A finite vocabulary of antibody-antigen interaction enables predictability of paratope-epitope binding

Rahmad Akbar[1], Jeliazko R. Jeliazkov[2], Philippe A. Robert[1], Igor Snapkov[1], Milena Pavlović[1,3], Andrei Slabodkin[1], Cédric R. Weber[4], Yana Safonova[5], Geir K. Sandve[3], and Victor Greiff[1,*]
[1] Department of Immunology, University of Oslo, Oslo, Norway
[2] Department of Biochemistry, University of Zürich, Zürich, Switzerland
[3] Department of Informatics, University of Oslo, Oslo, Norway
[4] Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[5] Computer Science and Engineering Department, University of California San Diego, La Jolla, USA
[*] Correspondence: Victor Greiff, victor.greiff@medisin.uio.no

Abstract

Antibody recognition of antigen relies on the specific interaction of amino acids at the paratope-epitope interface. A long-standing question in the fields of immunology and structural biology is whether paratope-epitope interaction is predictable. A fundamental premise for the predictability of paratope-epitope binding is the existence of structural units that are universally shared among antibody-antigen binding complexes. Here, we identified structural interaction motifs, which together compose a vocabulary of paratope-epitope binding that is shared among investigated antibody-antigen complexes. The vocabulary (i) is finite with less than 10^4 motifs, (ii) mediates specific and non-redundant interactions between paratope-epitope pairs, (iii) is immunity-specific (distinct from the motif vocabulary used by non-immune protein-protein interactions), and (iv) enables the machine learning prediction of paratope or epitope. The discovery of a vocabulary of paratope-epitope interaction demonstrates the learnability and predictability of paratope-epitope interaction.

Introduction

Antibody-antigen binding is mediated by the interaction of amino acids at the paratope-epitope interface of an antibody-antigen complex. A long-standing question in the fields of immunology and structural biology is whether paratope-epitope interaction is predictable. The predictability of paratope-epitope binding is a prerequisite for predicting antibody specificity and in silico antibody and vaccine design. So far, however, it remains unclear whether paratope-epitope binding is predictable (Brown et al., 2019; Raybould et al., 2019a; Sela-Culang et al., 2013).

The current paradigm is that antibody-antigen interaction is too high-dimensional of a problem to permit prediction. This long-standing belief has its roots in the works of Karl Landsteiner and colleagues showing that antibodies can recognize a nearly infinite number of antigens (Landsteiner, 1947). The complexity paradigm was reinforced over the subsequent decades by sequencing and crystallography studies suggesting that not only the interaction space is infinite, but also the sequence and structural space (with minor constraints on the antibody side) (Elhanati et al., 2015; Greiff et al., 2017a; Raybould et al., 2019a).

Antibody binding to the epitope is mainly formed by the three hypervariable regions called the complementarity-determining regions (CDRs) situated in both antibody heavy and light chains (Barlow et al., 1986; Inbar et al., 1972; Wu and Kabat, 1970). The hypervariability of the CDR3 is key to the immunological specificity of antibodies (Xu and Davis, 2000) and is generated by

somatic recombination of the variable (V), diversity (D, only for the heavy chain), and joining (J) genes of the B-cell genomic locus (Tonegawa, 1983). Combinatorial diversity from rearranged germline gene segments, somatic hypermutation, and antigen-driven selection steps enable antibodies to interact specifically with virtually any given antigen (Landsteiner, 1936; Padlan, 1977; Tonegawa, 1983).

The most reliable method for identifying paratope-epitope pairs is by solving the 3D structure of antigen-antibody complexes and determining which amino acids in the two partners make contact with each other (Van Regenmortel, 2014). Thus, the systematic mapping of interacting residues in antibody-antigen complexes is of great importance for establishing a rule-based understanding of antibody-antigen interaction. Over the last decades, the increasing amount of antibody-antigen crystal structures has enabled the first quantitative insight into the physicochemical features of antibody-antigen interaction (Benjamin et al., 1984; Berzofsky, 1985; Burkovitz et al., 2013; Dalkas et al., 2014; Lawrence and Colman, 1993; MacCallum et al., 1996; Ofran et al., 2008; Peng et al., 2014; Raghunathan et al., 2012; Sela-Culang et al., 2013; Sivalingam and Shepherd, 2012). For example, it has been observed repeatedly that paratopes localize mostly, but not exclusively, to CDRs (Kunik et al., 2012a), and that certain amino acids are preferentially enriched or depleted in the antibody binding regions (Mian et al., 1991; Nguyen et al., 2017; Ramaraj et al., 2012; Sela-Culang et al., 2013; Wang et al., 2018). For epitopes, several analyses have shown that their amino-acid composition is essentially indistinguishable from that of other surface-exposed non-epitope residues if the corresponding antibody is not taken into account (Kringelum et al., 2013; Kunik and Ofran, 2013; Ponomarenko and Bourne, 2007).

Recently, computational and machine learning approaches for the sequence-based prediction of paratopes (Deac et al., 2018; Kunik et al., 2012b; Liberis et al., 2018), epitopes (Kringelum et al., 2012) or paratope-epitope (antibody-antigen) interaction (Baran et al., 2017; Deac et al., 2018; Jespersen et al., 2019; Kilambi and Gray, 2017; Krawczyk et al., 2013) are accumulating (for a more complete list of references see here: (Brown et al., 2019; EL-Manzalawy et al., 2017; Esmaielbeiki et al., 2016; Raybould et al., 2019a)). While the accuracy for the sequence-based prediction of paratopes seems generally higher than that for epitopes, overall, prediction accuracy has remained suboptimal – especially with regard to sequence-based prediction of epitope sites and paratope-epitope interaction (Brown et al., 2019; Greenbaum et al., 2007). In brief, while there exists robust knowledge on the general amino acid and physicochemical features involved in antibody-antigen interaction, to our knowledge, no study has yet conclusively shown that antibody-antigen interaction is predictable and if so, based on what theoretical and biological grounds.

Recent reports have, however, provided preliminary evidence for the potential predictability of antibody-antigen interaction: (i) The antibody repertoire field has now established that antibody sequence diversity underlies predictable rules (Elhanati et al., 2015; Greiff et al., 2017b, 2017a). (ii) The presence of transferable “specificity units” between distinct antibody molecules was recently suggested by showing that tightly binding functional antibodies may be conceived by designing and improving seemingly unrelated paratopes (Nimrod et al., 2018).

Previous efforts towards predicting paratope-epitope interaction have been stifled by both a one-sided investigation of either exclusively the paratope or the epitope and the failure to break down the problem of antibody-antigen interaction into its *fundamental units.* The fundamental units of antibody-antigen interaction are the sequence regions on the antibody and the antigen that compose the paratope-epitope interface. We conjectured that the comparison of those interaction units across antibody-antigen crystal structures may lead to the discovery of a general vocabulary of antibody-antigen interaction. If a general vocabulary for antibody-antigen interaction existed, then paratope-epitope interaction must be predictable.

To prove our conjecture, we devised a domain-specific sequence language that encodes the units of paratope and epitope interaction as *structural interaction motifs* (henceforth termed interaction motif or simply motif). These structural interaction motifs (i) conserve the link between paratope and epitope and, at the same time, (ii) allow for the comparison of paratope-epitope structural interaction across all available crystallized antibody-antigen complexes.

Here, we identified a finite vocabulary of structural interaction motifs (398 paratope and 1594 epitope motifs) that govern antibody-antigen interaction. This interaction motif vocabulary has the following characteristics: (i) it is finite and immunity-specific. Specifically, we found that our dataset of 825 antibody-antigen (protein antigen) crystal structures – although tens of orders of magnitude below the sequence diversity of antibody receptors and antigen sequences – covers already 60% of all possible paratope and 30% of all possible epitope motifs. We excluded the possibility that the motif vocabulary discovered is a trivial subset of the protein-protein binding space. (ii) The paratope motif vocabulary shows mutually exclusive epitope reactivity. Specifically, while a small number of paratope motifs was found to bind to as many as 50% of all observed epitope motifs (paratope polyspecificity), the majority of paratope and epitope motifs interacted with very few motifs (oligospecificity) creating a two-tiered architecture of antibody-antigen interaction that is composed of motifs of broad reactivity on the one hand and very restricted reactivity on the other hand. (iii) The motif vocabulary is seemingly universal. For example, interaction motifs were shared across entirely unrelated antibody-antigen complexes (different antibody germline genes, different antigen classes). The existence of a vocabulary in antibody-antigen binding implies that antibody-antigen recognition is predictable. To confirm this claim, we leveraged deep learning to show that paratope-epitope interaction motif pairing can be predicted with an accuracy ranging from 60–70% (baseline accuracy: 20–30%).

Methods

Dataset of antibody-antigen crystal structures

A dataset of antibody-antigen complexes in the format of Protein Data Bank (PDB) (Fig. 1A) was obtained from Antibody Database (AbDb) (Berman et al., 2000; Ferdous and Martin, 2018). AbDb routinely crawls PDB to find existing non-redundant antibody-antigen (protein antigen) crystal structures and preprocesses them by (i) identifying the crystallized antibody (VH-VL, variable [V] heavy [H] and light [L] chain domains) and the corresponding co-crystallized ligand, (ii) annotating the antibody variable region (Fv) by consulting the Summary of Antibody Crystal Structure (SACS)

database (Allcorn and Martin, 2002) and (iii) applying a standardized numbering scheme for the antibody sequences. As of 15 June 2019, AbDb stored 825 complexes. Atoms labeled with PDB record type HETATM (non-protein atoms serving as co-factors) were removed from the complexes. The final curated dataset comprised 825 antibody-antigen (protein antigen only) pairs. All structures analyzed were validated to have a resolution of at least 4.0Å (*44*). To gain statistical power, we analyzed mouse and human antibody complexes as one entity. Mouse and human antibody-antigen complexes represent 90% of the AbDb (Suppl Fig. 11). Indeed, we found only minor differences in paratope motif spaces used by mouse and humans (>50% overlap, Suppl Fig. 4C, 14) and it was recently shown that neither CDR/FR length nor the distribution of interface residues in human and murine antibodies differ substantially (Collis et al., 2003; Henry and MacKenzie, 2018; Wang et al., 2018).

Annotations for 113 broadly neutralizing antibodies were obtained from the database bNAber (Eroshkin et al., 2014), 70 of these antibodies were represented as 24 non-redundant complexes in AbDb and were included in this study (Suppl Fig. S5). The remaining 38 (without antigens) and 5 missing structures were excluded.

Selection of antibody sequence numbering scheme

AbDb provides datasets with three numbering schemes: Kabat (Kabat et al., 1992), Chothia (Chothia and Lesk, 1987), and Martin (Abhinandan and Martin, 2008). These numbering schemes partition the antibody heavy and light chains into framework (FR1, FR2, FR3, FR4) and CDR (CDR1, CDR2, and CDR3) regions. In the Kabat numbering, gaps found within the alignment are based on the variability of the aligned sequences. As more three-dimensional (3D) structural information became available, Chothia and Lesk created a numbering scheme that takes spatial alignment into consideration. In particular, they corrected the positioning of the first CDR in both heavy and light chains. Abhinandan and Martin further refined the Chothia numbering scheme by making corrections, not only in the CDRs but also in the FRs. Here, we used the Martin numbering scheme to annotate the FRs and CDRs of antibodies as it was found to be suitable for structural and antibody engineering (Dondelinger et al., 2018) and it is the most recent of the presently available numbering schemes. Supplementary Table S1 summarizes the position of FR and CDR regions and the position of insertions according to the Martin numbering scheme. In the Martin numbering scheme, the CDR-H3 region excludes the V-gene germline part of the antibody gene (typically identified by the amino acid triplet CAR), as well as parts of the J-gene germline part (typically identified by “W”) as demonstrated in Supplementary Table S2.

Definition of interacting residues in antibody-antigen complexes

To identify interactions between amino acid residues in antibody-antigen complexes, a distance cutoff was set. Distance cutoffs between 4–6Å are routinely used when examining interactions between proteins or protein-ligand pairs as most noncovalent atomic interactions are short-range (e.g., hydrogen bonds and Van der Waals interaction range from 3–4Å (Esmaielbeiki et al., 2016; Lodish et al., 2000)). For instance, a recent study on contact-based protein structure networks by Viloria and colleagues found that a distance cutoff of <5Å (heavy atoms) is most sensitive to changes in residue interactions and variables such as force fields (Salamanca Viloria et al., 2017). Therefore, we defined paratope-epitope interacting residues by a distance cutoff of <5Å between

heavy atoms, i.e., amino acid residues are interacting if they have heavy atoms with a distance <5Å from each other (Fig. 1B). For completeness, we evaluated the variation of the total number of interacting residues, their distribution across FR and CDR regions as well as the overlap of interaction motifs (for explanation see below) for the three commonly used distance cutoffs <4Å, <5Å and <6Å in Suppl Figs. S3 and S7. We confirmed that the overall trends in per-region residue distribution, overlap between paratope and epitope interaction motifs (Suppl Fig. S3) and (dis)continuity (Suppl Fig. S7) hold across the three distance cutoffs tested indicating that our definition of interacting residues is appropriate and robust.

Definition of paratope, epitope, and paratope-epitope sequence interaction motifs

(i) A paratope is defined as the set of interacting amino acid residues within a particular FR or CDR region of an antibody (e.g., residues in salmon in Fig. 1B). (ii) An epitope is defined as the set of antigen amino acid residues that interact with a paratope. Epitopes are annotated according to the FR or CDR regions of the corresponding paratopes. (iii) The length of a paratope or epitope is defined as the number of amino acid residues constituting the paratope or epitope (see paratope/epitope length in Fig. 1D). (iv) A gap is defined as the number of non-interacting residues separating two paratope or epitope residues (Fig. 1B, 1D, and 2A). (v) A paratope or epitope structural interaction motif is composed of interacting paratope and epitope amino acid residues as well as non-interacting ones (gap). Interacting residues are encoded with the letter X and non-interacting residues are encoded with an integer quantifying gap size (number of non-interacting residues, Fig. 2A). For example, the string X1X encodes a paratope or epitope interaction motif of two interacting amino acid residues (X,X) separated by one non-interacting residue (gap).

Definition of interaction motif angle

The angle of an interaction motif was computed by defining two vectors spanning the midpoint of a motif and its start and end positions (see inlet in Suppl Fig. S8 for illustration), in a similar fashion to `AngleBetweenHelices`, a `Pymol` module for calculating the angle between helices (Schrödinger, LLC, 2015). Larger angles would indicate that the structure of the interaction motif is more extended whereas small angles indicate that the interaction motif would tend to form a loop. Protein 3D structures was rendered and visualized in `Pymol 2.1.0` (Schrödinger, LLC, 2015).

Diversity analysis of interaction sequence motifs

We estimated the relative coverage of the total respective paratope and epitope sequence motif space using species accumulation curves. The species accumulation curves were constructed as described previously (Greiff et al., 2017a). Briefly, we first defined the motif coverage (Coverage$_i$) of a given antibody-antigen complex (Complex$_i$) as the percentage of overlap of its set of unique paratope or epitope motifs (i) with the set of unique paratope or epitope motifs contained in the previous complexes: $j = i - 1$

$$\text{Coverage}_i = \left[\text{Motifs in Complex}_i \cap \bigcup_{1}^{j} \text{Motifs in Complex}_j\right] / \text{Motifs in Complex}_i,$$

where $i \in \{1, ..., T\}$, with T is the total number of antibody-antigen complexes. Coverage

calculations were performed 1000 times in different orderings of complexes to exclude ordering bias.To extrapolate the potential (total) paratope or epitope sequence diversity, we used the Chao1 estimator (Chao, 1984, 1987; Chao and Chiu, 2016), a non-parametric estimator of the lower bound of species richness [number of unique sequence motifs], as implemented in the R package `Fossil 0.3.7` (`chao1`).

Paratope-epitope amino acid contact map

Paratope(P)-epitope(E) amino acid contact maps were obtained by computing the log odds ratio, $L_x(P_i, E_j) = \log_2(P(P_i, E_j)/P(P_i)P(E_j))$, of the observed occurrence of each amino acid pair over the corresponding expected frequency as described in (Kunik and Ofran, 2013); where i is the paratope amino acid, j is the epitope amino acid, and x is the regions (FR/CDR in antibody-antigen complexes). Analogously, protein-protein amino acid contact maps were computed for inter- and intradomain PPI complexes.

Construction of bipartite paratope-epitope motif reactivity network

A paratope-epitope motif interaction network (*reactivity network*) was constructed by connecting each paratope motif to its corresponding epitope motif (undirected edge). The degree distribution, the distribution of the number of connections (edges) to a node (degree), of the resulting interaction network was tested to fit a power-law distribution by calculating a goodness-of-fit value with bootstrapping using the `poweRlaw R 0.70.2` package (Gillespie, 2015) as described by Clauset and colleagues (Clauset et al., 2009). Here, a network whose degree distribution fits a power-law distribution (exponent between 2 and 3) is defined as scale-free (Broido and Clauset, 2019). Networks and the corresponding visualizations were constructed using the network analysis and visualization suite `Cytoscape 3.7.1` (Shannon et al., 2003).

Analysis of sequential dependencies in interaction motifs

To quantify the sequential dependencies in paratope and epitope interaction motifs, we determined for each multi-residue motif, the 2-mer decomposition of each paratope/epitope sequence (sliding window) of the ensemble of paratope/epitope sequences mapping to the respective motif. For each motif, these sequential dependencies were visualized as directed networks where the 20 amino acids are the nodes and the edges are the frequency with which a 2-mer occurred (sequential dependency). Networks and the corresponding visualization were constructed using the network analysis and visualization suite `Cytoscape 3.7.1` (Shannon et al., 2003). Hierarchical clustering of the motifs' sequential dependencies was performed using the R package `pheatmap 1.0.12` (Kolde, 2019), distances between motifs were quantified by Euclidean distance or correlation and agglomeration was carried out using the complete-linkage method.

Dataset of protein-protein interaction and definition of protein-protein sequence interaction motifs

A dataset of protein-protein interactions (PPI) was sourced from 3did, a catalog of known three-dimensional structure domain-based interactions (Mosca et al., 2014). The database (i) collects high-resolution 3D-structures from PDB (`version 2019_1`) (Berman et al., 2000) and (ii)

annotates the structures according to the protein domain definitions provided by Pfam (version 32.0, Table S3 summarizes the top 10 protein domains in the latest version 3did) (El-Gebali et al., 2019). Interactions between domains originating from different chains were annotated as *interdomain* whereas interactions originating from the same chain as *intradomain*. Structures with Pfam domain description (i) immunoglobulin and (ii) Ig-like were excluded (as they overlap with structures from AbDb). The PPI dataset contained a total of 18,599,078 contact residue pairs (100,888 protein crystal structures), which is three orders of magnitude higher than the number of antibody-antigen contact residues (18,630 residue pairs, Fig. 1C). Protein-protein interaction motifs were constructed for each domain pair analogously to paratope-epitope interaction motifs (see previous section). Motifs with gap length larger than seven were excluded from the analysis (to match the largest gap size found in paratopes, Fig. 1). The final dataset comprises 23,403 interdomain and 8,311 intradomain complexes (Suppl Fig. S6).

Quantification of somatic hypermutation on antibody amino acid sequences

To quantify somatically hypermutated (SHM) amino acid residues in the dataset, we annotated the sequences with the corresponding species (here shown only human and mouse) and aligned the sequences against germline immunoglobulin V, D, and J genes sourced from IMGT.

The IMGT database (Lefranc et al., 1999) includes 570 (578), 34 (39), and 32 (26) human (mouse) germline immunoglobulin V, D, and J genes, respectively. We translated the nucleotide sequences of V and J genes according to their ORFs. As D genes can be truncated during recombination process, we used amino acid sequences corresponding to all three ORFs (excluding non-productive translations). To compute alignments, we used the following scoring scheme: match reward = 2, mismatch penalty = –1, gap opening penalty = –5, and gap extension penalty = –2. For each sequence, we selected germline V, D (if sequence corresponds to heavy chain), and J genes with the highest alignment scores. SHMs were defined as differences in the alignment between the antibody sequence and the selected germline genes. Exonucleolytic removals during V(D)J recombination lead to deterioration of the alignment quality at the end (start) of V (J) genes. To reduce their impact on SHM quantification, we discarded SHMs corresponding to three amino acid residues at the end (start) positions of V (J) genes in the alignment. Ralph and Matsen showed that three amino acids (up to 9 nt) correspond to the average lengths of exonucleolytic removals in V and J genes (Ralph and Iv, 2016). To reduce the impact of exonucleolytic removals in D genes, we considered only SHMs emerging between the first and the last matches in the alignments. Suppl Fig. S10 shows that the inferred SHMs localize around CDR1s and CDR2s and thus partially correlate with the paratopes positions centered in all three CDRs. Suppl Fig. S10 also shows only few SHMs in CDR3s. We argue that this is due to the limitations of our SHM quantification approach rather than the biological features of immunoglobulin sequences.

Machine-learning based prediction of paratope-epitope pairing

To demonstrate that paratope-epitope interactions are learnable, we trained a range of sequence to sequence deep learning models. Specifically a model for the translation of epitope to paratope (and vice versa) both at motif and sequence level was constructed based on Neural Machine

Translation (Luong et al., 2015), Here, pairs of input-output sequences were translated via a combination of two components: encoder and decoder with gated recurrent units (see Fig. 6 and the workflow graphic below). For example a paratope structural interaction motif XXX interacts with an epitope motif X2X resulting an input-output pair XXX-X2X, similarly, a paratope sequence NMA interacts with an epitope sequence RA giving an input-output pair NMA-RA. The datasets, for both motif and amino acid sequences, comprise a total of 5,327 input-output pairs. For the translation task of sequence motifs, we abstracted the gaps within a motif by replacing them with dashes, for example, all motifs of the form X*i*X (where *i is any integer*) were encoded as X-X. Table 1 summarizes the setup for both motif and sequence translations.

The dataset was split into 80% training and 20% test data. Numerical representation of the input pairs was learned by vector embedding. The (hyper) parameters (i) embedding dimension and (ii) number of units (hidden dimension) were optimized by the adaptive moment estimation (Adam) (Kingma and Ba, 2014). Embedding dimension is the length of the vector representing the input. Number of units is the number of cells in the gated recurrent unit (GRU) otherwise known as the length of the hidden dimension. The training procedure was carried out for 20 epochs and was replicated ten times. Each replicate comprises 121 models for a total of 1,210 models (see workflow). The final model from each replicate was used to generate a prediction on the test dataset. Discrepancy (error) between predictions and the true motifs (sequences) was measured by calculating (i) the correlation between the length of the predictions and the true motifs (sequences) as well as (ii) the normalized Levenshtein distance, $\mathrm{LD}_{\text{prediction vs truth}} / \left(\max(\mathrm{length}(\mathrm{prediction}), \mathrm{length}(\mathrm{truth}))\right)$, between the predicted motifs (sequences) and true motifs (sequences). Baseline prediction accuracies were calculated based on label-shuffled data where antibody and antigen-binding partners were randomly shuffled. Deep learning models were constructed in `TensorFlow 1.13.1` (Martín Abadi et al., 2015) with `Keras 2.2.4-tf` (Chollet and others, 2015) in `Python 3.6.4` (Van Rossum and Drake Jr, 1995), and computations were performed on the high-performance computing cluster Fram (Norwegian e-infrastructure for Research and Education `sigma2.no/fram`).

Table 1. Summary of deep learning tasks.

Incorporation of amino acid position	Direction	Data type
No	Paratope to epitope	Motif
Yes	Paratope to epitope	Motif
No	Epitope to paratope	Motif
Yes	Epitope to paratope	Motif
No	Paratope to epitope	Sequence
Yes	Epitope to paratope	Sequence

Machine learning workflow for the prediction of epitope motif (or sequence) from paratope motif (or sequence) and vice versa. At the start of the workflow, a dataset comprising 5327 pairs of paratope-epitope motifs (or sequences) was split into a train (80%) and a test (20%) datasets. A pair of hyperparameter (embedding dimension and GRU unit) was selected from a set of 121 pairs. Each pair defines a model and each model was trained for 20 epochs on the training dataset. Once trained, the model was evaluated on the dataset by calculating (i) the correlation between the length of the predictions the true motifs (sequences) as well as (ii) normalized Levenshtein distance (see above) between the predictions and true motifs (sequences). The procedure was replicated ten times. To accelerate the training process, parallelization was carried out in two stages: (i) jobs in the replication step ("repeat ten times") were sent to a set of nodes via jobarray in `SLURM` (Yoo et al., 2003) and (ii) the 121 models were scattered across 121 CPUs via `mpi4py` (Dalcin et al., 2011). Baseline prediction accuracies were calculated based on label-shuffled data where paratope and epitope-binding partners were randomly shuffled (random pairing of paratope-epitope pairs).

Generation of graphics

All non-network graphics were generated using the statistical programming environment `R 3.5.2` (R Core Team, 2018) with the grammar of graphics R package `ggplot2 3.1.0` (Wickham, 2016), the R package `VennDiagram 6.20` (Chen, 2018), and the ggplot2 theme `themeakbar 0.1.2` (Akbar, 2019). Figures were organized and schematics were designed in `Adobe Illustrator CC 2019`.

Dataset and code availability

The dataset, code and figures are available at:
`https://github.com/GreiffLab/manuscript_ab_epitope_interaction.`

Results

Antibody and antigen sequences in antibody-antigen crystal structure dataset are diverse

To gain a representative picture of antibody-antigen 3D interaction, we retrieved a dataset of all known 825 non-redundant antibody-antigen complexes (protein antigen only) archived in the Antibody Database (AbDb) (Berman et al., 2000; Ferdous and Martin, 2018) (Fig. 1A). Antibody sequences mapped to a diverse set of 64 (mouse)–104 (human) V genes (Suppl Fig. S1A) and antigen sequences belonged to a diverse set of antigen classes from both humans and mice (Suppl Fig. 11). Both antibody (Ab) and antigen (Ag) sequences showed a high median sequence

distance among either antibody framework (FR) and complementarity determining regions (CDR) (Suppl Fig. S1B) or antigen classes (Suppl Fig. S1C). Thus, the dataset is diverse and not overly biased to one type of antibody or antigen class or to sequences of high similarity.

The majority of paratope interacting residues are located in the antibody complementarity determining regions

We identified the set of interacting residues at the interface of antibody-antigen structures by using a heavy-atom distance cutoff of <5Å (Ostmeyer et al., 2019) (see Methods and see Suppl Fig. S3 and S7 for the robustness of the distance cutoff). Antibody-antigen amino acid pairs within this distance were designated as interacting residues (Fig. 1A–B). Together, the sets of antibody and antigen interacting residues form paratope-epitope pairs. In accord with previous reports (MacCallum et al., 1996; Stave and Lindpaintner, 2013), paratope residues mapped overwhelmingly to the complementarity-determining regions (CDR) 1–3 ($V_{H,CDR1-3}$: 89.5% and $V_{L,CDR1-3}$: 89.2%) and, consequently, were only rarely found in the framework regions (FR) 1–4 ($V_{H,FR1-4}$: 10.5% and $V_{L,FR1-4}$: 10.8 %) (Fig. 1C). While the percentage of interacting residues found in the CDR3 was consistent between heavy and light chains, only half as many interacting residues mapped to CDR-H1 (14.8%) as compared to CDR-L1 (32.5%), whereas the reverse was true for CDR-H2 (36.1%) and CDR-L2 (13.9%). Since we used the Martin numbering scheme for CDR and FR annotation (see Methods), which mostly excludes germline gene residues from the CDR3, the above numbers demonstrate that germline-gene residues surrounding the CDR3 (FR3, FR4) contribute relatively little to antibody-interaction and that CDR3 paratope-epitope interaction is essentially non-germline gene residue driven (Abhinandan and Martin, 2008; Dondelinger et al., 2018). Finally, we found that paratope interacting residues correlate positively with the sites of (inferred) somatic hypermutation (SHM) (Spearman (Pearson) correlation: 0.31–0.52 (0.44–0.58), Suppl Fig. S10) suggesting that interacting residues investigated herein are likely subject to antigen-driven selection.

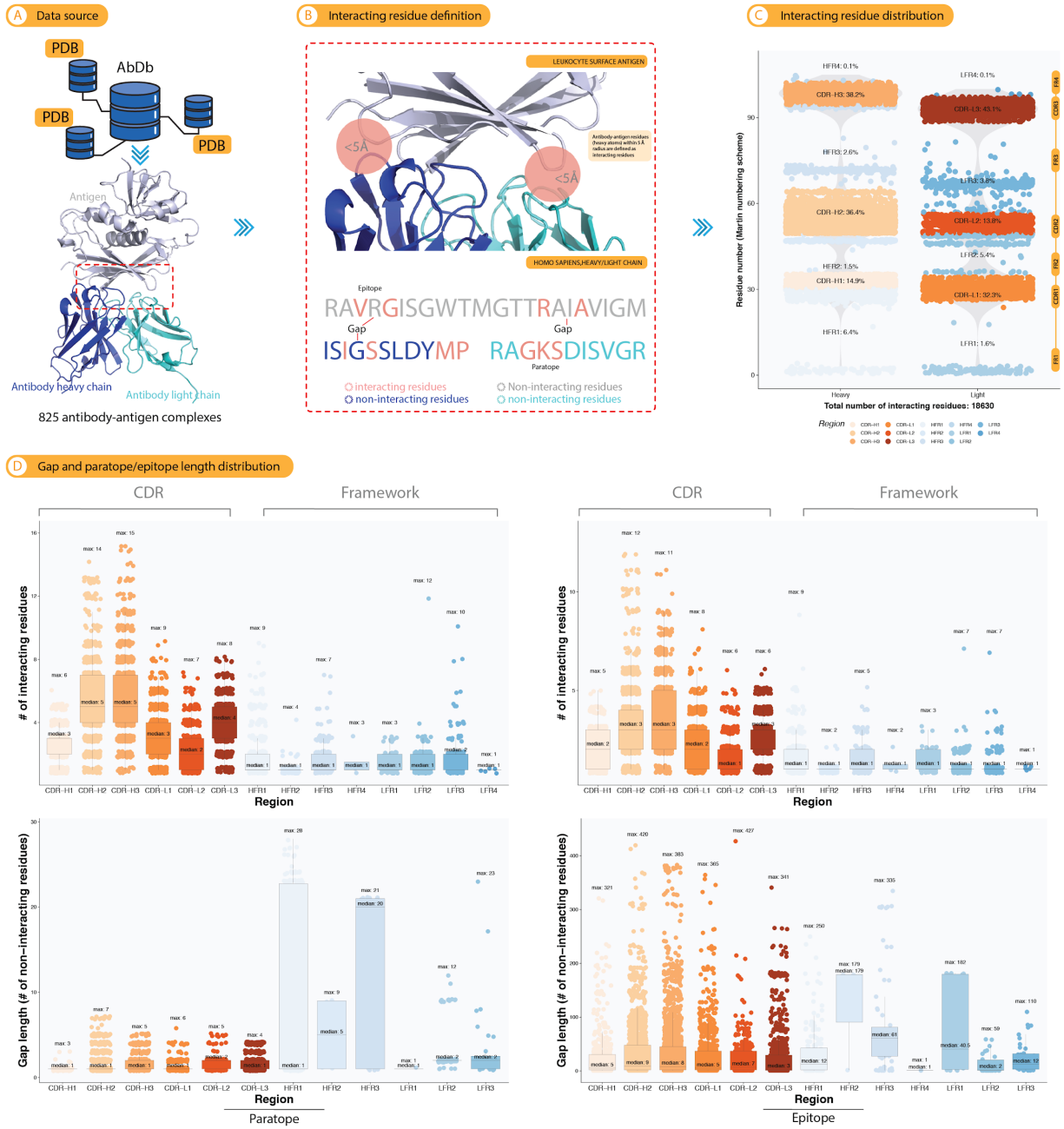


Figure 1. Characterization of interacting residues at the paratope-epitope interface. **(A)** We characterized antibody-antigen interaction using publicly available 3D structures deposited in the Antibody Database (AbDb) (see Methods), a total of 825 antibody-antigen complexes. Antigens are distributed across a wide range of antigen classes (Suppl Fig. 11) and antibodies are ≈90% of human and mouse origin (Suppl Fig. 11). **(B)** Interacting residues were defined as the set of residues in the antibody-antigen interface that were found within a radius of <5Å from each other. These residues are termed *paratope* in antibody and *epitope* in antigen. More specifically, a paratope was defined as the set of interacting amino acid residues within a particular FR or CDR region of an antibody. An epitope is defined as the set of antigen amino acid residues that interact with a paratope. Epitopes are annotated according to the FR or CDR regions of the corresponding paratopes. Gaps are defined as the non-interacting residues that lie between interacting residues. **(C)** The interacting residues mapped predominantly to the CDRs and less so to the framework regions (sequence annotation follows the Martin numbering scheme, see Methods). **(D)** The number of interacting residues ranged between 2–5 (CDR, median) and 1–2 (FR, median) in paratopes and 2–3 (CDR, median) and 1 (FR,

median) in epitopes. Gap lengths (number of non-interacting residues) ranged between 1–2 (CDR, median) and 1–20 (FR, median) in paratopes and 3–9 (CDR, median) and 1–179 (FR, median) in epitopes.

Paratope interacting residues are enriched in aromatic residues whereas epitope ones are enriched in polar and charged residues

The residues tyrosine (V_H: 18.2% and V_L: 20.0%), and serine (V_H: 11.2% and V_L: 15.7%) were the two most frequent interacting residues found in paratopes. For epitopes, lysine (V_H: 8.9% and V_L: 9.6%), and arginine (V_H: 8.7% and V_L: 11.2%) were found to be the most abundant residues at the paratope-epitope interface (Suppl Fig. S2A–B). The enrichment of aromatic residues (namely tyrosine) in paratope sequences and the abundance of polar and charged residues in epitopes (e.g., lysine and arginine) are in accord with published literature (Peng et al., 2014; Ramaraj et al., 2012) validating further the robustness of our definition of interacting residues. While overall the amino acid usage was more uniform in epitopes, both in paratope and epitope methionine and cysteine were heavily selected against. In fact, at least in the gapped residues of paratope and epitope, cysteine was more common (Suppl Fig. S2).

In terms of amino acid usage, the correlation between heavy chains and light chains of paratopes was very high (>0.80) as was the correlation between the heavy and light-chains of epitopes (>0.87, Suppl Fig. S6E). Paratope versus epitope residue usage correlation was far lower (0.13–0.49). We also investigated the correlation between paratope and epitope residue usage with that of non-immune protein-protein interaction (PPI) residues and found that PPI and paratope were not correlated (-0.03–0.07). PPI and epitope were only weakly correlated (0.16–0.55, Suppl Fig. 6D,E).

Finally, we investigated to what extent amino acid contact pairs of paratope and epitopes differ across CDR/FR regions and to what extent these amino-acid binding preferences overlap with those of non-immune protein-protein interaction (Suppl Fig. S15). We found that amino-acid level interaction preferences differed across CDR/FR regions with no discernable pattern and with substantially larger cross-type (type: charged, polar, amphiphatic, hydrophobic) interaction both at a local (CDR/FR region wise) and a global level (composite; Suppl Fig. S15A). In contrast, amino acid interaction preferences were much clearer for PPI where we found that amino acid levels preferred to interact within-type rather than across types.

Structural interaction motifs enable the comparison of paratope-epitope interfaces of unrelated antibody-antigen complexes

Paratopes and epitopes may manifest in two ways: (i) as a continuous stretch of interacting residues or (ii) discontinuously, separated by one or more non-interacting residues (gaps) due to protein folding (Barlow et al., 1986; Stave and Lindpaintner, 2013) (Fig. 1B). Indeed, we found that paratope lengths (number of interacting residues) ranged between 1 and 15 (median: 1–5) in the CDRs and 1–12 (median: 1–2) residues in the FR regions whereas the length of gaps (number of non-interacting residues) ranged between 1 and 7 (median: 1–2) residues in CDRs and 1–28 (median: 1–20) residues in the FR regions (Fig. 1D). Epitope can span up to 12 residues long with gaps up to 427 residues long (Fig. 1D).

The fundamental units of antibody-antigen interaction are the sequence regions on the antibody and the antigen (termed paratopes and epitopes) that compose the interacting and non-interaction residues at the paratope-epitope interface. As of now, it remains challenging to compare paratope-epitope interfaces across in unrelated antibody-antigen complexes. To this end, we devised a structural interaction motif notation that accounts simultaneously for gaps and residues in both paratopes and epitopes. A paratope or epitope structural interaction motif is composed of interacting paratope and epitope amino acid residues as well as non-interacting ones (gap). Specifically, we encoded any interacting residue as capital X and any gaps as integers. Here, the integer quantifies the number of non-interacting amino acid residues (Fig. 2A). The combination of amino acid and gap encodings is termed *structural interaction motif* (from here on: interaction motif or simply, motif). For instance, the paratope DGS (three consecutive interacting residues) is denoted as a XXX interaction motif and the epitope VRAG (two interacting residues, V and G, separated by two non-interacting residue, R and G) is denoted as X2Xmotif. Both motifs (XXX and X2X) have the same length: three. Our motif notation for antibody-antigen interaction places the paratope-epitope interface into a common coordinate system thereby preserving the link between paratope and epitope and enabling universal computational traceability of both continuous and discontinuous (structural) antibody-antigen interaction across antibody-antigen complexes.

Figure 2 | Structural interaction motifs represent a finite vocabulary for the composition of the paratope-epitope interface. (A) We devised a structural interaction motif notation that accounts simultaneously for gaps and residues in both paratopes and epitopes. A paratope or epitope structural interaction sequence motif is composed of interacting paratope and epitope amino acid residues as well as non-interacting ones (gap). Interacting residues are encoded with

the letter X and non-interacting residues are encoded with an integer quantifying gap size. For example, the string X1X encodes a paratope or epitope motif of two interacting amino acid residues (X,X) separated by one non-interacting residue (gap). **(B)** Length distribution of paratope and epitope motifs. The length of a motif is defined as the number of characters in a motif. For example, motifs X1X and X2X are both of length 3. **(C)** Pearson (range: 0.49–1.00) and Spearman (range: 0.47–1.00) correlation of paratope and epitope lengths. **(D)** Absolute and relative overlap of paratope and epitope motifs (Venn diagram). **(E)** Accumulation curves indicate that the motifs detected in the 825 AbDb crystal structures cover up to 67% (paratope) and 33% (epitope) of the potential motif diversity. Each boxplot indicates the median percentage of motifs of a given crystal structure that were already contained in the accumulation of all previous complexes. Boxplots summarize 1000 random crystal structure orderings. See *Methods* for details on the construction of the accumulation curves. **(F)** Extrapolation of the accumulation curves using the Chao1 estimator (see *Methods*) suggested that the respective potential diversity of paratope and epitope motifs remains below 10^4 distinct motifs.

Paratope and epitope motif lengths varied across FR and CDR regions but remained below a length of 10 (median length: 1–7, Fig. 2B). Motifs were generally shorter in the FR (median length: 1–2) compared to the CDR (median length: 3–7). Interestingly, motifs of CDR-H2 (median length: 7) were longer than those in the CDR-H3 (median length: 6). Paratope and epitope motif lengths showed consistently positive correlation in CDRs ($r_{\text{Pearson,CDR-H1–3 (L1–3)}}$ = 0.73–0.77 (0.5–0.84), $r_{\text{Spearman, CDR-H1–3 (L1–3)}}$ = 0.7–0.74 (0.48–86), Fig. 2C). Furthermore, motifs were substantially shared across CDR and FR in both paratope and epitope (although in epitopes to a lesser extent that in paratopes) and were thus not exclusive to a specific CDR or FR (Suppl Fig. S4A, B lower triangle). On average, three to four motifs were found per antibody chain (Suppl Fig. S9A).

With our motif language, we can now query key parameters of antibody-antigen recognition: (i) motif sequence diversity, (ii) structural diversity (motif angle and (dis)continuity), (iii) co-occurrence across complexes and (iv) predictability of paratope-epitope interaction. We found no difference between interaction motifs found in HIV broadly neutralizing antibodies (bNAb), widely regarded as hallmarks of unusual antibody-binding behavior, and non-bNAb (Suppl Fig. S5A). Furthermore, the combined set of paratope and epitope motifs were not overlapping with motifs found in non-immune protein-protein interaction (PPI). Specifically, (1) PPI motifs were generally a few characters larger than paratope and epitope motifs (median length 7 in paratopes vs 8–12 in PPI, Fig. 2B, Suppl Fig. S6B), and (2) while 70% of paratope motifs were found in PPI motifs, only 11.4% of epitope motifs were congruent with PPI motifs (Suppl Fig. S6A).

The diversity of paratope and epitope interaction motifs is restricted

We then asked to what extent interaction motifs of paratopes and epitopes overlap. Out of the total number of 1594 epitopes and 398 paratope motifs, only 80 unique motifs overlapped (Fig. 2D). 61% of the shared motifs were continuous and 39% were discontinuous.

Subsequently, we asked how much of the potential paratope and epitope motif diversity is covered by our dataset of 825 antibody-antigen crystal structures? To answer this question, we constructed a coverage accumulation curve for both paratope and epitope motifs and extrapolated the curve to obtain the estimate for the potential motif diversity using the Chao1 estimator – a method often used in ecology for estimating population sizes (Fig. 2E, see *Methods* for details). For paratopes, the set of unique motifs in our dataset covered about 60% of the potential diversity of all epitopes, whereas the set of unique epitope motifs in our dataset covered 33% of the total epitope diversity. The coverage is in accord with the estimates obtained from the Chao method

(Chao, 1984), where the total diversity was estimated to be 737 for paratopes and 10,507 for epitopes (Fig. 2F). Thus, our dataset captures a substantial portion of the total motif space indicating the restriction of the paratope-epitope interaction space. The estimated potential motif space is smaller ($<10^4$) by 10 orders of magnitude than the total number of antibody sequences ($>10^{14}$, (Briney et al., 2019; Elhanati et al., 2015)). The estimated potential size of the paratope motif space is one order of magnitude smaller than the theoretical size ($\approx 10^4$, see Suppl Text for analytical derivation).

The structure of paratope and epitope motifs differs across CDRs and FRs

Next, we asked whether interaction motifs differ structurally. To address this question, we measured the "angle" of each motif. The angle of an interaction motif was computed by defining two vectors spanning the midpoint of a motif and its start and end positions. Larger angles would indicate that the structure of the motif is more extended whereas small angles indicate that the motif would manifest as a loop (see more details in the *Methods* section). In general, median epitope motif angles were only maximally as high as paratope motif angles. We found that median motif angles in heavy and light chain of CDR3 did not substantially differ between paratopes and epitopes (Suppl Fig. S8A). This is in contrast to previous studies that found epitopes prefer to localize on the planar parts of the antigen (Rubinstein et al., 2008) manifesting as flat oblong elliptical shapes (Kringelum et al., 2013), at least for the CDR3 region. Differences of CDR motif angles among regions were more apparent for example for paratope (median): CDR-H1: 130º, CDR-H2: 42.5º, CDR-H3: 57º, reflecting the structural diversity of antibody-antigen interaction, specifically, CDR-H2 and CDR-H3 were "loopier" whereas CDR-H1 tended to be linear. In contrast to CDRs, median FR paratope angles were mostly larger than epitope angles. However, since the data points were much sparser for FR compared to CDR, these observations should be viewed with caution. Paratope and epitope angles correlated moderately positively in the majority of the regions (max $r_{\text{Pearson,CDR-H1–3}} = 0.57$, max $r_{\text{Spearman, CDR-H1–3}} = 0.55$, Suppl Fig. 8B).

Paratope motifs are shared across antibody-antigen complexes

Since our structural motif encoding of the paratope and epitope interface discriminates between continuous and discontinuous motifs, we asked to what extent paratopes or epitopes are continuous or discontinuous. To this end, we counted both the number of continuous and discontinuous motifs for each region (Fig. 3A) and the number of complexes that share identical motifs (Fig. 3B). We found that: (i) CDR-H3 is an obligate region for antibody-antigen interaction because the CDR-H3 is the only region that had interacting residues in each antibody-antigen complex investigated whereas all other CDR/FR regions were at least partially dispensable for antigen interaction, (ii) paratope motifs are predominantly continuous (except in CDR-H2, CDR-L1, and CDR-L3) and (iii) continuous paratope motifs, more so than discontinuous, are shared across antibody-antigen complexes; epitope motifs exhibited substantially less sharing and were more discontinuous. Specifically, when the motifs were filtered only for those which were found in 10% (or 82 in absolute numbers) or more complexes, ten paratope motifs (X, XXX, XX, X1X, XXXX1X, XXXX, XXXXX, X2X, XXXXXX, XXX1X, Suppl Fig. S4H) and five epitope motifs (X, XX, X1X, X2X, X3X, Suppl Fig. S4I) stood out. Six out of the ten top-shared paratope motifs were continuous, all of which are no more than seven residues long (including gaps) indicating that short motifs mediate a substantial proportion of antibody-antigen interactions. Indeed, the 20 top

shared paratope motifs made up 78% out of all encountered motifs and the 20 top shared epitope motifs made up 58.5% (Suppl Fig. 4D). Importantly, shared paratope interaction motifs were not specific to a given class of antigens (Suppl Fig. S4F), or to a specific germline gene (Suppl Fig. S4G) and 13 of the most shared interaction motifs were also found in HIV-bNAb (Suppl Fig. S5B) underlining the generality and diversity restriction of the interaction motifs here investigated.

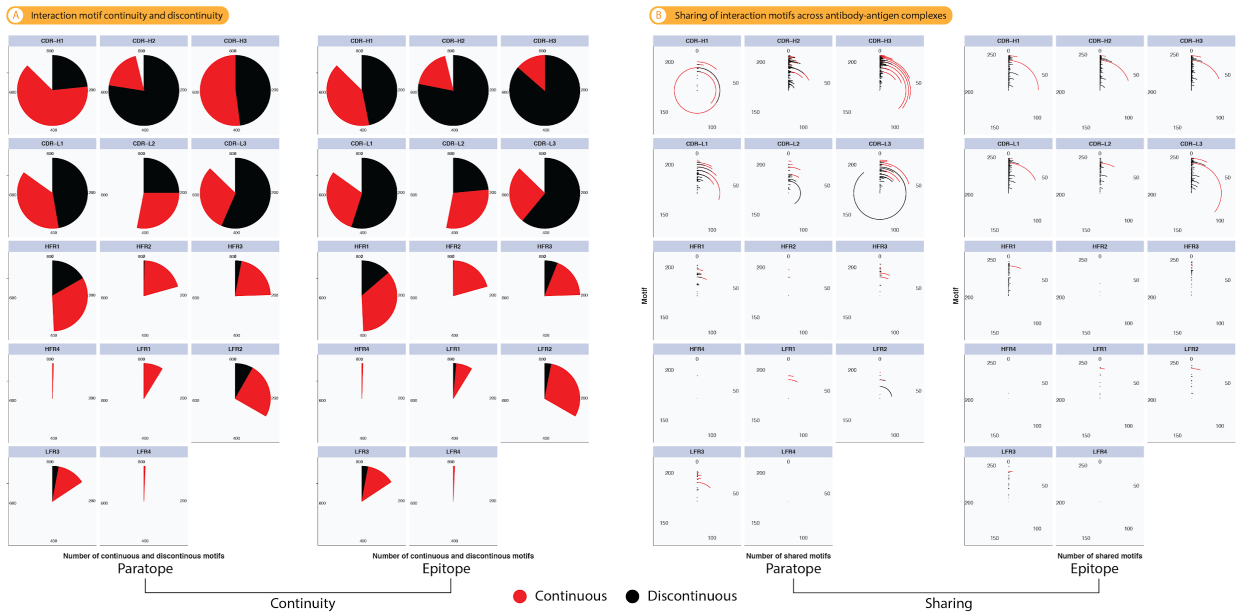


Figure 3 | Paratope interaction motifs show a higher extent of continuity and sharing across antibody-antigen complexes than epitope interaction motifs. (A) Ratio of continuous (absence of non-interacting residues) and discontinuous (presence of at least on non-interacting residue/gap) motifs paratope and epitope interaction motifs. **(B)** Absolute number of antibody-antigen crystal structures a given interaction motif is found in. The 20 most shared motifs are explicitly displayed in Suppl Fig. S4. Gaps in pie charts indicate that for a given region not all crystal structures showed interacting residues.

Each interaction motif has a unique sequential amino acid signature

To investigate the sequence dependencies within selected multi-residue (length>1) paratope and epitope interaction motifs, we determined the 2-mer decomposition of each paratope/epitope sequence (by a sliding window technique) of the ensemble of paratope/epitope sequences mapping to the respective motif (see Methods). We visualized thus computed sequential motif dependencies (for 4 interaction motifs, top shared 3 from both paratope [XXX, XX, X1X] and epitope [XXX, XX, X2X]) as directed networks where the 20 amino acids are the nodes and the edges are the frequency with which a 2-mer sequential dependency occurred (Fig. 4A).

Visually, paratope and epitope motif sequential signatures differed drastically from one another (Fig. 4A). To quantify this difference, we showed that sequential dependencies were largely non-overlapping (Fig. 4C) and if overlapping, were numerically different (Fig. 4B).

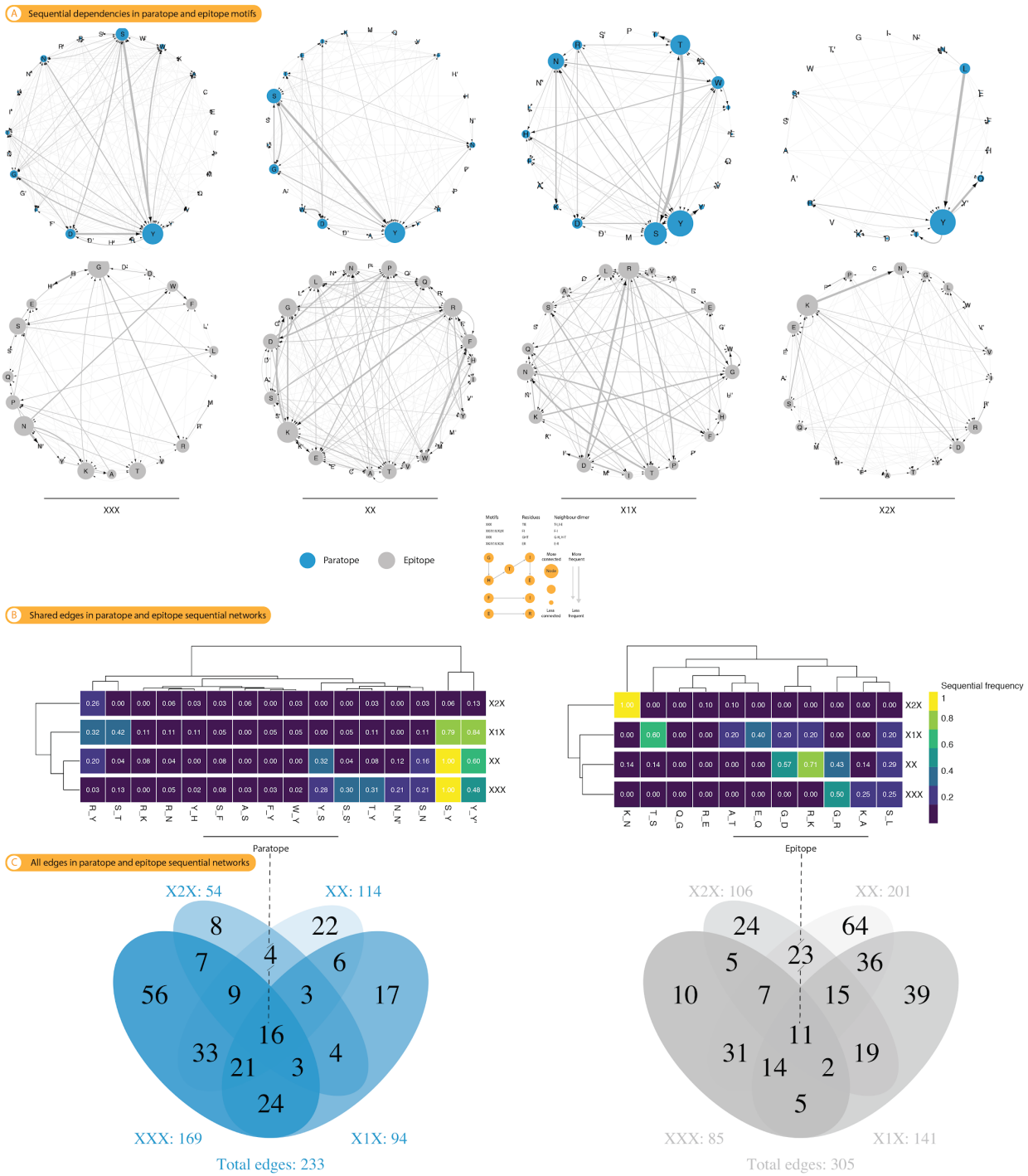


Figure 4 | Each interaction motif has a unique sequential dependency signature. (A) For each of the four most highly shared (across crystal structures) interaction motifs (Suppl Fig. 4), the sequential dependency signature was determined. Briefly, for the ensemble of paratope/epitope sequences mapping to a given interaction motif, the 2-mer decomposition of each paratope/epitope sequence was determined by sliding window (2-mer frequency distribribution). For each motif, these sequential dependencies were visualized as directed networks where the 20 amino acids are the nodes and the edges are the frequency with which a 2-mer occurred (sequential dependency). **(B)** Hierarchical clustering of sequential dependencies (network edges shown in (A)) that were shared between all four paratope or

epitope motifs. **(C)** Overlap of sequential dependencies (two-mers, network edges shown in (A)) shared across paratope or epitope motifs.

Finally, we compared the paratope and epitope sequence signatures to those of non-immune protein-protein interaction (PPI) and found that immune and non-immune interaction motifs, in particular that of paratope, have distinct sequence dependencies for the same motifs (Suppl. Fig. S13).

In summary, each motif, in paratopes as well as epitopes, has a unique sequential amino acid signature and the underlying sequence dependencies found in paratope-epitope interaction cannot be derived from non-immune protein-protein interaction.

The majority of paratope and epitope motifs are oligoreactive while a selected number of paratope motifs shows broad polyreactivity towards mutually exclusive epitope motif spaces

We next asked whether paratope and epitope motifs have preferred motif partners or whether interaction is mostly random. To answer this question, we built a paratope-epitope-motif network (a bipartite graph) by connecting each epitope motif to its cognate paratope motif (Fig. 5A). We call this network a *reactivity network.* In this network, we found that the top 7 connected motifs were paratope motifs (mostly continuous or with one gap) that made up 17% of all connections in the network (Fig. 5A,C). Even more importantly, together these top 7 paratope motifs made 829 connections to predominantly different epitopes (Fig. 5E, inlet) thereby collectively binding to ≈50% of all unique epitope motifs (Fig. 2D). Thus, while these paratope motifs showed broad polyreactivity, they bound to largely entirely different epitope motif groups. This finding is in line with the fact that each motif occupies a different sequential dependency space (Fig. 4). Apart from highly connected paratope (and very few epitope) motifs, paratope and epitope motifs were lowly connected and thus very specific with respect to their interaction partners. To statistically confirm this observation, we found that the degree distribution of the reactivity graph was power law-distributed and scale-free (p-value = 0.11, alpha = 2.04) (Clauset et al., 2009). To examine whether the connectivity patterns observed were simply due to the fact that there are more epitope than paratope motifs, we sampled 100 times 1000 paratope and epitope motifs and analyzed the resulting reactivity networks and degree distributions. In contrast to the observed reactivity network (Fig. 5A), the reactivity network obtained from random sampling showed (i) a more even distribution of connectivity for both paratopes and epitopes that was not power law (p-value<0.1) and with lower maximal node degrees and (ii) an increased overlap in bound partner motifs (Fig. 5B,D,E, inlet).

Furthermore, we asked whether the reactivity map observed for antibody-antigen interaction was immunity-specific. Therefore, we drew the PPI motif reactivity network for all available protein crystal structures (≈10^5). We found that PPI reactivity networks are exponentially distributed, lying between power-law and random networks (Suppl Fig. 12). Thus, similarly to the motif sequential dependencies and the amino acid usage, paratope-epitope reactivity cannot be derived from generic protein-protein interaction.

To summarize, the top-connected motifs in the reactivity network show polyreactivity towards epitope spaces that are non-overlapping. Most motifs, however, are oligoreactive (low number of partner interaction motifs) and thus highly specific. The observed reactivity network is fundamentally different from networks constructed from randomly sampling the observed paratope and epitope degree (connectivity) distributions and networks derived from generic protein-protein interaction.

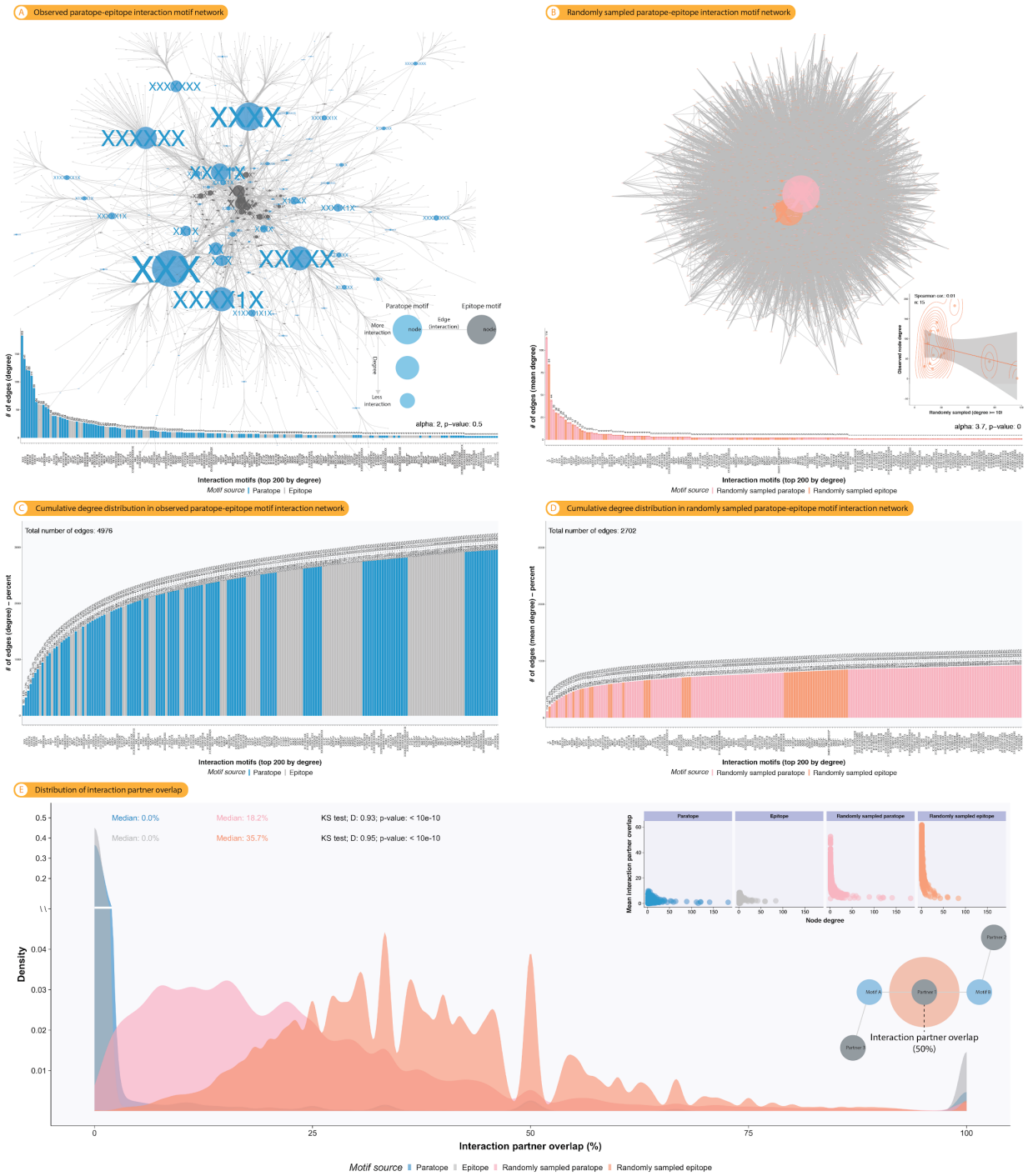


Figure 5 | The majority of paratope and epitope motifs are oligoreactive while a small number of paratope motifs shows broad polyreactivity towards mutually exclusive epitope motif spaces. **(A)** A bipartite reactivity network capturing paratope-epitope motif interaction was constructed by connecting each paratope motif to its corresponding epitope motif (undirected edge). Network vertices were scaled by their number of connections (degree). Only the largest connected portion of the network was visualized, for full network see Suppl Fig. 16. Degree distribution was tested to fit power-law distribution by calculating a goodness-of-fit value with bootstrapping using the `poweRlaw R 0.70.2` package (Gillespie, 2015) as described by Clauset and colleagues (Clauset et al., 2009). Here, a network whose degree distribution fits a power-law distribution (exponent between 2 and 3) is defined as scale-free (Broido and Clauset, 2019). A node's degree is the count of its connections from other nodes. **(B)** To confirm that the reactivity network architecture observed in (A) was unlikely to be observed by chance, we randomly sampled 100 times 1000 motifs from paratope and epitope motif distributions (Suppl Fig. 4D,E). Inlet: The Spearman correlation of node degree correlation of observed and randomly sampled networks is shown. For both networks, the respective node degree distribution is shown (for B, the standard error of the mean is also shown). Power-law fit was done as described in (A). **(C, D)** Cumulative degree distributions of networks (A) and (B**). (E)** Distribution of interaction partner overlap for networks (A,B). Briefly, for example for all paratopes in (A), the pairwise overlap of bound epitope motifs was calculated. The statistical significance of the difference between overlap distributions from (A) and (B) was computed using the Kolmogorov-Smirnov (KS) test. Inlet: the correlation of node degree and interaction partner overlap was determined.

Machine-learning prediction of paratope-epitope interaction

Predicting a cognate paratope for a given epitope (and vice versa) remains a challenge in the antibody field. When paratopes and epitopes are examined as motifs, the previous sections have illustrated that they possess signatures that are unique and distinct from each other. Briefly, paratope motifs tend to be continuous, shared, and enriched by a set of residues that are sequentially distinct yet complementary to epitope motifs. Hence, we asked whether these signatures can be leveraged to predict (translate) one paratope interaction motif (or sequence) into the cognate epitope interaction motif (or sequence) and vice versa. To this end, we trained a range of deep learning-based Neural Machine Translation models (121 models per use case with 10 replicates) on pairs of paratope-epitope interaction motif/sequence (Fig. 6, bottom panel) and evaluated their performance by comparing (i) the magnitude (length) and (ii) edit distance (error) of the predictions and the true motifs or sequences (see Methods). We opted to use Neural Machine Translation, a deep-learning sequence to sequence model, over classical classification models such as logistic regression or Naive Bayes for the reason that the prediction of one motif from another is inherently a sequence to sequence problem. That is, we aim to map a fixed-length input to a fixed-length output in which the length of the input and output may differ. For example, a paratope motif X2X pairs with an epitope motif XX, the former is of length 3 whereas the latter of length 2. The length of the input X2X is fixed at 3 characters and the output XX at 2 characters, however, the input-output length need not be identical (3≠2). A total of six use cases were trained (Table 1 and Fig. 6) and for each use case baseline prediction accuracy were calculated based on label-shuffled data where we trained models on randomly shuffled pairs of paratope-epitope with identical parameter combinations to the original translation tasks.

As the models approached the set of optimal parameters (hidden dimension and embedding dimension), we observed an increasingly positive correlation between the predicted lengths and the lengths of the true motifs (sequences) (Suppl Fig. 17A–B, $r_{pearson}$ 0.8–0.9). In contrast, models trained on randomized paratope-epitope pairs failed to recover the correct length indicating that the length for the motif or sequence is predictable. Although the length was equally predictable in both interaction motif and sequence task, the sequence content (indicated by edit distance/error) is not equally predictable (Fig. 6, top panel). The medians of prediction error of interaction motif use cases (0.3–0.4) were substantially lower than the medians of prediction error of sequence

use cases (0.9–1). The medians of prediction error for the control baseline use cases ranged from 0.7–0.8 and 1 for interaction motif and sequence, respectively. Together, these results indicate that the paired paratope-epitope interaction motif space, to a reasonable extent, is predictable whereas the sequence space remains challenging to predict.

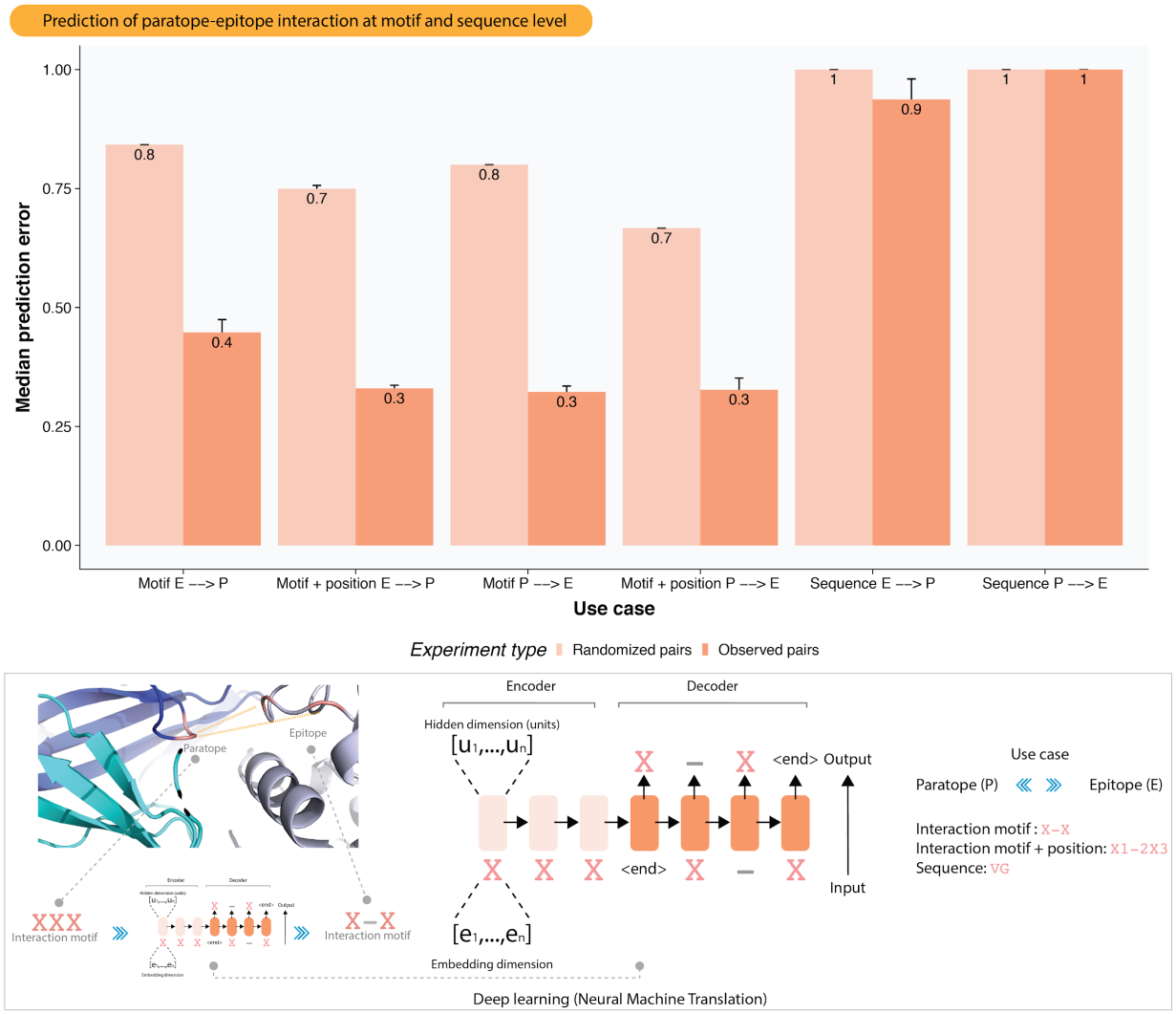


Figure 6 | Prediction of paratope-epitope interaction at interaction motif and sequence level. **(Bottom)** Schematic of the paratope-to-epitope and epitope-to-paratope prediction task. A deep learning model was trained on pairs of epitope-paratope interaction motif/sequence (and vice versa). Here, pairs of input-output sequences were translated via a combination of two components: encoder and decoder with gated recurrent units. The hyperparameters *hidden dimension* (1, 2^1, 2^2,..., 2^{10}) and *embedding units* (1, 2^1, 2^2 ,..., 2^{10}) were optimized during the 20 epochs-long training. A total of 121 models were trained for each use case with 10 replicates (see Methods for details). (**Top**) The median prediction error was obtained by calculating the median Levenshtein distance between the output and the predicted output for each use case. The distance ranges from 0 (perfectly matching output-predicted-output, high prediction accuracy) to 1 (fully dissonant output-predicted-output pairs, low prediction accuracy). Here shown is the mean of the medians from the replicates of each use case. Use cases cover the bidirectional prediction tasks (paratope to epitope as well as epitope to paratope) of motif to motif, motif with position to motif with position, and finally amino acid sequence to amino acid sequence (see Table 1). Baseline prediction accuracies (control) were calculated based on label-shuffled data where antibody and antigen-binding partners were randomly shuffled (randomized pairs).

Discussion

Summary

We discovered a finite vocabulary of antibody-antigen interaction in the form of structural interaction motifs. The existence of a finite vocabulary implies potential predictability of paratope-epitope binding. We have, therefore, provided evidence for the existence of sequence-based rules in 3D antibody-antigen interaction. By combining antibody-antigen interaction data with a (i) structural interaction motif language, (ii) network theory, and (iii) deep learning, we found that there exist paratope and epitope interaction motifs that are shared across antibody-antigen complexes. These motifs are predominantly simple (short and continuous), immunity-specific and their sequence diversity is restricted. We showed that each motif has unique sequential dependencies indicating that our motif definition captures underlying immunological principles. While most paratope and epitope motifs showed high specificity, we discovered a subset of paratope motifs with broad polyreactivity towards largely mutually exclusive epitope spaces thus partitioning the paratope-epitope interaction space. Via deep learning, we showed that knowing the epitope motif determines the paratope interaction motif, and vice-versa, with 70% certainty (median error 30%). To provide quantitative robustness to all of our findings, our study contains the most comprehensive statistical evaluation of antibody-antigen and non-immune protein-protein binding to date. Specifically, our study contains the largest volume of information on (i) the distribution of amino acid residues at the binding interface, (ii) the extent of binding interface (dis)continuity, (iii) quantification of interaction complex sequence similarity, (iv) the structural and thermodynamic definition of paratope and epitope interaction and (v) the relationship between somatic hypermutation sites and paratope-epitope contact residues. Together, our results demonstrate the existence of learnable sequence-based rules in 3D-antibody-antigen interaction. Below, we discuss the implications of our findings for the understanding of antibody specificity and the engineering of epitope-specific antibodies.

Antibody-antigen interaction operates via structural interaction motifs

In agreement with previous literature, we confirmed that most interacting residues are localized in the CDRs – albeit 10% of antibody-antigen contacts were localized outside the CDRs (Fig. 1C). Previous reports found that up to 20% of the residues lie outside the CDR region (*nota bene*: percentages are subject to numbering-scheme dependent variation) (Kunik et al., 2012a). Epitopes were discontinuous to a larger extent than paratopes (Fig. 3), which is in accord with previous literature (Barlow et al., 1986; Kringelum et al., 2013). Furthermore, we confirmed the previously observed enrichment of key amino acid residues in paratopes such as tyrosine and serine (Suppl Fig. S2). Importantly, the amino acid usage differed most between paratopes and non-immune PPI, whereas epitope amino acid usage was as similar to that of paratopes as to that of PPI (Suppl Fig. S6E).

Building on these confirmatory results, we showed that interaction motifs were shared across entirely different antibody-antigen complexes. The search for fingerprints of interaction has also

been recently reported for protein-protein interaction (Gainza et al., 2019). Thus, antibody-antigen interaction as a whole may be regarded as modular. Indeed, we found that only the VH-CDR3 region is an obligate region for antibody-antigen interaction. Seemingly, all other CDR and FR regions may be modularly added to fine-tune antibody recognition. The immune system is inherently a policing system that surveys both self and non-self for possible threats. Thus, it requires highly flexible recognition that is difficult to escape from via mutation – simple and small sequence motifs match this purpose. Interestingly, in HIV broadly neutralizing antibodies (bNAb), the same motifs as in non-bNAb were used thus underlining the generality of the motif vocabulary discovered and defined in this work (Chuang et al., 2019) (Suppl Fig. S5).

Importantly the discovery of interaction motifs crucially depended on the FR/CDR-focused definition of paratope and epitope (Fig. 1, see Methods). Whereas the common definition of paratope encompasses *all* interacting residues (Suppl Fig. S9), we opted for a *per CDR/FR* definition of paratope and epitope. In the future, once more antibody-antigen crystal structures become available, one may attempt to search for motifs based on the entire antibody and antigen (a given antibody V_H and V_L has a median of 3–4 motifs, Suppl Fig. S9).

Previously, Kunik and Ofran (Kunik and Ofran, 2013) showed the six antibody binding regions [ABR] (≈CDR-H/L1–3) differed significantly in their amino acid composition and that each ABR tends to bind different types of amino acids at the surface of proteins (Van Regenmortel, 2014). Van Regenmortel interpreted this as follows *"Because the entire accessible surface of a protein is a continuum of potential epitopes (Berzofsky, 1985), it could be argued that it would be advantageous for Abs to bind any protein surface patch without requiring specialized sites of increased stickiness. It seems that antibodies are in fact able to achieve this because they have evolved a set of ABRs where each ABR binds different types of amino acids, while the combined preference of the entire set is for epitopes that are indistinguishable from the rest of the protein surface"* (Van Regenmortel, 2014). While we were able to confirm that paratope-epitope amino level contact maps differ across CDR/FR regions (Suppl Fig. 15), we found that paratope interaction motifs were shared substantially across CDR/FR regions suggesting that binding spaces of CDR/FR regions are not as mutually exclusive as previously thought. Indeed, our reactivity network analysis suggested that binding spaces are partitioned on the motif level and not on the amino-acid level (Fig. 5). Specifically, structural interaction motifs encode geometric information of the local structure and are therefore linked to the angle of folding. Therefore, linking sequence to motifs and motifs to binding in a 2-step process may link in the future local folding to global specificity.

Interaction sequence motifs are immunity-specific and not biased to certain germline genes, antigen classes, or species

The 825 antibody-antigen complexes here investigated represent naturally only a mere fraction of the potential antibody-antigen interaction space. Although the collection of these complexes is biased towards antibody-antigen structures of past and current scientific interest, we verified that our dataset as well as the interaction motifs were not biased towards interaction definition-dependent or interaction-independent confounding factors. Specifically, we demonstrated that (i) our structural [distance cutoff-based] for calling interacting residues is robust, (ii) the 825 protein

antigens belonged to 84 different antigen classes (Suppl Fig. S11), (iii) shared paratope interaction motifs did not cluster by antigen class but instead were shared across antigen classes (Suppl Fig. S4), (iv) antibody and antigen sequences differed substantially in sequence across antibody-antigen complexes (Suppl Fig. S1), and (v) germline across germline V genes and (vii) species (human, mouse, Suppl Fig. S4).

Furthermore, we assembled a dataset of 100,888 protein-protein interaction structures (of which ≈30,000 structures had motifs similar to antibody-antigen data with respect to maximum gap size) to answer the question of whether the motifs in antibody-antigen structures are merely a subset of the larger class of protein-protein interaction (PPI) motifs. Indeed, it has been reported previously that, for example, protein-protein and antibody-antigen interaction do not differ in terms of shape complementarity (Kuroda and Gray, 2016) whereas other research suggested that antibody-antigen interaction differs from protein-protein interaction (Krawczyk et al., 2013; Raghunathan et al., 2012). Here, we found that paratope and epitope interaction motifs do not represent a subset of PPI motifs but are immunity-specific. Specifically, we found that protein-protein and paratope-epitope interfaces differ in terms of (i) amino acid acid composition and motif length (Suppl Fig. 6), (ii) sequential dependencies (Suppl Fig. 12) and (iii) global reactivity (Suppl Fig. 13).

Taken together, we validated in orthogonal controls that the rules of antibody-antigen interaction here described are universal and not merely a byproduct of potential confounding factors. Nevertheless, it is important to acknowledge that each crystal structure provides only a snapshot of the continuum of interactions occurring between antibody and antigen (*86*, *101*, *102*). Thus, future NMR-driven kinetic interaction data might reveal that motifs identified for a given antibody-antigen complex are actually part of a “quasi-species” motif space. An obvious extension of our approach is the addition of information on antibody-antigen affinity. While reliable affinity information was unavailable for the complexes studied herein, it will be of high interest to augment our interaction motif analysis with affinity data.

Epitope-specific recognition is not unilaterally encoded into the antibody sequence

Each of the paratope and epitope sequence interaction motifs possessed unique underlying amino acid sequence dependencies (Fig. 4) that are distinct of those found in non-immune protein-protein interaction (Suppl Fig. S6, S13). Furthermore, we showed that the polyreactive paratope motifs contact mutually exclusive epitope-motif spaces (Fig. 5). These findings support the notion that for the prediction of antibody-antigen interaction, both antibody (paratope) and antigen (epitope) information is crucial (Esmaielbeiki et al., 2016). It is important to note, however, that while *epitope*-motif spaces might be mutually exclusive across paratopes, *antigen* spaces might overlap since different epitope motifs may be contacted and located on the same antigen. Once more crystal structures become available, we will need to stratify epitope motifs by antigen class to better understand to what extent epitope motifs are exclusive to a given class. In fact, here we found that paratope motifs bound antigens of very diverse antigen classes (Suppl Fig. S4F). These further investigations will elucidate the question of whether paratope sequence motifs may be used as in silico scanning devices for epitope-specific sequence regions in high-

throughput antibody repertoire data for at least a subset of medically important antigens. Understanding the diversity and structure of antibody specificity on the repertoire level is also of importance for enhancing the precision of immune receptor-based immunodiagnostics where one of the most pressing current problems is a low signal-to-noise ratio (few antigen-specific sequences within a large pool of unrelated sequences) (Arora et al., 2019; Brown et al., 2019; Emerson et al., 2017; Ostmeyer et al., 2019).

Antibody-antigen recognition is overall oligoreactive with islands of high cross-reactivity: implications for humoral specificity

We identified not only predefined dependencies *within* paratope and epitope interaction motifs but also higher-order dependencies *among* paratope and epitope motifs. Specifically, we found that paratope-epitope interaction is power-law distributed (Fig. 5) with polyspecificity of few selected paratope motifs, general oligospecificity of the majority of paratope and epitope motifs. The highly polyreactive paratope motifs contacted mutually exclusive epitope space indicating an overall high degree of humoral specificity already on the motif level and not only on the amino acid level as previously thought. We did not observe a power law reactivity distribution in neither randomly sampled antibody-antigen data (Fig. 5) nor randomly sampled protein-protein interaction (Suppl Fig. S12). In fact, we found that protein-protein reactivity networks showed characteristics of exponential networks thus lie in between power law (antibody-antigen) and randomly sampled networks. Indeed, in agreement with previous extensive literature on protein-protein interaction networks, we confirmed the existence of hubs in protein-protein interaction (Kupiec, 2010; Nicholson, 2019; Stelzl et al., 2005). However, on the level of interaction motifs, such hubs were not as pronounced leading to a tree-branch-like structure (exponential degree distribution) as opposed to a power law degree distribution.

Power-law distributions have been repeatedly shown to occur in immune repertoires. We and others previously demonstrated that immune receptor repertoires store information in both the frequency as well as the sequence dimension in the form of power-law statistics (Greiff et al., 2015, 2017a; Miho et al., 2019). The interaction data here presented now suggest that paratope-epitope interaction may function as a third power-law governed layer of adaptive immunity. The great extent of organization in the paratope-epitope interaction space may be a corollary of the substantial antibody sequence bias observed in high-throughput antibody sequencing studies (recent high-throughput antibody sequencing studies (Briney et al., 2019; Miho et al., 2019; Soto et al.; Weinstein et al., 2009) have revealed high predictability and predetermination in both naive and antigen-specific antibody sequences). In future studies, we plan to investigate how sequence interaction motifs depend on the underlying immune receptor sequence generation probability landscape (Elhanati et al., 2015, 2018). While we do not yet understand the evolutionary selection pressures (Marchalonis et al., 2002; Stewart, 1992; Van Regenmortel, 2014) nor the functional significance of these power-law distributions in immune repertoires, their existence is a fundamental premise for the predictability of antibody-antigen interaction as they create a potential learnable space in which adaptive immunity operates.

Humoral specificity describes the capacity of the antibody immune system to selectively target a nearly infinite number of antigens. Given the large number of potential antigens, it is commonly

thought that antibody-antigen interaction is very challenging to predict. But if one approaches the challenge of understanding antibody-antigen interaction from a motif perspective, it breaks down the problem into a much lower dimensional task. Indeed, our extrapolation of motif diversity suggests that the potential epitope motif space is small (Fig. 2F). It is tempting to speculate as to the evolutionary advantage of short motifs in antigen recognition (non-immune protein-protein interaction has evolved substantially larger motifs, Suppl Fig. S6B,C). Short motifs decrease the potential escape space for antigen but also render self and non-self recognition more difficult. It would be interesting to investigate in how far, for example, autoimmunity and infections (bacterial, viral) occupy different motif spaces. So far, however, we observed that epitope motifs were shared consistently across antigen classes (Suppl Fig. S4F).

Implications for antibody discovery and engineering

Monoclonal antibodies are of substantial importance in the treatment of cancer and autoimmunity (Ecker et al., 2015). Thus, their efficient discovery is of particular interest. Our analyses suggest that the number of antibody binding motifs is relatively restricted (Fig. 1–3). Monoclonal antibody discovery is predominantly performed using synthetic antibody libraries. The number of developable hits of such libraries may be increased by tuning sequence diversity towards the interaction motifs (and their corresponding sequential bias) here discovered. Relatedly, engineering-driven computational optimization of antibody-antigen binding as well as docking algorithms might benefit from incorporating interaction-motif-based heuristics (Baran et al., 2017; Krawczyk et al., 2013; Kuroda and Gray, 2016; Mason et al., 2019; Sivasubramanian et al., 2009; Weitzner and Gray, 2016). Specifically, if we assume that the interaction motif sequential dependencies discovered here were evolutionarily optimized, they may be used to substitute for the lack of available multiple alignments that are used to calculate high-propensity interacting residues in protein-protein docking (Krawczyk et al., 2013). Furthermore, it will be of interest to investigate whether sequential dependencies are already predictive as to the antigen targeted (more paratope-epitope-paired data is needed for such investigations).

Furthermore, in contrast to earlier work (Raghunathan et al., 2012), we found that contact residues and somatic hypermutation are in fact correlated (Suppl Fig. S10). For antibody optimization this suggests that linking the antigen-contacting and somatically hypermutated positions (Suppl Fig. S10) in a high-throughput fashion and predicting whether the paratope prior to SHM was already binding or not, may enable, in theory, the construction of a hierarchy of evolutionary-driving driving SHM sites. Furthermore, it would be of interest to investigate in how far somatic hypermutation preserve binding motifs.

Finally, in the future, it may also be of interest to correlate interaction motifs with antibody developability parameters (Jain et al., 2017; Lecerf et al., 2019; Mason et al., 2019; Raybould et al., 2019b). Antibody developability depends on a multitude of parameters that are calculated based on the entire antibody complex (Andersen et al., 2011; Raybould et al., 2019b). Thus, all non-interacting residues also contribute to antibody developability calculations. Therefore, future studies have to delineate to what extent non-interacting residues correlate with specific interaction motifs. For example, for our dataset we did not see correlation between motif usage and germline gene (one proxy for non-interacting residue) usage (Suppl. Fig. S4).

Learnability and predictability of the paratope-epitope interface

Although there have been several attempts to predict paratope-epitope binding, prediction accuracy and generalization capacity have been generally suboptimal thus far (Brown et al., 2019). More fundamentally, the problem of antibody-antigen binding is generally regarded as too high-dimensional. We conjectured that if a general vocabulary for antibody-antigen interaction existed, then paratope-epitope interaction would be predictable. Indeed, we found that motif-based deep learning predicts paratope-epitope pairing reasonably well. Thus, we showed that paratope-epitope pairing is indeed predictable. Furthermore, we found that while interaction-motif pairing was predictable, the sequence interface of paratope-epitope interaction was not. Specifically, the paratope-epitope reactivity network (Fig. 5) allowed us to draw several conclusions and future directions with regards to the prediction of paratope-epitope binding. Given that there exist a few paratope motifs with broad epitope motif reactivity, motif-based prediction accuracy of paratope-epitope interaction cannot reach, by definition, 100%. However, as the epitope reactivity of the polyreactive paratopes was mutually exclusive, focusing prediction efforts on branches of the paratope-epitope reactivity network may improve the performance of sequence-based paratope-epitope prediction models – especially since we discovered that each motif has its own sequential dependency signature. Thus, the finite motif vocabulary of antibody-antigen interaction may help partition the high-dimensional antibody-antigen interaction space into smaller, less daunting prediction tasks thus rendering antibody-antigen interaction learnable.

Interaction sequence motifs provide ground truth for benchmarking immune receptor machine learning

Adaptive immune receptor repertoires represent a major target area for the application of machine learning in the hope that it may fast-track the in silico discovery and development of immune-receptor based immunotherapies and immunodiagnostics (Brown et al., 2019; Greiff et al., 2012; Mason et al., 2018, 2019; Miho et al., 2018). The complexity of sequence dependencies that determine antigen binding (Dash et al., 2017; Glanville et al., 2017), immune receptor publicity (Greiff et al., 2017b) and immune status (immunodiagnostics) (Ostmeyer et al., 2019; Thomas et al., 2014) represent a perfect application ground for machine learning analysis (Arora et al., 2019; Cinelli et al., 2017; Greiff et al., 2017b; Liu et al., 2019; Mason et al., 2019; Sidhom et al., 2018; Sun et al., 2017). As discussed extensively in a recent literature review by us (Brown et al., 2019) the development of ML approaches for immune receptor datasets was and is still hampered by the lack of ground truth datasets. Ground truth datasets are defined by the property that the link between the class of a sequence (class = disease/antigen specificity) or repertoire and the underlying sequence structure is known *a priori*. Thus, by definition, ground-truth datasets are those that are synthetically generated in silico (Weber et al., 2019a, 2019b). One of the current bottlenecks for generating nature-like synthetic datasets for predicting, for example, antigen-binding from the immune receptor sequence alone is the lack of knowledge on the complexity of paratope and epitope interaction motifs. Our research resolves this important knowledge gap. Specifically, our atlas structural interaction motifs now allows for the faithful simulation of native-like ground truth immune receptor datasets for the development of immune-receptor based machine learning methods (Weber et al., 2019a). For example, paratope and epitope sequence motifs may be implanted into synthetic 3D antibody-antigen structures (Robert et al., 2018) to

subsequently perform benchmarking of different ML architectures with regard to their capacity to (i) predict paratope-epitope binding and (ii) recover the exact sites of antibody-antigen interaction (native-like antigen-specific sequence motifs).

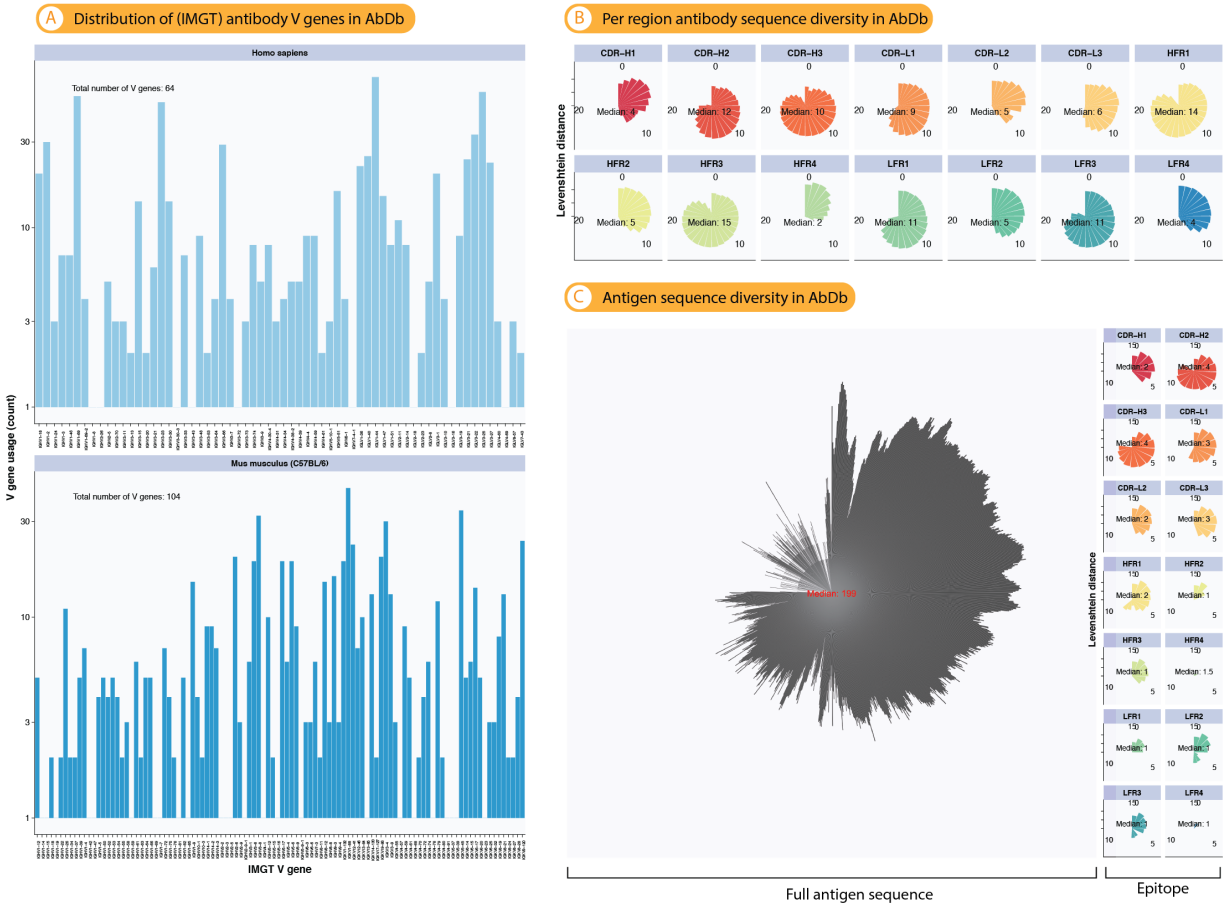
A word of caution: while we firmly maintain that for a profound understanding of antibody specificity, machine learning is of utmost necessity, machine learning alone will not suffice. Indeed, the discovery of general interaction sequence motifs was encoding-*dependent* and machine-learning-*independent*. As previously described at length (Brown et al., 2019), we believe that machine learning approaches adapted to the complexity and intricacies of immune receptor sequence data will have to be combined with sequence encodings that aim to answer specific questions. In fact, prediction tasks may be aided or complicated by a given encoding. In this work, for example, breaking down antibody-interaction into modular structural interaction motifs has greatly simplified machine learning tasks (Figure 6).

Concluding remarks

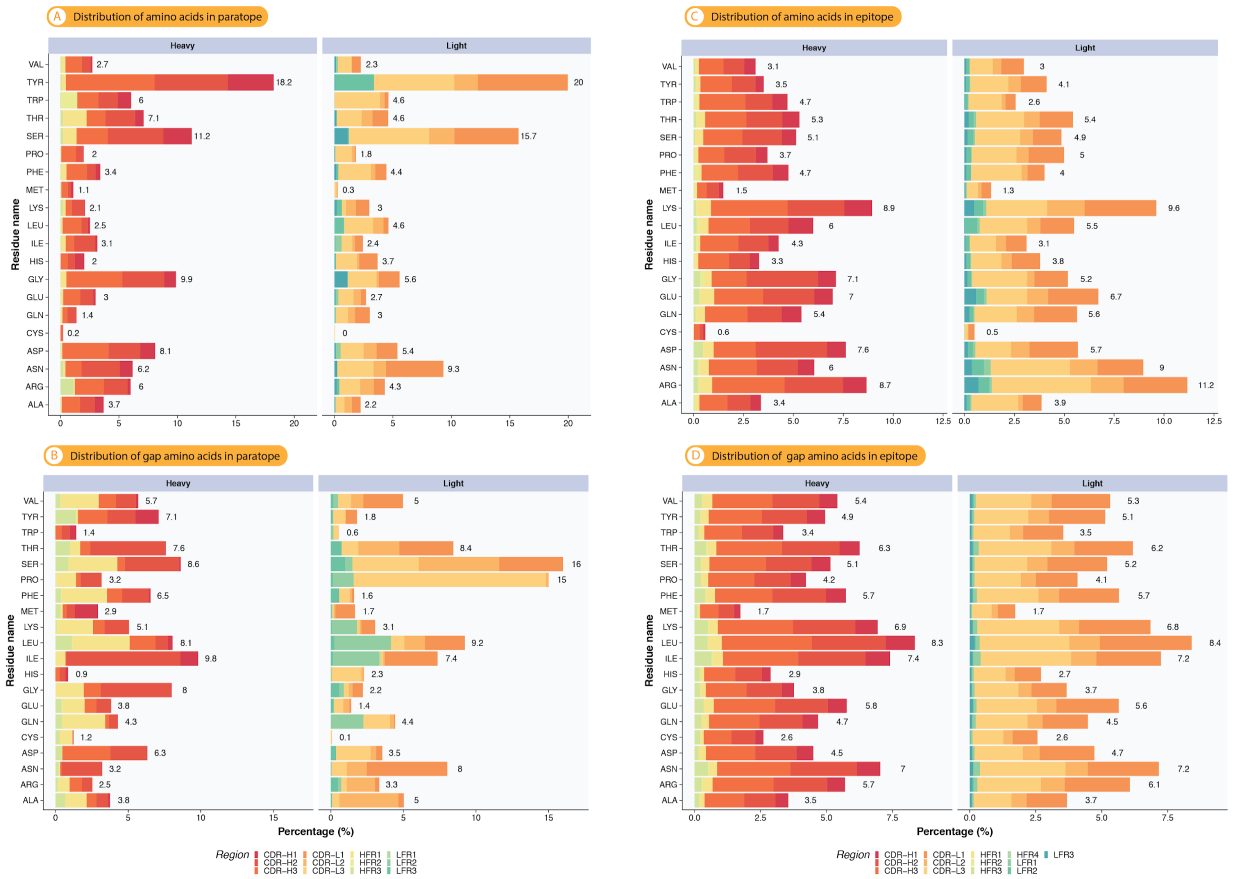
Antibodies constitute only half of the adaptive immune system. To our knowledge, a similar work on *TCR*-peptide interaction does not yet exist. Comparing motifs between TCR and antibody-antigen motifs would shed light on mechanistic similarities and differences in antibody and TCR antigen-interaction (Antunes et al., 2018; Bradley and Thomas, 2019; Dash et al., 2017; Glanville et al., 2017; Gowthaman and Pierce, 2018; Hellman et al., 2019, 2019; Lanzarotti et al., 2018; Ostmeyer et al., 2019; Riley and Baker, 2018; Turner et al., 2006). Furthermore, it remains to be investigated in how far the motif-based rules uncovered here for VH-VL-antigen complexes are transferable to scFV, nanobody, and other novel antibody constructs (Henry and MacKenzie, 2018; Mitchell and Colwell, 2018). Finally, we wish to state that, while one of the main aims of this paper was to advance our quantitative understanding of antibody-antibody recognition, the second main aim was to develop computational approaches that may help study antibody-antigen interaction in the years to come. Our systems-level approach of combining orthogonal statistical, network and machine learning approaches for the study of antibody-antigen interaction was necessary to reach the conclusions drawn in this work.

~ ~ ~ Fin ~ ~ ~

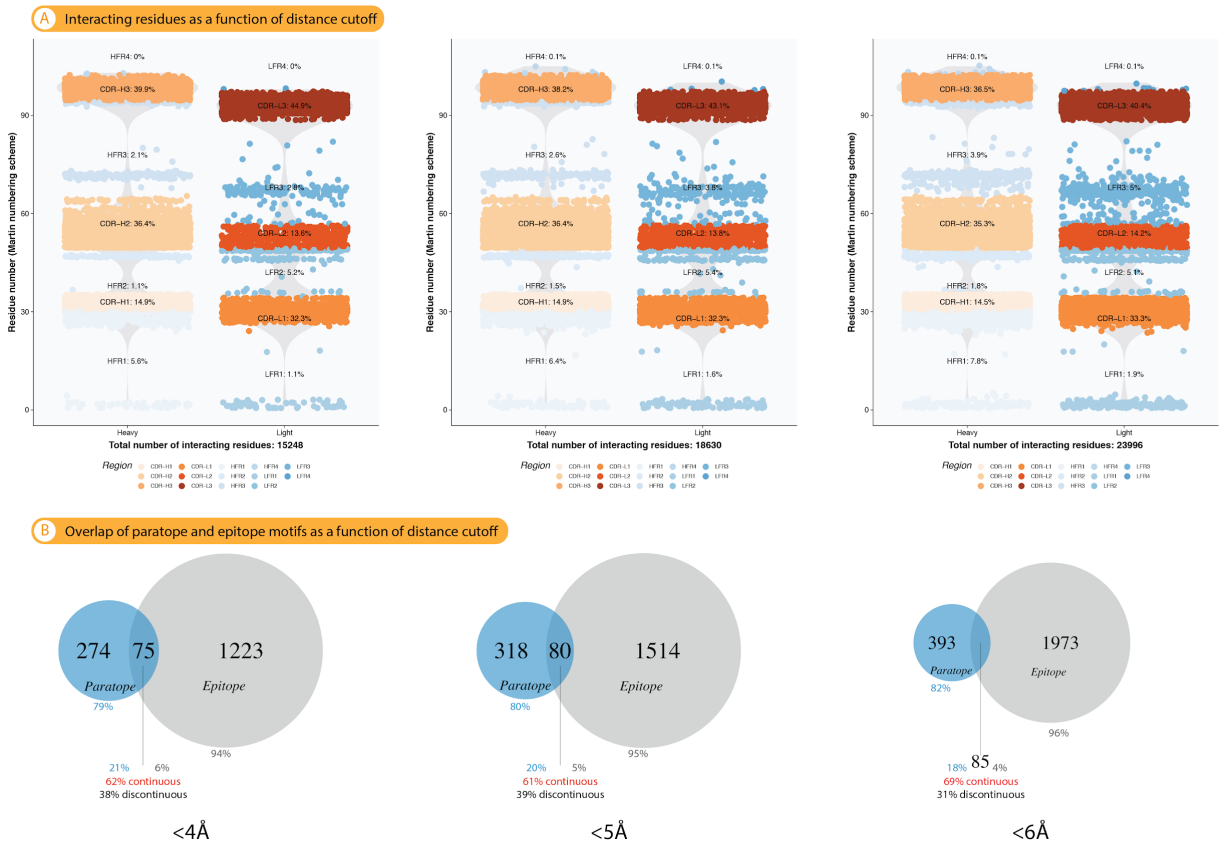
Supplementary Material



Supplemental Figure S1 | The AbDb dataset covers a large sequence diversity. (A) The 825 crystal structures show a large diversity in germline gene usage (human: 64, mouse: 104) covering a notable portion of the potential V germline gene diversity. **(B)** Distribution of pairwise sequence similarity across antibodies measured by Levenshtein distance. The median distance by region (CDR, FR) ranges between 2 (HFR4) and 14 (HFR1). **(C)** Distribution of pairwise sequence similarity across antigen sequences (both full and by epitope region) measured by Levenshtein distance. The median distance by region (CDR, FR) ranges between 1 (HFR2, HFR3, LFR1–4) to 4 (CDR-H3). The median pairwise distance of full length amino acid antigen sequences is 199.

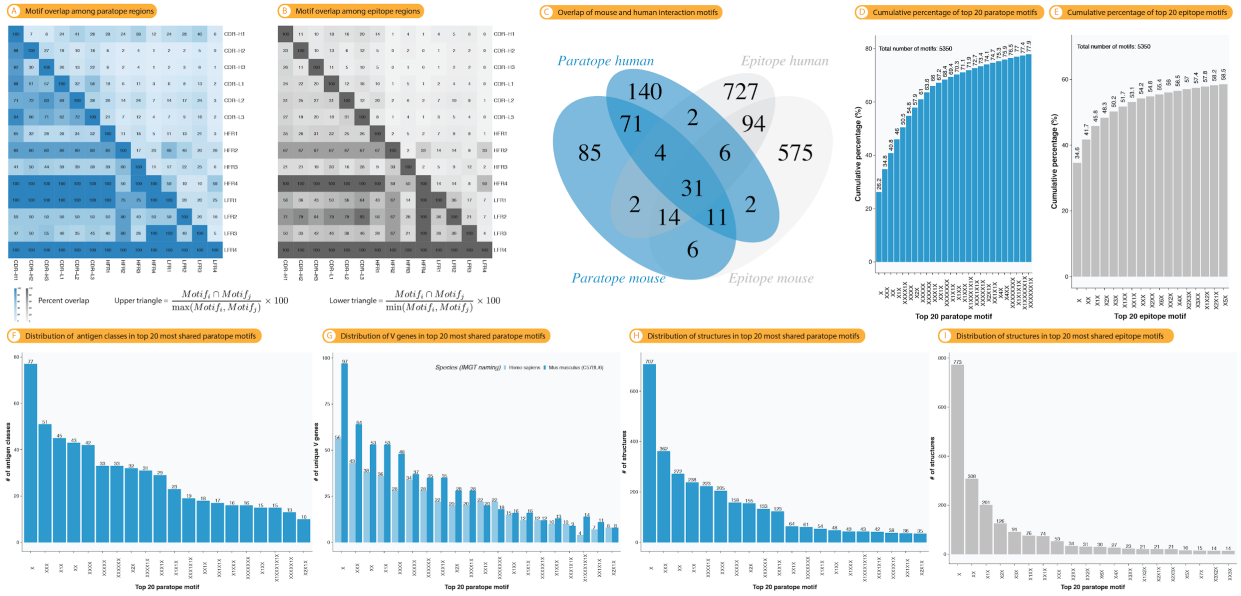


Supplemental Figure S2 | Amino acid residue distributions in paratope and epitope as well as gaps. (A) Relative residue distribution in paratopes. Aromatic residues namely tyrosine and serine are enriched in paratopes. Cysteine and methionine are the least used amino acids. **(B)** In contrast, gap residues in paratopes are more evenly distributed and in general enriched by a different set of residues. **(C)** In epitopes, polar and charged residues such as lysine and arginine are the two most frequently observed residues. In addition, epitope interacting residues are more evenly distributed than those in paratopes. **(D)** Similarly to paratopes (B), epitope gap residues are more evenly distributed and enriched by a different set of residues.

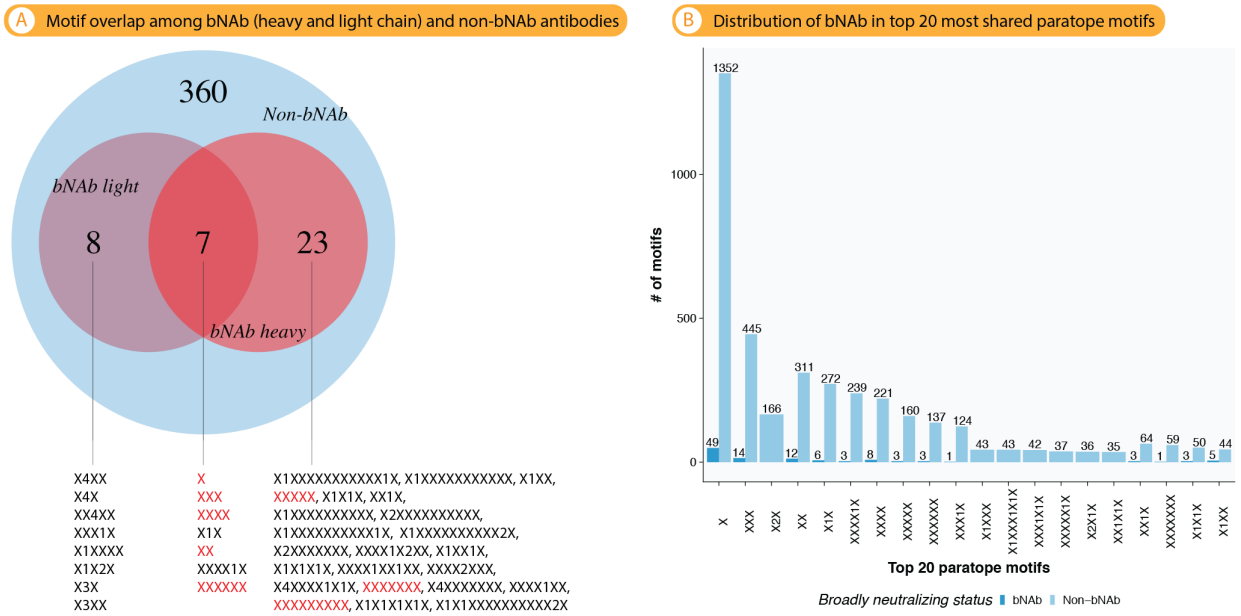


<4Å <5Å <6Å

Supplemental Figure S3 | Number of interacting residues and paratope and epitope motifs as a function of the distance cutoff that defines interacting residues. (A) The three panels show the total number of interacting residues as well as their relative fraction by CDR/FR for the three investigated and commonly used distance cutoffs: <4Å, <5Å, and <6Å. For all analyses in this study, we adopted the distance cutoff of <5Å (see explanation in Methods). The relative distribution of interacting residues across CDR/FR differed only marginally across distance cutoffs. **(B)** While the number of epitope motifs varied considerably across distance cutoffs, the number of overlapping motifs remained nearly constant.

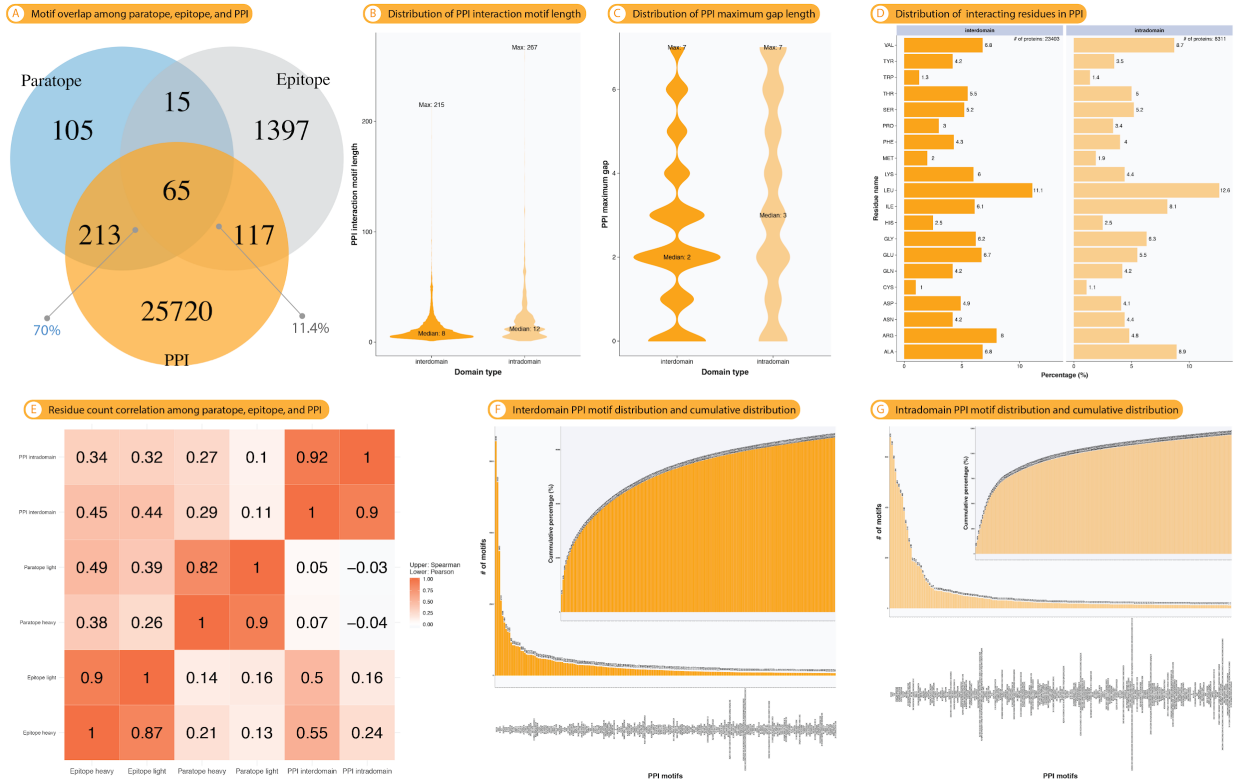


Supplemental Figure S4 | Interaction motifs shared across crystal structures cover a large antibody and antigen sequence space. Extent of motif sharing (overlap) across antibody regions (CDR/FR) in **(A)** paratopes and **(B)** epitopes. There is considerable overlap across paratope and epitope regions albeit to a slightly lesser extent in epitopes compared to paratopes. **(C)** Overlap of mouse and human paratope and epitope motifs. The cross-species relative overlap is higher for paratopes than for epitopes. **(D)** Cumulative motif distribution of paratopes and **(E)** epitopes ordered by frequency. Motif frequency denotes the number of times a motif has been detected within or across crystal structures. **(F)** Number of antigen classes, **(G)** V genes (separated by human and mouse), and **(H)** crystal structures paratope and **(I)** epitope motifs belong to. The top 20 most shared (across crystal structures) motifs are shown for subfigures E–I.

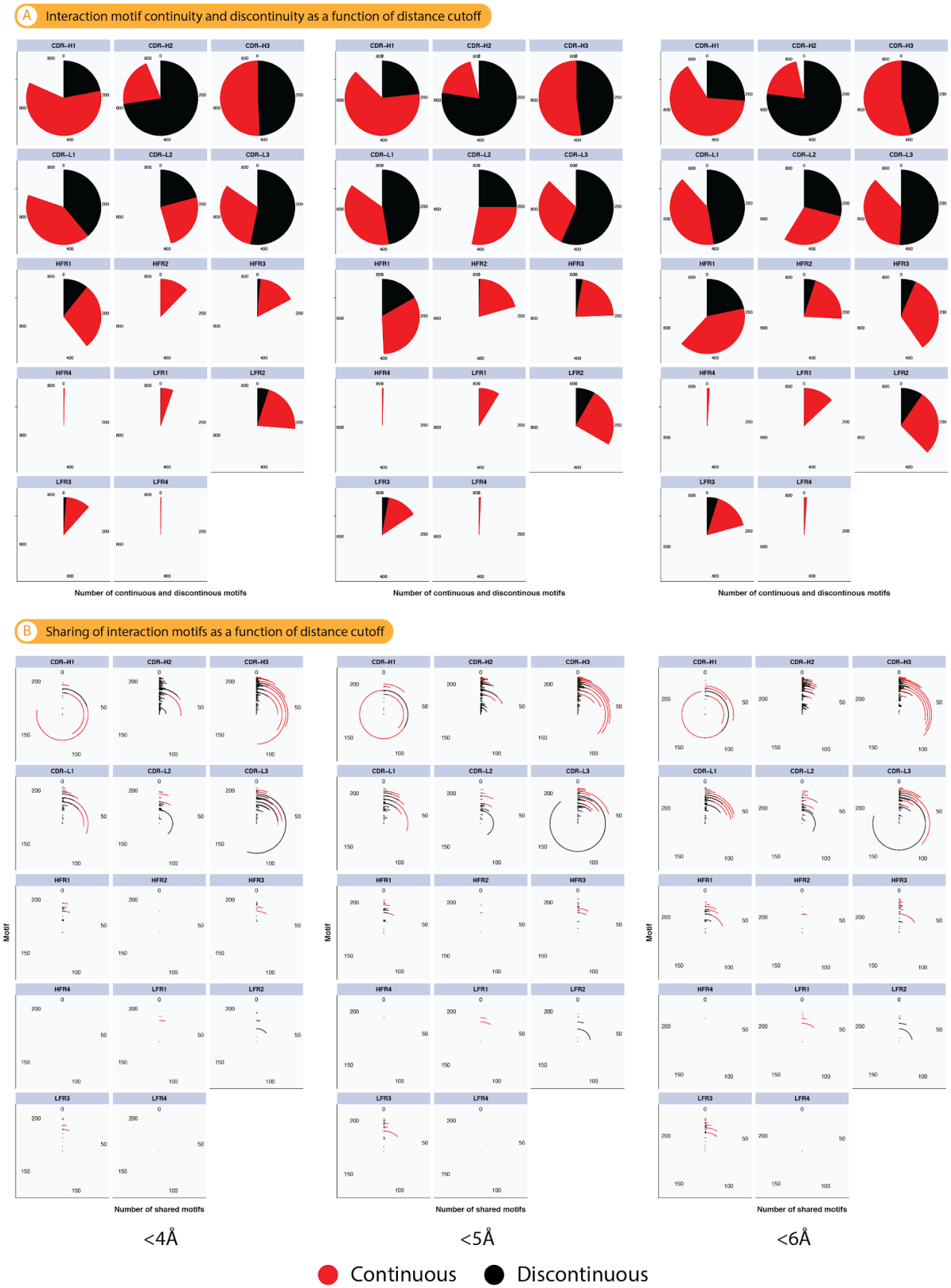


Supplemental Figure S5 | Interaction motifs of broadly neutralizing antibodies (bNAb) do not differ from those found in non-nbNAb antibodies. (A) All 38 motifs identified in the 24 HIV-bNAB crystal structures from the bnaber database (Eroshkin et al., 2014) (see Methods) are also found in non-HIV-bNAb crystal structures. Out of the 38 unique

interaction motifs identified, 8 are continuous and 30 are discontinuous. **(B)** Out of the 38 bNAb interaction motifs, 13 motifs were part of the 20 non-bNAb motifs that were shared most across cross-crystal structures (Suppl Fig. S4).

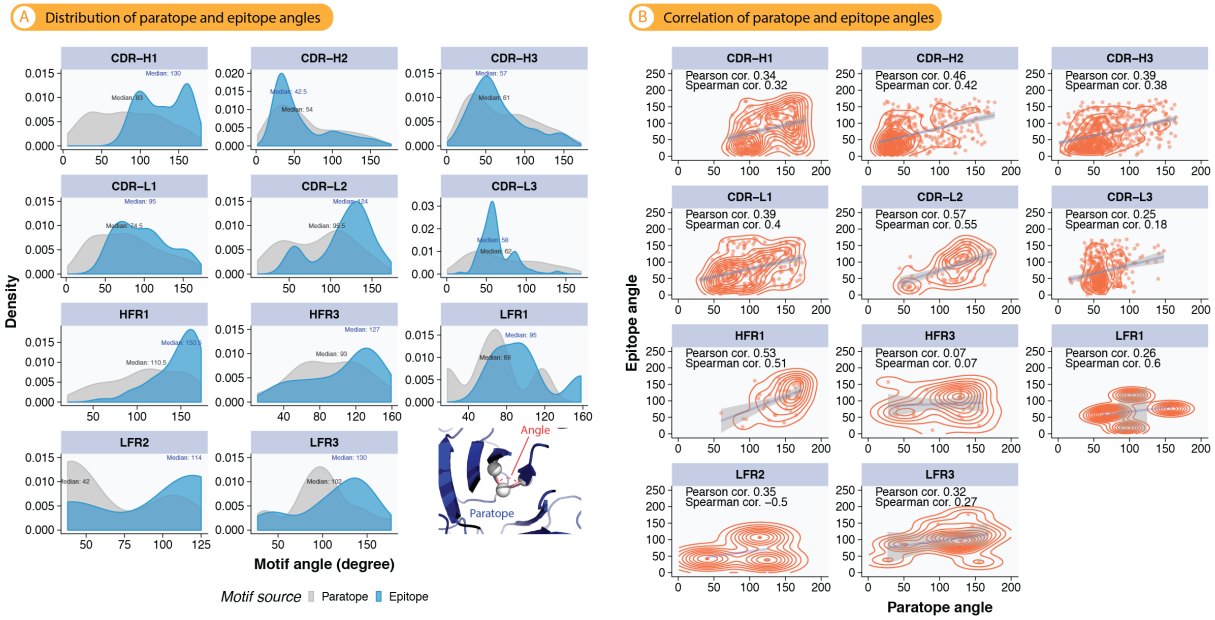


Supplemental Figure S6 | Comparison of interaction sequence motifs found in non-immune protein-protein interaction (PPI) with those found in antibody-antigen complexes. (A) Overlap of paratope, epitope and PPI interaction motifs. Whereas paratope motifs largely overlap with motifs found in PPI (70%), most epitope motifs do not overlap with that of PPI (11.4%). (**B**) Distribution of PPI interaction motif length. **(C)** Distribution of maximum gap lengths of PPI interaction motifs. **(D)** Amino acid distribution of interacting residues in PPI. **(E)** Pearson (lower triangular matrix) and Spearman (upper triangular matrix) correlation of the amino acid usage in paratope (heavy, light chain), epitope (heavy, light chain) and protein-protein interaction (intradomain and interdomain). **(F)** Inter- and **(G)** intradomain motif (cumulative) frequency distribution of the top 200 inter- and intradomain PPI motifs. Motif frequency denotes the number of times a motif has been detected within (intra) or across (inter) crystal structures (8'311 and 23'403, respectively).

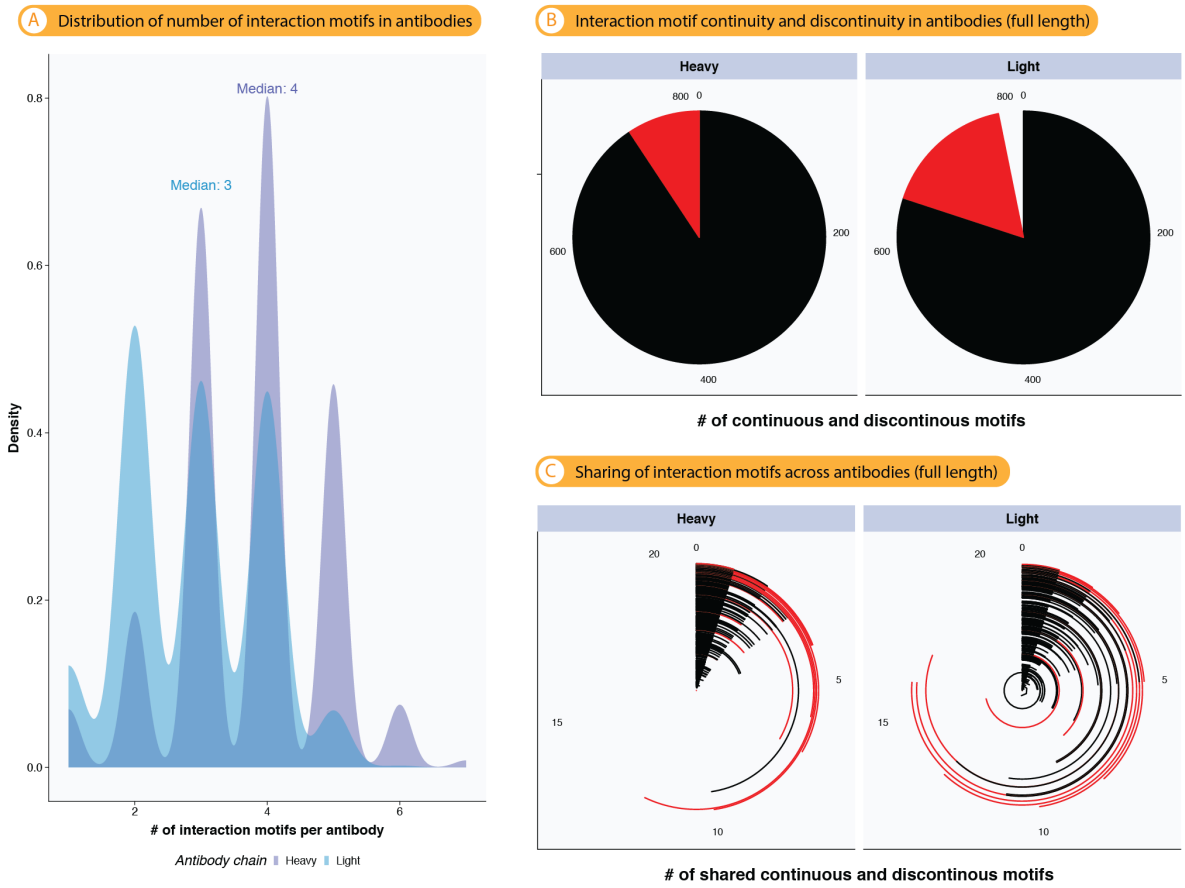


Supplemental Figure S7| The extent of motif continuity and sharing across crystal structures is robust to the distance cutoff chosen. (A) Relative motif continuity and discontinuity are shown as a function of the distance cutoff.

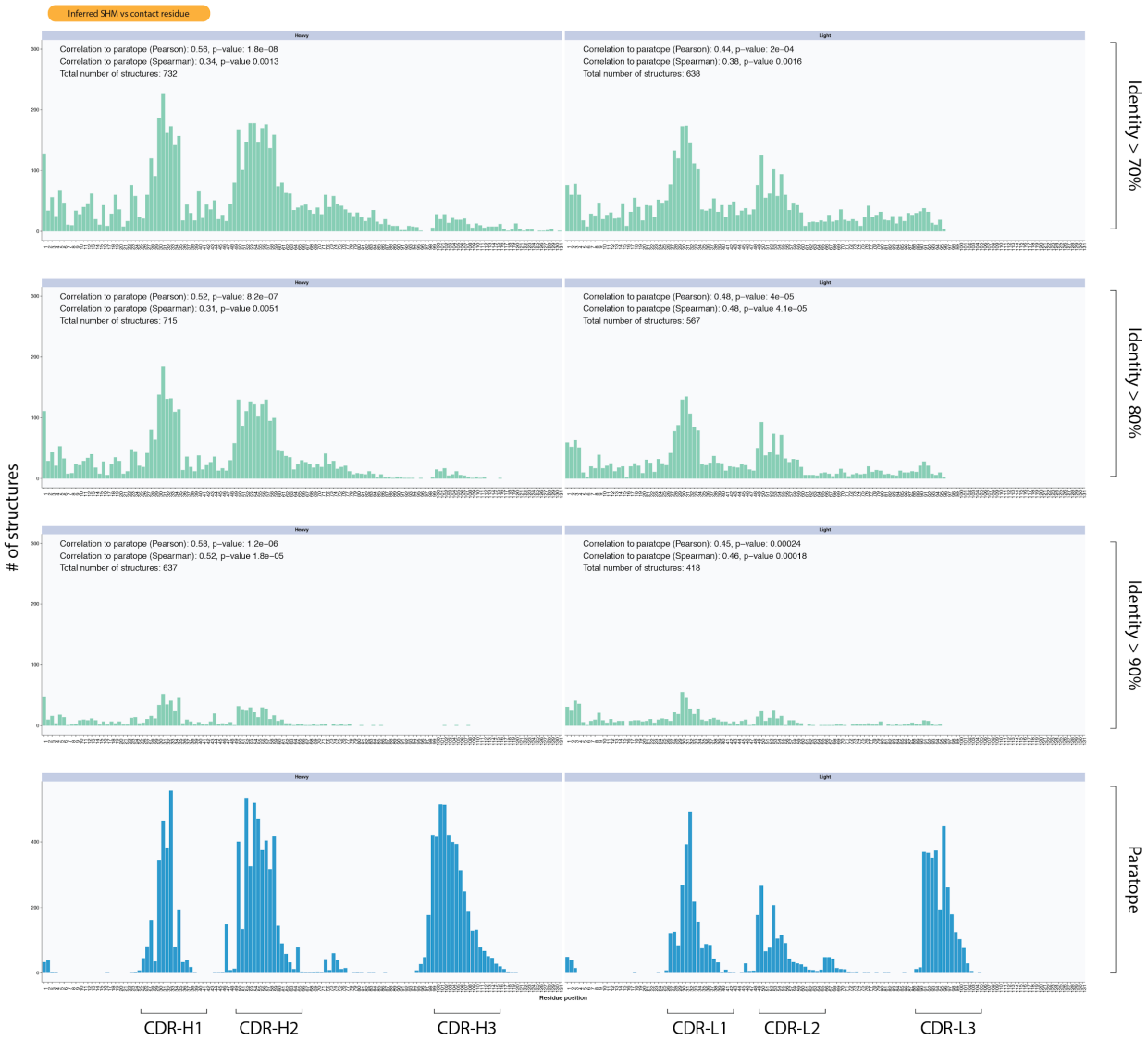
Ratios of continuous to discontinuous motifs are robust across distance cutoffs. **(B)** The extent of motif sharing across crystal structures as a function of the distance cutoff is shown. The sharing profiles are robust across distance cutoffs.



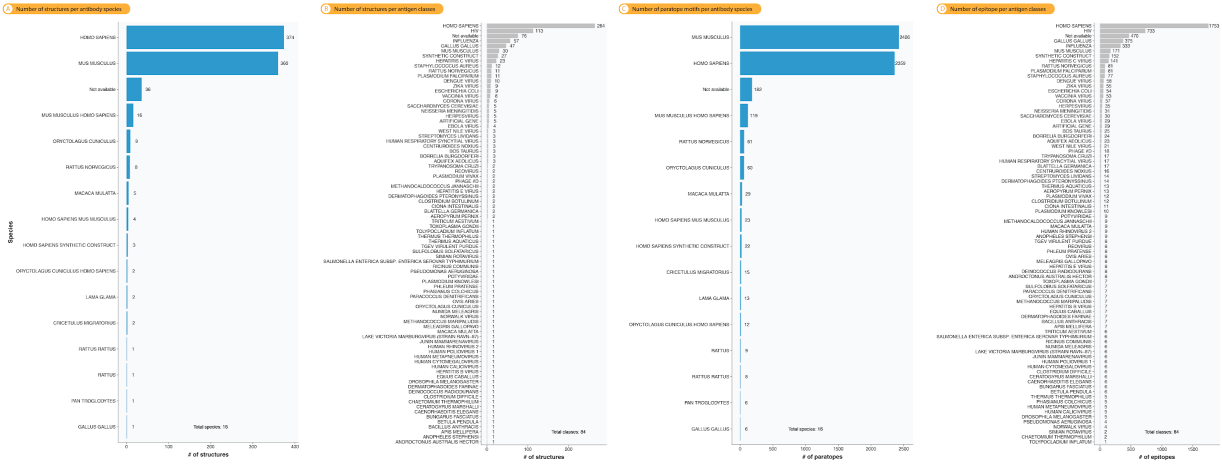
Supplemental Figure S8 | Angles of structural interaction motifs. (A) Motif angles in paratopes and epitopes (see Methods). For CDR-H2$_{VH,VL}$ and CDR-H3$_{VH,VL}$,the median angles in paratopes and epitopes are smaller than CDR-H1$_{VH,VL}$ indicating that the former are looped and the latter are extended. **(B)** Angles between paratope and epitopes correlate moderately in the majority of the regions examined.



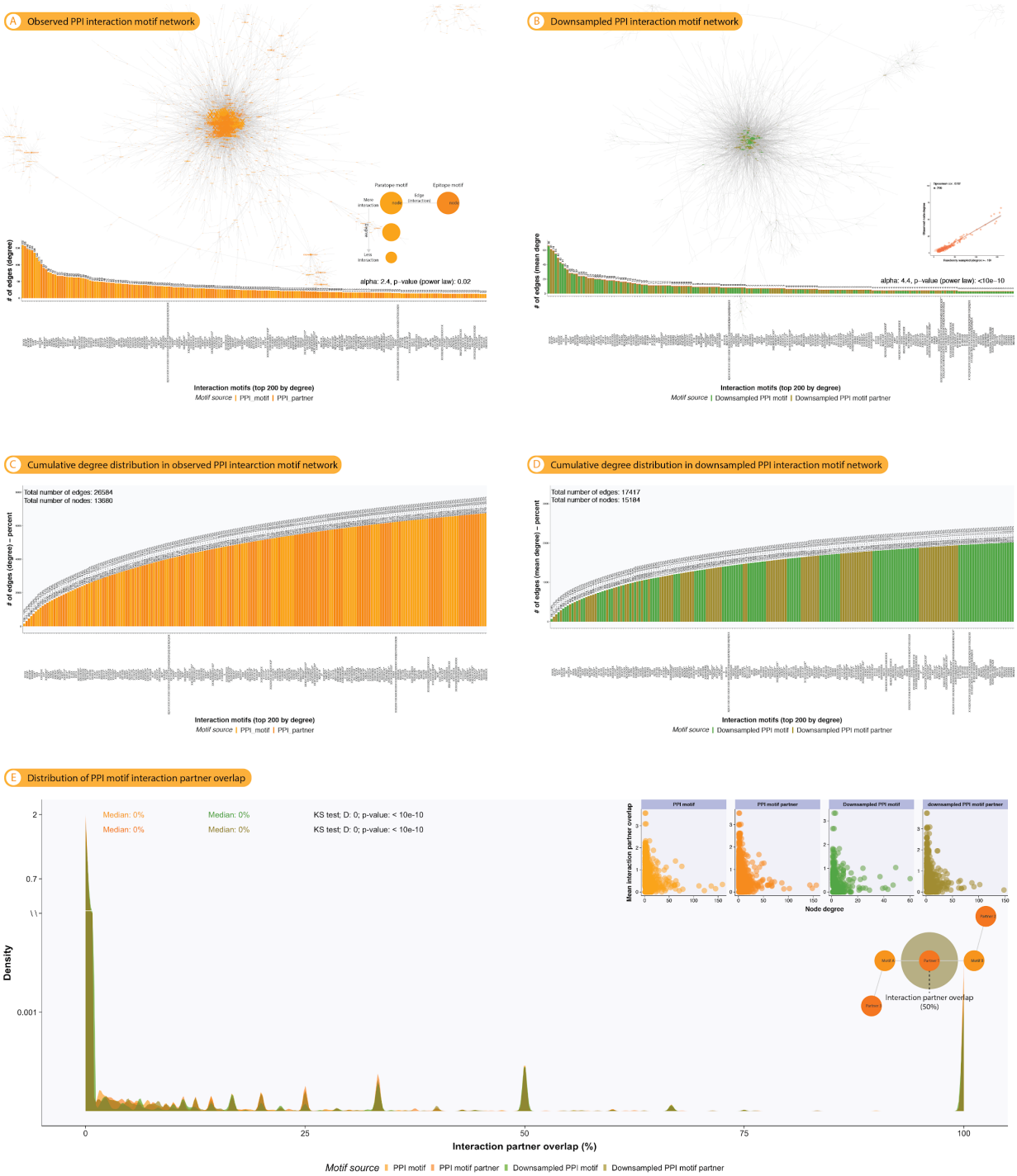
Supplemental Figure S9 | Analysis of interaction motifs on full length antibodies. (A) Number of paratope interaction motifs found in antibody heavy (median: 4) and light chains (median: 3). **(B)** Extent of ensembles of discontinuous (at least one discontinuous interaction motif per antibody heavy or light chain) or continuous (all motifs per antibody heavy or light chain are continuous) interaction motifs. **(C)** Sharing of discontinuous and continuous ensembles of interaction motifs across crystal structures (heavy-chain maximum: 12 crystal structures, light-chain maximum: 17 crystal structures).



Supplemental Figure S10 | Moderate concordance between somatic hypermutation sites and contact residues. Somatic hypermutation was inferred for all human and mouse antibodies (total: 732, Suppl Fig. S11, see Methods). Green graphs: for each position in the heavy and light chain, the number of structures mutated in that position was quantified. Positional mutability is shown as a function of identity to germline (>70%, >80%, >90%) where identity is calculated as: (length of the sequence – the total number of SHMs) / the length of sequence**.** The higher the identity, the more trustworthy the SHM count is. Blue graphs: number of structures with antigen-interacting residues in the given residue position. In each green-colored SHM graph, the correlation with the number of paratope contact sites per position is shown. Spearman (Pearson) correlation ranged between 0.31 (0.52) and 0.52 (0.58) for V_H and 0.38 (0.44) – 0.48 (0.48) for V_L, indicating moderate concordance between mutated and antigen-interacting residues.

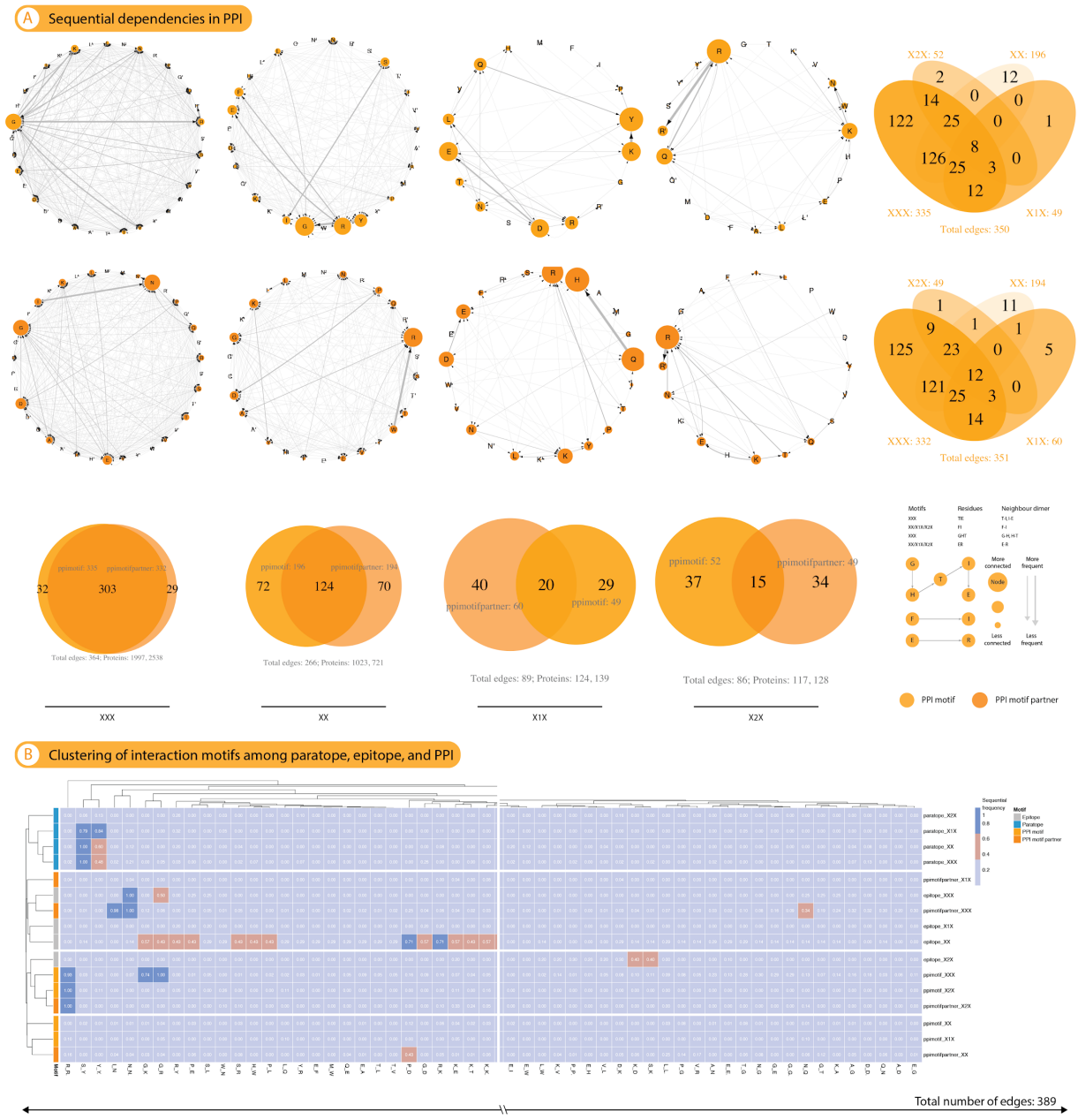


Supplemental Figure S11 | Crystal structure statistics: distribution of crystal structures across species and antigen classes. (A,C) Distribution of crystal structures (antibodies) and paratopes across species. The classes "human" and "mouse" make up together ≈90% of crystal structures here studied. **(B,D)** Distribution of crystal structures (antigens) and epitopes across antigen classes/species.



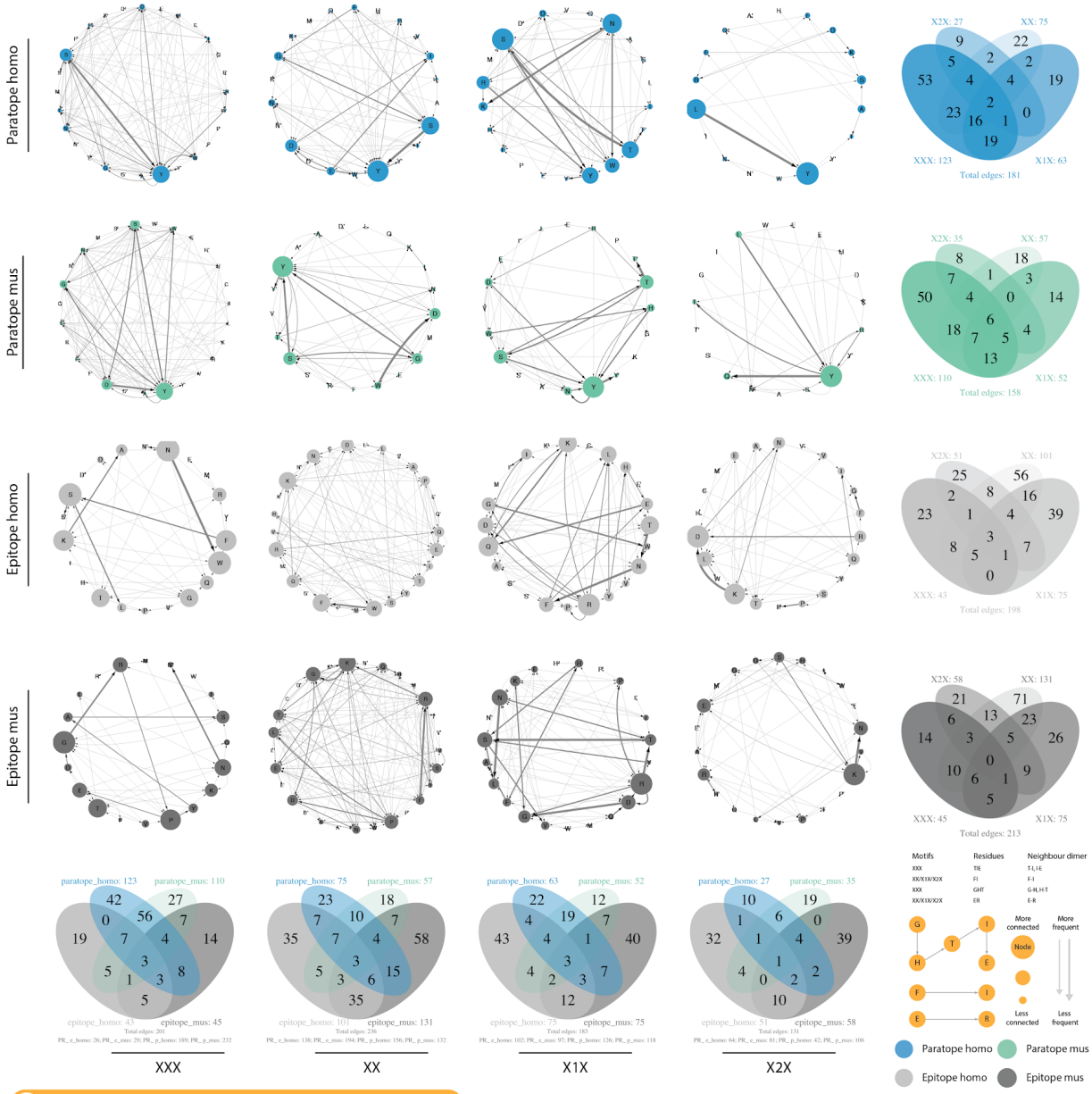
Supplemental Figure S12 | Protein-protein reactivity networks show exponential network architecture (no power law distribution as observed in antibody-antigen reactivity networks [Fig. 5A], but also no Poisson-distributed networks [Fig. 5B]). The figure replicates (with a few modifications explained below) Fig. 5. Only the largest connected portion of the network was visualized, for full network see Suppl Fig. 16. **(A)** A bipartite network capturing protein-protein binding from interdomain protein-protein protein interaction (Suppl Fig. S6) was constructed. The PPI interdomain reactivity network was constructed by connecting each protein motif to its corresponding partner protein motif (number of edges: 26,584; number of nodes: 13,680). We here used exclusively interdomain PPI (as opposed to intradomain PPI) as interdomain PPI is most similar to antibody-antigen interaction (Suppl Fig. S6E). Network vertices were scaled by their number of connections (edges). **(B)** To more directly compare PPI reactivity network architecture to antibody-antigen

interaction (Fig. 5A), we randomly sampled 100 times 5000 edges (similar to the number of edges found in the paratope-epitope reactivity network, Fig. 5A) from the reactivity network in (A). Inlet: The Spearman correlation of node degree correlation of observed and randomly sampled networks is shown. For both networks, the respective node degree distribution is shown (for B, the standard error of the mean is also shown). A node's degree is the count of its connections incoming from other nodes. **(C, D)** Cumulative degree distributions of networks (A) and (B). **(E)** Distribution of interaction partner overlap for networks (A,B). Briefly, for example for all protein binding partners in (A), the pairwise overlap of PPI motifs was calculated. The statistical significance of the difference between overlap distributions from (A) and (B) was computed using the Kolmogorov-Smirnov (KS) test. Inlet: the correlation of node degree and interaction partner overlap was determined.

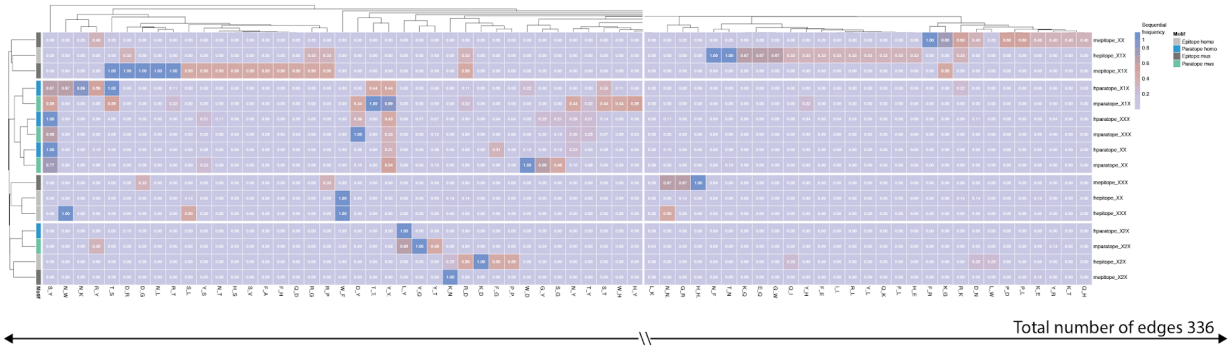


Supplemental Figure S13 | Sequential dependencies in interaction motifs of protein-protein interaction. (A) For each of the four most highly shared (across crystal antibody-antigen structures) motifs (Suppl Fig. 4), the sequential dependency signature of PPI motifs was determined. Briefly, for the ensemble of paratope/epitope sequences mapping to a given interaction motif, the 2-mer decomposition of each paratope/epitope sequence was determined by a sliding window. For each motif, these sequential dependencies were visualized as directed networks where the 20 amino acids are the nodes and the edges are the frequency with which a 2-mer occurred (sequential dependency). In addition, Venn diagrams show the overlap of sequential dependencies (two-mers, network edges) between motifs (vertically arranged Venn diagrams) or binding partners (horizontal Venn diagrams). **(B)** Hierarchical clustering of sequential dependencies (network edges shown in (A))**.**

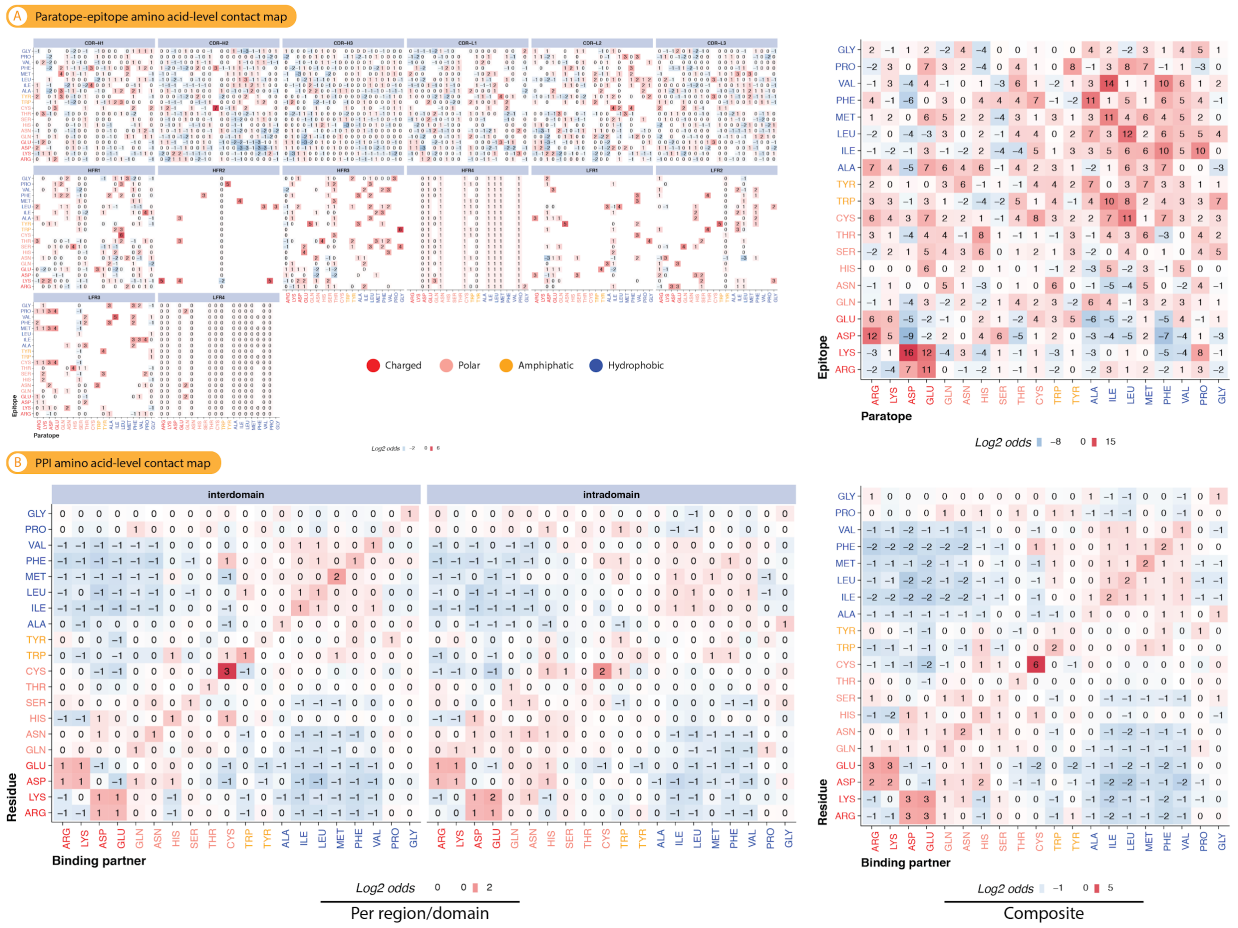
A Sequential dependencies of interaction motifs in human and mouse



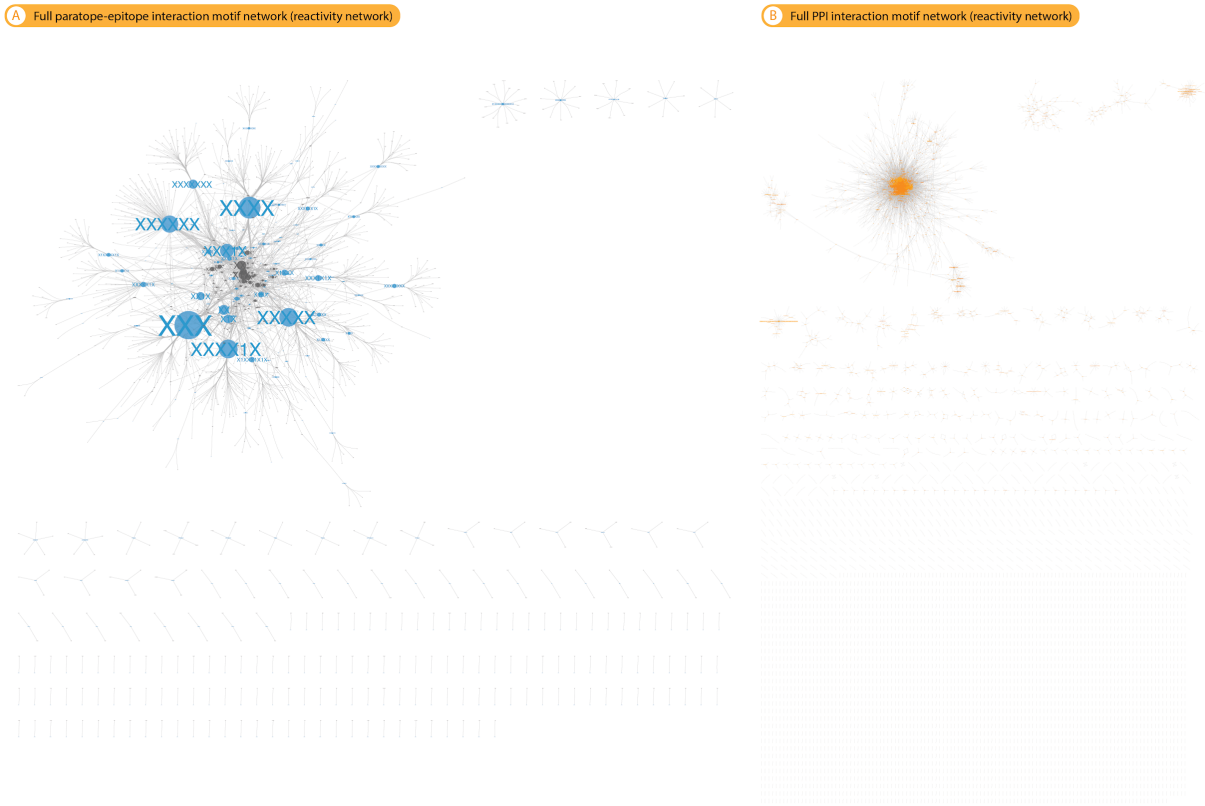
B Clustering of interaction motifs in human and mouse



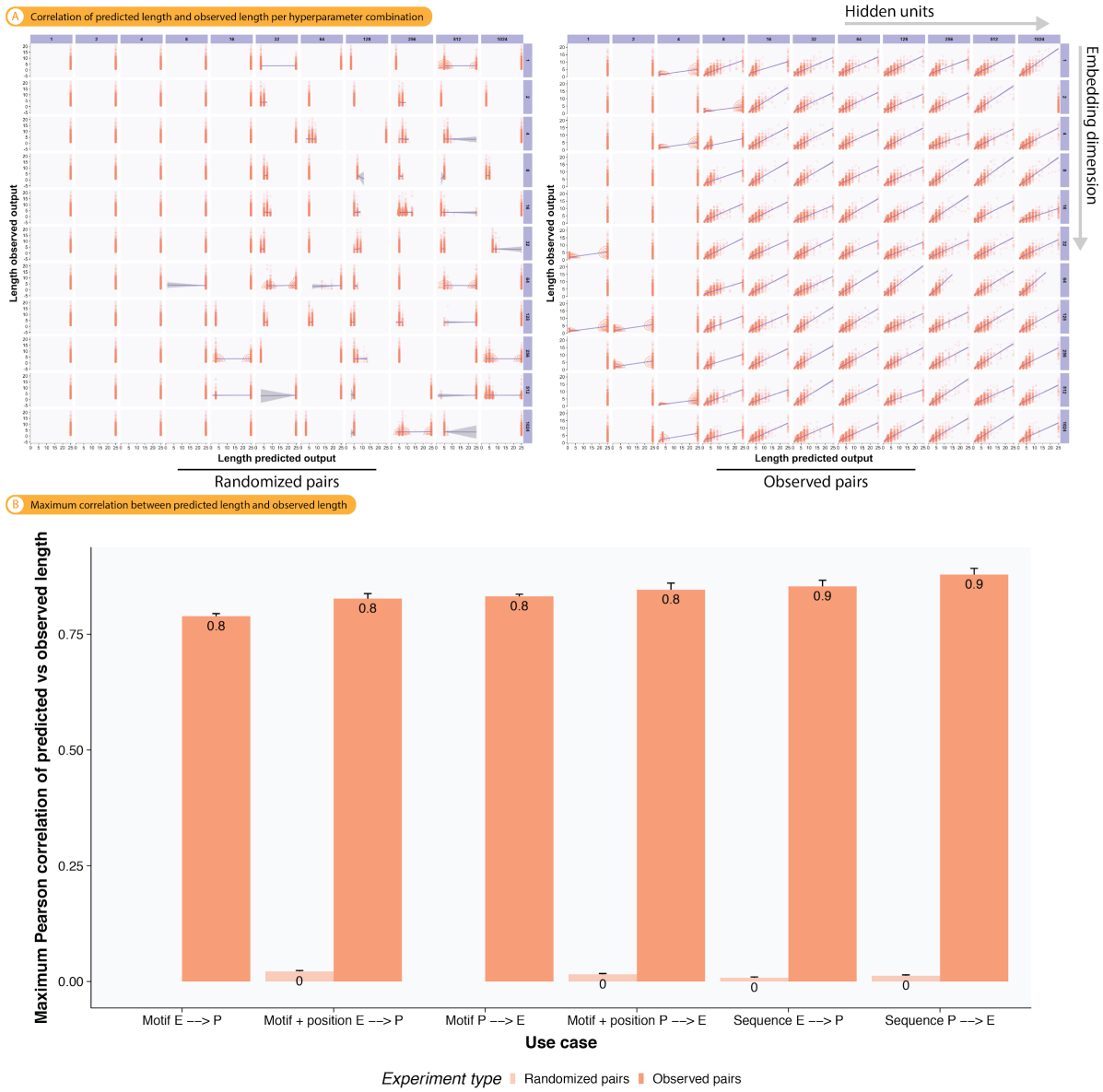
Supplemental Figure S14 | Sequential dependencies of interaction motifs in humans and mice are similar. (A) Sequential dependencies of mouse and human paratope and epitope interaction motifs were compared for the 4 most highly shared interaction motifs. Briefly, for the ensemble of paratope/epitope sequences mapping to a given interaction motif, the 2-mer decomposition of each paratope/epitope sequence was determined by sliding window. For each motif, these sequential dependencies were visualized as directed networks where the 20 amino acids are the nodes and the edges are the frequency with which a 2-mer occurred (sequential dependency). Overlap of two-mer combinations is shown by Venn diagram. **(B)** Hierarchical clustering of mouse and human paratope and epitope sequential dependencies (network edges shown in (A)).



Supplemental Figure S15 | Paratope-epitope amino acid contact maps. (A) The y-axis denotes epitope residues and the x-axis denotes the paratope residues. Left: Each panel represents one CDR/FR. Right: Composite map for V_H and V_L combined. The cells are color-coded according to the preference of the amino acid pairs to interact. A red square indicates that the interaction is favored, and a blue square indicates that it occurs less frequently than expected at random (log2 odds ratio, see Methods). Amino acids are color-coded by property: charged, polar, aromatic, amphiphatic, hydrophobic. **(B)** Identical to (A) except for non-immune protein-protein interaction (PPI). While there exist clear trends for PPI, namely off-diagonal amino acid interaction being disfavored and on-diagonal amino acid interaction being favored, paratope-epitope amino acid contact preferences (CDR/FR) cover a broader range in the interaction space as few amino acid interactions seem to be strictly disfavored (composite).



Supplemental Figure S16 | Full reactivity networks of paratope-epitope and protein-protein interaction are shown. (A) Paratope-epitope reactivity network. Please refer to Fig. 5A for a methods description. **(B)** Protein-protein reactivity network (interdomain only). Please refer to Suppl Fig. S12A for a methods description.



Supplemental Figure S17 | The lengths of interaction motif and sequence are predictable. (A) Example plot for the correlation between the predicted and observed length from one of the ten replicates in use case Sequence P→E as a function of embedding dimension and hidden units. Increasingly positive Pearson correlation values were recorded as the models approached optimal parameter pairs. **(B)**. Maximum predicted-observed length correlation. For the 10 replicates in each use case, the maximum correlation was obtained (the cell in (A) with the highest Pearson correlation) resulting in a total of 10 maximum correlation values. The mean of the maximum correlation values from each use case was computed and visualized as bar chart. The standard error of the mean of maximum correlation ($SD/\sqrt{n}$) was visualized as error bars. Use cases cover the bidirectional prediction tasks (paratope to epitope as well as epitope to paratope) of motif to motif, motif with position to motif with position, and finally amino acid sequence to amino acid sequence (see Table 1). Baseline prediction accuracies (control) were calculated based on label-shuffled data where antibody-antigen-binding partners were randomly shuffled (randomized pairs).

Supplemental Table S1. Antibody sequence numbering according to the Martin (Abhinandan and Martin, 2008) numbering scheme.

Segment	Residue number	Insertion
LFR1	1–23	10a
CDR-L1	24–34	30 (a–f)
LFR2	35–49	40a
CDR-L2	50–56	52 (a–e)
LFR3	57–88	68 (a–h)
CDR-L3	89–97	95 (a–f)
LFR4	98–110	-
HFR1	1–30	8 (a–d)
CDR-H1	31–35	31 (a–b)
HFR2	36–49	42
CDR-H2	50–65	52 (a–c)
HFR3	66–94	72 (a–c)
CDR-H3	95–102	100 (a–k)
HFR4	103–113	-

Supplemental Table S2. Examples of antibody chains numbered according to the Martin (Abhinandan and Martin, 2008) numbering scheme.

2EH8_1_H	2EH8_1_L	2LTQ_1_H	2LTQ_1_L	3JBA_1_H	3JBA_1_L
Q-GLN-1	D-ASP-1	E-GLU-1	D-ASP-1	S-SER-7	D-ASP-1
V-VAL-2	I-ILE-2	V-VAL-2	I-ILE-2	G-GLY-8	I-ILE-2
Q-GLN-3	Q-GLN-3	Q-GLN-3	V-VAL-3	G-GLY-9	V-VAL-3
L-LEU-4	M-MET-4	L-LEU-4	M-MET-4	G-GLY-10	M-MET-4
V-VAL-5	T-THR-5	V-VAL-5	S-SER-5	L-LEU-11	S-SER-5
Q-GLN-6	Q-GLN-6	E-GLU-6	Q-GLN-6	V-VAL-12	Q-GLN-6
S-SER-7	T-THR-7	S-SER-7	S-SER-7	K-LYS-13	S-SER-7
G-GLY-8	P-PRO-8	G-GLY-8	P-PRO-8	P-PRO-14	P-PRO-8
A-ALA-9	L-LEU-9	G-GLY-9	S-SER-9	G-GLY-15	S-SER-9
E-GLU-10	S-SER-10	G-GLY-10	S-SER-10	G-GLY-16	S-SER-10
V-VAL-11	L-LEU-11	L-LEU-11	L-LEU-11	S-SER-17	L-LEU-11
V-VAL-12	S-SER-12	V-VAL-12	A-ALA-12	L-LEU-18	A-ALA-12
K-LYS-13	V-VAL-13	K-LYS-13	V-VAL-13	K-LYS-19	V-VAL-13
P-PRO-14	T-THR-14	P-PRO-14	S-SER-14	L-LEU-20	S-SER-14
G-GLY-15	P-PRO-15	G-GLY-15	A-ALA-15	S-SER-21	V-VAL-15
A-ALA-16	G-GLY-16	G-GLY-16	G-GLY-16	C-CYS-22	G-GLY-16
S-SER-17	Q-GLN-17	S-SER-17	E-GLU-17	E-GLU-23	E-GLU-17
V-VAL-18	P-PRO-18	L-LEU-18	K-LYS-18	A-ALA-24	K-LYS-18
K-LYS-19	A-ALA-19	K-LYS-19	V-VAL-19	S-SER-25	V-VAL-19
V-VAL-20	S-SER-20	L-LEU-20	T-THR-20	G-GLY-26	T-THR-20
S-SER-21	I-ILE-21	S-SER-21	M-MET-21	F-PHE-27	M-MET-21
C-CYS-22	S-SER-22	C-CYS-22	S-SER-22	T-THR-28	S-SER-22
K-LYS-23	C-CYS-23	A-ALA-23	C-CYS-23	F-PHE-29	C-CYS-23
A-ALA-24	K-LYS-24	A-ALA-24	K-LYS-24	S-SER-30	K-LYS-24
S-SER-25	S-SER-25	S-SER-25	S-SER-25	S-SER-31	S-SER-25
G-GLY-26	S-SER-26	G-GLY-26	S-SER-26	Y-TYR-32	S-SER-26
Y-TYR-27	Q-GLN-27	F-PHE-27	Q-GLN-27	A-ALA-33	Q-GLN-27
A-ALA-28	S-SER-28	A-ALA-28	S-SER-28	M-MET-34	S-SER-28
F-PHE-29	L-LEU-29	F-PHE-29	L-LEU-29	S-SER-35	L-LEU-29
S-SER-30	L-LEU-30	S-SER-30	L-LEU-30	W-TRP-36	L-LEU-30
S-SER-31	Y-TYR-30	S-SER-31	N-ASN-30	V-VAL-37	Y-TYR-30
S-SER-32	S-SER-30	Y-TYR-32	S-SER-30	R-ARG-38	S-SER-30
W-TRP-33	N-ASN-30	D-ASP-33	R-ARG-30	Q-GLN-39	N-ASN-30
M-MET-34	G-GLY-30	M-MET-34	T-THR-30	T-THR-40	T-THR-30
N-ASN-35	K-LYS-30	S-SER-35	R-ARG-30	P-PRO-41	Q-GLN-30
W-TRP-36	T-THR-31	W-TRP-36	K-LYS-30	E-GLU-42	K-LYS-30
V-VAL-37	Y-TYR-32	V-VAL-37	N-ASN-31	K-LYS-43	N-ASN-31
R-ARG-38	L-LEU-33	R-ARG-38	Y-TYR-32	R-ARG-44	Y-TYR-32
Q-GLN-39	N-ASN-34	Q-GLN-39	L-LEU-33	L-LEU-45	L-LEU-33
A-ALA-40	W-TRP-35	T-THR-40	A-ALA-34	E-GLU-46	A-ALA-34
P-PRO-41	L-LEU-36	P-PRO-41	W-TRP-35	W-TRP-47	W-TRP-35
G-GLY-42	L-LEU-37	E-GLU-42	Y-TYR-36	V-VAL-48	Y-TYR-36
Q-GLN-43	Q-GLN-38	K-LYS-43	Q-GLN-37	A-ALA-49	Q-GLN-37
G-GLY-44	K-LYS-39	R-ARG-44	Q-GLN-38	S-SER-50	Q-GLN-38
L-LEU-45	P-PRO-40	L-LEU-45	K-LYS-39	I-ILE-51	K-LYS-39
E-GLU-46	G-GLY-41	E-GLU-46	P-PRO-40	S-SER-52	P-PRO-40
W-TRP-47	Q-GLN-42	W-TRP-47	G-GLY-41	S-SER-53	G-GLY-41
I-ILE-48	S-SER-43	V-VAL-48	Q-GLN-42	G-GLY-54	Q-GLN-42
G-GLY-49	P-PRO-44	A-ALA-49	S-SER-43	G-GLY-55	S-SER-43
R-ARG-50	K-LYS-45	Y-TYR-50	P-PRO-44	N-ASN-56	P-PRO-44
I-ILE-51	R-ARG-46	I-ILE-51	K-LYS-45	T-THR-57	K-LYS-45
Y-TYR-52	L-LEU-47	S-SER-52	L-LEU-46	H-HIS-58	L-LEU-46
P-PRO-52	I-ILE-48	S-SER-52	L-LEU-47	Y-TYR-59	L-LEU-47
G-GLY-53	Y-TYR-49	G-GLY-53	I-ILE-48	P-PRO-60	I-ILE-48
D-ASP-54	L-LEU-50	G-GLY-54	Y-TYR-49	D-ASP-61	Y-TYR-49

G-GLY-55	V-VAL-51	G-GLY-55	W-TRP-50	S-SER-62	W-TRP-50
D-ASP-56	S-SER-52	S-SER-56	A-ALA-51	V-VAL-63	A-ALA-51
T-THR-57	K-LYS-53	T-THR-57	S-SER-52	K-LYS-64	S-SER-52
N-ASN-58	L-LEU-54	Y-TYR-58	T-THR-53	G-GLY-65	T-THR-53
Y-TYR-59	D-ASP-55	Y-TYR-59	R-ARG-54	R-ARG-66	R-ARG-54
A-ALA-60	S-SER-56	P-PRO-60	E-GLU-55	F-PHE-67	E-GLU-55
Q-GLN-61	G-GLY-57	D-ASP-61	S-SER-56	T-THR-68	S-SER-56
K-LYS-62	V-VAL-58	T-THR-62	G-GLY-57	I-ILE-69	G-GLY-57
F-PHE-63	P-PRO-59	V-VAL-63	V-VAL-58	S-SER-70	V-VAL-58
Q-GLN-64	D-ASP-60	K-LYS-64	P-PRO-59	R-ARG-71	P-PRO-59
G-GLY-65	R-ARG-61	G-GLY-65	D-ASP-60	D-ASP-72	D-ASP-60
K-LYS-66	F-PHE-62	R-ARG-66	R-ARG-61	N-ASN-72	R-ARG-61
A-ALA-67	S-SER-63	F-PHE-67	F-PHE-62	A-ALA-72	F-PHE-62
T-THR-68	G-GLY-64	T-THR-68	T-THR-63	R-ARG-72	T-THR-63
L-LEU-69	S-SER-65	I-ILE-69	G-GLY-64	N-ASN-73	G-GLY-64
T-THR-70	G-GLY-66	S-SER-70	S-SER-65	I-ILE-74	S-SER-65
A-ALA-71	S-SER-67	R-ARG-71	G-GLY-66	L-LEU-75	G-GLY-66
D-ASP-72	G-GLY-68	D-ASP-72	S-SER-67	Y-TYR-76	S-SER-67
K-LYS-72	T-THR-69	N-ASN-72	G-GLY-68	L-LEU-77	G-GLY-68
S-SER-72	D-ASP-70	A-ALA-72	T-THR-69	Q-GLN-78	T-THR-69
T-THR-72	F-PHE-71	K-LYS-72	D-ASP-70	M-MET-79	D-ASP-70
S-SER-73	T-THR-72	N-ASN-73	F-PHE-71	S-SER-80	F-PHE-71
T-THR-74	L-LEU-73	T-THR-74	T-THR-72	S-SER-81	T-THR-72
A-ALA-75	K-LYS-74	L-LEU-75	L-LEU-73	L-LEU-82	L-LEU-73
Y-TYR-76	I-ILE-75	Y-TYR-76	T-THR-74	R-ARG-83	T-THR-74
M-MET-77	S-SER-76	L-LEU-77	I-ILE-75	S-SER-84	I-ILE-75
E-GLU-78	R-ARG-77	Q-GLN-78	S-SER-76	E-GLU-85	S-SER-76
L-LEU-79	V-VAL-78	M-MET-79	S-SER-77	D-ASP-86	S-SER-77
S-SER-80	E-GLU-79	S-SER-80	V-VAL-78	T-THR-87	V-VAL-78
S-SER-81	A-ALA-80	S-SER-81	Q-GLN-79	A-ALA-88	K-LYS-79
L-LEU-82	E-GLU-81	L-LEU-82	A-ALA-80	M-MET-89	A-ALA-80
R-ARG-83	D-ASP-82	K-LYS-83	E-GLU-81	Y-TYR-90	E-GLU-81
S-SER-84	V-VAL-83	S-SER-84	D-ASP-82	Y-TYR-91	D-ASP-82
E-GLU-85	G-GLY-84	E-GLU-85	L-LEU-83	C-CYS-92	L-LEU-83
D-ASP-86	V-VAL-85	D-ASP-86	A-ALA-84	A-ALA-93	A-ALA-84
T-THR-87	Y-TYR-86	T-THR-87	V-VAL-85	R-ARG-94	V-VAL-85
A-ALA-88	Y-TYR-87	A-ALA-88	Y-TYR-86	G-GLY-95	Y-TYR-86
V-VAL-89	C-CYS-88	M-MET-89	Y-TYR-87	L-LEU-96	Y-TYR-87
Y-TYR-90	V-VAL-89	Y-TYR-90	C-CYS-88	Y-TYR-97	C-CYS-88
F-PHE-91	Q-GLN-90	Y-TYR-91	K-LYS-89	Y-TYR-98	Q-GLN-89
C-CYS-92	G-GLY-91	C-CYS-92	Q-GLN-90	G-GLY-99	Q-GLN-90
A-ALA-93	T-THR-92	A-ALA-93	S-SER-91	Y-TYR-100	Y-TYR-91
R-ARG-94	H-HIS-93	R-ARG-94	Y-TYR-92	D-ASP-100	Y-TYR-92
E-GLU-95	F-PHE-94	P-PRO-95	N-ASN-93	E-GLU-100	S-SER-93
Y-TYR-96	P-PRO-95	D-ASP-96	L-LEU-94	G-GLY-100	Y-TYR-94
D-ASP-97	Q-GLN-96	A-ALA-97	Y-TYR-96	S-SER-100	P-PRO-95
E-GLU-98	T-THR-97	M-MET-98	T-THR-97	D-ASP-100	L-LEU-96
A-ALA-101	F-PHE-98	D-ASP-101	F-PHE-98	F-PHE-100	T-THR-97
Y-TYR-102	G-GLY-99	Y-TYR-102	G-GLY-99	D-ASP-101	F-PHE-98
W-TRP-103	G-GLY-100	W-TRP-103	G-GLY-100	Y-TYR-102	G-GLY-99
G-GLY-104	G-GLY-101	G-GLY-104	G-GLY-101	W-TRP-103	A-ALA-100
Q-GLN-105	T-THR-102	Q-GLN-105	T-THR-102	G-GLY-104	G-GLY-101
G-GLY-106	K-LYS-103	G-GLY-106	K-LYS-103	Q-GLN-105	T-THR-102
T-THR-107	V-VAL-104	T-THR-107	L-LEU-104	G-GLY-106	K-LYS-103
L-LEU-108	E-GLU-105	S-SER-108	E-GLU-105	T-THR-107	L-LEU-104
V-VAL-109	I-ILE-106	V-VAL-109	I-ILE-106	T-THR-108	E-GLU-105
T-THR-110	K-LYS-107	T-THR-110	K-LYS-107	L-LEU-109	L-LEU-106
V-VAL-111	R-ARG-108	V-VAL-111	A-ALA-108	T-THR-110	K-LYS-107
S-SER-112	T-THR-109	S-SER-112	D-ASP-109	V-VAL-111	R-ARG-108
S-SER-113	V-VAL-110	S-SER-113	A-ALA-110	S-SER-112	A-ALA-109

Supplemental Table S3. Top 10 protein domains and typical size sourced from Pfam (El-Gebali et al., 2019) and 3did (Mosca et al., 2014).

Domain name	Pfam ID*	Typical number of amino acids**	Description
Proteasome	PF00227	190	Protein degradation
C1-set	PF07654	80	Ig-like constant
Hemagglutinin	PF00509	300	Glycoprotein (Influenza)
V-set	PF07686	100	Ig-like variable
WD40	PF00400	40	Structural motif
adh_short_c2	PF13561	220	Reductase
lectin_legB	PF00139	220	Lectin proteins
EF-hand_7	PF13378	220	Enolase C-terminal domain-like
Ferritin	PF00210	140	Ferritin-like domain
Catalase	PF00199	380	Catalyzes the decomposition of hydrogen peroxide

*Pfam ID is a unique identifier to access an entry in the database of protein families (Pfam)
**The typical number of amino acids was obtained by mapping the Pfam domain to structures in Protein Data Bank (Berman et al., 2000) and sequences in UniProt (Bateman et al., 2017).

Acknowledgements

Funding: support was provided by The Helmsley Charitable Trust (#2019PG-T1D011, to VG), UiO World Leading Research Community (to VG), UiO:LifeSciences Convergence Environment Immunolingo (to VG and GKS), EU Horizon 2020 iReceptorplus (#825821) (to VG) and Stiftelsen Kristian Gerhard Jebsen (K.G. Jebsen Coeliac Disease Research Centre) (to GKS).

References

Abhinandan, K.R., and Martin, A.C.R. (2008). Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. Mol. Immunol. *45*, 3832–3839.
Akbar, R. (2019). themeakbar.
Allcorn, L.C., and Martin, A.C.R. (2002). SACS--self-maintaining database of antibody crystal structure information. Bioinformatics *18*, 175–181.
Andersen, J.T., Pehrson, R., Tolmachev, V., Daba, M.B., Abrahmsén, L., and Ekblad, C. (2011).

Extending Half-life by Indirect Targeting of the Neonatal Fc Receptor (FcRn) Using a Minimal Albumin Binding Domain. Journal of Biological Chemistry *286*, 5234–5241.
Antunes, D.A., Abella, J.R., Devaurs, D., Rigo, M.M., and Kavraki, L.E. (2018). Structure-based methods for binding mode and binding affinity prediction for peptide-MHC complexes. Current Topics in Medicinal Chemistry *18*, 2239.
Arora, R., Kaplinsky, J., Li, A., and Arnaout, R. (2019). Repertoire-Based Diagnostics Using Statistical Biophysics. BioRxiv 519108.
Baran, D., Pszolla, M.G., Lapidoth, G.D., Norn, C., Dym, O., Unger, T., Albeck, S., Tyka, M.D., and Fleishman, S.J. (2017). Principles for computational design of binding antibodies. PNAS *114*, 10900–10905.
Barlow, D.J., Edwards, M.S., and Thornton, J.M. (1986). Continuous and discontinuous protein antigenic determinants. Nature *322*, 747–748.
Bateman, A., Martin, M.J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., et al. (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res *45*, D158–D169.
Benjamin, D.C., Berzofsky, J.A., East, I.J., Gurd, F.R.N., Hannum, C., Leach, S.J., Margoliash, E., Michael, J.G., Miller, A., Prager, E.M., et al. (1984). The Antigenic Structure of Proteins: A Reappraisal. Annual Review of Immunology *2*, 67–101.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.
Berzofsky, J.A. (1985). Intrinsic and extrinsic factors in protein antigenic structure. Science *229*, 932–940.
Bradley, P., and Thomas, P.G. (2019). Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. 24.
Briney, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 1.
Broido, A.D., and Clauset, A. (2019). Scale-free networks are rare. Nature Communications *10*, 1017.
Brown, A.J., Snapkov, I., Akbar, R., Pavlović, M., Miho, E., Sandve, G.K., and Greiff, V. (2019). Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. Mol. Syst. Des. Eng. *4*, 701–736.
Burkovitz, A., Leiderman, O., Sela-Culang, I., Byk, G., and Ofran, Y. (2013). Computational Identification of Antigen-Binding Antibody Fragments. The Journal of Immunology *190*, 2327–2334.
Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics *11*, 265–270.
Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics *43*, 783–791.
Chao, A., and Chiu, C.-H. (2016). Species Richness: Estimation and Comparison. In Wiley StatsRef: Statistics Reference Online, (American Cancer Society), pp. 1–26.
Chen, H. (2018). VennDiagram: Generate High-Resolution Venn and Euler Plots.
Chollet, F., and others (2015). Keras.
Chothia, C., and Lesk, A.M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. Journal of Molecular Biology *196*, 901–917.
Chuang, G.-Y., Zhou, J., Acharya, P., Rawi, R., Shen, C.-H., Sheng, Z., Zhang, B., Zhou, T., Bailer, R.T., Dandey, V.P., et al. (2019). Structural Survey of Broadly Neutralizing Antibodies Targeting the HIV-1 Env Trimer Delineates Epitope Categories and Characteristics of Recognition. Structure *27*, 196-206.e6.
Cinelli, M., Sun, Best, K., Heather, J.M., Reich-Zeliger, S., Shifrut, E., Friedman, N., Shawe-Taylor, J., and Chain, B. (2017). Feature selection using a one dimensional naïve Bayes' classifier increases the accuracy of support vector machine classification of CDR3 repertoires.

Bioinformatics *33*, 951–955.
Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009). Power-Law Distributions in Empirical Data. SIAM Review *51*, 661–703.
Collis, A.V.J., Brouwer, A.P., and Martin, A.C.R. (2003). Analysis of the Antigen Combining Site: Correlations Between Length and Sequence Composition of the Hypervariable Loops and the Nature of the Antigen. Journal of Molecular Biology *325*, 337–354.
Dalcin, L.D., Paz, R.R., Kler, P.A., and Cosimo, A. (2011). Parallel distributed computing using Python. Advances in Water Resources *34*, 1124–1139.
Dalkas, G.A., Teheux, F., Kwasigroch, J.M., and Rooman, M. (2014). Cation–π, amino–π, π–π, and H-bond interactions stabilize antigen–antibody interfaces. Proteins: Structure, Function, and Bioinformatics *82*, 1734–1746.
Dash, P., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature *547*, 89–93.
Deac, A., Veličković, P., and Sormanni, P. (2018). Attentive cross-modal paratope prediction. Journal of Computational Biology.
Dondelinger, M., Filée, P., Sauvage, E., Quinting, B., Muyldermans, S., Galleni, M., and Vandevenne, M.S. (2018). Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition. Front Immunol *9*.
Ecker, D.M., Jones, S.D., and Levine, H.L. (2015). The therapeutic monoclonal antibody market. MAbs *7*, 9–14.
El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res *47*, D427–D432.
Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C.G., Mora, T., and Walczak, A.M. (2015). Inferring processes underlying B-cell repertoire diversity. Phil. Trans. R. Soc. B *370*, 20140243.
Elhanati, Y., Sethna, Z., Callan, C.G., Mora, T., and Walczak, A.M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunological Reviews *284*, 167–179.
EL-Manzalawy, Y., Dobbs, D., and Honavar, V.G. (2017). In Silico Prediction of Linear B-Cell Epitopes on Proteins. In Prediction of Protein Secondary Structure, Y. Zhou, A. Kloczkowski, E. Faraggi, and Y. Yang, eds. (New York, NY: Springer New York), pp. 255–264.
Emerson, R.O., DeWitt, W.S., Vignali, M., Gravley, J., Hu, J.K., Osborne, E.J., Desmarais, C., Klinger, M., Carlson, C.S., Hansen, J.A., et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat Genet *49*, 659–665.
Eroshkin, A.M., LeBlanc, A., Weekes, D., Post, K., Li, Z., Rajput, A., Butera, S.T., Burton, D.R., and Godzik, A. (2014). bNAber: database of broadly neutralizing HIV antibodies. Nucleic Acids Research *42*, D1133–D1139.
Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.-C., and Deane, C.M. (2016). Progress and challenges in predicting protein interfaces. Brief Bioinform *17*, 117–131.
Ferdous, S., and Martin, A.C.R. (2018). AbDb: antibody structure database—a database of PDB-derived antibody structures. Database (Oxford) *2018*.
Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Bronstein, M.M., and Correia, B.E. (2019). Deciphering interaction fingerprints from protein molecular surfaces. BioRxiv 606202.
Gillespie, C.S. (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. Journal of Statistical Software *64*, 1–16.
Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. Nature *547*, 94–98.
Gowthaman, R., and Pierce, B.G. (2018). TCRmodel: high resolution modeling of T cell

receptors from sequence. Nucleic Acids Research.
Greenbaum, J.A., Andersen, P.H., Blythe, M., Bui, H., Cachau, R.E., Crowe, J., Davies, M., Kolaskar, A.S., Lund, O., Morrison, S., et al. (2007). Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. Journal of Molecular Recognition *20*, 75–82.
Greiff, V., Redestig, H., Luck, J., Bruni, N., Valai, A., Hartmann, S., Rausch, S., Schuchhardt, J., and Or-Guil, M. (2012). A minimal model of peptide binding predicts ensemble properties of serum antibodies. BMC Genomics *13*, 79.
Greiff, V., Bhat, P., Cook, S.C., Menzel, U., Kang, W., and Reddy, S.T. (2015). A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Medicine *7*, 49.
Greiff, V., Menzel, U., Miho, E., Weber, C., Riedel, R., Cook, S., Valai, A., Lopes, T., Radbruch, A., Winkler, T.H., et al. (2017a). Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Reports *19*, 1467–1478.
Greiff, V., Weber, C.R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., and Reddy, S.T. (2017b). Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. The Journal of Immunology *199*, 2985–2997.
Hellman, L.M., Foley, K.C., Singh, N.K., Alonso, J.A., Riley, T.P., Devlin, J.R., Ayres, C.M., Keller, G.L.J., Zhang, Y., Vander Kooi, C.W., et al. (2019). Improving T Cell Receptor On-Target Specificity via Structure-Guided Design. Molecular Therapy *27*, 300–313.
Henry, K.A., and MacKenzie, C.R. (2018). Antigen recognition by single-domain antibodies: structural latitudes and constraints. MAbs *10*, 815–826.
Inbar, D., Hochman, J., and Givol, D. (1972). Localization of Antibody-Combining Sites within the Variable Portions of Heavy and Light Chains. Proc Natl Acad Sci U S A *69*, 2659–2662.
Jain, T., Sun, T., Durand, S., Hall, A., Houston, N.R., Nett, J.H., Sharkey, B., Bobrowicz, B., Caffry, I., Yu, Y., et al. (2017). Biophysical properties of the clinical-stage antibody landscape. PNAS *114*, 944–949.
Jespersen, M.C., Mahajan, S., Peters, B., Nielsen, M., and Marcatili, P. (2019). Antibody Specific B-Cell Epitope Predictions: Leveraging Information From Antibody-Antigen Protein Complexes. Front. Immunol. *10*.
Kabat, E.A., Wu, T.T., Foeller, C., Perry, H.M., and Gottesman, K.S. (1992). Sequences of Proteins of Immunological Interest (DIANE Publishing).
Kilambi, K.P., and Gray, J.J. (2017). Structure-based cross-docking analysis of antibody–antigen interactions. Sci Rep *7*, 1–15.
Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [Cs].
Kolde, R. (2019). pheatmap: Pretty Heatmaps.
Krawczyk, K., Baker, T., Shi, J., and Deane, C.M. (2013). Antibody i-Patch prediction of the antibody binding site improves rigid local antibody–antigen docking. Protein Eng Des Sel *26*, 621–629.
Kringelum, J.V., Lundegaard, C., Lund, O., and Nielsen, M. (2012). Reliable B Cell Epitope Predictions: Impacts of Method Development and Improved Benchmarking. PLoS Comput Biol *8*, e1002829.
Kringelum, J.V., Nielsen, M., Padkjær, S.B., and Lund, O. (2013). Structural analysis of B-cell epitopes in antibody:protein complexes. Molecular Immunology *53*, 24–34.
Kunik, V., and Ofran, Y. (2013). The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. Protein Engineering, Design and Selection.
Kunik, V., Peters, B., and Ofran, Y. (2012a). Structural Consensus among Antibodies Defines the Antigen Binding Site. PLoS Comput Biol *8*.

Kunik, V., Ashkenazi, S., and Ofran, Y. (2012b). Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. Nucleic Acids Res. *40*, W521-524.
Kupiec, J.-J. (2010). On the lack of specificity of proteins and its consequences for a theory of biological organization. Progress in Biophysics and Molecular Biology *102*, 45–52.
Kuroda, D., and Gray, J.J. (2016). Shape complementarity and hydrogen bond preferences in protein-protein interfaces: Implications for antibody modeling and protein-protein docking. Bioinformatics btw197.
Landsteiner, K. (1936). SEROLOGICAL REACTIONS. 189.
Landsteiner, K. (1947). The Specificity of Serological Reactions Revised Edition (Harvard University Press).
Lanzarotti, E., Marcatili, P., and Nielsen, M. (2018). Identification of the cognate peptide-MHC target of T cell receptors using molecular modeling and force field scoring. Mol Immunol *94*, 91–97.
Lawrence, M.C., and Colman, P.M. (1993). Shape Complementarity at Protein/Protein Interfaces. Journal of Molecular Biology *234*, 946–950.
Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S., and Dimitrov, J.D. (2019). Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. Molecular Immunology *112*, 338–346.
Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbié, V., and Chaume, D. (1999). IMGT, the international ImMunoGeneTics database. Nucl. Acids Res. *27*, 209–212.
Liberis, E., Veličković, P., Sormanni, P., Vendruscolo, M., and Liò, P. (2018). Parapred: antibody paratope prediction using convolutional and recurrent neural networks. Bioinformatics.
Liu, G., Zeng, H., Mueller, J., Carter, B., Wang, Z., Schilz, J., Horny, G., Birnbaum, M.E., Ewert, S., and Gifford, D.K. (2019). Antibody Complementarity Determining Region Design Using High-Capacity Machine Learning. BioRxiv 682880.
Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Noncovalent Bonds. Molecular Cell Biology. 4th Edition.
Luong, M.-T., Pham, H., and Manning, C.D. (2015). Effective Approaches to Attention-based Neural Machine Translation. ArXiv:1508.04025 [Cs].
MacCallum, R.M., Martin, A.C.R., and Thornton, J.M. (1996). Antibody-antigen Interactions: Contact Analysis and Binding Site Topography. Journal of Molecular Biology *262*, 732–745.
Marchalonis, J.J., Kaveri, S., Lacroix-Desmazes, S., and Kazatchkine, M.D. (2002). Natural recognition repertoire and the evolutionary emergence of the combinatorial immune system. The FASEB Journal *16*, 842–848.
Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.
Mason, D.M., Weber, C.R., Parola, C., Meng, S.M., Greiff, V., Kelton, W.J., and Reddy, S.T. (2018). High-throughput antibody engineering in mammalian cells by CRISPR/Cas9-mediated homology-directed mutagenesis. Nucleic Acids Res. *46*, 7436–7449.
Mason, D.M., Friedensohn, S., Weber, C.R., Jordi, C., Wagner, B., Meng, S., and Reddy, S.T. (2019). Deep learning enables therapeutic antibody optimization in mammalian cells. BioRxiv 617860.
Mian, I.S., Bradwell, A.R., and Olson, A.J. (1991). Structure, function and properties of antibody binding sites. Journal of Molecular Biology *217*, 133–151.
Miho, E., Yermanos, A., Weber, C.R., Berger, C.T., Reddy, S.T., and Greiff, V. (2018). Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. Frontiers in Immunology *9*.
Miho, E., Roškar, R., Greiff, V., and Reddy, S.T. (2019). Large-scale network analysis reveals

the sequence space architecture of antibody repertoires. Nature Communications *10*, 1321.
Mitchell, L.S., and Colwell, L.J. (2018). Analysis of nanobody paratopes reveals greater diversity than classical antibodies. Protein Eng Des Sel *31*, 267–275.
Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res *42*, D374–D379.
Nguyen, M.N., Pradhan, M.R., Verma, C., and Zhong, P. (2017). The interfacial character of antibody paratopes: analysis of antibody–antigen structures. Bioinformatics *33*, 2971–2976.
Nicholson, D.J. (2019). Is the cell really a machine? Journal of Theoretical Biology *477*, 108–126.
Nimrod, G., Fischman, S., Austin, M., Herman, A., Keyes, F., Leiderman, O., Hargreaves, D., Strajbl, M., Breed, J., Klompus, S., et al. (2018). Computational Design of Epitope-Specific Functional Antibodies. Cell Reports *25*, 2121-2131.e5.
Ofran, Y., Schlessinger, A., and Rost, B. (2008). Automated Identification of Complementarity Determining Regions (CDRs) Reveals Peculiar Characteristics of CDRs and B Cell Epitopes. J Immunol *181*, 6230–6235.
Ostmeyer, J., Christley, S., Toby, I.T., and Cowell, L.G. (2019). Biophysicochemical motifs in T cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocytes and adjacent healthy tissue. Cancer Res canres.2292.2018.
Padlan, E.A. (1977). Structural basis for the specificity of antibody–antigen reactions and structural mechanisms for the diversification of antigen-binding specificities. Quarterly Reviews of Biophysics *10*, 35–65.
Peng, H.-P., Lee, K.H., Jian, J.-W., and Yang, A.-S. (2014). Origins of specificity and affinity in antibody–protein interactions. Proc Natl Acad Sci U S A *111*, E2656–E2665.
Ponomarenko, J.V., and Bourne, P.E. (2007). Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Structural Biology *7*, 64.
R Core Team (2018). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
Raghunathan, G., Smart, J., Williams, J., and Almagro, J.C. (2012). Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. Journal of Molecular Recognition *25*, 103–113.
Ralph, D.K., and Iv, F.A.M. (2016). Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. PLOS Computational Biology *12*, e1004409.
Ramaraj, T., Angel, T., Dratz, E.A., Jesaitis, A.J., and Mumey, B. (2012). Antigen–antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics *1824*, 520–532.
Raybould, M.I.J., Wong, W.K., and Deane, C.M. (2019a). Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. Molecular Systems Design & Engineering 10.1039.C9ME00034H.
Raybould, M.I.J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A.P., Bujotzek, A., Shi, J., and Deane, C.M. (2019b). Five computational developability guidelines for therapeutic antibody profiling. PNAS 201810576.
Riley, T.P., and Baker, B.M. (2018). The intersection of affinity and specificity in the development and optimization of T cell receptor based therapeutics. Seminars in Cell & Developmental Biology *84*, 30–41.
Robert, P.A., Marschall, A.L., and Meyer-Hermann, M. (2018). Induction of broadly neutralizing antibodies in Germinal Centre simulations. Current Opinion in Biotechnology *51*, 137–145.
Rubinstein, N.D., Mayrose, I., Halperin, D., Yekutieli, D., Gershoni, J.M., and Pupko, T. (2008). Computational characterization of B-cell epitopes. Molecular Immunology *45*, 3477–3489.
Salamanca Viloria, J., Allega, M.F., Lambrughi, M., and Papaleo, E. (2017). An optimal distance cutoff for contact-based Protein Structure Networks using side-chain centers of mass. Sci Rep

7.
Schrödinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.
Sela-Culang, I., Kunik, V., and Ofran, Y. (2013). The structural basis of antibody-antigen recognition. Front. Immunol. *4*, 302.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res *13*, 2498–2504.
Sidhom, J.-W., Larman, H.B., Pardoll, D.M., and Baras, A.S. (2018). DeepTCR: a deep learning framework for revealing structural concepts within TCR Repertoire. BioRxiv 464107.
Sivalingam, G.N., and Shepherd, A.J. (2012). An analysis of B-cell epitope discontinuity. Molecular Immunology *51*, 304–309.
Sivasubramanian, A., Sircar, A., Chaudhury, S., and Gray, J.J. (2009). Toward high-resolution homology modeling of antibody Fv regions and application to antibody–antigen docking. Proteins: Structure, Function, and Bioinformatics *74*, 497–514.
Soto, C., Bombardi, R.G., Branchizio, A., Kose, N., Matta, P., Sevy, A.M., Sinkovits, R.S., Gilchuk, P., Finn, J.A., and Jr, J.E.C. High frequency of shared clonotypes in human B cell receptor repertoires. N A t U r e 18.
Stave, J.W., and Lindpaintner, K. (2013). Antibody and antigen contact residues define epitope and paratope size and structure. J. Immunol. *191*, 1428–1435.
Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., et al. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. Cell *122*, 957–968.
Stewart, J. (1992). Immunoglobulins did not arise in evolution to fight infection. Immunology Today *13*, 396–399.
Sun, Y., Best, K., Cinelli, M., Heather, J.M., Reich-Zeliger, S., Shifrut, E., Friedman, N., Shawe-Taylor, J., and Chain, B. (2017). Specificity, Privacy, and Degeneracy in the CD4 T Cell Receptor Repertoire Following Immunization. Front. Immunol. *8*.
Thomas, N., Best, K., Cinelli, M., Reich-Zeliger, S., Gal, H., Shifrut, E., Madi, A., Friedman, N., Shawe-Taylor, J., and Chain, B. (2014). Tracking global changes induced in the CD4 T cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. Bioinformatics.
Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575.
Turner, S.J., Doherty, P.C., McCluskey, J., and Rossjohn, J. (2006). Structural determinants of T-cell receptor bias in immunity. Nat Rev Immunol *6*, 883–894.
Van Regenmortel, M.H.V. (2014). Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. J. Mol. Recognit. *27*, 627–639.
Van Rossum, G., and Drake Jr, F.L. (1995). Python tutorial (Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands).
Wang, M., Zhu, D., Zhu, J., Nussinov, R., and Ma, B. (2018). Local and global anatomy of antibody-protein antigen recognition. Journal of Molecular Recognition *31*, e2693.
Weber, C.R., Akbar, R., Yermanos, A., Pavlović, M., Snapkov, I., Sandve, G.K., Reddy, S.T., and Greiff, V. (2019a). immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. BioRxiv 759795.
Weber, L.M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P.P., Boulesteix, A.-L., Saeys, Y., and Robinson, M.D. (2019b). Essential guidelines for computational method benchmarking. Genome Biology *20*, 125.
Weinstein, J.A., Jiang, N., White, R.A., Fisher, D.S., and Quake, S.R. (2009). High-Throughput Sequencing of the Zebrafish Antibody Repertoire. Science *324*, 807–810.
Weitzner, B.D., and Gray, J.J. (2016). Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint. J Immunol 1601137.
Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wu, T.T., and Kabat, E.A. (1970). An Analysis of the Sequences of the Variable Regions of Bence Jones Proteins and Myeloma Light Chains and Their Implications for Antibody Complementarity. Journal of Experimental Medicine *132*, 211–250.
Xu, J.L., and Davis, M.M. (2000). Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody Specificities. Immunity *13*, 37–45.
Yoo, A.B., Jette, M.A., and Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In Job Scheduling Strategies for Parallel Processing, D. Feitelson, L. Rudolph, and U. Schwiegelshohn, eds. (Springer Berlin Heidelberg), pp. 44–60.

1 Supplemental Text. Derivation of an analytical solution for the theoretical number of motifs within one CDR/FR region

1.1 Objectives and motif definition

Our aim is to determine the number of possible structural interaction motifs for any motif length. A given sequence motif is defined as follows:

- An amino acid is encoded as **X**.
- A gap is encoded as integer n where n quantifies the length of the gap.
- Each motif starts and ends with an amino acid **X**.
- There can be > 1 amino acids in sequential positions but not > 1 gaps.

Let us give two different definitions of motif length. By simply "motif length" we mean the number of **X**s in it plus the number of gaps, we note this lengths L. By "amino acid length", we mean the number of amino acids included in the sequence, i.e. the number of **X**s plus the sum of all gap lengths. Please, refer to the section 1.2 for a few examples.

As the interaction sequence cannot exceed the size of the CDR/FR it is located in, we need to add one more constraint:

- The amino acid length of the motif is not bigger then a predefined number.

Let us denote the number of unique motifs of lengths L and amino acid length A as $\mathrm{N}_{L,A}$ and the number of unique motifs of length L with amino acid length *not exceeding* A as $\bar{\mathrm{N}}_{L,A} = \sum_{A_1=L}^{A} \mathrm{N}_{L,A_1}$

1.2 Examples

To derive a formula for $\mathrm{N}_{L,A}$, we inspect a few examples first for intuition purposes.

$$
\begin{array}{lll}
L=1, A=1 & \rightarrow \mathbf{X} & \rightarrow \mathrm{N}_{1,1}=1 \\
L=2, A=2 & \rightarrow \mathbf{XX} & \rightarrow \mathrm{N}_{2,2}=1 \\
L=3, A=3 & \rightarrow \mathbf{XXX}, \mathbf{X1X} & \rightarrow \mathrm{N}_{3,3}=2 \\
L=3, A=n>3 & \rightarrow \mathbf{X}k\mathbf{X}, \ \ (k=n-2) & \rightarrow \mathrm{N}_{3,n}=1 \\
L=4, A=4 & \rightarrow \mathbf{XXXX}, \mathbf{X1XX}, \mathbf{XX1X} & \rightarrow \mathrm{N}_{4,4}=3 \\
L=4, A=n>4 & \rightarrow \mathbf{X}k\mathbf{XX}, \mathbf{XX}k\mathbf{X} \ \ (k=n-3) & \rightarrow \mathrm{N}_{4,n}=2 \\
L=5, A=5 & \rightarrow \mathbf{XXXXX}, \mathbf{X1XXX}, \mathbf{XX1XX}, \mathbf{XXX1X}, \mathbf{X1X1X} & \rightarrow \mathrm{N}_{5,5}=5 \\
L=5, A=n>5 & \rightarrow \mathbf{X}k\mathbf{XXX}, \mathbf{XX}k\mathbf{XX}, \mathbf{XXX}k\mathbf{X}, \mathbf{X}k_1\mathbf{X}k_2\mathbf{X} & \rightarrow \mathrm{N}_{5,n}=n-1
\end{array}
$$

In the last line, $k = n-4$ and $k_1 + k_2 = n-3$. Let us clarify this last line: there are only 3 motifs with a single gap, but if there are two gaps, their lengths can vary: $k_1 = 1, k_2 = n-4$; $k_1 = 2, k_2 = n-5$, ... , so that we have $n-4$ double-gapped motifs in total.

1.3 General formula

Now we can proceed to derive a general formula for $\mathrm{N}_{L,A}$. Let us note the number of **X**s in a motif as n_x and the number of gaps as n_g. We can count the motifs for fixed n_x and n_g and then we will just have to sum the results over all $n_x + n_g = L$. Thus, we have n_x **X**s and $n_x - 1$ slots for gaps – between any two neighbouring **X**s there can be a gap. First, we have to choose n_g slots: the number of ways to do this is

$$\binom{n_x - 1}{n_g}$$

Now we have n_g gaps of total amino acid length $A - n_x$, and we need to distribute the lengths between the gaps. In other words, we need to split the number $A - n_x$ into a sum of n_g nonzero terms. The number of ways to do this is the number of n_g-*compositions* of $A - n_x$, which equals

$$\binom{A - n_x - 1}{n_g - 1}$$

Now we can write down the formula for $\mathrm{N}_{L,A}$ as

$$\mathrm{N}_{L,A} = \sum_{n_g + n_x = L, n_g \leq n_x - 1} \binom{n_x - 1}{n_g} \binom{A - n_x - 1}{n_g - 1}$$

We did not take into account the all-**X** case, so for $A = L$ we should have

$$\mathrm{N}_{L,L} = 1 + \sum_{n_g + n_x = L, n_g \leq n_x - 1} \binom{n_x - 1}{n_g} \binom{A - n_x - 1}{n_g - 1}$$

Similarly, the formula for $\bar{\mathrm{N}}_{L,A}$ is

$$\bar{\mathrm{N}}_{L,A} = \sum_{A_1 = L}^{A} \mathrm{N}_{L,A_1} = 1 + \sum_{A_1 = L}^{A} \sum_{n_g + n_x = L, n_g \leq n_x - 1} \binom{n_x - 1}{n_g} \binom{A_1 - n_x - 1}{n_g - 1}$$

Figures 1 and 2 show the growth of $\bar{\mathrm{N}}_{L,A}$ for L in $1, \ldots, 10$. We set 10 as maximum motif length based on our observations (Fig. 2B).

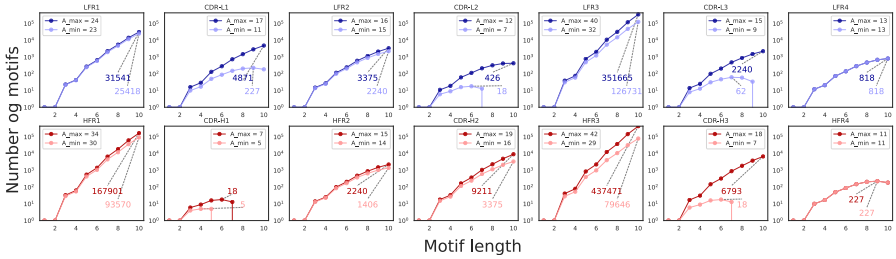


Figure 1: The number of unique motifs (Y axis) for a given motif length (X axis) that could be located in a certain FR/CDR (see possible FR/CDR lengths in Supplementary Table S1). The amino acid length of the motifs is bounded by the minimum and maximum possible region length (Supplementary Table S1).

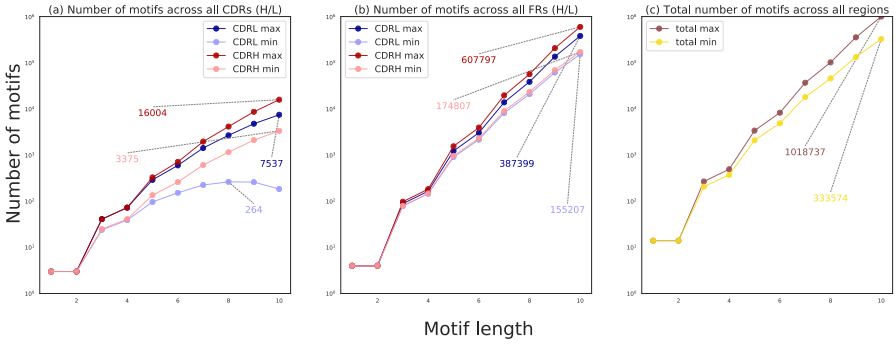


Figure 2: The total number of unique motifs (Y axis) for a given length (X axis) across all CDR-Ls and CDR-Hs **(a)**, across all LFRs and HFRs **(b)**, across all regions **(c)**.