bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Biochemical Patterns of Antibody Polyreactivity Revealed Through a Bioinformatics-Based Analysis of CDR Loops

Christopher T. Boughter[1], Marta T. Borowska[2], Jenna J. Guthmiller[3], Albert Bendelac[4], Patrick C. Wilson[3,4], Benoit Roux[2], and Erin J. Adams[2,4,*]

[1]Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL

[2]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL

[3]Department of Medicine, Section of Rheumatology, University of Chicago, Chicago, IL

[4]Committee on Immunology, University of Chicago, Chicago, IL

[*]To whom correspondence should be addressed. Email: ejadams@uchicago.edu

Abstract

Antibodies are critical components of adaptive immunity, binding with high affinity to pathogenic epitopes. Antibodies undergo rigorous selection to achieve this high affinity, yet some maintain an additional basal level of low affinity, broad reactivity to diverse epitopes, a phenomenon termed "polyreactivity". While polyreactivity has been observed in antibodies isolated from various immunological niches, the biophysical properties that allow for promiscuity in a protein selected for high affinity binding to a single target remain unclear. Using a database of nearly 1,500 polyreactive and non-polyreactive antibody sequences, we created a bioinformatic pipeline to isolate key determinants of polyreactivity. These determinants, which include an increase in inter-loop crosstalk and a propensity for an "inoffensive" binding surface, are sufficient to generate a classifier able to identify polyreactive antibodies with over 75% accuracy. The framework from which this classifier was built is generalizable, and represents a powerful, automated pipeline for future immune repertoire analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Antibodies are immunogenic proteins expressed by B cells that play a major role in the adaptive immune response against non-self. Upon recognition of target epitopes, these antibodies undergo multiple rounds of somatic hypermutation and affinity maturation inside a germinal center, whereby the amino acid sequence of the epitope-binding surface is selected for optimal binding to the target [1–3]. The longer this affinity maturation process extends, the higher the affinity and specificity of the antibodies towards their target antigen, primarily through mutagenesis of the six complementarity determining region (CDR) loops of the antibody [1]. Using a combination of affinity matured CDR loops, these antibodies bind strongly to the target and aid in invader neutralization. While the process of affinity maturation and somatic hypermutation of antibodies results in high-affinity and incredibly specific binders to a particular epitope, some antibodies have been shown to display signs of reactivity towards diverse off-target epitopes. This broad but low-affinity binding has been termed "polyreactivity".

Antibody polyreactivity has been hypothesized to be beneficial in the early stages of antibody maturation, acting as a pool of diverse binders ready to recognize novel antigens and initiate the more stringent selection process [4]. To this end, a majority of B cell receptors and antibodies which have not undergone somatic hypermutation, including those on immature B cells and early "natural" antibodies, have been found to be polyreactive to some extent and are suggested to have an innate-like response to pathogens [5, 6]. While these mostly unmutated polyreactive antibodies remain at low frequency in antigen-experienced individuals, a distinct population of polyreactive antibodies that have undergone selection are still expressed by mature B cells that circulate in blood [7]. In fact, some studies have found the polyreactivity status of an antibody is mostly independent of the number of somatic hypermutations in the antibody sequence [8, 9]. In line with this finding, only 5-10% of the repertoire of naive B cells circulating in the periphery are polyreactive, but this increases to 20-30% in the memory B cell compartment, showing a distinct capability of polyreactivity to survive selection [7, 10]. These results suggest that polyreactivity can persist, or perhaps even be selected for during the selection process within the germinal center.

In a few notable cases, polyreactivity may in fact augment the efficacy of a given immune response. Polyreactive IgA antibodies have been shown to have an inherent reactivity to microbiota in the mouse gut, with a predicted role in host homeostasis [11]. These previously identified antibodies so far have no known primary ligands, yet play a key role in facilitating the gut immune response to the plethora of exogenous antigens encountered in the dynamic dietary and microbial environment of the gut. This implies the existence of antibodies whose primary function is to act as polyreac-

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tive sentries in the gut, yet the downstream effects of polyreactive antibodies coating commensal bacteria is so far unclear. Similar polyreactive IgA and IgG mucosal antibodies were found in the gut of human immunodeficiency virus (HIV) infected patients, but these antibodies either had low affinity to the virus or lacked neutralization capabilities [12]. The benefit of singular antibody sequences with the ability to sample large portions of the commensal population may represent an improvement in efficiency of the homeostatic machinery of the gut.

While the precise role of these primarily polyreactive gut antibodies is still a topic of debate, polyreactivity has been suggested to augment the immune response in other immunological niches. Broadly neutralizing antibodies (bnAbs), which bind robustly to conserved epitopes on the surface glycoproteins of influenza viruses or HIV are more likely to be polyreactive [13–15]. In one study of HIV binding antibodies, over half of all tested bnAbs were found to be polyreactive [16]. These bnAbs have been the subject of intense study for their potential as the central components of an HIV treatment or as the byproduct of an immune response to a universal Influenza vaccine [15, 17–19]. One hypothesized mechanism for the capability of polyreactive antibodies to confer this broad neutralization in the face of a changing viral epitope is heteroligation, the ability of a single antibody to bind the primary target with one binding domain and use the other binding domain to bind in a polyreactive manner [8]. This heteroligation allows the antibody to take advantage of the significant avidity increase afforded by bivalent binding, despite the low envelope protein density of HIV or a geometry which does not readily lend itself to bivalent binding on the surface of influenza viruses [20].

Although polyreactivity may play a positive role in natural immune responses, oftentimes this same property is considered undesirable from the point of view of generating therapeutic antibodies with high specificity. Antibody-based treatments, which generally take the form of an intravenous transfusion, are sensitive to the accelerated systemic clearance of polyreactive antibodies [21–24]. In general, much work has focused on attempting to answer the question of optimizing “developability” of a given antibody. These efforts have been dedicated to determining the most critical components of developability through a large array of experimental assays, in silico structural prediction-based methods, sequence-based analysis and their correlations with clearance, sequence-based SASA predictions, and sequence-based aggregation propensity predictors [25–29]. In many of these studies polyreactivity or non-specificity in general was seen to be a negative indicator of the developability of a drug, suggesting that therapeutic antibodies should strive towards a drug-like specificity [30].

In line with this goal of understanding the predominant factors involved in the specificity of therapeutic antibodies, many researchers have worked to identify the biophysical underpinnings of polyre-

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

activity in natural immune responses. The most popular hypotheses for the primary biophysical predictors of polyreactivity have included CDR3 length [9], CDR3 flexibility [16], net hydrophobicity [31] and net charge [32]. More observational studies have found an increased prevalence of arginine and tyrosine in polyreactive antibodies [23, 33]. While these previous studies represent substantial advances in the study of polyreactivity, they have often been limited in scope, focusing on a singular antibody source and primarily focused on CDR3H. Comparing across these individual antibody sources highlights discrepancies between the proposed predictors of polyreactivity. The aforementioned properties determined to be key to polyreactivity in previous studies were found to be statistically insignificant in studies of HIV-binding and mouse gut polyreactive antibodies [8,11].

Clearly, a computational framework that would enable us to predict the polyreactivity of a given antibody *a priori*, whether evaluating the efficacy of a natural immune response or the potential fate of a therapeutic antibody, would be tremendously useful. Such a framework, for example, could be used to assist in the isolation of broadly neutralizing anti-viral antibodies, or speed up the process of therapeutic antibody screening. To achieve this goal, a thorough understanding of the molecular features behind polyreactive binding interactions is critical. Experimental approaches utilizing next-generation sequencing and ELISA allow for the identification of hundreds of polyreactive antibody sequences. However, the systematic characterization of these antibodies is difficult. More detailed biochemical studies of polyreactive antibodies via protein crystallography, quantitative binding experiments, and mutagenesis provide exceptional insight but are inherently low throughput. Structural modeling of these polyreactive antibodies represent a high throughput approach, but models of flexible loops are relatively unreliable, and are unlikely to capture nuances in side-chain placement [34]. A bioinformatics-based approach, centered around high throughput analysis that minimizes structural assumptions while maintaining positional context of amino acid sequences would provide a thorough, unbiased analysis of existing data and create a powerful pipeline for future studies.

In this study, we show that, using just the amino acid sequences of antibodies from a database of nearly 1,500 polyreactive and non-polyreactive sequences, unifying biophysical properties that distinguish polyreactive antibodies from non-polyreactive antibodies can be identified. We find that, while charge and hydrophobicity are in fact important determinants of polyreactivity, the characteristic feature of polyreactive antibodies appears to be a shift towards neutrality of the binding interface. In addition, loop crosstalk is more prevalent in the heavy chain of polyreactive antibodies than non-polyreactive antibodies. From these properties, a machine learning-based classification software was developed with the capability to determine the polyreactivity status of a given sequence. This software is generalizable and can be re-trained on any binary classification

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

problem and identify the key differences between two distinct populations of antibodies, T cell receptors, or MHC-like molecules at the amino acid level. As a test case, the same analysis was applied to a dataset of therapeutic antibodies, demonstrating the overall flexibility of the software generated in this study.

Results

Database

Our aggregate database of nearly 1,500 antibody sequences is compiled from our own previously published and new data, published studies by the Mouquet and Nussenzweig labs, and the therapeutic antibody database TheraSabDab (Table 1) [8, 11, 12, 14, 16]. Using an ELISA-based assay, the reactivity of each antibody is tested against a panel of 4-7 biochemically diverse target antigens: DNA, Insulin, lipopolysaccharide (LPS), flagellin, albumin, cardiolipin, and keyhole limpet hemocyanin (KLH). This panel has become increasingly prevalent in the literature for experimental measures of polyreactivity in antibodies [8, 9, 11, 12, 14–16, 25, 35, 36]. The ligands represent a diverse sampling of biophysical and biochemical properties; for example, enrichment in negative charge (DNA, insulin, LPS, albumin), amphipathic in nature (LPS, cardiolipin), exceptionally polar (KLH), or large in size (KLH, flagellin). From this panel, a general rating of "polyreactive" or "non-polyreactive" is given to 529 and 524 antibodies, respectively. For the purposes of this study, antibodies are determined to be polyreactive if the authors of the original studies determined a particular clone binds to two or more ligands in the panel. Those that bind to one or none of the ligands in the panel are deemed non-polyreactive. The nearly 500 therapeutic antibodies are treated separately, as many of these sequences either are not measured for polyreactivity or use a different metric as a measure of polyreactivity. The results presented below utilize this dataset of 1053 non-therapeutic antibody sequences, unless otherwise noted.

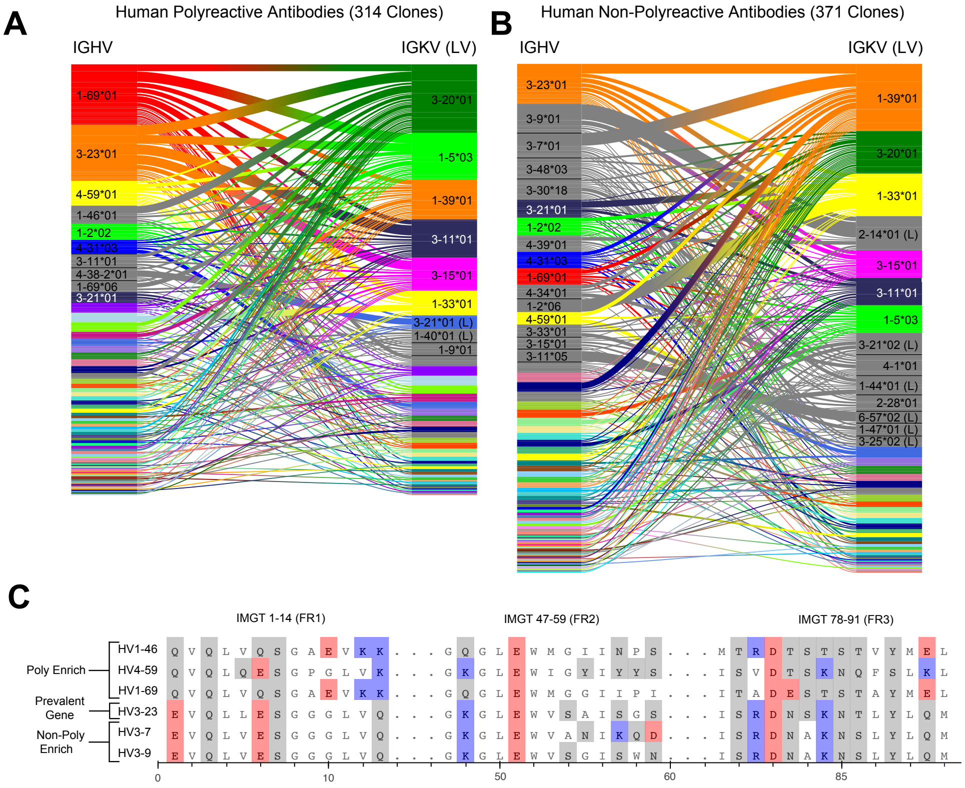
Dataset	Polyreactive	Non-Polyreactive	Total
Mouse IgA	205	240	445
HIV Reactive	172	124	296
Influenza Reactive	152	160	312
Therapeutics	-	-	434
	529	524	1487

Table 1: A quantification of the antibodies used in this study.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A Surface-Level Analysis of Polyreactive Antibody Sequences

As a first pass at the given dataset, we focus on the most simplistic of the possible explanations for differences between polyreactive and non-polyreactive antibodies, specifically the J- and V-gene usage of each group. Figure 1A and 1B, rendered with code adapted from the Dash et. al. derived program TCRdist [37], represents each antibody V-gene as a line connecting a single heavy and light chain gene for the human-derived antibodies (685 sequences).



D

Genes	Count Polyreactive	Percent Polyreactive	Count Non-Polyreactive	Percent Non-Polyreactive
HV3-23*01	41	13.06 %	29	7.82 %
HV3-9*01	3	0.96 %	21	5.66 %
HV3-7*01	5	1.59 %	18	4.85 %
HV1-69*01	44	14.01 %	12	3.23 %
HV4-59*01	18	5.73 %	9	2.43 %
HV1-46*01	13	4.14 %	5	1.35 %

Figure 1: **A comparative genetic analysis of human-derived polyreactive and non-polyreactive antibody sequences uncovers population level differences.** Gene usage diagrams comparing (A) human polyreactive and (B) non-polyreactive sequences show a qualitative difference in the VH gene usage. Shared colors indicate identical genes, grey indicates genes that are not seen in the other population at a level over 2%. Unlabeled genes are colored randomly to highlight genetic variation in the populations. (C) Sequence alignment of the most prevalent genes in the polyreactive and non-polyreactive populations compared to a reference gene common to each population. Hydrophobic amino acids are colored white, hydrophilic amino acids are colored grey, and positively or negatively charged amino acids are colored blue or red, respectively. (D) Percentage and raw count of observed gene usage for the polyreactive and non-polyreactive sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Direct comparisons between mouse and human derived antibodies is difficult at the gene usage level. A similar analysis highlighting differences between mouse polyreactive and non-polyreactive antibodies can be found in the supplement (Figure S1).

Genes are identified from nucleotide sequences using NCBI's IgBLAST command line tool [38]. Heavy and light chain genes that are shared between polyreactive and non-polyreactive sequences are colored for the top labelled instances. Genes which are labelled but not found above a 2% threshold in the opposite population are colored grey, while those that do not have a visible name are colored randomly to highlight variation in gene usage. From this comparison, it is clear that the variable gene usage is skewed between polyreactive and non-polyreactive sequences, with an enrichment of V_H1-69, V_H1-46, and V_H4-59 in the polyreactive population. In contrast, no qualitative differences in the J-gene usage are readily discernible between these two groups (Figure S2).

While the full alignment of these most used heavy chain variable genes shows a high degree of sequence similarity (Figure S3), Figure 1C highlights the regions of highest dissimilarity between the biophysical properties of amino acids in prevalent genes within each population. V_H3-23, the most prevalent gene in the non-polyreactive human dataset and the second most prevalent gene in the polyreactive human dataset, can be used as a reference for comparisons between genes enriched in each individual population. This reference gene shares a high degree of sequence similarity with the second and third most frequently occurring genes in the non-polyreactive dataset, V_H3-7 and V_H3-9, save for a lysine and glutamic acid pair in framework 2 of V_H3-7. The genes enriched in the polyreactive dataset, however, are quite different from this reference. All three of the polyreactive enriched genes have charged residues where the non-polyreactive enriched genes have hydrophilic residues (or vice versa) at IMGT positions 1, 13, and 88. These initial results hint at some systematic differences between the polyreactive and non-polyreactive antibody populations.

Figure 1D quantifies the extent of the difference in gene usage in each population by comparing these most prominent genes from our accumulated dataset of HIV- and influenza virus-reactive antibodies. While the two most common genes in the polyreactive dataset account for 27% of the human polyreactive antibodies in this study, the top three most common genes in the non-polyreactive dataset account for just over 17% of the total population. In addition to being the most prevalent gene in the polyreactive dataset, V_H1-69*01 has also been found historically to be more prevalent in broadly neutralizing antibodies against influenza viruses, in line with the previously mentioned overlap between bnAbs and polyreactivity [15, 36].

Overall, there is a noticeable difference between the gene usage frequency of polyreactive and non-

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

polyreactive antibodies, but the overlap in the usage of the two populations suggests that gene usage alone is not sufficient to distinguish the two groups. While there exist qualitative differences between framework sequences enriched in the polyreactive dataset compared to the non-polyreactive population, a look at the amino acid usage of the CDR loops of each group shows no significant differences (Figure S4). This implies that the positional context of a given amino acid is critical to tease out differences in antibody binding properties.

A Position Sensitive Matrix Representation of Sequences Provides Further Insights into Polyreactivity

To identify deeper trends in the biophysical properties of polyreactive antibodies, we utilize a new methodology to analyze and represent a range of different properties inherent to these sequences. While the framework regions of antibodies are highly conserved, the CDR loops vary significantly in length and show very low conservation between populations. This makes alignment of CDR loops difficult without creating subgroups for loops of identical length. To overcome this, the sequence data is re-organized into a matrix representation (Figure 2A). Each sequence is aligned by the center of each CDR loop, with spaces between the loops set to zero and each amino acid encoded as a number from 1 to 21. While this alignment method excludes the framework regions of the antibodies and slightly averages out some of the properties at the edge of the CDR loops, we reason that most of these differences are evident in the gene usage analysis of the previous section. From this simple alignment, no obvious patterns emerge separating polyreactive and non-polyreactive antibodies, however we can clearly see that mouse gut-derived IgA antibodies have generally shorter CDR3H loops, and more conserved CDR3L sequences when compared to the human-derived antibody sequences. All subsequent analysis is derived from this matrix representation of the sequences.

With this new positionally sensitive and quantitative alignment method, we are able to further dissect the differences in amino acid sequences presented in Figure 1. Figure 2B uses this positional sequence encoding to determine the amino acid frequency difference between polyreactive and non-polyreactive sequences. For example, phenylalanine is found at position 93 in roughly 40% of polyreactive sequences and nearly 60% of non-polyreactive sequences. Therefore position 93, amino acid F has an intensity of -0.2 in Figure 2B. From this panel it is evident that most of the major differences are in the germline encoded regions CDR1H and CDR2H, in line with the observations from Figure 1 that suggest polyreactive antibodies have a distinct gene usage when compared to non-polyreactive antibodies. Figure 2C further expands on these differences, showing the largest changes in amino acid frequencies between the two populations. We can see that there is a slight decrease of phenylalanine frequency in CDR1H of polyreactive antibodies, in favor of isoleucine. Additionally, there is a general shift towards hydrophobicity in CDR2H, as the hydrophilic residue

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

serine at matrix positions 78 and 82 is less prevalent in polyreactive antibodies, instead replaced by the more hydrophobic residues isoleucine and glycine.

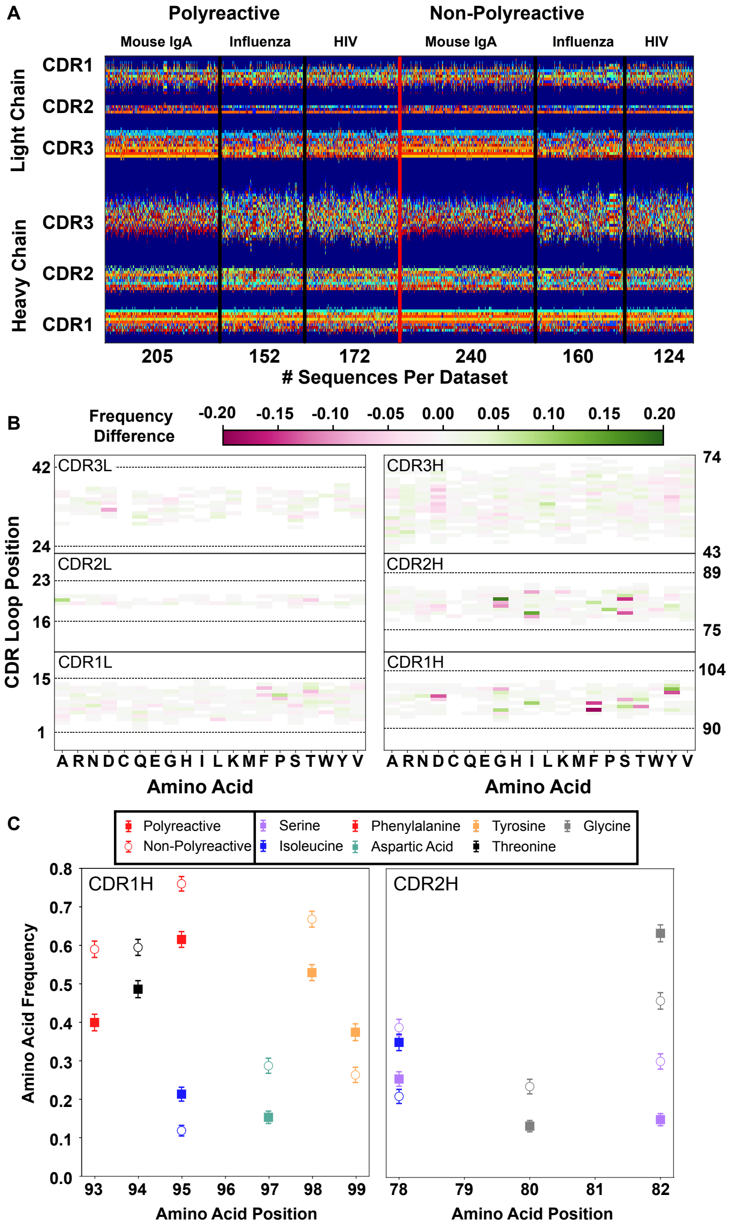


Figure 2: **A new representation of CDR loop sequences improves the position-sensitivity of quantitative antibody analysis.** (A) Matrix representation of the amino acid sequences used in this study provides a framework for further analysis. Each amino acid is encoded as a number from 1 to 21, represented by a distinct color in the matrix. A 0-value is used as a buffer between loops and is represented by the dark blue regions. The red line separates polyreactive and non-polyreactive sequences. (B) Amino acid frequency difference between polyreactive and non-polyreactive sequences for all six CDR loops. Residues more common in polyreactive sequences are shown in green, while those more common in non-polyreactive sequences are shown in pink. Loop positions correspond to the numerical position within the matrix of panel A. (C) An in-depth representation highlighting the amino acid frequencies used to create panel B. Only frequency changes greater than 10% are shown for clarity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This increased prevalence in loop hydrophobicity of polyreactive antibodies has been suggested before in the literature [16] along with a net increase in positive charge [32], so we next aimed to analyze this matrix systematically using biophysical properties inherent to the loops. A simple analysis of the full human and mouse-derived dataset investigating classical parameters explored previously by other groups (CDR loop length, net charge, net hydrophobicity, and gene usage) and some new properties (side chain flexibility, side chain bulk, and Kidera Factors [39]) show some significant differences between polyreactive and non-polyreactive antibodies (Figure 3A,B). The versatility of the positionally sensitive amino acid matrix allows for the application of multiple ”property masks” to tease out the specific regions of each CDR loop that contributes most to these significant differences. Given a property, amino acid charge for example, we can replace each simple 1-21 representation with a distinct representation based upon amino acid properties.

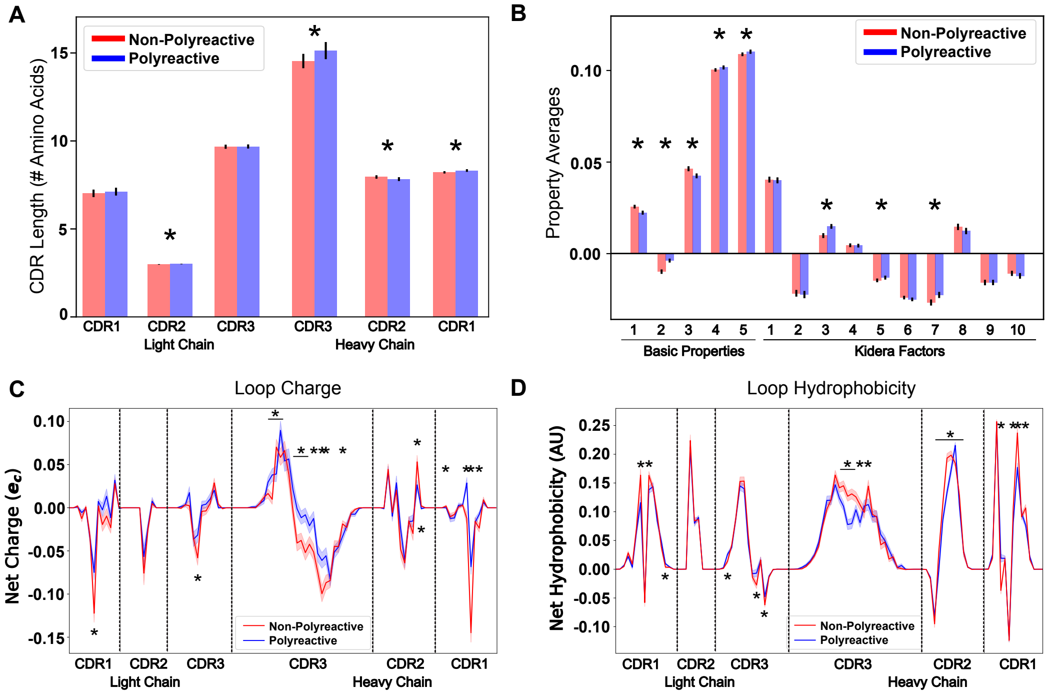


Figure 3: **Position-sensitive quantification of CDR loop properties of mouse and human antibody sequences highlights differences between polyreactive and non-polyreactive populations.** Plotting the average CDR loop lengths (A) and net antibody biophysical properties (B) show small but significant differences when analyzed in bulk. Basic properties 1-5 are hydrophobicity1, charge, hydrophobicity2, side chain flexibility, and side chain bulk. Plotting the average net charge (C) and hydrophobicity (D) as a function of position of polyreactive and non-polyreactive sequences highlights significant differences in CDR3H. Light shadow around lines represent bootstrap standard errors. All uncertainties obtained via bootstrapping. Stars indicate p-value ≤ 0.05 calculated via nonparametric Studentized bootstrap test. Bars with a single star above represent contiguous regions of significance.

In the matrix of Figure 2A leucine, histidine, and arginine are represented by the integers 3, 16, and 17. As an example, when the charge property mask is applied, the matrix representations

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of these three amino acids in all sequences is changed to 0.00, 0.091, and 1.00, respectively. We apply 62 such masks to this matrix, including simple metrics like charge, hydrophobicity, side chain flexibility, and side chain bulkiness to go along with more carefully curated metrics from the works of Kidera et. al. and Liu et. al [39, 40]. A complete description of these properties can be found in Supplemental Table 1. The application of these masks gives an entirely new matrix describing the localization of amino acids with a given property.

By averaging across all sequences in the polyreactive or non-polyreactive dataset when these masks are applied, we can readily see differences in charge patterning and hydrophobicity when comparing polyreactive and non-polyreactive sequences (Figure 3C,D). Including errors obtained via bootstrapping, we see that these differences are most pronounced in the center of CDR3H, with some differences also apparent in the remaining five loops. This analysis shows an overall bias towards neutrality (i.e. neither positively nor negatively charged, neither strongly hydrophilic nor hydrophobic) in these regions. These results also contextualize the findings of Figure 2C. The trend towards hydrophobic residues in CDR2H of polyreactive antibodies importantly does not make these regions net hydrophobic, but instead make these regions slightly less hydrophilic on average.

Systematic Determination of the Key Contributions to Polyreactivity

Along with simple property averaging, these masks also give a high dimensional space from which we can determine, in an unbiased way, the primary factors that discriminate polyreactive and non-polyreactive antibodies. As a first pass, we apply a principal component analysis (PCA) to the matrix of all antibody sequences in an attempt to separate the polyreactive or non-polyreactive populations along the axes of highest variation in the dataset. Unfortunately, the principal components of these data do not effectively distinguish between the two populations (Figure S5).

To further investigate the physical and sequence-based properties of polyreactivity in antibodies in a more targeted manner, we employ linear discriminant analysis (LDA), a common technique often applied in classification problems [41–43]. LDA works in a manner conceptually similar to PCA, reducing the dimensionality of a given dataset via a linear combination of the original dimensions. However, LDA takes one additional input, the label or class of each sequence. Whereas the objective of PCA is to identify the axes which maximize the variance in the dataset, LDA has the dual objective of maximizing the projected distance between two classes while minimizing the variance within a given class. While LDA is more well adapted for classifying two distinct populations, it is susceptible to overfitting, unlike PCA [44]. Here, we have labelled our two classes in the matrix

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with either a "1" for polyreactive, or "0" for non-polyreactive. In our application of LDA we parse down the large number of input vectors using either PCA or an algorithm which selects the vectors with the largest average differences between the two populations. This reduction in dimensionality ensures the data are not being overfit, and the tunable number of input vectors allows us to control for overfitting in each individual application.

Figure 4A shows the results of LDA when applied to a parsed dataset comprised of 311 polyreactive antibodies and 362 non-polyreactive antibodies. A limitation of the full human and mouse-derived polyreactivity dataset is that there exists an intermediate between the two classes. It is not immediately obvious where the line for polyreactivity should be drawn. An antibody that binds to 2-3 ligands may not necessarily achieve broad reactivity through the same mechanism as an antibody that binds 4 or more ligands from a panel of 6 or 7. To remove these ambiguities, in this parsed dataset we denote antibodies that bind 4-7 ligands as polyreactive, antibodies that bind 0 panel ligands as non-polyreactive, and those that bind 1-3 are removed from the analysis.

LDA analysis is versatile in its applications, and in this work we utilize the method in two distinct modes. In the first mode, all of the available data is used as input with the output vector representing the features that best distinguish between the two complete populations. Plots of the data projected onto this vector (as in Figure 4A) represent the maximum achievable separation between the two populations for a defined number of input components from the given biophysical property matrix. In the second mode, we utilize LDA as a more canonical classification algorithm separating the data randomly into training and test groups. In this classification mode of operation, a combination of correlation analysis coupled with maximal average differences is used to parse input features, and a support vector machine (SVM) is used to generate the final classifier from these features. Accuracy of the resultant classifiers is assessed via leave one out cross validation, these accuracies are shown in Figure 4B.

In the first mode, we find that the data can be split more effectively when the parsed dataset is broken up into the distinct "reactivity" groups, i.e. those antibodies specific for influenza viruses, HIV, or found in the mouse gut (Figure 4A). This suggests there may be some bias due to antigen specificity, or lack thereof, whereby influenza virus-specific antibodies take a slightly different path towards polyreactivity compared to HIV reactive or mouse gut IgA antibodies. However, when using the classification mode, the classification accuracy is roughly equivalent across all tested datasets (Figure 4B). Testing this classifier with a scrambled dataset, where the labels are randomly assigned, shows the expected decrease in classification accuracy for each individual dataset for all ranges of input features.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

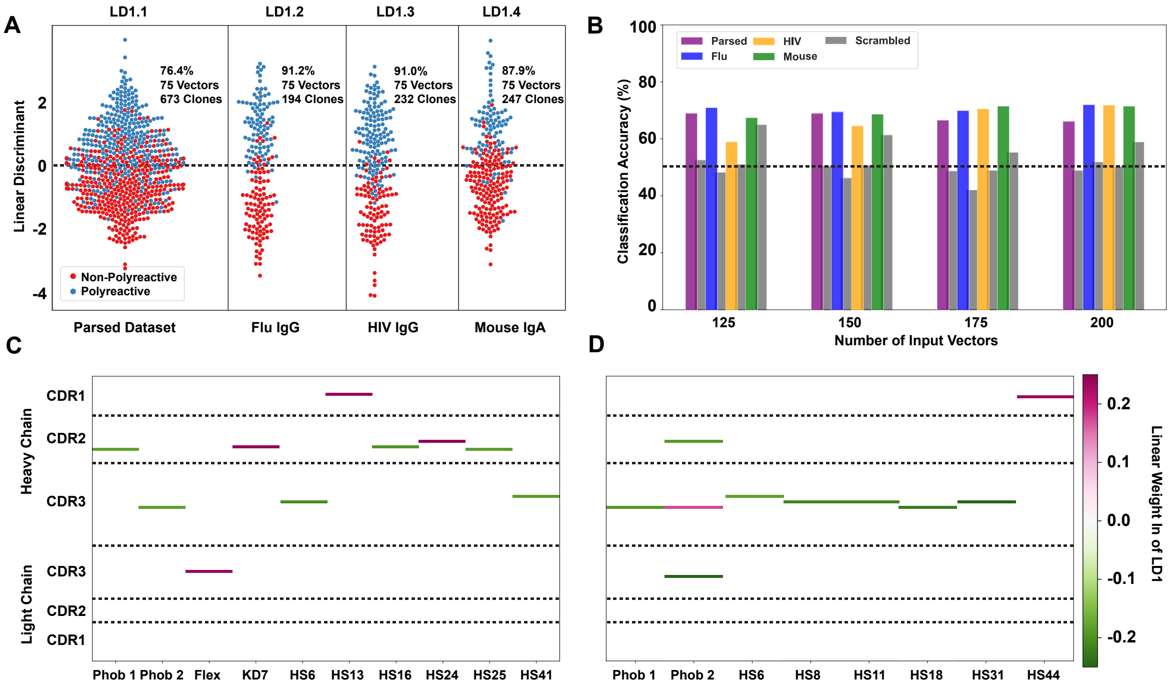


Figure 4: **Linear discriminant analysis (LDA) can meaningfully separate the two populations and these meaningful differences can be used to generate a polyreactivity classifier.** LDA applied individually to the complete parsed, Influenza, HIV, and mouse datasets. Percentages indicate the accuracy of the linear discriminant in labelling polyreactive and non-polyreactive antibodies. For these data, the plotted linear discriminants are comprised of different linear weights. (B) Accuracies of a polyreactivity classifier with a separate test and training dataset. Groupings in this figure are the same as those in panel A. A support vector machine is generated for each individual population, and the reported values are accuracies calculated through leave one out cross validation. Shown are test data and a scrambled dataset where the labels of "polyreactive" or "non-polyreactive" are applied randomly (grey bars). The dotted line indicates 50% accuracy threshold. (C) Property matrices highlighting the top 10 weights of the linear discriminants in panel A for the parsed dataset with 75 vectors (C) and the HIV dataset with 75 vectors (D). Color bar represents the normalized weight of each property, where pink rectangles represent properties positively correlated with increased polyreactivity, and green rectangles represent properties negatively correlated with decreased polyreactivity. For clarity, only the top ten linear weights are included. The full matrix of this data can be found in supplemental Figure S6.

When applying LDA in the first mode (Figure 4A), we can directly pull the linear weights of each component comprising linear discriminant 1 and reveal which biophysical properties at each CDR position best distinguish between the two populations. The differences in the linear weights from the heavy chain CDR loops comprising each discriminant show clear differences when comparing the complete parsed dataset (Figure 4C) to the HIV only dataset (Figure 4D). In the parsed dataset, the discriminating weights are heavily concentrated in CDR2H. Whereas in the HIV dataset, these weights are centered around the CDR3H loop. Only the top ten linear weights are shown in

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4C,D. The full matrix of linear weights can be found in Figure S6. The predominant discriminating factors between datasets might be due to the significant difference in CDR3H length between the mouse (IgA) and the human datasets, which confounds the analysis in this region. However, when examining each individual subset of the complete dataset we do find that there are common properties that seem to be the primary discriminators (i.e. largest linear weights). These are hydrophobicity 1, hydrophobicity 2, and hotspot variable 6 (a structural parameter related to alpha-helix propensity).

An Information Theoretic Approach

While analysis of the biophysical property differences between polyreactive and non-polyreactive sequences provides some insight into the molecular basis for the polyreactivity phenomenon, a broad unifying pattern which could discern the biophysical mechanism behind polyreactivity was not readily evident across all types of antibodies. To probe these polyreactive sequences in a quantitative yet more coarse manner, we applied the formalism of information theory to our dataset of antibody sequences. Information theory, a theory classically applied to communication across noisy channels, is incredibly versatile in its applications, with high potential for further applications in immunology [45–50]. In this work, we utilize two powerful concepts from information theory, namely Shannon entropy and mutual information.

Shannon entropy, in its simplest form, can be used as a proxy for the diversity in a given input population. This entropy, denoted as H has the general form:

$$H(X) = -\sum_{X} p(x) \log_2 p(x) \tag{1}$$

Where $p(x)$ is the occurrence probability of a given event, and X is the set of all events. We can then calculate this entropy at every position along the CDR loops, where X is the set of all amino acids, and $p(x)$ is the probability of seeing a specific amino acid at the given position. In other words, we want to determine, for a given site in a CDR loop, how much diversity (or entropy) is present. Figure 5A shows this Shannon entropy distribution for the full dataset of polyreactive and non-polyreactive antibodies. Given there are only 20 amino acids used in naturally derived antibodies, we can calculate a theoretical maximum entropy of 4.2 bits, which assumes that every amino acid occurs at a given position with equal probability. Although the observed entropy of the CDR3H loop approaches this theoretical maximum, it hovers below it (3.5 Bits) due to the relative absence of the amino acids cysteine and proline in the center of this loop. The difference in the entropy distributions in CDR1H are consistent with the bias in amino acid usage in this region, shown previously in Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Importantly, from this entropy we can calculate an equally interesting property of the dataset, namely the mutual information. Mutual information is similar, but not identical to, correlation. Whereas correlations are required to be linear, if two amino acids vary in any linked way, this will be reflected as an increase in mutual information. In addition, due to some of the highly conserved residues in the non-CDR3H loops, high covariance can be achieved for residues that have not been specifically selected for in the germinal center. Using this information theory framework, these conserved residues have a mutual information of 0. Overall, the mutual information can be used to identify patterns in antibody sequences that were not readily evident through the previous analysis in this or other studies. If there is some coevolution or crosstalk between residues undergoing some selection pressure in the antibody maturation process, it will be reflected as an increase in the mutual information.

In this work, mutual information $I(X;Y)$ is calculated by subtracting the Shannon entropy described above from the conditional Shannon entropy $H(X|Y)$ at each given position as seen in equations 2 and 3:

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \tag{2}$$

$$I(X;Y) = H(X) - H(X|Y) \tag{3}$$

To orient ourselves in physical space, Figure 5B gives an example crystal structure (PDB: 5UGY) [51] highlighting the lateral arrangements of the CDR loops. The matrix in Figure 5C shows that the mutual information between CDR loops on this binding surface is increased in the heavy chains of polyreactive antibodies over non-polyreactive ones, suggesting there exists more loop crosstalk in antibodies that exhibit polyreactivity. Interestingly, it appears that there is a corresponding decrease of loop crosstalk in the light chains of polyreactive antibodies. This observed crosstalk persists across all polyreactive antibodies within all subsets of our tested dataset and is evident both in intra-loop and inter-loop interactions. Figure 5D highlights some examples of the interesting significant differences of this crosstalk at distinct given positions within CDR1H and CDR3H. A complete plot of the statistically significant differences ($p \leq 0.05$) of Figure 5C (Figure S7) shows that a large portion of these differences are in fact significant.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

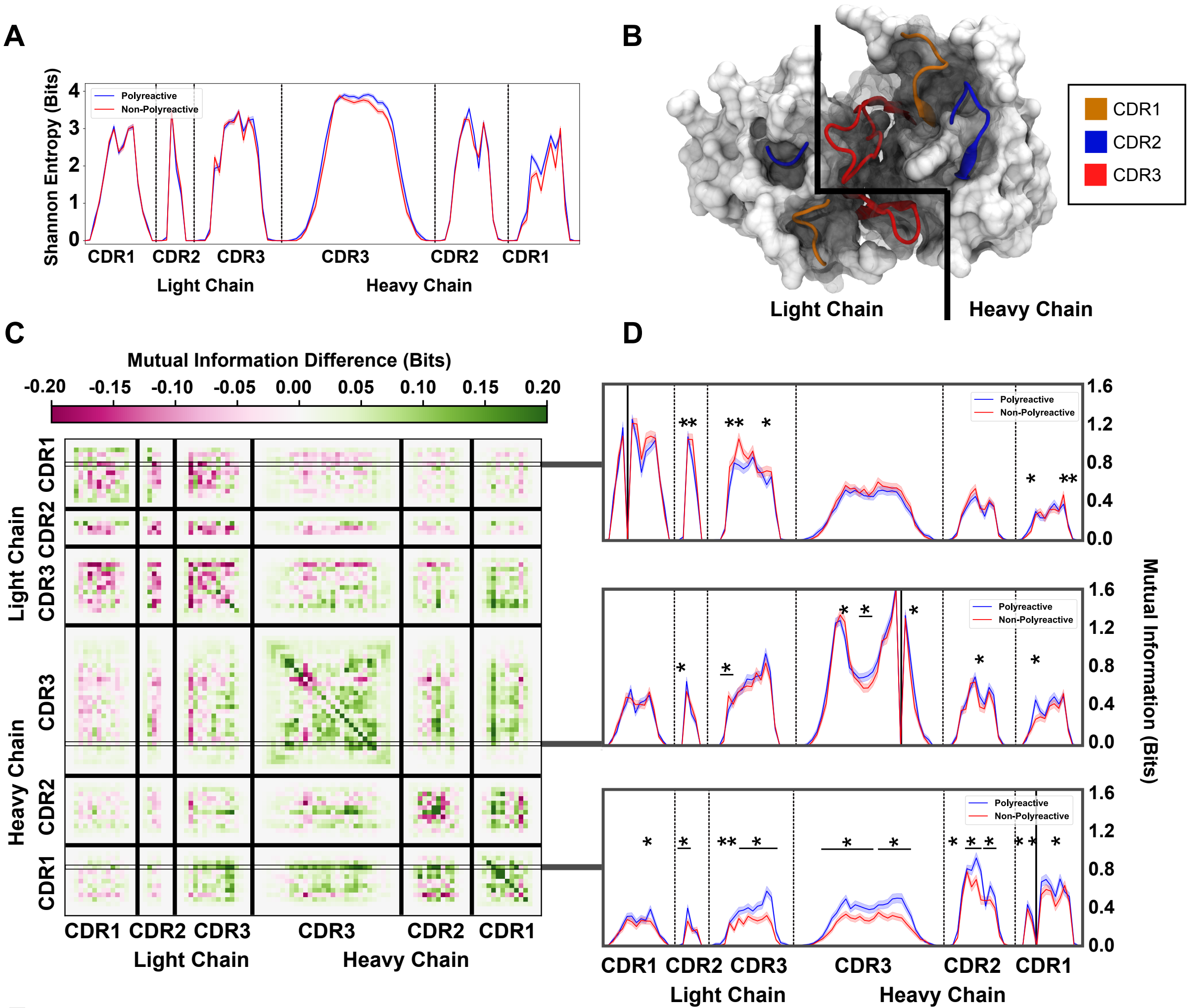


Figure 5: **An information theoretic analysis of antibody sequences shows an increase in polyreactive antibody loop crosstalk.** (A) The sequence diversity of the polyreactive and non-polyreactive datasets, quantified using Shannon Entropy, highlight similar diversities between the two groups. (B) A crystal structure (PDB: 5UGY) provides a visual representation of the lateral organization of the CDR loops on the antibody binding surface. (C) The difference in mutual information between polyreactive and non-polyreactive sequences shows that CDR loops of the heavy chain have more crosstalk in polyreactive antibodies. Each individual row represents the given condition, whereas each column gives the location the mutual information is calculated. (D) Singular slices of the mutual information show the data in (C), projected from the matrix onto a line, highlighting the significance of the differences at these particular locations. The positions of the "given" amino acid, i.e. the particular Y in $H(X|Y)$, are highlighted by grey boxes in panel C. Solid black lines indicate where on the X-axis this "given" amino acid is located. Stars indicate statistical significance ($p \leq 0.05$) calculated through a nonparametric permutation test. Bars with a single star above represent contiguous regions of significance.

The ordering of these entropy and information plots was chosen to reflect the spatial arrangement of the loops on the antibody surface; as such they show also that mutual information between loops drops off with physical distance between these loops. In other words, loops (and residues) that are located close to each other will have more of an effect on their direct neighbors as opposed to those

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

that are more physically distant. This increased mutual information suggests that in the heavy chains of polyreactive antibodies, there is enhanced cooperativity or co-evolution of the amino acids of intra- and inter-CDR loop pairs.

Application to Therapeutic Antibodies

As discussed previously, many studies on antibody repertoires specific to a given target have also revealed polyreactivity in these binders. Given the architecture of the software built around this bioinformatic analysis of polyreactivity in natural immune responses, the identical treatment of therapeutic antibodies is a logical next step. Using the published experimental tests of Jain & Sun et. al. and the extensive database provided by Thera-SAbDab we were able to compare the polyreactivity of a natural immune response with that seen in therapeutic antibodies [25, 52, 53].

Figure 6A shows the extent to which a linear discriminant trained on the parsed polyreactivity dataset can effectively discriminate approved and discontinued antibody therapeutics. From these plots we see that polyreactivity status of naturally-derived antibodies does not correlate well with the acceptance or discontinuation of a therapeutic antibody. Additionally, the polyreactivity status of naturally-derived antibodies correlates poorly with the reported polyreactivity of therapeutic antibodies (Figure S8). Importantly however, polyreactivity for these therapeutic antibodies is reported in a different manner compared to the other antibodies in this study. Rather than a count of the number of ligands the antibody reacts to, the polyreactivity is reported as an average score. Re-training the linear discriminant on these therapeutic antibodies (Figure 6B), shows an ability to split the approved and discontinued antibodies with an accuracy of 76% when using LDA mode 1 with 15 input vectors. While the software does seem able to effectively split approved and discontinued therapeutic antibodies to some extent, the biophysical properties which are effectively creating this split are not as obvious as in the case of polyreactive and non-polyreactive naturally derived antibodies.

Both the position-sensitive charge and hydrophobicity (Figure S9) show no significant differences between approved and discontinued antibodies. Plotting the linear weights of LD1 from Figure 6B, we can see that the primary discriminating factors between approved and discontinued antibodies are unsurprisingly centered around CDR3H. Significant differences can be seen in the CDR3H average value of Kidera Factor 7, a metric based upon side chain partial specific volume (Figure 6D). Overall, the software can meaningfully separate and analyze a binary split between groups, demonstrating its applicability to a broad array of sequence analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

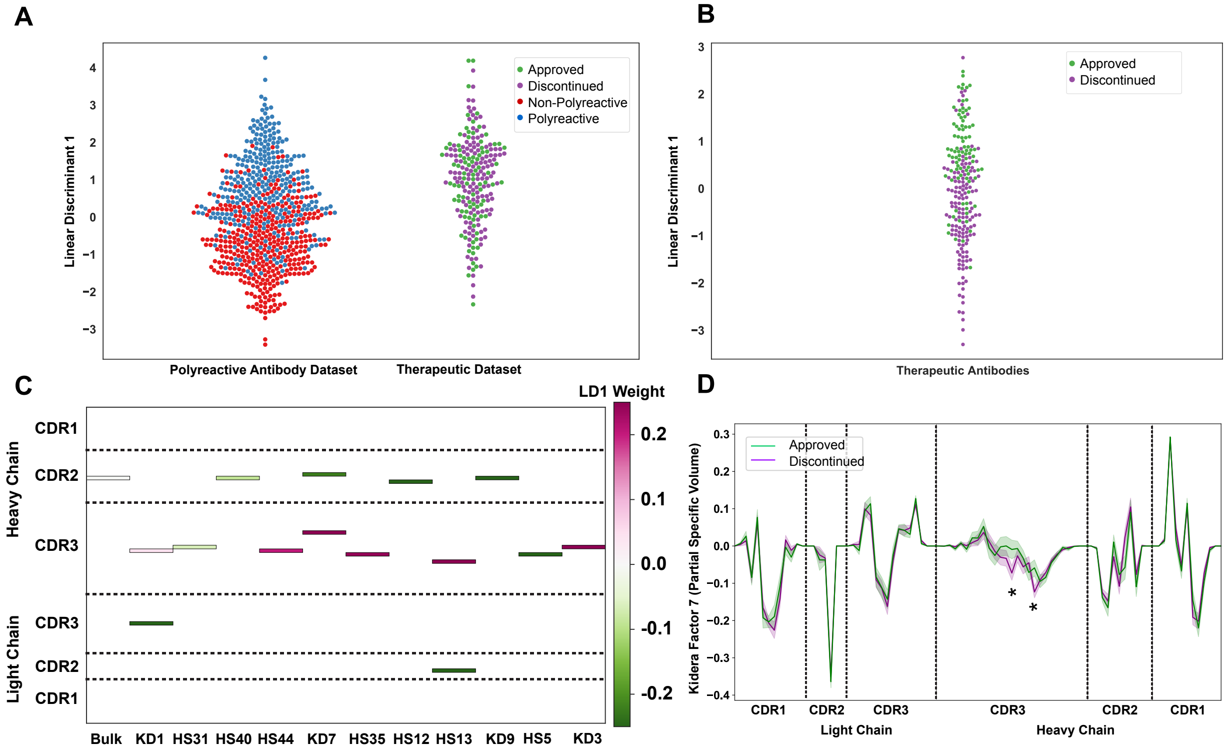


Figure 6: **An application of the linear discriminant analysis module of the software to therapeutic antibodies highlights the broad applicability of this analysis.** (A) A linear discriminant generated using the parsed naturally-derived antibody dataset applied to approved and discontinued therapeutic antibodies. (B) Projection of the approved and discontinued antibodies onto a linear discriminant trained on that data. (C) The location and intensity of the linear weights of the linear discriminant in panel B highlight the properties that best split the approved and discontinued antibodies. (D) Position sensitive plot of Kidera Factor 7, for approved and discontinued therapeutic antibodies. Stars indicate significance of $p \leq 0.05$, calculated via one-sided non-parametric bootstrap test. Error bars calculated using the bootstrapped standard deviation.

Discussion

Previous research has highlighted the importance of hydrophobicity, charge, and CDR loop flexibility on antibody specificity. In this work, we expand upon these previous results with a new bioinformatic and biophysical characterization of polyreactive antibodies. The software generated for this study provides a powerful computational tool which can be utilized by researchers interested in discerning differences between populations of adaptive immune molecules in broad contexts. Building off of the efforts of our own work and that of experimental collaborators, we were able to aggregate to date one of the largest publicly available datasets of antibodies tested for polyreactivity. Differences in the germline gene frequency and amino acid frequencies show there exists some underlying differences between polyreactive and non-polyreactive antibodies. A surface level

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analysis of this dataset is able to discriminate certain features of polyreactive and non-polyreactive antibodies, namely that on average, polyreactive antibodies are less strongly negatively charged, less hydrophilic, and have a higher prevalence of antibodies with longer CDR loops of the heavy chain. Importantly, however, these binding surfaces do not have a net positive charge nor are they net hydrophobic.

To dig deeper into the biophysical differences between polyreactive and non-polyreactive antibodies, we created an adaptable software for the automated analysis of large antibody datasets and the application of a new analysis pipeline for the study of polyreactive antibodies. Overall, the improvements of this software to the current state of antibody sequence analysis are sufficient to highlight key differences in the two populations with improved spatial resolution. The position sensitive sequence alignment is able to further parse through the genetic differences and show that in general, polyreactive antibodies have a tendency to have more hydrophobic residues in CDR2H, and a decreased preference for phenylalanine in CDR1H. While these observational differences provided some initial insight, a more rigorous biophysical treatment was necessary. With the addition of 62 biophysical properties analyzed using the position sensitive alignment, significant differences between the CDR3H loops in polyreactive and non-polyreactive antibodies became immediately evident, providing a more detailed depiction of the antigen binding surface of polyreactive antibodies.

These data suggest a movement towards neutrality or “inoffensive” residues in the CDR loops of polyreactive antibodies: amino acids that are neither exceptionally hydrophobic nor hydrophilic and with a net charge close to 0. Previous studies have suggested that polyreactive antibodies tend to have more hydrophobic CDR loops, such that low affinity Van der Waals interactions might be the primary means of polyreactive interactions [16, 30]. However, these studies counted the number of hydrophobic residues per sequence or averaged the hydrophobicity of all six CDR loops. While our results partially agree with these previous findings, our analysis extends much further into defining the biophysical basis of this phenomenon. For example, while our position sensitive representation of the sequences shows that CDR3H does become more hydrophobic in polyreactive sequences, it is still net hydrophilic on average. A highly hydrophobic binding surface would provide an avenue for non-specific interactions with other hydrophobic proteins, but it would occlude binding to highly hydrophilic ligands like DNA. A slightly hydrophilic, neutral-charged binding surface would permit weak interactions with a wide range of ligands.

Using these and other biophysical properties as input feature vectors, we were able to generate a generalizable protocol for binary comparisons between two distinct populations of Ig-domain se-

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

quences. This framework is able to successfully split all tested polyreactive and non-polyreactive antibody datasets. Care was taken to not overfit these data and a preliminary classifier built from this algorithm was able to identify the proper number of input vectors for each LDA application. While there are general features which best split the polyreactive and non-polyreactive antibodies in these datasets, including charge, hydrophobicity, and beta sheet propensity, these features alone are not sufficient to discriminate between the two populations. Instead, 75 vectors taken from the position-sensitive biophysical property matrix are necessary to properly split the groups, including both simple properties like charge, hydrophobicity, flexibility, and bulkiness and more carefully curated properties like the often used Kidera factors and the hotspot detecting variables of Liu et. al [39, 40, 54]. The inability to arrive at a core few biophysical properties that could effectively distinguish polyreactive and non-polyreactive antibodies necessitated the application of further approaches, namely information theory.

The tools provided by information theory proved to be effective in the present study. The classic approach to information theory considers some input, communication of this input across a noisy channel, and then reception of a meaningful message from the resultant output. We can think of the analogous case for these antibodies, whereby the sequence and structure of the antibodies can be seen as our input, the thermal noise inherent to biological systems can complicate biochemical interactions, and the necessary output is antigen recognition, i.e. binding between the antibody and the ligand. Focusing just on the antibody side of this communication channel, we determined the underlying loop diversity through the Shannon entropy of the polyreactive and non-polyreactive datasets. This diversity was found to be nearly equivalent while the mutual information, a metric of “crosstalk” across populations, between and within CDR loops was found to be increased in the heavy chain and decreased in the light chain of polyreactive antibodies. What this loop crosstalk entails physically is not immediately clear from these measurements.

The mutual information increase could come from gene usage being somehow coupled, amino acid usage coupling with the cognate ligand, or the amino acids directly interacting physically with each other. In some way, this crosstalk appears to be selected for in the polyreactive population. If this increase in mutual information manifests as an increase of charge-charge interactions, this could explain why there is a minimal change in net charge of antibodies between the two groups, yet a significant move towards neutrality in the CDR loops of polyreactive antibodies. The pairing of two charged groups would help move the binding surface of polyreactive antibodies towards a more “inoffensive” binding surface. A binding surface that is neither exceptionally hydrophobic nor hydrophilic, and lacks a significant positive or negative charge, would represent a relatively appealing binding interface for a low-affinity interaction with a large array of diverse ligands. A patchwork

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of hydrophobic and hydrophilic non-charged residues exposed to potential ligands would represent an ideal candidate polyreactive surface. The corresponding decrease in the mutual information between the light chain CDR loops of polyreactive antibodies could be caused by a de-emphasis in the involvement of these loops due to differential binding configurations of polyreactive ligands, as has been previously hypothesized [4, 55].

In addition to the insights into polyreactivity, the computational tools developed for this study are broadly applicable to future studies of large antibody or T cell receptor repertoires. One of the strengths of this approach is a decreased emphasis on structural information when crystal structures are unavailable. Computational prediction of loop conformation is difficult, and drawing inferences from incorrect models regarding side-chain interactions and positioning could be misleading. Reliable structural information on these polyreactive antibodies will be critical to a further understanding of the mechanisms of polyreactivity, including complex structures of antibodies bound to various ligands. In the high-throughput analysis of antibody sequences, our approach strikes a careful balance of the structural assumptions that should apply consistently across antibody populations.

This streamlined analysis allows for the generation of each figure in this study to be applied to thousands of sequences in a matter of minutes. The classification capabilities of the software could prove particularly useful when comparing binary classes, such as T cell receptors or antibody sequences derived from healthy and diseased tissue samples. To demonstrate this broad applicability, a database of nearly 500 therapeutic antibodies was analyzed using the linear analysis module of the software. This linear analysis highlighted the differences between polyreactivity of therapeutic antibodies and naturally derived antibodies. When applying this linear analysis to split approved and discontinued therapeutics, the biophysical property differences were less stark than those between polyreactive and non-polyreactive antibodies. This makes intuitive sense, as therapeutics can be discontinued for a myriad of reasons, not necessarily due just to non-specificity or instability of the antibody.

Those therapeutic antibodies that were tested for polyreactivity appeared have little overlap with the polyreactivity of the naturally derived antibodies central to this study. This could be due to fundamental differences between the biophysical determinants of polyreactivity arising from antibodies generated *in vivo* vs *in vitro*, or could be due to experimental differences in the reporting of polyreactivity. While a single metric for polyreactivity, as is sometimes reported, is convenient, information on the binding of each sequence to all tested ligands is important. It is not necessarily obvious a higher average ELISA score corresponds to increased polyreactivity. Is an antibody that

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

binds to three targets with high affinity more polyreactive than one that binds to seven ligands with somewhat lower affinity? These nuances require as much transparency as possible when reporting experimental results.

Further experimental assays will be necessary to more comprehensively identify the underlying mechanisms of polyreactivity, including further sequencing and biochemical analysis of polyreactive and non-polyreactive antibodies. Antibodies specific to other pathogens or those from other organisms tested for polyreactivity will help form a more complete picture and improve the generality of the results. As with any machine learning based approach, the classification algorithm is only as good as the data it is trained on. Adding further data in the training set, including more mutations and germline reversions that turn a polyreactive antibody non-polyreactive or vice-versa, will be critical for a comprehensive analysis of polyreactivity. Additionally, a more complete understanding of the germinal center and the selection processes inherent to the affinity maturation process will assist in the determination of whether polyreactivity is a byproduct or a purposeful feature of the affinity maturation process.

The software generated for this study is publicly available as a python application (see Methods). The unique aspect of this software is its hybrid approach to position-sensitive amino acid sequence analysis. Structural information is implicitly encoded by the alignment strategy employed, yet these assumptions are weaker than those imposed by explicit structural prediction. Downstream analysis from this positional encoder is streamlined and can be generalized to analyze any binary or higher order classification problems. Acceptable inputs are not restricted to CDR loops of immunoglobulins, and in fact the software has already been adapted for analyzing MHC-like molecules (data not shown). This software represents a strong addition to the existing toolkit for repertoire analysis of diverse molecular species.

Methods

Software

All analysis was performed in python, with code tested and finalized using Jupyter Notebooks [56]. Figures were generated with matplotlib [57] or seaborn [58], while the majority of data analysis was carried out using Pandas [59], SciPy [60], and SciKit-learn [61]. All code will become available at https://github.com/ctboughter/AIMS upon publication, including the original Jupyter Notebooks used to generate the data in this manuscript as well as generalized versions for analysis of novel datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Statistical Analysis

Error bars in all plots are provided by the standard deviation of 1000 bootstrap iterations. Statistical significance is calculated using either a two-sided nonparametric Studentized bootstrap or a two-sided nonparametric permutation test as outlined in "Bootstrap Methods and Their Application" [62]. For the Studentized bootstrap, the bootstrapped data are drawn from a resampling of the empirical distributions of each respective group with replacement. Practically, what this entails is a separation of the polyreactive and non-polyreactive antibodies into distinct matrices and using the Scikit-learn resample module to preserve the number of sequences in each population. From these resampled populations, all of the relevant properties used in this study were re-calculated. These 1000 iterations of each property were then compared to the empirical distribution to calculate a p-value using the relation:

$$p = \frac{1 + \sharp(z^2 \geq z_0^2)}{R+1} \tag{4}$$

Here, we calculate the p-value by counting the number of bootstrap iterations where z^2 is greater than or equal to z_0^2. z^2 and z_0^2 are Studentized test statistics taken from the bootstrap and empirical and distributions, respectively. R is the number of times this bootstrapping process is repeated. The general form of z is given by:

$$z = \frac{\bar{Y}_2 - \bar{Y}_1 - (\mu_2 - \mu_1)}{(\frac{\sigma_2^2}{n_2} - \frac{\sigma_1^1}{n_1})^{1/2}} \tag{5}$$

Where $\bar{Y}$ represents the bootstrapped sample mean, μ is the observed sample mean from the original data, σ is the bootstrapped sample standard deviation, and n is the number of samples. Sample 1 and 2 in this case correspond to polyreactive and non-polyreactive antibodies. To calculate z for the empirical distribution (z_0), the $\bar{Y}$ terms are set to 0 and all other values correspond to the empirical rather than bootstrapped values.

To calculate p-values for differences in mutual information, the permutation test was used rather than the Studentized bootstrap. Here, the test statistic t is set to a simple difference of means, and rather than sampling with replacement from the empirical distribution, we randomly permute the data into "polyreactive" or "non-polyreactive" bins. We then count the number of permutations where the randomly permuted test statistic is greater than or equal to the empirical test statistic. This count then replaces the count ($\sharp$) in the above equation for p.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

This work was supported by the National Science Foundation through grant MCB-1517221 (BR and CTB), NIH NIBIB Training Grant T32 EB009412 (CTB), and NIH NIAID R01 AI115471, R01 AI147954 (EJA and CTB). We especially want to thank Cyril Planchais, Hugo Mouquet, and Michel Nussenzweig, for their assistance in procuring many of the antibody sequences used in this work. CTB would like to acknowledge the University of Chicago's Research Computing Center Midway resource, where some of these calculations were carried out, as well as Kristof Nolan, Caitlin Castro, Ryan Duncombe, and Nabil Faruk for their insightful discussions of the research.

Competing Interests

The authors declare no competing interests.

References

[1] Gabriel D. Victora and Michel C. Nussenzweig. Germinal Centers. *Annual Review of Immunology*, 2012.

[2] Herman N. Eisen and Gregory W. Siskind. Variations in Affinities of Antibodies during the Immune Response. *Biochemistry*, 1964.

[3] D. McKean, K. Huppi, M. Bell, L. Staudt, W. Gerhard, and M. Weigert. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, 1984.

[4] Jordan D. Dimitrov, Cyril Planchais, Lubka T. Roumenina, Tchavdar L. Vassilev, Srinivas V. Kaveri, and Sebastien Lacroix-Desmazes. Antibody Polyreactivity in Health and Disease: Statu Variabilis. *The Journal of Immunology*, 2013.

[5] Adrian F. Ochsenbein, Thomas Fehr, Claudia Lutz, Mark Suter, Frank Brombacher, Hans Hengartner, and Rolf M. Zinkernagel. Control of early viral and bacterial distribution and disease by natural antibodies. *Science*, 1999.

[6] Hedda Wardemann, Sergey Yurasov, Anne Schaefer, James W. Young, Eric Meffre, and Michel C. Nussenzweig. Predominant autoantibody production by early human B cell precursors. *Science*, 2003.

[7] Thomas Tiller, Makoto Tsuiji, Sergey Yurasov, Klara Velinzon, Michel C. Nussenzweig, and Hedda Wardemann. Autoreactivity in Human IgG+ Memory B Cells. *Immunity*, 2007.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[8] Hugo Mouquet, Johannes F. Scheid, Markus J. Zoller, Michelle Krogsgaard, Rene G. Ott, Shetha Shukair, Maxim N. Artyomov, John Pietzsch, Mark Connors, Florencia Pereyra, Bruce D. Walker, David D. Ho, Patrick C. Wilson, Michael S. Seaman, Herman N. Eisen, Arup K. Chakraborty, Thomas J. Hope, Jeffrey V. Ravetch, Hedda Wardemann, and Michel C. Nussenzweig. Polyreactivity increases the apparent affinity of anti-HIV antibodies by heteroligation. *Nature*, 2010.

[9] Julie Prigent, Valérie Lorin, Ayrin Kök, Thierry Hieu, Salomé Bourgeau, and Hugo Mouquet. Scarcity of autoreactive human blood IgA+ memory B cells. *European Journal of Immunology*, 2016.

[10] Kristi Koelsch, Nai Ying Zheng, Qingzhao Zhang, Andrew Duty, Christina Helms, Melissa D. Mathias, Mathew Jared, Kenneth Smith, J. Donald Capra, and Patrick C. Wilson. Mature B cells class switched to IgD are autoreactive in healthy individuals. *Journal of Clinical Investigation*, 2007.

[11] Jeffrey J. Bunker, Steven A. Erickson, Theodore M. Flynn, Carole Henry, Jason C. Koval, Marlies Meisel, Bana Jabri, Dionysios A. Antonopoulos, Patrick C. Wilson, and Albert Bendelac. Natural polyreactive IgA antibodies coat the intestinal microbiota. *Science*, 2017.

[12] Cyril Planchais, Ayrin Kök, Alexia Kanyavuz, Valérie Lorin, Timothée Bruel, Florence Guivel-Benhassine, Tim Rollenske, Julie Prigent, Thierry Hieu, Thierry Prazuck, Laurent Lefrou, Hedda Wardemann, Olivier Schwartz, Jordan D. Dimitrov, Laurent Hocqueloux, and Hugo Mouquet. HIV-1 Envelope Recognition by Polyreactive and Cross-Reactive Intestinal B Cells. *Cell Reports*, 2019.

[13] Barton F. Haynes, Judith Fleming, E. William St. Clair, Herman Katinger, Gabriela Stiegler, Renate Kunert, James Robinson, Richard M. Scearce, Kelly Plonk, Herman F. Staats, Thomas L. Ortel, Hua Xin Liao, and S. Munir Alam. Immunology: Cardiolipin polyspecific autoreactivity in two broadly neutralizing HIV-1 antibodies. *Science*, 2005.

[14] Hugo Mouquet, Florian Klein, Johannes F. Scheid, Malte Warncke, John Pietzsch, Thiago Y.K. Oliveira, Klara Velinzon, Michael S. Seaman, and Michel C. Nussenzweig. Memory B cell antibodies to HIV-1 gp140 cloned from individuals infected with clade A and B viruses. *PLoS ONE*, 2011.

[15] Sarah F. Andrews, Yunping Huang, Kaval Kaur, Lyubov I. Popova, Irvin Y. Ho, Noel T. Pauli, Carole J.Henry Dunand, William M. Taylor, Samuel Lim, Min Huang, Xinyan Qu, Jane Hwei Lee, Marlene Salgado-Ferrer, Florian Krammer, Peter Palese, Jens Wrammert, Rafi Ahmed, and Patrick C. Wilson. Immune history profoundly affects broadly protective B cell responses to influenza. *Science Translational Medicine*, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[16] Julie Prigent, Annaëlle Jarossay, Cyril Planchais, Caroline Eden, Jérémy Dufloo, Ayrin Kök, Valérie Lorin, Oxana Vratskikh, Thérèse Couderc, Timothée Bruel, Olivier Schwartz, Michael S. Seaman, Oliver Ohlenschläger, Jordan D. Dimitrov, and Hugo Mouquet. Conformational Plasticity in Broadly Neutralizing HIV-1 Antibodies Triggers Polyreactivity. *Cell Reports*, 2018.

[17] Barton F. Haynes, Dennis R. Burton, and John R. Mascola. Multiple roles for HIV broadly neutralizing antibodies. *Science Translational Medicine*, 2019.

[18] Trevor A. Crowell, Donn J. Colby, Suteeraporn Pinyakorn, Carlo Sacdalan, Amélie Pagliuzza, Jintana Intasan, Khunthalee Benjapornpong, Kamonkan Tangnaree, Nitiya Chomchey, Eugène Kroon, Mark S. de Souza, Sodsai Tovanabutra, Morgane Rolland, Michael A. Eller, Dominic Paquin-Proulx, Diane L. Bolton, Andrey Tokarev, Rasmi Thomas, Hiroshi Takata, Lydie Trautmann, Shelly J. Krebs, Kayvon Modjarrad, Adrian B. McDermott, Robert T. Bailer, Nicole Doria-Rose, Bijal Patel, Robert J. Gorelick, Brandie A. Fullmer, Alexandra Schuetz, Pornsuk V. Grandin, Robert J. O'Connell, Julie E. Ledgerwood, Barney S. Graham, Randall Tressler, John R. Mascola, Nicolas Chomont, Nelson L. Michael, Merlin L. Robb, Nittaya Phanuphak, Jintanat Ananworanich, Julie A. Ake, Siriwat Akapirat, Meera Bose, Evan Cale, Phillip Chan, Sararut Chanthaburanun, Nampueng Churikanont, Peter Dawson, Netsiri Dumrongpisutikul, Saowanit Getchalarat, Surat Jongrakthaitae, Krisada Jongsakul, Sukalaya Lerdlum, Sopark Manasnayakorn, Corinne McCullough, Mark Milazzo, Bessara Nuntapinit, Kier On, Madelaine Ouellette, Praphan Phanuphak, Eric Sanders-Buell, Nongluck Sangnoi, Shida Shangguan, Sunee Sirivichayakul, Nipattra Tragonlugsana, Rapee Trichavaroj, Sasiwimol Ubolyam, Sandhya Vasan, Phandee Wattanaboonyongcharoen, and Thipvadee Yamchuenpong. Safety and efficacy of VRC01 broadly neutralising antibodies in adults with acutely treated HIV (RV397): a phase 2, randomised, double-blind, placebo-controlled trial. *The Lancet HIV*, 2019.

[19] Gui Mei Li, Christopher Chiu, Jens Wrammert, Megan McCausland, Sarah F. Andrews, Nai Ying Zheng, Jane Hwei Lee, Min Huang, Xinyan Qu, Srilatha Edupuganti, Mark Mulligan, Suman R. Das, Jonathan W. Yewdell, Aneesh K. Mehta, Patrick C. Wilson, and Rafi Ahmed. Pandemic H1N1 influenza vaccine induces a recall response in humans that favors broadly cross-reactive memory B cells. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[20] Joshua S. Klein and Pamela J. Bjorkman. Few and far between: How HIV may be evading antibody avidity. *PLoS Pathogens*, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[21] Isidro Hötzel, Frank Peter Theil, Lisa J. Bernstein, Saileta Prabhu, Rong Deng, Leah Quintana, Jeff Lutman, Renuka Sibia, Pamela Chan, Daniela Bumbaca, Paul Fielder, Paul J. Carter, and Robert F. Kelley. A strategy for risk mitigation of antibodies with fast clearance. *mAbs*, 2012.

[22] Ryan L. Kelly, Tingwan Sun, Tushar Jain, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Maximiliano Vásquez, K. Dane Wittrup, and Yingda Xu. High throughput cross-interaction measures for human IgG1 antibodies correlate with clearance rates in mice. *mAbs*, 2015.

[23] Ryan L. Kelly, Doris Le, Jessie Zhao, and K. Dane Wittrup. Reduction of Nonspecificity Motifs in Synthetic Antibody Libraries. *Journal of Molecular Biology*, 2018.

[24] Amita Datta-Mannan, Jirong Lu, Derrick R. Witcher, Donmienne Leung, Ying Tang, and Victor J. Wroblewski. The interplay of non-specific binding, targetmediated clearance and FcRn interactions on the pharmacokinetics of humanized antibodies. *mAbs*, 2015.

[25] Tushar Jain, Tingwan Sun, Stéphanie Durand, Amy Hall, Nga Rewa Houston, Juergen H. Nett, Beth Sharkey, Beata Bobrowicz, Isabelle Caffry, Yao Yu, Yuan Cao, Heather Lynaugh, Michael Brown, Hemanta Baruah, Laura T. Gray, Eric M. Krauland, Yingda Xu, Maximiliano Vásquez, and K. Dane Wittrup. Biophysical properties of the clinical-stage antibody landscape. *Proceedings of the National Academy of Sciences of the United States of America*, 2017.

[26] Matthew I.J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 2019.

[27] Vikas K. Sharma, Thomas W. Patapoff, Bruce Kabakoff, Satyan Pai, Eric Hilario, Boyan Zhang, Charlene Li, Oleg Borisov, Robert F. Kelley, Ilya Chorny, Joe Z. Zhou, Ken A. Dill, and Trevor E. Swartz. In silico selection of therapeutic antibodies for development: Viscosity, clearance, and chemical stability. *Proceedings of the National Academy of Sciences of the United States of America*, 2014.

[28] Tushar Jain, Todd Boland, Asparouh Lilov, Irina Burnina, Michael Brown, Yingda Xu, and Maximiliano Vásquez. Prediction of delayed retention of antibodies in hydrophobic interaction chromatography from sequence using machine learning. *Bioinformatics*, 2017.

[29] Olga Obrezanova, Andreas Arnell, Ramón Gómez De La Cuesta, Maud E. Berthelot, Thomas R.A. Gallagher, Jesús Zurdo, and Yvette Stallwood. Aggregation risk prediction for antibodies and its application to biotherapeutic development. *mAbs*, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[30] Charles G. Starr and Peter M. Tessier. Selecting and engineering monoclonal antibodies with drug-like specificity, 2019.

[31] Maxime Lecerf, Alexia Kanyavuz, Sébastien Lacroix-Desmazes, and Jordan D. Dimitrov. Sequence features of variable region determining physicochemical properties and polyreactivity of therapeutic antibodies. *Molecular Immunology*, 2019.

[32] Lilia A. Rabia, Yulei Zhang, Seth D. Ludwig, Mark C. Julian, and Peter M. Tessier. Net charge of antibody complementarity-determining regions is a key predictor of specificity. *Protein engineering, design & selection : PEDS*, 2018.

[33] Sara Birtalan, Yingnan Zhang, Frederic A. Fellouse, Lihua Shao, Gabriele Schaefer, and Sachdev S. Sidhu. The Intrinsic Contributions of Tyrosine, Serine, Glycine and Arginine to the Affinity and Specificity of Antibodies. *Journal of Molecular Biology*, 2008.

[34] Yasaman Karami, Julien Rey, Guillaume Postic, Samuel Murail, Pierre Tufféry, and Sjoerd J. De Vries. DaReUS-Loop: a web server to model multiple loops in homology models. *Nucleic Acids Research*, 2019.

[35] Karlynn E. Neu, Jenna J. Guthmiller, Min Huang, Jennifer La, Marcos C. Vieira, Kangchon Kim, Nai Ying Zheng, Mario Cortese, Micah E. Tepora, Natalie J. Hamel, Karla Thatcher Rojas, Carole Henry, Dustin Shaw, Charles L. Dulberger, Bali Pulendran, Sarah Cobey, Aly A. Khan, and Patrick C. Wilson. Spec-seq unveils transcriptional subpopulations of antibody-secreting cells following influenza vaccination. *Journal of Clinical Investigation*, 2019.

[36] Jens Wrammert, Dimitrios Koutsonanos, Gui Mei Li, Srilatha Edupuganti, Jianhua Sui, Michael Morrissey, Megan McCausland, Ioanna Skountzou, Mady Hornig, W. Ian Lipkin, Aneesh Mehta, Behzad Razavi, Carlos Del Rio, Nai Ying Zheng, Jane Hwei Lee, Min Huang, Zahida Ali, Kaval Kaur, Sarah Andrews, Rama Rao Amara, Youliang Wang, Suman Ranjan Das, Christopher David O'Donnell, Jon W. Yewdell, Kanta Subbarao, Wayne A. Marasco, Mark J. Mulligan, Richard Compans, Rafi Ahmed, and Patrick C. Wilson. Broadly cross-reactive antibodies dominate the human B cell response against 2009 pandemic H1N1 influenza virus infection. *Journal of Experimental Medicine*, 2011.

[37] Pradyot Dash, Andrew J. Fiore-Gartland, Tomer Hertz, George C. Wang, Shalini Sharma, Aisha Souquette, Jeremy Chase Crawford, E. Bridie Clemens, Thi H.O. Nguyen, Katherine Kedzierska, Nicole L. La Gruta, Philip Bradley, and Paul G. Thomas. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, 2017.

[38] Jian Ye, Ning Ma, Thomas L. Madden, and James M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic acids research*, 2013.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[39] Akinori Kidera, Yasuo Konishi, Masahito Oka, Tatsuo Ooi, and Harold A. Scheraga. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 1985.

[40] Quanya Liu, Peng Chen, Bing Wang, Jun Zhang, and Jinyan Li. Hot spot prediction in protein-protein interactions by an ensemble system. *BMC Systems Biology*, 2018.

[41] Matthew Barker and William Rayens. Partial least squares for discrimination. *Journal of Chemometrics*, 2003.

[42] Marli Tenório Cordeiro, Ulisses Braga-Neto, Rita Maria Ribeiro Nogueira, and Ernesto T.A. Marques. Reliable classifier to differentiate primary and secondary acute dengue infection based on IgG ELISA. *PLoS ONE*, 2009.

[43] Yuqian Ma, David Vilanova, Kerem Atalar, Olivier Delfour, Jonathan Edgeworth, Marlies Ostermann, Maria Hernandez-Fuentes, Sandrine Razafimahatratra, Bernard Michot, David H. Persing, Ingrid Ziegler, Bianca Törös, Paula Mölling, Per Olcén, Richard Beale, and Graham M. Lord. Genome-Wide Sequencing of Cellular microRNAs Identifies a Combinatorial Expression Signature Diagnostic of Sepsis. *PLoS ONE*, 2013.

[44] Zhihua Qiao, Lan Zhou, and Jianhua Z. Huang. Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG International Journal of Applied Mathematics*, 2009.

[45] Claude E. Shannon. The Mathematical Theory of Communication. *The Bell System Technical Journal*, 1948.

[46] Ramón Román-Roldán, Pedro Bernaola-Galván, and José L. Oliver. Application of information theory to DNA sequence analysis: A review. *Pattern Recognition*, 1996.

[47] Raymond Cheong, Alex Rhee, Chiaochun Joanne Wang, Ilya Nemenman, and Andre Levchenko. Information transduction capacity of noisy biochemical signaling networks. *Science*, 2011.

[48] Susana Vinga. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 2014.

[49] Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 2010.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[50] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences of the United States of America*, 2012.

[51] James R.R. Whittle, Ruijun Zhang, Surender Khurana, Lisa R. King, Jody Manischewitz, Hana Golding, Philip R. Dormitzer, Barton F. Haynes, Emmanuel B. Walter, M. Anthony Moody, Thomas B. Kepler, Hua Xin Liao, and Stephen C. Harrison. Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, 2011.

[52] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: The structural antibody database. *Nucleic Acids Research*, 2014.

[53] Matthew I.J. Raybould, Claire Marks, Alan P. Lewis, Jiye Shi, Alexander Bujotzek, Bruck Taddese, and Charlotte M. Deane. Thera-SAbDab: the Therapeutic Structural Antibody Database. *Nucleic acids research*, 2020.

[54] Mauno Vihinen, Esa Torkkila, and Pentti Riikonen. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, 1994.

[55] Dhruv K. Sethi, Anupriya Agarwal, Venkatasamy Manivel, Kanury V.S. Rao, and Dinakar M. Salunke. Differential Epitope Positioning within the Germline Antibody Paratope Enhances Promiscuity in the Primary Immune Response. *Immunity*, 2006.

[56] Thomas Kluyver, Benjamin Ragan-kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter Development Team. Jupyter Notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. 2016.

[57] John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 2007.

[58] Erik Ziegler, Yury V. Zaytsev, Michael T. Waskom, Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, Alistair Miles, Tom Augspurger, Tal Yarkoni, Tobias Megies, Luis Pedro Coelho, Daniel Wehner, and Michael Waskom. seaborn: v0.5.0. *zenodo*, 2014.

[59] Wes McKinney and PyData Development Team. Pandas - Powerful Python Data Analysis Toolkit. *Pandas - Powerful Python Data Analysis Toolkit*, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[60] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R.J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 2020.

[61] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[62] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. 1997.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

843 # Supplemental Figures

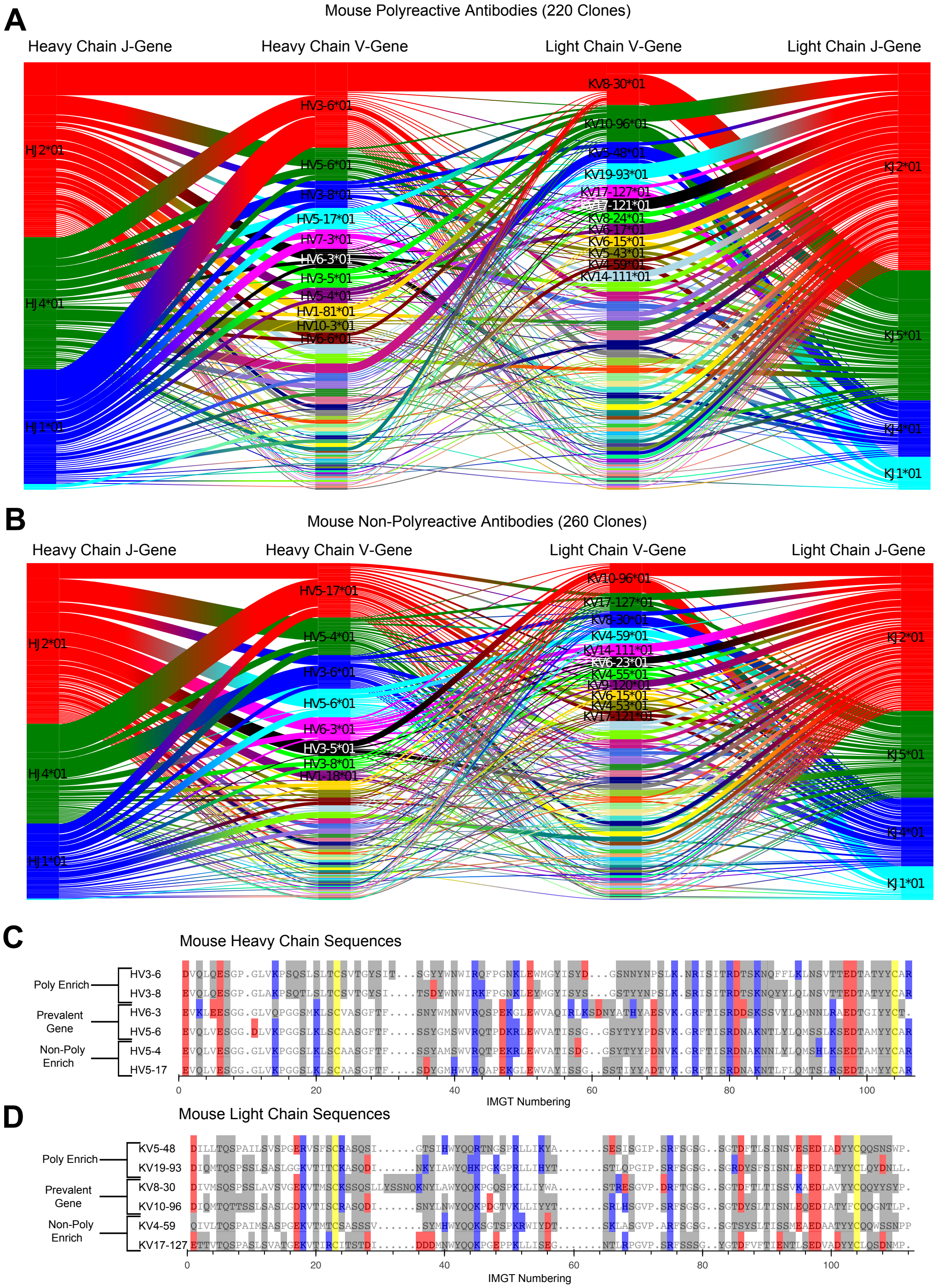


Figure S1: Gene usage plots comparing mouse polyreactive (A) and (B) non-polyreactive clones including J-gene usage. Colors represent the most commonly used genes in each individual dataset, with colors not necessarily consistent between panels. Sequence alignments comparing the amino acids of these most common genes for polyreactive and non-polyreactive mouse antibodies for the heavy chain (C) and the light chain (D). Prevalent genes are present in both populations. Cysteine is colored yellow, hydrophobic amino acids are colored white, hydrophilic amino acids are colored grey, and positively or negatively charged amino acids are colored blue or red, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

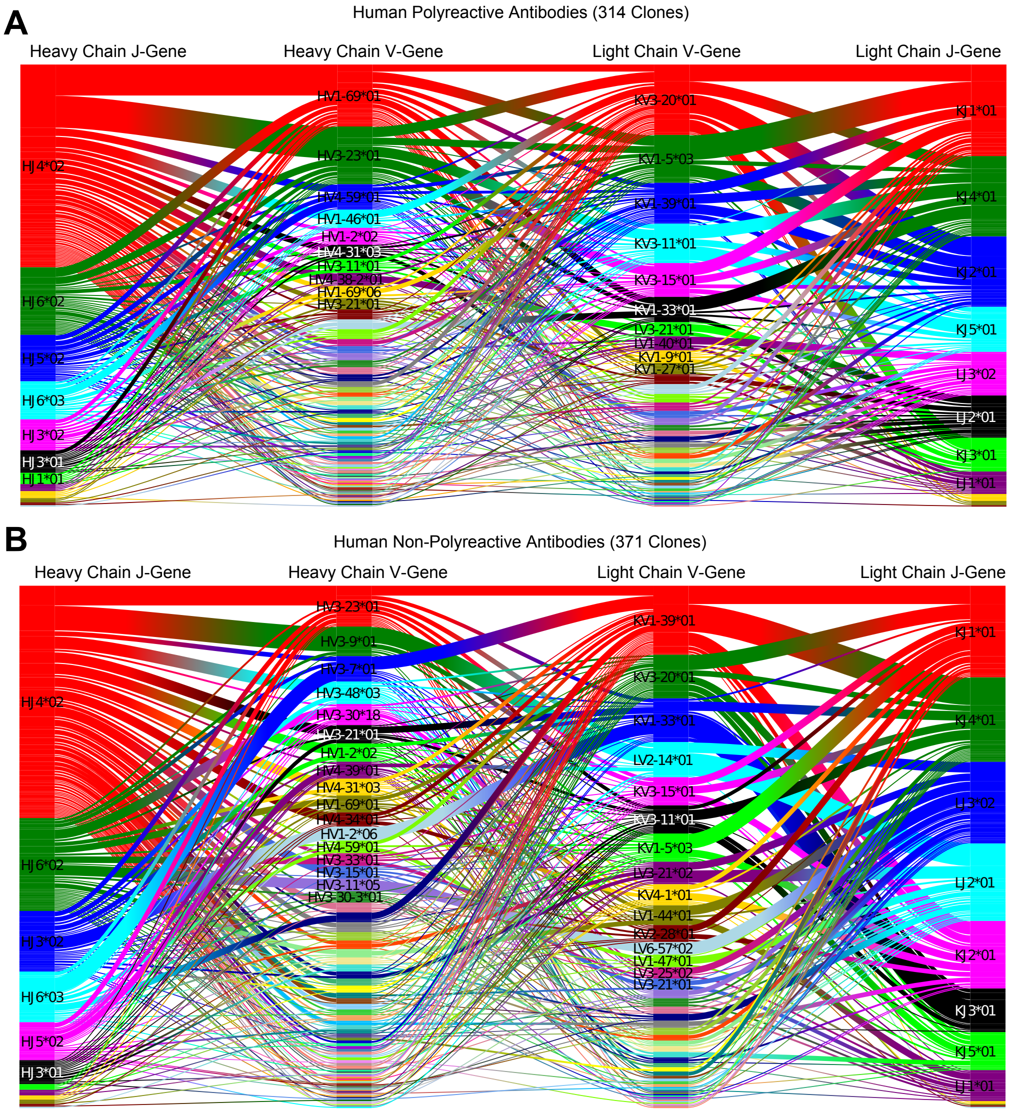


Figure S2: Gene usage plots comparing human polyreactive (A) and (B) non-polyreactive clones including J-gene usage. Data is the same as that in Figure 1A and 1B, with a different color scheme used for genes. Colors represent the most commonly used genes in each individual dataset, with colors not necessarily consistent between panels.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

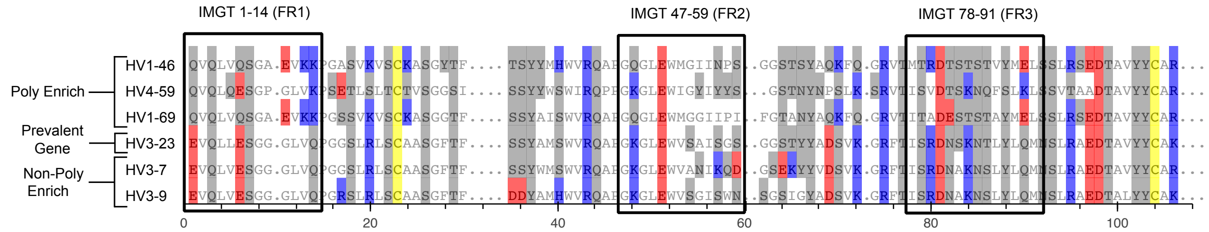

IMGT 1-14 (FR1)
IMGT 47-59 (FR2)
IMGT 78-91 (FR3)
Poly Enrich
Prevalent Gene
Non-Poly Enrich
HV1-46 QVQLVQSGA.EVKKPGASVKVSCKASGYTF....TSYYMHWVRQAPGQGLEWMGIINPS..GGSTSYAQKFQ.GRVTMTRDTSTSTVYMELSSLRSEDTAVYYCAR...
HV4-59 QVQLQESGP.GLVKPSETLSLTCTVSGGSI....SSYYWSWIRQPPGKGLEWIGYIYYS...GSTNYNPSLK.SRVTISVDTSKNQFSLKLSSVTAADTAVYYCAR...
HV1-69 QVQLVQSGA.EVKKPGSSVKVSCKASGGTF....SSYAISWVRQAPGQGLEWMGGIIPI..FGTANYAQKFQ.GRVTITADESTSTAYMELSSLRSEDTAVYYCAR...
HV3-23 EVQLLESGG.GLVQPGGSLRLSCAASGFTF....SSYAMSWVRQAPGKGLEWVSAISGS..GGSTYYADSVK.GRFTISRDNSKNTLYLQMNSLRAEDTAVYYCAK...
HV3-7 EVQLVESGG.GLVQPGGSLRLSCAASGFTF....SSYWMSWVRQAPGKGLEWVANIKQD..GSEKYYVDSVK.GRFTISRDNAKNSLYLQMNSLRAEDTAVYYCAR...
HV3-9 EVQLVESGG.GLVQPGRSLRLSCAASGFTF....DDYAMHWVRQAPGKGLEWVSGISWN..SGSIGYADSVK.GRFTISRDNAKNSLYLQMNSLRAEDTALYYCAK...
0
20
40
60
80
100

Figure S3: Sequence alignment of the most polyreactive genes compared to the most prevalent gene and the most non-polyreactive genes. Alignment uses IMGT numbering scheme and displays the entirety of the heavy chain variable gene's amino acid sequence. Boxes represent the sections highlighted in Figure 1C. Cysteine is colored yellow, hydrophobic amino acids are colored white, hydrophilic amino acids are colored grey, and positively or negatively charged amino acids are colored blue or red, respectively.

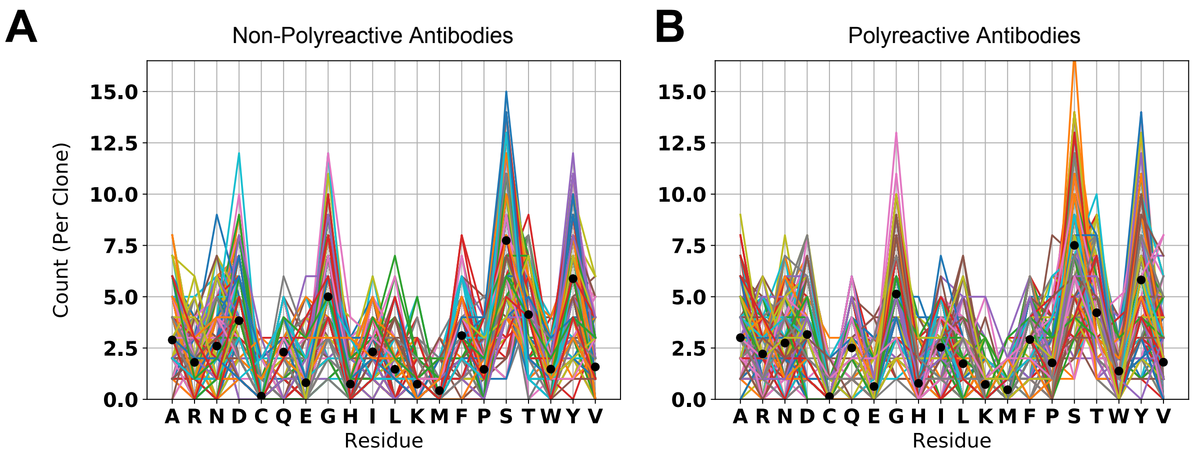


Figure S4: Amino acid usage plot highlighting the occurrence of each amino acid in non-polyreactive (A) and polyreactive (B) CDR loops. Each line represents an individual clone, and each point along the line represents the count of each amino in that given clone. Black dots represent the average counts per clone.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

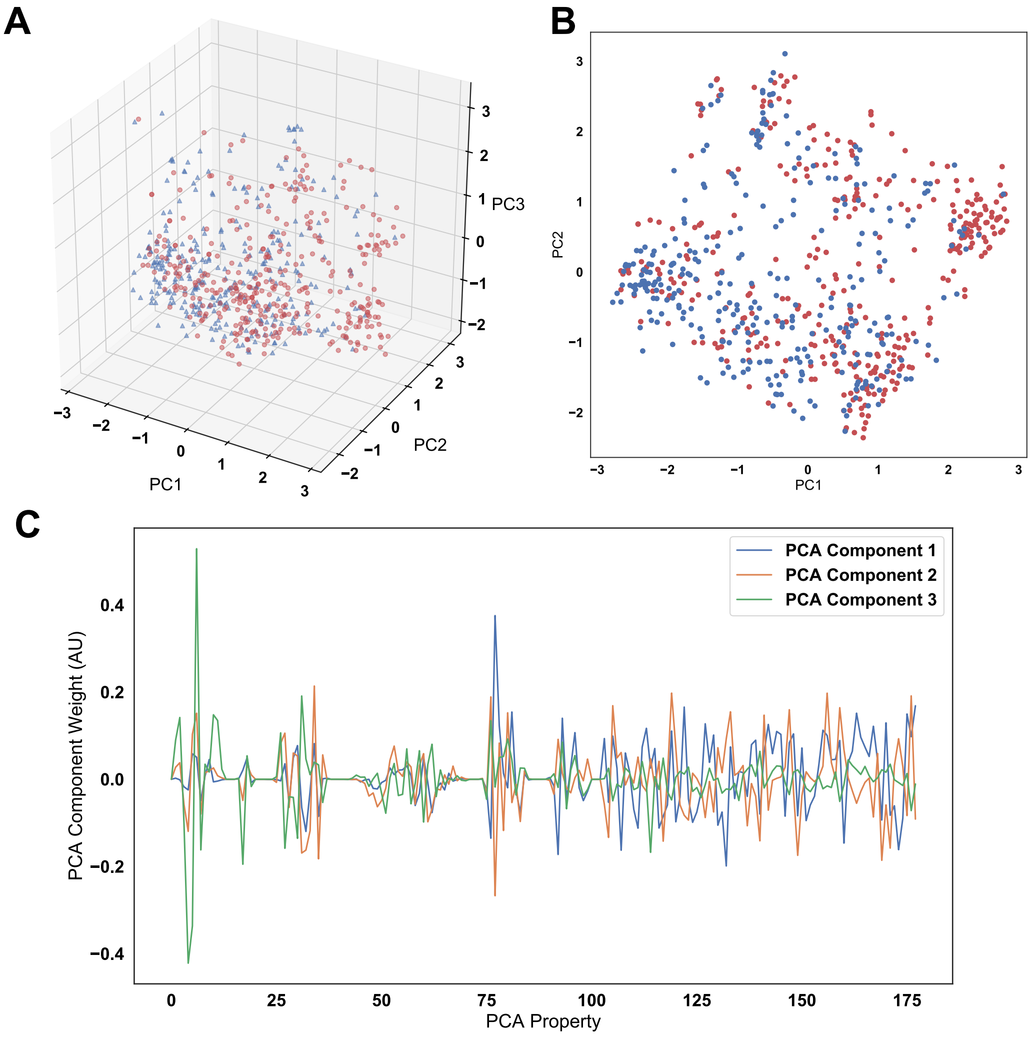


Figure S5: Principal component analysis (PCA) applied to the full amino acid usage matrix and the top 75 discriminating vectors used for linear discriminant analysis shows an inability to distinguish the two populations when showing the first three (A) and first two (B) principal components. (C) Examination of the weights of these first three components shows there is no one property disproportionately contributing to the variance in the dataset. The vector normal of each set of weights is equivalent to 1. The red dot represents the transition from the simple property-based representation of each set of CDR loops to the top 75 discriminating properties.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

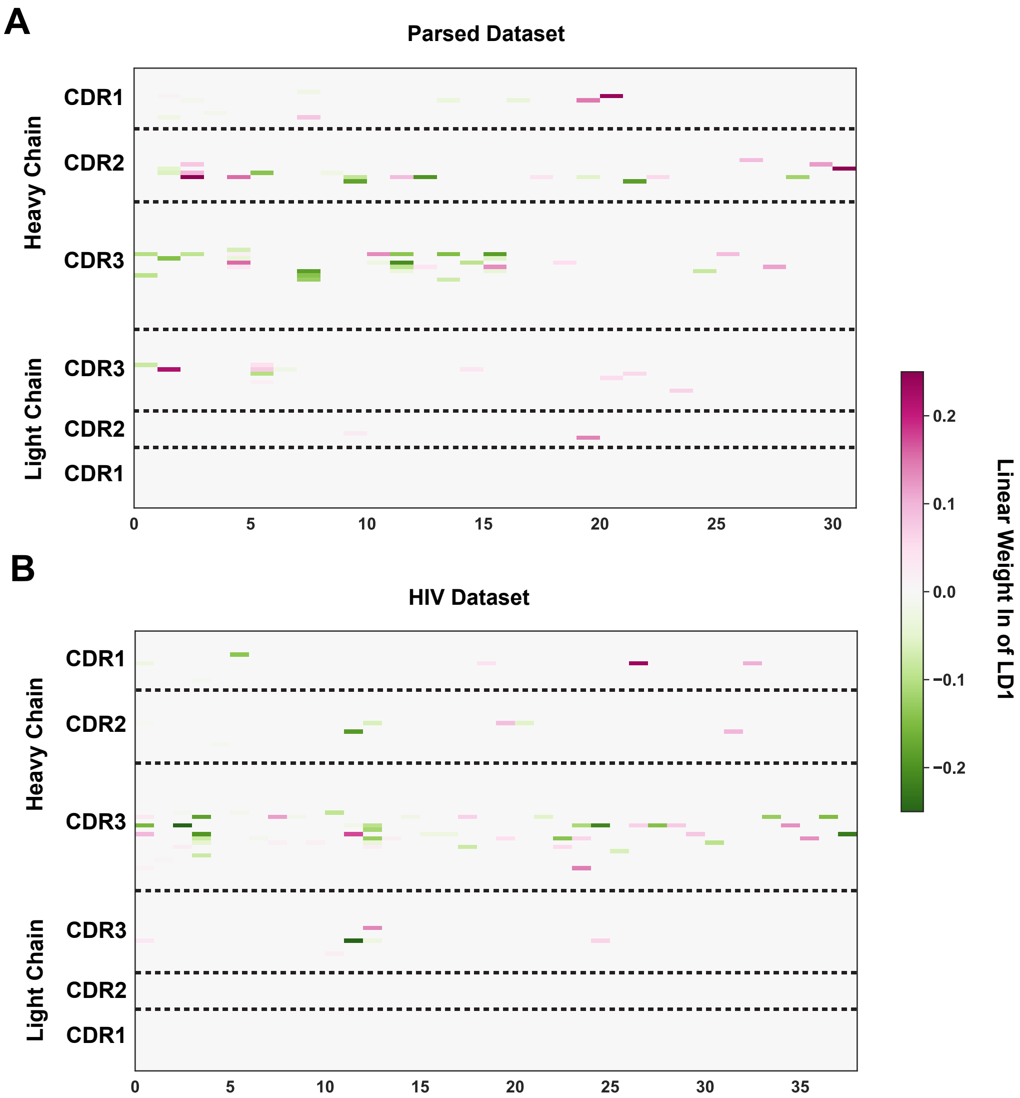


Figure S6: The complete representation of the 75 linear weights that most effectively separate polyreactive and non-polyreactive sequences in the parsed complete dataset (A) and the parsed HIV dataset (B). The x-axes each represent a single biophysical property selected after parsing down the full feature list using a maximal difference algorithm and a correlation analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

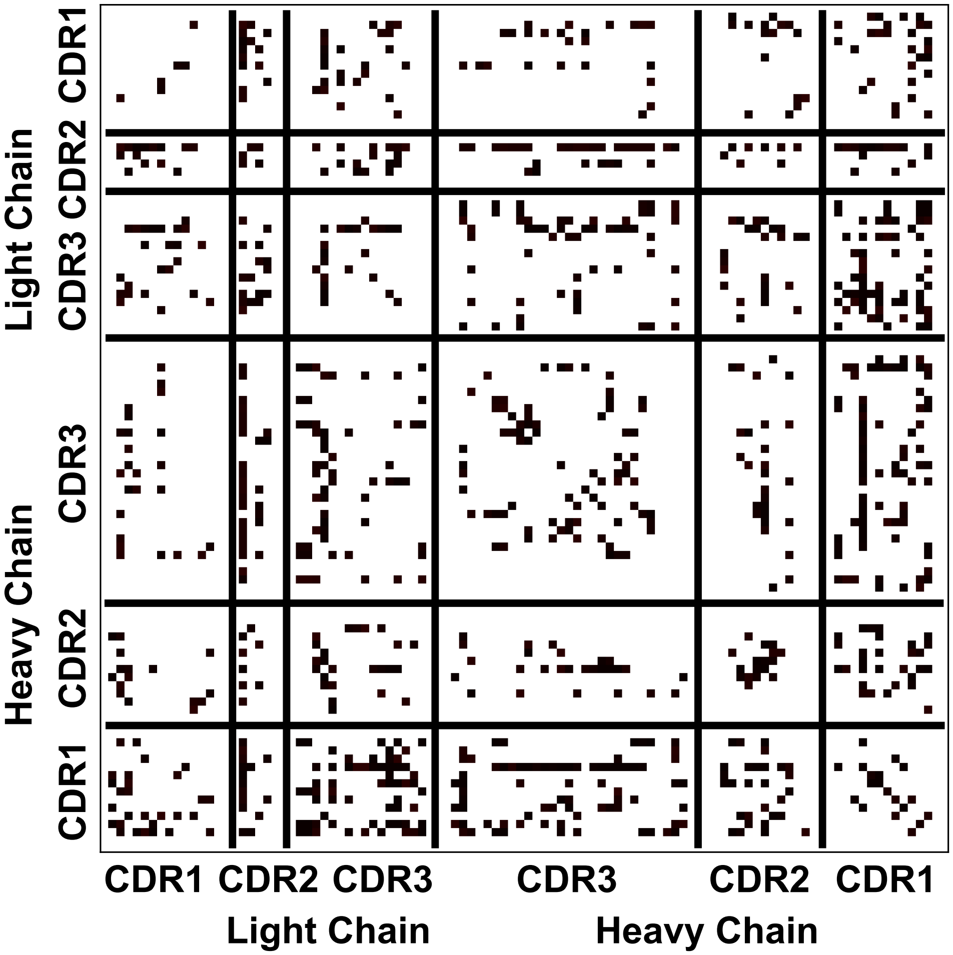


Figure S7: The statistical significance of the values reported in Figure 5C. Each black dot represents statistical significance (p $\leq$ 0.05) at that given location. Significance was calculated using a non-parametric permutation test.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

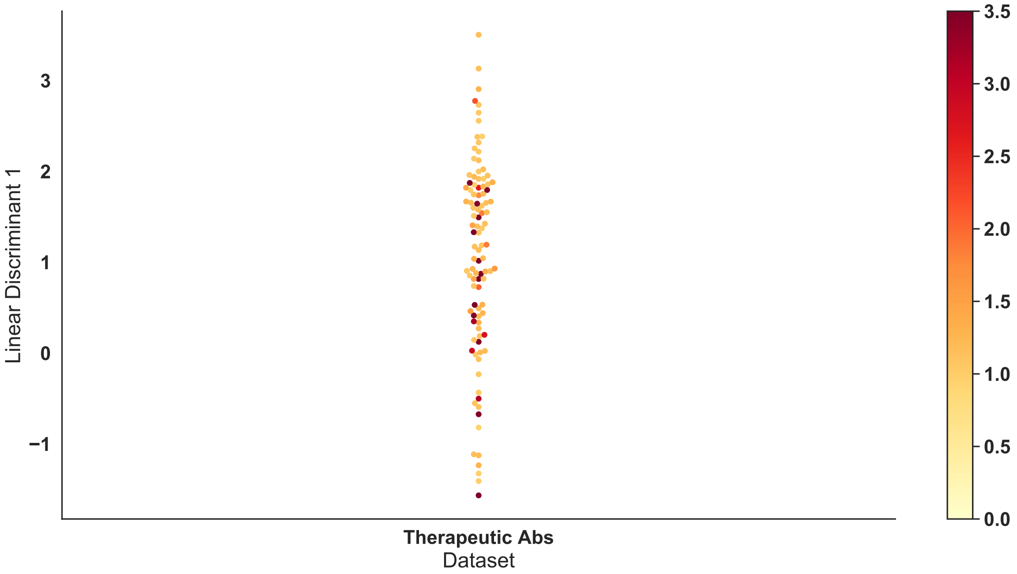


Figure S8: A mapping of the therapeutic antibodies tested for polyreactivity by Jain & Sun et. al. (PNAS 2017) onto the linear discriminant trained on the parsed dataset of naturally derived polyreactive antibodies. The linear discriminant here is identical to that in Figure 6A, while the sequences plotted above are subset of the "Therapeutic Antibodies" in that same panel. These therapeutic antibodies were tested for polyreactivity using an ELISA based assay aggregated into a single value reported in the original study. These values are represented in this plot by color, with the color bar providing the scale.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

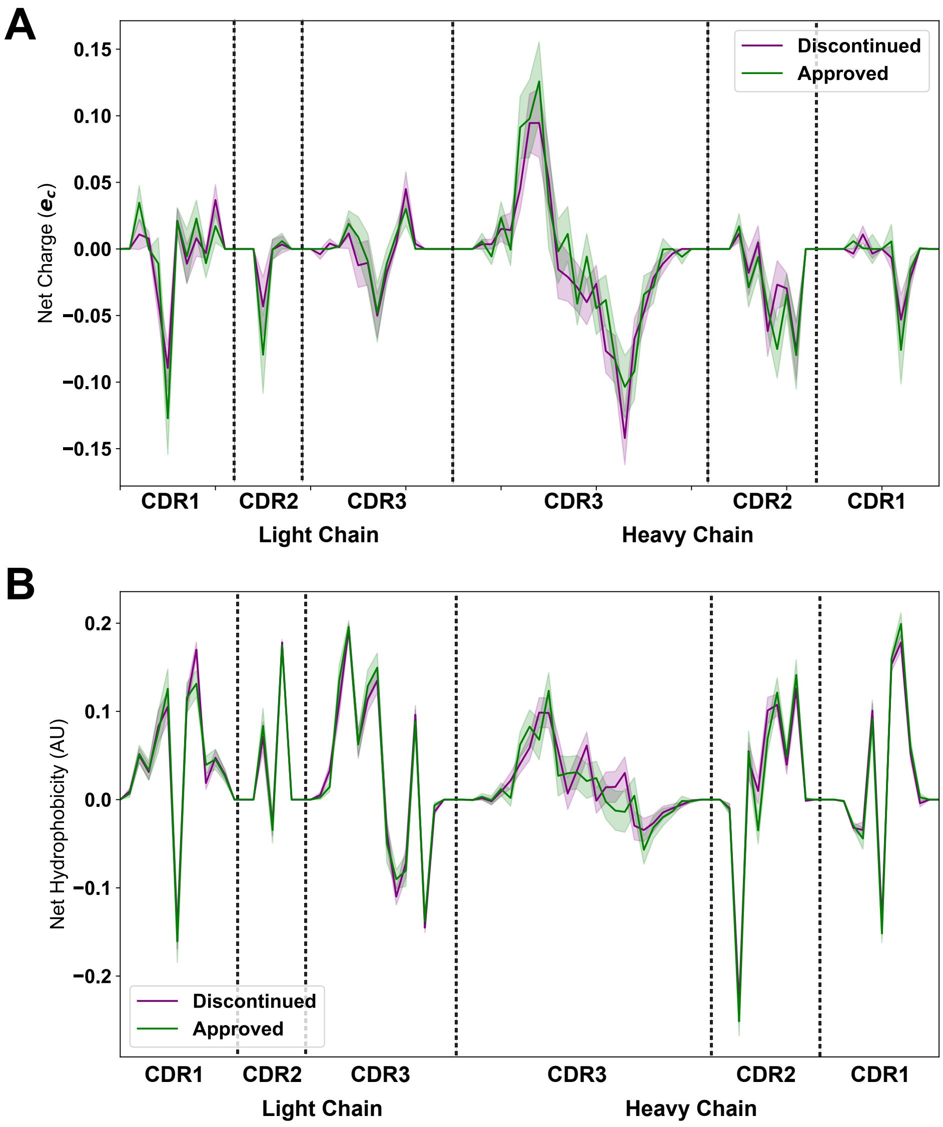


Figure S9: Plotting the average net charge (A) and net hydrophobicity (B) as a function of distance of discontinued and accepted therapeutic antibodies highlights a lack of significant differences. Light shadow around lines represents standard deviation obtained via bootstrapping.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table S1: List of all of biophysical properties used for this study. For hotspot detecting variables (HS) a simplified form of the description is used. For more in-depth descriptions, the original reference should be used.

Property Shorthand	Description
Phob1	Hydrophobicity Scale [-1,1]
Charge	Charge [ec]
Phob2	Octanol-Interface Hydrophobicity Scale
Bulk	Side-Chain Bulkiness
Flex	Side-Chain Flexibility
KD1	Helix/Bend Preference
KD2	Side-Chain Size
KD3	Extended Structure Preference
KD4	Hydrophobicity
KD5	Double-bend Preference
KD6	Flat Extended Preference
KD7	Partial Specific Volume
KD8	Occurrence in alpha-region
KD9	pK-C
KD10	Surrounding Hydrophobicity
HS1	Normalized Positional Residue Freq at Helix C-term
HS2	Normalized Positional Residue Freq at Helix C4-term
HS3	Spin-spin coupling constants
HS4	Random Parameter
HS5	pK-N
HS6	Alpha-Helix Indices for Beta-Proteins
HS7	Linker Propensity from 2-Linker Dataset
HS8	Linker Propensity from Long Dataset
HS9	Normalized Relative Freq of Helix End
HS10	Normalized Relative Freq of Double Bend
HS11	pK-COOH
HS12	Relative Mutability
HS13	Kerr-Constant Increments
HS14	Net Charge
HS15	Norm Freq Zeta-R
HS16	Hydropathy Scale
HS17	Ratio of Average Computed Composition

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.30.229013; this version posted July 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

HS18	Intercept in Regression Analysis
HS19	Correlation coefficient in Reg Anal
HS20	Weights for Alpha-Helix at window pos
HS21	Weights for Beta-sheet at window pos -3
HS22	Weights for Beta-sheet at window pos 3
HS23	Weights for coil at win pos -5
HS24	Weights coil win pos -4
HS25	Weights coil win pos 6
HS26	Avg Rel Frac occur in AL
HS27	Avg Rel Frac occur in EL
HS28	Avg Rel Frac occur in A0
HS29	Rel Pref at N
HS30	Rel Pref at N1
HS31	Rel Pref at N2
HS32	Rel Pref at C1
HS33	Rel Pref at C
HS34	Information measure for extended without H-bond
HS35	Information measure for C-term turn
HS36	Loss of SC hydropathy by helix formation
HS37	Principal Component 4 (Sneath 1966)
HS38	Zimm-Bragg Parameter
HS39	Normalized Freq of ZetaR
HS40	Rel Pop Conformational State A
HS41	Rel Pop Conformational State C
HS42	Electron-Ion Interaction Potential
HS43	Free energy change of epsI to epsEx
HS44	Free energy change of alphaRI to alphaRH
HS45	Hydrophobicity coeff
HS46	Principal Property Value z3 Wold et. al. 1987