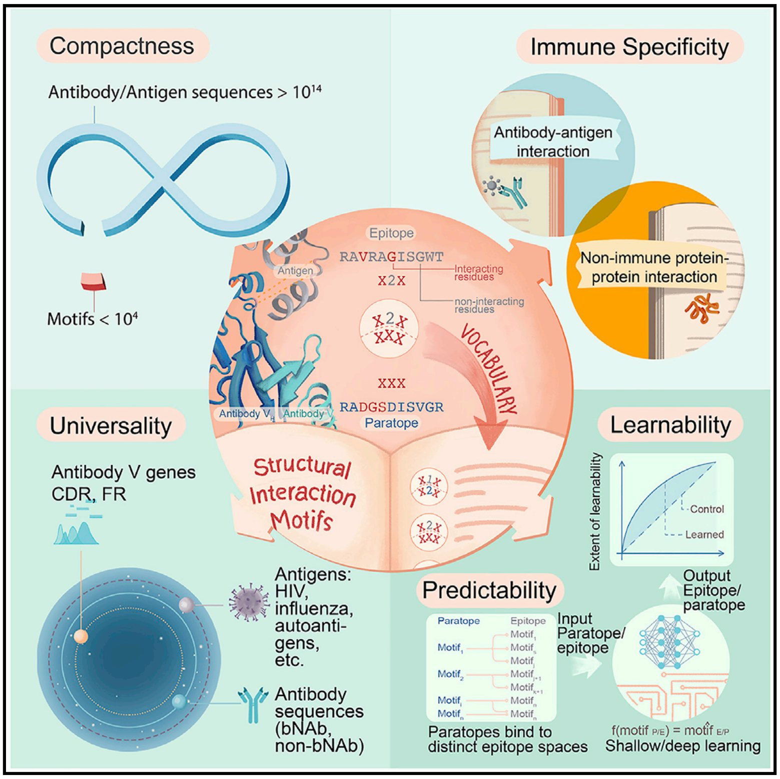
Cell Reports

Article

A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding

Graphical Abstract



Highlights

- Prediction of antibody-antigen binding is a central question in immunology
- A motif vocabulary of paratope-epitope interactions governs antibody specificity
- Proof of principle that antibody-antigen binding is predictable
- Implications for *de novo* antibody and (neo-)epitope design

Authors

Rahmad Akbar, Philippe A. Robert, Milena Pavlović, ..., Yana Safonova, Geir K. Sandve, Victor Greiff

Correspondence

rahmad.akbar@medisin.uio.no (R.A.),
victor.greiff@medisin.uio.no (V.G.)

In brief

Prediction of antibody-antigen binding is a central question in immunology and of high relevance for predictive antibody and vaccine design. Akbar et al. prove the predictability of antibody-antigen binding by discovering a universal, compact, and immunity-specific motif vocabulary of paratope-epitope interactions.

Akbar et al., 2021, Cell Reports *34*, 108856
March 16, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.celrep.2021.108856

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding

Rahmad Akbar,[1,*] Philippe A. Robert,[1,11] Milena Pavlović,[2,9,10,11] Jeliazko R. Jeliazkov,[3] Igor Snapkov,[1] Andrei Slabodkin,[1] Cédric R. Weber,[4] Lonneke Scheffer,[2,9] Enkelejda Miho,[5] Ingrid Hobæk Haff,[6] Dag Trygve Tryslew Haug,[7] Fridtjof Lund-Johansen,[1] Yana Safonova,[8] Geir K. Sandve,[2,9,10] and Victor Greiff[1,12,*]

[1]Department of Immunology, University of Oslo, Oslo, Norway
[2]Department of Informatics, University of Oslo, Oslo, Norway
[3]Department of Biochemistry, University of Zürich, Zürich, Switzerland
[4]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland
[5]Institute of Medical Engineering and Medical Informatics, School of Life Sciences, FHNW University of Applied Sciences and Arts Northwestern Switzerland, Muttenz, Switzerland
[6]Department of Mathematics, University of Oslo, Oslo, Norway
[7]Department of Philosophy, Classics, History of Arts and Ideas, Oslo, Norway
[8]Computer Science and Engineering Department, University of California, San Diego, La Jolla, CA, USA
[9]Centre for Bioinformatics, University of Oslo, Norway
[10]K.G. Jebsen Centre for Coeliac Disease Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[11]These authors contributed equally
[12]Lead contact
*Correspondence: rahmad.akbar@medisin.uio.no (R.A.), victor.greiff@medisin.uio.no (V.G.)

https://doi.org/10.1016/j.celrep.2021.108856

SUMMARY

Antibody-antigen binding relies on the specific interaction of amino acids at the paratope-epitope interface. The predictability of antibody-antigen binding is a prerequisite for *de novo* antibody and (neo-)epitope design. A fundamental premise for the predictability of antibody-antigen binding is the existence of paratope-epitope interaction motifs that are universally shared among antibody-antigen structures. In a dataset of non-redundant antibody-antigen structures, we identify structural interaction motifs, which together compose a commonly shared structure-based vocabulary of paratope-epitope interactions. We show that this vocabulary enables the machine learnability of antibody-antigen binding on the paratope-epitope level using generative machine learning. The vocabulary (1) is compact, less than 10^4 motifs; (2) distinct from non-immune protein-protein interactions; and (3) mediates specific oligo- and polyreactive interactions between paratope-epitope pairs. Our work leverages combined structure- and sequence-based learning to demonstrate that machine-learning-driven predictive paratope and epitope engineering is feasible.

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 antibody-antigen binding is predictable (Brown et al., 2019; Raybould et al., 2019a; Sela-Culang et al., 2013).

Antibody binding to the epitope is mainly formed by the three hypervariable regions termed 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 enables antibodies to interact specifically with virtually any 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). 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 (ABRs) (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 (Benjamin et al., 1984; Berzofsky, 1985; Burkovitz et al., 2013; Dalkas et al., 2014; Greiff et al., 2020; Jespersen et al., 2019; Kringelum et al., 2013; Kunik and Ofran, 2013; Lawrence and Colman, 1993; MacCallum et al., 1996; Mahajan et al., 2019; Ofran et al., 2008; Peng et al., 2014; Ponomarenko and Bourne, 2007; Raghunathan et al., 2012; Sela-Culang et al., 2013; Sivalingam and Shepherd, 2012).

Recently, computational and machine learning approaches for the sequence-based and structural prediction of paratopes (Deac et al., 2019; 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., 2019; Jespersen et al., 2019; Kilambi and Gray, 2017; Krawczyk et al., 2013) are accumulating (for a more complete list of references, see Brown et al., 2019; EL-Manzalawy et al., 2017; Esmaielbeiki et al., 2016; Norman et al., 2020; Raybould et al., 2019a; and Sanchez-Trincado et al., 2017). Although the accuracy for the prediction of paratopes seems generally higher than that for epitopes, it has not been conclusively shown that antibody-antigen interaction is *a priori* predictable and if so, based on what theoretical and biological grounds (Brown et al., 2019; Greenbaum et al., 2007).

Recent reports have provided preliminary evidence for the potential predictability of antibody-antigen interaction. First, the antibody repertoire field has now established that antibody sequence diversity underlies predictable rules (Elhanati et al., 2015; Greiff et al., 2017a, 2017b). Second, 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 toward 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. The 3D complex structure of an epitope typically emerges from different sub-peptides of the protein, folded in the same place. Therefore, the binding units reach beyond a single linear peptide, hindering the power of sequence-based prediction tools. We conjectured that the comparison of those interaction units across antibody-antigen complexes may lead to the discovery of a general vocabulary of antibody-antigen interaction. If a general compact (restricted) vocabulary for antibody-antigen interaction existed that unambiguously formed paratope-epitope pairs, then paratope-epitope interaction is *a priori* predictable. Here, we show that such a vocabulary exists.

RESULTS

The majority of paratope interacting residues are located in the antibody CDRs

To gain a representative picture of antibody-antigen 3D interaction, we compiled a diverse dataset of 825 non-redundant antibody-antigen complexes (protein antigen only) (Figure 1A). Antibody sequences mapped to a diverse set of V genes (Figure S1A), and antigen sequences belonged to a diverse set of antigen classes (Figures S1B–S1G). Thus, the dataset is neither biased to one type of antibody or antigen class nor to sequences of high similarity.

We identified the set of interacting residues at the interface of antibody-antigen structures by using a heavy-atom (non-hydrogen atoms) distance cutoff of <5 Å (Ostmeyer et al., 2019) (see STAR Methods and Figure S3 for an examination of the robustness of the distance cutoff). Antibody-antigen amino acid pairs within this distance were designated as interacting residues (Figures 1A and 1B). 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 CDRs 1–3 ($V_{H,CDR1–3}$: 89.5% and $V_{L,CDR1–3}$: 89.2%; Figure 1C). Because we used the Martin numbering scheme for CDR and framework region (FR) annotation (see STAR 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 the position of paratope interacting residues correlated significantly ($p < 0.05$) with sites of (inferred) somatic hypermutation hotspots (SHMs) (Spearman/Pearson correlation: 0.31–0.52/0.44–0.58; Figure S2E), suggesting that interacting residues investigated herein have been subjected to antigen-driven selection.

Paratopes are enriched in aromatic and polar residues, whereas epitopes are enriched in charged residues

We found an enrichment of aromatic residues (e.g., tyrosine) in paratope sequences and polar and charged residues in epitopes (e.g., lysine and arginine) (Figures S2A–S2D), in accord with published literature (Peng et al., 2014; Ramaraj et al., 2012) validating further the robustness of our definition of interacting residues.

Amino acid usage correlation was relatively low between paratope and epitope residues ($r_{Spearman/Pearson}$: 0.13–0.49; Figure S5E). Paratope and non-immune protein-protein interaction (PPI) residues were uncorrelated ($r_{Spearman/Pearson}$: 0.06–0.29), whereas PPI and epitope were moderately correlated ($r_{Spearman/Pearson}$: 0.57–0.71; Figure S5E).

Finally, we investigated amino acid contact pairs of paratope and epitopes across CDR/FR regions and non-immune PPI. We found substantial cross-type (type: charged, polar, aromatic, hydrophobic/nonpolar) interactions both at the local (CDR/FR region wise) and the global level (full sequence; Figure S7C), while in contrast, PPI amino acid interaction preferences were predominantly within type (Figure S7D).

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

The fundamental units of antibody-antigen interaction are the sequence regions on the antibody and the antigen that comprise the interacting (paratope, epitope) and non-interaction residues

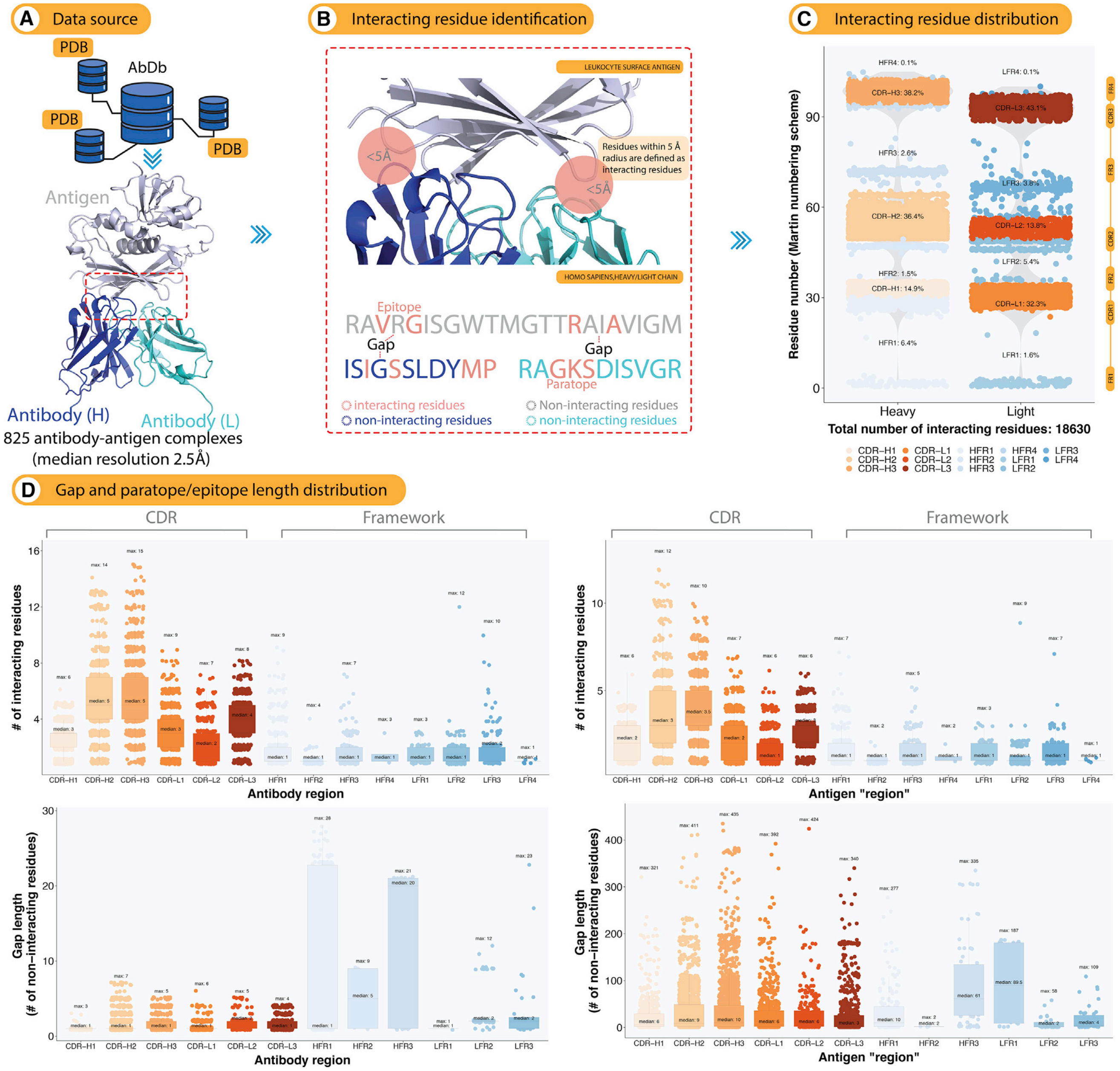


Figure 1. Characterization of interacting residues at the paratope-epitope interface

(A) We characterized antibody-antigen interaction using 825 publicly available 3D structures deposited in the Antibody Database (AbDb) (see STAR Methods).

(B) A paratope was defined as the set of interacting amino acid residues within a particular FR or CDR 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 of the corresponding paratopes (see antigen "region" in D; see STAR Methods). Gaps are defined as the non-interacting residues that lie in between interacting residues within each FR or CDR.

(C) The interacting residues mapped predominantly to the CDRs and less so to the FRs.

(D) Number of gap (non-interacting) and interacting residues by CDRs and FRs.

(gaps, Figure 1B). Paratope lengths ranged between 1 and 15 (median: 1–5) residues in the CDRs and 1 and 12 (median: 1–2) residues in the FR regions, whereas the length of gaps ranged between 1 and 7 (median: 1–2) residues in CDRs and 1 and 28 (median: 1–20) residues in the FR regions (Figure 1D). Epitopes can span up to 12 residues with gaps up to 435 residues (Figure 1D). Thus, the paratope-epitope interface cannot be described exclusively by continuous stretches of amino acids.

To compare structural patterns of paratope-epitope across unrelated antibody-antigen complexes, we devised a structural interaction motif notation that accounts simultaneously for gaps and residues in both paratopes and epitopes. A paratope or

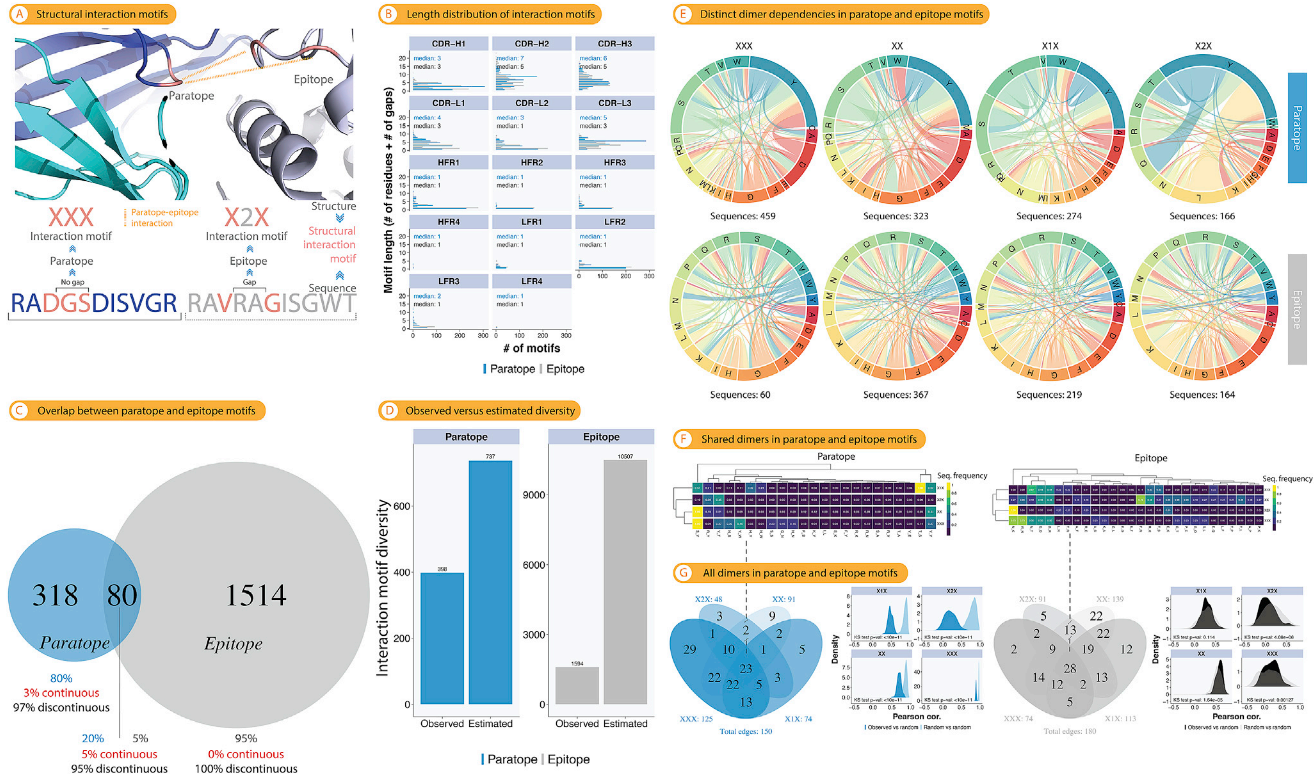


Figure 2. Structural interaction motifs represent a compact vocabulary for the composition of the paratope-epitope interface

(A) We devised a structural interaction motif notation that accounts simultaneously for gaps and interacting residues in both paratopes and epitopes.

(B) Length distribution of paratope and epitope motifs by FR/CDR.

(C) Absolute and relative overlap of paratope and epitope motifs (Venn diagram).

(D) Estimation of the potential (observed + unobserved) motif diversity using the Chao1 estimator (see STAR Methods).

(E) For each of the four most highly shared (across structures) interaction motifs (Figures S4A–S4I), the sequential (dimer [2-mer]) dependency signature was determined (see STAR Methods).

(F) Hierarchical clustering of sequential dependencies (2-mers) that were shared among all four paratope or epitope motifs.

(G) Venn diagrams: overlap of sequential dependencies (2-mers) shared across paratope or epitope motifs. Density plots: we tested whether the 2-mer distribution (sequential dependencies) observed in (F) for each of the four motifs could be caused by random effects. To this end, we sampled 100 times 2-mers from the number of 2-mers possible according to the number of sequences mapping to each motif (E) and calculated the correlation either among all randomly drawn 2-mer distributions (gray: epitope, light blue: paratope) or between an observed and randomly drawn one (black: epitope, dark blue: paratope). The significance in the difference between the distributions was tested using the Kolmogorov-Smirnov (KS) test.

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 (Figure 2A). The combination of amino acid and gap encodings is termed structural interaction motif (henceforth interaction motif or simply, motif). Therefore, motifs describe the spatial conformation of the binding and can be used in addition to residue information to characterize antibody-antigen binding. Our motif notation for antibody-antigen interaction places the paratope-epitope interface into a unified coordinate system that preserves the link between paratope and epitope and enables computational traceability of both continuous and discontinuous (structural) antibody-antigen interaction across antibody-antigen complexes. Thus, the motif definition now enables querying key parameters of antibody-antigen recognition: (1) motif sequence diversity, (2) structural diversity (motif angle and (dis)continuity), (3) co-occurrence across complexes, and (4) predictability and learnability of paratope-epitope interaction.

The combined set of paratope and epitope motifs was generally distinct from that found in non-immune PPI (Figure S5A). Paratope and epitope motif lengths varied across FR and CDR regions but remained below a maximum (max) length of 10 (median length: 1–7; Figure 2B). On average, three to four motifs were found per antibody heavy and light chain (Figure S4L).

The diversity of paratope and epitope interaction motifs is restricted (compact)

Out of 1,594 and 398 unique paratope and epitope motifs, only 80 motifs overlapped (Figure 2C). Therefore, we asked how much of the potential (observed + unobserved) paratope and epitope motif diversity is covered by our dataset of 825 antibody-antigen structures? To answer this question, we used the Chao1 estimator (Chao, 1984) (see STAR Methods) and found that for paratopes, the set of unique motifs in our dataset covered about 50% (398) of the potential diversity of all paratopes (estimated total: 737), whereas the set of unique epitope motifs in our dataset covered 15% (1,594) of the total epitope diversity (10,507; Figure 2D). The estimated total size of the paratope motif space is one order of magnitude smaller than the analytically derived theoretical size ($\approx 10^5$; see Methods S1). Of interest, the size of the potential epitope motif space is similar to that of the PPI motif space (Figure S5H).

To summarize, the estimated potential motif space is smaller ($<10^4$) than the total number of antibody sequences ($>10^{14}$) (Briney et al., 2019; Elhanati et al., 2015) by at least 10 orders of magnitude. Our dataset captures a substantial portion of the total motif space indicating the restriction of the paratope-epitope interaction motif space.

Interaction motifs have a unique sequential amino acid signature suggesting immunological function

Because structural interaction motifs retain association with their underlying paratope and epitope sequences, we were able to ask whether structural interaction motifs group sequences with common sequence signatures. If so, it would suggest that structural interaction motifs bear distinct immunological and biochemical function. To investigate the sequence dependencies within selected multi-residue (length > 1) paratope and epitope interaction motifs, we determined the 2-mer decomposition of the sequences mapping to the four most shared paratope/epitope motifs (see STAR Methods) and found that the sequential dependencies indeed differed among paratope and epitope motifs, respectively (Figures 2E–2G), and thus may have immunological function. Non-immune PPI structural interaction motifs differed from both paratope and epitope ones (Figures S1H, S1I, S7A, and S7B).

The structure of paratope and epitope motifs differs across CDR and FR regions

To address the question whether paratope and epitope interaction motifs differ structurally, we measured the ‘‘angle’’ of each motif (see STAR Methods). In general, median epitope motif angles were only maximally as high as paratope motif angles (Figure S6A). 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; Figure S6B). To further substantiate our structural motif analysis, we compared Ramachandran plot statistics (distribution of backbone dihedral angles) between paratope, epitope, and PPI motifs (Figure S6C). In addition to verifying that FR and CDR use different angles (Figure S6A), we found that PPI mostly manifests as alpha helix, whereas antibodies in antibody-antigen complexes mostly manifest as beta-strand/sheet, P_{II} spiral, and delta turn, thus underlining the uniqueness of immune protein interaction (Figure S6C).

Taken together, we showed paratope and epitope motifs vary structurally across FR and CDRs and are structurally distinct from PPI motifs and, thus, to a large extent are unique to antibody-antigen recognition.

Paratope motifs are shared across antibody-antigen complexes

We quantified both the number of continuous and discontinuous motifs for each FR/CDR region (Figure 3A) and the number of complexes that share identical motifs (Figure 3B). We found the following: (1) 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; (2) paratope motifs are predominantly continuous; and (3) continuous paratope motifs, more so than discontinuous, are shared across antibody-antigen complexes; epitope motifs exhibited substantially less sharing and were more discontinuous. Specifically, we found that only 10 paratope motifs and 5 epitope motifs (Figures S4D, S4E, S4H, and S4I) were present in at least 10% (or 82 in absolute numbers) or more complexes. Importantly, shared paratope interaction motifs were not specific to a given class of antigens (Figure S4F), nor to a specific germline gene (Figure S4G), and 13 of the most shared interaction motifs were also found in HIV broadly neutralizing antibodies (bNAbs) (Figures S4J and S4K), underlining the generality and diversity restriction of the interaction motifs investigated here.

More generally, 8% of all paratope motifs were shared across at least 20 complexes, whereas that was the case for only 1% of the epitopes (Figure 3C). Epitopes were similar in sharing behavior to PPI motifs, with only 3% being shared across 20 or more complexes (Figure S5P).

A selected number of paratope motifs show broad polyreactivity toward mutually exclusive epitope motif spaces demonstrating *a priori* predictability of antibody-antigen binding

We next asked whether paratope and epitope motifs have preferred motif partners, which would indicate *a priori* predictability of paratope-epitope binding. To answer this question, we constructed a paratope-epitope-motif network (a bipartite graph) by connecting each epitope motif to its cognate paratope motif (Figure 4A). We termed such a 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 (Figures 4A and 4C). Together, these top 7 paratope motifs made 829 connections to predominantly different epitopes (Figure 4E, inset), thereby collectively binding to ≈50% of all unique epitope motifs (Figure 2D). Thus, although these paratope motifs showed broad polyreactivity, they bound to largely entirely different epitope motif groups. We found that the degree distribution of the paratope-epitope reactivity graph was power-law-distributed and scale free (Figure 4A) (Clauset et al., 2009). To exclude the possibility that the connectivity patterns observed were simply due to the fact that there are more epitope motifs than paratope motifs (Figure 2C; Figure S7E), we demonstrated that random reactivity networks showed the following: (1) no power law ($p < 0.1$); and (2) an increased overlap of bound partner motifs, and thus significantly lower specificity (Figures 4B, 4D, and 4E). Finally, we found that non-immune protein-protein reactivity networks differed from paratope-epitope ones (Figures S5I–S5M).

To summarize, the top-connected paratope motifs in the reactivity network show polyreactivity toward distinct epitope spaces that are non-overlapping (polyreactive specificity). Most motifs, however, are oligoreactive and thus highly specific.

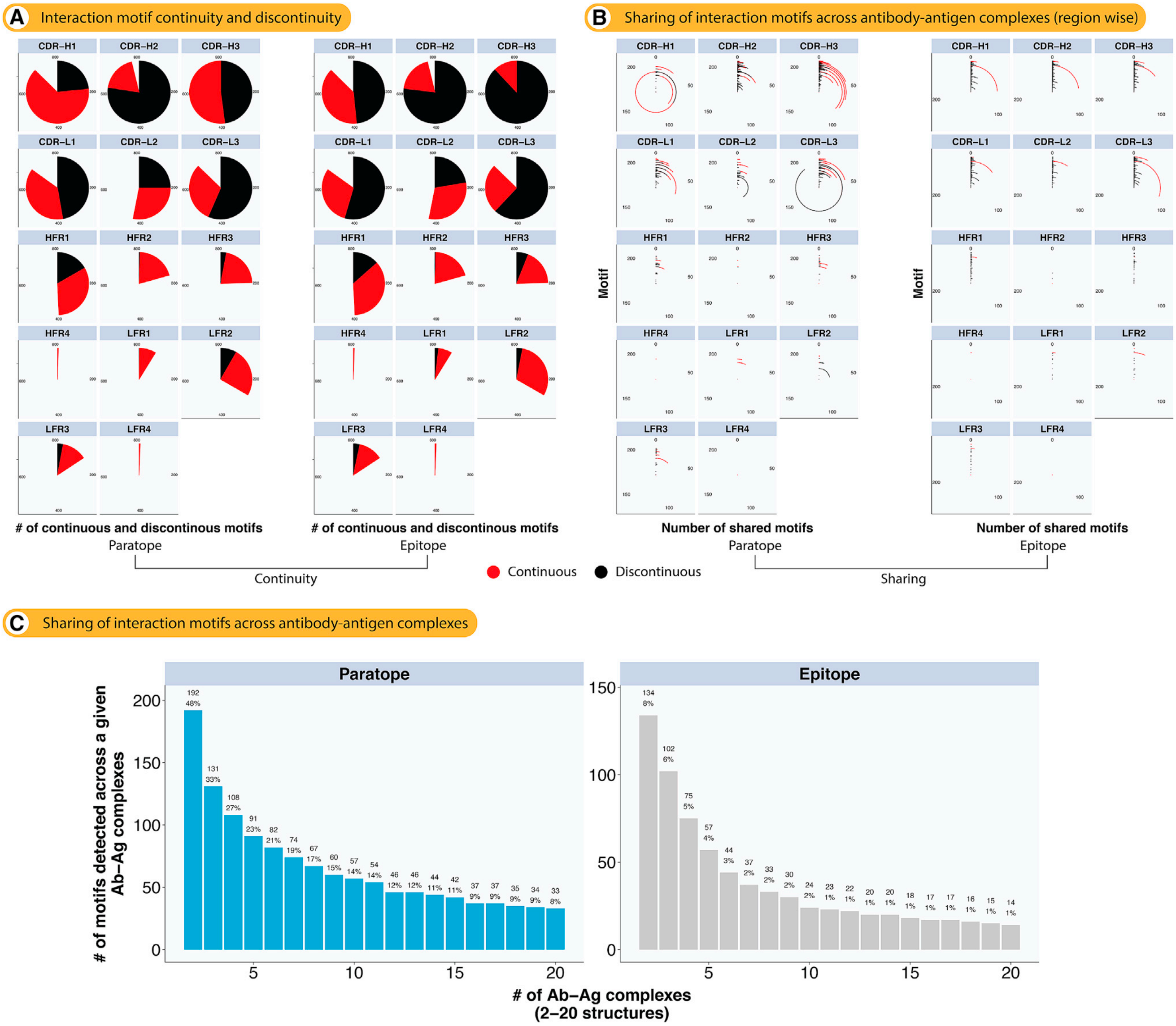


Figure 3. Paratope interaction motifs show a higher extent of continuity and sharing across antibody-antigen complexes than epitope interaction motifs

(A) Ratio of continuous (absence of non-interacting residues) and discontinuous (presence of at least one non-interacting residue/gap) paratope and epitope interaction motifs across antibody-antigen complexes. For example, for paratope CDR-H3, the pie chart signifies that in ≈50% of the complexes, CDR-H3 motifs are continuous and in 50% discontinuous. Gaps in pie charts indicate that for a given region not all structures showed interacting residues.

(B) Absolute number of antibody-antigen structures containing a given interaction motif by CDR/FR.

(C) Absolute and relative number of motifs found across at least 2–20 antibody-antigen complexes (x axis).

The combined high specificity and distinctiveness of paratope-epitope interaction indicates that paratope-epitope binding is *a priori* predictable.

Quantification of machine learnability of paratope-epitope interactions

The paratope-epitope reactivity map indicates *a priori* predictability of antibody-antigen binding (Figure 4A). To quantify the accuracy (learnability) with which one can predict (translate) one paratope interaction motif (or sequence) into the cognate epitope interaction motif (or sequence) and vice versa, we leveraged both shallow and deep learning (Figure 5, bottom panel). We evaluated model performance by comparing (1) the motif length and (2) edit distance (error) of the predictions and the true paratope motifs or sequences (see STAR Methods).

As the deep models scale with the complexity of the parameters (hidden dimension and embedding dimension), we observed an increasingly positive correlation between the lengths of prediction and true motifs (sequences) (Figures S6D

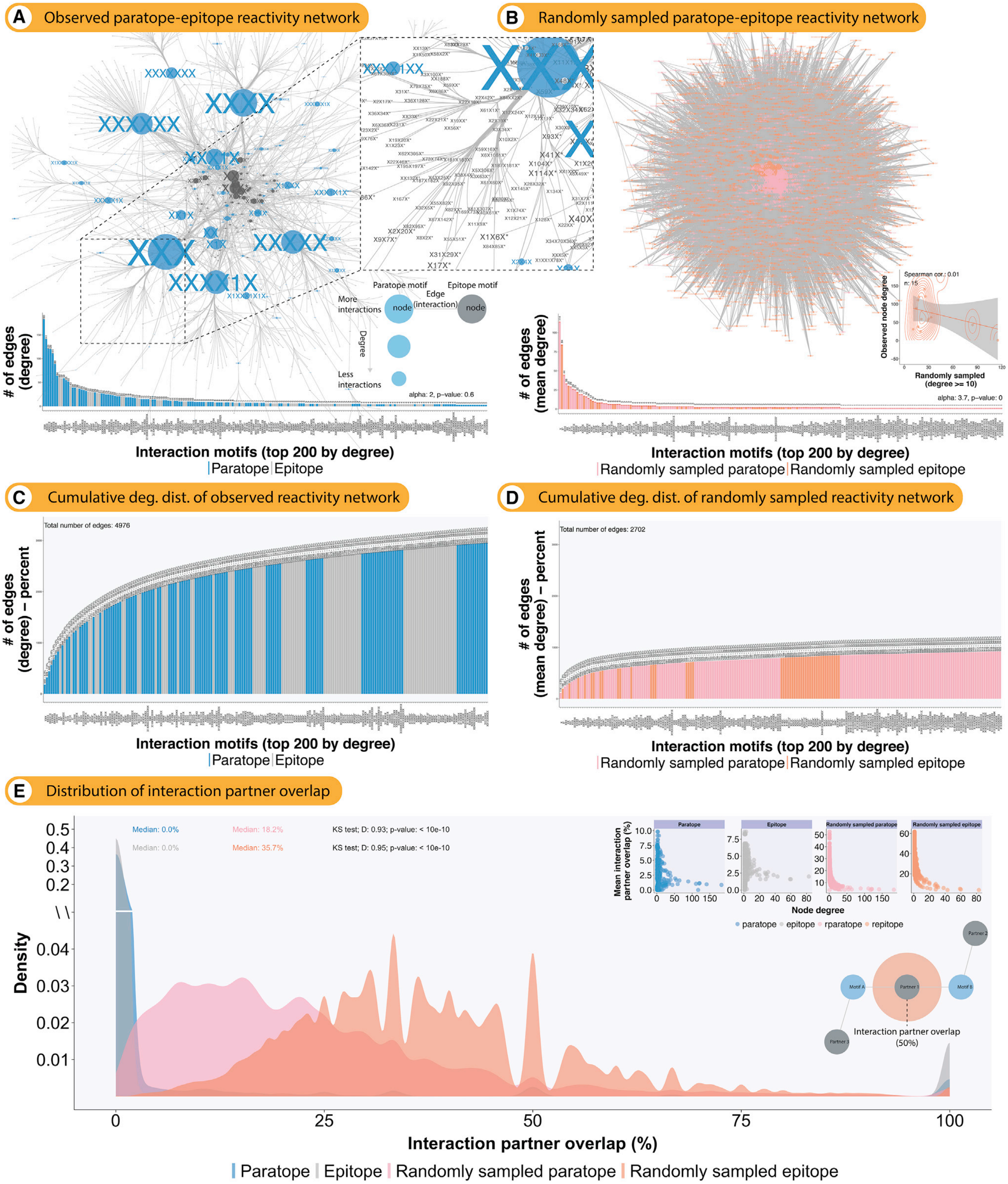


Figure 4. The majority of paratope and epitope motifs are oligoreactive, while a small number of paratope motifs show broad polyreactivity toward distinct epitope motif spaces

The reactivity network of paratope-epitope interaction motifs indicates *a priori* predictability of antibody-antigen binding.

(legend continued on next page)

and S6E, $r_{Pearson}$ 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 (Figure S6E).

For both shallow and deep learning models, the medians of prediction error of interaction motif use cases ranging between 0.25 and 0.42 (accuracy 58%–75%) were substantially lower than those of sequence 0.78–0.87 (accuracy 13%–22%) use cases (Figure 5). These results indicate that the paired paratope-epitope interaction motif space reaches reasonable accuracy, whereas the sequence space remained challenging to predict. We observed similar trends— prediction accuracy at interaction motif level is higher than at sequence level—when examining PPI data (Figure S6F).

Given that structural interaction motifs represent one of the layers of antibody-antigen binding, we asked whether integrating motif and sequence information improves sequence-based prediction. Indeed, when combining structural motifs and sequences to an “aggregate,” the prediction accuracy of the deep model, but not shallow models, improved by 2–7 percentage points as compared with the sequence-only use case (Figure 5). Thus, adding structural information improves the sequence-based prediction accuracy of antibody-antigen binding, possibly because it removes interaction ambiguity from the paratope-epitope reactivity space (Figures S7F and S7G).

DISCUSSION

Our results demonstrate the existence of learnable sequence and structural rules in 3D antibody-antigen interaction. We performed an unbiased search for binding fingerprints in a set of 825 curated antibody-antigen structures and discovered a compact vocabulary of antibody-antigen interaction in the form of structural interaction motifs. We showed that the motif vocabulary is a valuable feature for the development of paratope-epitope prediction tools. 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 suggesting that our motif definition captures underlying immunological principles. To provide quantitative robustness to our findings, our study contains, to our knowledge, one of the most comprehensive statistical evaluations of antibody-antigen and non-immune protein-protein with respect to the following: (1) the distribution of amino acid residues at the binding interface, (2) the extent of binding interface (dis)continuity, (3) quantification of interaction complex sequence similarity, (4) a range of structural definitions of paratope and epitope interaction, and (5) the relationship between somatic hypermutation sites and paratope-epitope contact residues (Stave and Lindpaintner, 2013). While one of the main aims of this work 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. Indeed, future studies may investigate alternative motif definitions (possibly identified by end-to-end machine learning) that could unveil further structure/patterns in the antibody-antigen interaction space. Finally, to our knowledge, similar work on TCR-peptide interaction has not been performed yet. 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; Lanzarotti et al., 2018; Ostmeyer et al., 2019; Riley and Baker, 2018; Turner et al., 2006).

Antibody-antigen interaction operates via structural interaction motifs

The discovery of shared interaction motifs crucially depended on the FR/CDR-focused definition of paratope and epitope (Figure 1; see STAR Methods) because these are the locales of the fundamental binding units of antibody-antigen binding. In the future, once more antibody-antigen structures become available, one may attempt to search for motifs based on the entire antibody and antigen. A given antibody V_H (variable heavy) and V_L (variable light) has a median of three to four motifs (Figure S4L). Relatedly, Kunik and Ofran (2013) showed the six ABRs ($\approx$CDR-H/L1–3) differed significantly in their amino acid composition, and that each ABR tends to bind different types of amino acid at the surface of proteins (Van Regenmortel, 2014). Although we were able to confirm that paratope-epitope amino acid level contact maps differ across CDR/FR regions (Figure S7), 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 at the motif level and not at the amino acid level (Figure 4; Figures S7E–S7G). Specifically, structural interaction motifs encode geometric

(A) A bipartite reactivity network capturing paratope-epitope motif interaction was constructed by connecting each paratope motif to its corresponding epitope motif (undirected edge); given that paratope and epitope motifs may occur more than once across antibody-antigen structures, paratope and epitope motifs may have multiple network connections. Network vertices were scaled by their number of connections (degree). Only the largest connected portion of the network was visualized. The network degree distribution was tested to fit a power-law distribution (Clauset et al., 2009). A p value > 0.1 means that a power law cannot be ruled out. Inset shows a zoomed-in section of a paratope motif (blue) connecting to a diverse set of epitope motifs colored in gray (polyreactivity).

(B) To confirm that the reactivity network architecture observed in (A) was unlikely to be observed by chance, we randomly sampled 100 times 1,000 motifs from paratope and epitope motif distributions (Figures 4D and 4E). Inset: 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). The power-law fit was done as described in (A).

(C and D) Cumulative degree distributions of networks (A) and (B).

(E) Distribution of interaction partner overlap for networks (A) and (B). In brief, 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 KS test. Inset: node degree as a function of interaction partner overlap.

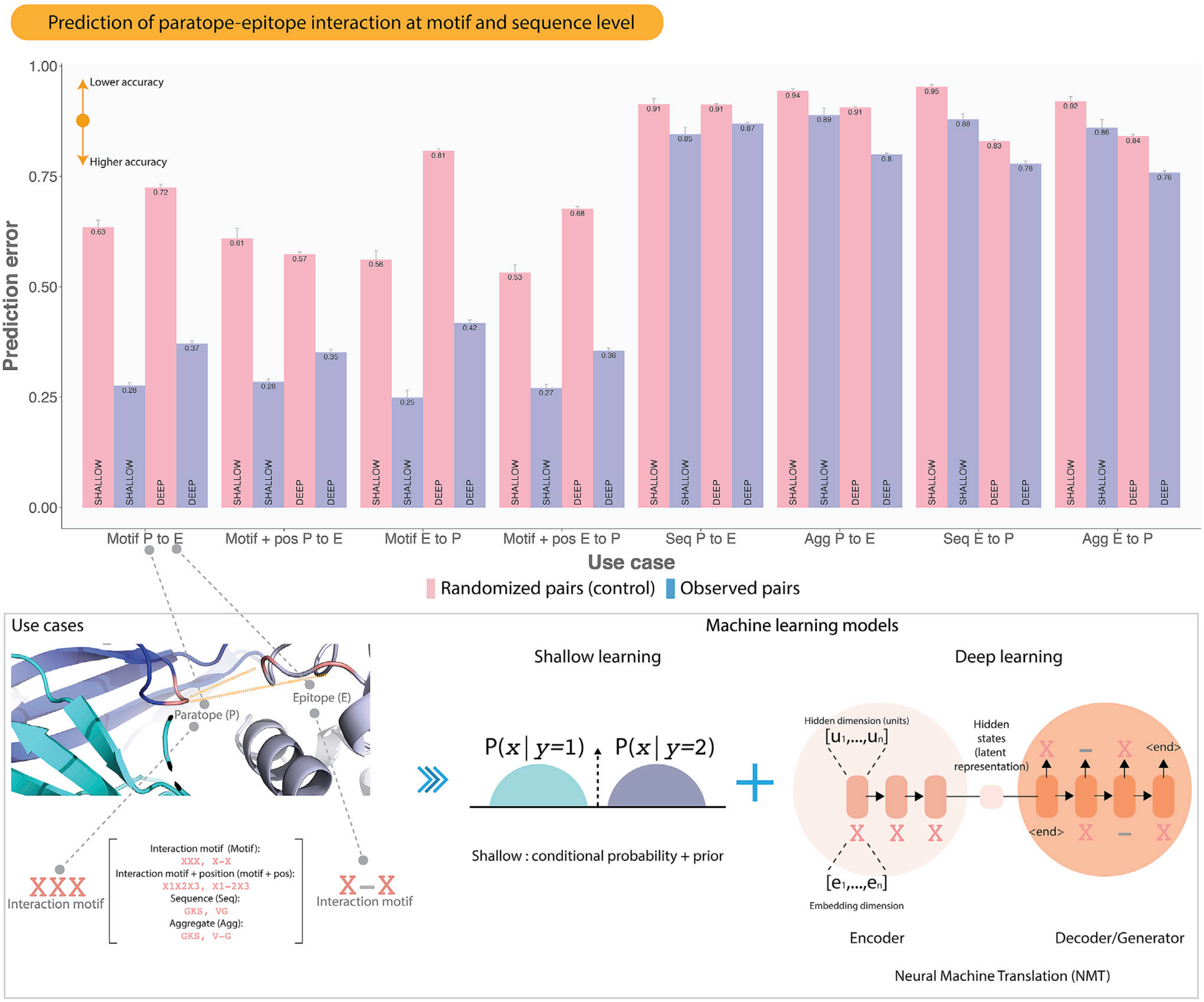


Figure 5. Quantification of machine learnability of paratope-epitope interactions at motif and sequence level

(Bottom) Schematic of the paratope-to-epitope and epitope-to-paratope prediction tasks (use cases). To quantify the learnability of antibody-antigen interactions at motif and sequence levels, four distinct use cases were used: (1) interaction motif; (2) interaction motif with positional index; (3) sequence; and finally, (4) motif and sequence aggregate (all use cases are explained with examples in STAR Methods). We leveraged both deep and shallow machine learning approaches. (Top) The median prediction error was obtained by calculating the median Levenshtein distance between the output and the predicted output for each use case across all parameters. The distance ranges from 0 (perfectly matching output-predicted-output, high prediction accuracy) to 1 (fully dissonant output-predicted-output pairs, low prediction accuracy). 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. Baseline prediction accuracies (control) were calculated based on label-shuffled data where antibody and antigen-binding partners were randomly shuffled (randomized pairs). Total unique pairs for motif, sequence, and aggregate levels are 2,847, 3,967, and 3,986, respectively. Error bar: ±2 × standard error $\left(SD/\sqrt{n}\right)$.

information of the local structure and are therefore linked to the angle of folding (Figures S6A and S6B). Thus, linking sequence to motifs and motifs to binding in a two-step process may connect, in the future, local folding to global specificity. Indeed, we found that merging motif and sequence information increased the prediction accuracy of paratope-epitope interaction (Figure 5). Interestingly, in HIV bNAbs (Figures S4J and S4K), the same motifs as in non-bNAbs were used, underlining the generality of the motif vocabulary here discovered and characterized (Chuang et al., 2019).

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

We identified not only predefined dependencies within sequences mapping to 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 (Figure 4) with polyreactivity of a few selected paratope motif ''hubs'' and general oligospecificity of the majority of paratope and epitope motifs. The highly polyreactive paratope motifs were more continuous and 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.

Humoral specificity describes the capacity of the antibody immune response 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 lower-dimensional task. It is tempting to speculate as to the evolutionary advantage of short motifs in antigen recognition (non-immune PPI has evolved substantially larger motifs; Figures S5B and S5C). Short motifs decrease the potential escape space for antigens but also render self- and non-self-recognition more difficult. It would be interesting to investigate whether, for example, autoimmunity and infections (bacterial, viral) occupy different motif spaces. So far, however, we observed that paratope motifs were shared across antigen classes (Figure S4F).

Predictability and learnability of the paratope-epitope interface

Paratope-epitope prediction is a task known in the structural bioinformatics field as binding site prediction and is typically formulated as the problem of finding the set of residues (or patches) on the protein surface likely to interact with other proteins (Akbar and Helms, 2018; Akbar et al., 2017; Hwang et al., 2016; Jordan et al., 2012; Northey et al., 2018; Porollo and Meller, 2007). This problem can be formalized as a binary classification task in which a model is trained to discriminate binders from non-binders at residue or sequence level. That is, given a sequence VGRAISPRAS, the model would assign a probability to each amino acid signifying its likelihood to bind to a partner residue (Pbind(V) = 0.3, Pbind(G) = 0.05, Pbind(R) = 0.7 etc.), or Pbind(VGRAISPRAS) = 0.6). Here, we went beyond a binary (binder or non-binder) classification setting toward a more nuanced multi-class setting. Specifically, we asked for a given paratope motif, the corresponding epitope motif (multi-class setting), instead of whether the motif binds or not (binary-class setting). Prediction for the multi-class classification setting proved reasonably successful at the motif level (Figure 5). Transitioning into a higher-dimensional multi-class setting, such as the transition between the sequence and aggregate encodings, adversely impacted our shallow model more so than the deep one, suggesting that as the class diversity tends to infinity, the deep model would increasingly outperform the shallow one.

More generally, a target-agnostic approach, such as Paratome (Kunik et al., 2012b), finds a set of regions from a structural alignment of antibody-antigen complexes and uses it to locate similar regions in new sequences. Antibody i-Patch (Krawczyk et al., 2013) utilizes contact propensity data to score each amino acid while at the same time leveraging information from neighboring residues (a patch of residues). To be applicable for antibody-antigen prediction, the original i-Patch (Hamer et al., 2010) algorithm, however, would have to be adjusted with respect to two central assumptions: (1) multiple sequence alignment (MSA) signifying evolutionary (sequence/structural) conservation in protein-interacting domains; and (2) protein-protein-derived amino acid propensity score, both of which are less pertinent for antibody-antigen complexes as antibodies (as does surviving antigens) constantly evolve obscuring many forms of conservations, structural and sequence alike. Specifically, CDRs in antibodies manifest as unstructured loops with minimal structural conservation across antibodies, and amino acid propensity differs between protein-protein and antibody-antigen complexes. This combination compounds the complexity in learning the rules that govern antibody-antigen interaction and necessitates a unique approach separate from the conventional approaches presently applied in the PPI field. The motifs discovered here fill this missing gap by capturing structural and sequence information in a single notation across antibody-antigen complexes and projecting antibody-antigen interaction onto substantially lower dimensions (10^2 paratope and 10^3 epitope motifs), which allowed us to observe conservation from a motif's perspective. For instance, we showed in Figures 3B and 3C and S4F–S4I that motifs are ''conserved'' (shared) across different antigen classes, V genes, and structures. Tools such as Antibody i-Patch may, for instance, leverage a motif-driven alignment in place of the missing MSA data because of sequence diversity of antibody-antigen complexes.

Beyond target-agnostic approaches, accumulating evidence has demonstrated the utility of integrating the information from the interacting partner in improving state-of-the-art performance (Ahmad and Mizuguchi, 2011). Townshend et al. (2019) achieved state-of-the-art performance for the prediction of PPI by training a model that comprises two separate convolutional neural networks (one from each interacting partner) and concatenating them to produce the final output. Similarly, Pittala and Bailey-Kellogg (2019) used an attention layer on top of two separate convolutional layers (one each for antibody and antigen) to produce superior predictions to target agnostic approaches, such as DiscoTope and Antibody i-Patch (Andersen et al., 2006; Krawczyk et al., 2013). Finally, Deac et al. (2019) eclipsed the performance of the target-agnostic Parapred approach by building a model that cross-modally attends antigen residues (Liberis et al., 2018). Although much more sophisticated in terms of model complexity and architecture in addition to ''target-aware''-ness, these models remain anchored to the problem of delineating binders and non-binders (binary prediction) and have yet to venture to a multi-class setting. We note as well that antibody-epitope prediction is typically treated separately (1) to predict residues or sequences in antibodies that bind to epitopes (paratope prediction), and (2) to predict residues or sequences in antigens that bind to paratopes (epitope prediction). In this dichotomy, paratope prediction typically fares several folds better than epitope prediction (for context, a state-of-the-art predictor from Pittala and Bailey-Kellogg (2019) yielded areas under the precision-recall curve [AUC-PR] of 0.7 and 0.212 for paratope and epitope prediction respectively). In contrast, we observed a notably less dramatic difference at least at the motif level, where accuracy ranges from 0.63 to 0.72 and 0.58 to 0.75 for paratope and epitope prediction, respectively (Figure 5). Thus, we speculate that "motif-awareness" may further extend the performance of these approaches similar to the added benefit of target-awareness earlier described and may bridge the dichotomy between paratope and epitope prediction.

Given that there exist a few paratope motifs with broad epitope motif reactivity, the motif-based prediction accuracy of paratope-epitope interaction cannot reach, by definition, 100% (as opposed to other potential paratope/epitope encodings) (Figure 4; Figures S7E–S7G). However, because the epitope reactivity of the polyreactive paratopes was mutually exclusive (distinct), focusing prediction efforts on branches of the paratope-epitope reactivity network, as well as increasing the amount of data to train and build the network, may improve the performance of sequence-based paratope-epitope prediction models, especially because we discovered the following: (1) motifs possess distinct sequential dependency signatures, and (2) motifs aid the sequence-based prediction of paratope-epitope pairing (sequence-motifs aggregates; Figure 5).

Implications for machine-learning-driven antibody, epitope, and vaccine engineering

Monoclonal antibodies are of substantial importance in the treatment of cancer and autoimmunity (Brown et al., 2019; Csepregi et al., 2020; Ecker et al., 2015). Thus, their efficient discovery is of particular interest. Given that our work is unbiased toward both paratope and epitope analysis, it demonstrated the feasibility of the reconstruction, via machine learning, of potential neo-epitopes for neo-epitope design or the discovery of neo-epitope-specific antibodies. Our analyses suggest that the number of antibody-binding motifs is relatively restricted (Figures 1 and 2). 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 toward the interaction motifs (and their corresponding sequential bias) discovered here (Amimeur et al., 2020; Chen et al., 2020). 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, 2017). 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 MSAs 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 by themselves as to the antigen targeted (more paratope-epitope-paired data are needed for such investigations) (Mason et al., 2019).

We found that contact residues and somatic hypermutation are in fact correlated (Figure S2E). For antibody optimization, this suggests that linking the antigen-contacting and somatically hypermutated positions 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 SHM sites. Furthermore, it would be of interest to investigate the extent to which somatic hypermutation preserves binding motifs or, relatedly, how a reversal to germline would change interaction motifs. The latter is a particularly important question because there is likely an overrepresentation of high-affinity antibodies in the dataset investigated here.

The antibodies studied here are diverse and can harbor specific structural features, such as glycans, for binding envelope proteins of HIV or influenza. Further, those antigens may also harbor glycans depending on the mode of protein synthesis before crystallization. Our dataset therefore inherently already contains the effect of post-translational modifications to antigen-antibody binding. Interestingly, we did not find substantial differences in the motif usage between bNAbs against viral glycoproteins, supporting that the vocabulary of motifs is shared among antigens with diverse structural features in the dataset. However, without more abundant experimental 3D-antibody-antigen binding data, we are unable to predict whether this holds true for proteins that cannot be crystallized or for unstructured loops of antigens, which are typically missing in structural databases.

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 will have to delineate to what extent non-interacting residues correlate with specific interaction motifs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - A dataset of non-redundant and diverse 3D antibody-antigen complexes
  - Selection of antibody sequence numbering scheme
  - Identification of interacting residues in antibody-antigen complexes
  - Definition of paratope, epitope, and paratope-epitope structural interaction motifs
  - Definition of interaction motif angle
  - Diversity analysis of interaction motifs
  - Paratope-epitope amino acid contact map
  - Construction of bipartite paratope-epitope and PPI reactivity networks at motif and sequence level
  - Analysis of sequential dependencies in interaction motifs
  - Dataset of protein-protein interaction and definition of protein-protein interaction motifs

- ○ Quantification of somatic hypermutation on antibody amino acid sequences
- ○ Ramachandran plot analysis
- ○ Machine-learning prediction of paratope-epitope and PPI at interaction motif, sequence and aggregate level
- ○ Use cases (encoding):
- ○ Deep learning
- ○ Shallow learning
- ○ Evaluation
- ○ Graphics
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2021.108856.

ACKNOWLEDGMENTS

We thank Andrew Macpherson (University of Bern, Switzerland), Anne Corcoran (Babraham Institute, UK), Ludvig M. Sollid (University of Oslo, Norway), Thérèse Malliavin (Institut Pasteur, Paris, France), Marieke Kuijjer (University of Oslo, Norway), and Luca Piccoli (Institute for Research in Biomedicine, Switzerland) for their comments and suggestions that have led to an improvement of the manuscript. We thank Yi Zheng (yizheng_art, fiverr.com) for the help with designing the graphical abstract. Financial support was provided by The Helmsley Charitable Trust (2019PG-T1D011 to V.G.); UiO World-Leading Research Community (to V.G.); UiO:LifeScience Convergence Environment Immunolingo (to V.G. and G.K.S.); EU Horizon 2020 iReceptorplus (825821 to V.G.); Research Council of Norway FRIPRO project (300740 to V.G.); and Stiftelsen Kristian Gerhard Jebsen (K.G. Jebsen Coeliac Disease Research Centre) (to G.K.S.).

AUTHOR CONTRIBUTIONS

Conceptualization, V.G. and R.A.; methodology, V.G., R.A., P.A.R., M.P., and G.K.S.; investigation, R.A., P.A.R., M.P., J.R.J., Y.S., A.S., and C.R.W.; writing – original draft, V.G. and R.A.; writing – review & editing, P.A.R., M.P., I.S., L.S., E.M., I.H.H., D.T.T.H., F.L.-J., Y.S., and G.K.S.; visualization, R.A. and P.A.R; funding acquisition, V.G.; resources, V.G. and G.K.S.; supervision, V.G. and G.K.S.

DECLARATION OF INTERESTS

E.M. declares holding shares in aiNET GmbH. V.G. declares advisory board positions in aiNET GmbH and Enpicom B.V.

Received: March 12, 2020
Revised: November 29, 2020
Accepted: February 22, 2021
Published: March 16, 2021

SUPPORTING CITATIONS

The following reference appears in the supplemental information: Bateman et al. (2017).

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv, 1603.04467.

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.

Ahmad, S., and Mizuguchi, K. (2011). Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS ONE *6*, e29104.

Akbar, R. (2019). themeakbar (Version 0.1.2). Zenodo. https://doi.org/10.5281/zenodo.3362026.

Akbar, R., and Helms, V. (2018). ALLO: A tool to discriminate and prioritize allosteric pockets. Chem. Biol. Drug Des. *91*, 845–853.

Akbar, R., Jusoh, S.A., Amaro, R.E., and Helms, V. (2017). ENRI: A tool for selecting structure-based virtual screening target conformations. Chem. Biol. Drug Des. *89*, 762–771.

Allcorn, L.C., and Martin, A.C.R. (2002). SACS—self-maintaining database of antibody crystal structure information. Bioinformatics *18*, 175–181.

Amimeur, T., Shaver, J.M., Ketchem, R.R., Taylor, J.A., Clark, R.H., Smith, J., Citters, D.V., Siska, C.C., Smidt, P., Sprague, M., et al. (2020). Designing Feature-Controlled Humanoid Antibody Discovery Libraries Using Generative Adversarial Networks. bioRxiv, 2020.04.12.024844.

Andersen, P.H., Nielsen, M., and Lund, O. (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. *15*, 2558–2567.

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. J. Biol. Chem. *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. Curr. Top. Med. Chem. *18*, 2239–2255.

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. Proc. Natl. Acad. Sci. USA *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.; The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45* (*D1*), 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. Annu. Rev. Immunol. *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. Annu. Rev. Immunol. *37*, 547–570.

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature *566*, 393–397.

Broido, A.D., and Clauset, A. (2019). Scale-free networks are rare. Nat. Commun. *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. J. Immunol. *190*, 2327–2334.

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. Scand. J. Stat. *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. https://rdrr.io/cran/VennDiagram/.

Chen, X., Dougherty, T., Hong, C., Schibler, R., Zhao, Y.C., Sadeghi, R., Matasci, N., Wu, Y.-C., and Kerman, I. (2020). Predicting Antibody Developability from Sequence using Machine Learning. bioRxiv, 2020.06.18.159798.

Chollet, F. (2015). Keras. https://keras.io/.

Chothia, C., and Lesk, A.M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. J. Mol. Biol. *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.

Clauset, A., Shalizi, C.R., and Newman, M.E.J. (2009). Power-Law Distributions in Empirical Data. SIAM Rev. *51*, 661–703.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

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. J. Mol. Biol. *325*, 337–354.

Csepregi, L., Ehling, R.A., Wagner, B., and Reddy, S.T. (2020). Immune Literacy: Reading, Writing, and Editing Adaptive Immunity. iScience *23*, 101519.

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 *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. (2019). Attentive Cross-Modal Paratope Prediction. J. Comput. Biol. *26*, 536–545.

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*, 2278.

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* (*D1*), D427–D432.

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. (Springer New York), pp. 255–264.

Elhanati, Y., Sethna, Z., Marcou, Q., Callan, C.G., Jr., Mora, T., and Walczak, A.M. (2015). Inferring processes underlying B-cell repertoire diversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140243.

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 Res. *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*, bay040.

Gillespie, C.S. (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. J. Stat. Softw. *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.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT Press).

Gowthaman, R., and Pierce, B.G. (2018). TCRmodel: high resolution modeling of T cell receptors from sequence. Nucleic Acids Res. *46* (*W1*), W396–W401.

Greenbaum, J.A., Andersen, P.H., Blythe, M., Bui, H.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. J. Mol. Recognit. *20*, 75–82.

Greiff, V., Menzel, U., Miho, E., Weber, C., Riedel, R., Cook, S., Valai, A., Lopes, T., Radbruch, A., Winkler, T.H., and Reddy, S.T. (2017a). Systems Analysis Reveals High Genetic and Antigen-Driven Predetermination of Antibody Repertoires throughout B Cell Development. Cell Rep. *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. J. Immunol. *199*, 2985–2997.

Greiff, V., Yaari, G., and Cowell, L.G. (2020). Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. Curr. Opin. Syst. Biol. *24*, 109–119.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812.

Hamer, R., Luo, Q., Armitage, J.P., Reinert, G., and Deane, C.M. (2010). i-Patch: interprotein contact prediction using local network information. Proteins *78*, 2781–2797.

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. Mol. Ther. *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.

Hollingsworth, S.A., Lewis, M.C., Berkholz, D.S., Wong, W.-K., and Karplus, P.A. (2012). $(\varphi,\psi)_2$ motifs: a purely conformation-based fine-grained enumeration of protein parts at the two-residue level. J. Mol. Biol. *416*, 78–93.

Hwang, H., Petrey, D., and Honig, B. (2016). A hybrid method for protein-protein interface prediction. Protein Sci. *25*, 159–165.

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. USA *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. Proc. Natl. Acad. Sci. USA *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*, 298.

Jordan, R.A., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2012). Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics *13*, 41.

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*, 8145.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv, 1412.6980.

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. Mol. Immunol. *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 Eng. Des. Sel. *26*, 599–609.

Kunik, V., Peters, B., and Ofran, Y. (2012a). Structural consensus among antibodies defines the antigen binding site. PLoS Comput. Biol. *8*, e1002388.

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–W524.

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 *32*, 2451–2456.

Landsteiner, K. (1936). Serological reactions. 189. The specificity of serological reactions (New York: Dover Publications, Inc.).

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. J. Mol. Biol. *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. Mol. Immunol. *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. Nucleic 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 *34*, 2944–2950.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). Noncovalent Bonds, Fourth Edition (Molecular Cell Biology).

Luong, M.-T., Pham, H., and Manning, C.D. (2015). Effective Approaches to Attention-based Neural Machine Translation. arXiv, 1508.04025.

MacCallum, R.M., Martin, A.C.R., and Thornton, J.M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. J. Mol. Biol. *262*, 732–745.

Mahajan, S., Yan, Z., Jespersen, M.C., Jensen, K.K., Marcatili, P., Nielsen, M., Sette, A., and Peters, B. (2019). Benchmark datasets of immune receptor-epitope structural complexes. BMC Bioinformatics *20*, 490.

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. https://doi.org/10.1101/617860.

McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference, S. van der Walt and J. Millman, eds., pp. 51–56.

Mian, I.S., Bradwell, A.R., and Olson, A.J. (1991). Structure, function and properties of antibody binding sites. J. Mol. Biol. *217*, 133–151.

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.

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 Rep. *25*, 2121–2131.e5.

Norman, R.A., Ambrosetti, F., Bonvin, A.M.J.J., Colwell, L.J., Kelm, S., Kumar, S., and Krawczyk, K. (2020). Computational approaches to therapeutic antibody design: established methods and emerging trends. Brief Bioinform. *21*, 1549–1567.

Northey, T.C., Barešic, A., and Martin, A.C.R. (2018). IntPred: a structure-based predictor of protein-protein interaction sites. Bioinformatics *34*, 223–229.

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 lymphocyte and adjacent healthy tissue. Cancer Res. *79*, 1671–1680.

Padlan, E.A. (1977). Structural basis for the specificity of antibody-antigen reactions and structural mechanisms for the diversification of antigen-binding specificities. Q. Rev. Biophys. *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. USA *111*, E2656–E2665.

Pittala, S., and Bailey-Kellogg, C. (2019). Learning Context-aware Structural Representations to Predict Antigen and Antibody Binding Interfaces. bioRxiv. https://doi.org/10.1101/658054.

Ponomarenko, J.V., and Bourne, P.E. (2007). Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct. Biol. *7*, 64.

Porollo, A., and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins *66*, 630–645.

R Core Team (2018). R: A Language and Environment for Statistical Computing (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. J. Mol. Recognit. *25*, 103–113.

Ralph, D.K., and Matsen, F.A., 4th. (2016). Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. PLoS Comput. Biol. *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. Biochim. Biophys. Acta *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. Mol. Syst. Des. Eng. *4*, 679–688.

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. Proc. Natl. Acad. Sci. USA *116*, 4025–4030.

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. Semin. Cell Dev. Biol. *84*, 30–41.

Rodrigues, J.P.G.L.M., Teixeira, J.M.C., Trellet, M., and Bonvin, A.M.J.J. (2018). pdb-tools: a swiss army knife for molecular structures. F1000Res. *7*, 1961.

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*, 2838.

Sanchez-Trincado, J.L., Gomez-Perosanz, M., and Reche, P.A. (2017). Fundamentals and Methods for T- and B-Cell Epitope Prediction. J. Immunol. Res. *2017*, 2680160.

Schrödinger (2015). The PyMOL Molecular Graphics System, Version 1.8 (Schrödinger).

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.

Sivalingam, G.N., and Shepherd, A.J. (2012). An analysis of B-cell epitope discontinuity. Mol. Immunol. *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 *74*, 497–514.

Stave, J.W., and Lindpaintner, K. (2013). Antibody and antigen contact residues define epitope and paratope size and structure. J. Immunol. *191*, 1428–1435.

Stein, A., Russell, R.B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. Nucleic Acids Res. *33*, D413–D417.

Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature *302*, 575–581.

Townshend, R.J.L., Bedi, R., Suriana, P.A., and Dror, R.O. (2019). End-to-End Learning on 3D Protein Structure for Interface Prediction. arXiv, 1807.01297.

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, F.L., Jr. (1995). Python Tutorial (Centrum voor Wiskunde en Informatica Amsterdam).

Vavrek, M.J. (2011). fossil: palaeoecological and palaeogeographical analysis tools. Palaeontol. Electronica *14*, 16.

Wang, M., Zhu, D., Zhu, J., Nussinov, R., and Ma, B. (2018). Local and global anatomy of antibody-protein antigen recognition. J. Mol. Recognit. *31*, e2693.

Weitzner, B.D., and Gray, J.J. (2017). Accurate Structure Prediction of CDR H3 Loops Enabled by a Novel Structure-Based C-Terminal Constraint. J. Immunol. *198*, 505–515.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).

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. J. Exp. Med. *132*, 211–250.

Xu, J.L., and Davis, M.M. (2000). Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity *13*, 37–45.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PDB (accession date: Jul 2019)	Berman et al., 2000	https://www.rcsb.org/
3did (accession date: Jul 2019)	Stein et al., 2005	https://3did.irbbarcelona.org/
IMGT (accession date: Feb 2019)	Lefranc et al., 1999	http://www.imgt.org/
Antibody Database (AbDb) (accession date: Jul 2019)	Ferdous and Martin, 2018	http://www.abybank.org/abdb/
bNAber (accession date: Jul 2019)	Eroshkin et al., 2014	http://bigd.big.ac.cn/databasecommons/database/id/301
Preprocessed datasets	This manuscript	https://github.com/GreiffLab/manuscript_ab_epitope_interaction
Raw ML files	This manuscript	[https://archive.sigma2.no]: https://doi.org/10.11582/2020.00060
Software and algorithms		
Python 3.6.4	Van Rossum and Drake, 1995	https://www.python.org/
R 3.5.2	R Core Team, 2018	https://www.r-project.org/
ggplot2 3.1.0	Wickham, 2016	https://ggplot2.tidyverse.org/
VennDiagram 6.20	Chen, 2018	https://cran.r-project.org/web/packages/VennDiagram/index.html
TensorFlow 1.13.1	Abadi et al., 2015	https://www.tensorflow.org/versions
Keras 2.2.4-tf	Chollet, 2015	https://keras.io/
pandas 0.25.1	McKinney, 2010	https://pandas.pydata.org/
Biopython 1.74	Cock et al., 2009	https://biopython.org/
pdb-tools 2.0.0	Rodrigues et al., 2018	http://www.bonvinlab.org/pdb-tools/
poweRlaw R 0.70.2	Gillespie, 2015	https://cran.r-project.org/web/packages/poweRlaw/index.html
Cytoscape 3.7.1	Shannon et al., 2003	https://cytoscape.org/
Circlize 0.4.8	Gu et al., 2014	https://cran.r-project.org/web/packages/circlize/index.html
pheatmap 1.0.12	Kolde, 2019	https://www.rdocumentation.org/packages/pheatmap/versions/1.0.12/topics/pheatmap-package
Fossil 0.3.7	Vavrek, 2011	https://cran.r-project.org/web/packages/fossil/index.html
Pymol 2.1.0	Schrödinger, 2015	https://pymol.org/2/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Victor Greiff (victor.greiff@medisin.uio.no).

Materials availability

This study did not generate new reagents.

Data and code availability

Preprocessed datasets, code, and results figures are available at: https://github.com/GreiffLab/manuscript_ab_epitope_interaction. The accession number for the unprocessed deep learning models, checkpoints, and output files reported in this paper is [https://archive.sigma2.no]: https://doi.org/10.11582/2020.00060.

METHOD DETAILS

A dataset of non-redundant and diverse 3D antibody-antigen complexes

A dataset of 866 antibody-antigen complexes in the format of Protein Data Bank (PDB) (Figure 1A) was obtained from the Antibody Database (AbDb) (Berman et al., 2000; Ferdous and Martin, 2018) [download date 27 July 2019]. AbDb routinely crawls PDB to find existing antibody-antigen structures and preprocesses them by (i) identifying the antibody (VH-VL, variable [V] heavy [H] and light [L] chain domains) and the corresponding 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. To obtain non-redundant structures, AbDb performs pairwise comparisons across the structures (both heavy and light chains). Specifically, (i) PDB records with the type ATOM were used to extract the amino acid sequence of an antibody, (ii) each antibody pair is compared with respect to residue position and amino acid, for example, if the residue position 42 is present in both antibodies and the amino acid is different, then the two antibodies were regarded as non-redundant, and (iii) if there are missing residues in one antibody and not the other, these positions are ignored (we refer to the section ‘redundancy processing’ in the original paper for more details: Ferdous and Martin, 2018). Structures comprising different amino acid residues in the same position are considered non-redundant. From the initial dataset of 866 antibody-antigen complexes, we removed atoms labeled with PDB record type HETATM (non-protein atoms serving as co-factors) and structures with a resolution larger than 4.0Å (*44*). The final curated dataset comprises 825 antibody-antigen (protein antigen only) complexes with a median resolution of 2.5A (Figure 1A). 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 (Figure S1D) and we found large overlaps in paratope motif spaces used by mice and humans (Figures S1H, S1I, and S4C). Additionally, it was recently reported that neither CDR/FR length nor the distribution of interface residues in human and murine antibodies differs substantially (Collis et al., 2003; Henry and MacKenzie, 2018; Wang et al., 2018).

Annotations for 113 bNAbs were obtained from the database bNAber (Eroshkin et al., 2014). 70 of these antibodies were represented as 24 non-redundant (see above and the section ‘redundancy processing’ in the original AbDb paper for more details: Ferdous and Martin, 2018) complexes in AbDb and were included herein (Figures S4J and S4K). The remaining 43 were excluded (38 without antigens and 5 because of unavailable structures).

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 FR: FR1, FR2, FR3, and FR4; and complementary determining region (CDR): CDR1, CDR2, and CDR3. In the Kabat scheme, 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 previously determined to be suitable for structural and antibody engineering (Dondelinger et al., 2018). It is also the most recent of the presently available numbering schemes. 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 shown in Table S2.

Identification 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 4–6A 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 interactions range from 3–4A; (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 < 5A (heavy atoms) is most sensitive to changes in residue interactions and variables such as force fields (Salamanca Viloria et al., 2017). Therefore, we defined interacting paratope-epitope residues by a distance cutoff of < 5A between heavy atoms. In other words, amino acid residues are considered to be interacting if they have heavy atoms with a distance < 5A from each other (Figure 1B). We used the function NeighborSearch of the module Bio.PDB in Biopython (Cock et al., 2009) to identify neighboring amino acid residues within the distance cutoff. 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 < 4A, < 5A and < 6A in Figures S3A and S3D. We confirmed that the overall trends in per-region residue distribution, overlap between paratope and epitope interaction motifs (Figures S3A and S3B), and (dis)continuity (Figures S3C and S3D) 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 structural 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 colored in salmon in Figure 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 Figure 1D). (iv) A gap is defined as the number of non-interacting residues separating two paratope or epitope residues (Figures 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, Figure 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 (1). The number of interaction motifs in antibodies, motif continuity/sharing, correlation between paratope-epitope motifs lengths is shown in Figures S4L–S4Q.

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 inset in Figure S6A for illustration), in a similar fashion to AngleBetweenHelices, a Pymol module for calculating the angle between helices (Schrödinger, 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 were rendered and visualized in Pymol 2.1.0 (Schrödinger, 2015).

Diversity analysis of interaction motifs

To estimate the potential (observed + unobserved) 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) (Vavrek, 2011).

Paratope-epitope amino acid contact map

Paratope(P)-epitope(E) amino acid contact maps were obtained by computing the log odds ratio, $L_x(P_i,E_j) = \log_2\left(P(P_i,E_j)/P(P_i)P(E_j)\right)$, 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 region (FR/CDR in antibody-antigen complexes). Analogously, protein-protein amino acid contact maps were computed for inter- and intradomain in non-immune protein-protein complexes (PPI).

Construction of bipartite paratope-epitope and PPI reactivity networks at motif and sequence level

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). Reactivity networks for sequence and aggregate encoding, see machine learning use cases (encoding) below, as well as PPI reactivity networks, were constructed as above described and are shown in Figures S5 and S7, respectively.

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 (bidirectional sliding window) of the ensemble of paratope/epitope sequences mapping to the respective motif (non-interacting residues were not taken into account). For each motif, these sequential dependencies were visualized as Chord diagrams where the 20 amino acids form the segments in a track (the outermost ring) and the links indicate the frequency with which a 2-mer sequential dependency occurred (sequential dependency). Chord diagrams were constructed using Circlize 0.4.8 (Gu et al., 2014). 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 interaction motifs

A dataset of protein-protein interactions (PPI) was sourced from 3did, a catalog of three-dimensional structure domain-based interactions (Stein et al., 2005). 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). As of 2 July

2019, 3did comprised a total of 18,599,078 contact residue pairs (100,888 protein structures), which is three orders of magnitude larger than the number of antibody-antigen contact residues (18,630 residue pairs, Figure 1C). Protein-protein interaction motifs were constructed for each domain pair analogously to paratope-epitope interaction motifs (see the previous section). Motifs with gap lengths larger than seven were excluded from the analysis (to match the largest gap size found in paratopes, Figure 1) as well as complexes larger than 300 residues long. The final non-immune PPI dataset comprises 9621 interdomain and 1043 intradomain complexes for a total of 299,141 contact residues (Figure S5).

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/alleles, respectively (accession date: Feb 2019). We translated the nucleotide sequences of V and J genes according to their ORFs (open reading frame). As D genes can be truncated during the 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 the sequence corresponds to the 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 as it was shown previously that three amino acids (up to 9 nt) correspond to the average lengths of exonucleolytic removals in V and J genes (Ralph and Matsen 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. Figure S2E shows inferred SHMs localize around CDR1s and CDR2s and thus partially correlate with the paratopes positions centered in all three CDRs. We found only few SHMs in the CDR3s (Figure S2E). We caution that this may be a reflection of the limitation of our SHM quantification approach and not necessarily a biological feature of the immunoglobulin sequences here studied.

Ramachandran plot analysis

Ramachandran angles (Phi-Psi pairs) were extracted from PDB files using the package PDB in Biopython 1.74 (Cock et al., 2009). The package pdb-tools 2.0.0 was used to preprocess PDB files and extract the chains/regions of interest (Rodrigues et al., 2018). We examined six different groups: (i) residues in the CDR regions of the heavy or light chains of antibody structures (CDR); (ii) residues in the framework regions of heavy and light chains of antibody structures (FR); (iii) residues binding to the antigen (paratope), from the FR and CDR regions i.e., only the ‘X’ in the motifs (AbDb interacting residues); (iv) binding residues from the PPI dataset in inter-and intra-chain interactions (PPI interacting residues); (v) residues that belong to a motif (including gaps) in AbDb antibody structures, for instance, X-X–X leads to 6 angles (AbDb motifs), and (vi) residues that belong to a motif in intra- or inter-chain interactions in the PPI dataset (PPI motifs). Finally, following Hollingsworth and colleagues (Hollingsworth et al., 2012), we classified the Phi-Psi pairs into groups of secondary structure types (also known as Ramachandran regions).

Machine-learning prediction of paratope-epitope and PPI at interaction motif, sequence and aggregate level

To quantify the extent to which paratope-epitope/non-immune protein-protein interaction is learnable with the available dataset, we leveraged both deep and shallow learning approaches using several encodings of the input (see below). The shallow learning approach directly predicts the cognate (epi/para)-tope (or PPI binding partner) as an atomic unit. In contrast, the deep learning method generates the cognate (epi/para)-tope (or PPI binding partner) character by character (more details below). It thus represents a generative approach to prediction, although in a different sense than the typical meaning of generative machine learning (learning a joint distribution of independent and dependent variables) (Goodfellow et al., 2016).

Use cases (encoding):

Four levels of encoding in both directions, namely paratope to epitope and epitope to paratope (or PPI binding partner to PPI binding partner), were used. (i) Structural motif level: a paratope structural motif XXX interacting with an epitope motif X2X yields an input-output pair XXX–X2X. (ii) Position-augmented structural motif level: a paratope structural motif XXX interacting with an epitope motif X2X yields an input-output pair $X_1X_2X_3$–$X_12_2X_3$, the positions index each character in the sequence consecutively. (iii) Sequence level: a paratope sequence NMA interacting with an epitope sequence RA yields an input-output pair NMA–RA. (iv) Finally, an aggregate representation that simultaneously takes into account amino acid information and motif by replacing the abstraction character ‘X’ with the corresponding residue: a paratope-epitope interaction defined by the paratope sequence GR and motif X1X together with the epitope sequence LLW and motif XX1X yields an input-output pair G-R–LL-W. The antibody-antigen (PPI) datasets comprise a total of 5,327 (25,921) input-output pairs.

Deep learning

We leveraged a model based on Neural Machine Translation (Luong et al., 2015) to learn an epitope from (to) a paratope at motif, sequence and aggregate levels (same for PPI but instead of paratope/epitope, respective PPI binding partners). Specifically, pairs of input-output sequences were translated via a combination of two components: encoder and decoder with gated recurrent units (GRU, see Figures 5, bottom panel, and S8). During the decoding phase via an attention layer, a context vector is derived to capture relevant input-side information necessary for the prediction of an output. Utilizing the context vector, the decoder part of our deep model generates each paratope or epitope motif/sequence character by character. For the translation task, 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 simply as X-X. The dataset was split into 80% training and 20% test set. The numerical representation of the input pairs was learned by vector embedding. Pairwise parameter combination: (i) embedding dimension (1, 2^1, 2^2,..., 2^{10}) and (ii) number of units (hidden dimension) (1, 2^1, 2^2,..., 2^{10}) was used to parameterize the models. Here, the embedding dimension is the length of the vector representing the input whereas the number of units is the number of cells in the GRU otherwise known as the length of the hidden dimension. The training procedure was carried out for 20 epochs with Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba, 2014) and was replicated ten times. Each replicate comprises 121 models for a total of 1,210 models (121 × 10, see workflow). The model from the last epoch of each replicate was used to generate predictions on the test dataset.

Shallow learning

The shallow model takes into account the conditional probability of the output with respect to the input and a prior corresponding to the output with the highest marginal probability (the most frequent class).

Evaluation

Discrepancy (error) between predictions and the true motifs (sequences) was determined by the normalized Levenshtein distance, $\mathrm{LD}_{\text{prediction vs truth}} \big/ (\max(\text{length}(\text{prediction}), \text{length}(\text{truth})))$, between the predicted motifs (sequences) and true motifs (sequences). Baseline prediction accuracies were calculated based on label-shuffled data where antibody and antigen-binding (or PPI) partners were randomly shuffled. To ensure robustness when evaluating the deep models, instead of showing the error obtained from the "best model" in each replicate, we showed the mean of median error across all replicates and pairwise parameter combinations. Ratios of training and test datasets, as well as error computation for the shallow model, were identical to the above-described computation for deep models except for input motifs that were not present in the training dataset where the error was set to 1 (maximum error).

Deep learning models were constructed in TensorFlow 1.13.1 (Abadi et al., 2015) with Keras 2.2.4-tf (Chollet, 2015) in Python 3.6.4 (Van Rossum and Drake, 1995), while the statistical (shallow) model was constructed using pandas 0.25.1 (McKinney, 2010). Computations for deep models were performed on the high-performance computing cluster Fram (Norwegian e-infrastructure for Research and Education https://sigma2.no/fram).

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 using Adobe Illustrator CC 2019.

QUANTIFICATION AND STATISTICAL ANALYSIS

All tests for statistical significance were performed using R 3.5.2 (R Core Team, 2018). Statistical difference between distributions was computed using the Kolmogorov-Smirnov (KS) test. Details are described in the Figure Legends and method details.