Bioinformatics, 38(2), 2022, 377–383
doi: 10.1093/bioinformatics/btab660
Advance Access Publication Date: 21 September 2021
Original Paper

Structural bioinformatics

DLAB: deep learning methods for structure-based virtual screening of antibodies

Constantin Schneider [1], Andrew Buchanan[2], Bruck Taddese[3,†] and Charlotte M. Deane [1,*]

[1]Department of Statistics, University of Oxford, Oxford, OX1 3LB, UK, [2]Antibody Discovery & Protein Engineering, R&D, AstraZeneca, Cambridge, CB2 0AA, UK and [3]Discovery Sciences, R&D, AstraZeneca, Cambridge, CB2 0AA, UK

*To whom correspondence should be addressed.
†Bruck Taddese has left AstraZeneca since the completion of the work detailed in this manuscript.
Associate Editor: Alfonso Valencia

Received on February 24, 2021; revised on August 3, 2021; editorial decision on September 1, 2021; accepted on September 1, 2021

Abstract

Motivation: Antibodies are one of the most important classes of pharmaceuticals, with over 80 approved molecules currently in use against a wide variety of diseases. The drug discovery process for antibody therapeutic candidates however is time- and cost-intensive and heavily reliant on *in vivo and in vitro* high throughput screens. Here, we introduce a framework for structure-based deep learning for antibodies (DLAB) which can virtually screen putative binding antibodies against antigen targets of interest. DLAB is built to be able to predict antibody–antigen binding for antigens with no known antibody binders.

Results: We demonstrate that DLAB can be used both to improve antibody–antigen docking and structure-based virtual screening of antibody drug candidates. DLAB enables improved pose ranking for antibody docking experiments as well as selection of antibody–antigen pairings for which accurate poses are generated and correctly ranked. We also show that DLAB can identify binding antibodies against specific antigens in a case study. Our results demonstrate the promise of deep learning methods for structure-based virtual screening of antibodies.

Availability and implementation: The DLAB source code and pre-trained models are available at https://github.com/oxpig/dlab-public.

Contact: deane@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Antibodies are the most successful class of biotherapeutics, with 85 monoclonal antibody drugs on the US market at the time of writing. Global sales in monoclonal antibody therapeutics reached an estimated $98 billion in 2018 (Grilo and Mantalaris, 2019), making them one of the fastest growing and largest segments of the pharmaceutical industry. The potential and use of antibodies as therapeutics for a wide range of diseases is due to the high specificity and affinity of their binding, facilitated through the variability in their complementarity determining regions (CDRs) (Liu, 2014).

The development process for novel antibody therapeutics already benefits from computational tools predicting properties of the antibody and the antibody–antigen complex (Raybould *et al.*, 2019a). Here, we demonstrate that machine learning approaches using structural antibody data can enable large-scale computational screening in the antibody development pipeline.

To develop a successful therapeutic antibody, several features have to be optimized. Antibodies need to bind with high efficacy, specificity and affinity to the target of interest, while at the same time avoiding an immune reaction in the patient (Suscovich and Alter, 2015) as well as avoiding properties that lead to poor developability, such as self-association, viscosity or immunogenicity (Raybould *et al.*, 2019b). To achieve these goals, large-scale experimental screens are usually used in the pre-clinical stages of antibody–drug development (Almagro *et al.*, 2017; Suscovich and Alter, 2015).

Typically, initial leads for therapeutic human antibodies are generated using either *in vitro* display platforms or *in vivo* transgenic animals. Further improvement of these initial hits is then achieved through affinity maturation, either via generation and screening of further, hit-based mutagenesis libraries or via rational engineering. High-affinity antibodies generated in this way can be further engineered to achieve desirable properties for antibody therapeutics, for example changes to the constant region to modulate the effector

functions as well as *in vivo* half-life and the improvement of developability (Chiu and Gilliland, 2016).

These experimental methods are often effective at generating high-affinity antibodies for downstream development, but are cost- and time-intensive. Furthermore, they do not as standard generate insights into the binding mode of the generated antibodies, or if they are binding to the target epitope on the antigen.

The development pipeline detailed above can be supplemented using *in silico* methods, particularly once initial binding candidates have been identified. Computational tools have been used to rationally engineer the binding site in several different ways. For a recent comprehensive review of computational tools used in antibody engineering, see Norman *et al.* (2020). As antibody binding affinity is defined by the three-dimensional structure of the antigen-binding region on the antibody (the paratope, which is mainly composed of the CDRs), antibody modelling tools can aid the design process by rapidly generating models of antibodies [e.g. ABodyBuilder (Leem *et al.*, 2016), Rosetta (Weitzner *et al.*, 2017) and Kotai Antibody Builder (Yamashita *et al.*, 2014)]. Some methods combine antibody model generation with docking against target antigens and engineering of the antibody [e.g. the Rosetta antibody design tool (Adolf-Bryfogle *et al.*, 2018)].

The antibody models generated by those methods, though commonly high quality, are often not accurate across the CDRs, particularly the CDR H3 (Almagro *et al.*, 2014; Leem *et al.*, 2016). This presents a challenge for rigid-body docking methods, which do not allow flexibility of the binding partners, to be able to recapitulate the true antibody–antigen interface. More accurate interface models can be generated using computationally intensive docking approaches (Weitzner *et al.*, 2017), however, those are too slow to be used on high-throughput screens, where thousands of interface models need to be generated (Raybould *et al.*, 2019a).

The caveats described above (CDR model inaccuracy and the speed/accuracy trade-off for interface prediction) mean that currently, no effective structure-driven computational tools are available for the early high-throughput screening stages of the antibody development pipeline.

Machine learning approaches in the field of antibody therapeutics discovery have so far mainly focused on sequence rather than on structure as input (Bujotzek *et al.*, 2015; Liberis *et al.*, 2018; Mason *et al.*, 2021; Olimpieri *et al.*, 2013). These machine learning approaches have been shown to be highly efficient for prediction tasks which depend only on the antibody (rather than on the antibody–antigen interface) or are specific to one antigen: random forests have been used both for paratope prediction (Olimpieri *et al.*, 2013) and for the prediction of VH-VL interface angles in antibody modelling (Bujotzek *et al.*, 2015). A current state-of-the art paratope prediction tool, Parapred, is based upon a recurrent neural network approach using the variable region amino acid sequence of the antibody as input data (Liberis *et al.*, 2018). Another recent study has highlighted the potential of large scale, high throughput machine learning approaches for early stage antibody therapeutics development (Mason *et al.*, 2021). A large-scale (50k samples) mutagenesis study on one antibody target was used to train a recurrent neural network, which was then used to retrieve new antibody sequences from a sub-sample of antibody sequence space, all of which showed binding affinity *in vitro*. For a review of deep learning approaches in antibody research see Graves *et al.* (2020).

Sequence-based machine learning approaches have yet to provide generalizable predictions across different antigens in one model. An approach in this direction has recently been explored in a study by Akbar *et al.* (2021), in which machine learning accessible descriptions of the binding interface were generated using structure-derived interaction motifs. Further, graph convolutional neural networks have recently been used to encode structural information for the prediction of antibody and antigen interface residues (Pittala and Bailey-Kellogg, 2020), demonstrating the ability of machine learning approaches to successfully utilize structural information derived from both the antibody and the antigen.

2 Approach

Here, we describe a structure-based deep learning approach for early-stage virtual screening of antibody therapeutics, when an epitope target of interest is known but no viable hit antibodies have yet been identified. Our approach is able to make generalizable predictions across different antigens. Adopting a three-dimensional gridding method which has been used successfully alongside convolutional neural network methods to predict small molecule/protein binding (Imrie *et al.*, 2018; Ragoza *et al.*, 2017), we implement a similar convolutional neural network trained on rapidly generated rigid-body docking poses of modelled antibody structures in complex with antigen epitopes. We use this **D**eep **L**earning approach for **A**nti**B**ody screening (**DLAB**) to both improve the ranking of docks from the ZDock docking algorithm (Pierce *et al.*, 2011) and, in combination with docking scores generated by ZDock, for the prediction of antibody–antigen binding.

3 Materials and methods

3.1 Crystal structure dataset

The structural antibody database (SAbDab) (Dunbar *et al.*, 2014) contains an up-to-date collection of all antibody structures deposited in the PDB (Berman *et al.*, 2000). We selected a dataset of structures of VH-VL paired antibodies in complex with protein or peptide antigens with a resolution of less than 3 Å from a snapshot of the SAbDab, downloaded on December 19, 2018. The dataset consisted of 1216 pdb files of antibody–antigen complexes, of which 759 were non-redundant. Here, we considered an antibody to be non-redundant if its CDR sequence (the concatenated sequence across all CDR regions according to IMGT definition) was only present once in the dataset. In the following, this dataset of 759 complexes is called the crystal structure dataset. The PDB accession codes for the crystal structure dataset can be found in Supplementary File S1.

3.2 Model dataset

All antibody structures in the crystal structure dataset were modelled using ABodyBuilder (Leem *et al.*, 2016). Only non-identical template structures were used for modelling. All sidechains were modelled using PEARS (Leem *et al.*, 2018). This set of modelled antibodies is referred to as the model dataset. Model quality was assessed by calculating the Cα root mean square deviation (RMSD) of each modelled antibody to the crystal structure of the antibody, either across the entire Fv region or across the antibody CDRs by aligning the modelled and the crystal structure antibody by their framework regions and calculating RMSD across the CDR Cα atoms. The antibody models generated using this workflow were of comparable quality to previously published studies (Leem *et al.*, 2016) (see Supplementary Figs S1A and S2).

3.3 Binding and non-binding antibody/antigen examples

We considered every antibody–antigen pair in the crystal structure dataset as a binding pair. Since antibodies bind with high specificity, we generated non-binding antibody–antigen pairs by randomly sampling 50 non-cognate antibodies per antigen from the crystal structure/model antibody dataset. Our definition of non-cognate required the sampled antibodies to share less than 90% of their CDR sequence with the binding antibody.

3.4 Docking pose generation

For both the crystal structure and model dataset, docked antibody–antigen pairs were generated using ZDOCK (Pierce *et al.*, 2011). For each pair, 500 poses (distinct structures of the antibody–antigen complex) were generated. To aid the docking process, the paratope and epitope were identified and residues not belonging to either the paratope or the epitope were excluded from the interaction site using the standard ZDOCK pipeline as described in the ZDOCK documentation.

The antigen epitope was defined by separating the antibody and antigen in the crystal structure and calculating surface exposed residues for each binding partner separately using the PSA algorithm (Lee and Richards, 1971). All atoms belonging to surface residues on the antigen less than 4 Å from a surface exposed residue on the antibody were considered part of the epitope. Further, all atoms belonging to a surface exposed residue on the antigen within 4 Å of the defined epitope were included in the allowed docking interface to model a realistic level of access to the epitope.

The paratope of crystal structure antibodies was defined in the same way, using the interacting residues from the crystal structure.

For modelled antibody structures, the paratope was defined using the IMGT CDR definition, marking the IMGT defined CDRs and two residues to either side of each CDR as the paratope.

3.5 Clustering for train-test splits

To avoid similarity between binding modes in the train and test sets, CD-Hit (Li and Godzik, 2006) was used to cluster antibodies by CDR sequence identity as defined above, using a clustering cutoff of 90% sequence identity. For all learning tasks set out below, train-test splits were performed using clustered cross-validation, assigning all members of a cluster to either the train or the test set.

3.6 Data input for convolutional neural networks

Following the method of Ragoza *et al.* (2017), the docking poses were prepared for input into convolutional neural networks (CNNs) by discretising the atom information into four-dimensional grids, where three dimensions describe the spatial arrangement of the interaction site and the fourth dimension is used to indicate atom types (see Supplementary Fig. S3).

The centre of the interaction site of docking poses was calculated using the PSA algorithm by averaging the coordinates of all surface-exposed atoms within 4 Å of the interaction partner on both the antibody and the antigen and taking the mean of the two centre points. Poses which after docking had no interactions under 4 Å were discarded.

The grid contained only atoms that were within 24 Å of the interaction centre. These interaction site specifications were found to cover on average 96% of all interacting atoms on both antibody and antigen (see Supplementary Fig. S4). The grid resolution was set to 0.5 Å, leading to a total grid size of 96^3 voxels.

3.7 Reranking ZDOCK docking poses with DLAB-Re

To improve ZDOCK docking pose ranking, we created a machine learning method to identify and correctly rank good docking poses. This method, termed DLAB-Re(scoring), is a CNN (architecture shown in Supplementary Fig. S5) which predicts the *fnat* score The *fnat* score is the fraction of contacts between the interaction partners in the crystal structure that are recapitulated in the docked pose. The network generates a probability distribution over 11 *fnat* intervals (10 steps of size 0.1 over [0, 1] and one bin for *fnat* 0.0 poses), which are used to generate a *fnat* prediction via weighted averaging. For model training and testing, the top 500 poses as ranked by ZDOCK for each pairing in the binder set were annotated with their respective *fnat* score interval. For any given antibody–antigen pairing, there are considerably more poses with low *fnat* scores in the top 500 ZDOCK poses than high *fnat* poses. To avoid biasing the network towards predicting all poses into low *fnat* intervals, we used a stratified sampling scheme, sampling poses from each interval at the same rate during training (but not during testing).

During training, the input data were augmented by random rotation around the interaction centre, followed by random translations along the *x*, *y and z* axis between –2 and 2 Å. Models were trained for 200 000 parameter update steps using categorical cross-entropy and the rectified Adam optimizer. Since we wanted to use the improved ranking performance of DLAB-Re on the downstream virtual screening task, it was used to rerank the top 500 poses for all antibody–antigen pairings used during training and evaluation of DLAB-VS. For this experiment, we used the model weights derived during cross-validated training. In the case of cognate and non-cognate antibody–antigen pairings, the DLAB-Re model used was not trained on either the pairings or on the antigen.

To identify antibody–antigen pairings with low-quality docking poses, we determined the highest DLAB-Re score given to any of the top 500 poses generated by ZDOCK for each antibody–antigen pairing. This score (DLAB-Re-max) was used to discard particular pairings by ranking all pairings by their DLAB-Re-max score and discarding the bottom 40%, 60% or 80%. To contrast the performance of ZDOCK on the same task, this score thresholding was also applied to the ZDOCK output score of the top pose as ranked by ZDOCK.

3.8 Virtual screening with DLAB-VS and ZDOCK

The goal of virtual antibody screening is to discern binding antibodies against a given epitope from a pool of candidate antibodies. To generate a classification model able to accomplish this task, we trained an ensemble of CNN models (architectures depicted in Supplementary Fig. S5), which we termed DLAB-VS (virtual screening), a binder/non-binder classifier for individual docking poses of antibody–antigen complexes.

The input poses for training were selected as follows. For non-binding pairs, the highest ranked pose after DLAB-Re rescoring was selected as a non-binding pose. For binding pairs, we selected up to 50 poses with $fnat > 0.7$ where those were available. Further, following the approach taken by Scantlebury *et al.* (2020), five poses of the same binding pair with $fnat < 0.1$ were selected as non-binding poses to force the networks to learn from the interaction between antibody and antigen by providing for the same antibody–antigen pairing both good and bad binding poses.

Data augmentation was performed in the same manner as for DLAB-Re. Models were trained for 50 000 parameter update steps using the rectified Adam optimizer (as the smaller input dataset resulted in earlier convergence). A validation set comprising 10% of the total dataset was created using the same CD-Hit clustering as for the training set creation. The validation set was used to select a snapshot of the model during training by choosing the model snapshot with the highest average precision on the validation set. To counteract class imbalance, binder and non-binder poses were sampled so that each batch contained equal numbers of both classes.

For each train/test split and network architecture, two different validation sets were used to train an ensemble of four models per fold.

At test time, the DLAB-VS scores of the top 10 poses were averaged. As described above, we used DLAB-Re reranked poses for this purpose. Where an ensemble of models was used, the output scores by the ensemble members were averaged to arrive at the DLAB-VS score. For each antigen target, the antibodies docked against that target (correct and decoys) were ranked by their respective DLAB-VS score.

For the ZDOCK-based classifier, the ZDOCK score of the top-ranked binding pose was used to rank the antibodies docked against a particular target.

For the DLAB-VS+ZDOCK model, the DLAB-VS output scores and the ZDOCK output score were normalized per antigen target via minmax scaling and averaged to arrive at the final scores for each target antigen.

The DLAB-Re-max score, output from the reranking method, of each antibody–antigen pairing was used to discard antigens for which the binding antibody was not predicted to have produced any satisfactory docking poses.

3.9 DOVE rescoring

We compared the DLAB-Re results to the DOVE method for CNN-based docking pose ranking (Wang *et al.*, 2020). Input file preparation and score generation were performed according to the tutorials on the author's github page. As detailed on the author's github page, only the GOAP and ATOM20/ATOM40 scores were used, as the IT-scores were unavailable.

3.10 Additional test sets

To create an unseen test set, which was not used at any point during model choice and hyperparameter optimization, we used SAbDab entries deposited after the snapshot used for the training dataset creation to create an unseen test set. This dataset, referred to in the following as the post-snapshot model dataset, contained 222 antibody/protein antigen complexes, which formed 173 CDR clusters after clustering the CDR-sequences using CD-Hit at 90% identity. On this test set, we performed modelling, docking, rescoring and binder classification as described above, using the models trained on the model dataset, with the exception of the ensemble DLAB-VS score calculation, for which we combined all 40 previously trained models into one ensemble from which the DLAB-VS score for each pairing was averaged. The PDB accession codes for this dataset can be found in Supplementary File S2.

We created a SARS-CoV2 dataset (the SARS-CoV2 variant dataset) by extracting all antibodies from the Coronavirus Antibody Database (CoVAbDab) (Raybould *et al.*, 2020) which were confirmed to bind to the SARS-CoV2 wild-type RBD while also being confirmed not to bind to at least one SARS-CoV2 variant and for which an experimentally determined complex structure was available from which the epitope could be determined as described above. We used ABodyBuilder to model the antibodies as described above. We created structural models of the variant antigens using Foldx5 (Schymkowitz *et al.*, 2005), using the PDB files listed in Supplementary Table S1 as templates and copied the epitope definition for docking purposes from the templates onto the variant models. We then docked each antibody model, defining the paratope as described above, against its epitope on both the wild-type RBD and the confirmed non-cognate variant RBDs and performed rescoring and binder classification as described above, using the 40-model ensemble. As for most of the epitopes in this set, only one antibody was docked, score normalization was performed over the entire dataset instead of on a per-epitope basis.

3.11 Statistical testing

For statistical significance testing of the difference between means of the best *fnat* in top 10 ranked poses, we used the two-tailed t-test implementation in the scipy python package (Virtanen *et al.*, 2020). For statistical significance testing of the ranking performance of DLAB-Re, we considered the ratio of antibody–antigen pairs for which a pose with a specific *fnat* is found in the top 10 poses a Poisson rate and calculated *P*-values using the implementation of the test described in Gu *et al.* (2008) in the statsmodels python package (Seabold and Perktold, 2010). Correspondingly, we approximated the standard deviation of the ratio as $\frac{\sqrt{c}}{n}$, where c is the count of antibody–antigen pairings for which a pose with a specific *fnat* is found in the top X poses and n is the total number of pairings assessed.

4 Results

4.1 Crystal structure docking yields high-quality poses

In order to establish a baseline for ZDOCK performance on antibody and antigen crystal structures, we re-docked the complexes in the crystal structure dataset and quantified the docking performance through the *fnat* score of the docking poses (Supplementary Figs S1B, C and S6). ZDOCK yielded high-quality docking poses, ranking at least one pose with *fnat* > 0.5 in the top ten poses for 93% of the pairings.

4.2 Model docking yields low-quality poses

Docking the model dataset antibodies against their cognate antigens on the other hand yields lower-quality docking poses. Here, ZDOCK created a pose with *fnat* > 0.5 in the top ten poses for only 44% of antibody–antigen pairings but 70% of pairings had a pose with *fnat* > 0.5 in their 500 highest ranked poses (Supplementary Figs S1C and S6).

4.3 DLAB-Re can improve ZDOCK docking pose ranking

Given these results, the first stage was to create a method that is able to identify good docking poses and rank these correctly. DLAB-Re is a CNN trained to predict the *fnat* of docking poses of antibody–antigen pairings. To determine the ability of our method, DLAB-Re, to improve docking pose ranking, we ranked the top 500 docking poses generated by ZDOCK for each binder pair by the predicted *fnat* value, using the clustered, cross-validated train-test procedure set out in Section 3 (compare Supplementary Fig. S1E).

This rescoring procedure improves upon the performance of ZDOCK ranking. On the crystal structure dataset, DLAB-Re recapitulates the ranking performance of native ZDOCK. On the model dataset, DLAB-Re significantly increased the number of antibody–antigen pairings for which a pose with *fnat* > 0.5 is ranked in the top ten poses by 16% ($P = 0.047$) (see Fig. 1C), significantly increasing both the mean best *fnat* in the top 10 poses from 0.46 to 0.50 ($P = 0.002$). In Supplementary Figure S7, we show two antibody–antigen pairings for which DLAB-Re strongly increases the *fnat* of the best pose in the top 10 ranked poses.

4.4 DLAB-Re enables identification of successfully docked antibody–antigen pairings

Using the maximum predicted *fnat* score generated by DLAB-Re, we can discard poorly docked antibody–antigen pairings (see Section 3 and Fig. 1B and D). Choosing thresholds so that 40%, 60% or 80% of pairings, respectively, are discarded, the remaining pairings are increasingly enriched both in pairings for which the docking poses are ranked well by DLAB-Re as well as in pairings which have at least one pose with a high *fnat* score in the top 500 poses. Discarding 80% of the pairings in the model dataset raises the proportion of pairings for which a pose with at least 0.5 *fnat* was ranked by DLAB-Re in the top ten poses from 51% to 84%, meaning that using this thresholding approach eliminated 93.4% of the antibody–antigen pairings for which DLAB-Re did not manage to rank a pose with at least 0.5 *fnat* in the top ten poses while retaining 33% of the pairings for which it did.

Using this approach with the ZDOCK output scores does not yield the same improvement, only raising the proportion of pairings for which a pose with at least 0.5 *fnat* was ranked by ZDOCK in the top ten poses from 44% to 57% (see Supplementary Fig. S8).

4.5 A CNN docking rescoring tool trained on crystal structure data does not replicate the DLAB-Re performance

We compared the performance of DLAB-Re with the DOVE tool developed by Wang *et al.* (2020). DOVE is a CNN-based docking pose evaluation tool, which is designed to predict docking pose quality according to CAPRI criteria on crystal structure-based general protein–protein docking poses. We used the publicly available ATOM40+GOAP model to test whether this training generalizes to the antigen—model antibody docking case. Using the DOVE ATOM40+GOAP score to rerank antibody–antigen docking poses, DOVE performs considerably worse than both ZDOCK and DLAB-Re. This holds for both the crystal structure and the model dataset (see Supplementary Figs S1D, S9A and B).

DOVE is trained to classify docking poses into one of two classes: CAPRI acceptable and not CAPRI acceptable (Wodak and Méndez, 2004; Wong *et al.*, 2019), where CAPRI acceptable poses have *fnat* > 0.1 and interface RMSD < 4A or ligand RMSD < 10A. Using this classification to evaluate both DOVE and DLAB-Re results, DOVE still performed worse than ZDOCK and DLAB-Re (see Supplementary Fig. S9C). These results highlight the added value from training DLAB-Re both on a domain specific (antibody–antigen) as well as task specific (modelled antibodies) dataset.

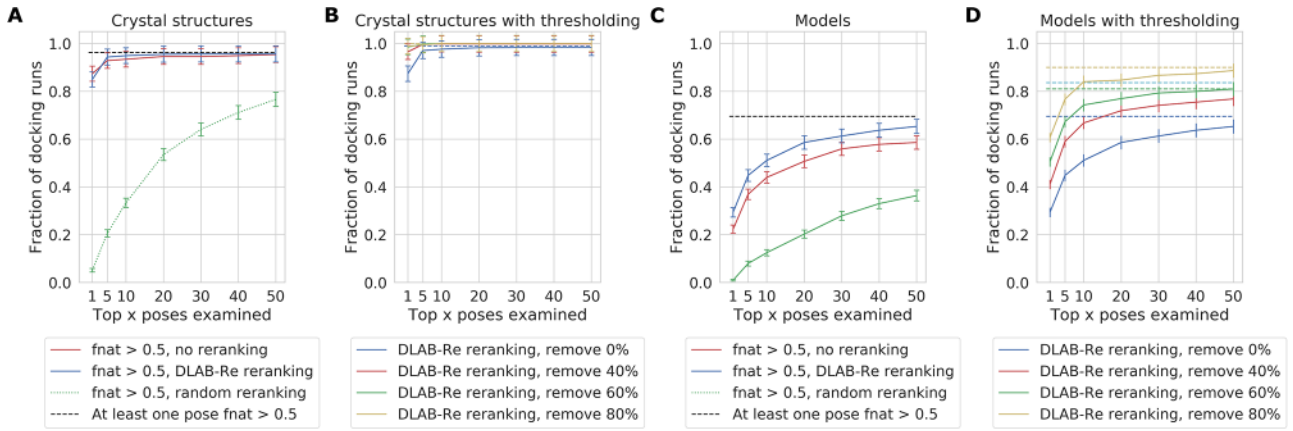

Fig. 1. DLAB-Re improves docking performance on the crystal structure dataset (**A, B**) and the model dataset (**C, D**). On crystal structure data, ZDOCK ranking and DLAB-Re ranking perform similarly and well. On models, the ZDOCK baseline performance is considerably worse and DLAB-Re significantly improves ranking performance. (A, C) DLAB-Re ranks the top 500 poses generated by ZDOCK better than ZDOCK, enriching the ratio of pairings with *fnat* > 0.5 poses ranked highly. The dashed line indicates the fraction of docking runs in which a pose with *fnat* > 0.5 is present in the 500 assessed poses.(B, D) Using the DLAB-Re-max score to remove 40%, 60% or 80% of the antibody–antigen pairings, respectively, can remove antibody–antigen pairings which did not yield high-*fnat* poses. This selects for pairings for which *fnat* > 0.5 poses exist in the top 500 poses generated by ZDOCK (dashed line) and for which DLAB-Re ranks the top 500 poses well (solid line). Error bars are ±one standard deviation, approximated, as described in Section 3

4.6 ZDock easily retrieves binders from the crystal structure dataset but not from the model dataset

Virtual screening for antibody discovery aims to find binding antibodies for a given epitope from a large set of potential binders. We attempted to retrieve correct binders from the crystal structure dataset by docking both the cognate antibody crystal structure as well as 50 non-cognate antibody crystal structures against each antigen crystal structure in the dataset. Using ZDock, the cognate binder is found in the top 2% (i.e. top ranked) of antibodies for 49.7% of antigen targets, and in the top 10% in 65.6% of antigen targets, outperforming the random baseline. On the model dataset, using the same procedure, ZDock ranks the binder in the top 2% of antibodies for only 5.5% of antigen targets, and in the top 10% for only 18.8% of antigen targets (see Supplementary Fig. S10).

4.7 DLAB-VS and ZDock can be combined to improve performance on the model dataset

To improve our ability to perform virtual screening on modelled antibodies, we trained a new model, termed DLAB-VS (virtual screening) to classify antibody–antigen pairings as binders or non-binders, as detailed in Section 3. For a large proportion of model antibody structures, the docking pipeline does not yield high-quality complex structures (see Fig. 1). Therefore, at train time, only docking poses with *fnat* > 0.7 were shown to the network as positive binders and poses with *fnat* < 0.1 were shown to the network as non-binders. Furthermore, at test time, we averaged the network output over the ten highest-ranked poses for each antibody–antigen pairing. Finally, rather than training a single model for each train/test fold, we trained four models using two different architectures for each of ten clustered cross-validation folds as detailed in Section 3 and used the averaged output as the DLAB-VS score. Using this training approach, the DLAB-VS model achieved classification performance comparable to ZDock: on the model dataset, the cognate antibody was ranked in the top 2% of antibodies for 4.7% of antigens and in the top 10% for 16.4% of antigens (see Supplementary Fig. S10).

However, the two approaches, ZDock and DLAB-VS, did not perform equally across antigen targets. It was possible to improve classification performance using the mean of the two scores to rank putative binders (see Section 3 and Supplementary Fig. S1E and F). Using this approach, termed DLAB-VS+ZDock, the binder was ranked in the top 2% of antibodies for 6.4% of antigen targets, and in the top 10% in 19.7% of antigen targets (see Fig. 2). In the following, we use this DLAB-VS+ZDock approach.

4.8 Using DLAB-Re to discard antigen targets enables selection of well-performing models

The performance of the DLAB-VS+ZDock model is highly dependent on the quality of the docking poses from which the score is derived as well as the quality of the antibody model (see Supplementary Figs S1G and S11). As described above, DLAB-Re enables the selection of well-docked antibody–antigen pairings, therefore, to further improve the performance of the DLAB-VS+ZDock classifier, we used the output from DLAB-Re to identify antibody models with high likelihood of being well docked from the model dataset. To test if this would improve results, for each of the training cross-validation folds, we only considered antigen targets for which the DLAB-Re-max score of the cognate antibody was within the top 20% of DLAB-Re-max scores of antibody–antigen pairings within that fold. On these targets, where the cognate antibody was predicted to be well docked by DLAB-Re, the cognate antibody was ranked in the top 2% for 17.6% of antigens and in the top 10% for 40.8% of antigens (see Fig. 2A). This improvement was reliant on using the combined DLAB-VS+ZDock model, using the same approach while ranking by the ZDock output scores alone only increased the classification performance marginally (see Supplementary Fig. S10).

4.9 DLAB-VS+ZDock performs well on the post-snapshot model dataset

To test the performance of the DLAB pipeline on a completely unseen test set (see Supplementary Fig. S1H), we ran the pipeline on the post-snapshot model dataset (see Section 3). On this set, an ensemble of all 40 DLAB-Vs models trained during the clustered cross-validation training on the model dataset and the ZDock output scores were used to rank the cognate antibody model as well as 50 non-cognate antibody models. The ensemble achieved higher performance on this test set than on the previous test cases (both with and without the DLAB-Re selection criterion).

To calculate how different the post-snapshot model dataset was from the training set and the potential for this to influence model performance, we clustered the antibody CDR sequences in both the model dataset and the post-snaphot model dataset at 90% identity. Of the new CDR sequences, 17.3% clustered with at least one CDR sequence in the snapshot. On the cognate antigen targets for those antibodies, the large ensemble performs exceptionally well both before and after DLAB-Re thresholding (binder in top 2% for 18% and 57% of antigen targets respectively, see Supplementary Fig. S12). On the subset of new additions to SAbDab without overlap to the snapshot, the performance using the 40-model DLAB-

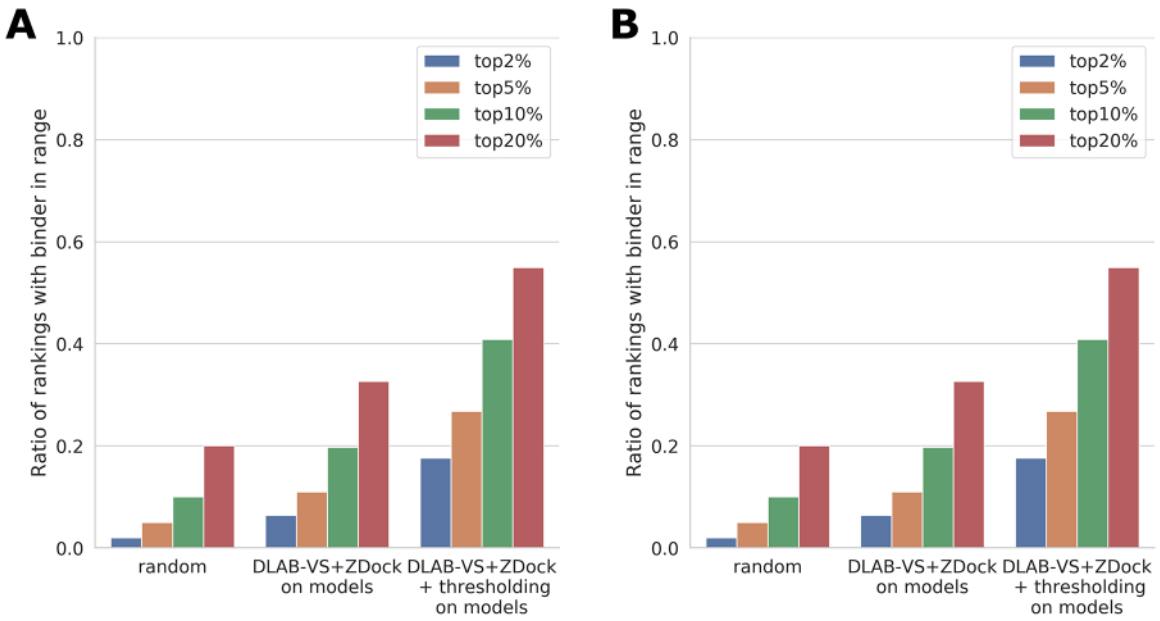

Fig. 2. DLAB-VS and ZDock binder classification performance. For each approach, the ratio of pairings for which the binding antibody was ranked in the top 2%, top 5%, top 10% and top 20% respectively is shown. (**A**) Comparison of the performance of ZDock and DLAB-VS binder classification on the model dataset to the random expectation ('random') of finding the binder in the top N%. Using the combination of DLAB-VS and ZDock scores ('DLAB-VS+ZDock') detailed in Section 3 and supplementing it with the DLAB-Re-max thresholding ('DLAB-VS+ZDock + thresholding'), the classification performance on the model dataset can be improved significantly. (**B**) Performance on the post-snapshot model dataset after removing any CDR sequences with overlap to the model dataset (at 90% CDR sequence identity as defined above) both without ('DLAB-VS+ZDock') and with ('DLAB-VS+ZDock + thresholding') DLAB-Re-max score thresholding. Performance for the CDR sequences clustering with sequences in the model dataset is shown in Supplementary Figure S12

VS+ZDock ensemble is similar to the performance on the snapshot using 10 folds of 4-model ensembles (see Fig. 2B). The generalization performance of DLAB-VS is therefore not based on overlap between the training set and the post-snapshot test set. We further demonstrated this by assessing the performance of the DLAB pipeline on the set of post-snapshot targets with cognate antibodies with at most 85%, 80%, 75% or 70% CDR sequence identity to any antibody in the training set. We observed no significant change in performance between the different thresholds, demonstrating that the predictive performance of DLAB is generalizable (see Supplementary Fig. S13).

4.10 DLAB-VS can distinguish binding and non-binding antigen variants

A use case of interest is determining whether mutations in the antigen can disrupt antibody binding (see Supplementary Fig. S1I). To test whether this task is accessible to structure-based deep learning tools, we created a dataset of antibodies confirmed to bind against the SARS-CoV2 wild-type receptor binding domain (RBD) while also being confirmed not to bind against at least one SARS-CoV2 RBD variant. We ran the dataset through the DLAB pipeline as described in Section 3.10 and assessed whether the DLAB-VS+ZDOCK output score consistently scored the antibody–wild-type pair higher than the antibody variant pair. For the 14 antibody-variant pairs in the dataset, the DLAB-VS+ZDOCK score of the antibody–wild-type pair was higher than the score of the antibody-variant pair in 13/14 cases. This result indicates that the variant classification problem is accessible to structure-based deep learning tools.

5 Discussion

One of the major shortcomings of current computational antibody drug discovery is the lack of structure-based, early-pipeline screening tools to identify promising candidate antibodies.

Here, we have shown how DLAB, our structure-based deep learning approach, can be used to improve pose selection in antibody–antigen docking experiments and can enable the identification of antibody–antigen pairings for which accurate poses have been generated and selected. DLAB-Re is able to identify pairings for which a pose with *fnat* > 0.5 is in the top ten poses for 84% of the pairings, which can be used to improve binder classification performance downstream.

We have furthermore demonstrated that our DLAB tool can identify putative binders to a given epitope in several different settings. The complete DLAB pipeline of docking followed by DLAB-Re and DLAB-VS enriched binders both against the background of non-binding SAbDab-deposited sequences as well as in a more realistic usage scenario against H3 length-matched antibody sequences drawn from antibody repertoire data.

On the crystal dataset with highly accurate antibody structures and docking poses, both DLAB-VS and ZDOCK are able to strongly enrich binders. In the case of model antibodies docked to antigens, where both model and docking quality have to be considered, ZDOCK and DLAB-VS approaches fail to achieve strong discrimination between binders and non-binders. However combining the two scores improved performance. These results are in line with previously published work on the ability to classify cognate antibodies through cross-docking analysis (Kilambi and Gray, 2017).

On a realistic use case, using the SARS-CoV2 receptor binding domain as the target antigen, we have demonstrated the utility of the DLAB pipeline, correctly scoring antibody escape variants lower than the cognate epitopes for 13 of 14 antibody-variant pairs.

The DLAB pipeline has been trained specifically on a combination of ABodyBuilder and ZDOCK. The use of a different input pipeline would likely require additional finetuning of the weights of both DLAB-Re and DLAB-VS. One natural extension would be the use of flexible docking approaches, which could improve the input docking poses but would be computationally expensive given the scale of experiments needed in a high-throughput setting.

We have demonstrated the applicability of structure-based deep learning approaches both to antibody research in general and to the virtual screening task specifically. Methods such as DLAB will improve with increasing availability of structural antibody data as well as improved antibody modelling and improved fast docking

methods. DLAB demonstrates the potential of structure-based deep learning approaches to supplement traditional experimental screening approaches and sets a course for structure-based virtual screening methods for antibody drug discovery.

Funding

This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and the Medical Research Council (MRC) [EP/L016044/1]; and AstraZeneca.

Conflict of Interest: none declared.

Data availability

The DLAB source code and pre-trained models are available at https://github.com/oxpig/dlab-public. The data used to create train and test sets are available in the online supplementary material or were derived from a source in the public domain: SAbDab (http://opig.stats.ox.ac.uk/webapps/newsabdab/sabdab/).

References

Adolf-Bryfogle,J. *et al.* (2018) RosettaAntibodyDesign (RAbD): a general framework for computational antibody design. *PLoS Comput. Biol.*, **14**, e1006112.

Akbar,R. *et al.* (2021) A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, **34**, 108856.

Almagro,J.C. *et al.* (2014) Second Antibody Modeling Assessment (AMA-II). *Proteins Struct. Funct. Bioinf.*, **82**, 1553–1562.

Almagro,J.C. *et al.* (2017) Progress and challenges in the design and clinical development of antibodies for cancer therapy. *Front. Immunol.*, **8**, 1751.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bujotzek,A. *et al.* (2015) Prediction of VH-VL domain orientation for antibody variable domain modeling. *Proteins Struct. Funct. Bioinf.*, **83**, 681–695.

Chiu,M.L. and Gilliland,G.L. (2016) Engineering antibody therapeutics. *Curr. Opin. Struct. Biol.*, **38**, 163–173.

Dunbar,J. *et al.* (2014) SAbDab: the structural antibody database. *Nucleic Acids Res.*, **42**, D1140–D1146.

Graves,J. *et al.* (2020) A review of deep learning methods for antibodies. *Antibodies*, **9**, 12.

Grilo,A.L. and Mantalaris,A. (2019) The increasingly human and profitable monoclonal antibody market. *Trends Biotechnol.*, **37**, 9–16.

Gu,K. *et al.* (2008) Testing the ratio of two Poisson rates. *Biometrical J.*, **50**, 283–298.

Imrie,F. *et al.* (2018) Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *J. Chem. Inf. Model.*, **58**, 2319–2330.

Kilambi,K.P. and Gray,J.J. (2017) Structure-based cross-docking analysis of antibody–antigen interactions. *Sci. Rep.*, **7**, 8145.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.

Leem,J. *et al.* (2016) ABodyBuilder: automated antibody structure prediction with data–driven accuracy estimation. *mAbs*, **8**, 1259–1268.

Leem,J. *et al.* (2018) Antibody side chain conformations are position-dependent. *Proteins Struct. Funct. Bioinf.*, **86**, 383–392.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Liberis,E. *et al.* (2018) Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics*, **34**, 2944–2950.

Liu,J.K.H. (2014) The history of monoclonal antibody development – progress, remaining challenges and future innovations. *Ann. Med. Surg.*, **3**, 113–116.

Mason,D.M. *et al.* (2021) Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, **5**, 600–612.

Norman,R.A. *et al.* (2020) Computational approaches to therapeutic antibody design: established methods and emerging trends. *Brief. Bioinf.*, **21**, 1549–1567.

Olimpieri,P.P. *et al.* (2013) Prediction of site-specific interactions in antibody–antigen complexes: the proABC method and server. *Bioinformatics*, **29**, 2285–2291.

Pierce,B.G. *et al.* (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS One*, **6**, e24657.

Pittala,S. and Bailey-Kellogg,C. (2020) Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*, **36**, 3996–4003.

Ragoza,M. *et al.* (2017) Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.*, **57**, 942–957.

Raybould,M.I. *et al.* (2019a) Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Mol. Syst. Des. Eng.*, **4**, 679–688.

Raybould,M.I.J. *et al.* (2019b) Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl. Acad. Sci. USA*, **116**, 4025–4030.

Raybould,M.I.J. *et al.* (2020) CoV-AbDab: the coronavirus antibody database. *Bioinformatics*, **37**, 734–735.

Scantlebury,J. *et al.* (2020) Data set augmentation allows deep learning-based virtual screening to better generalize to unseen target classes and highlight important binding interactions. *J. Chem. Inf. Model.*, **60**, 3722–3730.

Schymkowitz,J. *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.

Seabold,S. and Perktold,J. (2010) statsmodels: econometric and statistical modeling with python. In: *9th Python in Science Conference*. Austin, Texas, USA.

Suscovich,T.J. and Alter,G. (2015) In situ production of therapeutic monoclonal antibodies. *Exp. Rev. Vaccines*, **14**, 205–219.

Virtanen,P. *et al.*; SciPy 1.0 Contributors. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, **17**, 261–272.

Wang,X. *et al.* (2020) Protein docking model evaluation by 3D deep convolutional neural networks. *Bioinformatics*, **36**, 2113–2118.

Weitzner,B.D. *et al.* (2017) Modeling and docking of antibody structures with Rosetta. *Nat. Protoc.*, **12**, 401–416.

Wodak,S.J. and Méndez,R. (2004) Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.*, **14**, 242–249.

Wong,W.K. *et al.* (2019) Comparative analysis of the CDR loops of antigen receptors. *Front. Immunol.*, **10**, 2454.

Yamashita,K. *et al.* (2014) Kotai antibody builder: automated high-resolution structural modeling of antibodies. *Bioinformatics*, **30**, 3279–3280.