

Supplementary Materials

VCAb: A web-tool for structure-guided antibody engineering

Dongjun Guo^{1,2,+}, Joseph Chi-Fung Ng^{1,+}, Deborah K. Dunn-Walters³, and Franca Fraternali^{1,*}

¹*Institute of Structural and Molecular Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, United Kingdom*

²*Randall Centre for Cell and Molecular Biophysics, School of Basic and Medical Biosciences, King's College London, London SE1 1UL, United Kingdom*

³*School of Biosciences and Medicine, University of Surrey, Guildford GU2 7XH, United Kingdom*

⁺*Shared first authorship.*

^{*}*Corresponding author. f.fraternali@ucl.ac.uk*

Contents

S1 Description of functionalities	S2
S1.1 Visualization of antibody information	S2
S1.1.1 Sequence	S2
S1.1.2 Species and isotype	S3
S1.1.3 Antibody structure metadata	S3
S1.1.4 Antibody three-dimensional structures	S3
S1.2 Analysis of CH1-CL interface	S3
S1.2.1 Visualization of CH1-CL interface residues	S3
S1.2.2 Search structures by CH1-CL interface similarity	S3
S1.3 Search structures by antibody sequence similarity	S4
S1.4 Map repertoire sequence to experimental structure	S4
S1.5 Access pre-computed <i>in silico</i> mutational scanning data	S7
S1.5.1 Visualization of the <i>in silico</i> mutational scanning results	S7
S2 Case studies	S7
S2.1 Interface Investigation	S7
S2.2 <i>In silico</i> mutational scanning profiles	S9
S2.3 Isotype sensitivity analysis using <i>in silico</i> mutational scanning data	S9
S3 Methods	S10
S3.1 Data collection and annotation	S10
S3.1.1 Collecting Antibodies	S10
S3.1.2 Species, isotype and structural coverage annotation	S10
S3.1.3 Antibody C region numbering	S12
S3.1.4 CH1-CL interface analysis	S13
S3.1.5 Elbow angle calculation	S14
S3.1.6 Disulphide bond calculation	S14
S3.2 The generation of <i>in silico</i> mutational scanning profiles	S14
S3.2.1 Rosetta point mutant scan	S14
S3.2.2 AntiBERTy pseudo-log likelihoods	S15
S3.2.3 AlphaMissense pathogenicity score	S16
S3.3 VCAb webserver	S17

S1 Description of functionalities

S1.1 Visualization of antibody information

Each Antibody entry in the VCAb containing features of four aspects: sequence, species, isotype, structure metadata, and three-dimensional structure.

User can customise the table view to include columns being displayed, with bulk downloads of search results and/or the entire VCAb database in Comma Separated Value (CSV) format enabled.

User can also query VCAb by filtering features of their interest. This could be the starting point of antibody engineering for the selection of potential structures as the template for further modification. Entries matching the user's query are displayed in the antibody information panel. User can download these compiled data tables, visualise and inspect the selected antibody structure and interactively browse the CH1-CL interface residues in the 3D structure viewer, with all CH1-CL interface residues tabulated, Fab contact matrix displayed, and *in silico* mutational scanning profiles are shown.

All the features listed below can be accessed via VCAb webserver. A list of detailed information included in VCAb can be found in the online documentation accessible in the "About" page of VCAb.

S1.1.1 Sequence

The sequences for each antibody entry are provided and downloadable to the user. We also list the sequence represented in the PDB ATOM records. This reflects the true coverage of the PDB structure, and is therefore relevant in structural investigations (e.g. in prioritising templates to be used to model antibody structures). Users can compare the domain coverage of the sequence downloaded from PDB API and the sequence representing PDB ATOM records by clicking the "Sequence Coverage" tab (**Figure S1**).

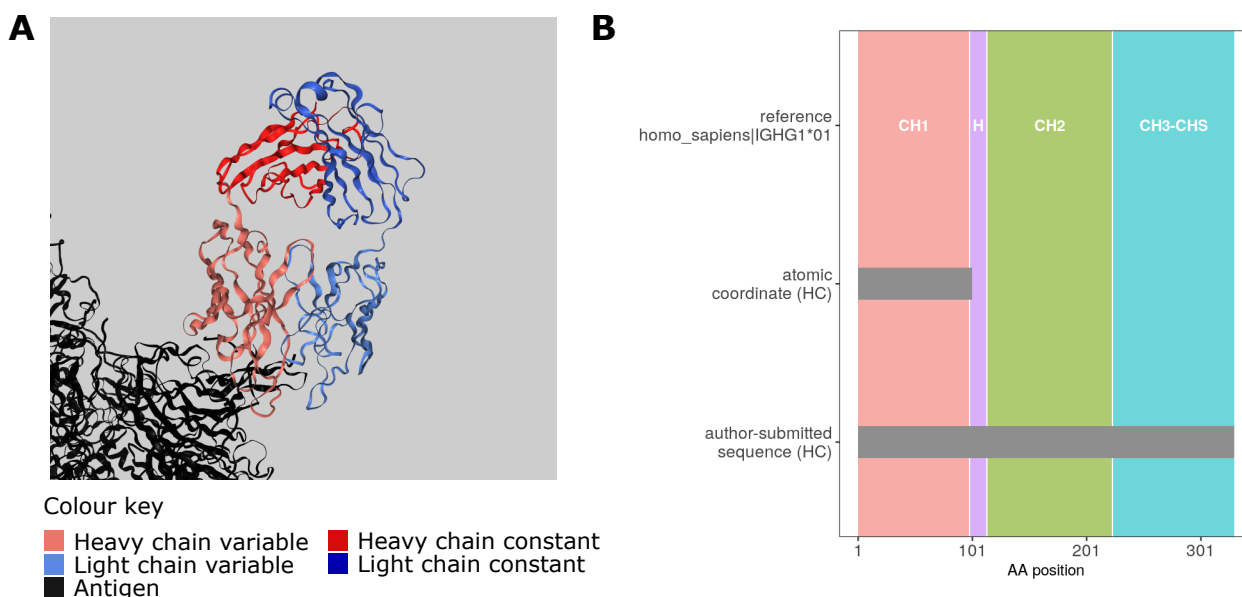


Figure S1: SEQRES records represents the structural coverage of the antibody. (A) Antibody 7c2l displayed in the 3D structural viewer of VCAb. The heavy chain only contains two domains, with one being constant region. (B) The SEQRES records for the C region are compared with the author-submitted sequence, with both sequences aligned to the reference allele. By downloading the H chain sequence of 7c2l from the Protein Data Bank (PDB), one can see that includes 3 constant domains (CH1, CH2, and CH3, H is the flexible hinge connecting CH1 and CH2). In contrast, the SEQRES records only cover two domains, reflecting its actual structural coverage as depicted in (A).

Each sequence is numbered based on IMGT scheme (Lefranc *et al.*, 2005) by ANARCIvc package we developed based on ANARCI (Dunbar and Deane, 2016). The numbered sequence is displayed in the "Sequence numbering" tab panel, allowing for the download of the numbered sequence accessible. The user can hover on the residue in the figure to show the numbering information, including residue name, IMGT numbering and the position of the residue (V or C region, and the fragment where this residue belonging to, such as framework or CDR loops). The user can switch between heavy and light chain sequences by selecting the options under "Choose the sequence to display".

S1.1.2 Species and isotype

We annotate the most likely species and isotype for each VCAb structure as described in section S3.1.2. This information together with the mapped allele(s) are listed when users query entries in VCAb. Alternative chain types (i.e. isotype and light chain type) would be listed in a separate column if they have the same BLAST identity with the top hit.

S1.1.3 Antibody structure metadata

The VCAb webserver provides tabulated details of experimental methods and resolution (downloaded from PDBe API), structural coverage (Fab or full-antibody), H-L interface, disulphide bonds, geometric angles (CH1-CL angles and elbow angles, section S3.1.4, S3.1.5), and whether the structural entry contains antigen. Hyperlinks for the corresponding PDB or SAbDab webpages are also included for further inspection.

S1.1.4 Antibody three-dimensional structures

Three-dimensional structural viewer is integrated in VCAb by using NGLViewer (version 1.3.2) in R. User can rotate, zoom in/out the antibody structures in the three-dimensional structure viewer. By default, only one heavy/light chain pairs would be presented. Other chains can be presented by clicking the “view all the chains with the same PDB ID” button, with the antigen chain (if any) colored in black (section S3.1.1) and other ligand(s) (colored cyan).

S1.2 Analysis of CH1-CL interface

S1.2.1 Visualization of CH1-CL interface residues

VCAb supports interactively investigation of interface residues by two approaches. By selecting a region of interest on the domain contact map of Fab region calculated for each antibody structure, interface residues in the selected box can be automatically zoomed in, highlighted in the structural viewer and allowing for a closer inspection (**Figure S2**). User can also select individual interface residues tabulated in the “H-L interface residues” tab to achieve the same effect.

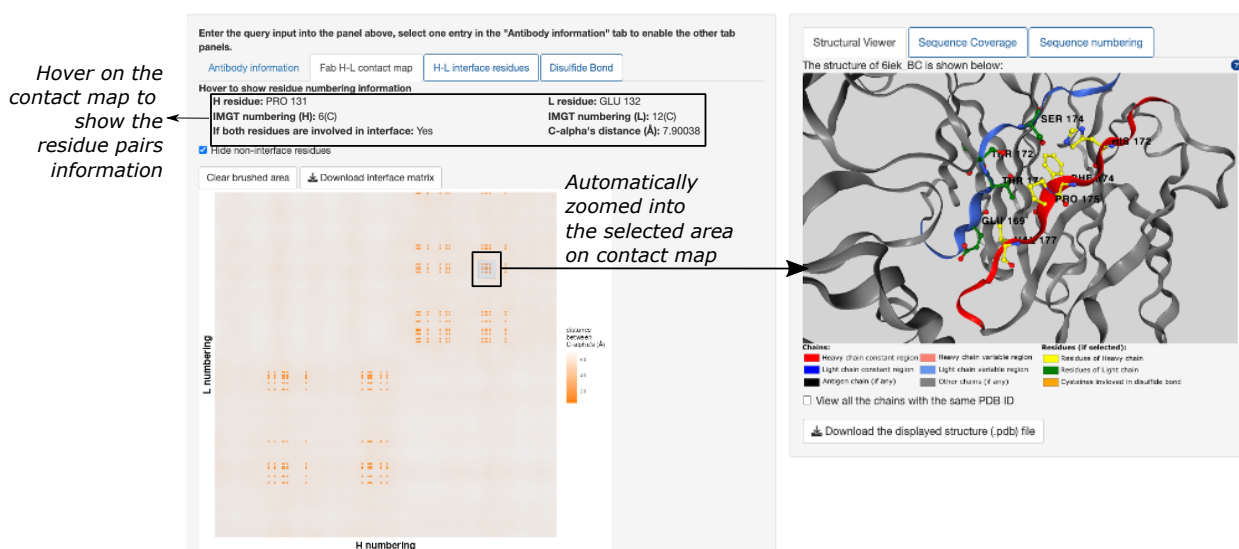


Figure S2: Interactive domain contact map and three-dimensional structural viewer. For each VCAb structure, both the Fab H-L contact map and the three dimensional structure are displayed. The x and y axis in Fab H-L contact map represent heavy and light chain residues numbered conforming to IMGT scheme, respectively. Users can highlight residue pairs where both of them are involved in the interface by clicking “hide non-interface residues” and show the residue information by hovering onto the contact map. The 3D structural viewer will be automatically zoomed into the selected area on the Fab contact map.

S1.2.2 Search structures by CH1-CL interface similarity

Taking advantage of the the CH1-CL interface analysis we performed over all VCAb structures (section S3.1.4), users can search VCAb according to CH1-CL interface similarity. In “CH1-CL Interface” panel,

users can input the PDB code and chain IDs of their antibody structure of interest. VCAb entries will then be ranked by calculated IntDiff score (equation 4), with the smaller value indicating more similar CH1-CL interface.

S1.3 Search structures by antibody sequence similarity

VCAb web tool enables the filtering of antibody structures by sequence similarity, using BLAST (Altschul *et al.*, 1990) provided in the rBLAST package (version 0.99.2). This search is flexible to the region of interest (V region or both V & C regions) and both paired and unpaired H/L chain sequences. As described in section S3.1.2, sequences for V regions are extracted from the BLAST alignments and collected. Together with the sequences of the VCAb entries covering both V and C regions, four databases are formed, containing the sequences of all VCAb entries covering (i) VH domains, (ii) VL domains, (iii) full-length H chains and (iv) L chains, respectively. The webserver automatically chooses the database for the purpose of BLAST searches, according to the region of interest and the searching mode (paired or unpaired H and L chains) selected by the user (**Table S1**).

Table S1: The selection of database for BLAST implemented in the VCAb application. “VH or VL” means selecting VH or VL VCAb sequences as the BLAST database, according to the chain type (H or L) selected/labeled by user. “VH and VL” means the database combining VH and VL sequences.

	Unpaired		Paired	
	V region	Full sequence (V & C)	V region	Full sequence (V & C)
Search individual sequence	VH or VL	H or L	VH or VL	H or L
Search in batch	VH and VL	H and L	VH or VL	H or L

VCAb supports input of single sequence and batch uploads of sequences. If individual sequence is provided, VCAb entries will be ranked by sequence identity, and the IMGT numbering for the user-inputted sequence would also be displayed. If an individual sequence of the V region is inputted, the C region of selected structural hits can be appended for further antibody design and/or lab investigations, with the appended sequence numbered under the IMGT-scheme and downloadable.

VCAb webserver enables batch searching of sequence similarity, by allowing the user to upload a FASTA file containing multiple sequences (maximum 200 per batch). For each query sequence in the uploaded file, the top three VCAb entries given by BLAST would be listed. Under the paired mode, users need to indicate the pairing between individual H and L chains by following the format to specify the FASTA header of the sequences: AbName-HorL, where AbName should be the name of the antibody, and HorL should be either letter “H” or “L” to specify H or L chain. Paired chains can contain only one H and one L chain. Unpaired sequences would not be annotated; they are tabulated below the results of searching the paired chains.

S1.4 Map repertoire sequence to experimental structure

Currently, VCAb supports the upload of repertoire files and searching of matching experimentally determined antibody structures based on sequence homology. Uploaded repertoire data can be standard (AIRR, 10X Genomics CellRanger) or customised formats. When repertoire data in customised format is provided, columns holding the amino acid sequence, identity of the chain (heavy or light chain), and cell barcode should be specified, while these columns are automatically selected when standard repertoire formats are uploaded (**Table S2**). Repertoire file can be uploaded to VCAb through multiple ways: a) browsing for a local repertoire file and upload it to the server, b) live imports from online analyses on the BRepertoire (Margreitter *et al.*, 2018) webserver, or c) uploading a remote repertoire file by providing a http address. User can select different mode in searching: unpaired (individual sequence selected from sampled repertoire) or paired mode (paired heavy and light chains of the same antibody). In the unpaired mode, the selected sequence entry will be entered into the BLAST(Altschul *et al.*, 1990) pipeline, while the latter mode automatically picks up the other paired H/L chain indicated by the same cell barcode and inputs the paired chains into the BLAST process.

Structural hits will be ranked by percentage identity of the amino acid sequence to the selected individual or paired chains in uploaded repertoire. Antibody and structural information for each structural hits will be displayed, with its three-dimensional structure visualised and inspected (section S1.1).

As an example, we showed structure matches of the two sequences we inputted from a published antibody sequence repertoire (Stewart *et al.*, 2022):

Table S2: Columns required for repertoire search for structural match

Columns	AIRR	10X Genomics CellRanger	customised formats
Amino acid sequence	automatically selected <i>Column: sequence_alignment_aa</i>	automatically selected <i>Columns: fwr1, cdr1, fwr2, cdr2, fwr3, cdr3, fwr4</i> ; sequence fragments in these columns are automatically concatenated together for search	Multiple columns can be selected here, sequence fragments in these columns will be concatenated according to the input order
Identity of the chain	automatically selected <i>Column: locus</i>	automatically selected <i>Column: chain</i>	column values can only contain "IGH", "IGK", or "IGL"
Cell ID/ Cell barcode <i>* this column is only required under paired searching mode</i>	automatically selected <i>Column: cell_id</i>	automatically selected <i>Column: barcode</i>	column values contain unique cell ID for each cell

- The top structural match (PDB 7tn0) of the repertoire sequence with identifier "CV222D2_M_10050314_TAACCTTCTGTTATAT_RD39" indicates antibody recognise RBD epitope distinct from ACE2 binding site.
- Another sequence with identifier "CV227D9_G1_10089371_TTCGGTCACGTGAATT_RD30" is also inputted to VCAb to search for structural match (PDB 7bz5), which shows this antibody directly blocks ACE2 binding.

A

Query for antibody structures

PDB Features Sequence CH1-CL Interface Repertoire

Upload repertoire table

Browse... sample_repertoire_COVID.txt Upload complete

Or input a http address for repertoire file here:

Load the example file for different format: AIRR example Cell ranger example customized repertoire file

Clear uploaded repertoire file (url)

Please select the format of the repertoire file:

AIRR

Cell ranger

Customized

Please select the searching mode:

Unpaired chains

Paired H-L chains

Choose column(s) holding amino acid sequences:

ASeq

Choose column indicating the identity of the chain (IGH/IgK/IgL):

Vfamily

Please select one entry (by clicking the row of the table) of the uploaded repertoire file to start search

Search

Show 10 entries

Seq_ID	AA-JUNCTION	Class	Subclass	CellType	Vfamily	Vgene	Jfamily	Dfamily	Dgen
5691	CV222D2_M_10047871_TAAGTATATGCGCT_RD52	M	IgM	IgMemory	IGHV3	IGHV3-74	IGHJ4	IGHD6	IGHD13
5692	CV222D2_M_10050314_TAACCTTCTGTTATAT_RD39	M	IgM	IgMemory	IGHV3	IGHV3-13	IGHJ4	IGHD3	IGHD22
5693	CV222D2_A1_10069919_TCGAATAACTGCGAT_RD18	A	IgA1	IgA1	IGHV3	IGHV3-21	IGHJ3	IGHD2	IGHD6
5694	CV222D2_A1_10199943_TAAATCATTAAAT_RD57	A	IgA1	IgA1	IGHV3	IGHV3-23	IGHJ2	IGHD2	IGHD15
5695	CV222FU_A1_10003561_TTTATTGGATCGTCT_RD58	A	IgA1	IgA1	IGHV3	IGHV3-49	IGHJ4	IGHD3	IGHD6
5696	CV222FU_A2_10118548_TTGGATATGTTGAAT_RD106	A	IgA2	IgA2	IGHV3	IGHV3-64D	IGHJ4	IGHD6	IGHD19
5697	CV222FU_G1_10058462_TACAGTAGTGTGTT_RD35	G	IgG1	IgG1	IGHV3	IGHV3-30	IGHJ6	IGHD3	IGHD6
5698	CV222D0_G1_10022507_TTTATTCTCGAAT_RD28	G	IgG1	IgG1	IGHV3	IGHV3-30	IGHJ6	IGHD3	IGHD6
5699	CV222D0_A1_10023924_TGGATTTGATGTTAT_RD77	A	IgA1	IgA1	IGHV3	IGHV3-9	IGHJ4	IGHD6	IGHD19
5700	CV222D0_A1_10103732_TCGGCTCGTACTTT_RD10	A	IgA1	IgA1	IGHV3	IGHV3-74	IGHJ4	IGHD6	IGHD19

Showing 5,691 to 5,700 of 7,818 entries

Previous 1 ... 569 570 571 ... 782 Next

B

Enter the query input into the panel above, select one entry in the "Antibody information" tab to enable the other tab panels.

Antibody information Fab H-L contact map H-L interface residues Disulfide Bond

Select the additional column(s) you want to display

antigen_chain antigen_description

Download search results Select this to download pdb files of VCAb entries (.zip)

Show 10 entries

SubjectID	Perc Ident	E	iden_code	Structure	antigen_chain	antigen_description	Htype
1	7tn0_MN-HV_seq	95.833	2.13e-83	7tn0_MN	S	Spike protein S1	IgG1(homo_sapiens) GHG1*01, Per.Ident: 100.0
2	7tn0_HG-HV_seq	95.833	2.13e-83	7tn0_HG	I	Spike protein S1	IgG1(homo_sapiens) GHG1*01, Per.Ident: 100.0
3	70n_MN-HV_seq	95.833	2.13e-83	70n_MN	S	Spike protein S1	IgG1(homo_sapiens) GHG1*01, Per.Ident: 100.0
4	70n_HL-HV_seq	95.833	2.13e-83	70n_HL	R	Spike protein S1	IgG1(homo_sapiens) GHG1*01, Per.Ident: 100.0

Structural Viewer Sequence Coverage Sequence numbering

The structure of 7tn0_MN is shown below:

Chains: Heavy chain constant region, Heavy chain variable region, Light chain constant region, Light chain variable region, Antigen chain (if any), Other chains (if any)

Residues (if selected): Residues of Heavy chain, Residues of Light chain, Cysteines involved in disulfide bond

View all the chains with the same PDB ID

Display options: Color antigen chain in this pdb, Show other ligand(s) in this pdb, Zoom in to the heavy-light chain pair of the selected VCAb entry

Download the displayed structure (.pdb) file

C

Enter the query input into the panel above, select one entry in the "Antibody information" tab to enable the other tab panels.

Antibody information Fab H-L contact map H-L interface residues Disulfide Bond

Select the additional column(s) you want to display

antigen_chain antigen_description

Download search results Select this to download pdb files of VCAb entries (.zip)

Show 10 entries

SubjectID	Perc Ident	E	iden_code	Structure	antigen_chain	antigen_description	Htype
1	7bz5_HL-HV_seq	95.652	3.17e-79	7bz5_HL	A	Spike protein S1	IgG1(homo_sapiens) GHG1*03, Per.Ident: 100.0
2	7chs_HL-HV_seq	95.726	3.91e-78	7chs_HL	E	Spike protein S1	IgG1(homo_sapiens) GHG1*04, Per.Ident: 100.0
3	8cww_HL-HV_seq	94.017	1.46e-77	8cww_HL	A	Spike protein S1	IgG1(homo_sapiens) GHG1*03, Per.Ident: 100.0
4	6xc3_BA-HV_seq	94.017	1.46e-77	6xc3_BA	C	Spike protein S1	IgG1(homo_sapiens) GHG1*04, Per.Ident: 100.0

Structural Viewer Sequence Coverage Sequence numbering

The structure of 7bz5_HL is shown below:

Chains: Heavy chain constant region, Heavy chain variable region, Light chain constant region, Light chain variable region, Antigen chain (if any), Other chains (if any)

Residues (if selected): Residues of Heavy chain, Residues of Light chain, Cysteines involved in disulfide bond

View all the chains with the same PDB ID

Display options: Color antigen chain in this pdb, Show other ligand(s) in this pdb, Zoom in to the heavy-light chain pair of the selected VCAb entry

Download the displayed structure (.pdb) file

Figure S3: Investigation of COVID-19 repertoire from a structural perspective. (A) The user can upload the repertoire file and search for structural match. (B) Top structural match (PDB 7tn0) of the repertoire sequence "CV222D2_M_10050314_TAACCTTCTGTTATAT_RD39" indicates antibody recognise RBD epitope distinct from ACE2 binding site. (C) Repertoire sequence "CV227D9_G1_10089371.TTCGGTCACGTGAATT_RD30" is inputted to VCAb to search for structural match (PDB 7bz5). The top structural match shows this antibody directly blocks ACE2 binding.

S1.5 Access pre-computed *in silico* mutational scanning data

S1.5.1 Visualization of the *in silico* mutational scanning results

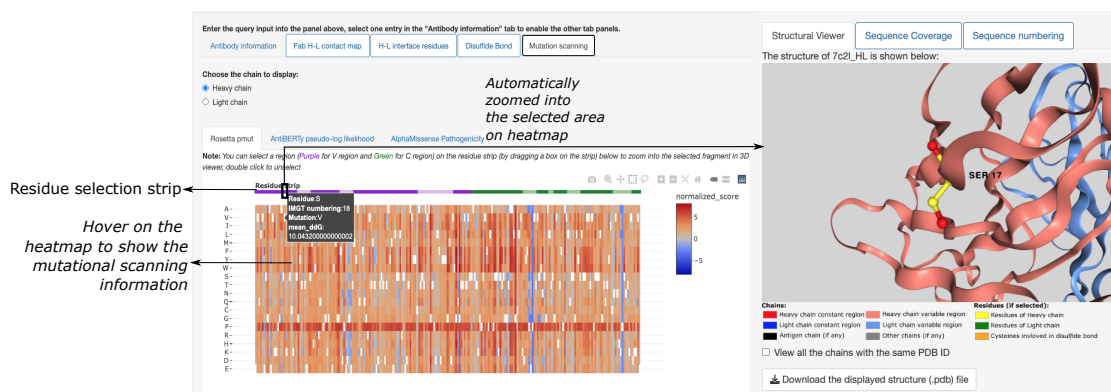


Figure S4: *in silico* mutational scanning profiles and three-dimensional structural viewer.

For each structural hit, pre-computed *in silico* mutational scanning profiles for the heavy and light chains have been generated by Rosetta pmut scan, AntiBERTy, and AlphaMissense. For each wild-type (WT) residue on the sequence, its mutation to every one of the other 19 amino acids has been evaluated using these methods to yield a score measuring mutational impact. These data are visualised as heatmaps which are displayed in separate tabs. Each *in silico* mutational scanning heatmap is accompanied by a residue selection strip positioned above the heatmap (**Figure S4**). Each dot on the residue strip represents one amino acid on the displayed structure, corresponding to one column in the heatmap. Residue selection strip is designed to allow the user to select residues of their interest to highlight on structural viewer. CDRs/loops and frameworks are highlighted by alternating dark/light colours on the selection strip. The x-axis for the heatmap represents the antibody sequence, and the y-axis represent the 20 amino acids. Red pixels in the heatmap represents mutations which are not preferable, and blue pixels represent preferable mutations. The WT residue for each position is marked by white. A hovered box is displayed for each point in the heatmap, showing the WT and mutant, the IMGT numbering of the position and the score representing the mutational impact as evaluated using the method selected by the user.

S2 Case studies

S2.1 Interface Investigation

Analysis of the CH1-CL interface is important in our understanding of the principles behind the pairing between H and L chains, and the design of bispecific antibodies for therapeutic purposes (Lewis *et al.*, 2014; Liu *et al.*, 2015; Mazor *et al.*, 2015; Krah *et al.*, 2017). With VCAb encompassing accurate annotations of the C region of antibody structures, we hypothesise that the database can be used to compare the residue composition of the CH1-CL interface under different types of H-L combination. Here, we give an example of comparing CH1-CL interface residues when different types of light chain (κ or λ) is paired with the same type of heavy chain (IgG1). The PDB entry 6rlm is an IgG1- κ antibody, while PDB 5y9k is IgG1- λ antibody. As expected, the CH1 sequences represented in these two structures are identical (**Figure S5A**). However, our analysis suggests that in these structures the κ and λ chains use different β -strands in engaging with the IgG1 H chain, involving different residues: in the IgG1- κ structure Thr198 of the heavy chain is in the interface with the CL domain (**Figure S5B**), while this threonine residue in the IgG1- λ structure is not (**Figure S5A**). We observe the opposite scenario for a nearby serine residue at the same β -strand: Ser189 in the IgG1- λ structure is involved in the CH1-CL interface (**Figure S5C**) while the serine residue at the same position in the IgG1- κ structure is not (**Figure S5A**). This indicates that in these two cases, the heavy chain of the same sequence in IgG1 pairs with the κ and λ chains using a slightly different configuration, with a certain degree of compensation from amino acids with similar physicochemical characteristics in the vicinity of one another.

Beyond this single example, VCAb provides the opportunity to investigate such problems at a larger scale with its accurate annotation of such structural features, periodically updated to include newly resolved antibody structures. We compare the residue contact patterns at the CH1-CL interface (**Figure S5D&E**) across all human VCAb entries. We performed pairwise comparison of the residue contacts between any two

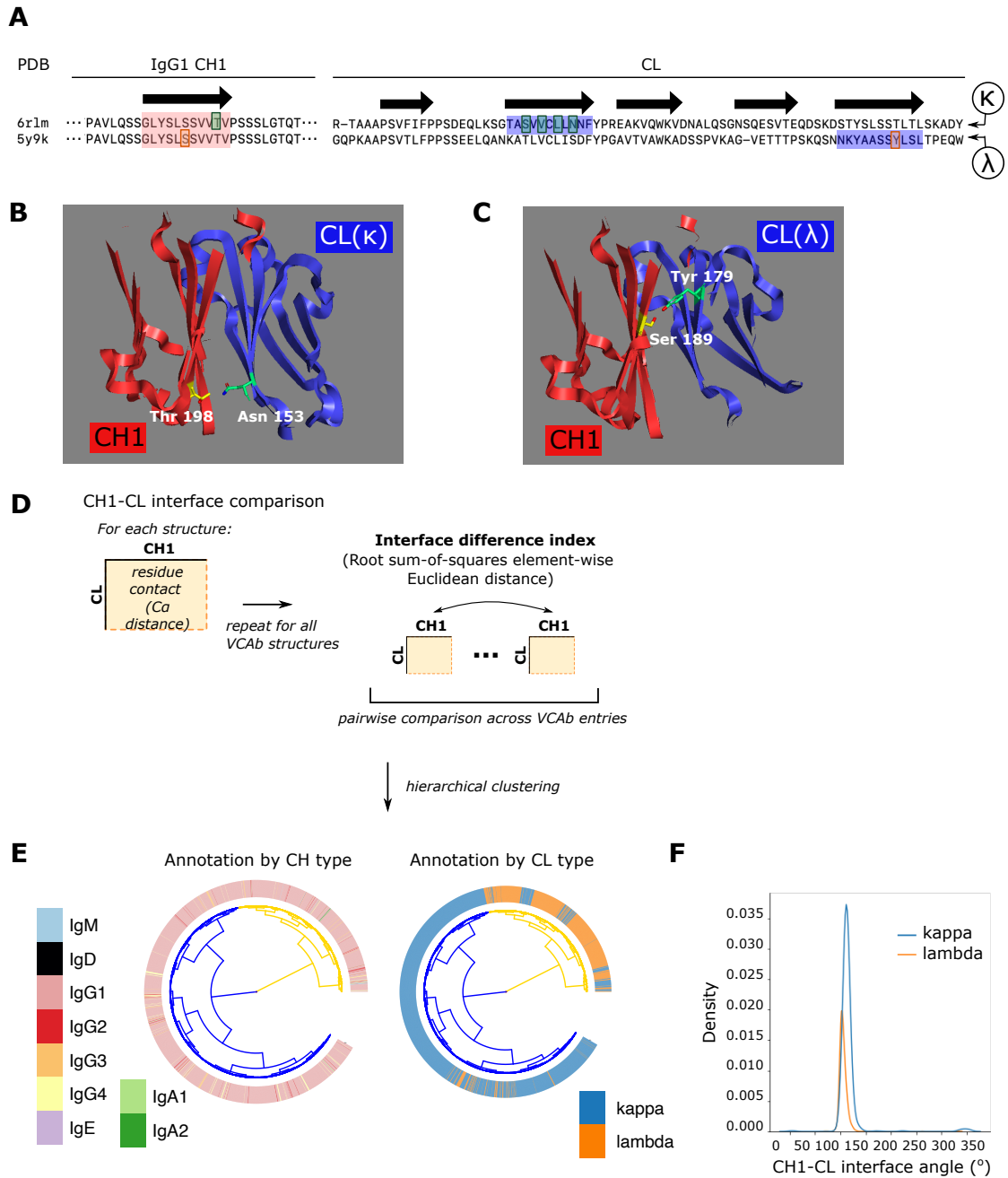


Figure S5: CH1-CL interface analysis. (A) The sequence alignment for the CH1 and CL sequences of PDB 6r1m (IgG1- κ) and PDB 5y9k (IgG1- λ). Arrows at the top of the sequence represents the position of the beta strand in the domain. Regions involved in the CH1-CL interface are highlighted in red (CH1) and blue (CL), with specific interface residues indicated with dark coloured boxes. (B and C) The display of the interacting residues of interest between CH1 and CL(κ) from PDB 6r1m (B), and CH1 and CL(λ) from PDB 5y9k (C). (D) Schematic explaining the process to compare CH1-CL interface across VCAb entries, by comparing residue contacts between CH1 and CL to generate what we term the “interface difference index” which compares CH1-CL $C\alpha$ contacts between any pair of VCAb structures. (E) Hierarchical clustering analysis of pairwise interface difference index across all human VCAb entries reveals 2 clusters of CH1-CL residue contact patterns. Each structure is annotated by the CH (left) and CL (right) types. (F) Distribution of CH1-CL interface angle for VCAb entries using the κ (blue) and λ (orange) light chains.

Fab structures (formula defined in **equation 4**) to generate a distance matrix. Hierarchical clustering (using Ward’s distance in R) was applied to this distance matrix to generate the dendrogram shown in **Figure S5E**, using the ggtree package. This comparison reveals two CH1-CL residue contact patterns; we found that whilst these patterns do not seem to depend on heavy chain type (primarily because of the sparsity of structures other than those of the IgG1 subtype in the data), the interface residue contact patterns can

be separated in accordance with the light chain types of antibody: most antibodies with κ are within one cluster, while most antibodies with λ are clustered together (**Figure S5E**), suggesting the choice of light chain partner influences CH1-CL packing. We further investigate the geometry of this CH1-CL packing by comparing the CH1-CL interface angles for antibodies with κ and λ light chains (**Figure S5F**). The difference between κ and λ paired antibodies is small (median for κ structures $\approx 112^\circ$, median for $\lambda \approx 104^\circ$). We also considered the elbow angles and they display a similar pattern (**Figure S5F**). In conclusion, CH1-CL interface patterns appear to be influenced by the light chain, utilising different residue contacts for such packing without altering the packing angles governing Fab geometry. In the VCAb web-server, users can query an antibody structure to search for other structures with similar CH1-CL residue contacts; these structures can again be interactively visualised and downloaded via the webserver.

S2.2 *In silico* mutational scanning profiles

Melting temperature measured for the experimental validation of the *in silico* mutational scanning profiles are collected from the publication and listed in the table below.

Table S3: Melting temperatures for the validation of *in silico* mutational scanning profiles

Antibody	Mutations	T _{mWT}	T _{mmutant}	ΔT _m
MEDI-UCA	G95P (L)	84.06 ± 0.012	72.01 ± 0.026	-12.05
mAb114	A68T-E72D-S79Y-I113T (H)/V43A (L)	74.25 ± 0.708	81.85 ± 0.023	7.6

S2.3 Isotype sensitivity analysis using *in silico* mutational scanning data

Rosetta pmut scan results can be downloaded from the VCAb webserver, for further analysis of positions in VH domain where the stability changes significantly upon isotype change. We reason that these positions potentially represent regions sensitive to switching the CH1 domains from one isotype to another. Isotype switching is an important process for B cell maturation *in vivo*, and also essential in the design of antibodies to accomplish specific effector functions. To mimic this process *in silico*, we selected as case study two antibody structures with the same V region sequence but different isotype (IgA1: PDB 3qnx; IgG1: PDB 3qo0). We additionally selected another IgG1 with the same V region (PDB 3qo1) as a negative control. This set of structures allow us to perform two comparisons: IgA1 v.s. IgG1 and IgG1 v.s. IgG1. The goal here is to find positions that produce significant changes in IgA1-IgG1 compared with IgG1-IgG1.

Within each comparison (i.e. IgA1-IgG1 or IgG1-IgG1), the difference between the mutation scores of two structures are calculated (Δmut), which is used to generate the Volcano plot in the main text (Figure 4). For each wild-type residues, Student's t-test is performed to calculate the p-values by comparing the 19 values of Δmut in group IgA1-IgG1 ($\Delta mut_{A1,G1}$) and group IgG1-IgG1 ($\Delta mut_{G1,G1}$). The fold change (FC) for each wild-type residue (r) is defined as equation 5, i.e. obtaining the excess mutational impact (represented by the average mutational difference over all 19 mutations for a given residue) in the IgG1-IgA1 comparison ($\overline{\Delta mut_{A1,G1}}$) over that from the IgG1-IgG1 comparison ($\overline{\Delta mut_{G1,G1}}$).

We note further that the FC_r value can be either positive or negative: since high mutation scores from Rosetta pmut represent low stability, positions with FC larger than zero indicates mutations here tends to be IgG1 favorable and FC smaller than zero indicates IgA1 favorable. We therefore normalise these FC_r values by an inverse hyperbolic sine transformation (i.e. $arcsinh(FC)$). In the volcano plot (main text Figure 4C), the x-axis depicts $arcsinh(FC)$ while the y-axis represents $-\log_{10}$ p-value. Each dot represents a wild-type residue; those with $|FC|$ larger than 2 and p value smaller than 0.01 are highlighted.

$$FC_r = \frac{\overline{\Delta mut_{A1,G1}} - \overline{\Delta mut_{G1,G1}}}{|\overline{\Delta mut_{G1,G1}}|} \quad (5)$$

Code for the analysis above can be reproduced by the jupyter notebook (https://github.com/Fraternallilab/VCAb/blob/main/isotype_sensitivity.ipynb).

S3 Methods

S3.1 Data collection and annotation

S3.1.1 Collecting Antibodies

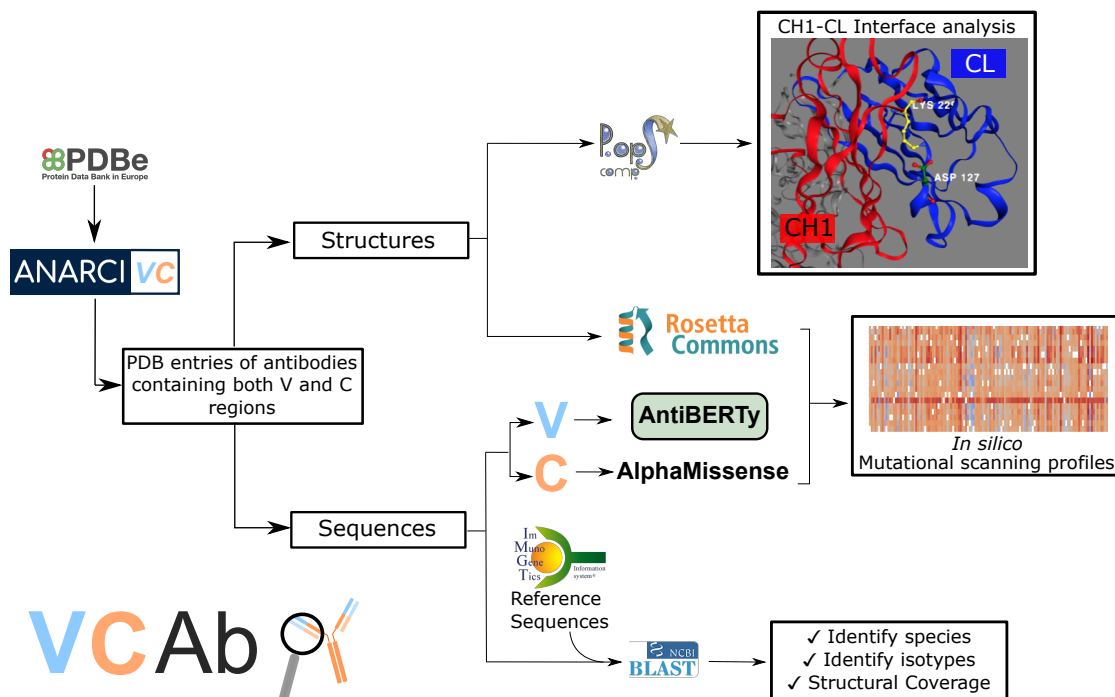


Figure S6: Schematic of the data collection, data annotation, and structure analysis in VCAb.

Protein sequences are downloaded (download date: 18th March, 2024) from worldwide PDB archive (Burley *et al.*, 2019) and inputted into ANARCIvc (section S3.1.3), a package modified from ANARCI (Dunbar and Deane, 2016), to identify all the antibody sequences containing both V and C regions in PDB (**Figure S6**). Heavy and light chains in each PDB entry are paired by applying the distance constraint of 22 Å between the conserved cysteines at position 104 of IMGT antibody V region numbering (Dunbar *et al.*, 2014; Lefranc *et al.*, 2005). Antigen chains are recorded as the chain bearing fragments within 7.5 Å to the CDR loops of the antibody (Dunbar *et al.*, 2014).

We note that the sequence directly accessed from PDB archive does not necessarily match the sequence represented in the PDB ATOM records and this could be misleading when selecting template for further modelling. We therefore downloaded the structural files (mmCIF format) and extracted the ATOM records for accurate feature annotation. Residues missing ATOM records are omitted in this case. Sequences downloaded directly from PDB archive and sequences extracted from the ATOM records are used for different purpose (see section S3.1.2).

To remove redundancy due to multiple copies of the same antibody sequence being represented in the PDB structures, antibody entries are collapsed based on whether they have the same PDB identifier and the same coordinate sequences (i.e. the sequence represented in the PDB ATOM records) of both H and L chains. We give a unique identification code for each antibody entry, in the format of `pdb_HL`, where `pdb` is the 4-character PDB code, H and L are the chain IDs of respective heavy and light chain from the first H-L pair. Duplicated antibody chains found in the same assembly are recorded in the collapsed database separately with all matching chain names.

S3.1.2 Species, isotype and structural coverage annotation

We noticed that species annotation for antibody sequences can be highly complex, and in some cases the single species nominated in the annotations provided by the PDB can be misleading. For example, some antibodies annotated as human in PDB might contain sequence fragments from other species, such as chimeric antibodies (i.e. sequences of V and C regions are from different species) or humanised antibodies (i.e. antigen-binding loops and fragments are grafted onto a human immunoglobulin framework). These antibodies are reflected by the low percentage identity when aligning to the human reference sequences for either/both of the C and

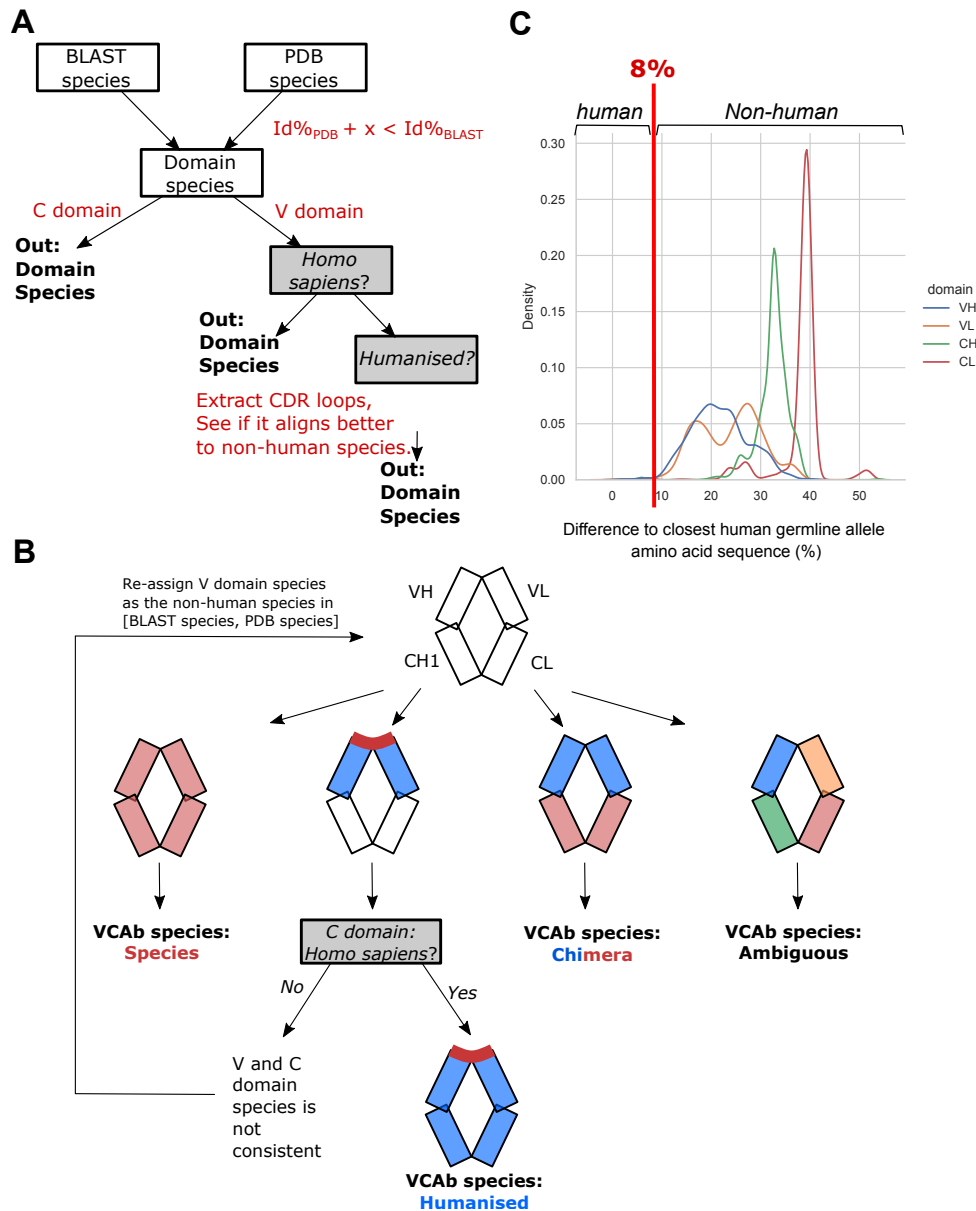


Figure S7: Species assignment in VCAb. **(A)** The process to assign species to domains. BLAST species is the species which give the highest percentage identity when aligned to the reference gene, PDB species is the species annotation in Protein Data Bank. We compare the percentage identity of the top hit from the BLAST species ($Id\%_{BLAST}$) to that of the top hit from the PDB species ($Id\%_{PDB}$). To allow for mutations and sequence polymorphisms, we only overwrite the PDB species as the designated species of the domain if the percentage identity for the BLAST species is higher than a certain value (x) compared with the $Id\%_{PDB}$. For V domains, an additional step is applied to identify whether it is humanised, by extracting CDR loops (section S3.1.2). **(B)** Procedure to assign the species for the entire antibody structure. Species are assigned to each domain first by the process shown in panel A, then all the domains species are considered together to give the species for the antibody. Different colors indicate different domain species, the dark red lines indicate the CDR. **(C)** The distributions for non-human antibodies of the difference between $Id\%_{BLAST}$ and the $Id\%$ when aligning to human reference genes. These sequences are from non-human antibodies where PDB species and BLAST species are consistent, and is trying to mimic the situation when they are mistakenly annotated as human. Different colors show the distribution of different domains. Majority shows the distribution higher than 8%. We therefore set x as illustrated in panel A as 8 in our pipeline.

V regions. To flag these cases, sequences of V and C regions for both H and L chains are extracted and passed as input to separate BLAST runs, to compare with the reference alleles from different species of V and C regions (downloaded from IMGT, date accessed: 1st July, 2022), and search for the alleles with the best sequence similarity. Species is firstly assigned for V and C regions separately (**Figure S7a**), then the final decision of the species for the entire antibody is made from the species annotation for regions (**Figure S7b**).

By doing so, we flag chimeric antibodies as those with different species annotations for the V and C regions. If an antibody is identified as human, then the CDR loops (IMGT definitions) would be extracted to compare its alignments to reference alleles from other species versus human, to identify humanised antibodies (**Figure S7ac**). Species annotation provided by PDB will be overwritten by the species of the best BLAST hit when the percentage identity for the best BLAST hit is higher than the hit corresponding to the PDB species by a certain cut-off (**Figure S7A**). To determine this cut-off, we extracted the non-human antibodies where the species annotations derived from BLAST and PDB are consistent. Sequences of non-human antibodies are aligned with the human reference sequences, to mimic the situation if the non-human antibodies are annotated as human in PDB. Then the percentage identity (Id%) for species from the best BLAST hit are compared with the Id% when aligned to human references. We find that a cut-off of 8% effectively distinguishes non-human and human antibodies (**Figure S7C**).

To identify the isotype of antibody chains found in the collated VCAb entries, we search sequences directly derived from PDB SEQRES records against IMGT (Lefranc *et al.*, 2015) reference CH and CL alleles of the species we assigned based on the procedure detailed above, using the BLAST program in the command line via Bio.Blast.Applications module in Biopython (version 1.79) (**Figure S6**). Alleles of isotypes downloaded from IMGT are first collapsed by their amino acid sequences to build a non-redundant BLAST sequence database. The best BLAST alignment with highest percentage identity is considered to assign the CH/CL type of the query chain.

The isotype of the antibody is listed in the `Htype` column of the VCAb database. In this column, following the name of isotype, the allele with the highest percentage identity in the BLAST result is shown. Alleles with identical amino acid sequence are also listed. For cases where the sequence is similarly close to multiple isotypes to give the same percentage identity or score as joint top hits in BLAST, they are listed in the `Alternative_Htype` column for users' reference. The light chain type of antibody is identified in the same process likewise.

To identify the structural coverage of the antibody, the heavy chain sequence represented in PDB ATOM records (instead of the sequence from PDB SEQRES records) are searched using BLAST against the IMGT reference IGHC sequences for the assigned C region species (**Figure S6**). Based on the start and end positions of the BLAST alignments, together with the domain boundaries described in the FASTA header of relevant CH alleles in IMGT, we classify the structural coverage of the antibody as Fab (the heavy chain contains only VH and CH1, but no other domains) or full antibody (the heavy chain contains every CH domain relevant for the given isotype). Since the light chain is immaterial in the classification into Fab and full antibodies (in both cases VL and CL should be present), this classification is based on the domain coverage of the heavy chain sequence with ATOM records extracted from PDB files. A second check is applied here to make sure all the antibodies in VCAb contain both V and C regions, by excluding antibodies identified only containing V regions using the sequence observed in PDB ATOM records; users can download the list of structures which are excluded from VCAb due to failures to meet these listed criteria on the VCAb web application.

The sequences of antibodies are separated into V and C regions, based on identifying the beginning of the alignment to the CH/CL alleles. These sequences for V region are collected into a flat-file which can be used as a sequence database for searching VCAb based on the V sequence similarity (see section S3.3, "VCAb webserver").

S3.1.3 Antibody C region numbering

Antibody sequence numbering is a process to assign, for every observed residue in the antibody sequence, an identifier based on a scheme that can be applied to every antibody sequence. The use of such universal scheme would allow cross-comparisons of residues between structures, for convenient extraction of its structural position (e.g., with respect to secondary structure elements, or for structurally conserved positions such as cysteines which precede CDR loops) and comparisons of these features across different antibody sequences. Numbering schemes for the V region have been well-established with an array of computational tools already available that automate the process to number a user-supplied V region sequence. Although a numbering scheme is also curated for the C region (IMGT)(Lefranc *et al.*, 2005), to our knowledge there are no tools which allow users to automatically number C region antibody sequence. We reason this is essential to generate annotations in VCAb which can be interpreted for sequence characterisation: to identify antibodies with C region sequences (see section S3.1.1, "Collecting antibodies"), to define reference points for CH1-CL angle calculation, "CH1-CL interface angle"), and to compare CH1-CL interfaces across VCAb (see section S3.1.4).

We address this issue by modifying the tool ANARCI (Dunbar and Deane, 2016). ANARCI built Hidden Markov Models (HMMs) of V region germline sequences (Dunbar and Deane, 2016), which can subsequently be used to number a user-supplied V region sequence using the `hmmscan` function in HMMER (Dunbar and

Deane, 2016). We expand ANARCI to ANARCIvc to enable the numbering of both V and C region sequences. For the numbering of C sequences, similarly to ANARCI, a set of HMM profiles are generated for reference sequences of CH1 and CL domains from different species automatically downloaded from IMGT. The source code of ANARCI is modified in order to implement the IMGT numbering scheme for the C region, and apply this scheme to number the C region observed in any antibody structure during the VCAb compilation process. The IMGT C region numbering scheme is built on the basis of V region numbering scheme, where it inherits the V region numbering for structurally equivalent β strands when structures for V and C domains are superimposed, remove the numbering for extra beta strands in V domains, and place insertions for additional residues in C domains (Lefranc *et al.*, 2005). The implementation and HMM models for V region numbering has been wholly inherited from ANARCI without modification. ANARCIvc source code is freely available as a github repository forked from ANARCI: https://github.com/FraternaliLab/ANARCI_vc.

S3.1.4 CH1-CL interface analysis

H-L interface residues POPSComp (Kleinjung and Fraternali, 2005) is applied to identify the residues involved in the interface between heavy and light chains, by evaluating the difference in the solvent accessible surface area (SASA) upon the formation of the H-L complex (**Figure S6**). This SASA difference (Δ SASA) is the area buried due to the interaction between the H and L chains. Residues with Δ SASA larger than 15 \AA^2 is considered as part of the H-L interface and listed as a table in the VCAb webserver; this Δ SASA cut-off has previously been applied to study protein-protein interaction interfaces across the PDB (Fornili *et al.*, 2013).

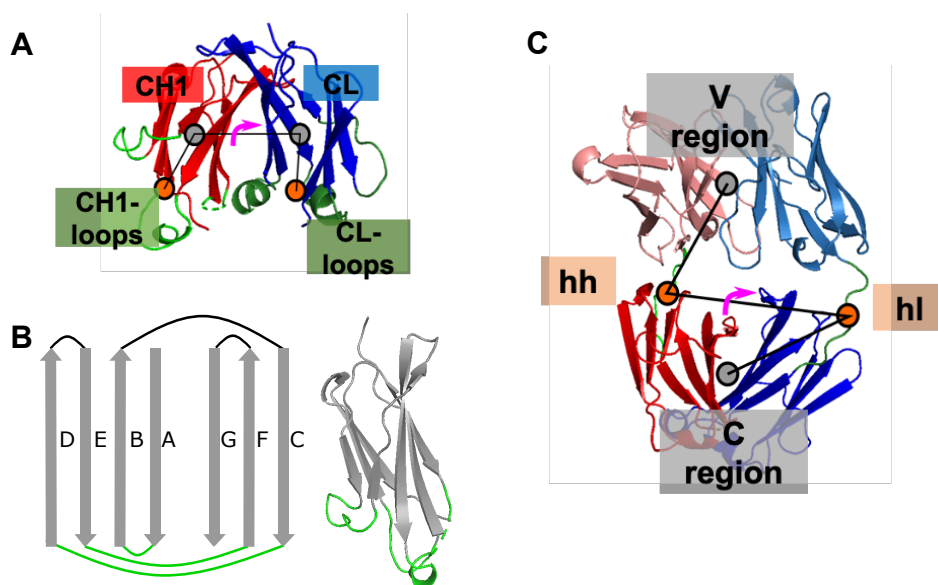


Figure S8: Angles calculated in VCAb. (A) CH1-CL interface angle defined by then center of mass (COM) of CL loops, CL domain, CH1 domain and CH1 loops. (B) CH1 and CL loops is defined as the loops between β -strand A&B, C&D, E&F for both CH1 and CL, and is highlighted in green. (C) The elbow angle is calculated as the torsion angle of the COM of V region, the COM of the loop linking VH and CH1 domain (hh), COM of the loop linking VL and CL domain (hl), and COM of the C region of the Fab.

CH1-CL interface angle calculation CH1-CL interface angles is defined as the torsion angle formed by the center of mass (COM) of the CL loops, COM of the CL domain, COM of the CH1 domain, and COM of the CH1 loops (Fernández-Quintero *et al.*, 2020) (**Figure S8A**). The CH1 and CL domains are extracted based on the domain boundaries identified during the BLAST step (see section S3.1.2). The region of CH1 loops and CL loops are defined as the fragments linking adjacent β -strands at the C-terminal of the C domain, i.e. fragments between β -strand A&B, C&D, E&F for both CH1 and CL (**Figure S8B**, (Chiu *et al.*, 2019)). The extraction of loops are based on loop positions suggested by antibody C region numbering for each VCAb entry. After these, the COM for CL loops, CL domain, CH1 domain and CH1 loops are calculated using Bio.PDB package (Bio version 1.79), for the calculation of CH1-CL angle.

CH1-CL interface similarity comparison To compare patterns of residue contacts between the CH1 and CL domains across all VCAb entries, we first generate a matrix r storing pairwise $C\alpha$ distances between

residues in the CH1 and CL domains; in the formula below, h and l refer to residues from the CH1 and CL domains, respectively.

$$r_{h,l} = \begin{cases} \text{distance between } \alpha\text{-carbons of } h \text{ and } l, & \text{if both } h \text{ and } l \text{ are interface residues} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Interface residues are taken from POPSCOMP analysis. In order to keep the dimensions of every contact matrix constant across all VCAb entries, sequences of CH1 and CL domains for all VCAb entries are numbered by ANARCIvc, and the resultant IMGT-gapped sequences are used for the two dimensions of the residue contact matrix.

We then compute a distance metric to quantify the difference between any two given residue contact matrices from two VCAb entries. The ‘‘interface difference index’’ ($\text{IntDiff}(A, B)$) is calculated for every pair of VCAb entries using the formula below, where A, B are residue contact matrices of VCAb entries:

$$\text{IntDiff}(A, B) = \sqrt{\sum_{h=1}^n \sum_{l=1}^n (a_{h,l} - b_{h,l})^2} \quad (4)$$

This amounts to a root sum-of-squares element-wise difference between two matrices A and B , which is used to search for antibody structures with similar CH1-CL contact in VCAb web server, or potentially for large-scale analysis (section S2.1).

S3.1.5 Elbow angle calculation

Elbow angle is defined as the torsion angle set by the center of mass (COM) of the V region (VH and VL domains), COM of the loop linking VH and CH1 domain, COM of the loop linking VL and CL domain, and COM of the C region of the Fab (CH1 and CL domains) (Fernández-Quintero *et al.*, 2020) (**Figure S8C**). V region and C region of the Fab are extracted according to the domain boundaries identified in BLAST (see section S3.1.2). The boundaries of loops linking VH and CH1, VL and CL are set by finding the nearest residues involved in adjacent β -strands with respect to the VH-CH1, VL-CL boundaries identified in BLAST. The extraction of these regions and the calculation of COM are performed using Bio.PDB package (Bio version 1.79).

	Mean	Inter-quartile range
CH1-CL interface angle		
κ	110.69	106.09 - 110.61
λ	102.90	98.74 - 106.03
Elbow angle		
κ	169.71	157.88 - 181.75
λ	180.49	152.53 - 209.51

Table S4: Summary of elbow angles and CH1-CL interface angles for VCAb structures.

S3.1.6 Disulphide bond calculation

The distance between $C\alpha$ of cysteines involved in disulphide bond is below 7.5 Å in most cases (Gao *et al.*, 2020). Distances between $C\alpha$ s of all the cysteines from each VCAb entry are calculated, and 7.5 Å is used as threshold to consider cysteines as forming a disulphide bridge.

S3.2 The generation of *in silico* mutational scanning profiles

The *in silico* mutational scanning profiles are generated by mutating each residue in the sequence to all the other 19 amino acids. The effect of these mutations are then scored by different methods. Currently, VCAb incorporates three *in silico* methods: Rosetta point mutant scan applications, AntiBERTy pseudo-log likelihoods, and AlphaMissense pathogenicity score.

S3.2.1 Rosetta point mutant scan

Antibody structures are inputted into the Rosetta point mutant scan (Rosetta pmut) pipeline, with a list of mutant containing all the possible mutations for this antibody provided (length of the list: 19 mutations * length of antibody sequence). Rosetta pmut will then calculate the energy changes upon this mutation

Table S5: The comparison between *in silico* mutational scanning profiles generated by different methods

	Input	Region	Meaning
Rosetta point mutant scan	Antibody structure	V and C regions	The energy changes when this mutation occurs
AntiBERTy	Antibody V region sequence	V region	How likely the mutated amino acids will appear at each position
AlphaMissense	Antibody C region sequence	C region	How likely the mutant would be pathogenic.

occurs (Leaver-Fay *et al.*, 2011). This energy changes will be the value in the *in silico* mutational scanning matrix, with negative being stabilizing (blue), and positive being destabilizing (red).

S3.2.2 AntiBERTy pseudo-log likelihoods

AntiBERTy mutational scanning score is derived from the pseudo-log likelihood of each possible amino acids being presented at every position, as determined by the AntiBERTy language model trained over 26 million antibody sequences (Ruffolo *et al.*, 2021). We scaled the raw AntiBERTy score by subtracting the AntiBERTy score for the wild type (WT) residue at each position, such that they can be interpreted as relative changes in amino acid preference compared to WT (equation 1). A positive scaled AntiBERTy score means this mutation is preferable at this position and vice versa. Since these likelihood scores are in the logarithm space, the mutation score of multiple co-existing point mutations is calculated by summing up scores corresponding to the individual mutations (equation 2, **Figure S9**).

$$Score_{pmut_i} = \log P(Mut) - \log P(WT) = \log \frac{P(Mut)}{P(WT)} \quad (1)$$

$$Score_{mmut} = \sum_i Score_{pmut_i} = \sum_i \log \frac{P(Mut_i)}{P(WT_i)} = \log \frac{\prod_i P(Mut_i)}{\prod_i P(WT_i)} \quad (2)$$

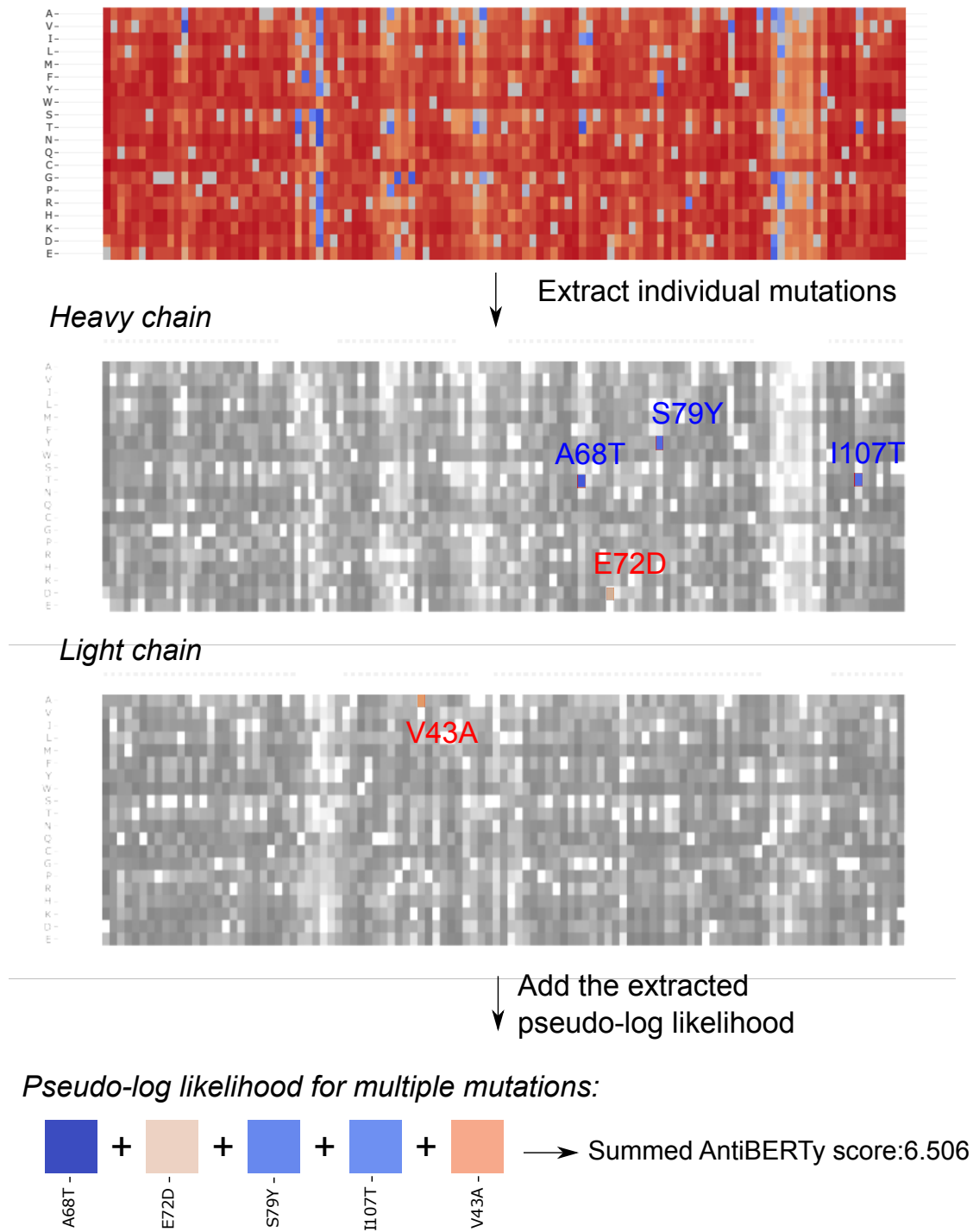


Figure S9: AntiBERTy mutational score can be added together to represent the mutational impact of multiple co-existing amino acid substitutions.

S3.2.3 AlphaMissense pathogenicity score

AlphaMissense calculates the pathogenicity score for all the possible variants in human proteome. Therefore, we downloaded the AlphaMissense pathogenicity prediction for all the heavy chain and light chain reference sequences. These predictions are then mapped to the corresponding heavy/light chain types.

S3.3 VCAb webserver

The website version of the VCAb is built to allow public access for academic research purposes. Shiny (version 1.6.0) in the R statistical computing environment (version 4.1.1) is used as the basis to build the website. NGLViewer (version 1.3.2) is used for visualisation of antibody structures. The VCAb webserver is available at: <https://fraternalilab.cs.ucl.ac.uk/VCAb/>. The underlying database is updated monthly, with the automatic pipeline scheduled to start at 07:00 Greenwich Mean Time on the last Monday of every calendar month. The pipeline begins by download all the protein sequences in PDB archive, and perform all the aforementioned analyses to populate VCAb. The source code for generating and hosting the shiny VCAb webserver is also made available (<https://github.com/Fraternalilab/VCAb>) for users to host a local version of the VCAb database.

The VCAb web server provides visualisation functionalities for the queried 3D structures, structural coverage, *in silico* mutational scanning profiles, antibody numbering information for both V and C regions, tabulated details of CH1-CL interface and disulphide bond, as well as bulk download of search results and the entire VCAb database in Comma Separated Value (CSV) format. For the data tables in the “Antibody information” panel, the following columns are displayed by default: **iden_code**, **Structure** (button to display in the 3D structure viewer and links to external databases), **Htype** (assignment of CH allele), **Ltype** (assignment of CL allele), **Structural.Coverage** (indication of Fab/full antibody represented in the atomic coordinates); users can customise this view by selecting other columns (such as **special_cases**) accessible by selecting the corresponding names in the “Select the additional column(s) you want to display” box. All columns as described in the online documentation can be accessed in the downloadable CSV files.

References

- Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410.
- Burley S. K., Berman H. M., Bhikadiya C., Bi C., Chen L., et al. (2019). Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic Acids Research*, **47**, D520–D528.
- Chiu M. L., Goulet D. R., Teplyakov A., and Gilliland G. L. (2019). Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies 2019, Vol. 8, Page 55*, **8**, 55.
- Dunbar J. and Deane C. M. (2016). Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, **32**, 298.
- Dunbar J., Krawczyk K., Leem J., Baker T., Fuchs A., et al. (2014). SAbDab: the structural antibody database. *Nucleic Acids Research*, **42**, D1140–D1146.
- Fernández-Quintero M. L., Kroell K. B., Heiss M. C., Loeffler J. R., Quoika P. K., et al. (2020). Surprisingly Fast Interface and Elbow Angle Dynamics of Antigen-Binding Fragments. *Frontiers in Molecular Biosciences*, **7**, 339.
- Fornili A., Pandini A., Lu H. C., and Fraternali F. (2013). Specialized Dynamical Properties of Promiscuous Residues Revealed by Simulated Conformational Ensembles. *Journal of Chemical Theory and Computation*, **9**(11), 5127–5147.
- Gao X., Dong X., Li X., Liu Z., and Liu H. (2020). Prediction of disulfide bond engineering sites using a machine learning method. *Scientific Reports*, **10**, 1–9.
- Kleinjung J. and Fraternali F. (2005). POPSCOMP: an automated interaction analysis of biomolecular complexes. *Nucleic Acids Research*, **33**, W342.
- Krah S., Sellmann C., Rhiel L., Schröter C., Dickgiesser S., et al. (2017). Engineering bispecific antibodies with defined chain pairing. *New Biotechnology*, **39**, 167–173.
- Leaver-Fay A., Tyka M., Lewis S. M., Lange O. F., Thompson J., et al. (2011). Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In M. L. Johnson and L. Brand, editors, *Methods in Enzymology*, volume 487 of *Computer Methods, Part C*, pages 545–574. Academic Press.
- Lefranc M. P., Pommié C., Kaas Q., Duprat E., Bosc N., et al. (2005). Imgt unique numbering for immunoglobulin and t cell receptor constant domains and ig superfamily c-like domains. *Developmental and comparative immunology*, **29**, 185–203.
- Lefranc M. P., Giudicelli V., Duroux P., Jabado-Michaloud J., Folch G., et al. (2015). IMGT[®], the international ImMunoGeneTics information system[®] 25 years on. *Nucleic Acids Research*, **43**, D413–D422.
- Lewis S. M., Wu X., Pustilnik A., Sereno A., Huang F., et al. (2014). Generation of bispecific IgG antibodies by structure-based design of an orthogonal Fab interface. *Nature Biotechnology*, **32**, 191–198.
- Liu Z., Leng E. C., Gunasekaran K., Pentony M., Shen M., et al. (2015). A Novel Antibody Engineering Strategy for Making Monovalent Bispecific Heterodimeric IgG Antibodies by Electrostatic Steering Mechanism. *Journal of Biological Chemistry*, **290**, 7535–7562.
- Margreitter C., Lu H. C., Townsend C., Stewart A., Dunn-Walters D. K., and Fraternali F. (2018). BRepertoire: a user-friendly web server for analysing antibody repertoire data. *Nucleic acids research*, **46**(W1), W264–W270.
- Mazor Y., Oganessian V., Yang C., Hansen A., Wang J., et al. (2015). Improving target cell specificity using a novel monovalent bispecific IgG design. *mAbs*, **7**, 377–389.
- Ruffolo J. A., Gray J. J., and Sulam J. (2021). Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv:2112.07782 [cs, q-bio].
- Stewart A., Sinclair E., Ng J. C. F., O’Hare J. S., Page A., et al. (2022). Pandemic, Epidemic, Endemic: B Cell Repertoire Analysis Reveals Unique Anti-Viral Responses to SARS-CoV-2, Ebola and Respiratory Syncytial Virus. *Frontiers in Immunology*, **13**, 1530.