scientific data

OPEN

DATA DESCRIPTOR

PPB-Affinity: Protein-Protein Binding Affinity dataset for AI-based protein drug discovery

Huaqing Liu[1,6], Peiyi Chen[1,6], Xiaochen Zhai[2], Ku-Geng Huo[3], Shuxian Zhou[1], Lanqing Han[1,4] ✉ & Guoxin Fan[5] ✉

Prediction of protein-protein binding (PPB) affinity plays an important role in large-molecular drug discovery. Deep learning (DL) has been adopted to predict the changes of PPB binding affinities upon mutations, but there was a scarcity of studies predicting the PPB affinity itself. The major reason is the paucity of open-source dataset with PPB affinity data. To address this gap, the current study introduced a large comprehensive PPB affinity (PPB-Affinity) dataset. The PPB-Affinity dataset contains key information such as crystal structures of protein-protein complexes (with or without protein mutation patterns), PPB affinity, receptor protein chain, ligand protein chain, etc. To the best of our knowledge, this is the largest publicly available PPB affinity dataset, and we believe it will significantly advance drug discovery by streamlining the screening of potential large-molecule drugs. We also developed a deep-learning benchmark model with this dataset to predict the PPB affinity, providing a foundational comparison for the research community.

Background & Summary

Protein-based drugs, including cytokines, enzymes, antibodies, and vaccines, continue to be a major research focus due to their impact on a wide range of diseases including cancer, cardiovascular disease, hepatitis, gastrointestinal disease, autoimmune disease, and transplant rejection. However, there are still many challenges in the development of protein drugs, one of which is the screening efficiency of drug candidates. The prediction of protein-protein binding (PPB) affinity is a crucial step in the screening process of protein drugs. Protein drugs usually exert their effects by binding to specific receptors, frequently other proteins. Thus, a higher PPB affinity translates to stronger binding between the drug and the receptor, which in many cases may lead to improved therapeutic efficacy. Prior studies have primarily focused on predicting the changes of PPB binding affinity upon mutations[1–9], which only allows the discovery of structure-similar protein drugs. The datasets for such studies typically include SKEMPI v2.0 dataset[10], AB-Bind[11], and some deep mutagenesis datasets[12–18]. Only a few studies endeavored to predict PPB affinity, most of which focused in predicting antigen-antibody binding affinity or TCR-pMHC binding affinity[19–26], but the prediction accuracy has fallen short of practical application.

The difficulty of predicting PPB affinity lies in the development of prediction algorithms and the sources of data, and data scarcity hinders algorithm development. Currently available datasets include SKEMPI v2.0, AB-Bind, SAbDab[27–29], PDBbind v2020 (http://pdbbind.org.cn/)[30–35], Affinity Benchmark v5.5[25,26,36], and ATLAS[37]. The SKEMPI v2.0 dataset contains 7085 samples of affinity changes upon mutations in protein-protein complexes, including data such as crystal structures of wild-type complexes, protein mutation patterns, and the magnitude of affinity changes. However, the SKEMPI v2.0 dataset only included 345 crystal structures, because most of the samples are mutant type. The AB-Bind dataset is included within the SKEMPI v2.0 dataset. The SAbDab dataset is a collection of antibody crystal structure data, containing over 7000 antibody-antigen binding crystal structures, but only a small portion of these samples records the PPB affinities. The PDBbind v2020 dataset contains 23,496 crystal structures of biomolecular complexes and their affinity data, covering protein-small

[1]Artificial Intelligence Innovation Center, Research Institute of Tsinghua, Pearl River Delta, Guangzhou, 510700, China. [2]Cyagen Biosciences (Suzhou) Inc., Guangzhou, 215000, China. [3]Cyagen Biosciences (Guangzhou) Inc., Guangzhou, 510700, China. [4]Cyagen Biomodels (Guangzhou) Co., Ltd, Guangzhou, 510700, China. [5]Department of Pain Medicine, Shenzhen Nanshan People's Hospital, Shenzhen University Medical School, Shenzhen, 518056, China. [6]These authors contributed equally: Huaqing Liu, Peiyi Chen. ✉e-mail: hanlance@tsinghua-gd.org; fanguoxin@email.szu.edu.cn

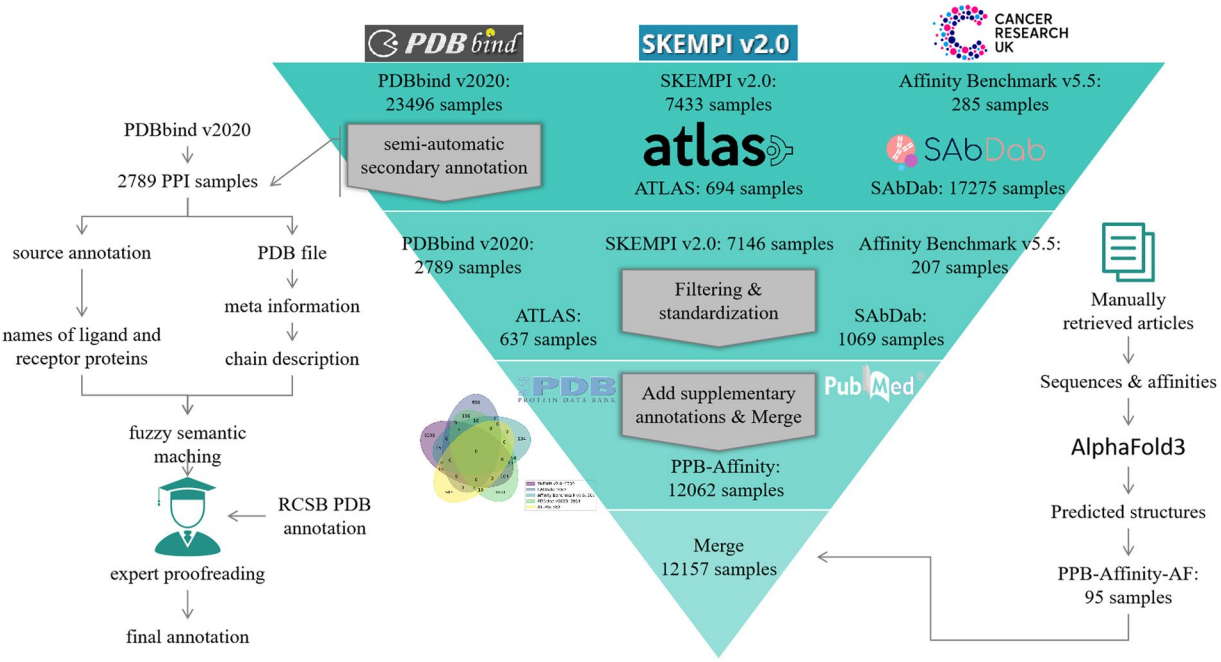

Fig. 1 Schematic workflow of creating the PPB-Affinity database.

molecule ligand complexes, protein-protein complexes, protein-nucleic acid complexes, and nucleic acid-small molecule complexes. There are 2789 samples of protein-protein complexes in the PDBbind v2020 dataset, but it does not explicitly indicate the receptor protein chain and ligand protein chain in the complexes. The Affinity Benchmark v5.5 dataset provides crystal structures of protein-protein complexes and their affinities, but it only consists of 207 protein-protein samples. ATLAS provides affinity data for TCR and its antigen (peptide-MHC complex) binding, affinity changes upon mutations, and complex crystal structures, but it has only 694 samples, with only 112 complex crystal structures obtained from biological experiments. Additionally, the above public databases might fail to include some PPB affinity data promptly, which were experimentally measured and published by most recent studies[38–40]. In summary, there is a paucity of comprehensive public datasets with experimentally measured information for the AI prediction of PPB affinity.

The quality and sample size of experimental data, as well as the diversity of samples, are essential to achieve accurate and highly generalized PPB prediction. Although machine learning (ML) algorithm, especially deep learning (DL) technique has been validated in optimizing multiple steps of drug discovery, there is a paucity of ML or DL studies predicting PPB affinity. To address this gap, we introduce the PPB-Affinity dataset, the largest publicly available comprehensive dataset meticulously integrated and processed from all currently available public data with protein-protein complex crystal structures and their affinities. We believe PPB-Affinity will significantly enhance drug discovery efficiency in the pharmaceutical industry by streamlining the screening of potential large-molecule drugs.

Methods

This study has been exempted by the local ethics committee because the original data are extracted from the publicly available online datasets and then semi-automatic processed.

Data processing and quality assessment. We integrated and processed multiple related open-source datasets through a combination of automatic processing and manual verification (Fig. 1). The source datasets utilized in this study were SKEMPI v2.0, SAbDab (by July 25th, 2024), PDBbind v2020, Affinity Benchmark v5.5, and ATLAS. The crucial information for PPB-Affinity dataset includes the experimentally measured affinity values, the crystal structures of protein-protein complexes (particularly the crystal structures of the binding interfaces), ligand protein chains, and receptor protein chains. To be noted, many samples involve mutant protein complexes, where only the crystal structures of their corresponding wild-type complexes are available. For these samples, we also provide information on the mutation patterns. Additional supplementary information includes the source dataset, experimental temperature for affinity measurement, affinity measurement methods, crystal structure determination methods, and relevant references.

Affinity. Different source datasets provide affinity information in various formats. For instance, SKEMPI v2.0 provides dissociation constant (KD) values in molar units (M), while some samples from SAbDab present both KD values and change in Gibbs free energy upon binding (ΔG) values. PDBbind v2020, on the other hand, offers KD, inhibition constant (Ki), or half maximal inhibitory concentration (IC50) values, with units ranging from millimolar (mM), micromolar (uM), to nanomolar (nM). ATLAS, meanwhile, provides KD values in mM units. To ensure consistency, we have standardized the affinity representation across all samples, expressing them

uniformly as KD values in molar units (M). Notably, IC50 values cannot be directly converted to KD values, and we have excluded them from our dataset as these samples constitute a minority. This approach allows for a more coherent and comparable analysis of PPB affinity across different datasets.

Ligand and receptor types. We only included samples from the source dataset where both receptors and ligands are proteins or polypeptides.

Protein chains. The affinity of PPB physically refers to the change in free energy before and after the binding of receptor proteins and ligand proteins. The greater the reduction in free energy after binding, the stronger the affinity. Therefore, computational methods, including ML, must distinguish between receptor proteins and ligand proteins. Our dataset annotates the ligand protein chains and receptor protein chains in each protein-protein complex. Some source datasets, such as SKEMPI v2.0, ATLAS, and Affinity Benchmark v5.5, have already annotated ligand protein chains and receptor protein chains. For these source datasets, we directly incorporate their annotations and correct a few mislabeled ones. For example, in the Affinity Benchmark v5.5 dataset, the ligand and receptor labeled for 3A4S are chains A and D, respectively. However, there is no binding interface between these two chains in PDB. The correct annotation should be chains B and C. For 4FQI in the Affinity Benchmark v5.5 dataset, the originally labeled receptor and ligand chains are H, L and A, B, E, F, C, D, respectively. Nevertheless, chains E, F, and D do not exist in the PDB. Although chain C exists, it is not an amino acid chain and is far from the binding interface, so it is also excluded. The corrected annotation is chains H and L as ligands, and chains A and B as receptors. Although the protein-protein subset of PDBbind v2020 annotates the names of ligand proteins and receptor proteins in the complex, it does not clearly annotate the chains of ligand proteins and receptor proteins. We used the following semi-automatic annotation method for these samples:

(1) For each sample, read the text names of ligand proteins and receptor proteins annotated by PDBbind v2020;
(2) Read the meta-information of the corresponding PDB file for the sample and obtain the descriptive text for each protein chain;
(3) Use a fuzzy semantic matching method to calculate the matching degree between the ligand and receptor protein name texts and any protein chain description text;
(4) Based on the matching degree, identify the most likely ligand protein chain and receptor protein chain;
(5) Expert proofreading: Structural biology experts observe the crystal structure and sequence of the PDB file, combined with annotation information from the RCSB Protein Data Bank (RCSB PDB) to manually proofread the protein chains found through semantic matching;
(6) Splitting: A PDB file may contain multiple identical ligand-receptor complexes. Structural biology experts observe the crystal structure and sequence in the PDB file, referring to information from the RCSB PDB database, to split and annotate each independent complex. Through this method, we annotated ligand protein chains and receptor protein chain information for 2788 samples.

It is noted that the SAdDab database annotates antigen chains, antibody heavy chains, and antibody light chains. For these samples, we annotate the antigen chain as the receptor chains and the antibody heavy and light chains as the ligand chains. The ATLAS dataset includes affinity data for TCR and p-MHC binding and annotates limited information such as mutated TCR chains and peptide sequences. We identify TCR, peptide-MHC molecules in PDB files through structural biology expert observation, and annotate TCR as the ligand chain and peptide-MHC as the receptor chain. To trace the original data, we also annotate the dataset from which each sample originates and record the original reference as completely as possible. Some records exist in two or more different source datasets, and when merging the source datasets, we only kept one of these records and deleted the rest of the duplicates. To address the cross-source consistency issue, we assign a unique Complex ID for each record in the source dataset, defined by the following formula using the PDB code, ligand and receptor chain codes (sorted), mutation information (sorted), and PubMed ID (the reference for measuring the affinity):

Complex ID = f"{pdb}:{chains}:{mutations}:PMID = {affinity_PMID}"

Samples with the same Complex ID were considered the same sample, which means that their structures and sequences are consistent, and the affinity measurement values come from the same experiment. Therefore, when merging different source datasets based on the Complex ID, we only keep one of the samples with the same Complex ID. In this way, we merged the 5 source datasets as the PPB-Affinity dataset. In other words, there are a few samples that have the same PDB code, ligand, and receptor chain codes (sorted), and mutations, but they have different PubMed IDs, indicating that they come from different references, and the conditions of the affinity measurement experiments may vary, as may the affinity measurement values. As a result, the PPB-Affinity dataset regarded these samples as different samples, and annotated the affinity measurement experimental conditions (including the measurement method, environmental temperature, pH value, etc.) and the size of the affinity for each sample. Additionally, some unqualified data were deleted during manual screening, and the deletion details of samples are listed as follows:

(1) Lack of annotated affinity: There are 57 records without annotated affinity in the SKEMPIv2.0 dataset; the Affinity Benchmark v5.5 has 78 records without annotated affinity; the SAbDab dataset has 14,148 records without annotated affinity.
(2) Records that are not protein-protein complexes: 6 non-protein-protein complexes in the PDBbind v2020 dataset and 46 such records in the SAbDab dataset.

(3) Errors or inability to annotate protein chains: We used a semi-automatic method plus expert review to annotate ligand chain IDs and receptor chain IDs for each record in the PDBbind v.2020 dataset, and for 62 samples, we could not clearly annotate them, so they were not included in our dataset. In addition, 95 records in SAbDab have no antigen chain (receptor chain) annotation; there are 8 records with chain annotation errors because the annotated (antibody heavy chain, light chain, or antigen chain) does not exist in the corresponding PDB file.
(4) The annotated affinity cannot be converted into KD values: There are 62 such samples in PDBbind v2020. Specifically, their affinities are not KD values, Ki values, or ΔG values, but are IC50/EC50 values, etc. Although IC50 or EC50 can be mathematically converted into KD values, the conversion cannot be completed due to the lack of some necessary variable information in the formula (such as substrate concentration, etc.).

We also created a sub-dataset called "PPB-Affinity-AF", which were collected from most recent studies publishing the experimentally measured affinity, and their three-dimensional structure of the complex were created with AphaFold3. The including criteria of the collected samples were: 1. The ΔG or KD of the protein has been determined through experiments, 2. The protein sequence has been published, but there is no crystal structure of the protein complex determined by instruments. These data are scattered in different papers, so we tried our best to collect them from different papers that we could identify. Then we manually organized the protein sequence, affinity, and other information for each sample. A total of 95 samples were finally included, and we used AlphaFold3 to predict the corresponding three-dimensional structure of the complex and saved them as PDB files.

Data Record

The PPB-Affinity dataset[41] is available on https://zenodo.org/doi/10.5281/zenodo.11070823. The dataset is now accessible under the Creative Commons Attribution 4.0 International, which supports its use for educational and research purposes. Users should cite this paper when they incorporate the dataset into their projects. The presented dataset consists of two parts. The first part is an excel spreadsheet (summary.xlsx) that summarizes the annotation information of all samples in the dataset. The second part is a folder containing the crystal structure (PDB files) of all samples in the dataset.

Data format.

(1) PDB: pdb code of the crystal structure of the protein-protein complex
(2) Source dataset: refer to the sample source of the original datasets
(3) Model: refer to which model in the PDB that the target complex belongs to, the default is 0
(4) Mutations: refer to the mutations of ligand proteins and receptor proteins and the format is [PDB code]_[reference animo acid][mutation site][alteration animo acid]
(5) Ligand Chains: refer to the ligand protein chains
(6) Recaptor Chains: refer to the receptor protein chains
(7) Ligand Name: refer to the receptor protein name
(8) Receptor Name: refer to the receptor protein name
(9) KD(M): the KD value of affinity and the unit is M
(10) ΔG(kcal/mol): The change difference of the Gibbs free energy when the receptor and the ligand binds together, which is transformed from the KD value (Environment temperature is 25°C, and the unit is kcal/mol).
11) Affinity Method: measurement methods of affinity
12) Structure Method: measurement methods of crystal structures
13) Temperature(K): experiment temperature when measuring the affinity, and the unit is K
14) Resolution(Å): resolution of the crystal structure, and the unit is Å
15) PDB PubMed ID: PubMed ID of the literature that disclose the crystal structure
16) PDB Release Date: release date of the crystal structure
17) Affinity PubMed ID: PubMed ID of the literature that disclose the affinity of the proteins

Metadata. We identified a total of 3032 unique PDB codes and the intersection situation among different source datasets was presented with Venn diagrams (Fig. 2A). We also identified a total of 12062 unique samples, which referred to unique PDB, ligand chains, receptor chains, and mutations (Fig. 2B). We used histograms to present the PDB resolution (Fig. 2C) and the distribution of sample affinity (Fig. 2D) of the whole dataset. We summarized the proportion of different structural determination methods for all samples in the dataset and the distribution of structural resolution by different structural determination methods (Fig. 3A). Of the 306 samples with crystal structures determined using solution NMR, only two of them had recorded resolution values. The average resolution of the structures determined by X-RAY diffraction was slightly higher than that of electron microscopy. Additionally, the number of different affinity determination methods for each sample was also presented. As Fig. 3B presented, surface plasmon resonance (SPR) was the main method (43.6%) of determining affinity in samples with known affinity measurement methods. We also summarized the distribution of mutant patterns for the whole dataset. As Fig. 3C presented, 54.6% of PDBs associated with single samples (wild-type), 24.3% associated with double samples (1 wild-type and 1 mutant-type), 16.4% associated with multiple samples (1 wild-type and 1–9 mutant-type), and 4.7% associated with large number of samples (1 wild-type and more than

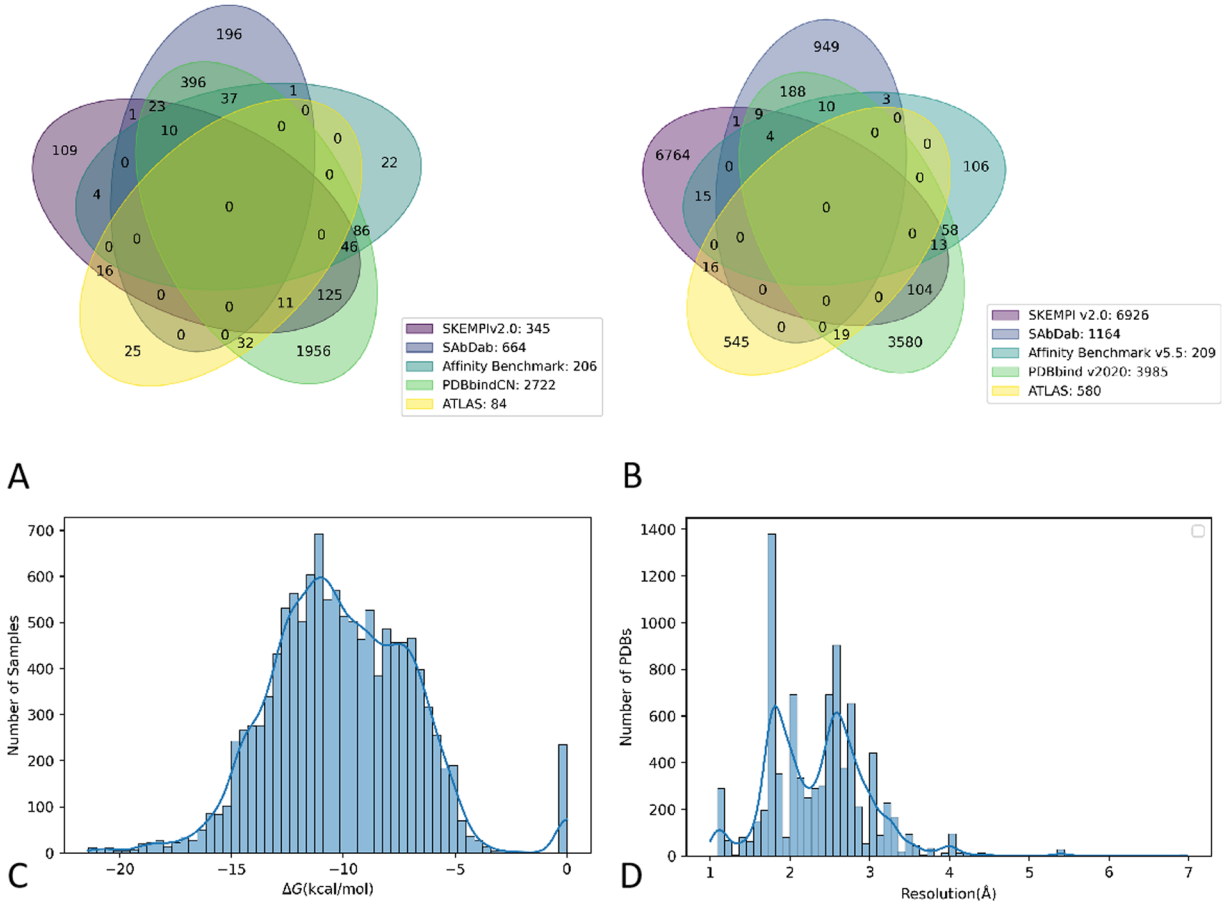

Fig. 2 Distribution summary of the whole dataset. (**A**) unique PDB codes distributed among different source datasets; (**B**) unique samples distributed among different source datasets; (**C**) sample affinity of the whole dataset; (**D**) PDB resolution of the whole dataset.

10 mutant-type). As shown in Fig. 3D,E, the affinity of TCR-pMHC binding was primarily concentrated between −8 to −5.5 kcal/mol, which was weaker than the affinities observed in antigen-antibody interactions and other common protein interactions. This phenomenon could be explained by the monitor and regulate role of TCRs in the immune system. As illustrated in Fig. 3F,G, both the antibody-antigen subgroup and the TCR-pMHC subgroup contained a higher proportion of mutant samples compared to the entire dataset.

Data diversity. In the analysis of protein-protein interactions, sequence alignment is often performed on complexes to determine homology/similarity. Homologous proteins are defined as those with a similarity score greater than 50% and at least 30% sequence identity. Given that affinity is highly correlated with the three-dimensional structure of the binding interface, we adopted a novel method called IDist[9] to present the similarity network of binding interface of all included complex. In this method, binding sites are defined as locations where the Euclidean distance between any CA atom in the receptor protein and any CA atom in the ligand protein is less than 10 Å. The SE(3) feature vectors of these binding sites are then calculated using the IDist algorithm and then the representation of the binding interfaces were obtained by integrating SE(3) features of all binding spots. If the Euclidean distance between the representations of any two binding interfaces is less than 0.05, these interfaces are defined as similar, and in the graph, this similarity is represented by connecting the corresponding nodes.

In Fig. 4A, each node represents a complex, with blue nodes representing antibody-antigen complexes, red nodes representing TCR-pMHC complexes, and gray nodes representing other proteins. Edges indicate that the two connected nodes have similar protein-protein binding interfaces. It was noted that the TCR-pMHC subgroup exhibits a high degree of clustering. This is likely due to the high similarity in the binding interfaces of diverse TCR-pMHC complexes. On the other hand, antibody-antigen complexes also showed sort of clustering, but the degree of aggregation was significantly lower than that of the TCR-pMHC subgroup. This could be explained by the fact that although the variable region of antibodies had a certain degree of conserved structure, its diversity was higher than that of TCRs. More importantly, antigens exhibit a wide range of sequences and structures, leading to a higher diversity in the binding interfaces of antibody-antigen complexes.

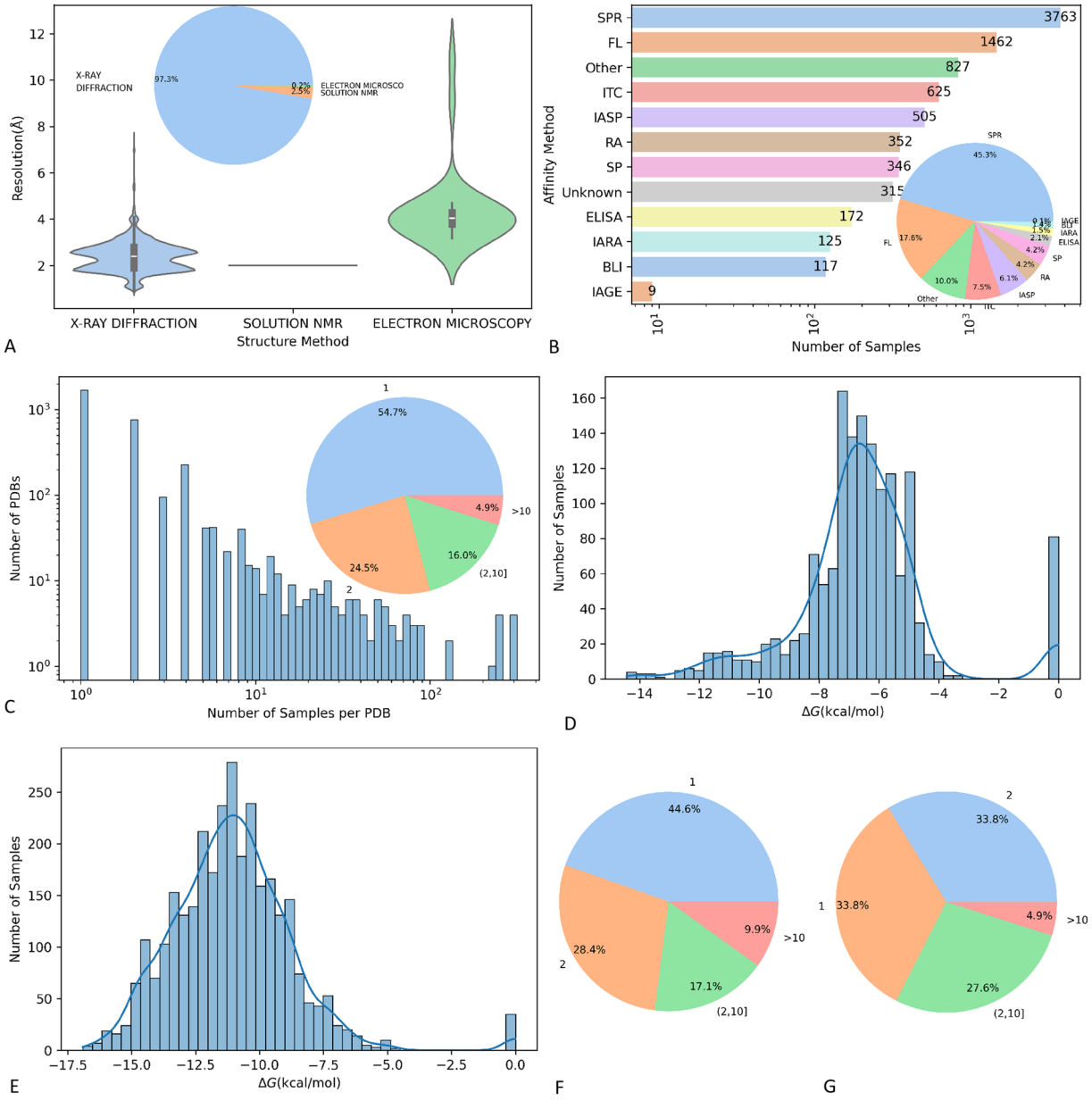

Fig. 3 Meta-data information of the entire dataset and subgroup distribution. (**A**) Summary of structural determination methods; (**B**) Summary of affinity determination methods.; (**C**) distribution of mutant patterns for the entire dataset. (**D**) affinity distribution of TCR-pMHC subgroup; (**E**) affinity distribution of antibody-antigen subgroup; (**F**) mutant patterns of TCR-pMHC subgroup; (**G**) mutant patterns of antibody-antigen subgroup.

Technical Validation

A benchmark affinity prediction. We proposed a benchmark algorithm based on geometric deep learning method (Fig. 4B). The core idea is to use the IPA (invariant point attention)[42] method to extract features from the crystal structure of protein-protein complexes to predict the affinity magnitude. Firstly, 128 amino acid residues at the ligand-acceptor binding interface and its immediate neighbors were intercepted. The amino acid residue sequence was input to the residue encoder for residue encoding. The extractor was designed to extract features from both individual amino acids and pairs of amino acids. Specifically, the type, position, dihedral angle and other information of amino acids serve as note features, while the type, relative position, distance, virtual dihedral angle and other information of amino acid pairs constitute edge features. These features were collectively referred to as residue feature X, since some amino acids in X may undergo mutated. To account for these mutations, a mutation mask was introduced to indicate any animo acid alterations. This mask was then embedded into a vector, denoted asβ, and incorporated to X. Subsequently, the residue feature X and its 3-dimensional spatial coordinates were fed into the spatially invariant point attention module. This module was expected to extract

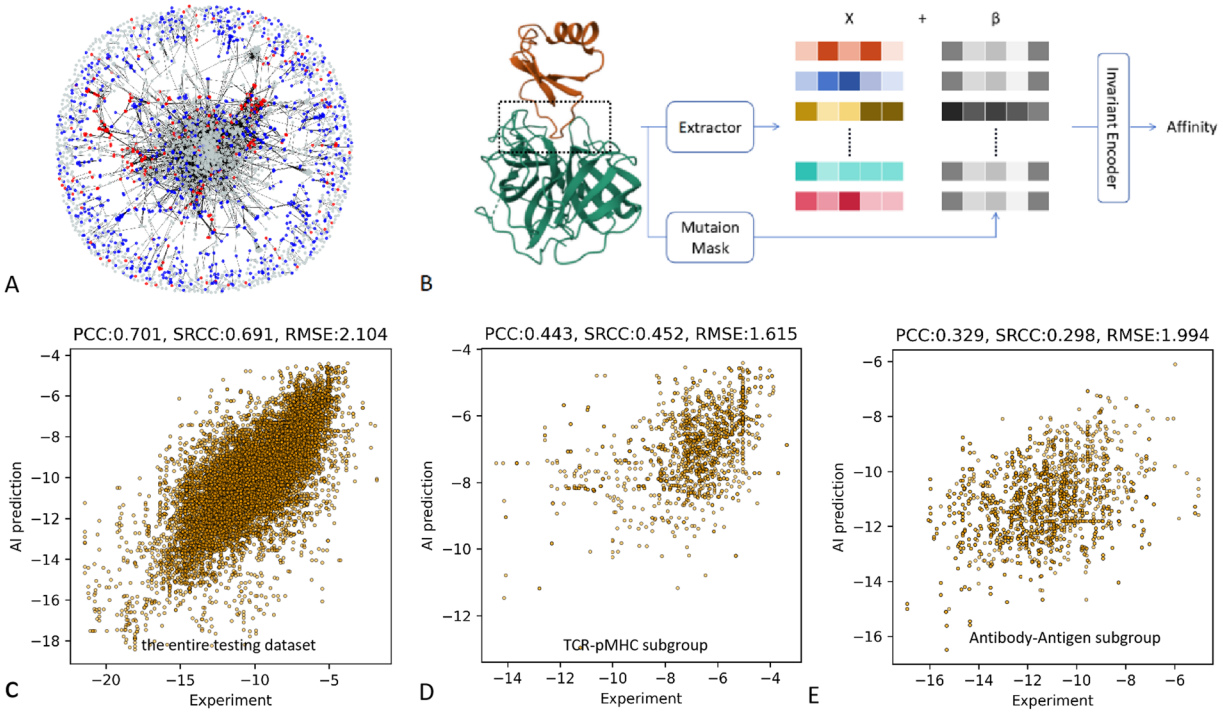

Fig. 4 Visualization of data diversity and benchmark affinity prediction. (**A**) similarity network of binding interface, while blue nodes representing antibody-antigen complexes, red nodes representing TCR-pMHC complexes, and gray nodes representing other proteins; (**B**) design of benchmark algorithm for affinity prediction; (**C**) performance of the benchmark model for the entire testing dataset; (**D**) performance of the benchmark model for the TCR-pMHC subgroup; (**E**) performance of the benchmark model for the antigen-antibody subgroup.

the rotation and translational invariant spatial features of the residues. Finally, the spatial characteristics of the residues were integrated to predict the affinity magnitude through the prediction head.

Whole dataset. As shown in Fig. 4C, the Benchmark algorithm achieved a Pearson correlation coefficient of 0.701 and a Spearman correlation coefficient of 0.691 on this dataset. However, there were some weak horizontal lines visible in the chart. This was because the benchmark model might have difficulty capturing the differences between certain mutants and their wild types, hence it predicted similar or identical affinity values. The main reasons are: 1) the dataset does not provide the crystal structures of the mutant-type samples, and 2) the benchmark algorithm itself is weak to capture the characteristics of the mutant-type crystal structures based on the wild-type crystal structures and their mutation information.

Antibody-antigen subgroup. In the antibody-antigen subgroup, the correlation between the predicted values by the benchmark model and the actual values was significantly lower than that of the entire dataset (Fig. 4D). The main reasons for this discrepancy are likely as follows: 1) One of the reasons for this discrepancy may be due to the high structural similarity of antibodies, making it difficult for current technologies or algorithms to distinguish such structurally similar proteins; 2) Another reason could be the diversity and uncertainty of antigens and antigenic epitopes; 3) compared to the overall dataset, there is a higher proportion of mutant samples in the antibody-antigen subgroup, whose complex crystal structures are unknown. The benchmark algorithm lacks the capability to capture the characteristics of mutated structures.

TCR-pMHC subgroup. Similarly, the correlation between the predicted values and the actual values for the TCR-pMHC subgroup by the benchmark model was lower than that of the overall dataset, and slightly higher than that of the antibody-antigen subgroup (Fig. 4E). The reasons may be attributed to the fact that not only the structures of TCR is highly similar, but also the peptide-MHC complex looks alike.

Source datasets. We also demonstrated the cross-dataset performance of the benchmark affinity prediction. It seemed that the testing samples from the Affinity Benchmark v5.5 dataset, the PDBBind v2020 dataset and the SKEMPI v2.0 dataset showed higher accuracy than the other two source datasets (Fig. 5A–E). Additionally, we trained the same algorithm with various percentage of the entire dataset to demonstrate the advantages of the current dataset. Results demonstrated that same algorithm trained with more and more training data did showed improved prediction performance (Fig. 5F).

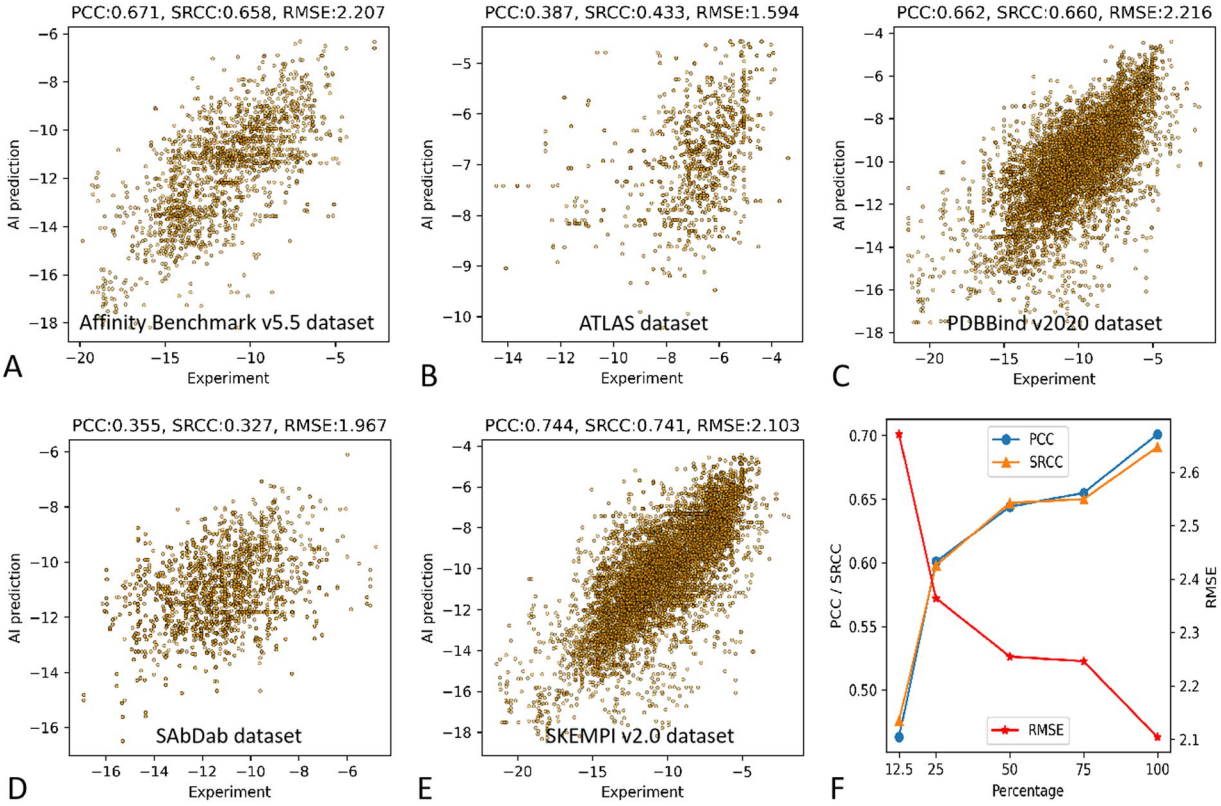

Fig. 5 Benchmark affinity prediction for the source datasets; (**A**) performance of the benchmark model for the Affinity Benchmark v5.5 dataset; (**B**) performance of the benchmark model for the ATLAS dataset; (**C**) performance of the benchmark model for the PDBBind v2020 dataset; (**D**) performance of the benchmark model for SAbDab dataset; (**E**) performance of the benchmark model for the SKEMPI v2.0 dataset; (**F**) improved prediction performance with more training data.

PPB-Affinity-AF subdataset. As the complex structure were not experimentally measured, we demonstrated distribution of the interface predicted template modeling (ipTM) of the complex from the PPB-Affinity-AF subdataset (Fig. 6A), which indicated the prediction reliability of the complex. Although the affinity prediction of the entire PPB-Affinity-AF subdataset seemed to be unsatisfied (Fig. 6B), the affinity prediction accuracy got further improved when the ipTM of samples were from 0.5 to 0.75 (Fig. 6C,D), which further indicate the reliability of the benchmark algorithm and the importance of complex structures.

Usage Notes

Instructions for use supplement, including:

The dataset can be publicly accessed when you agree to cite the DOI of the collection for your publication resulting from this dataset.

Potential uses of the dataset. We hope that this dataset will help researchers, and AI algorithm engineers in the field of protein drug discovery to develop more powerful protein-protein binding affinity prediction algorithms.

Limitations of datasets. First, all samples are currently from other publicly available datasets. Thus, new research papers may sporadically publish experimentally measured protein-protein binding affinity with their crystal structures, but these data might not be promptly collected into the PPB-Affinity dataset. Second, it should be noted that many mutant protein complexes did not have their experimentally measured crystal structures, where only the crystal structures of their corresponding wild-type complexes are available. Third, we only provided the benchmark prediction performance with one mainstream algorithm. As a dataset article, we hope this dataset can attract more researchers to develop more creative algorithms to promote drug discovery in this field.

The meaning of the dataset and the future work of our team: The unique features of our dataset could be summarized as follows: (1) all of the samples have unified affinity indicator, recorded experimental condition and annotated ligand chain and receptor chain, which are not always available for all the source datasets; (2) we also have a sub-dataset called "PPB-Affinity-AF", which were collected from most recent studies publishing the experimentally measured affinity, and their three-dimensional structure of the complex were created with AphaFold3. The future work will include and develop more experimental data and crystal structure of PPB affinity.

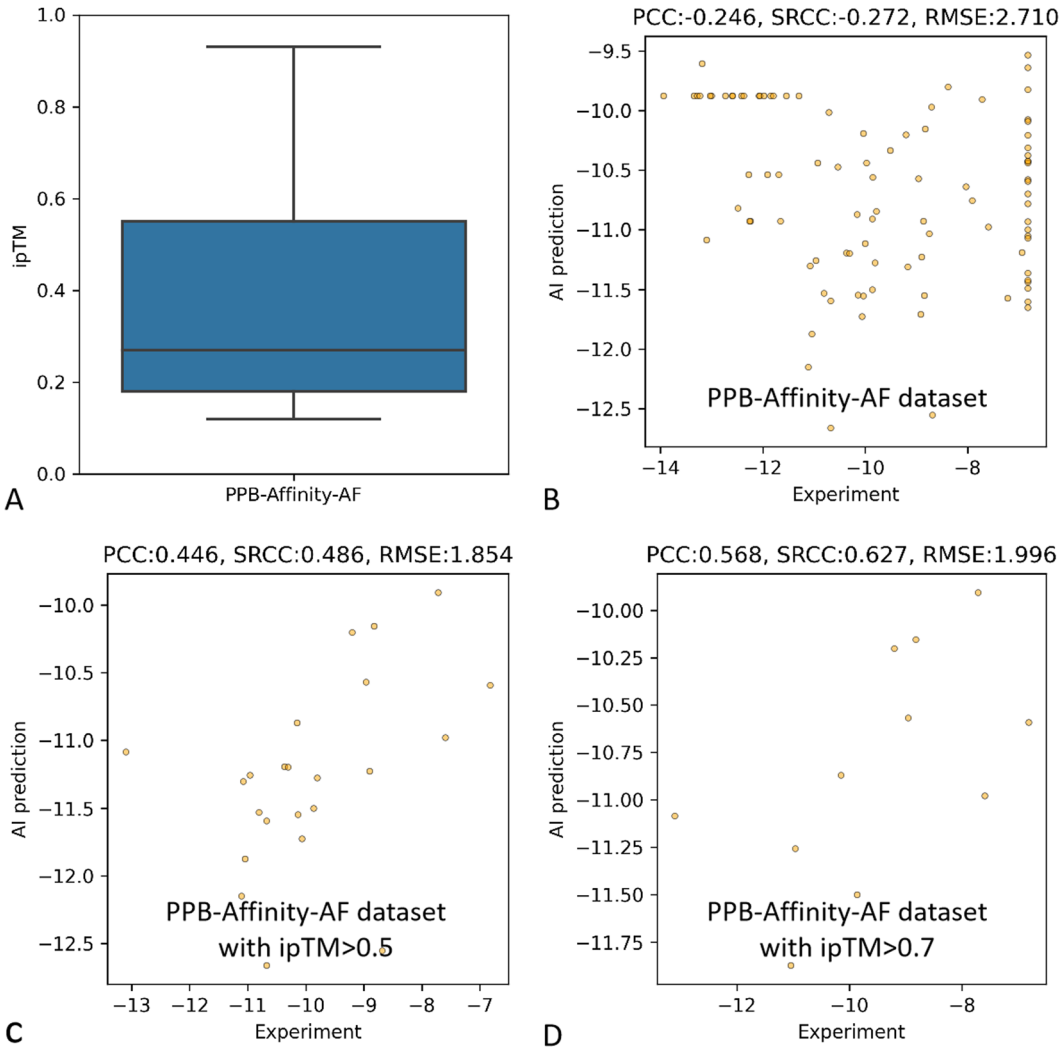

Fig. 6 Benchmark affinity prediction for the PPB-Affinity-AF subdataset. (**A**) distribution of the interface predicted template modeling (ipTM) of the complex from the PPB-Affinity-AF subdataset; (**B**) affinity prediction results of all samples from the PPB-Affinity-AF subdataset; (**C**) affinity prediction results of samples with ipTM over 0.5; (**D**) affinity prediction results of samples with ipTM over 0.75.

Code availability

Codes for PPB-Affinity database preparation is disclosed at https://github.com/Huatsing-Lau/PPB-Affinity-DataPrepWorkflow. Codes for benchmark algorithm is disclosed at https://github.com/ChenPy00/PPB-Affinity.

Received: 24 May 2024; Accepted: 11 October 2024;
Published online: 03 December 2024

References

1. Hummer, A. M., Schneider, C., Chinery, L. & Deane, C. M. Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen ΔΔG Prediction. *bioRxiv*, 2023.2005. 2017.541222 (2023).
2. Mohseni Behbahani, Y., Laine, E. & Carbone, A. Deep Local Analysis deconstructs protein–protein interfaces and accurately estimates binding affinity changes upon mutation. *Bioinformatics* **39**, i544–i552 (2023).
3. Luo, S. *et al*. Rotamer Density Estimator is an Unsupervised Learner of the Effect of Mutations on Protein-Protein Interaction. *bioRxiv*, 2023.2002. 2028.530137 (2023).
4. Shan, S. *et al*. Deep learning guided optimization of human antibody against SARS-CoV-2 variants with broad neutralization. *Proc Natl Acad Sci USA* **119**, e2122954119, https://doi.org/10.1073/pnas.2122954119 (2022).
5. Liu, X., Luo, Y., Li, P., Song, S. & Peng, J. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS Comput Biol* **17**, e1009284, https://doi.org/10.1371/journal.pcbi.1009284 (2021).
6. Myung, Y., Rodrigues, C. H. M., Ascher, D. B. & Pires, D. E. V. mCSM-AB2: guiding rational antibody design using graph-based signatures. *Bioinformatics* **36**, 1453–1459, https://doi.org/10.1093/bioinformatics/btz779 (2020).
7. Myung, Y., Pires, D. E. V. & Ascher, D. B. mmCSM-AB: guiding rational antibody engineering through multiple point mutations. *Nucleic Acids Res* **48**, W125–w131, https://doi.org/10.1093/nar/gkaa389 (2020).
8. Wang, M., Cang, Z. & Wei, G.-W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nature Machine Intelligence* **2**, 116–123 (2020).

9. Bushuiev, A. *et al*. Learning to design protein-protein interactions with enhanced generalization. *arXiv preprint arXiv:2310.18515 v3* (2023).
10. Jankauskaite, J., Jiménez-García, B., Dapkunas, J., Fernández-Recio, J. & Moal, I. H. SKEMPI 2.0: an updated benchmark of changes in protein-protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **35**, 462–469, https://doi.org/10.1093/bioinformatics/bty635 (2019).
11. Sirin, S., Apgar, J. R., Bennett, E. M. & Keating, A. E. AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Sci* **25**, 393–409, https://doi.org/10.1002/pro.2829 (2016).
12. Starr, T. N. *et al*. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424, https://doi.org/10.1126/science.abo7896 (2022).
13. Starr, T. N. *et al*. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295–1310.e1220, https://doi.org/10.1016/j.cell.2020.08.012 (2020).
14. Chan, K. K. *et al*. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265, https://doi.org/10.1126/science.abc0870 (2020).
15. Cao, Y. *et al*. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663, https://doi.org/10.1038/s41586-021-04385-3 (2022).
16. Kowalsky, C. A. & Whitehead, T. A. Determination of binding affinity upon mutation for type I dockerin-cohesin complexes from Clostridium thermocellum and Clostridium cellulolyticum using deep sequencing. *Proteins* **84**, 1914–1928, https://doi.org/10.1002/prot.25175 (2016).
17. Liu, L. *et al*. Striking antibody evasion manifested by the Omicron variant of SARS-CoV-2. *Nature* **602**, 676–681, https://doi.org/10.1038/s41586-021-04388-0 (2022).
18. Wang, R. *et al*. Analysis of SARS-CoV-2 variant mutations reveals neutralization escape mechanisms and the ability to use ACE2 receptors from additional species. *Immunity* **54**, 1611–1621.e1615, https://doi.org/10.1016/j.immuni.2021.06.003 (2021).
19. Myung, Y., Pires, D. E. V. & Ascher, D. B. CSM-AB: graph-based antibody-antigen binding affinity prediction and docking scoring function. *Bioinformatics* **38**, 1141–1143, https://doi.org/10.1093/bioinformatics/btab762 (2022).
20. Lei, Y. *et al*. A deep-learning framework for multi-level peptide-protein interaction prediction. *Nat Commun* **12**, 5465, https://doi.org/10.1038/s41467-021-25772-4 (2021).
21. Yang, Y. X., Wang, P. & Zhu, B. T. Binding affinity prediction for antibody-protein antigen complexes: A machine learning analysis based on interface and surface areas. *J Mol Graph Model* **118**, 108364, https://doi.org/10.1016/j.jmgm.2022.108364 (2023).
22. Romero-Molina, S. *et al*. PPI-affinity: A web tool for the prediction and optimization of protein–peptide and protein–protein binding affinity. *Journal of Proteome Research* **21**, 1829–1841 (2022).
23. Yuan, Y., Chen, Q., Mao, J., Li, G. & Pan, X. DG-Affinity: predicting antigen-antibody affinity with language models from sequences. *BMC Bioinformatics* **24**, 430, https://doi.org/10.1186/s12859-023-05562-z (2023).
24. Yang, Y. X., Huang, J. Y., Wang, P. & Zhu, B. T. AREA-AFFINITY: A Web Server for Machine Learning-Based Prediction of Protein-Protein and Antibody-Protein Antigen Binding Affinities. *J Chem Inf Model* **63**, 3230–3237, https://doi.org/10.1021/acs.jcim.2c01499 (2023).
25. Guest, J. D. *et al*. An expanded benchmark for antibody-antigen docking and affinity prediction reveals insights into antibody recognition determinants. *Structure* **29**, 606–621.e605, https://doi.org/10.1016/j.str.2021.01.005 (2021).
26. Kastritis, P. L. *et al*. A structure-based benchmark for protein–protein binding affinity. *Protein Science* **20**, 482–491 (2011).
27. Schneider, C., Raybould, M. I. J. & Deane, C. M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res* **50**, D1368–d1372, https://doi.org/10.1093/nar/gkab1050 (2022).
28. Dunbar, J. *et al*. SAbDab: the structural antibody database. *Nucleic Acids Research* **42**, D1140–D1146, https://doi.org/10.1093/nar/gkt1043 (2013).
29. Raybould, M. I. J. *et al*. Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic Acids Research* **48**, D383–D388, https://doi.org/10.1093/nar/gkz827 (2019).
30. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry* **47**, 2977–2980, https://doi.org/10.1021/jm030580l (2004).
31. Liu, Z. *et al*. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Accounts of chemical research* **50**, 302–309, https://doi.org/10.1021/acs.accounts.6b00491 (2017).
32. Liu, Z. *et al*. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics (Oxford, England)* **31**, 405–412, https://doi.org/10.1093/bioinformatics/btu626 (2015).
33. Li, Y. *et al*. Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *Journal of chemical information and modeling* **54**, 1700–1716, https://doi.org/10.1021/ci500080q (2014).
34. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling* **49**, 1079–1093, https://doi.org/10.1021/ci9000053 (2009).
35. Wang, R., Fang, X., Lu, Y., Yang, C. Y. & Wang, S. The PDBbind database: methodologies and updates. *Journal of medicinal chemistry* **48**, 4111–4119, https://doi.org/10.1021/jm048957q (2005).
36. Vreven, T. *et al*. Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* **427**, 3031–3041, https://doi.org/10.1016/j.jmb.2015.07.016 (2015).
37. Borrman, T. *et al*. ATLAS: A database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes. *Proteins* **85**, 908–916, https://doi.org/10.1002/prot.25260 (2017).
38. Adler, A. S. *et al*. Rare, high-affinity mouse anti-PD-1 antibodies that function in checkpoint blockade, discovered using microfluidics and molecular genomics. *mAbs* **9**, 1270–1281, https://doi.org/10.1080/19420862.2017.1371386 (2017).
39. Kang-Pettinger, T. *et al*. Identification, binding, and structural characterization of single domain anti-PD-L1 antibodies inhibitory of immune regulatory proteins PD-1 and CD80. *The Journal of biological chemistry* **299**, 102769, https://doi.org/10.1016/j.jbc.2022.102769 (2023).
40. Porebski, B. T. *et al*. Rapid discovery of high-affinity antibodies via massively parallel sequencing, ribosome display and affinity screening. *Nature biomedical engineering* **8**, 214–232, https://doi.org/10.1038/s41551-023-01093-3 (2024).
41. Liu, H. *et al*. PPB-Affinity dataset. *Zenodo*. https://doi.org/10.5281/zenodo.13067409 (2024).
42. Jumper, J. *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589, https://doi.org/10.1038/s41586-021-03819-2 (2021).

Acknowledgements

We thank Prof. Xiang Liao and Dr. Chaobo Feng in Shenzhen Nanshan People's Hospital for suggestions of context layout and figure editing. We acknowledged funding from the National Natural Science Foundation of China (82102640), Medical Scientific Research Foundation of Guangdong Province of China (A2023195), Nanshan District Health Science and Technology Major Project (NSZD2023023) and Nanshan District Health Science and Technology Project (NS2023044) for Guoxin Fan.

Author contributions

Huaqing Liu contributed to the study conception, manuscript drafting, algorithm development and data creation. Peiyi Chen contributed to the manuscript drafting, algorithm development, data creation, and data analysis. Xiaochen Zhai contributed to the manuscript revision, material preparation, and data analysis. Ku-Geng Huo contributed to the manuscript revision, and data analysis. Shuxian Zhou contributed to the data creation. Guoxin Fan contributed to the study conception, study design and manuscript writing. Lanqing Han contributed to the study conception, manuscript revision and project supervision.

Competing interests

We disclosed no conflicts of interest that might be perceived to influence the results and/or discussion reported in this paper.

Additional information

Correspondence and requests for materials should be addressed to L.H. or G.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.