RESEARCH

DG-Affinity: predicting antigen–antibody affinity with language models from sequences

Ye Yuan[1*†], Qushuo Chen[2†], Jun Mao[2], Guipeng Li[2] and Xiaoyong Pan[1*]

†Ye Yuan and Qushuo Chen are co-authors.

*Correspondence:
yuanye_auto@sjtu.edu.cn;
2008xypan@sjtu.edu.cn

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China
[2] DigitalGene, Ltd, Shanghai 200240, China

Abstract

Background: Antibody-mediated immune responses play a crucial role in the immune defense of human body. The evolution of bioengineering has led the progress of antibody-derived drugs, showing promising efficacy in cancer and autoimmune disease therapy. A critical step of this development process is obtaining the affinity between antibodies and their binding antigens.

Results: In this study, we introduce a novel sequence-based antigen–antibody affinity prediction method, named DG-Affinity. DG-Affinity uses deep neural networks to efficiently and accurately predict the affinity between antibodies and antigens from sequences, without the need for structural information. The sequences of both the antigen and the antibody are first transformed into embedding vectors by two pre-trained language models, then these embeddings are concatenated into an ConvNeXt framework with a regression task. The results demonstrate the superiority of DG-Affinity over the existing structure-based prediction methods and the sequence-based tools, achieving a Pearson's correlation of over 0.65 on an independent test dataset.

Conclusions: Compared to the baseline methods, DG-Affinity achieves the best performance and can advance the development of antibody design. It is freely available as an easy-to-use web server at https://www.digitalgeneai.tech/solution/affinity.

Keywords: Affinity, Deep learning, Sequence embedding, Antibody–antigen interaction

Background

Antibody-mediated immune response is a central component of human immune system. Antibodies are a special protein that can specifically recognize invading antigens, such as viruses, by binding to epitopes on the antigens through the two ends of their Y-shaped structure, known as the complementarity-determining regions (CDRs) [1–3]. Due to the high diversity of CDRs, they show binding specificity toward specific antigens [4]. The biopharmaceutical industry has utilized this specificity to develop monoclonal antibodies (MAbs) as therapeutic drugs, which have high success rates and efficacy for diseases. In addition, they suffer to minimal side effects [5–8]. With the advancement of biotechnology techniques, such as antibody–drug conjugates (ADCs), even traditional

"undruggable" targets of diseases can be targeted. Antibodies can be used to treat various cancers, as well as autoimmune diseases like rheumatoid arthritis, attracting huge research attention and development efforts [8–12]. Since the approval of the first monoclonal antibody, antibodies have become popular drugs, occupying more than half of the therapeutic market [13]. The latest application of monoclonal antibodies is the treatment of the 2019 coronavirus disease (COVID-19), since some patients may not be suitable for vaccination due to severe allergic reactions or inability to generate protective immune responses from the vaccine. Recently, monoclonal antibodies against SARS-CoV-2, such as bebtelovimab, tixagevimab and cilgavima, have been approved by the FDA for the treatment or pre-exposure prevention of COVID-19, demonstrating that monoclonal antibodies can be an effective complement to vaccines against COVID-19 [14–21].

Determining the affinity of antibody–antigen interactions is an important step in antibody development. Experimental methods for affinity determination include radio-immunoassay (RIA), enzyme-linked immunosorbent assay (ELISA), surface plasmon resonance (SPR), and bio-layer interferometry (BLI) [22–26]. However, some of these experimental methods are resource-intensive and time-consuming. Moreover, these experimental methods are not suitable for large-scale high-throughput antibody screening [27]. Fortunately, extensive immunological databases from experiments have been established, generating a wealth of experimental affinity data for antigen and antibody studies [28–33]. With the advancement of artificial intelligence technologies, especially, deep learning performs better than traditional machine learning methods on large datasets. For example, ConvNeXt outperforms the Swin-T model in multiple classification and recognition tasks. The model with ConvNeXt as the backbone has also achieved good results in fields such as medical imaging and traditional Chinese medicine. It has become possible to build predictive models based on these collected data and deep learning methods to predict antibody–antigen affinity [34–36] with high accuracy. For example, PIPR is a sequence-based method and employs a residual RCNN [37] to predict binding affinity using information from antigen–antibody pairs. It achieves good generalization performance on various tasks. The RCNN framework in PTPR adopts a bidirectional gated recursive unit module (GRU), however, GRU has the drawbacks of slow learning efficiency and convergence speed [38]. Another model is the CSM-AB model [39], it first requires docking of antibody and antigen structures or utilizes known complex structures, and then obtains geometric information of the contact interface to establish a predictive model using Extra Trees algorithm. Recently, the AREA-AFFINITY was developed to predict antibody–antigen binding affinity [40]. It built different models including linear model, neural network, random forest and mixed model. The mixed model yields the best performance than other compared methods. Similar to CSM-AB, the AREA-AFFINITY is also a structure-based model. However, the limitation lies in the requirement for antigen–antibody complex structure information, which is difficult to acquire.

In this study, we propose a sequence-based method DG-Affinity for predicting antibody–antigen binding affinity. It is trained on a larger and more comprehensive dataset than CSM-AB, and only utilizes sequence information to predict the affinity between antibodies and antigens. DG-Affinity combined two pre-trained embeddings (TAPE for antigen sequences and Ablang for antibody sequences) on an antibody–antigen interaction dataset,

Yuan *et al. BMC Bioinformatics*   (2023) 24:430

Page 3 of 12

and used a ConvNeXt framework [41] to learn the relationship of antibody–antigen binding affinity. DG-Affinity outperforms other existing methods in an independent test dataset.

Materials and methods

Benchmark datasets

The benchmark antigen–antibody data comes from two primary sources. One is the sdAb-DB database (http://www.sdab-db.ca) [42], which is a freely available repository that collects antibody–antigen data. The other is the Round A data from the Baidu PaddlePaddle 2021 Global Antibody Affinity Prediction Competition. In this study, we combine the two datasets by removing the shared antibody–antigen interactions to construct the benchmark dataset. The benchmark dataset comprises 1,673 entries involving 448 distinct antibody–antigen complexes [43].

We divided the data into five equal parts. Four parts were used for training and one for validation, take turns to conduct five times of cross-validation. Then, we utilized an independent test set to evaluate the model's generalizability. This independent test set came from (https://github.com/piercelab/antibody_benchmark) [33], which contains structural files for 42 antigen–antibody complexes with affinity values not overlapping with the training data. We selected structures where both the antibody and the antigen are single chains, totaling 26, and used the PDB module in Biopython to extract sequence information from the complex structures as the independent test set [44].

We processed the data by treating each antibody–antigen pair and its corresponding affinity label separately. As shown in Fig. 1a, the value range of the original dissociation constant (kd) value of the binding affinity is from -2 to -16, and was then transformed by taking the negative logarithm, as used in [45], and dividing by 10 for normalization as follows:

$$y_{label} = \frac{-(y_{kd} * 1000)/591.282}{log_e(10)/10}$$

where y_{kd} is the original value and 591.82 is from the study [45].

As shown in Fig. 1b, it can be clearly seen that the original kd value was successfully converted to the range of 0–1, and only a small number of abnormal data values were mapped beyond 1.
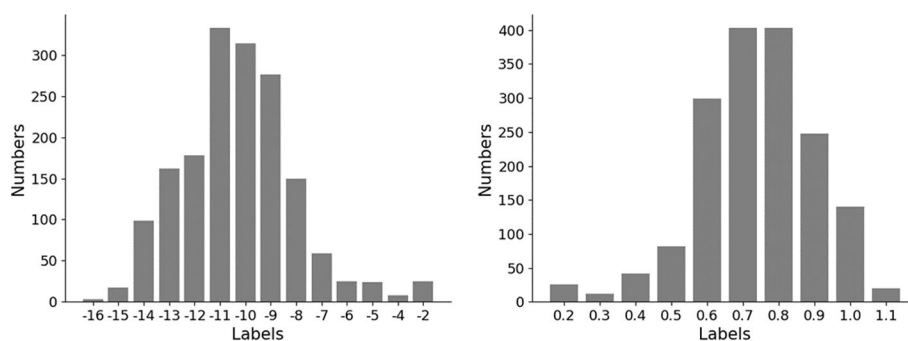

Fig. 1 Distribution of affinity values before and after data preprocessing. **a** Before preprocessing and **b** after preprocessing. The x-axis represents the data label value and the y-axis represents the numbers of data within this value range

Sequence embedding of antibody and antigen

For the antigen sequence embeddings, we used TAPE's pre-trained model to obtain [46], a protein language model, the embeddings. A protein language model is a type of language model designed for the protein sequences. It is trained on protein sequences and learns underlying biochemical properties, secondary and tertiary structures, and intrinsic functional patterns.

TAPE uses bi-directional encoder from the Transformers model, and is trained on 31 million protein sequences from Pfam53 [47]. The model's effectiveness was validated across six downstream tasks including remote homology detection, contact prediction, and protein engineering tasks.

Considering that antibody is different from general proteins, for antibody embeddings, we used AbLang for embeddings [48]. AbLang is an antibody-specific language model trained on the Observed Antibody Space (OAS) database [49, 50], which contains about 71.98 million sequence data (52.89 million unpaired heavy chains and 19.09 million unpaired light chains). It can be used for antibody residue repair, sequence-specific predictions, and residue sequence embedding representations, and AbLang provides more accurate antibody representation than ESM-1b [51, 52]. Interestingly, AbLang requires separate embeddings for the heavy and light chains of the antibody, as two AbLang models were trained, one for the heavy chain and the other for the light chain. One potential reason for training separate models is that heavy and light chains have different components: the light chain has two such immunoglobulin domains, whereas the heavy chain of the antibody contains four.

ConvNeXt backbone

The ConvNeXt network is composed purely of convolutional layers and inspired by the architecture of vision transformer and ResNet [53–55]. ConvNeXt mainly improves the model performance in the following aspects: (1) Macro design (2) ResNeXt (3) Reverse bottleneck (4) Large kernel size (5) Various layered micro designs. Overall, the ConvNeXt network has four stages, with a block stacking ratio of 1:1:3:1 for each stage, and a convolutional layer with the same kernel size of 4 and a step size as the Swin-T network. The ConvNeXt network has also designed an anti-bottleneck structure based on "fine end coarse medium" as a reference, replacing RELU with more commonly used GELU activation function, resulting in a slight improvement in model performance. ConvNeXt not only reduces the use of regularization functions, but also replaces Batch Norm with Layer Norm. These two operations slightly improve the accuracy of the model. In this study, we removed the final MLP layer of the original ConvNeXt backbone as one of the modules for DG-Affinity.

Architecture of DG-Affinity

Our DG-Affinity's architecture is shown in Figs. 2 and 3, three parallel ConvNeXt backbones accept three different features obtained from TAPE or Ablang feature extractors, which are antibody features, antigen features, and antibody–antigen concatenated features. After passing through the ConvNeXt backbone for feature
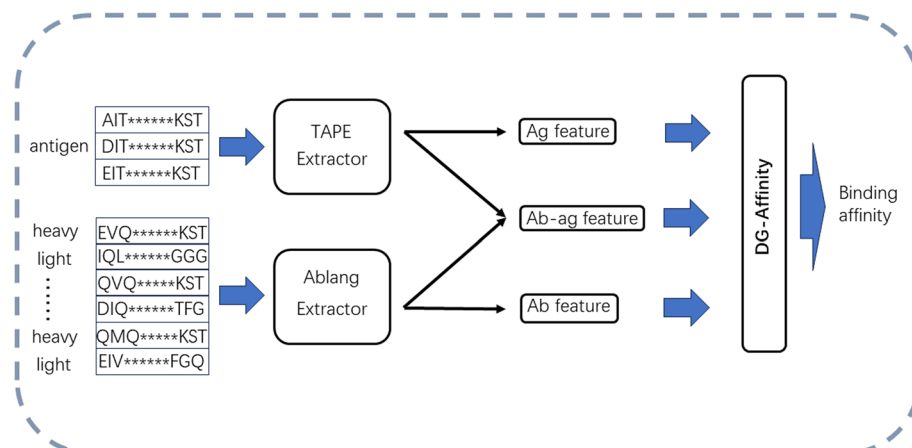
Fig. 2 The workflow of DG-Affinity. Among them, Ab represents antibodies and Ag represents antigens. The antibody and antigen sequences are respectively fed into the corresponding embedding extractors, and three features are obtained: Ab features, Ag features, and Ab-ag features. Then, these three features are concatenated into the MLP to predict binding affinity values
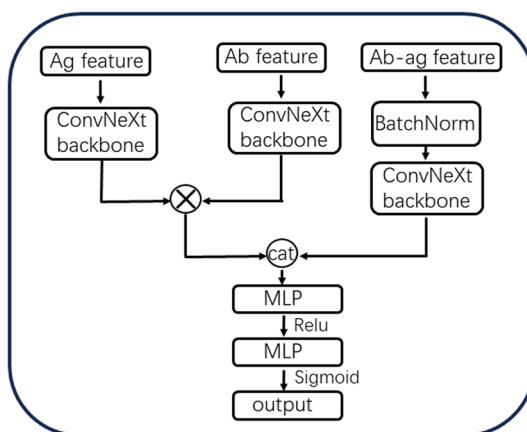

Fig. 3 Network architecture of DG-Affinity, "cat" symbol represents concatenation, "×" symbol represents element-wise multiplication, "Ag feature" is antigen feature, "Ab feature" is antibody feature, "Ab–ag feature" is concatenated feature of antigen feature and antibody feature and MLP is a single linear layer

representation learning, the antibody representation and antigen representation are multiplied element-wise and concatenated with the antibody–antigen feature representation learned through the ConvNeXt backbone. Finally, the new representation is passed through two layers of MLP to predict the affinity value, among them, each layer of MLP is composed of a linear function followed by relu or sigmoid activate function. In this study, we used the open-source code to construct an ConvNeXt backbone (https://github.com/facebookresearch/ConvNeXt). Considering the capacity and performance of the model, we chose the tiny ConvNeXt version. The neural network was built and trained using the Pytorch library [56], and we trained the model using 50 epochs with a learning rate of 0.000001, using the ADAM optimizer [57]. The network output is a regression value prediction of the affinity.

Performance metrics

Our model's predictive ability was measured using the Pearson correlation coefficient (R), R-squared (R2), Root Mean Square Deviation (RMSD) and Mean Absolute Error (MAE). To better evaluate our model, we also tested other well-known structure-based antigen–antibody affinity prediction model (e.g. CSM-AB and AREA-AFFINITY). We input the 26 complex structures from the independent set into their online prediction website and manually calculated the R, R2, RMSD and MAE values based on the true and predicted binding affinity values for each method, respectively.

Baseline methods

CSM-AB It is the first scoring function specifically designed for antibody antigen docking and binding affinity prediction. By adjusting the graph-based structure, this method can capture close-contact features and surrounding structural information related to antibody antigen binding.

AREA-AFFINITY It integrates 60 effective area-based protein–protein affinity prediction models and 37 effective area models for antibody protein antigen binding affinity prediction.

LISA It is an empirical affinity function based on the atomic-atomic contact model and a radial function based on the density functional theory.

CIPS It is a new pair potential combining interface composition with residue–residue contact preference.

Prodigy It is based on the counting of atom–atom contacts at the interface and on the charge distribution at the non-interacting surface of the complex.

NIS It considers distinguished physico-chemical properties of residues lying on the complex surface.

CCharPPI It integrates over a hundred tools into a single web server, including electrostatic, solvent removal, and hydrogen bonding models, as well as interface filling and complementarity scores, empirical potentials of various resolutions, docking potentials, and composite scoring functions.

We follow the comparison protocol of CSM-AB and LISA [58]. In the experiment, LISA, CIPS [59], NIS [60] and PRODIGY [61] were standalone scripts, AREA-AFFINITY [40] and CSM-AB were one-line webserver, other models or tools (e. g. FIREDOCK) were calculated using CCharPPI webserver based on physical potentials and composite descriptors [62]. We collected structure data from SAbDab database and Protein Data Bank since most of the existing comparative models are based on structure [63, 64].

Results

Comparison DG-Affinity with other baseline methods

To demonstrate the advantages of DG-Affinity, we evaluated other methods on the training set, the results in Fig. 4a and Additional file 1: Table S1 show that DG-Affinity significantly outperformed all of them. At the same time, we also evaluate the stability and consistency of our model's performance using tenfold cross validation, 15-fold cross validation, and 20-fold cross validation. The results are stable and consistent (shown in Additional file 1: Figure S1), showing that there is no significant sampling bias for
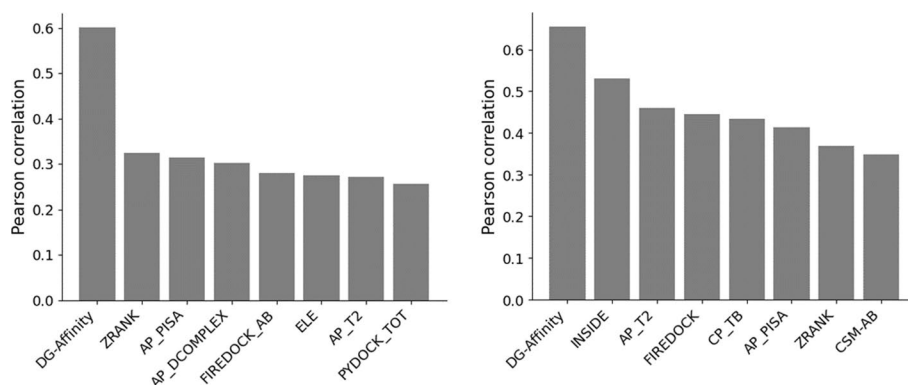
Fig. 4 performance distribution of the top 8 models on the independent set (**a**) and the training set (**b**), respectively

DG-Affinity. After manually adjusting the model parameters to the best performance on the validation set, we compared it with baseline methods on the independent test set. As shown in Fig. 4b and Additional file 1: Table S2, DG-Affinity achieves the best R of over 0.6556 in the independent set. Most of the baseline models yield an R from 0.3 to 0.5, which is much lower than that of DG-Affinity. Moreover, two models have negative Pearson's correlation, such as CP_PIE (-0.3332) and AREA-AFFINITY (-0.2019). In addition, our method outperforms other methods in all the metrics (Additional file 1: Table S2).

Comparing the effectiveness of different architectures in DG-Affinity

DG-Affinity uses ConvNeXt as the backbone network, to demonstrate the effectiveness, we compare it with the other 12 widely used backbone networks, i.e. convolutional network, transformer, and several classic models. The source codes for these architectures are downloaded from the corresponding github repository (https://github.com/weiaicunzai/pytorch-cifar100), and slightly modified to remove the final MLP layer and replace it with the global mean pooling. These backbones replace all ConvNeXt modules in DG-Affinity, and the replaced model structure is trained and validated using the aforementioned training set and independent test set. For detailed hyperparameter information, please refer to Additional file 1: Table S3 in the supporting material. The results are shown in Fig. 5 and Additional file 1: Table S4, S5. Of the 13 network backbones, the ConvNeXt achieves relatively better performance than other backbone networks in the fivefold cross validation and in the independent test dataset.

Exploration of model ablation studies

In this section, we investigated the contribution of each module in DG-Affinity. During the experiment, we maintain the parameter consistency. After respectively removing the ConvNeXt module for learning antigen features, antibody features, and concatenated features antibody and antigen embeddings, we concatenate the output representation vectors of these remaining ConvNeXt modules and input them into the MLP layer to make regression. As shown in the Fig. 6, it is evident that the lack of the module for learning antigen features has a significant impact on the performance of DG-Affinity. The results show that antigen information may be more closely related to the interactions than antibodies. We
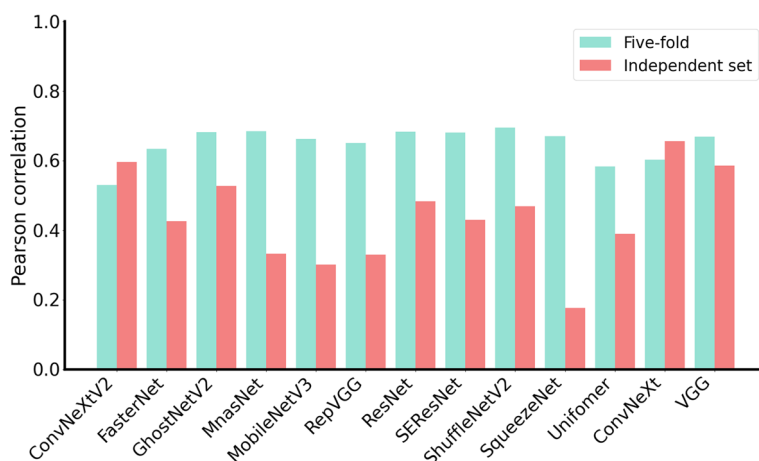
Fig. 5 Performance of different network backbones in DG-Affinity on the training set (five-fold) and independent set
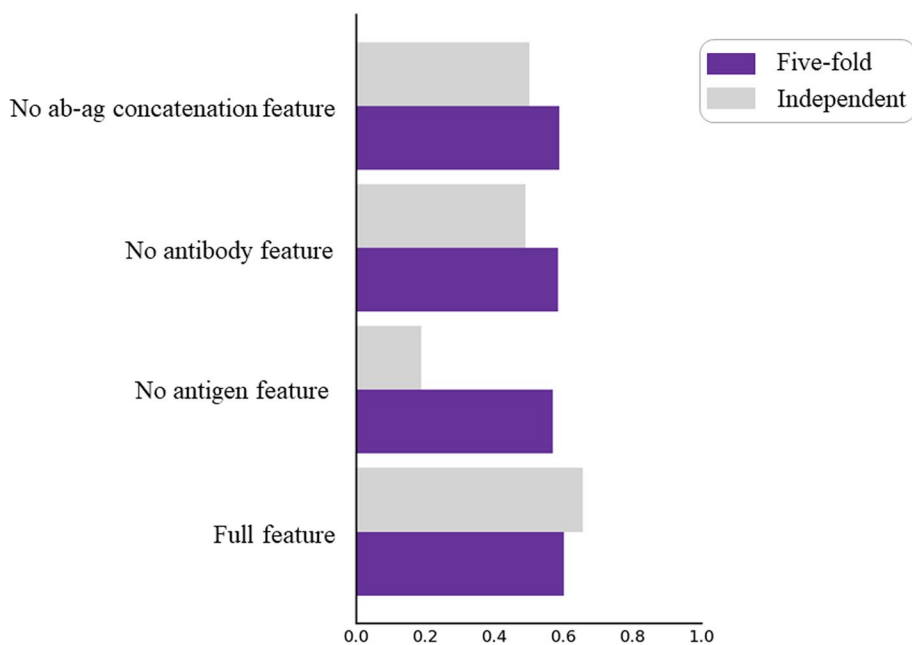

Fig. 6 The ablation experiments of our model on the training set (five-fold) and independent set. The results of model in training and independent sets are represented by purple and gray bars

have also found that when using full feature, the performance on the independent datasets in Fig. 5 is higher than that of five-fold. One potential reason is that the independent datasets are not large enough and the samples are imbalanced, a more number of features potentially introduce better generalizability, resulting in a better performance on the independent test set.

Discussion

The development of bioengineering has led the development of antibody drugs, demonstrating good therapeutic effects in the treatment of cancer and autoimmune diseases. The key step in this development process is to obtain the affinity between antibodies and their binding antigens. However, the limitation of current methods lies in the requirement for structural information of antigen antibody complexes, which is difficult to obtain. To address the above issues, in this study, we propose a deep learning based model for predicting antigen–antibody binding affinity with pre-trained emebddings from sequences, our model has achieved promising performance due to the followings: (1) the training dataset is larger, while traditional structure-based methods have much smaller structure data set than sequences, making them prone to overfitting and low generalization performance. (2) The ConvNeXt framework has recently been widely used in various fields and has been proven to achieve good prediction results. (3) The pre-trained emeddings on the large unlabeled sequences for proteins and antibody.

Conclusion

In this study, we propsoed a new sequence-based antigen–antibody binding affinity prediction method, named DG-Affinity, based on protein and antibody language models. Antigen and antibody sequences are first transformed into the embedding vectors through two pre-trained methods (TAPE and Ablang), then a ConvNeXt-backbone based network is used to learn the affinity relationship between antigen and antibody. The results on benchmark datasets indicate that DG-Affinity outperforms existing methods, including the popular structure-based antigen antibody affinity prediction methods as well as the traditional tools, for both the fivefold and independent validation, achieving a Pearson correlation coefficient of over 0.65 on the independent test datasets. In addition, we developed an easy-to-use website version of DG-Affinity. It can be expected that our method DG-Affinity will advance the progress and development of antibody drug.

Abbreviations

CDRs	Complementarity-determining regions
MAbs	Monoclonal antibodies
ADCs	Antibody–drug conjugates
RIA	Adioimmunoassay
ELISA	Enzyme-linked immunosorbent assay
SPR	Surface Plasmon resonance
BLI	Bio-layer interferometry
GRUs	Bidirectional gated recursive unit module
PIPR	Protein–Protein Interaction Prediction Based on Siamese Residual RCNN
OAS	Observed Antibody Space
ESM	Evolutionary Scale Modeling

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05562-z.

Additional file 1: Table S1. Performance comparison between DG-Affinity and other models on training set. **Table S2.** Performance comparison between DG-Affinity and other models on independent set. **Table S3.** Parameter settings for each model backbone on DG Affinity. **Table S4.** Performance of DG-Affinity with different backbones on 5-fold cross validation. **Table S5.** Performance of DG-Affinity with different backbones on independent datasets.

Fig. S1. Performance distribution of DG-Affinity per fold under different validation schemes including, 5-fold, 10-fold and 20-fold cross validation, Prove the consistency and robustness of this method.

Availability of data and materials
The website is provided at https://www.digitalgeneai.tech/solution/affinity, this study does not include any human involvement or human data which is not publicly available.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References
1. Oostindie SC, Lazar GA, Schuurman J, Parren PWHI. Avidity in antibody effector functions and biotherapeutic drug design. Nat Rev Drug Discov. 2022;21(10):715–35.
2. Hviid L, Lopez-Perez M, Larsen MD, Vidarsson G. No sweet deal: the antibody-mediated immune response to malaria. Trends Parasitol. 2022;38(6):428–34.
3. Rascio F, Pontrelli P, Netti GS, Manno E, Infante B, Simone S, Castellano G, Ranieri E, Seveso M, Cozzi E, Gesualdo L, Stallone G, Grandaliano G. IgE-mediated immune response and antibody-mediated rejection. Clin J Am Soc Nephrol. 2020;15(10):1474–83.
4. Kapingidza AB, Kowal K, Chruszcz M. Antigen–antibody complexes. Subcell Biochem. 2020;94:465–97.
5. Bayer V. An overview of monoclonal antibodies. Semin Oncol Nurs. 2019;35(5):150927.
6. Posner J, Barrington P, Brier T, Datta-Mannan A. Monoclonal antibodies: past, present and future. Handb Exp Pharmacol. 2019;260:81–141.
7. Castelli MS, McGonigle P, Hornby PJ. The pharmacology and therapeutic applications of monoclonal antibodies. Pharmacol Res Perspect. 2019;7(6):e00535.
8. Le Basle Y, Chennell P, Tokhadze N, Astier A, Sautou V. Physicochemical stability of monoclonal antibodies: a review. J Pharm Sci. 2020;109(1):169–90.
9. Hafeez U, Parakh S, Gan HK, Scott AM. Antibody-drug conjugates for cancer therapy. Molecules. 2020;25(20):4764.
10. Ponziani S, Di Vittorio G, Pitari G, Cimini AM, Ardini M, Gentile R, Iacobelli S, Sala G, Capone E, Flavell DJ, Ippoliti R, Giansanti F. Antibody-drug conjugates: the new frontier of chemotherapy. Int J Mol Sci. 2020;21(15):5510.
11. Baah S, Laws M, Rahman KM. Antibody-drug conjugates-a tutorial review. Molecules. 2021;26(10):2943.
12. Jin Y, Schladetsch MA, Huang X, Balunas MJ, Wiemer AJ. Stepping forward in antibody-drug conjugate development. Pharmacol Ther. 2022;229:107917.
13. Lu RM, Hwang YC, Liu IJ, Lee CC, Tsai HZ, Li HJ, Wu HC. Development of therapeutic antibodies for the treatment of diseases. J Biomed Sci. 2020;27(1):1.
14. Corti D, Purcell LA, Snell G, Veesler D. Tackling COVID-19 with neutralizing monoclonal antibodies. Cell. 2021;184(12):3086–108.
15. Taylor PC, Adams AC, Hufford MM, de la Torre I, Winthrop K, Gottlieb RL. Neutralizing monoclonal antibodies for treatment of COVID-19. Nat Rev Immunol. 2021;21(6):382–93.
16. Hwang YC, Lu RM, Su SC, Chiang PY, Ko SH, Ke FY, Liang KH, Hsieh TY, Wu HC. Monoclonal antibodies for COVID-19 therapy and SARS-CoV-2 detection. J Biomed Sci. 2022;29(1):1.
17. Bakkari MA, Moni SS, Sultan MH, Madkhali OA. Monoclonal antibodies and their target specificity against SARS-CoV-2 Infections: perspectives and challenges. Recent Pat Biotechnol. 2022;16(1):64–78.
18. Cruz-Teran C, Tiruthani K, McSweeney M, Ma A, Pickles R, Lai SK. Challenges and opportunities for antiviral monoclonal antibodies as COVID-19 therapy. Adv Drug Deliv Rev. 2021;169:100–17.

19. Tabll AA, Shahein YE, Omran MM, Elnakib MM, Ragheb AA, Amer KE. A review of monoclonal antibodies in COVID-19: Role in immunotherapy, vaccine development and viral detection. Hum Antib. 2021;29(3):179–91.
20. Asdaq SMB, Rabbani SI, Alkahtani M, Aldohyan MM, Alabdulsalam AM, Alshammari MS, Alajlan SA, Binrokan A, Mohzari Y, Alrashed A, Alshammari MK, Imran M, Nayeem N. A patent review on the therapeutic application of monoclonal antibodies in COVID-19. Int J Mol Sci. 2021;22(21):11953.
21. Focosi D, McConnell S, Casadevall A, Cappello E, Valdiserra G, Tuccori M. Monoclonal antibody therapies against SARS-CoV-2. Lancet Infect Dis. 2022;22(11):e311–26. https://doi.org/10.1016/S1473-3099(22)00311-5. (**Erratum in: Lancet Infect Dis. 2022;22(9): e239**).
22. Liao J, Madahar V, Dang R, Jiang L. Quantitative FRET (qFRET) technology for the determination of protein–protein interaction affinity in solution. Molecules. 2021;26(21):6339.
23. Tabatabaei MS, Ahmed M. Enzyme-linked immunosorbent assay (ELISA). Methods Mol Biol. 2022;2508:115–34.
24. Sparks RP, Jenkins JL, Fratti R. Use of surface plasmon resonance (SPR) to determine binding affinities and kinetic parameters between components important in fusion machinery. Methods Mol Biol. 2019;1860:199–210.
25. Rhea K. Determining the binding kinetics of peptide macrocycles using bio-layer interferometry (BLI). Methods Mol Biol. 2022;2371:355–72.
26. Mir DA, Mehraj U, Qayoom H, et al. Radioimmunoassay (RIA) (2020).
27. Guo Z, Yamaguchi R. Machine learning methods for protein-protein binding affinity prediction in protein design. Front Bioinform. 2022;2:1065703.
28. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 2004;47(12):2977–80.
29. Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J. A structure-based benchmark for protein-protein binding affinity. Protein Sci. 2011;20(3):482–91.
30. Moal IH, Fernández-Recio J. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics. 2012;28(20):2600–7.
31. Jankauskaite J, Jiménez-García B, Dapkunas J, Fernández-Recio J, Moal IH. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. Bioinformatics. 2019;35(3):462–9.
32. Sirin S, Apgar JR, Bennett EM, Keating AE. AB-Bind: antibody binding mutational database for computational affinity predictions. Protein Sci. 2016;25(2):393–409.
33. Guest JD, Vreven T, Zhou J, Moal I, Jeliazkov JR, Gray JJ, Weng Z, Pierce BG. An expanded benchmark for antibody–antigen docking and affinity prediction reveals insights into antibody recognition determinants. Structure. 2021;29(6):606-621.e5.
34. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
35. Ruffolo JA, Chu LS, Mahajan SP, Gray JJ. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. Nat Commun. 2023;14(1):2389.
36. Abanades B, Georges G, Bujotzek A, Deane CM. ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation. Bioinformatics. 2022;38(7):1877–80.
37. Chen M, Ju CJ, Zhou G, Chen X, Zhang T, Chang KW, Zaniolo C, Wang W. Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics. 2019;35(14):i305–14.
38. Lee M. Recent advances in deep learning for protein–protein interaction analysis: a comprehensive review. Molecules. 2023;28(13):5169.
39. Myung Y, Pires DEV, Ascher DB. CSM-AB: graph-based antibody–antigen binding affinity prediction and docking scoring function. Bioinformatics. 2022;38(4):1141–3.
40. Yang YX, Huang JY, Wang P, Zhu BT. AREA-AFFINITY: a web server for machine learning-based prediction of protein–protein and antibody-protein antigen binding affinities. J Chem Inf Model. 2023;63(11):3230–7.
41. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. arXiv e-prints.
42. Wilton EE, et al. sdAb-DB: the single domain antibody database. ACS Synth Biol. 2018;7:2480–4.
43. Wilton EE, Opyr MP, Kailasam S, Kothe RF, Wieden HJ. sdAb-DB: the single domain antibody database. ACS Synth Biol. 2018;7(11):2480–4.
44. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.
45. Cruz VL, Souza-Egipsy V, Gion M, Perez-Garcia J, Cortes J, Ramos J, Vega JF. Binding affinity of trastuzumab and pertuzumab monoclonal antibodies to extracellular HER2 domain. Int J Mol Sci. 2023;24:12031.
46. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, Abbeel P, Song YS. Evaluating protein transfer learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689–701.
47. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):D427–32.
48. Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. Bioinform Adv. 2022;2(1):046.
49. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol. 2018;201(8):2502–9.
50. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. 2022;31(1):141–6.

Yuan *et al. BMC Bioinformatics*   (2023) 24:430

Page 12 of 12

51. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA. 2021;118(15):e2016239118.
52. Rao R, Meier J, Sercu T, et al. Transformer protein language models are unsupervised structure learners. Biorxiv. 2020;15:422761.
53. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T et al. An image is worth 16×16 words: transformers for image recognition at scale. In: Proceedings of the 9th international conference on learning representations, OpenReview.net, Vienna, 2021; 3–7
54. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):87–110.
55. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. IEEE; 2016.
56. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library; 2019.
57. Kingma DP, Ba J. Adam: a method for stochastic optimization. Available from: http://arxiv.org/abs/1412.6980
58. Raucci R, Laine E, Carbone A. Local interaction signal analysis predicts protein-protein binding affinity. Structure. 2018;26(6):905-915.e4.
59. Nadalin F, Carbone A. Protein-protein interaction specificity is captured by contact preferences and interface composition. Bioinformatics. 2018;34(3):459–68.
60. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein-protein complexes. Elife. 2015;4:e07454.
61. Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics. 2016;32(23):3676–8.
62. Moal IH, Jiménez-García B, Fernández-Recio J. CCharPPI web server: computational characterization of protein-protein interactions from structure. Bioinformatics. 2015;31(1):123–5.
63. Schneider C, Raybould MIJ, Deane CM. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. Nucleic Acids Res. 2022;50(D1):D1368–72.
64. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.

Publisher's Note