Nucleic Acids Research, 2023, 1–7
https://doi.org/10.1093/nar/gkad1056
Database issue

The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures

Brennan Abanades [1,†], Tobias H. Olsen [1,†], Matthew I.J. Raybould [1], Broncio Aguilar-Sanjuan [1], Wing Ki Wong [2], Guy Georges [2], Alexander Bujotzek [2] and Charlotte M. Deane [1,*]

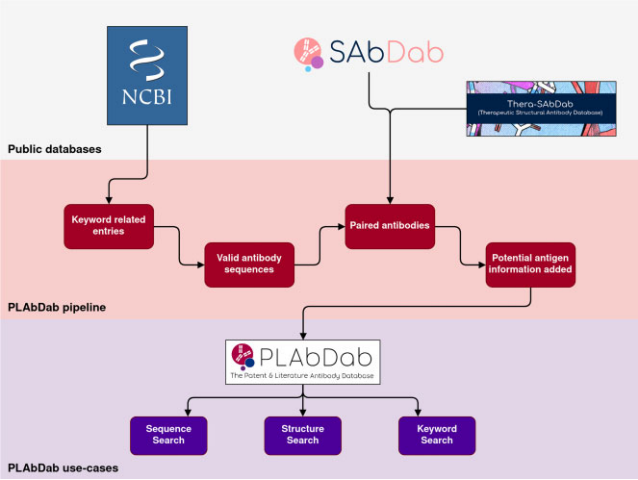
[1]Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX1 3LB, UK
[2]Large Molecule Research, Roche Pharma Research and Early Development, Roche Innovation Center Munich, DE-82377 Penzberg, Germany

[*]To whom correspondence should be addressed. Tel: +44 1865 272860; Email: deane@stats.ox.ac.uk
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Antibodies are key proteins of the adaptive immune system, and there exists a large body of academic literature and patents dedicated to their study and concomitant conversion into therapeutics, diagnostics, or reagents. These documents often contain extensive functional characterisations of the sets of antibodies they describe. However, leveraging these heterogeneous reports, for example to offer insights into the properties of query antibodies of interest, is currently challenging as there is no central repository through which this wide corpus can be mined by sequence or structure. Here, we present PLAbDab (the Patent and Literature Antibody Database), a self-updating repository containing over 150,000 paired antibody sequences and 3D structural models, of which over 65 000 are unique. We describe the methods used to extract, filter, pair, and model the antibodies in PLAbDab, and showcase how PLAbDab can be searched by sequence, structure, or keyword. PLAbDab uses include annotating query antibodies with potential antigen information from similar entries, analysing structural models of existing antibodies to identify modifications that could improve their properties, and facilitating the compilation of bespoke datasets of antibody sequences/structures that bind to a specific antigen. PLAbDab is freely available via Github (https://github.com/oxpig/PLAbDab) and as a searchable webserver (https://opig.stats.ox.ac.uk/webapps/plabdab/).

Graphical abstract



Introduction

Antibodies are by far the most successful type of biotherapeutic, with over 100 approved by the FDA and many more in the advanced stages of clinical development (1,2). Their high specificity and affinity also make them a valuable tool in many areas of medical and scientific research. For example, antibodies are used routinely in diagnostic assays (3), and to better understand the effects of vaccination on the immune system (4).

The variable region in antibodies (Fv) that is responsible for antigen binding, has a conserved global structure. It is formed by two immunoglobulin domains, the heavy (VH) and light (VL) chain variable domains. The binding site is divided between both chains, and is concentrated in six hypervariable

Received: July 26, 2023. **Revised:** October 20, 2023. **Editorial Decision:** October 20, 2023. **Accepted:** October 30, 2023

loops, three on each chain, known as the complementarity-determining regions (CDRs). Among the CDR loops, the third CDR of the heavy chain (CDR-H3) is the most diverse in sequence and structure and often the primary contributor to antigen binding (5). However, the other five CDR loops and the relative orientation of the heavy and light chain have also been shown to affect binding (6).

Next generation sequencing (NGS) has enabled researchers to take snapshots of the immune repertoire of an individual at a given point in time, leading to the generation of vast amounts of single chain antibody sequences. Efforts to compile this data has led to the creation of datasets such as Unpaired OAS (7) and iReceptor (8) which contain the sequence of the heavy or light variable domains for billions of antibodies. Paired VH–VL sequence data is more expensive to generate and currently just over a million paired antibody sequences can be found in Paired OAS (9). However, with the binding site in antibodies sitting across both chains, paired data gives a more complete picture of how and to what an antibody binds (10,11).

Although NGS data has proven invaluable to compare repertoires between individuals, it provides little information on the functions of individual sequences within a repertoire. However, there also exists a large number of smaller scale studies, each one dedicated to investigating a small number of antibodies. When combined, the antibodies from these studies amount to a large number of sequences with rich metadata. There are a number of databases that aim to compile subsets of this data, for example SAbDab (12,13) for antibodies with resolved crystal structures, Thera-SAbDab for antibody therapeutics (2), CoV-AbDab for COVID-19 binding antibodies (14) or PAD for unpaired antibody sequences from patents (15). Paired antibody sequences with information on their epitope can also be obtained from IEDB (16).

Here we present PLAbDab, a database containing 150 000 paired antibody sequences from over 10 000 small scale studies. PLAbDab is larger than any other non-NGS database of paired antibody sequences by at least an order of magnitude. We make the data freely available and provide methods to rapidly search it by either sequence identity using KA-search (17), structural similarity (18,19), or by keywords in the title of the study. Each of the sequences comes with a direct link to its source material, making it easy to obtain additional information about any antibody of interest.

Materials and methods

Collecting unpaired antibody sequences

The majority of data in PLAbDab is extracted from the Protein database of NCBI (20). The BioPython Entrez module (21) is used to query the database for entries containing the words 'antibody' , 'antibodies', 'immunoglobulin', 'scfv' or 'bcr' in any of their fields. Due to the lack of an automated method to accurately differentiate humanised nanobodies from human antibody heavy chains, entries containing the words nanobody or nanobodies are removed at this stage. Entries with sequences longer than 1000 amino acids or shorter than 70 are also removed. Around 2.5 million entries were returned from this search.

Sequences for each of these 2.5 million entries were then searched for antibody variable domain sequences using ANARCI (22). Sequences missing residues in the CDR regions were removed at this stage. This resulted in around 530 000 potential antibody variable domain sequences from around 13 000 different sources.

Creating paired antibody sequences from the unpaired data

For a large number of entries the metadata provides enough information to pair the heavy and light variable domains. For entries from the same literature source, VH-VL pairing was attempted using the following heuristics:

(1) 'Same entry': If there is only one VH and only one VL within a single entry, these are paired together [20.5% of entries].
(2) 'Unique word': If there is a unique non-common word that is found in the description of exactly one VH and one VL, these are paired together. Uncommon words are defined as words found in the description of less than 20 entries in the unpaired database [0.8% of entries].
(3) 'Unique source': Some entries contain information on the experimental source, such as the isolate or clone. If there is a unique VH and VL from the same experimental source they are paired together [12.8% of entries].
(4) 'Patent text': Entries from patents have sequence IDs by which they are referred to in the patent text. If the patent text mentions a VH and VL within the same paragraph, these are paired together. If a paragraph in the patent text mentions the same number of VH and VL entries, these are paired in the same order as they are mentioned [50.5% of entries].
(5) 'Unique chain': If there is a single VH (or VL) from one source, this entry is paired with all other VL (or VH) entries from that source [2.5% of entries].
(6) 'Ordered entries': If there are the same number of VH and VL entries from one source, they are paired in order [4.9% of entries].

The described strategies for pairing VH–VL sequences have varying levels of accuracy. Twenty entries paired by each of the methods described above were randomly selected and manually checked to estimate the accuracy of each pairing method. Methods 1–4 achieved perfect accuracy on the twenty entry test set. We therefore refer to these methods as pairing with 'high confidence'. Method 5 incorrectly paired one entry and Method 6 incorrectly paired two entries. Each paired entry is given a flag to indicate how it was paired, making it easy for users to filter for less accurate pairing methods. To further increase the coverage of the dataset, sequences from both SAbDab [7.4% of entries] and Thera-SAbDab [0.6% of entries] were also added to PLAbDab.

Labelling antibodies with potential antigen information

The target of the majority of antibodies in PLAbDab can be inferred from a manual review of their source literature. Although possible, this process can be time-consuming. To reduce the number of entries to review, we provide a 'targets_mentioned' column that lists common antigens mentioned in the source material.

For patent entries, potential antigens were identified by searching the title, abstract, and claims sections for mentions of commonly recognised antigen names. When a term in the source literature was prefixed with 'anti-', it was also included

in the list of possible antigens. For non-patent entries, only the title was searched for potential targets.

The use of the 'targets_mentioned' label exhibits high recall: when an antibody's antigen is known, it is often mentioned in the analysed text. However, its accuracy is moderate. Accurate labelling of each entry with antigen information requires manual inspection. Nonetheless, this approach offers a fast and straightforward method for generating an initial reference set.

Searching PLAbDab

To allow users to rapidly search the database for antibodies with a similar sequence, we implemented KA-Search (17). KA-search is able to carry out rapid and exhaustive sequence identity searches, allowing users to find similar sequences over the whole variable domain, the CDR loops, the CDR-H3 or a user defined region. The entire database can be queried with KA-Search in under 5 s on 5 CPUs.

Structurally similar antibodies can have similar functions, even if they are distantly related in sequence (18,19). To enable CDR structure-based searching, we model all paired sequences in the database using ABodyBuilder2 (23). Entries missing more than eight residues at the start of the sequence were restored using AbLang (24) prior to modelling. Antibody sequences with non-standard residues, or for which ABodyBuilder2 was not capable of generating a refined model, are labelled as 'FAILED' under the 'Model' column. Entries with the same sequence were only modelled once, in total 64,000 unique antibody models were generated.

To structurally search PLAbDab, query sequences are first modelled using ABodyBuilder2, without refinement. The refinement step was skipped as in order to boost speed with minimal compromise on backbone model quality. The framework of the predicted structure is then aligned to the structure of all other entries in PLAbDab with the same CDR loop lengths. Lastly, the carbon-alpha (C_α) root-mean squared deviation (RMSD) over all CDR residues is computed and used to rank entries. Searching the database in this way takes around 10 s on 5 CPUs.

Finally, as each sequence in PLAbDab is linked to a patent or publication, we provide users the ability to search fields, such as the title of the study, using regular expressions.

Results

Database statistics

Currently the total number of entries in PLAbDab is around 150 000, over 90% of which are paired with high confidence using methods 1–4, or comes from therapeutic or crystal structure entries. As can be seen in Figure 1A, the number of antibody sequences that could be collected by the PLAbDab methodology has been steadily growing since the early 2000s, with somewhere between 10 000 and 30 000 new antibody sequences being published each year for the last 5 years.

Figure 1B shows the distribution of PLAbDab entries by source, around three quarters of entries come from antibody sequences described in patents, while less than 20% are derived from the scientific literature. This may be due to patents following a more standardised way of depositing antibody sequences. The rest of the entries in PLAbDab come either from solved structures in SAbDab (12), or from additional sequences in Thera-SAbDab (2).

Patent applicants are not obligated to upload species information about their sequences to NCBI. This means that the majority of entries in PLAbDab lack a species annotation. Unfortunately, computationally predicting the species origin for unannotated patent antibodies in a systematic way is likely to be ineffective due to the extent of non-natural engineering these sequences undergo. However, species information does exist for the majority of the entries sourced from the academic literature. Across the sequences with species information, most entries are labelled as human, with a smaller number of sequences annotated as mouse, macaque or rabbit. (Figure 1C).

It has been observed that therapeutic antibodies tend to have shorter CDR-H3 loops than those seen in large repertoire studies, an observation that may be due to longer CDR loops leading to issues during therapeutic antibody development (25,26). The average CDR-H3 loop length from antibodies in PLAbDab (around ∼14.0) falls somewhere between the average CDR-H3 loop length of OAS (around ∼15.6) and Thera-SAbDab (around ∼12.9) (Figure 1 D).

Searching PLAbDab

As described in the methods, PLAbDab allows users to search the database using a variety of methods. To demonstrate the sequence and structure search options, we compared the results of searching the database in four ways:

- For antibodies with a sequence identity of over 90% over the VH (VH identity).
- For antibodies with a sequence identity of over 90% for both the VH and VL (VH+VL identity).
- For antibodies with a C_α RMSD over the CDR loops of under 1.25Å to an ABodyBuilder2 model of the query (CDR structure).
- For antibodies with a C_α RMSD over the CDR loops of under 1.25Å to an ABodyBuilder2 model of the query and sequence identity over the CDR loops of over 80% (CDR structure+identity).

We performed the above searches for four antibodies each of which target a different antigen. Antibody one binds programmed cell death protein 1 (PD-1), and was taken from the patent 'Anti-PD-1 antibodies and methods of use thereof.'; antibody two binds Respiratory Syncytial Virus (RSV), sourced from the paper 'Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors' (27); antibody three is the therapeutic antibody Samalizumab, which binds the OX2 membrane glycoprotein (CD200); and antibody four is the therapeutic antibody Foralumab, that targets the CD3-epsilon peptide (CD3e). For each of the four queries we took the sequence and built an ABodyBuilder2 model for use in the structure searches. The result of using each search method described above to query PLAbDab with these antibodies is shown in Table 1.

The PD-1 case study shows the benefits of having the paired heavy and light chain variable domain sequence. While a sequence identity search over the VH finds antibodies binding to the same antigen around 25% (61 out of 258) of the time, including the VL improves the accuracy to around 75% (16 out of 22). The fact that there are many highly identical heavy chain sequences that bind to different antigens suggests that, for this antibody, the light chain may contribute significantly to binding. To validate this, we analysed the crystal structure

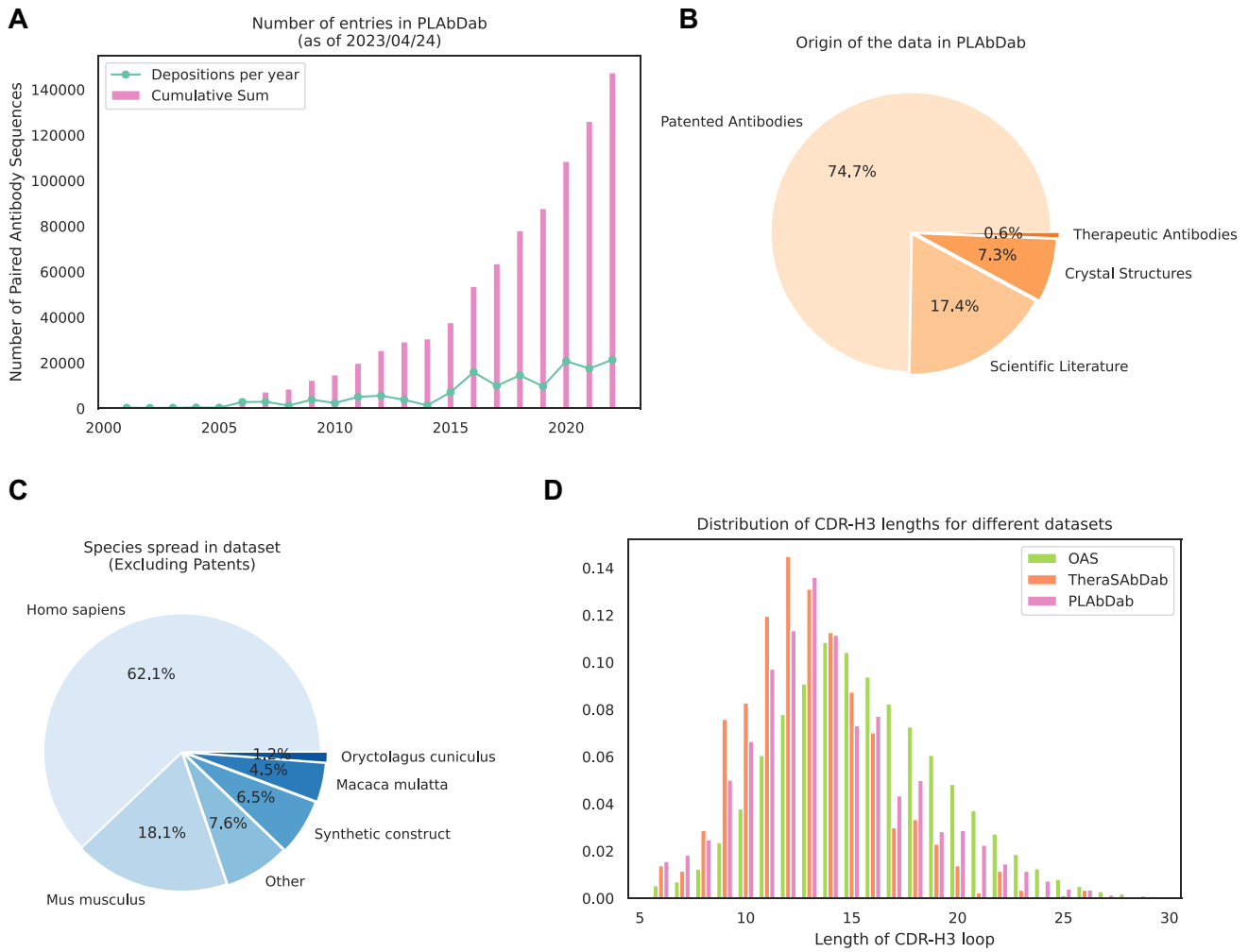


Figure 1. Summary statistics for PLAbDab. (**A**) Number of new antibody sequences each year and total number of antibody sequences that could be collected by the PLAbDab methodology. (**B**) The proportion of data in PLAbDab by type of source. (**C**) The proportion of antibodies from each species in non-patent entries. (**D**) Comparison of the distribution of CDR-H3 lengths in PLAbDab, Thera-SAbDab (2) and OAS (9).

of an antibody bound to RSV which is returned by all four search methods (PDB ID: 5GGR). According to Arpeggio (28), the paratope is evenly split between the heavy and light chain, with eight heavy chain and seven light chain residues found to interact with the antigen.

Searching PLAbDab for entries similar to our PD-1 target by CDR structure returns antibodies that target the same antigen around 40% of the time (19 out of 46), but many of them are from different studies than those found by sequence search. All the entries returned by a CDR structure plus sequence identity search target the same antigen as our query.

For the RSV binding antibody, there are only two entries in PLAbDab with >90% sequence identity over the VH. Searching PLAbDab for entries with a similar CDR structure returns 23 unique antibodies which bind the same epitope. This suggests that this antibody may require a very specific CDR loop conformation to bind the RSV Fusion Glycoprotein (30) at that specific site. Two of the retrieved entries belong to antibodies with resolved crystal structures (PDB IDs: 7LUC and 7LUD), one of which is bound to the antigen. Figure 2 shows the similarity of these structures to the ABodyBuilder2 predicted structure of the RSV binding antibody query.

Only two unique antibody sequences in PLAbDab have a VH sequence identity of over 90% to Samalizumab. A structure search using Samalizumab as a query results in over a thousand antibodies that bind a very diverse set of epitopes. In Figure 3, we show a selection of antibodies with very similar CDR backbone structures to Samalizumab binding to four very different epitopes on different antigens. This indicates that the backbone structure of the CDR loops is likely not the primary driver of binding specificity for this antibody. However, when adding an 80% sequence identity cutoff over the CDR loops to the structure search, we are left with five unique antibody sequences all of which bind the same antigen.

For the Foralumab case study, only two out of nine sequences with a VH identity of over 90% were found to also bind CD3e. Adding an additional 90% sequence identity filter over the VL reduces the number of non-CD3e binders returned to two. Searching PLAbDab for entries that share a similar CDR structure to Foralumab expands the number of unique sequences retrieved to 34, out of which 14 were found to bind CD3e. The CDR structure plus sequence identity search achieves perfect accuracy, but reduces the number of retrieved sequences to two.

Table 1. Results from searching PLAbDab with four different queries using four methods

Search method	Retrieved (cons.)	Sources (cons.)	Unique (cons.)
PD-1			
VH identity	576 (222)	180 (41)	258 (61)
VH+VL identity	155 (132)	39 (28)	22 (16)
CDR structure	227 (168)	60 (26)	46 (19)
CDR structure+identity	127 (127)	29 (29)	14 (14)
RSV			
VH identity	2 (2)	1 (1)	2 (2)
VH+VL identity	2 (2)	1 (1)	2 (2)
CDR structure	29 (29)	4 (4)	23 (23)
CDR structure+identity	4 (4)	1 (1)	4 (4)
CD200			
VH identity	13 (13)	4 (4)	2 (2)
VH+VL identity	13 (13)	4 (4)	2 (2)
CDR structure	1175 (95)	259 (15)	440 (50)
CDR structure+identity	37 (37)	5 (5)	5 (5)
CD3e			
VH identity	26 (15)	9 (4)	9 (2)
VH+VL identity	18 (15)	6 (4)	4 (2)
CDR structure	94 (51)	40 (15)	34 (14)
CDR structure+identity	15 (15)	4 (4)	2 (2)

For each method, the number of retrieved entries, the number of different sources the entries come from and the number of unique antibodies found is given. The value given in parenthesis is the number of entries or sources that are functionally consistent with the query. Details on the four query antibodies and the four searching methods are given in the main text.

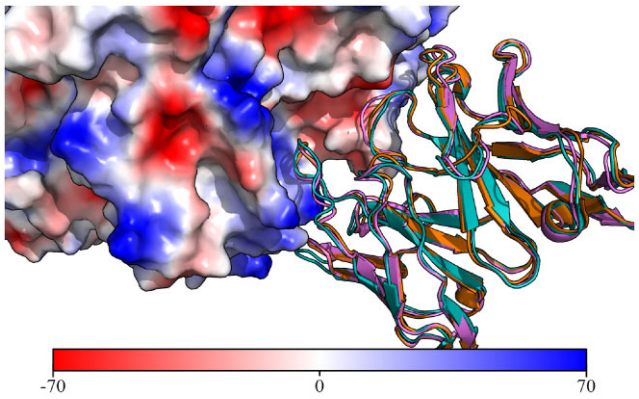


Figure 2. Crystal structures of anti-RSV antibodies aligned to the structural model of the query. A cartoon representation of the ABodyBuilder2 model for the query antibody is shown in purple with the crystal structure for 7LUC shown in blue and the crystal structure for 7LUD shown in orange. The surface of the RSV antigen is coloured based on electrostatics using PyMol (29) (units in $K_b T/e_c$).

These four case studies highlight the utility of PLAbDab's data and various search methods to identify antibodies with similar function to a query. The case studies also highlight that any computational search should be further investigated via both the meta data in PLAbDab and, if possible, experimental means.

Using PLAbDab to generate new datasets

As PLAbDab contains a large number of antibody sequences targeting a diverse set of antigens, it can be used to facilitate the generation of antigen-specific antibody libraries. For example, searching PLAbDab for entries containing the keywords 'ebov' or 'ebola' in their title returns almost 1500 unique antibody sequences from 56 sources. Using the keywords 'hiv' or 'immunodeficiency' finds over 6200 entries with over 3800 unique antibody sequences from over 500 sources, while 'autoimmune' and 'autoantibody' returns over 400 unique antibody sequences from over 90 sources. Not every antibody sequence retrieved from a patent or paper will necessarily bind the searched target, but such prefiltered sets provide a very valuable starting point towards generating an antigen specific library to a target of interest.

To benchmark how effective searching the titles of the sources of PLAbDab entries by keywords is at retrieving relevant antibodies, we manually inspected the results of a search by expression 'hiv|immunodeficiency'. By manually inspecting the source literature from a random set of 100 antibodies from different studies, we found that 88 were true HIV binders. Of the non-HIV binders, four were anti-idiotype antibodies, three bound CD4, three bound other antigens and two were incorrectly paired. Using the search terms 'covid|coronal|sars', we inspected another 100 randomly-selected antibodies from different studies. Of these, 98 targeted a coronavirus, one targeted ACE2, and the specificity of the last could not be verified. This highlights how PLAbDab can aid in the creation of antigen-specific antibody datasets, but also shows that manual inspection of the extracted sequences is still required to remove false positives.

Discussion

We present PLAbDab, a large database of paired antibody sequences from patents and papers. Each entry in the database contains a link to its original source material, which relates the paired sequences to useful functional information. PLAbDab can be automatically updated without any manual input, making it straightforward to keep up to date with the latest publications and patents.

To demonstrate the utility of PLAbDab, we explore different search methods for identifying antibodies with similar binding properties to a given antibody of interest. Our results demonstrate that for different antibodies, different types of search yield the best results. In all three cases investigated, CDR structure alongside CDR sequence identity gave the most accurate results, but in most cases, this also missed functionally cognate antibodies. The inclusion of the light chain in any search always improved accuracy, in agreement with recent studies that have highlighted the importance of the light chain residue motifs for antigen binding (11), and restricted heavy and light chain gene pairs observed among functional antibodies (10). In one case searching by structure provides a link that is not immediately apparent from sequence search alone. This finding is consistent with recent work which uses structural information for epitope binning (18,19).

One of the main challenges in developing PLAbDab, and the area with most potential for improvement, is the pairing of heavy and light chains based on metadata. Although the strategy described in this paper accurately pairs a large number of antibody sequences, there are still many for which pairing was not possible. The PLAbDab generation code also relies on authors publishing their antibodies to databases such as NCBI (20) or the PDB (31), and although many authors do submit their sequences to these databases, there remains a large body of antibody sequence data shared in the text, figures or tables of papers.

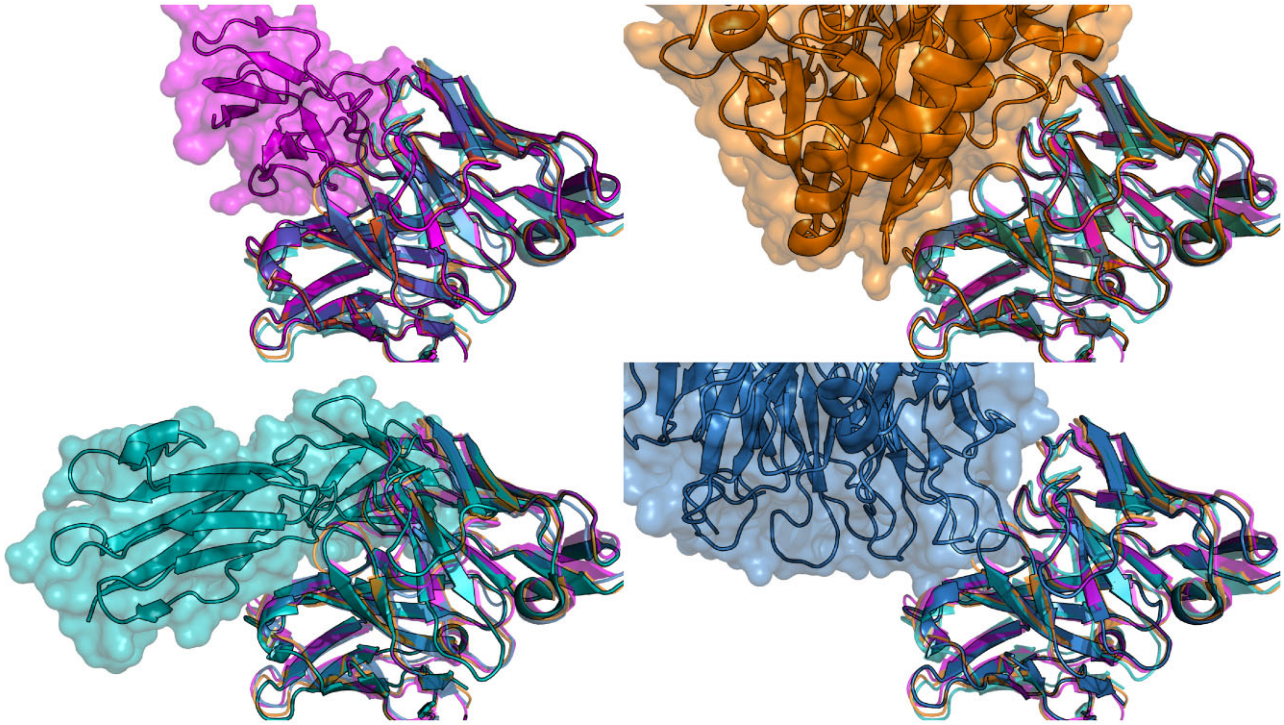


Figure 3. Crystal structure of four antibodies with very similar structure that bind very different antigens. The crystal structure of all four aligned antibodies is shown in each figure (4AL8 in purple, 7XY8 in green, 7LA4 in blue, and 7BBJ in orange). On the top left they are shown in complex with the antigen of 4AL8 (Dengue virus DII protein); top right 7BBJ (CD73); bottom left 7XY8 (Emmprin CD147) and bottom right 7LA4 (Integrin $\alpha_{IIb}\beta_3$).

Nevertheless, PLAbDab already contains over 60k unique annotated antibody sequences providing an invaluable resource to the antibody research community. PLAbDab can also be used to facilitate the generation of antigen specific libraries that could be used to train novel machine learning models or as a starting point for the development of novel therapeutics.

Data availability

PLAbDab is freely available to query at https://opig.stats.ox.ac.uk/webapps/plabdab/. The code is available to download via GitHub (https://github.com/oxpig/PLAbDab) or Zenodo (https://doi.org/10.5281/zenodo.10034277).

Funding

F. Hoffmann-La Roche; GlaxoSmithKline; Engineering and Physical Sciences Research Council [EP/S024093/1]. Funding for open access charge: University of Oxford.

Conflict of interest statement

None declared.

References

1. Kaplon,H., Crescioli,S., Chenoweth,A., Visweswaraiah,J. and Reichert,J.M. (2023) Antibodies to watch in 2023. *Mabs*, **15**, 2153410.
2. Raybould,M.I., Marks,C., Lewis,A.P., Shi,J., Bujotzek,A., Taddese,B. and Deane,C.M. (2020) Thera-SAbDab: the therapeutic structural antibody database. *Nucleic Acids Res.*, **48**, D383–D388.
3. Espejo,A.P., Akgun,Y., Al Mana,A.F., Tjendra,Y., Millan,N.C., Gomez-Fernandez,C. and Cray,C. (2020) Review of current advances in serologic testing for COVID-19. *Am. J. Clin. Pathol.*, **154**, 293–304.
4. Pollard,A.J. and Bijker,E.M. (2021) A guide to vaccinology: from basic principles to new developments. *Nat. Rev. Immunol.*, **21**, 83–100.
5. Regep,C., Georges,G., Shi,J., Popovic,B. and Deane,C.M. (2017) The H3 loop of antibodies shows unique structural characteristics. *Proteins*, **85**, 1311–1318.
6. Gordon,G.L., Capel,H.L., Guloglu,B., Richardson,E., Stafford,R.L. and Deane,C.M. (2023) A comparison of the binding sites of antibodies and single-domain antibodies. *Front. Immunol.*, **14**, 1231623.
7. Kovaltsuk,A., Leem,J., Kelm,S., Snowden,J., Deane,C.M. and Krawczyk,K. (2018) Observed Antibody Space: a resource for data mining next-generation sequencing of antibody repertoires. *J. Immunol.*, **201**, 2502–2509.
8. Corrie,B.D., Marthandan,N., Zimonja,B., Jaglale,J., Zhou,Y., Barr,E., Knoetze,N., Breden,F.M., Christley,S., Scott,J.K., *et al.* (2018) iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.*, **284**, 24–41.
9. Olsen,T.H., Boyles,F. and Deane,C.M. (2022) Observed Antibody Space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.*, **31**, 141–146.
10. Jaffe,D.B., Shahi,P., Adams,B.A., Chrisman,A.M., Finnegan,P.M., Raman,N., Royall,A.E., Tsai,F., Vollbrecht,T., Reyes,D.S., *et al.* (2022) Functional antibodies exhibit light chain coherence. *Nature*, **611**, 352–357.
11. Shrock,E.L., Timms,R.T., Kula,T., Mena,E.L., West,A.P. Jr, Guo,R., Lee,I.-H., Cohen,A.A., McKay,L.G., Bi,C., *et al.* (2023) Germline-encoded amino acid–binding motifs drive immunodominant public antibody responses. *Science*, **380**, eadc9498.

12. Dunbar,J., Krawczyk,K., Leem,J., Baker,T., Fuchs,A., Georges,G., Shi,J. and Deane,C.M. (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.
13. Schneider,C., Raybould,M.I. and Deane,C.M. (2022) SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res.*, **50**, D1368–D1372.
14. Raybould,M.I., Kovaltsuk,A., Marks,C. and Deane,C.M. (2021) CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, **37**, 734–735.
15. Krawczyk,K., Buchanan,A. and Marcatili,P. (2021) Data mining patented antibody sequences. *mAbs*, **13**, 1892366.
16. Fleri,W., Paul,S., Dhanda,S.K., Mahajan,S., Xu,X., Peters,B. and Sette,A. (2017) The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol.*, **8**, 278.
17. Olsen,T.H., Abanades,B., Moal,I.H. and Deane,C.M. (2023) KA-Search, a method for rapid and exhaustive sequence identity search of known antibodies. *Sci. Rep.*, **13**, 11612.
18. Robinson,S.A., Raybould,M.I., Marks,C., Schneider,C., Wong,W.K. and Deane,C.M. (2021) Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies. *PLoS Comput. Biol.*, **17**, e1009675.
19. Spoendlin,F.C., Abanades,B., Raybould,M. I.J., Wong,W.K., Georges,G. and Deane,C.M. (2023) Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind the same epitope. *Front. Mol. Biosci.*, **10**, 1237621.
20. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S., *et al.* (2021) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **50**, D20–D26.
21. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
22. Dunbar,J. and Deane,C.M. (2016) ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298–300.
23. Abanades,B., Wong,W.K., Boyles,F., Georges,G., Bujotzek,A. and Deane,C.M. (2023) ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. *Commun. Biol.*, **6**, 575.
24. Olsen,T.H., Moal,I.H. and Deane,C.M. (2022) AbLang: an antibody language model for completing antibody sequences. *Bioinform. Adv.*, **2**, vbac046.
25. Raybould,M.I., Marks,C., Krawczyk,K., Taddese,B., Nowak,J., Lewis,A.P., Bujotzek,A., Shi,J. and Deane,C.M. (2019) Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 4025–4030.
26. Raybould,M.I., Turnbull,O.M., Suter,A., Guloglu,B. and Deane,C.M. (2023) Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. bioRxiv doi: https://doi.org/10.1101/2023.06.28.546839, 30 June 2023, preprint: not peer reviewed.
27. Gilman,M.S., Castellanos,C.A., Chen,M., Ngwuta,J.O., Goodwin,E., Moin,S.M., Mas,V., Melero,J.A., Wright,P.F., Graham,B.S., *et al.* (2016) Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors. *Sci. Immunol.*, **1**, eaaj1879.
28. Jubb,H.C., Higueruelo,A.P., Ochoa-Montaño,B., Pitt,W.R., Ascher,D.B. and Blundell,T.L. (2017) Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.*, **429**, 365–371.
29. DeLano,W.L. (2002) Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.*, **40**, 82–92.
30. Mukhamedova,M., Wrapp,D., Shen,C.-H., Gilman,M.S., Ruckwardt,T.J., Schramm,C.A., Ault,L., Chang,L., Derrien-Colemyn,A., Lucas,S.A. and et,al. (2021) Vaccination with prefusion-stabilized respiratory syncytial virus fusion protein induces genetically and antigenically diverse antibody responses. *Immunity*, **54**, 769–780.
31. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Received: July 26, 2023. **Revised:** October 20, 2023. **Editorial Decision:** October 20, 2023. **Accepted:** October 30, 2023