MABS
2025, VOL. 17, NO. 1, 2507950
https://doi.org/10.1080/19420862.2025.2507950

Taylor & Francis
Taylor & Francis Group

REPORT

Data-driven analyses of human antibody variable domain germlines: pairings, sequences and structural features

Clarissa A. Seidler[a], Vera A. Spanke[a], Jakob Gamper[a]*, Alexander Bujotzek[b], Guy Georges[b], and Klaus R. Liedl[a]

[a]Department of General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innsbruck, Austria; [b]Roche Pharma Research and Early Development, Large Molecule Research, Roche Innovation Center Munich, Penzberg, Germany

ABSTRACT

The Observed Antibody Space provides the most abundant collection of annotated paired antibody variable domain sequences, thus offering a unique platform for the systematic investigation of the factors governing the pairing of antibody heavy and light chains. By examining a range of characteristics, including amino acid conservation, structural features, charge distribution, and interface residue identity, we challenge the prevailing assumption that pairing is random. Our findings indicate that specific physicochemical properties of single amino acid residues may influence the compatibility and affinity of heavy and light chain combinations. Further structural analyses based on antibody Fv fragments deposited in the Protein Data Bank (PDB) provide insights into the underlying structural features driving these pairing preferences, including a novel definition for the residues constituting the V_H-V_L interface, based on a collection of over 3500 structures. These results have significant implications for understanding antibody assembly and may guide the rational design of therapeutic antibodies with desired properties. Moreover, we provide a complete description and reference characterizing the various human germlines.

ARTICLE HISTORY
Received 25 February 2025
Revised 29 April 2025
Accepted 14 May 2025

KEYWORDS
Antibody pairing; germline V-gene; V_H-V_L interface; antibody assembly; database analysis

Introduction

Antibodies are indispensable elements of the human immune system that play a crucial role in defending against a vast array of pathogens. Their extraordinary capacity to recognize and neutralize an almost limitless variety of antigens has long been a subject of scientific fascination.[1,2] By the middle of the 20th century, researchers had elucidated the mechanisms through which this diversity is generated, including gene rearrangement, somatic hypermutation, and class switch recombination, which occur in diverse stages of B cell development.[3–5] All mammalian antibodies, as well as the majority of antibodies observed in other species, are constructed from two heavy chains and two light chains, which form Y-shaped proteins.[6] The C-terminal ends of the heavy chains form the lower part of the antibody, which can be bound to surfaces *via* Fc receptors, or remain free in solution.[7] The domains that constitute the N-terminal end of these polypeptide chains are referred to as variable regions (F_v). These regions are composed of both a heavy and a light chain variable domain, named V_H and V_L, respectively, and are responsible for recognizing and binding antigens. Together with the subsequent domain of each chain, C_H1 and C_L respectively, the "arms" of the antibody are referred to as antigen binding fragments (Fab). The proteins are assembled from genes encoding the antibody sequences through a complex process involving gene rearrangement. Gene rearrangement entails the random recombination of DNA segments, which are subsequently transcribed and translated into polypeptide chains. In humans, the light chains are encoded on two distinct chromosomes: kappa light chains on chromosome 2 at the *locus* 2p11.2 and lambda light chains on chromosome 22 at the *locus* 22q11.2. The heavy chains, on the other hand, are encoded on chromosome 14 at the position 14q32.33. Each gene *locus* contains multiple variable (V), joining (J), and constant (C) gene segments. Additionally, diversity (D) genes serve to enhance the diversity of the complementarity-determining region (CDR) H3 loop, which is situated at the center of the heavy chain binding region.[8] A schematic representation of the different genes, and how they are arranged on the different *loci* is shown in Figure 1a. In the case of heavy chains, the V, D, and J segments undergo recombination to form the variable region. In the case of light chains, recombination occurs solely between the V and J segments. This process results in the generation of a vast repertoire of antibody variable regions. The different segments are theorized to encode a specific sequence of amino acids, which are not perfectly conserved due to somatic hypermutations, insertions, and deletions.[9–12] Nevertheless, the human antibody repertoire is based on a limited number of so-called germline genes, which assemble in order to encode the entire antibody sequence repertoire. The term 'germline' is used to describe the status of a gene, indicating that it has not yet undergone rearrangement. A total of 56 germline V-genes are responsible for encoding the heavy chains, according to the International Immunogenetics Information System (IMGT), while 45

CONTACT Klaus R. Liedl Klaus.Liedl@uibk.ac.at Department of General, Inorganic and Theoretical Chemistry, University of Innsbruck, Innsbruck, Austria
*Current address: Qubit Pharmaceuticals, 29 rue du Faubourg Saint Jacques, 75014 Paris, France.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/19420862.2025.2507950

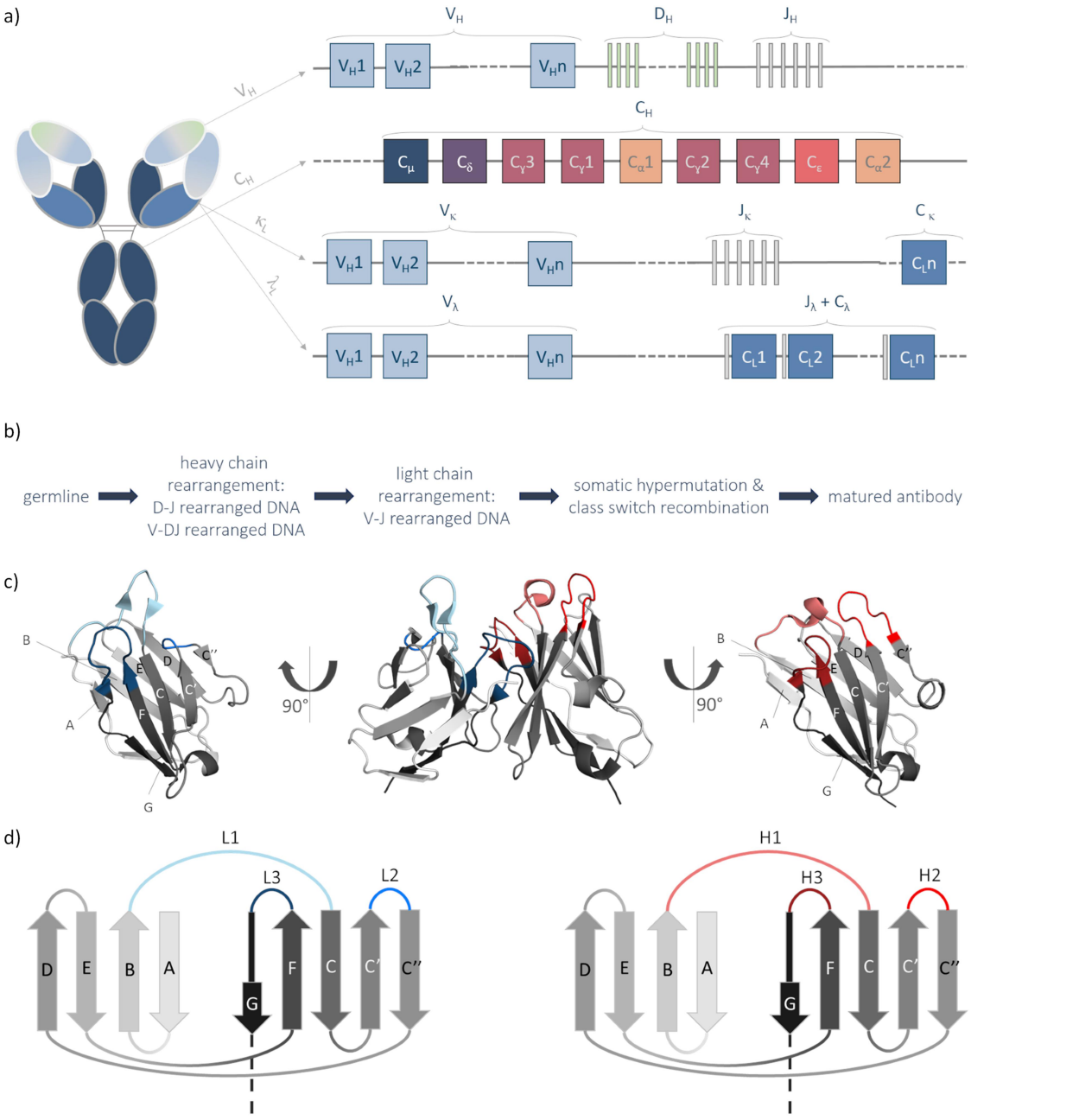

Figure 1. Overview of antibody germline genes and schematic/structural overview of the resulting antibody Fv domain. Subfigure a) illustrates an exemplary schematic antibody, accompanied by the respective germline genes which encode the different domains. The variable region of the heavy chain, encoded by V, D and J segments, is displayed in the first row. The constant domains are contingent upon the specific type of C genes, which ultimately determine the isotype. With regard to the light chains in humans, two distinct possibilities exist: the kappa light chains, which are encoded by a combination of V and J genes, as well as a single C gene; and the lambda light chains, which are encoded by V genes and a combination of J and C genes. Subfigure b) provides a schematic illustration of the process by which the germline genes are converted into the final mature antibody. This process involves the rearrangement of the heavy chain, including the D-J and subsequent V-DJ rearrangements, which occur prior to the rearrangement of the light chain, V-J. Subsequently, the process entails somatic hypermutations and class switch recombination. The mature DNA sequence is ultimately converted into the polypeptide sequence of the mature antibody. Subfigure c) illustrates an antibody variable domain, with color-coded CDRs. The light chain loops are depicted in blue tonalities, while the heavy chain loops are shown in shades of red. The domains are displayed on the left and right sides with the interface rotated 90° to the front (to the left and right, respectively). The sheets are labelled. Subfigure d) depicts a schematic representation of the antibody Fv domain Ig-fold and the labelling of the beta sheets and CDR loops. The color coding of the schematic structure corresponds to that of the structure in subfigure c).

V-genes encode kappa light chains and 39 V-genes encode lambda light chains.[13] These genes can be further classified into subgroups and groups, or further subdivided into alleles, which are polymorphic variants of a single gene. For the purposes of this study, only a distinction up to the gene variants has been considered. These genes can combine to produce a vast repertoire of antibodies *via* the aforementioned V(D)-J rearrangement. Following recombination and assembly of light and heavy chains, somatic hypermutation introduces mutations into the antibody genes after an antigen is encountered. This allows for the fine-tuning of antigen specificity and the selection of antibodies with higher affinity for their target antigens, a process known as affinity maturation.[12] Class switch recombination (CSR) allows B cells to switch the

antibody isotype by altering the constant region and maintaining the specificity of the variable domains. While CSR significantly affects the effector functions of antibodies, it does not directly affect the pairing of heavy and light chains within the Fv region, which is the primary focus of this study. This process from the germline gene to the matured antibody is briefly described in Figure 1b).

The prevailing hypothesis posits that all heavy and light chains randomly combine to form functional antibodies.[6,14,15] However, our study challenges this prevailing view, suggesting that specific properties of the individual chains may influence pairing preferences. This nonrandom pairing may be driven by factors such as amino acid sequence, structural features, or complementary electrostatic interactions.[16–18]

To further investigate this hypothesis, we analyzed the largest available dataset of paired antibody sequences, namely the Observed Antibody Space (OAS) Database. This is a compilation of nearly two million cleaned, annotated, and translated sequences of human immunoglobulin variable domains from 10 different studies. Prior to the introduction of the 10×Genomics sequencing technology, it was not feasible to screen a substantial corpus of paired antibody sequences due to the limitations of maintaining the pairing after sequencing the blood samples. This challenge has been addressed by the aforementioned technology, which has been used to achieve the large amount of data that has been collected and stored in the OAS database.[19,20]

The availability of this comprehensive dataset permits a systematic investigation and enables the identification of statistically significant properties and indicators associated with pairing preferences. The objective is to ascertain whether the pairing is random and to eventually identify the factors that contribute to nonrandom pairing, with the aim of gaining insight into their implications for antibody formation and engineering. By examining the conservation of amino acid residues, structural features, and charge distributions, the underlying mechanisms of pairing preferences are elucidated.

The analyses are conducted on the variable domains of the antibody's light and heavy chains, as illustrated in Figure 1c. In sub Figure 1d, schematic representation of these domains is presented, with the beta-sheets labeled. Each of the antibody chains contains three CDRs, which are hypervariable regions that play a pivotal role in antigen recognition.[21] The framework regions (FRs) represent the fragments that combine the highly variable loop regions. A human variable antibody domain comprises four framework regions, which can be further subdivided and structurally classified into the beta-sheets A, B, C, C', C'', D, E, F, and G for each chain.[22] Accordingly, sheets A and B are assigned to FR1, C and C' to FR2, C'', D, E and F to FR3, and sheet G to FR4.[23] Despite the prevailing view that the CDR loops constitute a significant portion of the interface and play a crucial role in heavy-light chain interactions, our hypothesis is that framework regions may also contribute to this process. The objective of this study is to identify patterns indicating nonrandom pairing in both regions, namely CDRs and FRs, through the analysis of paired sequences and structures.

The antibody variable heavy and light chains are assembled and remain in contact via hydrophobic interactions.[24] As previously described, the single domains are composed of several β-sheets assembled in a two-layer sandwich structure, known as the immunoglobulin fold. The amino acids that constitute the beta sheets are oriented alternately, with one side pointing inward toward the interface and the other oriented toward the interior of the domain.[6,22]

While efforts have been made, the precise definition of residues mediating interface formation remains elusive.[25–27] Potential candidates include those amino acids, which directly contact the counterpart domain. In order to characterize these interface residues, the three-dimensional structures of the antibodies are required. For this purpose, an additional database, the Structural Antibody Database (SAbDab),[28] is employed to characterize the Fv regions structurally. This database contains all antibody structures deposited in the Protein Databank (PDB), annotated and sorted in a manner that allows certain properties to be pre-set.

The identification of residues that affect binding, stability, degradation, pairing, and aggregation is crucial for the optimization of antibody development. A comprehensive understanding of the key positions for point mutations is essential for the engineering of therapeutic antibodies with the desired properties.[6] This study contributes to advancing our knowledge in these areas.

Results

In accordance with the methodology outlined in the "Materials and Methods" section, a total of 70 light chain V-genes (kappa and lambda combined) and 52 heavy chain V-genes were obtained, resulting in a total of 3,148 possible pair combinations. Each combination that indeed occurs in the OAS, is represented by a single dot in Figure 2. In this figure, both light and heavy chains are represented in descending order of occurrence, from left to right and bottom to top. This indicates that the pairs observed in the lower left angle of the graphical representation are also paired between chains that occur most frequently in the dataset. The frequency of occurrence of each pair within the entire dataset is represented by the shading of the color in the plot. It is evident that the pair counts exhibit a tendency to decline with the reduction in occurrence of the individual chains. However, there are exceptions that do not adhere to this pattern.

Figure 2b illustrates the observed pair counts in relation to the expected pairing occurrences. In the heatmap, a blue dot represents pairs that occur less frequently than expected, whereas red points indicate a count that exceeds the theoretically calculated pair count. As our calculations are constrained to a maximum of −100% on the negative half of the axis, but values exceeding + 100% are permitted, we elected to represent a pairing in excess of + 100% above the anticipated count with a black dot. In the event that the anticipated and actual counts were to align, the color of the plot would be white for all data points. In the present case, substantial deviations from expectations indicate a nonrandom pairing.

The statistical significance of deviations from random pairing was rigorously assessed using a combination of

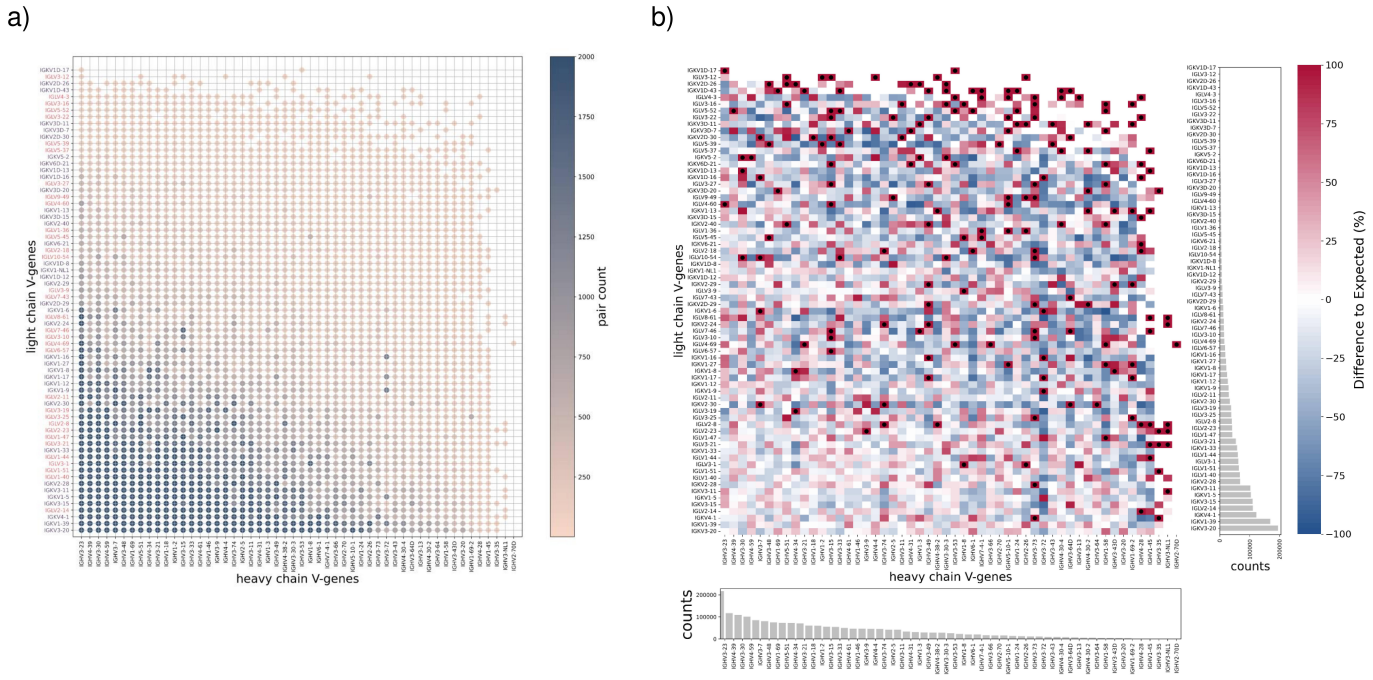

Figure 2. Preferences for pairing. In the scatter plot depicted in subfigure a), the pair count of each heavy chain with each light chain germline V-gene is represented as a dot, colored according to the color bar. The germlines are sorted according to their occurrences in the paired OAS, bottom to top and left to right in descending order, thereby indicating that frequent germline genes tend to pair more often. Subfigure b) illustrates the discrepancy between the actual and theoretical pair counts. In this heatmap, the pair counts derived from the paired OAS are contrasted with the theoretically calculated counts. The genes are ordered according to their occurrence in the database, as illustrated in the subplots beneath and beside the heatmap. A pair count exceeding expectations is indicated by red shading, while a pair count below expectations is indicated by blue tonality. Additionally, black dots within red squares indicate a deviation exceeding 100% from the expected count. This visualization suggests that more statistics may indicate more accurate pairing frequency predictions.

approaches, the details of which are further elaborated in the Supplementary Information. In order to ensure robustness against potential biases from varying pair counts, Monte Carlo simulations were conducted. The simulations consistently yielded extremely low empirical p-values ($p < 0.001$), thereby reinforcing the conclusion that the observed pairing patterns are highly unlikely to occur by chance. While complementary analyses using Chi-squared tests and generalized power divergence tests likewise indicated robust nonrandom pairing ($p < 0.001$ in almost all cases), the Monte Carlo simulations furnish especially persuasive evidence by demonstrating the robustness of this finding even when accounting for imbalances in the data. Additionally, Kolmogorov–Smirnov (KS) tests on the complete dataset revealed a statistically significant discrepancy between the observed and expected distributions of pairing frequencies ($p = 0.011$). In order to maintain the reliability of the statistical analyses, the focus was placed on V_H-V_L pairs with observed and expected counts exceeding 10. The consistent and significant results, especially from the Monte Carlo simulations, provide strong evidence for systematic and nonrandom V_H-V_L pairing in human antibodies.[29–31]

The consistency of VH-VL pairing preferences across different studies was evaluated by examining the percentage difference from expected pairing frequencies for the top 10 most abundant pairs (Figure 3). The left panel of Figure 3 illustrates the distribution of these percentage differences for each study, colored according to the study of origin. The crossed markers represent the overall deviation across all studies. As demonstrated in this panel, while the general trend (indicated by the crossed markers) reveals a consistent direction of preference (either over- or under-representation) for the majority of the top pairs, the degree of this preference varies across the individual studies.

The right panel of Figure 3 further accentuates this variability by coloring the data points based on whether the percentage difference is above (red) or below (blue) zero. This visualization provides evidence that for the majority of the top VH-VL pairs, the direction of deviation from random pairing is largely conserved across the datasets. However, the distribution of points for each pair indicates that the magnitude of this preference is not uniform, suggesting some level of inter-individual or inter-study variation in the strength of these pairing biases. For instance, while a particular VH-VL pair might be consistently over-represented across studies, the extent of this over-representation can range from a modest increase to a substantial one. This analysis underscores the presence of consistent directional preferences in VH-VL pairing, while also acknowledging a degree of variability in the strength of these preferences across different antibody repertoires.

The strongest and most long-ranging interactions within biomolecules are of electrostatic nature. One of our theories regarding the pairing preferences of heavy and light chains is thus based on the occurrence of charged residues in key positions within the domains. In order to gain a deeper understanding of these interactions and to verify the aforementioned assumption, the charges of the two domains were calculated. In the initial stage of the analysis, the entire polypeptide chain was considered, with the charges of the individual sequences

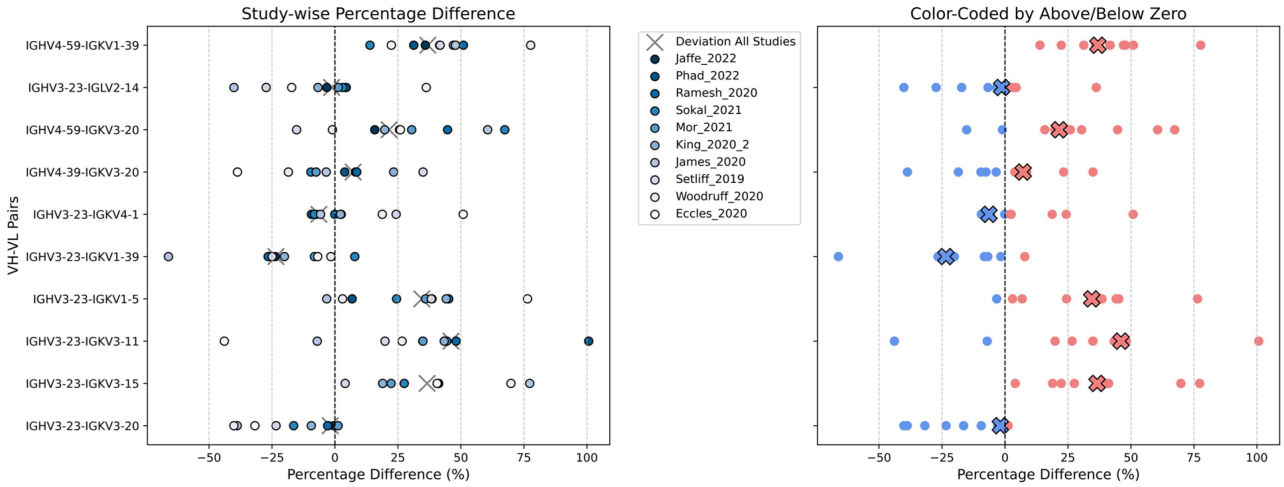

Figure 3. Percentage difference from expected pairing frequencies for the overall top 10 most abundant VH-VL pairs. The left panel shows study-specific deviations, while the right panel highlights deviations from expected above or below zero. The studies are ordered and color coded according to their number of sequences (darker colors represent larger studies). Crossed markers indicate the overall deviation across all studies.

and their resulting pairs being compared. We observed that there was a pronounced tendency for the light-chain charge pair + 1 to form together with the heavy chain charge + 2, resulting in antibody variable domains with a total charge of + 3. There are alternative possibilities for charge formation, but it is evident that the most prevalent number of antibody variable pairs are positively charged, as shown in Figure 4. The statistical relevance of the preferred pairing of certain charges has been demonstrated by statistical tests shown in the supplementary information file.

Subsequently, an in-depth examination of the charge numbers was conducted on a fragment basis. The objective of this investigation was to identify fragments of opposing charges that might exhibit an attractive force. Consequently, the charges on each fragment were calculated individually, and the residues in close proximity to each other were identified. As illustrated in Figure 4, the sole distributions exhibiting a maximum on the negative side are sheets A, E and F for the light chains. For the heavy chains, sheet F was the only one that shifted toward negative values. Within the fragments that shifted to either the positive or negative side, fragments A and E are situated far away from the core residues that interact within the two chains. This observation is supported by the structural representation depicted in Figure 1c and the schematic view of the beta-sheets presented in Figure 1d. It is important to note that both sheets are part of the outer layer of the domain. Conversely, sheet F is positioned on the inner side of the beta sheet layer, comprising residues that interact with the second domain. These residues will be examined in greater detail in Figure 7. The CDR loops are all on average neutral.

Furthermore, when examining the paired charges of the individual germlines, it becomes evident that the majority of pairings results in the formation of slightly positive antibodies. These findings are illustrated in Figure 4b. It is notable that some outliers are present, but most of these can be attributed to the less frequent germlines.

In order to gain insight into the primary distinctions between the various germlines, we conducted a comprehensive sequence comparison and utilized color coding to highlight the variations within a given germline. This is illustrated in Figure 5 for each of the three chain types separately. For all three subplots, the germline V-genes are ordered according to their occurrence in the OAS, with the most rarely occurring germlines displayed at the top and the most frequently occurring ones at the bottom. It is evident that the largest differences within a germline are observed in the regions proximal to the CDR loops and within the CDR loops themselves. Furthermore, the CDR loop regions exhibit the greatest degree of variation in length, particularly the CDR H3 loop. This is why the sequence alignment displays a considerable number of gaps in the central positions of the CDR loops. Nevertheless, we sought to incorporate the loop regions into our sequence alignment. Moreover, the sequences of FR4 are also presented, although these are independent of the V genes and are derived only from a very limited number of J genes, namely 6 functional genes for the heavy chains and 9–10 functional J genes for the κ and λ light chains together.[13] However, a high degree of identity was observed, particularly in this region. This is due to the fact that the J gene primarily influences the final residues of the CDR 3 loops, while the latter region is highly conserved. Consequently, the J genes are also responsible for determining the length of the CDR 3 loops and the final residues of these loops. Sequence logos are presented below the germline sequences, which combine all germlines and, thus, all sequences of the respective chain type from the paired OAS.

To gain further insight into the residues that facilitate domain pairing we calculated a median distance matrix derived from over 3500 experimentally determined antibody Fv structures on a per-residue level, which is represented in Figure 6. The interpolation was performed on the residue-level distance matrices, ensuring a smooth representation of residue–residue distances and providing a clearer perspective on the spatial relationships that contribute to domain interactions. This analysis required obtaining the relevant structural data from the SAbDab database. Figure 6 reveals that the primary points of contact within heavy and light chain are

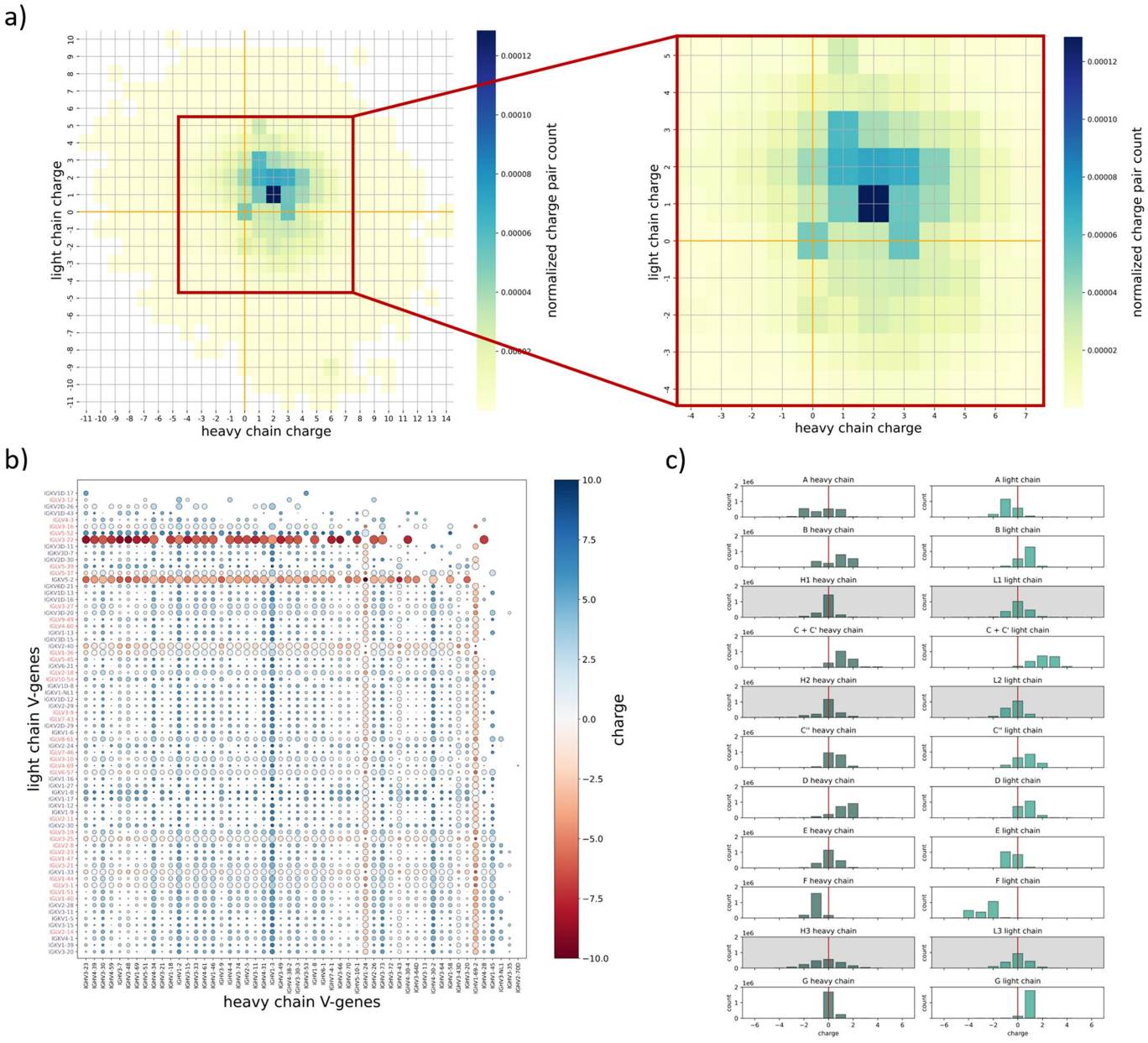

Figure 4. Preferred charge combinations, total charges and fragment-wise charge distributions. The preferred charge combinations occurring in the paired OAS are indicated by the color-coded heatmap in subfigure a). The most prevalent charge observed for the heavy chain was +2, while the most common charge for the light chain was +1. This leads to the conclusion that the most frequently occurring antibody variable domain charge is +3 in total. In subfigure b), the results are broken down on individual germline genes. The data indicate that the majority of pairs result in a slightly positive charge. The size of the dots in this plot provides an additional indication of the absolute difference between the heavy and light chain charges. The larger the dot, the greater the difference between the charges of the two chains. Subfigure c) shows the charge distributions of the heavy and light chains fragment-wise. On average, the majority of the sheets are positive, with sheet A, E and F of the light chains and sheet F of the heavy chains, being the only exceptions exhibiting a distribution shifted towards the negative side. CDR loops are highlighted with gray backgrounds.

located in four regions: the C and C' sheets and the loop connecting these strands, as well as the areas surrounding both CDR 3 loops. These positions facilitate contacts between the two chains. The interacting spots are clearly depicted via green and blue spots in Figure 6. As deduced from Figure 4c, the FR2 of one chain shows the closest spatial proximity (as seen in Figure 6) to regions on the opposing chain that correspond to the strands A and strand F (the portion right before the CDR 3 loop), as well as to the FR2 of the other chain and slightly to the tail region (beginning of the FR1). While Figure 4 c indicates that FR2 of one chain is slightly positively charged, the fragments on the opposing chain in contact with it have also been shown (via inference from Figure 4) to be slightly positively charged. We believe that, despite these like charges at the interface, other factors contribute to the stability of the pairing. The observed pattern of closest inter-chain contacts, where the FR2 region of one chain interacts most closely with the regions flanking the CDR3 loop and the FR2 of the other chain, along with the observation of similar charge distribution along these proximal regions (inferred from Figure 4), suggests a potential structural motif that contributes to stable heavy-light chain pairing. The prevalence of this

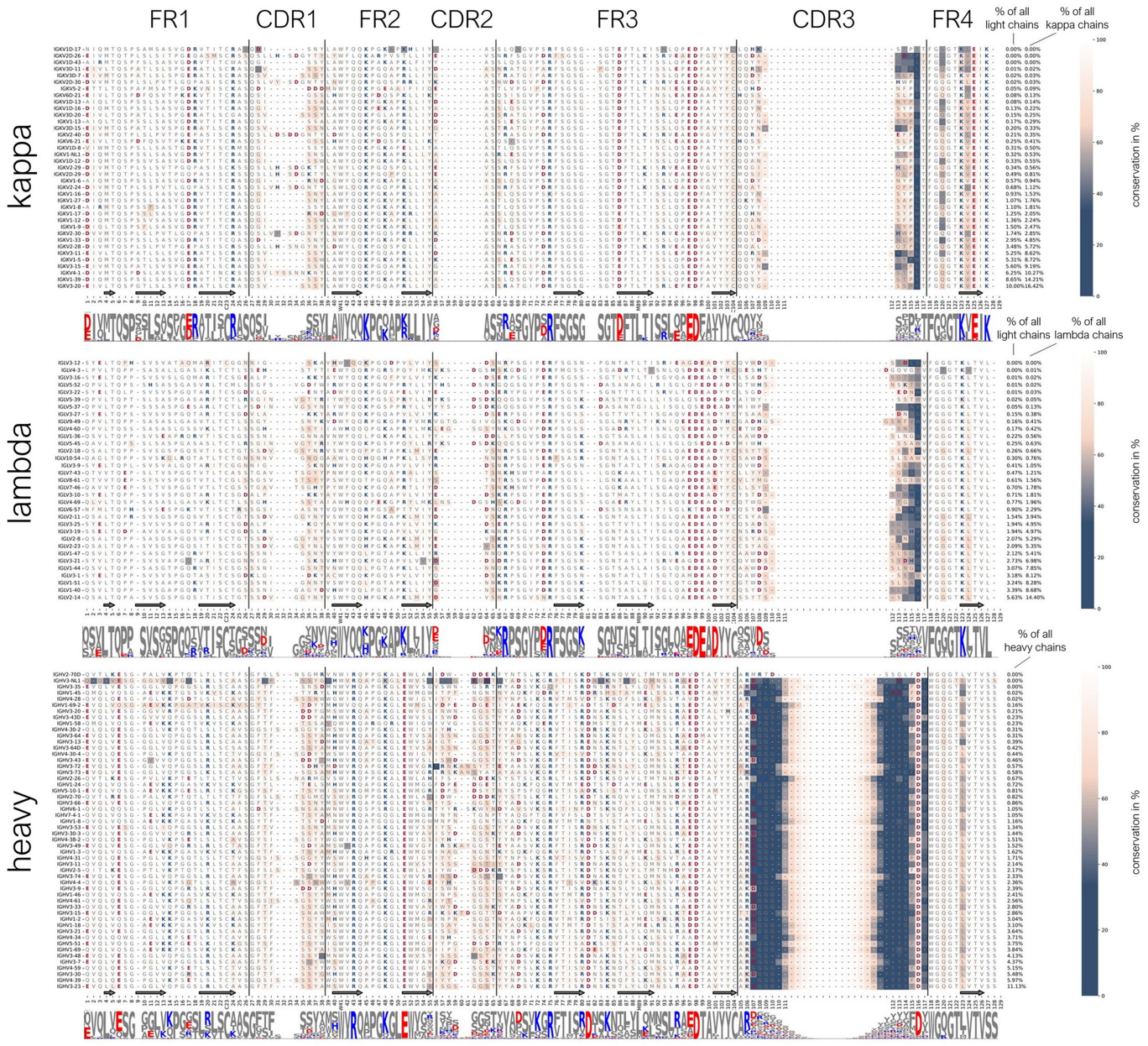

Figure 5. Sequence alignments of the entire variable domains, derived from the individual germline V-genes. Each row shows the consensus sequence over all sequences contained in the OAS having the respective germline V-gene. The alignment is divided into three sections, each displaying a distinct set of V-genes: kappa, lambda, and heavy chain germline V-genes. The sequence logos for all germlines are presented collectively beneath the sequences. These logos show the frequency of occurrence of amino acids on a position-wise basis. The presence of charged residues is indicated by the use of color coding, with basic residues displayed in blue and acidic residues in red. The color coding of the background serves to illustrate the degree of conservation at each amino acid position within a given germline. Beta-sheet structure elements are indicated by arrows beneath the sequences. The various framework and CDR loop regions are subdivided by separating black vertical lines and described above. Furthermore, for each germline gene, the percentage prevalence is given. For the light chains, the percentage of all light chains, as well as the occurrence within the single light chain types, is given next to the row of the corresponding germline. These are further ordered according to their occurrence in the OAS.

pattern within our extensive dataset provides evidence for its potential significance as a feature in heavy-light chain pairing.

Finally, we aimed to identify the amino acids that comprise the interface by calculating the closest residues within the two domains. The interacting residues within all cleaned structures were calculated, and the resulting contact plot is presented in Figure 7. In this plot, the heavy chain is displayed in the upper section, while the light chain is displayed in the lower section. The color coding of the lines indicates the percentage of structures in which the interaction is present. The analysis identified a total of 16 residues that are within 4Å, thereby constituting the interface between the two domains. The positions are additionally associated with the sequence logos derived from all analyzed crystal structures, displayed above and below the plot. These represent contacts occurring in over 50% of the analyzed structures, and are also listed in the table within the figure.

In Figure 7b, the positions highlighted in the interaction plot are shown on the schematic structure of the antibody domains. Our analysis revealed that most of the interface

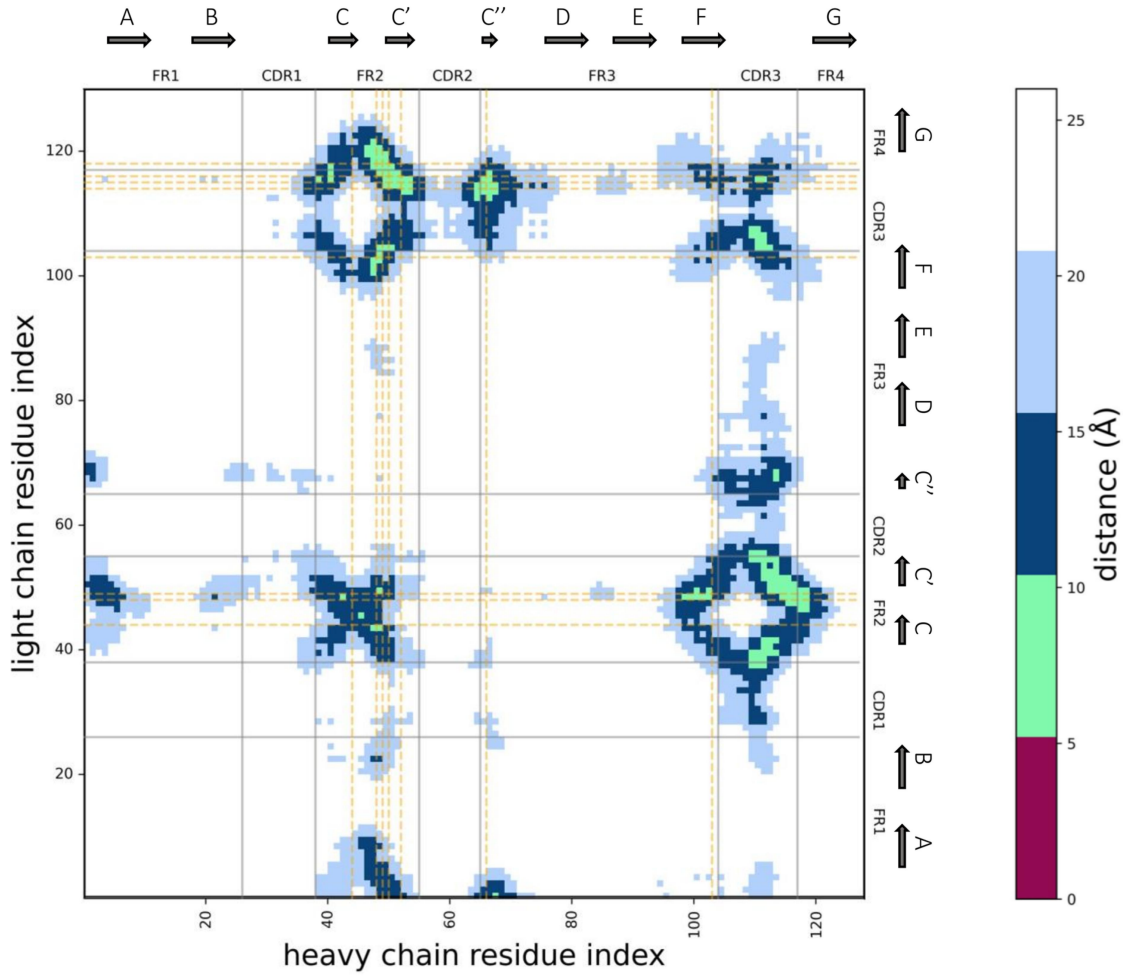

Figure 6. Interpolated median distance matrix over a sample of more than 3500 experimentally determined structures. The primary interaction points are evident in the region of the heavy chain C and C′ strands (equivalent to the FR2) with the L3 loop of the light chain, the H3 loop with the L3 loop, the C and C′ strands of the light chain (FR2) with the H3 loop, and the two C, C′ sheets of the two domains. Additionally, the regions adjacent to the CDR loop regions show slight contacts with the N-terminal ends of the two chains. The FR and CDR regions are highlighted with gray lines, while the orange dotted lines represent the closest interface residue contacts.

residues are located in the loop between the two strands C and C', while for the light chain additional residues were found to be situated at the terminal part of the CDR L3 loop. Another notable aspect is the nature of the interface residues: many of the residues are bulky, such as tyrosines and tryptophans, and in the light chain also a conserved phenylalanine was identified. The interface residues of both chains are highlighted as orange dotted lines in the distance matrix of Figure 6. The points where the lines intersect correspond to the interface contacts between the two chains.

Discussion

The assembly of antibody heavy and light chains is a critical step in the generation of functional antibodies. Proper pairing of these chains is essential for the formation of antigen-binding sites with high specificity and affinity. While the immune system generates a vast repertoire of antibody diversity through V (D)J recombination, somatic hypermutation and class switch recombination, the factors influencing the pairing of specific heavy and light chains remain incompletely understood. A deeper understanding of the factors influencing this pairing could significantly benefit the development of therapeutic antibodies, e.g., by providing a biophysically optimal pair of human acceptor frameworks for a given nonhuman antibody that needs to be humanized. Currently, the industry appears to use a limited subset of potential germline variants, often selecting them seemingly at random. This could be due to a reliance on historical assessments of germline usage frequency, leading to a bias toward certain germlines while neglecting others that are more rarely observed. The variability of the therapeutics is shown in Figure 8. In this figure, the distribution of therapeutic antibodies that are listed in Thera-SAbDab[32] (20.12.2024) and are either currently in clinical trials (Phase 1, 2, or 3) or approved are shown over the OAS distribution. The chains of each antibody are mapped to their closest human germline V-gene pairs using ANARCI[33] for the sequence alignment. The broad dispersion of antibodies suggests a lack of clear preference for specific germline combinations. The size of the dots corresponds to the clinical development stage of the antibody. Smaller dots represent early-phase clinical trials, while larger dots indicate later-stage or approved therapies. Some germlines, like IGHV3–23, appear frequently, likely due to their abundance. However, IGHV1–46, used in 83 of 817 antibodies, warrants

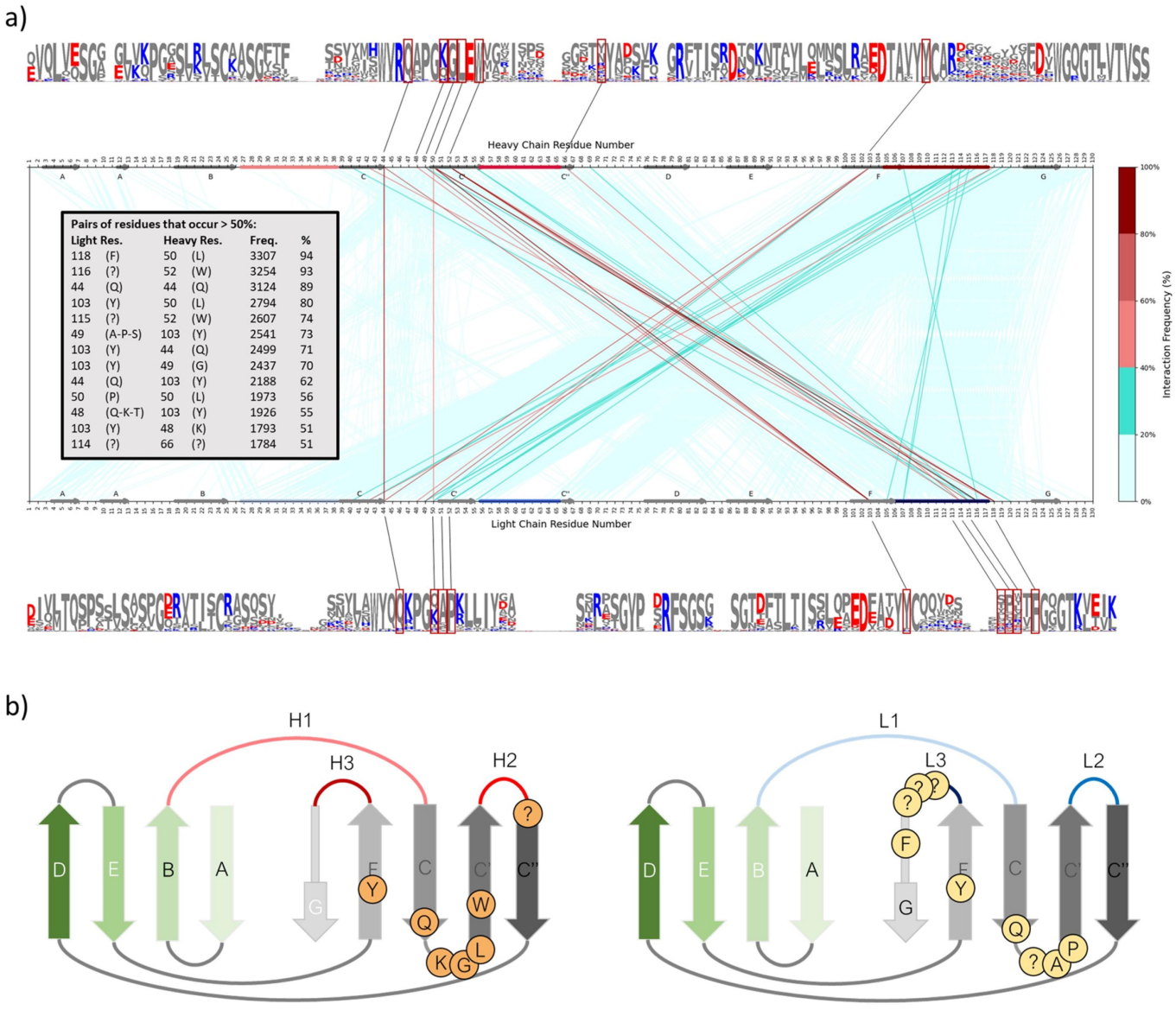

Figure 7. Interactions between the heavy and light chains within 4Å. In subfigure a), the positions that constitute the interface residues in over 50% of the analyzed structures are highlighted in the respective sequence logos. Additionally, subfigure b) depicts the positions in a schematic representation of the Ig-folded domains, with the interface residues highlighted. If the position contains a highly conserved residue, it is displayed in the structure; otherwise, the position is marked by a question mark.

closer investigation. Further analysis revealed that many of these antibody therapeutics are genetically not fully human, suggesting the use of a frequent animal-derived germline in their production.

In order to better understand the pairing, this study provides a comprehensive analysis of the most frequent human V-gene germline sequences, their resulting V_H-V_L pairs, the charge distributions across variable domains, and pairing preferences. Figure 2 reveals a clear trend of frequent germline pairings, although certain outliers deviate from this pattern. For instance, the heavy chain germline IGHV3–72 exhibits a higher-than-expected pairing frequency with the light chain germlines IGKV9–1, IGKV1–17, and IGKV1–16. Conversely, IGHV4–34 pairs less frequently with IGLV3–1. This discrepancy may be attributed to the unique charge distribution of IGLV3–1, which is rich in charged residues, while IGHV4–34 lacks certain conserved charged residues (Figure 4b). Figure 3b highlights numerous additional examples of over- and under-represented pairings, which can be further explored through the sequence alignments in Figure 5. The prevailing scientific literature suggests slightly positively charged antibodies, which is in accordance with our findings, as visible from Figure 4b. Nevertheless, certain germline sequences, such as IGLV3–22, IGKV5–2, IGHV1–3, IGHV1–24, and IGHV1-69-2, consistently exhibit atypical pairing behavior, suggesting unique charge-based interactions. It is worth noting, that most outliers deviate from the median charge of + 3 in a negative direction. This suggests that extremely positive charges might be less tolerated in therapeutic antibodies, potentially due to factors like increased clearance rates, which are known to cause developability problems.[34–36] We have attempted to identify regions of opposite charged residues, and could identify the sheets A, E and F of the light chains, and sheet F of the heavy chains as mainly negatively charged. All these regions are not situated in

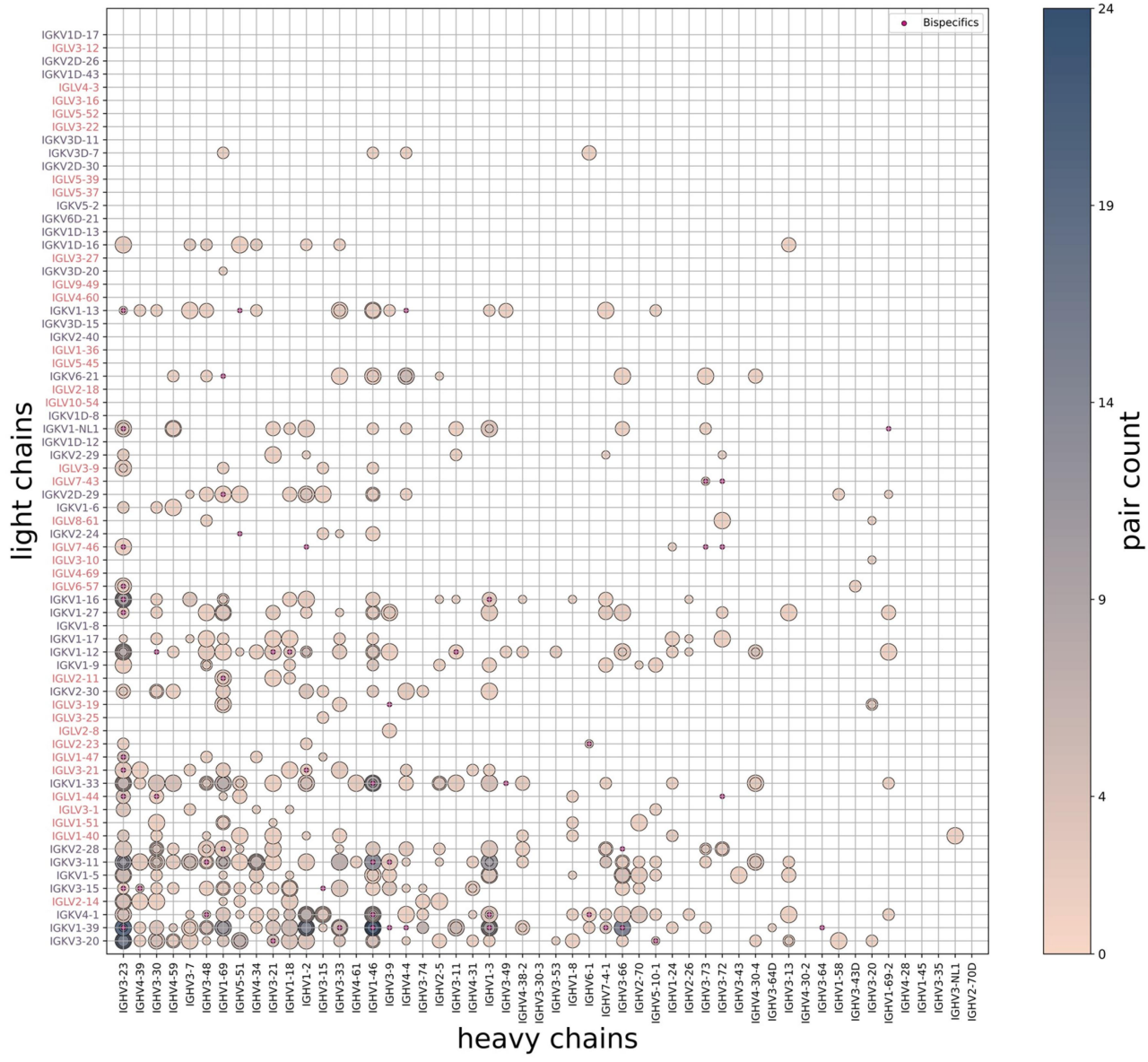

Figure 8. Distribution of human germline V-genes most closely related to the variable regions of therapeutic antibodies currently in clinical trials or approved for market use. Data was sourced from the Thera-SAbDab database. Color-coding highlights V-gene pairs that are frequently utilized in multiple therapeutic antibodies. The size of the data points is indicative of the progression of the respective clinical trials, with the smallest data points indicating Phase 1 and the largest data points indicating approved pharmaceuticals.

the core region of the interface, according to our interface definition shown in Figure 6. However, due to their long-range nature, the electrostatic interactions may exert an influence on the surrounding area over a longer distance.

Further analysis of the differences in charge distributions of the heavy and light chains for kappa and lambda light chains was conducted individually. This data is displayed in the supplementary information document as Figure S1. Our analysis revealed significant disparities in the charge distributions between kappa and lambda light chains, primarily within sheets A, E, and F. Specifically, while kappa light chains predominantly exhibited a negative charge in sheet A, lambda chains demonstrated an average charge of neutrality. In a similar manner, the F sheet displays a more negative charge in lambda light chains in comparison to kappa light chains. It is notable that the F strand of the light chain interacts primarily with the region between the CDR loops 1 and 2 of the heavy chain. It was also observed that some heavy chain germlines possess additional positive charges in this region. This finding suggests a potential selective advantage for these heavy chain germlines in pairing with the more negatively charged lambda light chains. Indeed, analysis of the IGHV3–30 germline (as a random example), which exhibits this additional positive charge, showed a slightly elevated pairing propensity for lambda light chains compared to the overall database distribution. These findings underscore the necessity of incorporating light chain isotype considerations into the analysis of heavy-light chain pairing preferences.

While the present study emphasizes the role of electrostatic interactions in heavy-light chain pairing, it is important to consider the potential influence of the light chain isotype. As demonstrated in Figure 7a, the frequency of residue pairs at the

heavy-light chain interface is illustrated across the entire dataset of antibody structures, without distinguishing between kappa and lambda light chains. Consequently, a direct comparison of kappa versus lambda interactions cannot be made from the original Figure 7a. In order to address this issue, a separate analysis was performed and sequence logos for the variable domains of kappa and lambda light chains were generated (see Figure S3 in the Supplementary Information). These logos reveal several differences in amino acid usage between kappa and lambda light chains, including variations in charge properties. However, at the main interacting residue positions frequently observed in our structural data (represented in Figure 7a), the sequence logos do not show clear or substantial deviations in preferred amino acid types between kappa and lambda light chains. This finding indicates that, at these pivotal contact points, driving frequent pairing, the underlying amino acid preferences might be largely conserved, irrespective of the light chain isotype. Furthermore, plots were created to illustrate the interacting residues for kappa and lambda light chains on an individual basis. It was noted that the PDB structures in question contain 70% kappa and 30% lambda light chains. This analysis of overall sequence preferences, categorized according to isotype, offers a more comprehensive understanding of potential differences between kappa and lambda light chain interactions.

Consequently, while electrostatic interactions involving charged residues undoubtedly play a role in heavy-light chain pairing, the evidence suggests a complex interplay where light chain isotype and other factors, including hydrophobic interactions, hydrogen bonds, and potentially aromatic interactions, also contribute to the overall affinity and stability of the paired antibody variable domains.

Our analysis also highlights the role of sequence variability in shaping antibody properties. The sequence analysis reveals a high degree of conservation across germline sequences. However, rarer germline sequences exhibit slightly greater variability. It remains unclear whether this increased variability is a cause or consequence of lower expression. Notably, the highest variability is observed in regions flanking and within the CDR loops. This is expected, as CDR loop sequences are influenced not only by V-genes, but also by J and D genes, as well as somatic hypermutation, deletion, and insertion, which are not considered in our analyses.

The most abundant interactions within the interface involve a phenylalanine following the CDR L3 loop interacting with a leucine at the beginning of the C' sheet, a tryptophan in a similar position on the heavy chain interacting with an unconserved residue on the light chain, and a well-established glutamine–glutamine interaction between the two C-C' loops. These interactions are shown in Figure 9, and their percentages over all calculated structures can be seen in Figure 7a.

One possible bottleneck in in-depth understanding of antibody V_H-V_L pairing is the limited data available. While the dataset used in this analysis is currently the most extensive, a larger dataset of paired sequences would significantly enhance our understanding of the underlying mechanisms. Such a larger dataset would allow for a more comprehensive analysis, potentially revealing subtle patterns and correlations

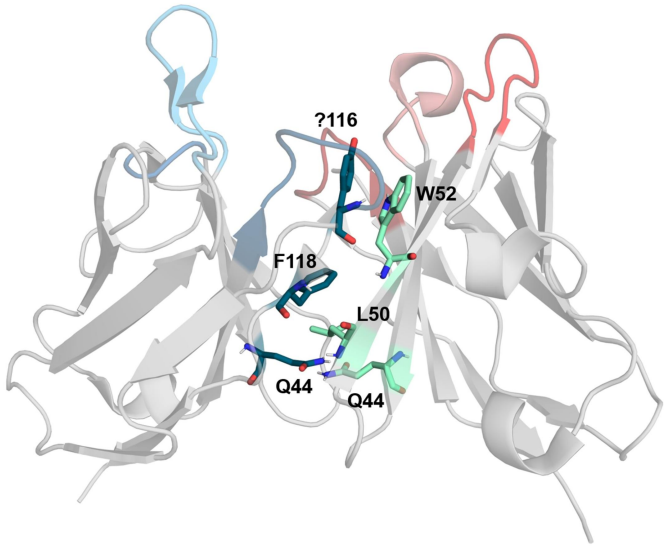

Figure 9. The three most frequent interface interactions. This figure highlights the three most common interactions observed within the heavy-light chain interface. Interacting residues, labeled according to IMGT nomenclature, are depicted as sticks. Light chain residues are colored mint, while heavy chain residues are shown in dark teal.

that are currently obscured by the limited sample size. This would also provide insight into whether the improved distribution of data points (*i.e.*, a pattern closer to the expected counts) in Figure 2b) is due to limited data availability or the intrinsic physicochemical properties of these germlines.

Additionally, for many of the analyzed questions, it is unclear whether the observed higher abundance of certain sequences is a direct consequence of their influence on pairing, or rather a result of their increased expression or stability. Further investigation is necessary to elucidate the precise mechanisms driving these abundance patterns.

Conclusion

This study examined the intricate relationship between human antibody heavy and light chain variable domains derived from different germlines, challenging the long-held assumption of random pairing. By analyzing a substantial dataset of paired sequences from the Observed Antibody Space (OAS) and structural information from the SAbDab database, we were able to gain valuable insights. A comprehensive analysis of germline sequences yielded valuable information regarding their frequency, variability, and potential pairing preferences.

The analysis yielded evidence of nonrandom pairing patterns, with specific heavy and light chain combinations displaying a proclivity for preferential associations. However, there are also indications that, in particular for more abundant germlines, the statistical distributions indicate a pairing with all chains, provided that they are present in high quantity.

Furthermore, electrostatic interactions appear to play a significant role in the pairing process. It is evident that there is a tendency for the pairing of the heavy chain with a charge of + 2 with light chains of a charge of + 1. However, there are also instances where this rule is not followed.

A total of 16 crucial residues at the heavy-light chain interface were identified and defined as "interface residues", which appear to play an important role in complex formation.

Our findings have significant implications for the understanding of antibody formation and the rational design of therapeutic antibodies. Future research and the development of computational models to predict pairing outcomes based on sequence and structural features would be a valuable tool for antibody engineering. By unraveling the intricacies of antibody pairing, we can accelerate the development of innovative antibody-based therapies and advance our understanding of the immune system.

Materials and methods

Most of this work is based on data extracted from the Observed Antibody Space (OAS) database created by Charlotte Deane and Co-workers at Oxford University (http://opig.stats.ox.ac.uk/webapps/oas/).[20] Besides a large number of unpaired antibody variable sequences, this database houses paired antibody variable domain sequences generated through 10×Genomics B-cell receptor repertoire sequencing. We accessed the web server (10.05.2024) and filtered for exclusively human and paired heavy-light chain sequences. The resulting 1,954,079 filtered sequences from 10 studies were downloaded, retaining only information relevant to our analysis, such as amino acid sequences and framework/CDR region delimiters based on the IMGT numbering scheme. This scheme allowed us to directly identify the different regions of interest in both heavy and light chains.[13,20]

Following download, the data was further processed to exclude mis-paired entries (e.g., heavy-heavy or light-light chain pairings). Of the remaining 1,954,070 entries, we also grouped different alleles from the same germline gene by eliminating the extension after the '*' symbol in the gene names. Data underwent extensive analysis using Python libraries. NumPy (v1.21)[37] was employed for efficient numerical operations on the multi-dimensional array data, while Pandas (v2.0)[38] facilitated data manipulation, reshaping, and merging. Matplotlib[39] was used to generate visualizations that supported the analysis.

Pair counts

In order to calculate the pairing preferences, the prepared dataset was used to enumerate the occurrences of V genes of each unique light chain, each unique heavy chain, and each pair. It should be noted that a chain designated as "unique" does not necessarily display 100% sequence identity to all other germline genes of the same type. Further, for most of our calculations we did not consider the J and D genes. Both of them were shown to be responsible mainly for the diversification of the CDR loops H3 and L3.[12] For the initial analyses of pairing occurrences, the same germline entries were grouped together without further differentiation. The occurrence of each unique pair was counted and visualized. In addition to the pairing occurrences counted, the pairing occurrences for the ideal case of equally probable pairing were calculated. The expected pair count for each heavy-light chain pair is given by the following formula:

$$expected = \frac{(light\ chain\ count_i * heavy\ chain\ count_j)}{tot.\ number\ of\ pairs}$$

Investigation of charges

The most pervasive and robust non-bonded interactions in proteins are electrostatic interactions. Consequently, we investigated the charges of the protein chains. In a first place, the calculation of the charges was performed for the entire chains. To this end, we did a straightforward addition of counting positively charged amino acids (arginine, lysine, and histidine). Due to the dual protonation state of histidine at a physiological pH, we conducted this analysis twice: once counting the positive charge of the histidines and once assuming the histidines to be neutral. Due to the rare occurrence of this amino acid type, the results did not significantly change. We then subtracted one charge for each negative residue (glutamic acid and aspartic acid). The charges were calculated for the pairs of germlines, as well as for the individual sequences and for the single fragments assembling to the whole Fv sequence.

Sequence analysis

The primary distinctions between the different framework regions, as well as CDR loop sequences, were then subjected to comprehensive examination. Accordingly, our analysis was primarily based on the sequence alignment of the annotated data set. The alignment was performed by use of the Python implementation of Clustal Omega.[40] In a first step, all regions were limited to a maximum number of residues, specifying a maximum number of CDR loop and FR residues. Longer sequences were eliminated as wrongly annotated outliers or as exceptions having an unusual number of insertions. Therefore, the following maximum numbers of residues were chosen for all gene variants (heavy chains, kappa light chains and lambda light chains). This decision was made in order to consistently obtain equally long chains for further analyses and representations.

FR1	CDR1	FR2	CDR2	FR3	CDR3	FR4
26	12	17	10	38	30	11

This led to a reduction of the dataset of about 2.7%, mainly due to wrongly annotated or especially long loop regions, resulting in 1,901,959 paired sequences. From these pairs, 1,157,977 were found to be paired with kappa light chains, and the remaining 743,982 with lambda light chains. In the next step, a sequence alignment of all sequences was performed by use of the Clustal Omega algorithm, and the gaps added as "-" signs.

The resulting sequence length of the heavy chains resulted in 92 framework residues, on which a position-wise comparison of the amino acids was performed. Given our understanding that the kappa, lambda, and heavy chains may exhibit significant sequence divergence, we conducted this

analysis separately for each. We then grouped the unique germline sequences and quantified the variability within each group. For the determination of the secondary structure elements, a subset of 22,074 sequences were modeled and the secondary structure elements determined with the DSSP algorithm implemented in cpptraj.[41] The amino acids appearing in at least 90% of all structures as beta-sheet structure elements were marked on the sequence plots of Figure 5 as gray arrows. Moreover, sequence logos for the entire domain sequences were constructed for the heavy chains, kappa light chains, and lambda light chains individually. These were generated through the use of the Logomaker[42] Python package.

Structural analyses/interface residues

For structural data analysis, we utilized the SAbDab database.[28] To focus on human antibodies, we filtered the database to include only structures containing the Fv region of human antibodies, by specifying "Fv" as the antibody type and "Human" as the species. This resulted in a dataset of 3,801 experimentally determined structures.

To prepare the dataset for analysis, we removed water molecules and symmetric replicas from the structures. We ensured that both heavy and light chains were annotated according to IMGT nomenclature and standardized chain identifiers to 'H' for heavy chains and 'L' for light chains.[13] Additionally, we eliminated ligands and antigens from the structures. After this preprocessing, we obtained a dataset of 3505 antibody structures containing only correctly annotated variable domains.

To calculate distances between heavy and light chain residues and identify interface residues, we employed Biopython libraries (Bio.PDB and Bio.SeqUtils)[43] to parse each structure. For each heavy-light chain residue pair, we calculated the minimum distance between their closest atoms and stored these distances in a distance matrix.

Using the distance matrices, we identified residue pairs between both chains within a distance cutoff of 4 Å, and defined these positions as interface residues. The resulting interface residue contacts and their frequencies were visualized in the Results section in Figure 6.

To gain further insights into the structural features of antibody-antigen interactions, we calculated a median distance matrix by interpolating the individual distance matrices. The interp2d function from the SciPy library was used for interpolation.[44] The resulting median distance matrix was visualized in Figure 7.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) [https://doi.org/10.55776/P34518].

ORCID

Clarissa A. Seidler http://orcid.org/0000-0001-7859-9234
Alexander Bujotzek http://orcid.org/0000-0001-5052-0221
Klaus R. Liedl http://orcid.org/0000-0002-0985-2299

Author contributions

The manuscript was discussed and written through contributions of all authors. All authors have given approval to the final version of the manuscript.

References

1. Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody structure and function: the basis for engineering therapeutics. Antibodies. 2019;8(4):55. doi: 10.3390/antib8040055.
2. Sela-Culang I, Kunik V, Ofran Y. The structural basis of antibody-antigen recognition. Front Immunol. 2013;4:302. doi: 10.3389/fimmu.2013.00302.
3. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. Annu Rev Biochem. 2007;76(1):1–22. doi: 10.1146/annurev.biochem.76.061705.090740.
4. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983;302(5909):575–581. doi: 10.1038/302575a0.
5. Tonegawa S. Somatic generation of immune diversity (nobel lecture). Angew Chem Int Ed In English. 1988;27(8):1028–1039. doi: 10.1002/anie.198810281.
6. Strohl WR, Strohl LM. Therapeutic antibody engineering: current and future advances driving the strongest growth area in the pharmaceutical industry. Amsterdam, Netherlands: Elsevier; 2012. ISBN 1-908818-09-3.
7. Megha K, Mohanan P. Role of immunoglobulin and antibodies in disease management. Int J Biol Macromolecules. 2021;169:28–38. doi: 10.1016/j.ijbiomac.2020.12.073.
8. Xu JL, Davis MM. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. Immunity. 2000;13(1):37–45. doi: 10.1016/S1074-7613(00)00006-6.
9. Lefranc M-P. Nomenclature of the human immunoglobulin kappa (IGK) genes. Exp Clin Immunogenet. 2001;18(3):161–174. doi: 10.1159/000049195.
10. Lefranc M-P. Nomenclature of the human immunoglobulin lambda (IGL) genes. Exp Clin Immunogenet. 2002;18(4):242–254. doi: 10.1159/000049203.
11. Lefranc M-P. Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp Clin Immunogenet. 2001;18(2):100–116. doi: 10.1159/000049189.
12. Lefranc M-P, Lefranc G. The immunoglobulin factsbook. Cambridge, Massachussets, USA: Academic Press; 2001.
13. Lefranc M-P, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, et al. IMGT®, the international ImMunoGeneTics information system®. Nucleic Acids Res. 2009;37(Database):D1006–D1012. doi: 10.1093/nar/gkn838.
14. Brezinschek H-P, Foster SJ, Dörner T, Brezinschek RI, Lipsky PE. Pairing of variable heavy and variable κ chains in individual naive and memory B cells. J Immunol. 1998;160(10):4762–4767. doi: 10.4049/jimmunol.160.10.4762.
15. De Wildt RM, Hoet RM, van Venrooij WJ, Tomlinson IM, Winter G. Analysis of heavy and light chain pairings indicates that receptor editing shapes the human antibody repertoire. J Mol Biol. 1999;285(3):895–901. doi: 10.1006/jmbi.1998.2396.
16. Chothia C, Novotný J, Bruccoleri R, Karplus M. Domain association in immunoglobulin molecules: the packing of variable domains. J Mol Biol. 1985;186(3):651–663. doi: 10.1016/0022-2836(85)90137-8.
17. Jayaram N, Bhowmick P, Martin AC. Germline VH/VL pairing in antibodies. Protein Eng Des Sel. 2012;25(10):523–530. doi: 10.1093/protein/gzs043.
18. Abhinandan K, Martin AC. Analysis and prediction of VH/VL packing in antibodies. Protein Eng Des Sel. 2010;23(9):689–697. doi: 10.1093/protein/gzq043.

19. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed antibody space: a resource for data mining next-generation sequencing of antibody repertoires. J Immunol. 2018;201(8):2502–2509. doi: 10.4049/jimmunol.1800708.
20. Olsen TH, Boyles F, Deane CM. Observed antibody space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Sci. 2022;31(1):141–146. doi: 10.1002/pro.4205.
21. Collis AV, Brouwer AP, Martin AC. 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. 2003;325(2):337–354. doi: 10.1016/S0022-2836(02)01222-6.
22. Bork P, Holm L, Sander C. The immunoglobulin Fold: structural classification, sequence patterns and common core. J Mol Biol. 1994;242(4):309–320. doi: 10.1016/S0022-2836(84)71582-8.
23. Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and ig superfamily V-like domains. Dev Comp Immunol. 2003;27(1):55–77. doi: 10.1016/S0145-305X(02)00039-3.
24. Tan PH, Sandmaier BM, Stayton PS. Contributions of a highly conserved VH/VL hydrogen bonding interaction to scFv folding stability and refolding efficiency. Biophys J. 1998;75(3):1473–1482. doi: 10.1016/S0006-3495(98)74066-4.
25. Herold EM, John C, Weber B, Kremser S, Eras J, Berner C, Deubler S, Zacharias M, Buchner J. Determinants of the assembly and function of antibody variable domains. Sci Rep. 2017;7(1):12276. doi: 10.1038/s41598-017-12519-9.
26. Vargas-Madrazo E, Paz-García E. An improved model of association for VH–VL immunoglobulin domains: asymmetries between VH and VL in the packing of some interface residues. J Mol Recognit. 2003;16(3):113–120. doi: 10.1002/jmr.613.
27. Chatellier J, Van Regenmortel MH, Vernet T, Altschuh D. Functional mapping of conserved residues located at the VL and VH domain interface of a fab. J Mol Biol. 1996;264(1):1–6. doi: 10.1006/jmbi.1996.0618.
28. Dunbar J, Krawczyk K, Leem J, Baker T, Fuchs A, Georges G, Shi J, Deane CM. SAbDab: the structural antibody database. Nucleic Acids Res. 2014;42(D1):D1140–D1146. doi: 10.1093/nar/gkt1043.
29. Zar JH. Biostatistical analysis. Chennai, Tamil Nadu, India: Pearson Education India; 1999. ISBN 81-7758-582-7.
30. Agresti A. Categorical Data Analysis. Hoboken, New Jersey, USA: John Wiley & Sons; 2013. ISBN 1-118-71094-0.
31. Massey FJ Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. J Am Stat Assoc. 1951;46(253):68–78. doi: 10.1080/01621459.1951.10500769.
32. Raybould MI, Marks C, Lewis AP, Shi J, Bujotzek A, Taddese B, Deane CM. Thera-SAbDab: the therapeutic structural antibody database. Nucleic Acids Res. 2020;48(D1):D383–D388. doi: 10.1093/nar/gkz827.
33. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298–300. doi: 10.1093/bioinformatics/btv552.
34. Kelly RL, Yu Y, Sun T, Caffry I, Lynaugh H, Brown M, Jain T, Xu Y, Wittrup KD. Target-Independent variable region mediated effects on antibody clearance can Be FcRn independent. Proceedings of the MAbs; 2016. Oxfirdshire, United Kingdom: Taylor & Francis; p. 1269–1275.
35. Schoch A, Kettenberger H, Mundigl O, Winter G, Engert J, Heinrich J, Emrich T. Charge-mediated influence of the antibody variable domain on FcRn-dependent pharmacokinetics. Proceedings of the National Academy of Sciences; Washington, DC, USA. 2015. p. 5997–6002.
36. Liu S, Verma A, Kettenberger H, Richter WF, Shah DK. Effect of variable domain charge on in vitro and in vivo disposition of monoclonal antibodies. Proceedings of the MAbs; 2021. Oxfirdshire, United Kingdom: Taylor & Francis; p. 1993769.
37. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, et al. Array programming with NumPy. Nat. 2020;585(7825):357–362. doi: 10.1038/s41586-020-2649-2.
38. McKinney W. Data structures for statistical computing in Python. Proceedings of the SciPy (Austin, Texas, USA); 2010. p. 51–56.
39. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9(3):90–95. doi: 10.1109/MCSE.2007.55.
40. Sievers F, Higgins DG. Clustal omega, accurate alignment of very large numbers of sequences. Mul Seq Align Meth. 2014:105–116.
41. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. J Chem Theory Comput. 2013;9(7):3084–3095. doi: 10.1021/ct400341p.
42. Tareen A, Kinney JB, Valencia A. Logomaker: beautiful sequence logos in Python. Bioinformatics. 2020;36(7):2272–2274. doi: 10.1093/bioinformatics/btz921.
43. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422. doi: 10.1093/bioinformatics/btp163.
44. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–272. doi: 10.1038/s41592-019-0686-2.