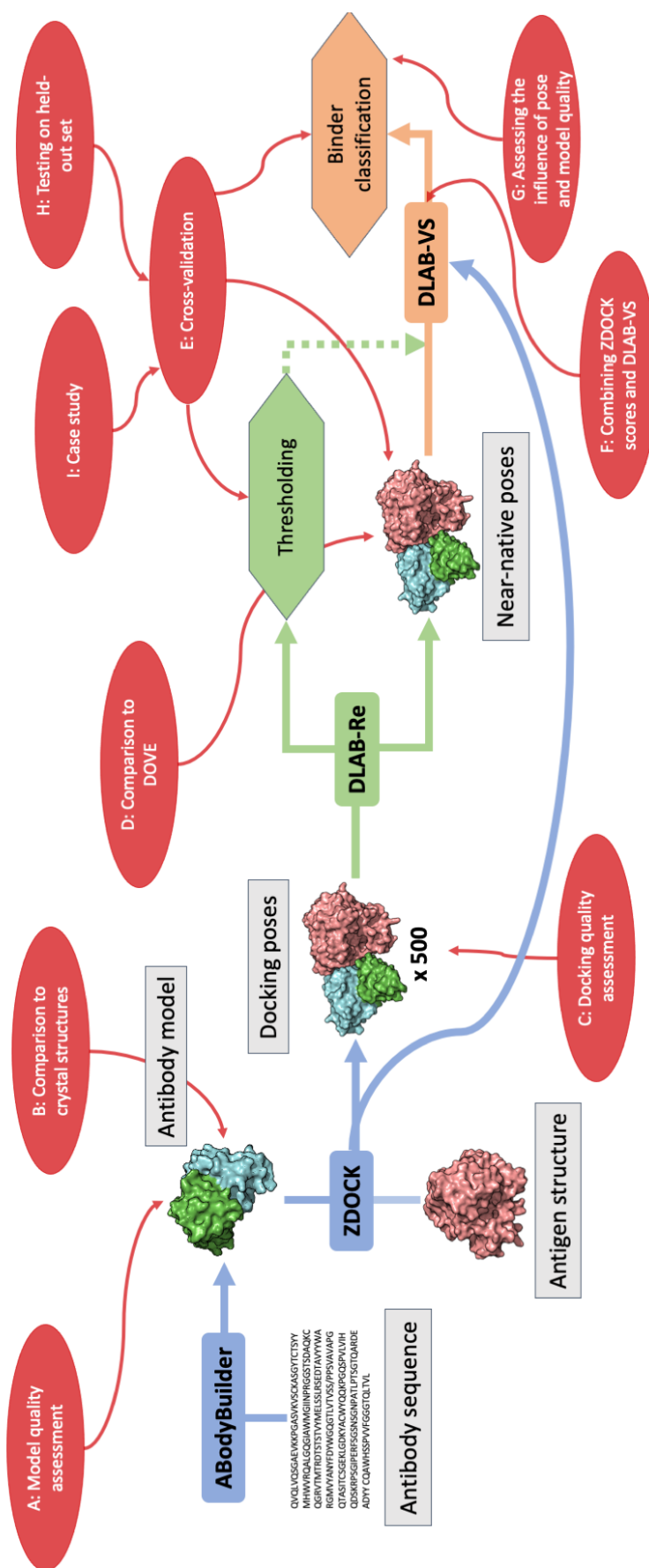
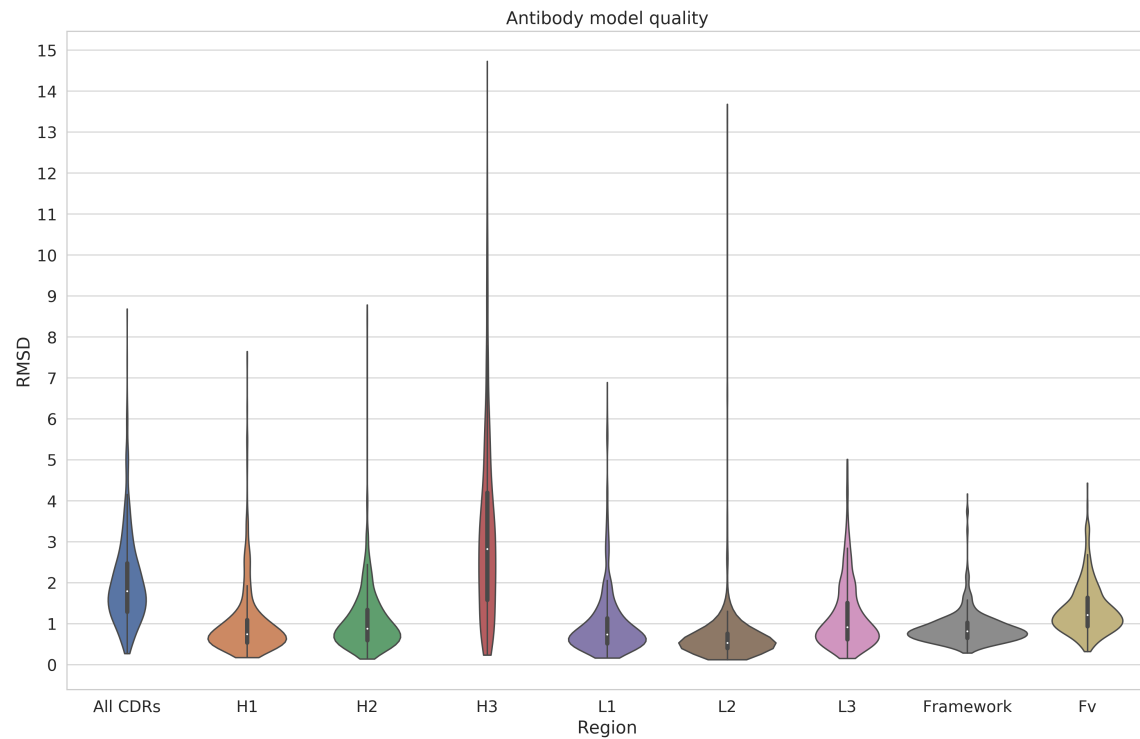
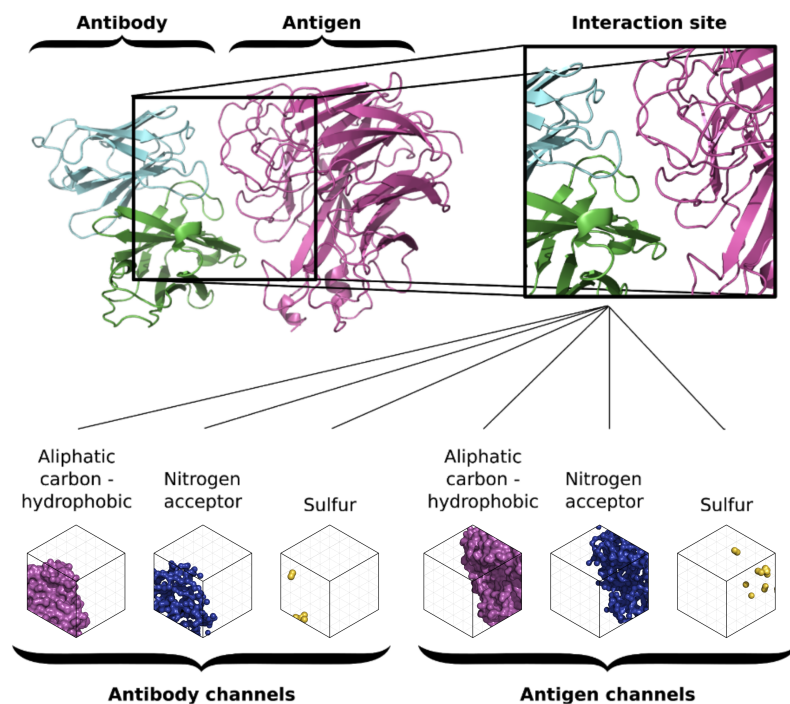

Supplementary Figures



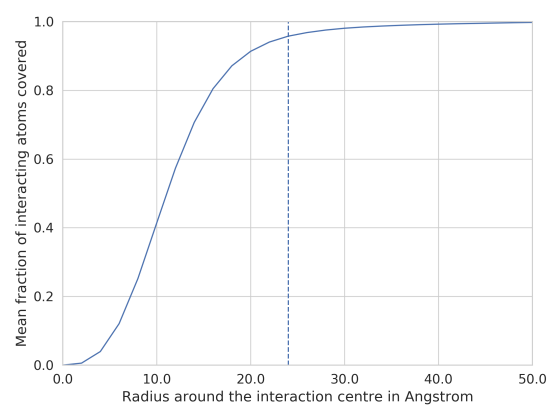
Supplementary Figure 1: Flowchart depicting the DLAB pipeline and considerations during the development of the final DLAB ensemble model. Blue boxes/arrows indicate external tools, green boxes/arrows indicate where DLAB-Re fits in the pipeline and similarly, orange boxes/arrows indicate DLAB-VS. Red boxes/arrows and the corresponding labels highlight the points of the pipeline analysed during different stages of the model development process and are referred to throughout the main text body.



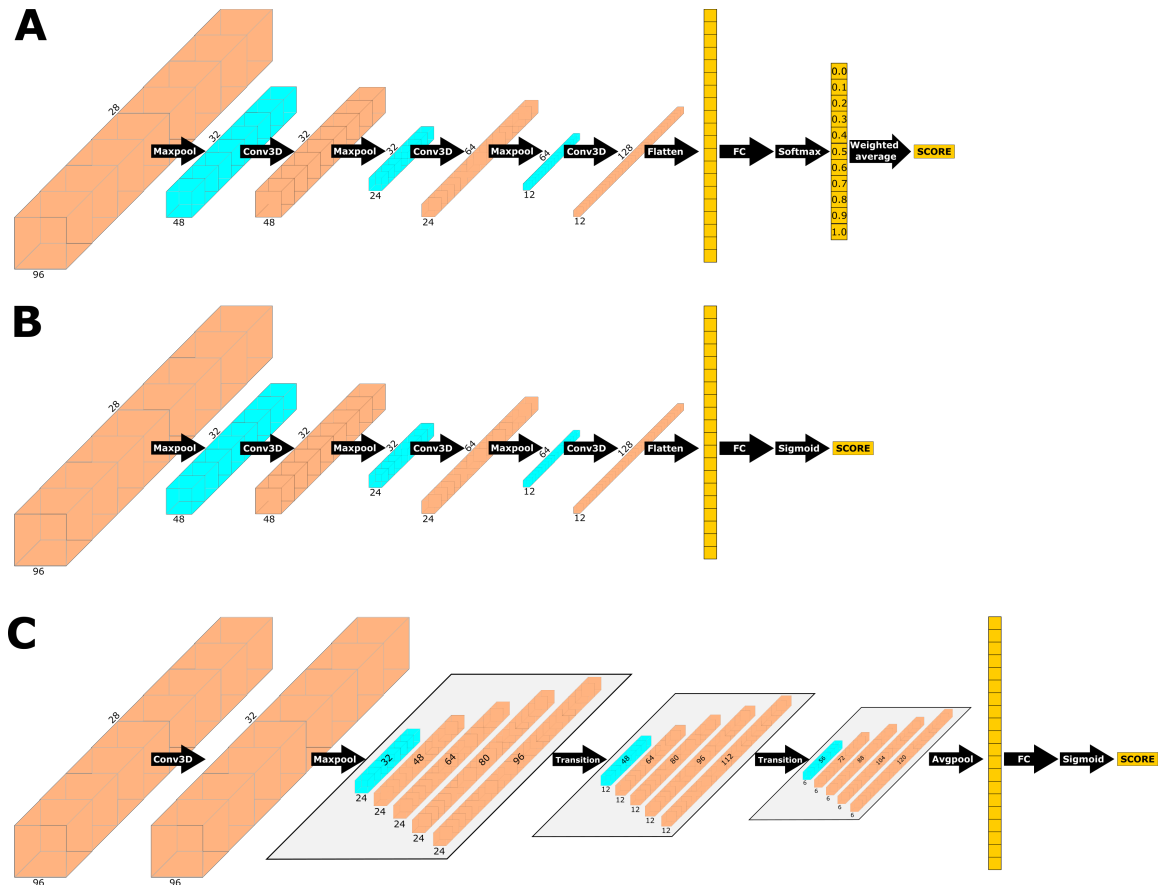
Supplementary Figure 2: Assessment of the ABodyBuilder generated antibody models. RMSD values across regions of interest on the antibody are calculated as described in the Methods section of the main paper.



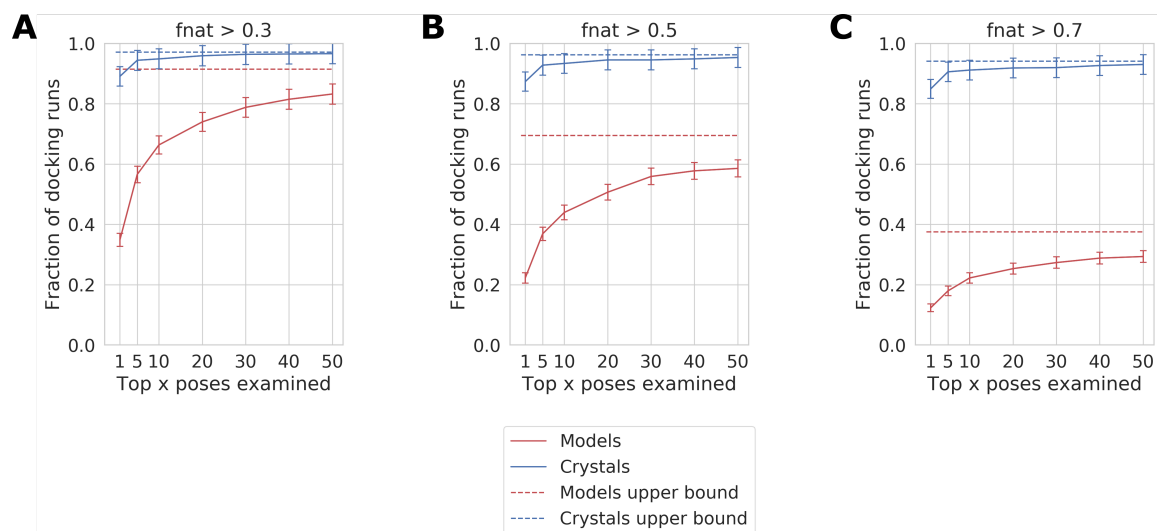
Supplementary Figure 3: The binding interface discretisation algorithm. (**Top**) The interface is defined as a 48 Å cube centered on the interaction center between antibody and antigen. (**Bottom**) The atoms occupying the interface in the docking pose are discretised using the libmolgrid python API into 3D arrays of atom type densities (14 types for both antibody and antigen, of which 3 are depicted here as an example).



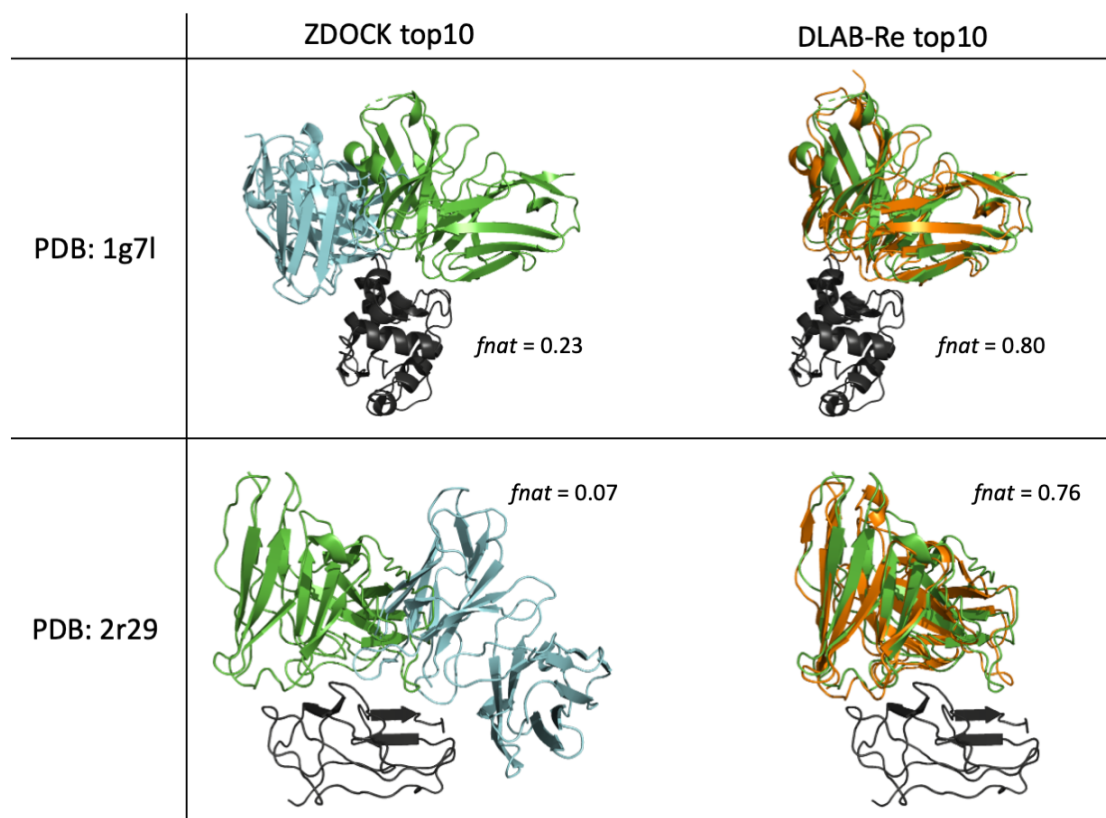
Supplementary Figure 4: The average fraction of interacting atoms (between the antibody and antigen) included in the interaction box depending on the radius around the interaction center. The dashed vertical line indicates 24 Å, the radius used in this study.



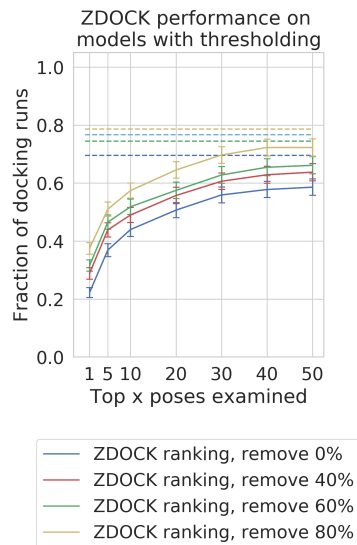
Supplementary Figure 5: CNN architectures used in the study. **(A)** DLAB-Re architecture. A 3-layer CNN followed by a fully connected layer is used to predict membership of one of 11 *fnat* intervals. To make *fnat* predictions, the output of the softmax layer is transformed into a single score via weighted averaging. **(B, C)** The DLAB-VS architectures forming the ensemble of DLAB-VS models. Both architectures output a binary binder/non-binder classification for an antibody/antigen pairing. **(B)** The same CNN architecture as in (A) with a different output layer. **(C)** A CNN architecture consisting of a single convolutional layer followed by 3 Denseblocks, using the same output layer as (B).



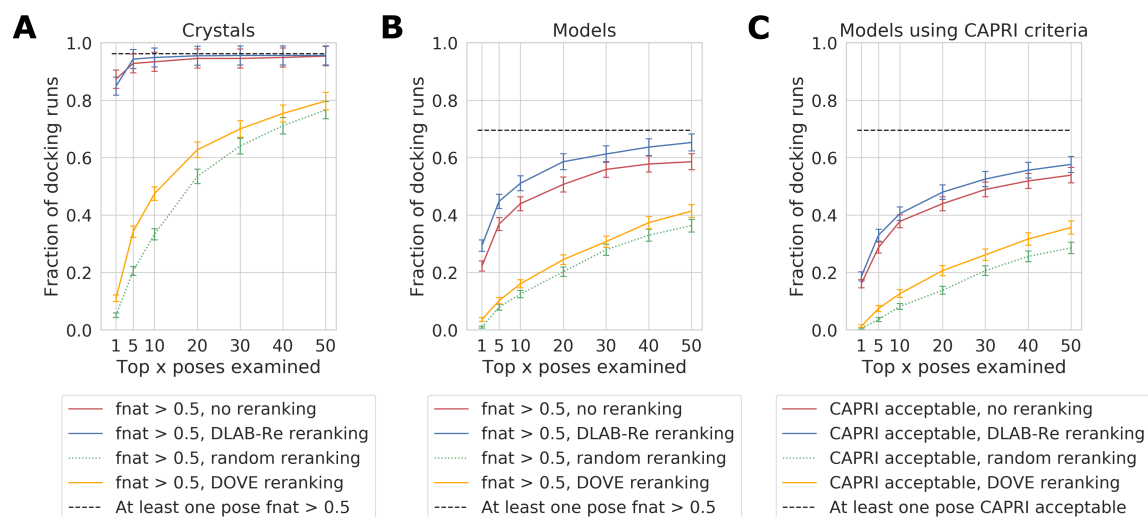
Supplementary Figure 6: Comparison of ZDock runs on crystal structures and models. Each of the three plots shows the fraction of antibody-antigen pairings which have at least one pose meeting the threshold $fnat$ score 0.3 (A), 0.5 (B) or 0.7 (C) in their x highest ranked docks. The dashed line indicates the fraction of docks which have such a pose in their top500 poses (the upper bound of the pose ranking algorithm).



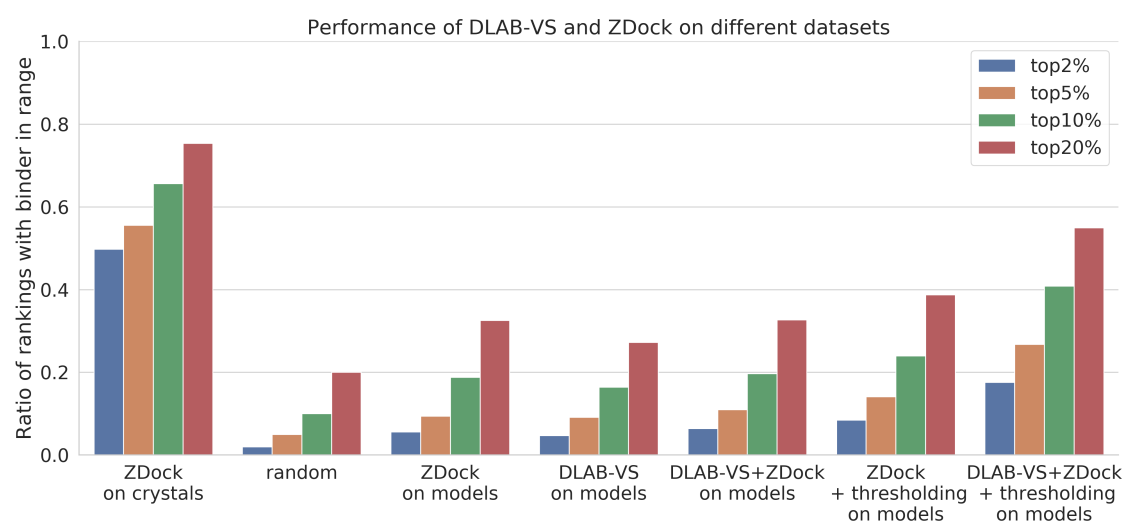
Supplementary Figure 7: Two examples for which DLAB-Re strongly improves pose ranking for modelled antibodies. (Left) Pose with the highest *fnat* in the top 10 poses as ranked by ZDOCK. (Right) Pose with the highest *fnat* in the top 10 poses as ranked by DLAB-Re. The target antigen is depicted in black, the experimentally determined bound antibody structure in green, the pose with the highest *fnat* in the top 10 poses as ranked by ZDOCK in cyan and the pose with the highest *fnat* in the top 10 poses as ranked by DLAB-Re in orange.



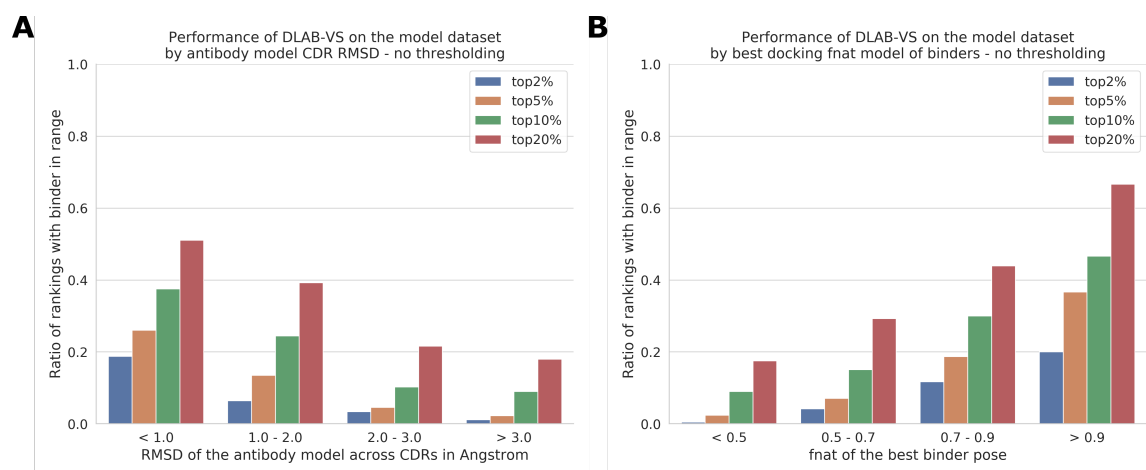
Supplementary Figure 8: Performance of ZDock-based max score thresholding on ranking performance. While higher overall ZDock scores correlate with improved ranking performance, the effect is much less pronounced than using DLAB-Re-max scores. The solid lines indicate the fraction of antibody-antigen pairings which have at least one pose with *fnat* over 0.5 in their *x* highest ranked docks, the dotted line indicates the fraction of pairings for which a pose with *fnat* over 0.5 exists in top 500 poses.



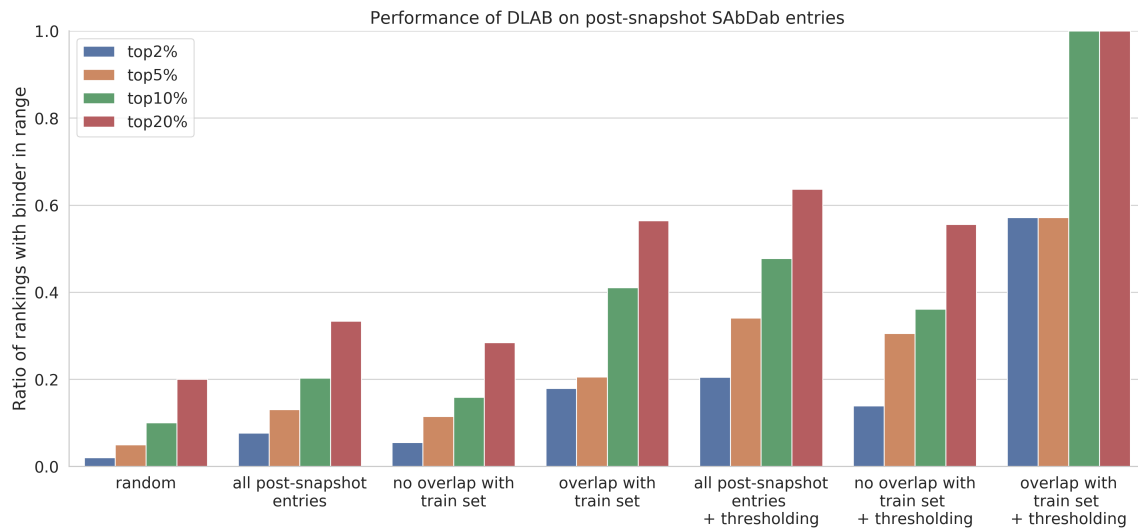
Supplementary Figure 9: Comparison of DLAB-Re performance to DOVE performance. (A, B) Comparison using the pose f_{nat} scores on the crystal structure data set (A) and the model data set (B). The fraction of antibody-antigen pairings which have at least one pose with f_{nat} over 0.5 in their x highest ranked docks for ZDock ranking ("no reranking") as well as DOVE and DLAB-Re reranking is shown in solid lines, the baseline of randomly shuffling the top 500 poses is shown with a dotted green line and the ratio of pairings with a pose with f_{nat} over 0.5 is indicated with a dotted black line. (C) Comparison using the CAPRI acceptability criterion for each pose. Legend explanation see (A, B). In all three cases (A, B, C), DLAB-Re achieves an improvement over native ZDock, while DOVE does not replicate the ranking performance achieved by ZDock.



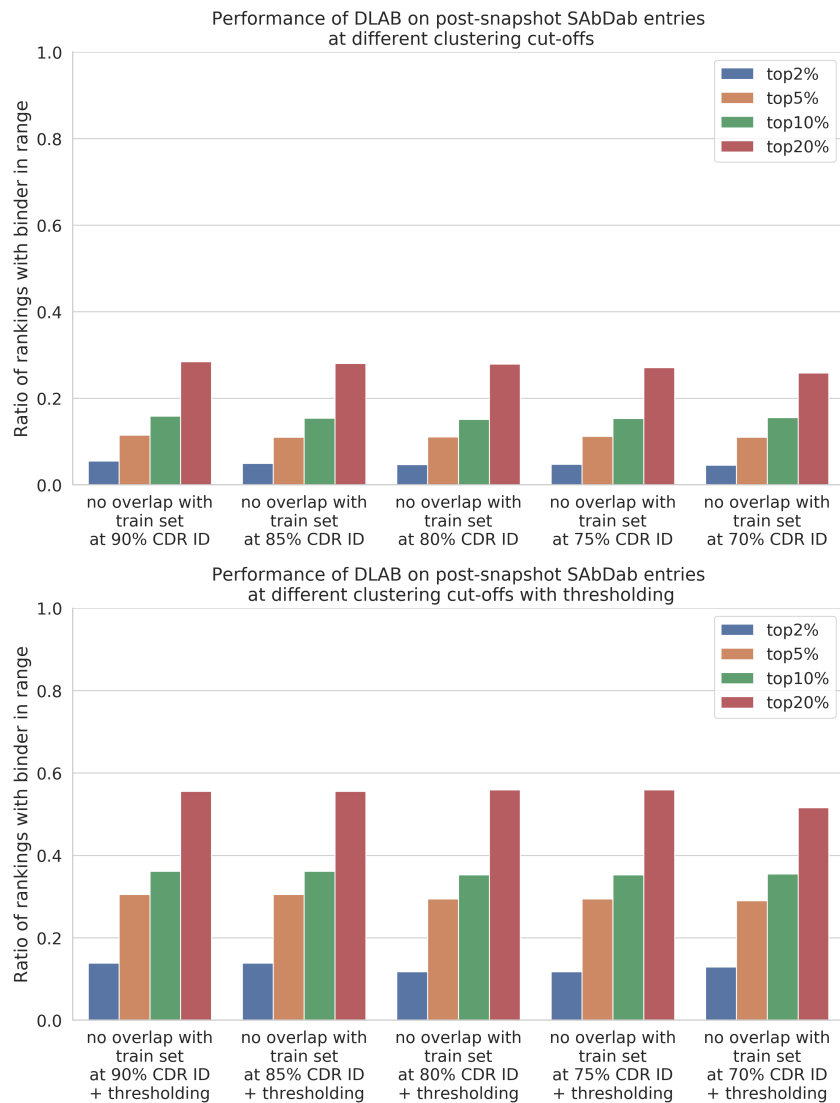
Supplementary Figure 10: DLAB-VS and ZDock binder classification performance on different data sets. For each approach, the ratio of pairings for which the binding antibody was ranked in the top 2%, top 5%, top 10% and top 20% respectively is shown. Comparison of the performance of ZDock binder classification on the crystal data set to the random expectation ("random") of finding the binder in the top N% and the performance of ZDock, DLAB-VS and the combination of ZDock and DLAB-VS generated as detailed in the Methods section ("DLAB-VS+Zdock") on the model data set, both with and without using the DLAB-Re-max thresholding approach (" + thresholding").



Supplementary Figure 11: Dependence of the performance of DLAB-VS+ZDock on antibody model quality (A), measured via RMSD of the CDR regions to the corresponding crystal structure, and docking quality (B), measured via the highest fnat achieved in the top500 docking poses generated by ZDock. Both good antibody models and high quality docking poses correlate with high classification performance.



Supplementary Figure 12: Full overview of the performance of DLAB-VS+ZDock on the post-snapshot model data set. As in the main text figure, the bars show the fraction of antigens within the data set for which the correct binder is ranked by DLAB-VS within the top 2%, top 5%, top 10% or top 20% respectively. We compare the random expectation value ("random") to the whole post-snapshot model set ("all post-snapshot entries"), the antigen targets within the set for which the binding antibody does ("overlap with train set") or does not ("no overlap with train set") cluster with at least one antibody from the model data set at 90% CDR sequence identity. For each of these three options, the improved performance upon using the DLAB-Re-max score to discard 80% of antigen targets as described above is shown as well (" + thresholding").



Supplementary Figure 13: Performance of DLAB-VS+ZDOCK on the post-snapshot dataset at different percentage CDR sequence identity cutoffs for overlap with the training set without (Top) and with (Bottom) DLAB-Re thresholding. The performance of the DLAB-VS+ZDOCK model only marginally declines when limiting the allowed overlap to training set antibodies from 90% CDR ID to 70% CDR sequence identity, demonstrating that the model performance is generalisable.

Antibody name	Non-cognate RBD variant mutations	PDB code of complex	Reference
AB1	E406W N487R K417N, E484K, N501Y	7mjl	Sun <i>et al.</i> (2021)
S2X259	G504D	7m7w	Tortorici <i>et al.</i> (2021)
BG10-19	K417N, E484K, N501Y	7m6e	Scheid <i>et al.</i> (2021)
BG1-22	K417N, E484K, N501Y	7m6f	Scheid <i>et al.</i> (2021)
BG4-25	K417N, E484K, N501Y	7m6d	Scheid <i>et al.</i> (2021)
S2-M28	L18F, R246I, D80A	7ly0	McCallum <i>et al.</i> (2021)
S2-L28	L18F, R246I, D80A, D215G L18F R246G D253G D253Y	7lxx	McCallum <i>et al.</i> (2021)
S2-X333	L18F, R246I, D80A	7lxx	McCallum <i>et al.</i> (2021)

Supplementary Table 1: Overview of the SARS-CoV2 variant binding dataset. Note that 7mjl was backmutated Y501N to restore wild-type, as no complex structure against the wild-type RBD was available. The triple mutant K417N, E484K, N501Y characterises the variant of concern B.1.351