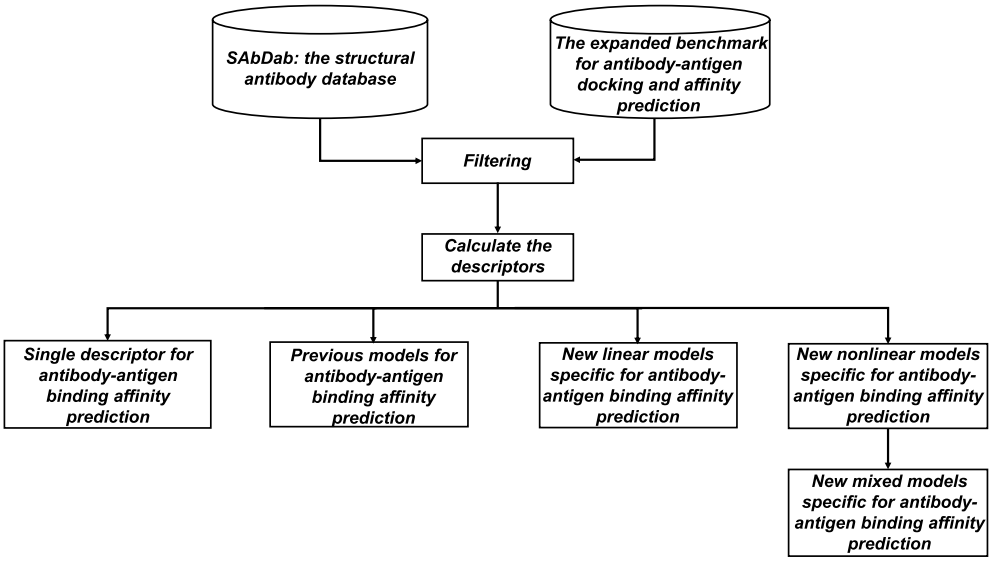

Fig. 1. Flowchart of the antibody–antigen binding affinity prediction. The essential steps include: ***1***) filtering of the original data; ***2***) calculation of the descriptors (area-based and contacts-based); and ***3***) training or constructing models specific for antibody–antigen binding affinity prediction. For comparison, some of the previously-reported models are also used to predict the antibody–antigen binding affinity.

linear and constructed nonlinear models had the explicit linear and nonlinear formations, respectively; the neural network and random forest models did not have explicit formations; the mixed models were the linear combination of the representative neural network or random forest models.

2.2.3. Set usage

When training linear model, neural network model, random forest and mixed models, *Set1* was divided into the training and validation sets evenly and stochastically, and *Set2* was used as the test set. The set division was repeated 100 times for linear formations and 30 times for

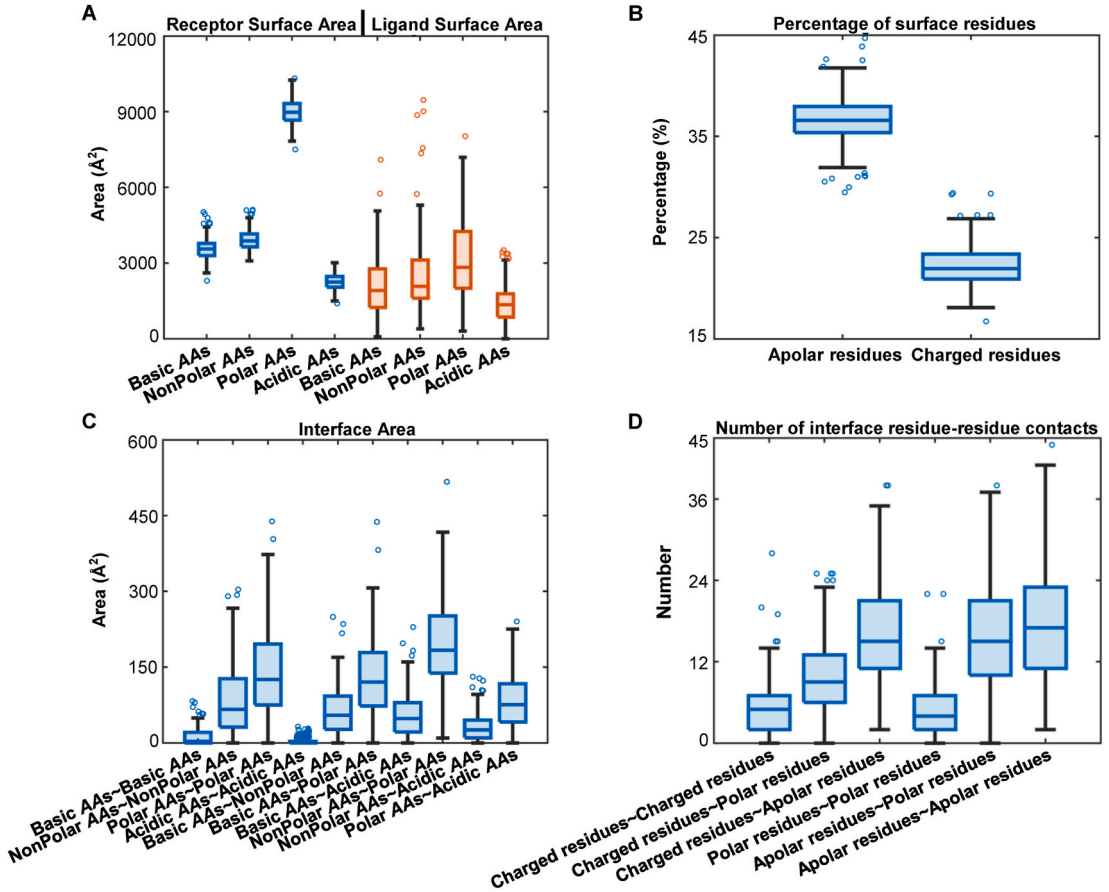

Fig. 2. Boxplots of the 18 area-based descriptors and 8 contacts-based descriptors in the structures of antibody–antigen complexes. A. Eight types of surface area. **B.** Percentages of nonpolar and charged surface residues. **C.** Ten types of interface area. **D.** Number of the 6 types of interface residue-residue contacts.

neural network and random forest models, respectively.

For the constructed nonlinear model, no specific training set existed. When generating *a* and *b* in equation (3) (shown in Table 1), the training set was *Set2* with 33 reliable binding affinities from the expanded benchmark [20].

2.2.4. Methods for generating explicit linear, nonlinear formations or training neural network, random forest models

The linear formations (equations (1), (3) and (4) are shown in Table 1) are generated using the least square method [35]. The explicit formation in the nonlinear model was constructed based on the assumption, *i.e.*, $\log(K) \propto -x_1^{k1}\, x_2^{k2} \ldots x_n^{kn}$. The possible power exponents of the variable were taken from −1, −0.5, 0.5, 1.

The neural network models were trained using back propagation neural network [36,37]. In the architecture in the neural network, one or two hidden layers was (were) used with number of nodes from 1 to 20. Because of the stochasticity of initial weight assignment of neural network model, 100 models were trained when the number of hidden layers and nodes were fixed.

In the random forest algorithm, the different random subsets of regression trees were combined to make each decision [38,39]. Here, an enhanced random forest algorithm, *i.e.*, bootstrap-aggregated (bagged) decision trees, was employed to alleviate the overfitting effects. In the architecture in the random forest, the number of trees is chosen from 50, 100, …, 500, and the minimum number of observations per tree leaf is chosen from 5, 10, 15, 20. Because of the stochasticity of adopted decision trees in the random forest model, 100 models were trained when the number of trees and leaves was fixed.

2.3. Metrics for evaluating the performances of different models

The predictive powers of different models were evaluated using two metrics: Pearson's correlation coefficient (*R*) and root mean square error (RMSE) between the experimental and predicted binding affinity values. They were calculated using the following equations:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{6}$$

Here, n is the number of complexes with known binding affinities in the set; x_i and y_i are the experimental and predicted values of the ith protein–protein binding affinity, respectively; $\bar{x}$ and $\bar{y}$ are the averages of the experimental and predicted binding affinities in the set, respectively.

3. Results

Fig. 1 is the flowchart of the antibody–antigen binding affinity prediction, which involves three steps: ***1***) filtering of data from SAbDab [32] and the expanded benchmark [20]; ***2***) calculation of the area-based and contacts-based descriptors; ***3***) training or construction of new models specific for antibody–antigen binding affinity prediction. Based on the data sets (after filtering) used in this study, some of the previously-established area-based models [29] and contacts-based models [30,31] for general protein–protein binding affinity prediction were also studied for comparison of their performances in antibody–antigen binding affinity prediction. For convenience of description, the sets composed of the antibody–antigen binding affinity data from SAbDab [32] and the expanded benchmark [20] are named *Set1* (containing 262 complexes) and *Set2* (containing 33 complexes), respectively.

3.1. Eighteen area-based descriptors and 8 contacts-based descriptors in the structures of antibody–antigen complexes

To compare different descriptors, the boxplots (Fig. 2) are plotted based on the 18 area-based descriptors and 8 contacts-based descriptors in the structures of all 295 antibody–antigen complexes from the two sets after filtering (**Supplementary File 1**). According to the positions of the descriptors in the structure of an antibody–antigen complex, the descriptors were further categorized into surface descriptors and interface descriptors.

Surface descriptors. Surface contributions to binding affinity were estimated using different types of surface area between the receptor and ligand in a complex along with the surrounding water molecules [29]. As shown in Fig. 2A, the magnitudes of the medians of different types of surface area were 1000 Å^2. The size relationship of the medians of different types of surface area was polar *AA*s > nonpolar *AA*s > basic *AA*s > acidic *AA*s for the receptor or ligand. The medians of the 4 types of surface area in the receptor were larger than the corresponding medians in the ligand. The ranges of the 4 types of surface area in the receptor were smaller than the ones in the ligand, which reflected the structural diversity of the protein antigen ligands. Contacts-based surface descriptors employed by PRODIGY were the percentages of surface residues rather than the surface contacts with water directly, and the receptor and ligand were not considered separately [30,31]. As shown in Fig. 2B, most of the nonpolar surface residues ranged from 31% to 42%, and most of the charged surface residues ranged from 18% to 27%. The percentage of nonpolar surface residues was higher than the percentage of charged surface residues, which was related to the numbers of nonpolar and charged amino acid residues.

Interface descriptors. Interface contribution to binding affinity was evaluated using the interface area between the receptor and ligand [29]. As shown in Fig. 2**C**, the medians of different types of interface area were 0–100 Å^2. The size relationship of the medians of different types of interface area was nonpolar *AA*s ~ polar *AA*s > polar *AA*s ~ polar *AA*s ≈ basic *AA*s ~ polar *AA*s > polar *AA*s ~ acidic *AA*s > nonpolar *AA*s ~ nonpolar *AA*s > basic *AA*s ~ nonpolar *AA*s > basic *AA*s ~ acidic *AA*s > other types of interface area. There were more types of interface area associated with polar *AA*s than with other *AA*s, which reflected the importance of the areas of polar *AA*s on the interface.

PRODIGY, a contacts-based model, estimates the interface contribution to binding affinity using interface contacts between the receptor and ligand [30,31]. As shown in Fig. 2**D**, the numbers of the 6 types of interface residue-residue contacts ranged from 0 to 45. The median of interface contact numbers was nonpolar residues ~ nonpolar residues > nonpolar residues ~ nonpolar residues = charged residues ~ nonpolar residues > charged residues ~ polar residues > charged residues ~ charged residues > polar residues ~ polar residues, which may imply that the order of importance of the 3 types of residues on the interface is nonpolar residues > charged residues > polar residues.

It should be noted that the area-based descriptors were based on the 4-group division of the 20 amino acid residues, while the contacts-based descriptors were based on the 3-group division of the 20 amino acid residues. The conclusions derived from area-based and contacts-based descriptors are not always in agreement with each other. The predictive powers of the models using area-based and contacts-based descriptors on binding affinity were compared below to assess the relative effectiveness of these two classes of descriptors.

3.2. Performances of the single descriptors for antibody–antigen binding affinity prediction

The effectiveness of the single area-based or contacts-based descriptors on antibody–antigen binding affinity prediction was estimated using the Pearson's correlation coefficient (*R*) in *Set1* and *Set2*. The area-based descriptors contained the following 6 classes: 4 types of receptor surface area, 4 types of ligand surface area, 10 types of interface area,

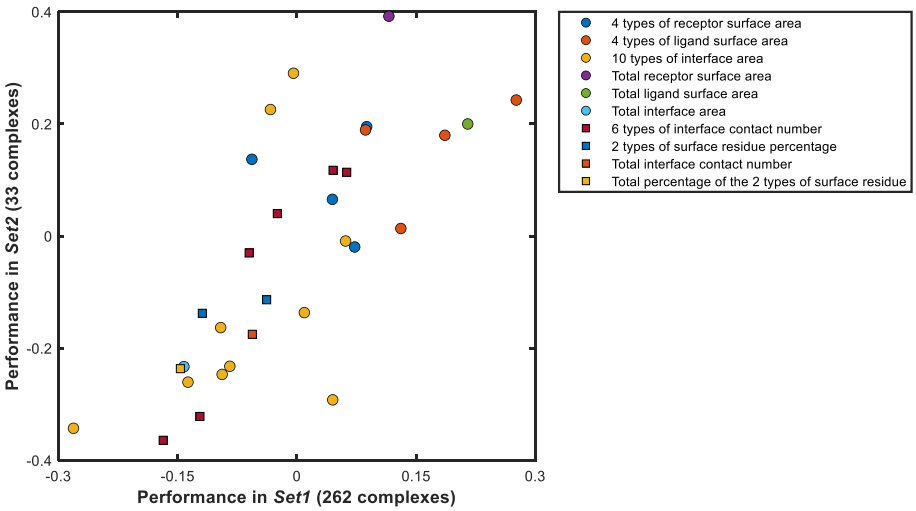

Fig. 3. Performance of the single descriptors for antibody–antigen binding affinity prediction. The area-based and contacts-based single descriptors are represented as dots and squares, respectively. The performance is estimated using the Pearson's correlation coefficient (*R*).

total receptor surface area in the complex, total ligand surface area in the complex, and total interface area in the complex. The contacts-based descriptors contained the following 4 classes: 6 types of the interface contact number, 2 types of the surface residue percentage, total interface contact number, and total percentage of the 2 types (nonpolar and charged) of surface residues. The *R* values of different descriptors are listed in **Supplementary File 1**, and the performances of different single descriptors in *Set1* and *Set2* are shown in Fig. 3.

Among the 21 area-based descriptors, there were 6 descriptors with absolute *R* values greater than 0.1 and 0.2 in *Set1* and *Set2*, respectively. Based on the absolute *R* values for the combined set (295 complexes) composed of *Set1* (262 complexes) and *Set2* (33 complexes), the order of these area-based descriptors is: interface area of basic *AA*s ~ nonpolar *AA*s (−0.29) > polar *AA*s of *LSA* (0.27) > total ligand surface area (0.21) > interface area of basic *AA*s ~ acidic *AA*s (−0.15) = total interface area (−0.15) > total receptor surface area (0.14). The corresponding *R* values for *Set1* (262 complexes) and *Set2* (33 complexes) are −0.28 and −0.34 for interface area of basic *AA*s ~ nonpolar *AA*s; 0.28 and 0.24 for polar *AA*s of LSA; 0.22 and 0.20 for total ligand surface area; −0.14 and −0.26 for interface area of basic *AA*s ~ acidic *AA*s; −0.14 and −0.23 for total interface area; 0.12 and 0.39 for total receptor surface area. Although the performance of total receptor surface area in *Set1* (0.12) was not as good as those of the other 5 descriptors, its performance in *Set2* (0.39) was the best among all the descriptors. Different performances in these two sets reflected the influence of quantity (large or small) and quality (high or low) of the data sets.

Based on the absolute *R* values, the component (basic *AA*s ~ nonpolar *AA*s (−0.29)) of the interface area was superior to the total interface area (−0.15), and the component (polar *AA*s of *LSA* (0.27)) of the ligand surface area was superior to the total ligand surface area in the complex (0.21). This implies that certain interface (or surface) areas are more important than other interface (or surface) regions for the association and disassociation of the antibody–protein antigen complexes, and thus exhibit higher correlations with the experimental binding affinities. The less important interface (or surface) regions can either increase or decrease the affinities in different environments, and accordingly they have lower absolute *R* values.

Among the 10 contacts-based descriptors, only 3 descriptors had absolute *R* values higher than 0.1 and 0.2 in *Set1* and *Set2*, respectively. Based on the absolute *R* values of the combined set (295 complexes) composed of *Set1* and *Set2*, the order of the 3 contacts-based descriptors was: the number of interface contacts of charged residues ~ nonpolar

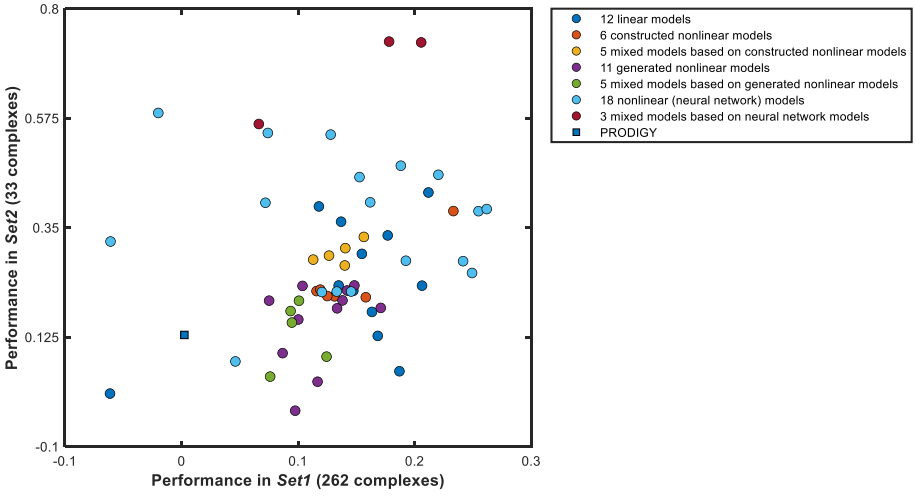

Fig. 4. Performances of the previous 60 representative area-based models [29] and PRODIGY [30] on antibody–antigen binding affinity prediction. The 60 area-based models and PRODIGY (contacts-based model) [30] are represented as dots and square, respectively. The performance is estimated using the Pearson's correlation coefficient (*R*).

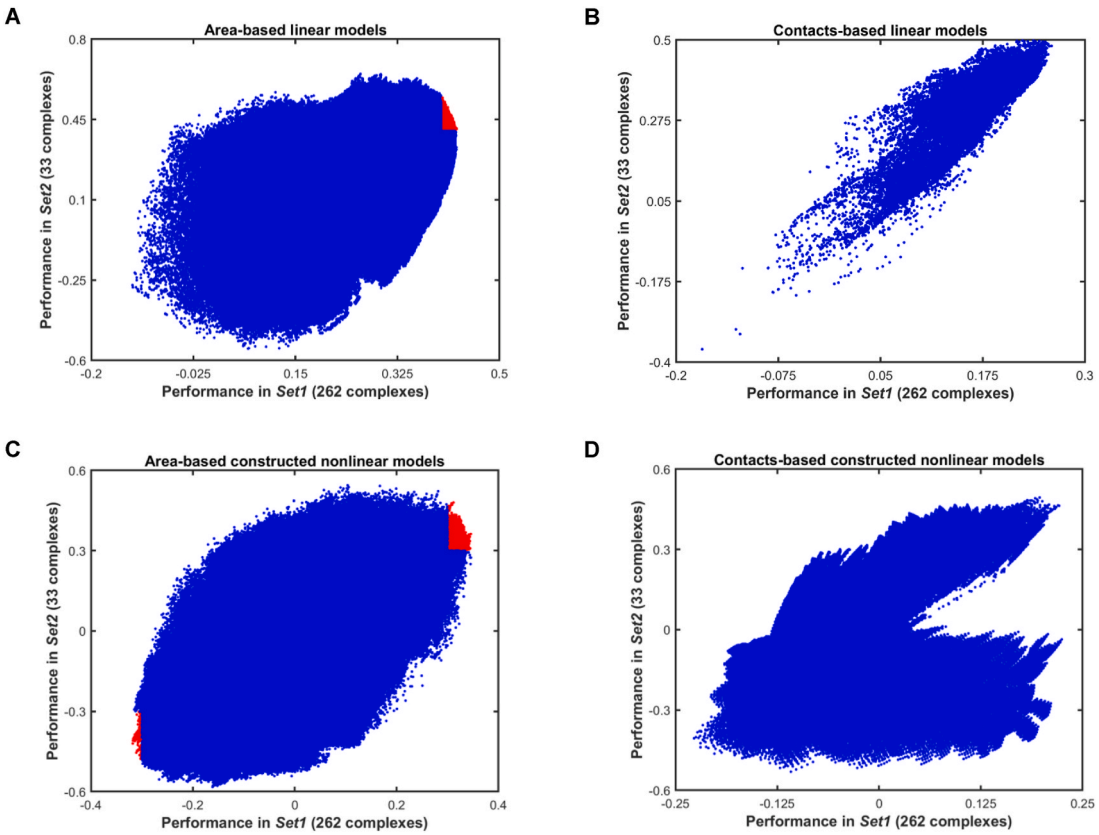

Fig. 5. Performances of all the area-based and contacts-based linear and constructed nonlinear models specific for antibody–antigen binding affinity prediction. A. Area-based linear models. **B.** Contacts-based linear models. **C.** Area-based contructed nonlinear models. **D.** Contacts-based contructed nonlinear models. The performance is estimated using the Pearson's correlation coefficient (*R*). One dot represents one model. A model is regarded as good when the absolute values of *R* in *Set1* (262 complexes) and *Set2* (33 complexes) are both higher than a given cutoff value. The good models are colored red, and the other models are colored blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

residues (−0.19) > the percentage of the nonpolar and charged surface residues (−0.16) > the number of the interface contacts of charged residues ~ charged residues (−0.15). The corresponding *R* values for *Set1* (262 complexes) and *Set2* (33 complexes) were −0.17 and −0.36 for the number of interface contacts of charged residues ~ nonpolar residues; −0.15 and −0.24 for the percentage of nonpolar and charged surface residues; −0.12 and −0.32 for the number of the interface contacts of charged residues ~ charged residues. The interface contacts of charged residues ~ nonpolar residues and charged residues ~ charged residues contributed more significantly to the binding affinity than other types of interface contacts, while the interface contacts of charged residues ~ charged residues were not in the same ranks of those contacts with the largest numbers. This demonstrates the relative importance of different amino acid types in binding affinity prediction.

In the case of single descriptors, the performance of the best area-based single descriptor in *Set1* and *Set2* (interface area of basic *AAs* ~ nonpolar *AAs* (0.29)) was slightly better than or comparable to the performance of the best contacts-based single descriptor (number of charged ~ nonpolar interface contacts (−0.19)). The predictive powers of the combinations of these two classes of descriptors on antibody–antigen binding affinity prediction are discussed below.

3.3. Performances of the existing area-based models and PRODIGY for protein–protein and antibody–protein antigen binding affinity prediction

In our previous work, 60 representative area-based models were developed for protein–protein binding affinity prediction [29]. Additionally, PRODIGY, a contacts-based simple model, is very effective for protein–protein binding affinity prediction [30,31]. The predictive powers of these two models on antibody–antigen binding affinity prediction were compared based on their performances in *Set1* and *Set2*. The predicted binding affinities and the *R* values of the 61 models (including PRODIGY) in the two sets are listed in **Supplementary File 2**.

Among the previous 60 area-based models, there are 12 linear models, 6 constructed nonlinear models, 5 mixed models based on constructed nonlinear models, 11 generated nonlinear models, 5 mixed models based on generated nonlinear models, 18 nonlinear (neural network) models, and 3 mixed models based on neural network models [29]. The performances of these models and PRODIGY are shown in Fig. 4. There were 6 models with *R* values higher than 0.2 and 0.3 in *Set1* and *Set2*, respectively. These models were linear model 3 (0.21, 0.42), constructed nonlinear model 1 (0.23, 0.38), nonlinear model (neural network) 9 (0.25, 0.38), nonlinear model (neural network) 17 (0.26, 0.39), nonlinear model (neural network) 18 (0.22, 0.46) and mixed model (based on neural network models) 1 (0.21, 0.73). The formations of linear model 3 and constructed nonlinear model 1 were $\log(K) = -0.000344614*A_3 + 0.000498127*A_4 + 0.000496231*A_7 - 0.013381318*A_{10} - 0.006715831*A_{11} - 0.004391993*A_{14} - 0.007269689*A_{15} - 5.276589910$ and $\log(K) = -0.304285849*A_{21}/\mathrm{sqrt}(A_{20}) - 4.313861057$, respectively. There were no explicit formations for the other 4 models. The *R* values of all the 6 models were smaller than 0.3 in *Set1* and 0.5 in *Set2* except the mixed model 1 (based on neural network models) (0.21, 0.73). The *R* values of PRODIGY in *Set1* and *Set2* were 0.00 and 0.13, respectively. It is apparent that the earlier models for protein–protein binding affinity prediction are unable to provide comparable predictive powers for antibody–protein antigen binding affinities, and new effective area-based and contacts-based models should be constructed and

Table 2
Numbers and percentages of different area-based and contacts-based models with relatively good performances.

Model type	Cutoff of good models[a]	Area-based models			Contacts-based models		
		Total number	Number of good models	Percentage of good models (%)	Total number	Number of good models	Percentage of good models (%)
Linear model	0.40	26214300	90425	0.34	25500	0	0
Constructed nonlinear model	0.30	22458100	2861	0.01	9765624	0	0
Neural network (nonlinear) model	0.40	1260000	5741	0.46	1260000	346	0.03
Mixed model (based on neural network models)	0.70	1638300	568602	34.71	1638300	8427	0.51
Random forest (nonlinear) model	0.50	120000	12537	10.30	120000	503	0.42
Mixed model (based on random forest models)	0.70	1638300	7142	0.44	1638300	0	0

[a] Good models: the good models are those with the absolute values of Pearson's correlation coefficients *R* higher than a given cutoff in both *Set1* and *set2*.

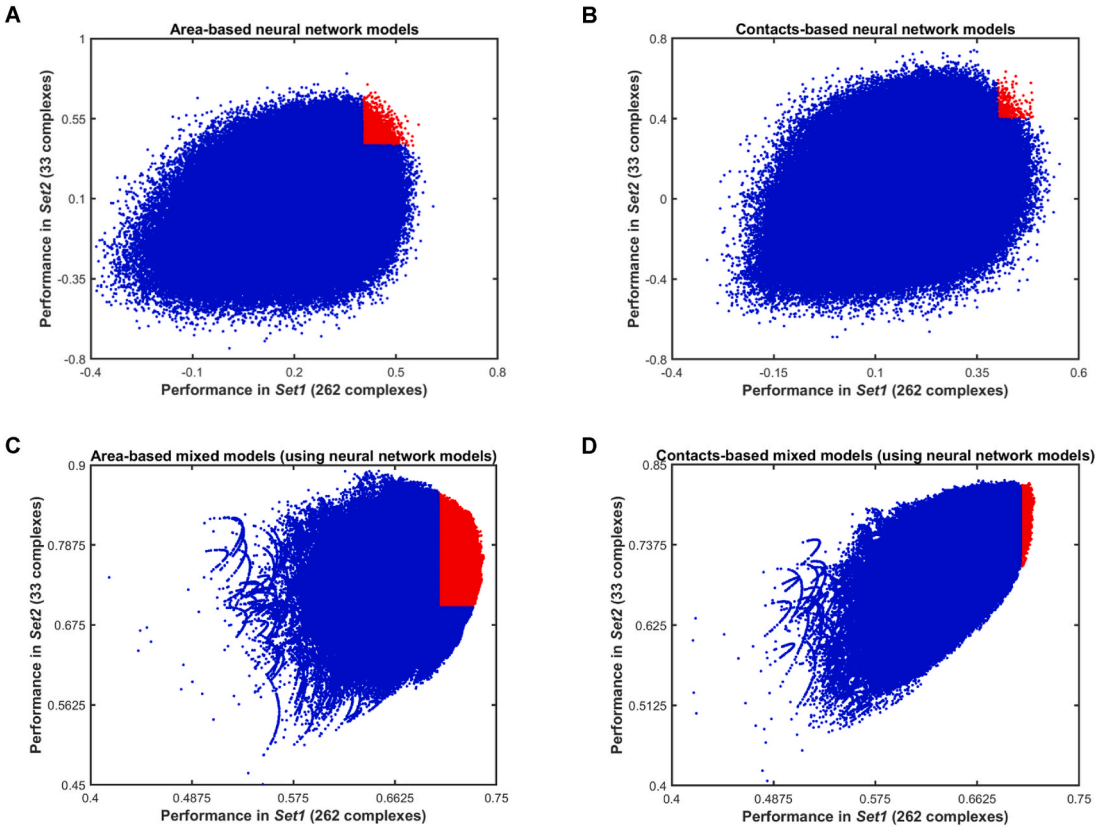

Fig. 6. Performances of all the area-based and contacts-based neural network and mixed models specific for antibody–antigen binding affinity prediction. A. Area-based neural network models. **B.** Contacts-based neural network models. **C.** Area-based mixed models (using neural network models). **D.** Contacts-based mixed models (using neural network models). The performance is estimated using the Pearson's correlation coefficient (*R*). One dot represents one model. A model is regarded as good when the absolute values of *R* in *set1* (262 complexes) and *set2* (33 complexes) are both higher than a given cutoff value. The good models are colored red, and the other models are colored blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

trained for this purpose.

3.4. Performances of new area-based and contacts-based models specific for antibody–antigen binding affinity prediction

Numerous models specific for antibody–antigen binding affinity prediction were constructed or trained in this study employing the area-based and contacts-based descriptors. The *R* values of all the models in *Set1* and *Set2* are shown in Fig. 5, **6** and **7**, and the overall results of the new models are shown in Table 2. In this study, a model is regarded as good if the absolute *R* values of the model are higher than a given cutoff in *Set1* and *Set2*, simultaneously. The cutoffs are 0.40 for linear models, 0.30 for constructed nonlinear models, 0.40 for neural network models, 0.70 for mixed models (based on neural network models), 0.50 for random forest models, and 0.70 for mixed models (based on random forest models). The criteria for setting the cutoffs are: ***1***) they are sufficiently high for eliminating a large number of models with bad performances; ***2***) they are also small enough for retaining a reasonably small number of models with good performances; ***3***) they are integral multiples of 0.10 for ease of calculation. The percentages (%) of good models were 0.34 and 0 for area-based and contacts-based linear models (Fig. 5**A** and **B**); 0.01 and 0 for area-based and contacts-based constructed nonlinear models (Fig. 5**C** and **D**); 0.46 and 0.03 for area-based and contacts-based neural network models (Fig. 6A and **B**); and 34.71

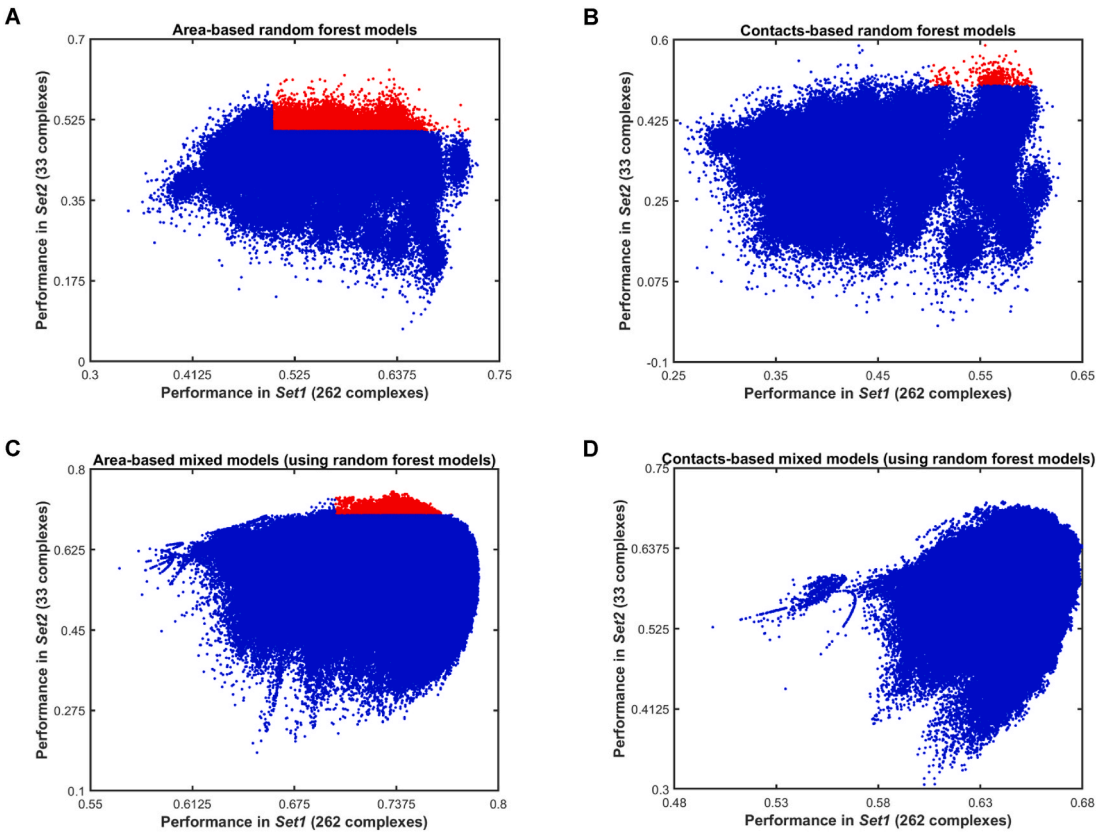

Fig. 7. Performances of all the area-based and contacts-based random forest and mixed models specific for antibody–antigen binding affinity prediction. **A.** Area-based random forest models. **B.** Contacts-based random forest models. **C.** Area-based mixed models (using random forest models). **D.** Contacts-based mixed models (using random forest models). The performance is estimated using the Pearson's correlation coefficient (*R*). One dot represents one model. A model is regarded as good when the absolute values of *R* in *set1* (262 complexes) and *set2* (33 complexes) are both higher than a given cutoff value. The good models are colored red, and the other models are colored blue. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Formations of the 2 representative area-based liner models and 3 representaive constructed nonlinear models.

Model	Formation
Linear model 1	$\log(K) = 0.000162011*A_1 + 0.000256584*A_3 - 0.000144866*A_5 + 0.000245891*A_7 - 0.009747517*A_9 - 0.001788572*A_{10} + 0.006572632*A_{12} - 0.006780979*A_{13} - 0.000407818*A_{14} - 0.003204142*A_{15} + 0.002389057*A_{17} - 0.001399766*A_{18} - 10.786964184$
Linear model 2	$\log(K) = 0.000310207*A_3 - 0.000269738*A_5 + 0.000342587*A_7 - 0.003031625*A_9 - 0.002816392*A_{10} - 0.000401154*A_{11} - 0.005831421*A_{13} - 0.006310465*A_{15} + 0.002184578*A_{17} - 0.002782307*A_{18} - 10.593420773$
Constructed nonlinear model 1	$\log(K) = 0.000000002*A_3*A_7*A_{19}/A_{21} - 9.638763114$
Constructed nonlinear model 2	$\log(K) = 0.000013728*A_3*A_7*\mathrm{sqrt}(A_{20}/A_5)/A_{21} - 9.557850214$
Constructed nonlinear model 3	$\log(K) = 0.000335444*A_3*A_7*\mathrm{sqrt}(A_{16}/A_{20})/A_{21} - 10.123693002$

Table 4
Performances of the 2 representative area-based liner models and 3 representaive constructed area-based nonlinear models.

Model	Pearson's correlation coefficient (*R*)			
	Training set (131 complexes)	Validation set (131 complexes)	*Set1* (262 complexes)	*Set2* (33 complexes)
Linear model 1	0.43	0.40	0.41	0.50
Linear model 2	0.40	0.44	0.42	0.46
Constructed nonlinear model 1	——	——	0.33	0.39
Constructed nonlinear model 2	——	——	0.35	0.36
Constructed nonlinear model 3	——	——	0.31	0.48

and 0.51 for area-based and contacts-based mixed models (using neural network models) (Fig. 6**C** and **D**), 10.30 and 0.42 for area-based and contacts-based random forest models (Fig. 7A and **B**), and 0.44 and 0 for area-based and contacts-based mixed models (using random forest models) (Fig. 7**C** and **D**) (Table 2). The percentages of good models with the same cutoff were employed to compare the overall performances of the contacts-based over area-based models. Overall, the effectiveness of contacts-based descriptors adopted by PRODIGY [30,31] was not as good as that of area-based descriptors. It should be noted that this is just a relative comparison based on the percentages of the good models with the same types of models and the same cutoffs.

Among the new models specific for antibody–antigen binding affinity prediction, 37 area-based models and 29 contacts-based models were selected to represent the good models. The 37 area-based models were composed of 2 linear models, 3 constructed nonlinear models, 14 neural network (nonlinear) models, 3 mixed models (based on neural network models), 14 random forest models, and 1 mixed model (based on random forest models). The 29 contacts-based models incorporated 14 neural network (nonlinear) models, 1 mixed model (based on neural

Table 5
Performances of the 14 representative area-based neural network (nonliner) models and 3 representaive area-based mixed models.

Model	Pearson's correlation coefficient (*R*)			
	Training set (131 complexes)	Validation set (131 complexes)	*Set1* (262 complexes)	*Set2* (33 complexes)
Neural network (nonliner) model 1	0.72	0.24	0.48	0.58
Neural network (nonliner) model 2	0.78	0.34	0.57	0.52
Neural network (nonliner) model 3	0.73	0.22	0.45	0.65
Neural network (nonliner) model 4	0.72	0.34	0.53	0.56
Neural network (nonliner) model 5	0.66	0.28	0.45	0.67
Neural network (nonliner) model 6	0.76	0.29	0.50	0.58
Neural network (nonliner) model 7	0.67	0.20	0.42	0.74
Neural network (nonliner) model 8	0.78	0.28	0.51	0.54
Neural network (nonliner) model 9	0.73	0.27	0.44	0.64
Neural network (nonliner) model 10	0.74	0.39	0.54	0.47
Neural network (nonliner) model 11	0.70	0.24	0.44	0.67
Neural network (nonliner) model 12	0.80	0.33	0.55	0.45
Neural network (nonliner) model 13	0.72	0.28	0.48	0.62
Neural network (nonliner) model 14	0.70	0.32	0.49	0.60
Mixed model 1	0.70	0.73	0.71	0.85
Mixed model 2	0.74	0.71	0.72	0.84
Mixed model 3	0.70	0.77	0.73	0.79

Table 6
Performances of the 14 representative contacts-based neural network (nonliner) models and 1 representaive contacts-based mixed model.

Model	Pearson's correlation coefficient (*R*)			
	Training set (131 complexes)	Validation set (131 complexes)	*Set1* (262 complexes)	*Set2* (33 complexes)
Neural network (nonliner) model 1	0.66	0.34	0.48	0.48
Neural network (nonliner) model 2	0.66	0.28	0.48	0.53
Neural network (nonliner) model 3	0.70	0.17	0.42	0.63
Neural network (nonliner) model 4	0.62	0.29	0.44	0.61
Neural network (nonliner) model 5	0.75	0.24	0.48	0.58
Neural network (nonliner) model 6	0.65	0.33	0.47	0.48
Neural network (nonliner) model 7	0.71	0.24	0.42	0.53
Neural network (nonliner) model 8	0.66	0.24	0.42	0.60
Neural network (nonliner) model 9	0.73	0.27	0.48	0.46
Neural network (nonliner) model 10	0.71	0.30	0.46	0.52
Neural network (nonliner) model 11	0.75	0.27	0.48	0.41
Neural network (nonliner) model 12	0.71	0.32	0.48	0.42
Neural network (nonliner) model 13	0.64	0.34	0.49	0.49
Neural network (nonliner) model 14	0.72	0.22	0.42	0.50
Mixed model 1	0.73	0.70	0.71	0.79

network models), and 14 random forest models. Because all the *R* values for contacts-based mixed models (based on random forest models) were smaller than 0.70 in *Set1* (Fig. 7**D**), no representative models were selected from these models. The predicted binding affinities of these representative models are summarized in **Supplementary File 3**. The formations of the 2 area-based linear models and 3 area-based constructed nonlinear models are shown in Table 3, and the performances of the representative models are shown in Table 4, **5**, **6**, **7** and **8**.

As shown Tables 3 and 4, there are 12 descriptors in area-based linear model 1 with *R* value equal to 0.41 in *Set1* and 0.50 in *Set2*, and 11 descriptors in area-based linear model 2 with *R* value equal to 0.42 in *Set1* and 0.46 in *Set2*. The products of the values of the descriptors and the corresponding coefficients reflect the varying degrees of contribution of different areas to the antibody–antigen binding affinity. The constructed nonlinear terms in the 3 representative models were $A_3{*}A_7{*}A_{19}/A_{21}$, $A_3{*}A_7{*}\mathrm{sqrt}(A_{20}/A_5)/A_{21}$, and $A_3{*}A_7{*}\mathrm{sqrt}(A_{16}/A_{20})/A_{21}$, respectively. The corresponding *R* values in *Set1* and *Set2* were 0.33 and 0.39 for constructed nonlinear model 1; 0.35 and 0.36 for

Table 7
Performances of the 14 representative area-based random forest (nonliner) models and 1 representaive area-based mixed model.

Model	Pearson's correlation coefficient (*R*)			
	Training set (131 complexes)	Validation set (131 complexes)	*Set1* (262 complexes)	*Set2* (33 complexes)
Random forest (nonliner) model 1	0.88	0.18	0.61	0.61
Random forest (nonliner) model 2	0.88	0.18	0.61	0.61
Random forest (nonliner) model 3	0.89	0.25	0.63	0.60
Random forest (nonliner) model 4	0.89	0.22	0.63	0.63
Random forest (nonliner) model 5	0.90	0.24	0.65	0.58
Random forest (nonliner) model 6	0.90	0.23	0.64	0.56
Random forest (nonliner) model 7	0.80	0.24	0.57	0.58
Random forest (nonliner) model 8	0.89	0.30	0.64	0.59
Random forest (nonliner) model 9	0.90	0.34	0.66	0.59
Random forest (nonliner) model 10	0.90	0.29	0.65	0.57
Random forest (nonliner) model 11	0.90	0.26	0.63	0.61
Random forest (nonliner) model 12	0.80	0.38	0.60	0.59
Random forest (nonliner) model 13	0.77	0.36	0.58	0.62
Random forest (nonliner) model 14	0.80	0.35	0.59	0.59
Mixed model 1	0.77	0.71	0.74	0.75

Table 8
Performances of the 14 representative contacts-based random forest (nonliner) models.

Model	Pearson's correlation coefficient (*R*)			
	Training set (131 complexes)	Validation set (131 complexes)	*Set1* (262 complexes)	*Set2* (33 complexes)
Random forest (nonliner) model 1	0.86	0.19	0.53	0.55
Random forest (nonliner) model 2	0.85	0.21	0.55	0.56
Random forest (nonliner) model 3	0.85	0.15	0.52	0.54
Random forest (nonliner) model 4	0.88	0.19	0.55	0.59
Random forest (nonliner) model 5	0.86	0.22	0.57	0.56
Random forest (nonliner) model 6	0.84	0.18	0.54	0.56
Random forest (nonliner) model 7	0.81	0.22	0.52	0.54
Random forest (nonliner) model 8	0.85	0.19	0.58	0.55
Random forest (nonliner) model 9	0.86	0.28	0.59	0.55
Random forest (nonliner) model 10	0.85	0.30	0.58	0.57
Random forest (nonliner) model 11	0.86	0.31	0.60	0.54
Random forest (nonliner) model 12	0.86	0.31	0.60	0.53
Random forest (nonliner) model 13	0.86	0.30	0.58	0.58
Random forest (nonliner) model 14	0.86	0.31	0.60	0.52

constructed nonlinear model 2; 0.31 and 0.48 for constructed nonlinear model 3. The common descriptors in the 3 constructed nonlinear terms were *RSA* of polar *AA*s (A_3), *LSA* of polar *AA*s (A_7), and total interface area (A_{21}). The surface and interface areas were adopted simultaneously in each of the 5 area-based linear or nonlinear models with explicit formations, which demonstrated that the surface and interface contributions to antibody–antigen binding affinity should be jointly considered to achieve more predictive power.

Among the 14 area-based neural network models, 4 models had *R* values equal to or higher than 0.5 in both *Set1* and *Set2* (neural network models 2, 4, 6 and 8 in Table 5). The mixed models were the linear combinations of the representative neural network models. The 3 area-based mixed models were superior to the single area-based neural network model (Table 5). One neural network model only grasped one aspect of the antibody–antigen binding affinity, which is analogous to the role of van der Waals, electrostatic or hydrogen bonding energy in the total nonbonding energy. Overall, the performances of 14 representative contacts-based neural network models (Table 6) were not as good as the ones of the representative area-based neural network models (Table 5). However, the performance of the contacts-based mixed model (*R* values in *Set1* and *Set2* are 0.71 and 0.79, respectively) (Table 6) was comparable to the performances of the area-based mixed models (*R* values in *Set1* and *Set2* are 0.71 and 0.85, respectively, for area-based mixed model 1; 0.72 and 0.84 for area-based mixed model 2; 0.73 and 0.79 for area-based mixed model 3) (Table 5).

The performances of the representative random forest and mixed models are shown in Tables 7 and 8. The *R* values of the 14 area-based random forest models in *Set1* or *Set2* were from 0.56 to 0.66 (Table 7), and those of the contacts-based random forest models were from 0.52 to 0.60 (Table 8). The representative area-based mixed model (based on random forest models) (Table 7, *R* values in *Set1* and *Set2* are 0.74 and 0.75, respectively) was superior to the single random forest models. The performances of the representative random forest models were generally better than those of the representative neural network models for both area-based and contacts-based models (Table 5 vs Table 7, Table 6 *vs* Table 8). Nevertheless, the predictive powers of the representative mixed models based on neural network models were comparable to the performances of the representative mixed models based on random forest models.

It is of note that the performances of single neural network or random forest models in the stochastic validation sets are not as good as those in the stochastic training sets (Table 5, **6**, **7** and **8**). Among all the single neural network or random forest models, there exists only one

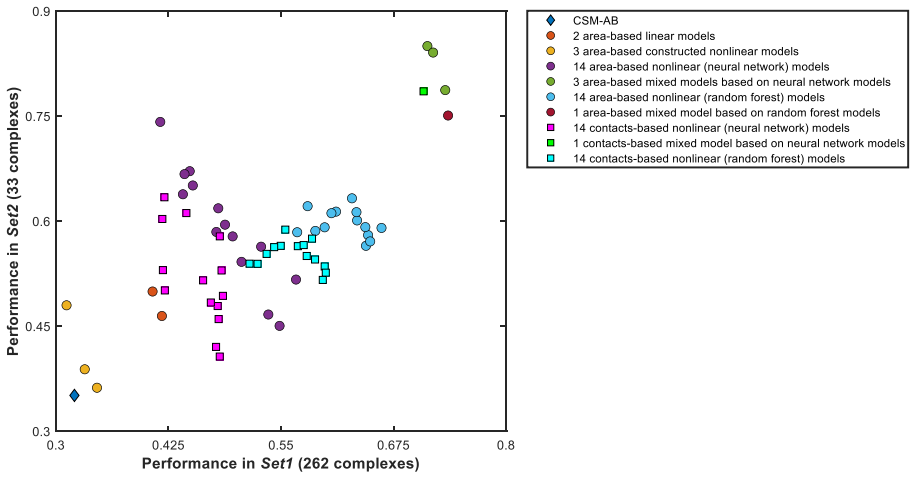

Fig. 8. Performances of different representative area-based, contacts-based models and the CSM-AB model [27]. The CSM-AB model [27], the 37 area-based models and the 29 contacts-based models are represented as diamond, dots and squares, respectively. The performance is estimated using the Pearson's correlation coefficient (*R*).

model for which the *R* values in the stochastic training set, stochastic validation set, *Set1* and *Set2* are consistently higher than 0.45. The model is an area-based random forest model with *R* values of 0.70 (stochastic training set), 0.46 (stochastic validation set), 0.58 (*Set1*) and 0.55 (*Set2*). In the random forest model, the number of trees and the minimum number of observations per tree leaf are 50 and 15, respectively. In order to generate models with the global minima of prediction errors in different sets, the same training sets, number of trees, minimum number of observations per tree leaf are adopted to train 10000 new random forest models. The performances of these models in the training and validation sets are shown in Supplementary Fig. 1A, and those in *Set1* and *Set2* are shown in Supplementary Fig. 1B. There were 3 models with *R* values higher than 0.45 in all the sets (Supplementary Table 1), which are not better than the initial random forest model, and are also not better than the best mixed models which had *R* values higher than 0.70 in all the sets (Table 5, **6**, and **7**).

3.5. Comparison of the performances of different models

The predictive powers of the CSM-AB model [27], the newly-developed 37 area-based models and the 29 contacts-based models were compared based on their performances in *Set1* and *Set2*. As shown in Fig. 8, the relative predictive powers of these models (according to their performances) were the CSM-AB model [27] < area--based constructed nonlinear models < area-based linear models < contacts-based nonlinear (neural network) models < area-based nonlinear (neural network) models ≤ contacts-based nonlinear (random forest) models < area-based nonlinear (random forest) models < area-based and contacts-based mixed (based on neural network or random forest) models.

In addition, the performances of the CSM-AB model [27], the best contacts-based model (mixed model 1 in Table 6) and the area-based model (mixed model 3 in Table 5) are shown in Fig. 9. The *R* values in *Set1* and *Set2* were (0.32, 0.35) for the CSM-AB model [27], (0.71, 0.79) for the best contacts-based model, and (0.73, 0.79) for the best area-based model. This result shows that our models have better predictive ability than the CSM-AB model.

4. Discussion

In the present study, different models were constructed to predict antibody–protein antigen binding using area-based and contacts-based descriptors through constructing and training different predictive models. The best area-based and contacts-based models developed in this study for antibody–antigen binding affinity prediction are superior to the previous area-based and contacts-based models constructed for general protein–protein binding affinity prediction [29–31] and the graph-based CSM-AB model specifically designed for antibody–antigen binding affinity prediction [27]. These models may aid in the antibody design and shed lights on the mechanism of the binding interaction between an antibody and a protein antigen.

While the previous predictive models were trained based on protein–protein binding affinity data, the new models developed in this study were trained solely based on the antibody–protein antigen binding affinity data. The performances of the previous models and the new models on antibody–protein antigen binding affinity prediction are shown in Figs. 4 and 8, respectively. The Pearson's correlation coefficients (*R*) in *Set1* (262 complexes) and *Set2* (33 complexes) for previous models (except two area-based mixed models) are below 0.30 and 0.60, respectively (Fig. 4 and **Supplementary File 2**), with best *R* values in *Set1* and *Set2* being 0.26 and 0.73, respectively. In comparison, the *R* values in *Set1* (262 complexes) and *Set2* (33 complexes) for all new models developed in this study are higher than 0.30 (Fig. 8, **Supplementary File 3** and **Supplementary File 4**), with the best *R* values in *Set1* and *Set2* being 0.74 and 0.85, respectively. The improvements in performance clearly indicate the importance of including exclusively the antibody–protein antigen binding affinity datasets as the training models.

One of the key questions in antibody-protein antigen binding affinity prediction is to determine what are the most suitable descriptors from both theoretical and practical points of views. In this work, the contacts-based and area-based descriptors were employed to predict antibody–protein antigen binding affinity, in an attempt to better understand the geometry–affinity relationship for antibody–protein antigen interactions. While the performances of the simple linear or nonlinear models with explicit and explainable formations were not better than the existing models (which was expected), the formations of these models still aid in the better understanding of the antibody-protein antigen binding interactions, and this information may also be of some value in the design of effective models for antibody-protein antigen binding affinity prediction.

It is of note that several platforms or protocols have been proposed for antibody design in the past few decades [5,7,10,11,40–43]. The importance and effects of different types of amino acids on antibody–protein antigen binding affinity/energy have been analyzed, and the knowledge gained from these studies helped the design of more effective antibodies [44–52]. For instance, the role of tyrosine in antibody–antigen binding interaction was noted in several earlier studies [44–48,

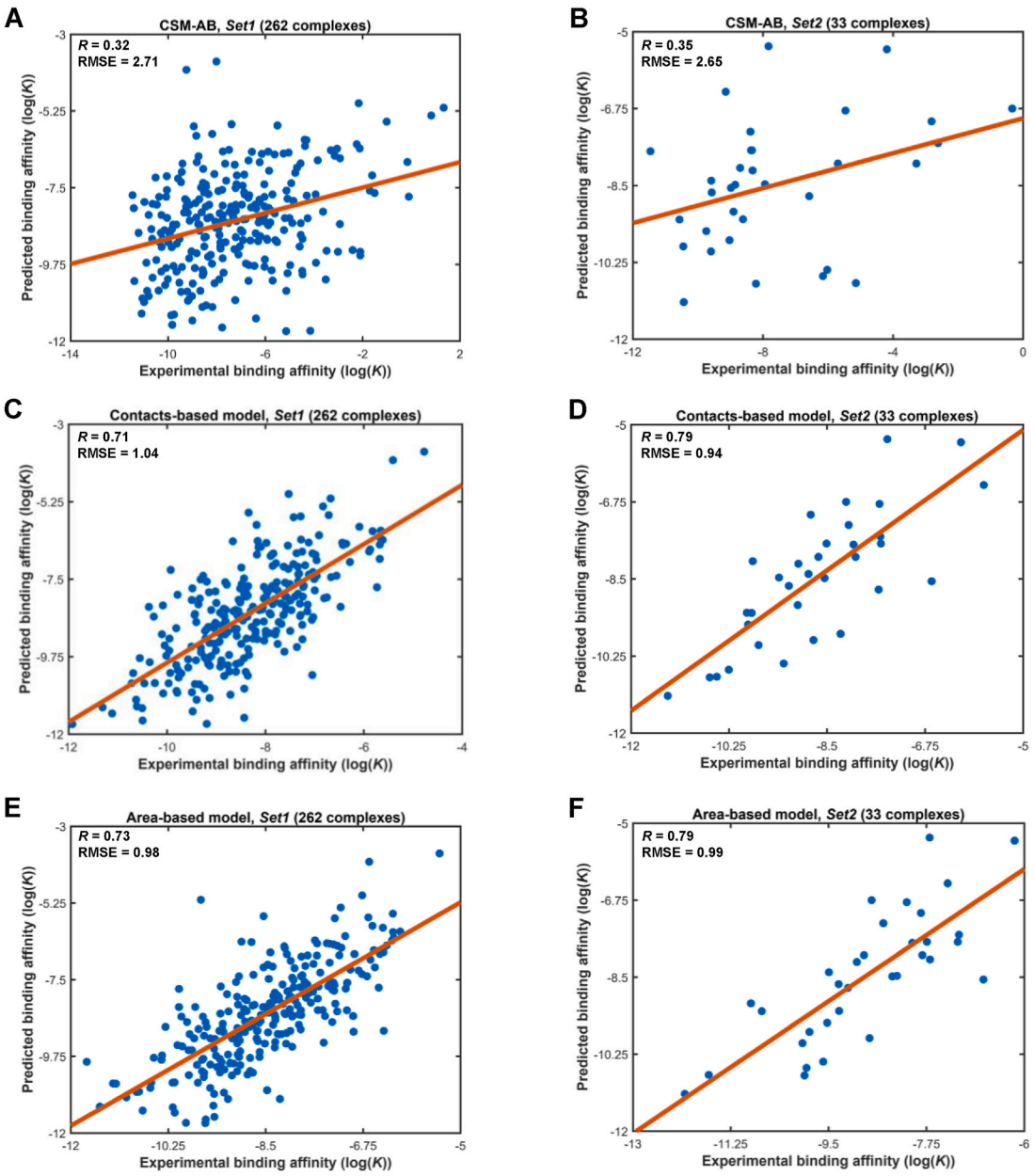

Fig. 9. Performances of the best area-based, contacts-based models and the CSM-AB model [27]. **A.** The CSM-AB model [27] in *Set1* (262 complexes). **B.** The CSM-AB model [27] in *Set2* (33 complexes). **C.** Contacts-based model in *Set1* (262 complexes). **D.** Contacts-based model in *Set2* (33 complexes). **E.** Area-based model in *Set1* (262 complexes). **F.** Area-based model in *Set2* (33 complexes). *R* and RMSE are the Pearson's correlation coefficient and root mean square error, repsetively.

50], which may be related to the biochemical characteristics of tyrosine (*e.g.*, aromaticity, polar but uncharged, and a bulky side chain). Another example is related to AA residues involved in electrostatic interactions. While some of the previous studies have shown that electrostatic interactions can improve antibody affinity and specificity [53–55], there were also studies reporting that arginine does not contribute to naïve antibody binding affinity, but it may enhance non-specific interactions [47,48]. Tiller et al. demonstrated that the effects of arginine mutations in CDRs are context-dependent [51]. Arginine mutations in hydrophobic portions of CDRs can lead to high-specificity antibody binding, where those in hydrophilic parts result in low-specificity binding [51]. Different influences of the same amino acid type may result from the structural environments with varying physicochemical properties which correspond to different configurations for antibody–antigen interactions. For example, diverse conformations of the CDR-H3 loop are very important for the specific binding of antibodies to different protein antigens [56–59].

According to the explicit formations of the five linear and nonlinear representative models in this work (Table 3), *RSA* of polar *AA*s (A_3) and *LSA* of polar *AA*s (A_7) are the common descriptors, which imply that the polar *AA*s (polar but uncharged amino acids: CYS, ASN, GLY, SER, GLN, TYR and THR) present on the antibody and protein antigen surface areas may play an important role in the antibody–protein antigen interactions. The positive coefficients of A_3 and A_7 signify that the larger surface area of the polar *AA*s is associated with the higher log(*K*) value (*i.e.*, a weaker binding affinity). The larger surface areas of polar *AA*s may increase the stabilities of antibody and protein antigen monomers in water solvents, thereby reducing the binding interaction between the antibody and the protein antigen. It is of note that these results are different from the results of the representative linear models for general protein–protein interactions described in our recent work in which the coefficients of A_3 and A_7 were negative and positive, respectively [29]. It

appears that the antibody–protein antigen interactions may have different characteristics from the general protein–protein interactions. In area-based protein–protein binding affinity prediction [29], RSA of polar AA (A_3) and the interface area of polar *AA*s ~ polar *AA*s (A_{11}) are the most important variables. RSA of polar AA (A_3) is a common important variable in both protein–protein binding affinity prediction and also antibody-protein antigen binding affinity prediction. It is possible that RSA of polar AA (A_3) is critical for sustaining the structural stability of the receptor (or antibody) and the complex in solvent. More reliable experimental data are required to confirm or correct the results reported in this work or in our recent work [29].

It is of note that the random forest is better than neural network in terms of the best performances of the trained models. The random forest is one of the ensemble learning methods, which integrates the predictive powers of different subsets of decision trees [38,39]. The mixed models in this work also belong to the ensemble learning methods. The potential drawbacks of the ensemble learning methods include increase of information redundancy and decrease of simplicity. Here it is worth mentioning that the variations in experimental data (such as binding affinity) represent an important intrinsic factor affecting the performance of the approximated models. As such, the predictive powers could not be improved infinitely using ensemble learning. With the increasing complexity of ensemble learning, the performance can be improved to some extent, but the general applicability of the model may decrease. As the effectiveness of the area-based and contacts-based descriptors for antibody–antigen binding affinity prediction is closely associated with the complexities of the models, the simpler models, in general, would be much more preferred if the same level of predictive power is achieved. Apparently, improvements are needed in future to develop simpler and yet more powerful models for the prediction of the antibody–antigen binding affinities.

5. Concluding remarks

Accurate antibody–antigen binding affinity prediction is of great practical value to the success of rational antibody design. In the present study, the contacts-based and area-based descriptors are adopted for prediction of the antibody-protein antigen binding affinities. Overall, the area-based descriptors show slightly better performance than contacts-based descriptors. Some representative models are selected from a large number of trained models, which incorporate 15 contacts-based and 22 area-based models specific for antibody–protein antigen binding affinity prediction. The performances of these representative models are better than those of the general models widely used for predicting the protein–protein interactions and a recently-developed method specifically designed for predicting antibody–protein antigen interactions. The results of this work may offer insights into the mechanisms of the antibody–antigen binding interactions and may also facilitate the antibody design in practice.

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.

Data availability

Data will be made available on request.

Acknowledgements

This work is supported by research grants from Shenzhen Key Laboratory of Steroid Drug Discovery and Development (No. ZDSYS20190902093417963) and Shenzhen Peacock Plan (No. KQTD2016053117035204). The computing exercises presented in this work were carried out on the High-Performance Computing Portal, which is under the administration of the Information Technology Services Office (ITSO) of The Chinese University of Hong Kong, Shenzhen.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmgm.2022.108364.

References

[1] L.B. Nicholson, The immune system, Essays Biochem. 60 (2016) 275–301.
[2] A.T. Huang, B. Garcia-Carreras, M.D.T. Hitchings, B. Yang, L.C. Katzelnick, S. M. Rattigan, B.A. Borgert, C.A. Moreno, B.D. Solomon, L. Trimmer-Smith, V. Etienne, I. Rodriguez-Barraquer, J. Lessler, H. Salje, D.S. Burke, A. Wesolowski, D.A.T. Cummings, A systematic review of antibody mediated immunity to coronaviruses: kinetics, correlates of protection, and association with severity, Nat. Commun. 11 (2020) 4704.
[3] R.A. Norman, F. Ambrosetti, A. Bonvin, L.J. Colwell, S. Kelm, S. Kumar, K. Krawczyk, Computational approaches to therapeutic antibody design: established methods and emerging trends, Briefings Bioinf. 21 (2020) 1549–1567.
[4] L. Gentiluomo, H.L. Svilenov, D. Augustijn, I. El Bialy, M.L. Greco, A. Kulakova, S. Indrakumar, S. Mahapatra, M.M. Morales, C. Pohl, A. Roche, A. Tosstorff, R. Curtis, J.P. Derrick, A. Norgaard, T.A. Khan, G.H.J. Peters, A. Pluen, A. Rinnan, W.W. Streicher, C.F. van der Walle, S. Uddin, G. Winter, D. Roessner, P. Harris, W. Friess, Advancing therapeutic protein Discovery and development through comprehensive computational and biophysical characterization, Mol. Pharm. 17 (2020) 426–440.
[5] D. Baran, M.G. Pszolla, G.D. Lapidoth, C. Norn, O. Dym, T. Unger, S. Albeck, M. D. Tyka, S.J. Fleishman, Principles for computational design of binding antibodies, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) 10900–10905.
[6] Z. Elgundi, M. Reslan, E. Cruz, V. Sifniotis, V. Kayser, The state-of-play and future of antibody therapeutics, Adv. Drug Deliv. Rev. 122 (2017) 2–19.
[7] R.J. Pantazes, C.D. Maranas, OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding, Protein Eng. Des. Sel. 23 (2010) 849–858.
[8] G.D. Lapidoth, D. Baran, G.M. Pszolla, C. Norn, A. Alon, M.D. Tyka, S.J. Fleishman, AbDesign: an algorithm for combinatorial backbone design guided by natural conformations and sequences, Proteins 83 (2015) 1385–1406.
[9] A. Sircar, E.T. Kim, J.J. Gray, RosettaAntibody: antibody variable region homology modeling server, Nucleic Acids Res. 37 (2009) W474–W479.
[10] R. Chowdhury, M.F. Allan, C.D. Maranas, OptMAVEn-2.0: de novo design of variable antibody regions against targeted antigen epitopes, Antibodies 7 (2018).
[11] T. Liang, H. Chen, J. Yuan, C. Jiang, Y. Hao, Y. Wang, Z. Feng, X.Q. Xie, IsAb: a computational protocol for antibody design, Briefings Bioinf. 22 (2021).
[12] L.A. Rabia, A.A. Desai, H.S. Jhajj, P.M. Tessier, Understanding and overcoming trade-offs between antibody affinity, specificity, stability and solubility, Biochem. Eng. J. 137 (2018) 365–374.
[13] O. Obrezanova, A. Arnell, R.G. de la Cuesta, M.E. Berthelot, T.R. Gallagher, J. Zurdo, Y. Stallwood, Aggregation risk prediction for antibodies and its application to biotherapeutic development, mAbs 7 (2015) 352–363.
[14] R. van der Kant, A.R. Karow-Zwick, J. Van Durme, M. Blech, R. Gallardo, D. Seeliger, K. Assfalg, P. Baatsen, G. Compernolle, A. Gils, J.M. Studts, P. Schulz, P. Garidel, J. Schymkowitz, F. Rousseau, Prediction and reduction of the aggregation of monoclonal antibodies, J. Mol. Biol. 429 (2017) 1244–1261.
[15] P. Sormanni, L. Amery, S. Ekizoglou, M. Vendruscolo, B. Popovic, Rapid and accurate in silico solubility screening of a monoclonal antibody library, Sci. Rep. 7 (2017) 8200.
[16] K.R. Abhinandan, A.C. Martin, Analyzing the "degree of humanness" of antibody sequences, J. Mol. Biol. 369 (2007) 852–862.
[17] C.H. Norn, G. Lapidoth, S.J. Fleishman, High-accuracy modeling of antibody structures by a search for minimum-energy recombination of backbone fragments, Proteins 85 (2017) 30–38.
[18] B.D. Weitzner, J.R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R.L. Dunbrack Jr., J.J. Gray, Modeling and docking of antibody structures with Rosetta, Nat. Protoc. 12 (2017) 401–416.
[19] F. Ambrosetti, B. Jimenez-Garcia, J. Roel-Touris, A. Bonvin, Modeling antibody-antigen complexes by information-driven docking, Structure 28 (2020) 119–129 e112.
[20] J.D. Guest, T. Vreven, J. Zhou, I. Moal, J.R. Jeliazkov, J.J. Gray, Z. Weng, B. G. Pierce, An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants, Structure 29 (2021) 606–621 e605.
[21] K. Yugandhar, M.M. Gromiha, Protein-protein binding affinity prediction from amino acid sequence, Bioinformatics 30 (2014) 3583–3589.
[22] S. Pittala, C. Bailey-Kellogg, Mixture of experts for predicting antibody-antigen binding affinity from antigen sequence, bioRxiv (2019), 511360.
[23] C. Ye, W. Hu, B. Gaeta, Machine Learning Prediction of Antibody-Antigen Binding: Dataset, Method and Testing, bioRxiv, 2021, 2021.2003.2019.435772.
[24] Y. Kang, D. Leng, J. Guo, L.J.A. Pan, Sequence-based Deep Learning Antibody Design for in Silico Antibody Affinity Maturation, abs/2103.03724, 2021.

[25] T. Sulea, V. Vivcharuk, C.R. Corbeil, C. Deprez, E.O. Purisima, Assessment of solvated interaction energy function for ranking antibody-antigen binding affinities, J. Chem. Inf. Model. 56 (2016) 1292–1303.
[26] S. Marillet, M.P. Lefranc, P. Boudinot, F. Cazals, Novel structural parameters of ig-Ag complexes yield a quantitative description of interaction specificity and binding affinity, Front. Immunol. 8 (2017) 34.
[27] Y. Myung, D.E.V. Pires, D.B. Ascher, CSM-AB: graph-based antibody-antigen binding affinity prediction and docking scoring function, Bioinformatics 38 (4) (2022) 1141–1143.
[28] J.M. Choi, A.W.R. Serohijos, S. Murphy, D. Lucarelli, L.L. Lofranco, A. Feldman, E. I. Shakhnovich, Minimalistic predictor of protein binding energy: contribution of solvation factor to protein binding, Biophys. J. 108 (2015) 795–798.
[29] Y.X. Yang, P. Wang, B.T. Zhu, Importance of interface and surface areas in protein-protein binding affinity prediction: a machine learning analysis based on linear regression and artificial neural network, Biophys. Chem. 283 (2022), 106762.
[30] A. Vangone, A.M. Bonvin, Contacts-based prediction of binding affinity in protein-protein complexes, Elife 4 (2015), e07454.
[31] L.C. Xue, J.P. Rodrigues, P.L. Kastritis, A.M. Bonvin, A. Vangone, PRODIGY: a web server for predicting the binding affinity of protein-protein complexes, Bioinformatics 32 (2016) 3676–3678.
[32] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, C. M. Deane, SAbDab: the structural antibody database, Nucleic Acids Res. 42 (2014) D1140–D1146.
[33] T.B. Fischer, J.B. Holmes, I.R. Miller, J.R. Parsons, L. Tung, J.C. Hu, J. Tsai, Assessing methods for identifying pair-wise atomic contacts across binding interfaces, J. Struct. Biol. 153 (2006) 103–112.
[34] J. Ribeiro, C. Rios-Vera, F. Melo, A. Schuller, Calculation of accurate interatomic contact surface areas for the quantitative analysis of non-bonded molecular interactions, Bioinformatics 35 (2019) 3499–3501.
[35] Y. Nievergelt, A tutorial history of least squares with applications to astronomy and geodesy, J. Comput. Appl. Math. 121 (2000) 37–72.
[36] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536.
[37] J. Li, J.-h. Cheng, J.-y. Shi, F. Huang, Brief introduction of back propagation (BP) neural network algorithm and its improvement, in: D. Jin, S. Lin (Eds.), Advances in Computer Science and Information Engineering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 553–558.
[38] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[39] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, 1983.
[40] V. Vivcharuk, J. Baardsnes, C. Deprez, T. Sulea, M. Jaramillo, C.R. Corbeil, A. Mullick, J. Magoon, A. Marcil, Y. Durocher, M.D. O'Connor-McCourt, E. O. Purisima, Assisted design of antibody and protein therapeutics (ADAPT), PLoS One 12 (2017).
[41] J. Adolf-Bryfogle, O. Kalyuzhniy, M. Kubitz, B.D. Weitzner, X. Hu, Y. Adachi, W. R. Schief, R.L. Dunbrack Jr., RosettaAntibodyDesign (RAbD): a general framework for computational antibody design, PLoS Comput. Biol. 14 (2018), e1006112.
[42] D.C.P. Lee, R. Raman, N.A. Ghafar, Y. Budigi, An antibody engineering platform using amino acid networks: a case study in development of antiviral therapeutics, Antivir. Res. 192 (2021).
[43] D. Prihoda, J. Maamary, A. Waight, V. Juan, L. Fayadat-Dilman, D. Svozil, D. A. Bitton, BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning, mAbs 14 (2022), 2020203.
[44] I.S. Mian, A.R. Bradwell, A.J. Olson, Structure, function and properties of antibody binding sites, J. Mol. Biol. 217 (1991) 133–151.
[45] K. Tsumoto, K. Ogasahara, Y. Ueda, K. Watanabe, K. Yutani, I. Kumagai, Role of Tyr residues in the contact region of anti-lysozyme monoclonal antibody HyHEL10 for antigen binding, J. Biol. Chem. 270 (1995) 18551–18557.
[46] F.A. Fellouse, P.A. Barthelemy, R.F. Kelley, S.S. Sidhu, Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code, J. Mol. Biol. 357 (2006) 100–114.
[47] S. Birtalan, Y. Zhang, F.A. Fellouse, L. Shao, G. Schaefer, S.S. Sidhu, The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies, J. Mol. Biol. 377 (2008) 1518–1528.
[48] S. Birtalan, R.D. Fisher, S.S. Sidhu, The functional capacity of the natural amino acids for molecular recognition, Mol. Biosyst. 6 (2010) 1186–1194.
[49] A. Gonzalez-Munoz, E. Bokma, D. O'Shea, K. Minton, M. Strain, K. Vousden, C. Rossant, L. Jermutus, R. Minter, Tailored amino acid diversity for the evolution of antibody affinity, mAbs 4 (2012) 664–672.
[50] G. Robin, Y. Sato, D. Desplancq, N. Rochel, E. Weiss, P. Martineau, Restricted diversity of antigen binding residues of antibodies revealed by computational alanine scanning of 227 antibody-antigen complexes, J. Mol. Biol. 426 (2014) 3729–3743.
[51] K.E. Tiller, L. Li, S. Kumar, M.C. Julian, S. Garde, P.M. Tessier, Arginine mutations in antibody complementarity-determining regions display context-dependent affinity/specificity trade-offs, J. Biol. Chem. 292 (2017) 16638–16652.
[52] A. Fukunaga, S. Maeta, B. Reema, M. Nakakido, K. Tsumoto, Improvement of antibody affinity by introduction of basic amino acid residues into the framework region, Biochem. Biophys. Rep. 15 (2018) 81–85.
[53] N. Sinha, S. Mohan, C.A. Lipschultz, S.J. Smith-Gill, Differences in electrostatic properties at antibody-antigen binding sites: implications for specificity and cross-reactivity, Biophys. J. 83 (2002) 2946–2968.
[54] S.M. Lippow, K.D. Wittrup, B. Tidor, Computational design of antibody-affinity improvement beyond in vivo maturation, Nat. Biotechnol. 25 (2007) 1171–1176.
[55] M. Kiyoshi, J.M. Caaveiro, E. Miura, S. Nagatoishi, M. Nakakido, S. Soga, H. Shirai, S. Kawabata, K. Tsumoto, Affinity improvement of a therapeutic antibody by structure-based computational design: generation of electrostatic interactions in the transition state stabilizes the antibody-antigen complex, PLoS One 9 (2014), e87099.
[56] B. North, A. Lehmann, R.L. Dunbrack Jr., A new clustering of antibody CDR loop conformations, J. Mol. Biol. 406 (2011) 228–256.
[57] A. Teplyakov, G. Obmolova, T.J. Malia, J. Luo, S. Muzammil, R. Sweet, J. C. Almagro, G.L. Gilliland, Structural diversity in a human antibody germline library, mAbs 8 (2016) 1045–1063.
[58] C. Regep, G. Georges, J. Shi, B. Popovic, C.M. Deane, The H3 loop of antibodies shows unique structural characteristics, Proteins: Struct., Funct., Bioinf. 85 (2017) 1311–1318.
[59] M.L. Fernandez-Quintero, J. Kraml, G. Georges, K.R. Liedl, CDR-H3 loop ensemble in solution - conformational selection upon antibody binding, mAbs 11 (2019) 1077–1088.