Cell Reports, Volume *43*

Supplemental information

Human antibody polyreactivity is governed primarily by the heavy-chain complementarity-determining regions

Hsin-Ting Chen, Yulei Zhang, Jie Huang, Manali Sawant, Matthew D. Smith, Nandhini Rajagopal, Alec A. Desai, Emily Makowski, Giuseppe Licari, Yunxuan Xie, Michael S. Marlow, Sandeep Kumar, and Peter M. Tessier

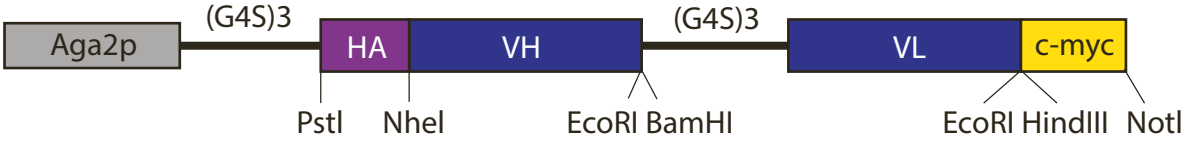


Figure S1. Schematic of yeast surface display plasmid used in this study. Related to STAR Methods. Single chain (scFv) antibodies expressed on the surface of yeast as a C-terminal fusion protein to Aga2p subunit. Aga2p: subunit of the a-agglutinin yeast cell surface anchor protein. (G4S)3: flexible linker, GGGGSGGGGSGGGGS. HA: N- terminal tag, YPYDVPDYA. VH: heavy chain, VL: light. c-myc: C-terminal tag, EQKLISEEDL.

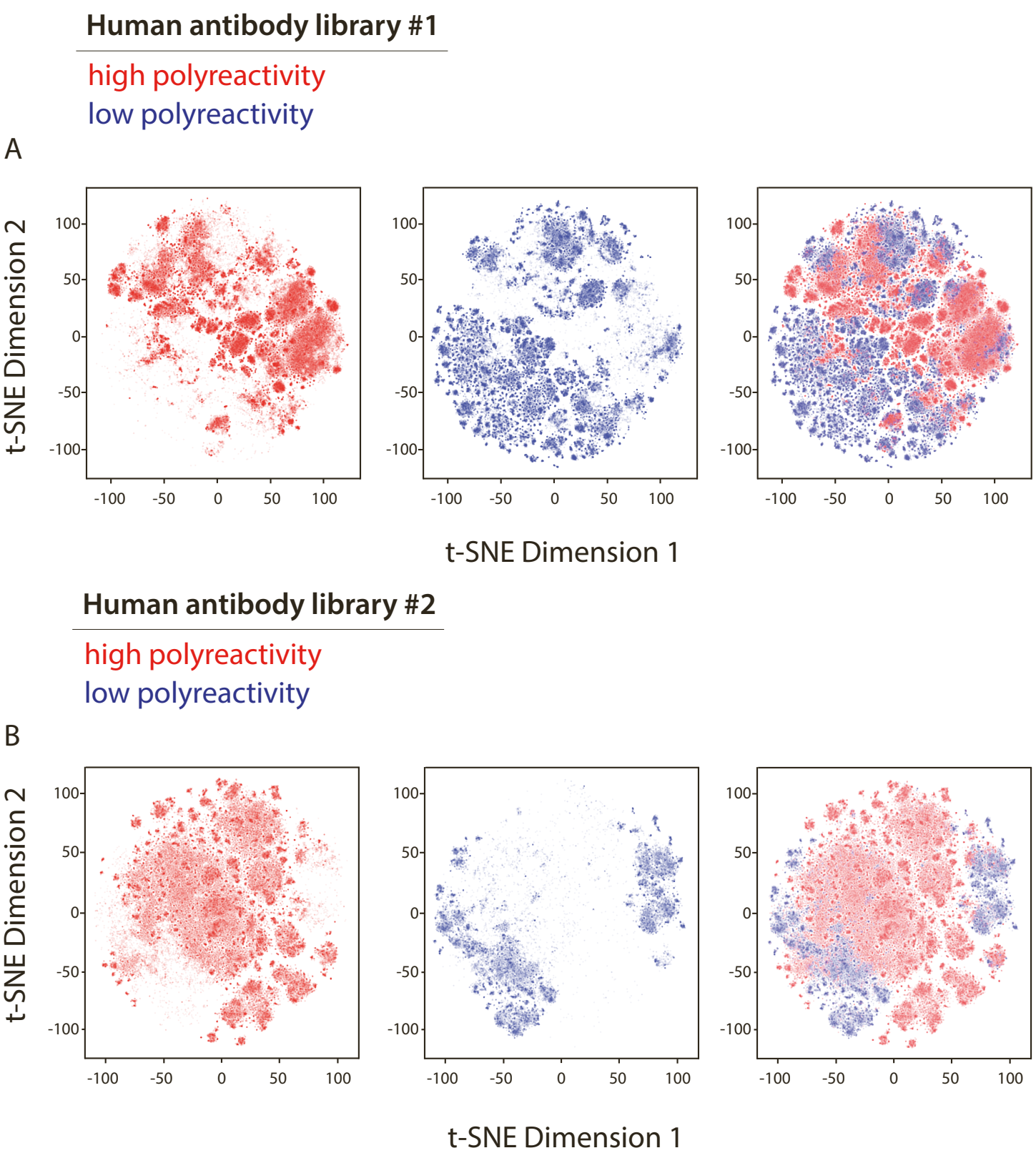
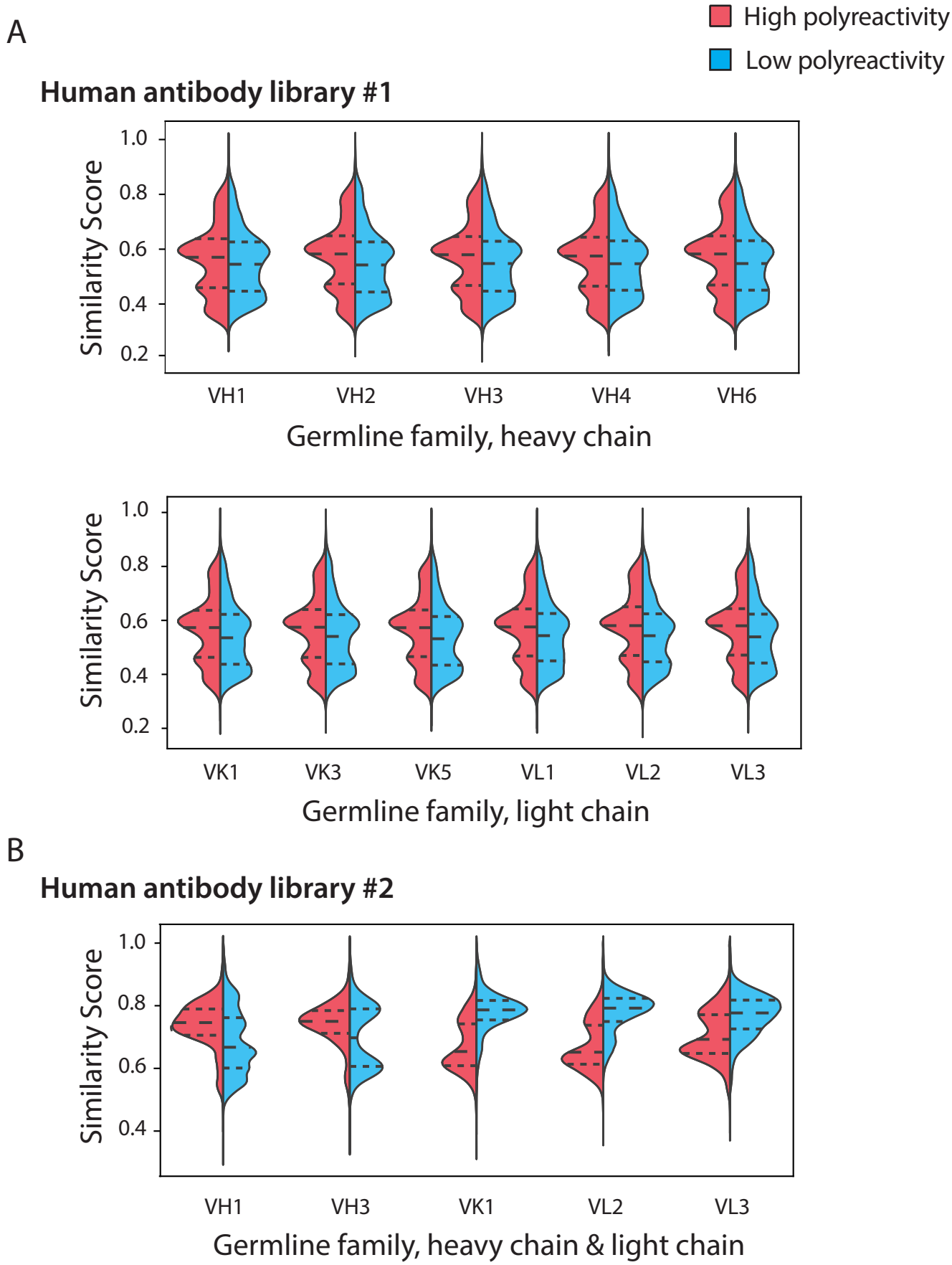


Figure S2. Analysis of sequence differences between high and low levels of polyreactivity for each two libraries using protein language model embeddings. Related to Figure 2. The sequences for (A) library #1 and (B) library #2 were represented by embeddings generated by ESM-2, subjected to dimensionality reduction via principal component analysis, and embedded into two-dimensional space for visualization with t-distributed stochastic neighbor embedding (t-SNE).



Firgure S3. The distribution of sequence similarities for high and low polyreactivity antibodies per germline family in antibody library #1 and #2. Related to Figure 2. The sequence similarity score is calculated by averaging the similarities between randomly-selected 1000 sequences from each of the most common germline families for (A) library #1 and (B) library #2. The most common germline families are shown if there are at least 400 sequences in both high and low polyreactivity antibodies.

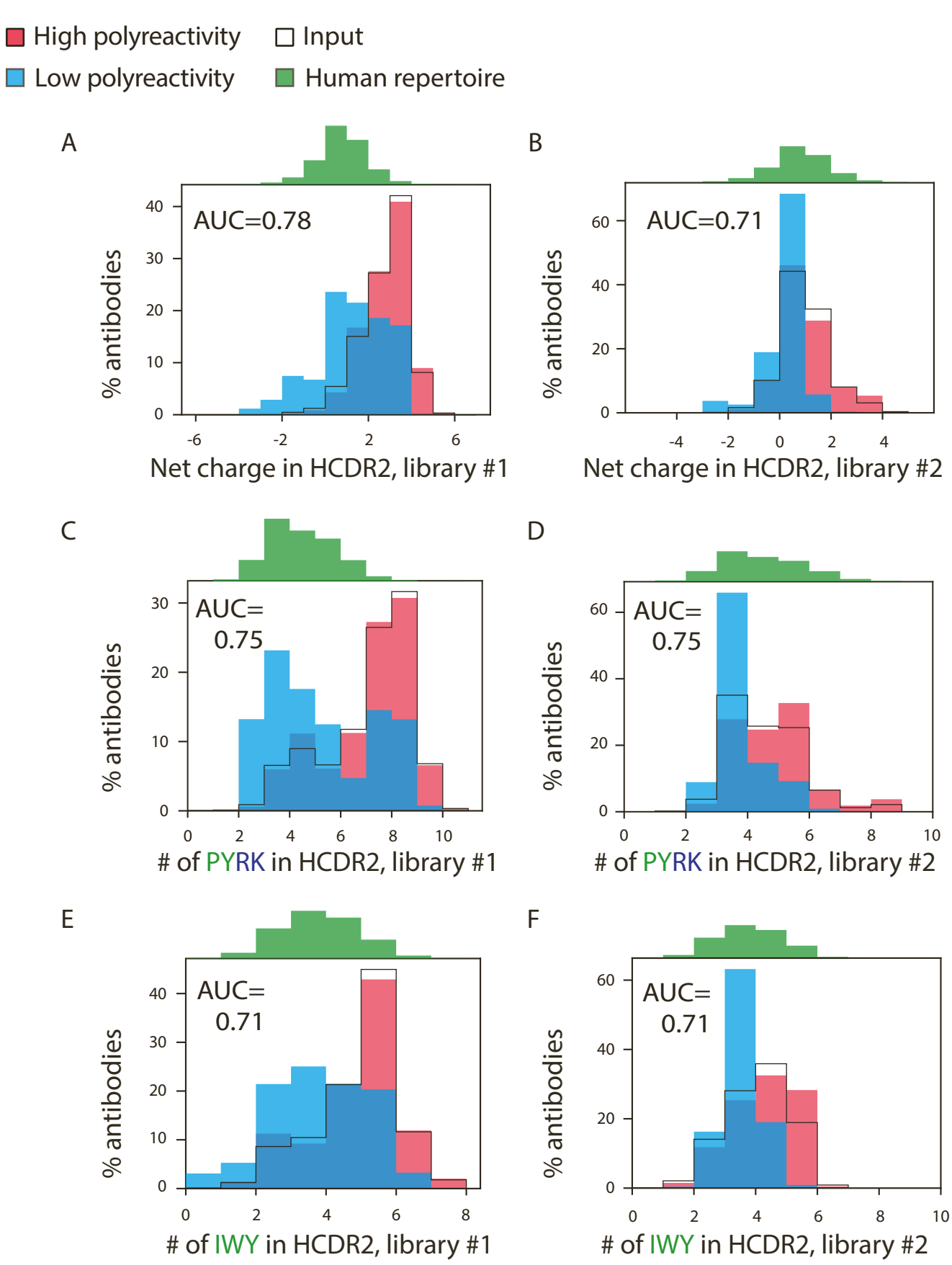
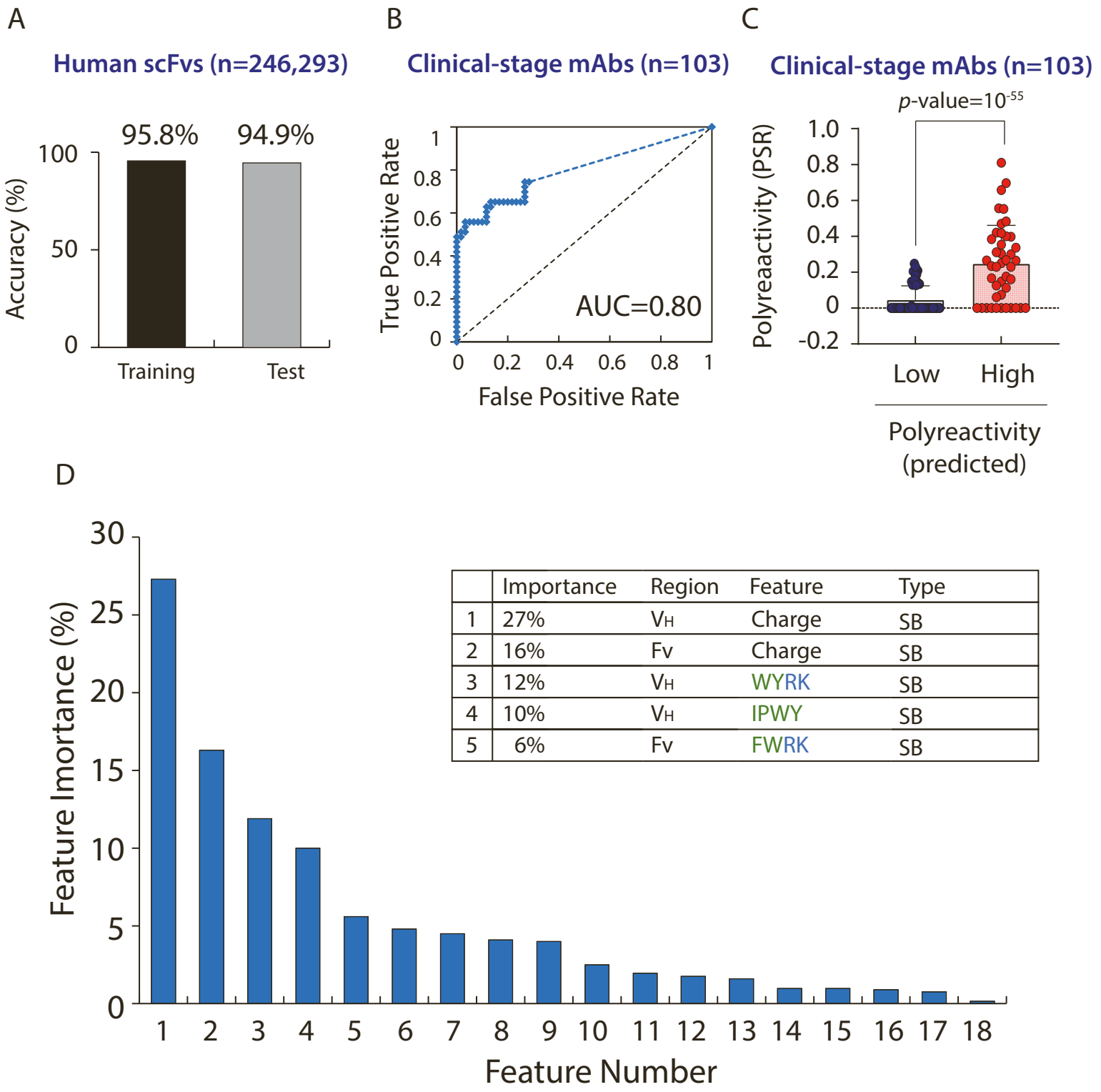


Figure S4. Charge and hydrophobicity features in HCDR2 that differentiate between antibodies with high and low levels of polyreactivity. Related to Figure 4. (A-F) Distributions of features in HCDR2 linked to polyreactivity for human antibodies (libraries #1 and #2) and their corresponding AUC values. The same features for input of each library and human repertoire were also calculated. (A, B) Net charge (pH 7.4) distributions for (A) library #1 and (B) library #2. (C, D) Distributions of the number of proline, tryptophan, arginine, and lysine residues for (C) library #1 and (D) library #2. (E, F) Distributions of the number of isoleucine, tyrosine, and tryptophan residues for (E) library #1 and (F) library #2.



	Importance	Region	Feature	Type
1	27%	VH	Charge	SB
2	16%	Fv	Charge	SB
3	12%	VH	WYRK	SB
4	10%	VH	IPWY	SB
5	6%	Fv	FWRK	SB

Figure S5. Training performance of random forest model for predicting human antibody polyreactivity. Related to Table 1. (A) Validation (training) and test performance of the model trained on 80% of antibody set #1 (tenfold cross validation) and tested on the other 20% of the antibodies. (B) ROC curve for logistic regression analysis of previously reported measurements of non-specific (PSR) binding for clinical-stage antibodies (antibody set #5 [S1]) relative to the model predictions. (C) Comparison of non-specific binding (PSR) values for clinical-stage antibodies and model predictions. (D) The relative importance of each of the 18 molecular features used during model training. In (B) and (C), the 103 clinical-stage antibodies (antibody set #5) and the non-specific binding (PSR) measurements were used during model training. In (D), the relative importance of the features was evaluated using the training set of antibodies (80% of antibody set #1), and the net charge feature was evaluated at pH 7.4. The list of the 18 features and their relative importance values are given in Table S11. In (C), the number of biological replicates is unknown. The *p*-value is calculated using Anderson-Darling test.

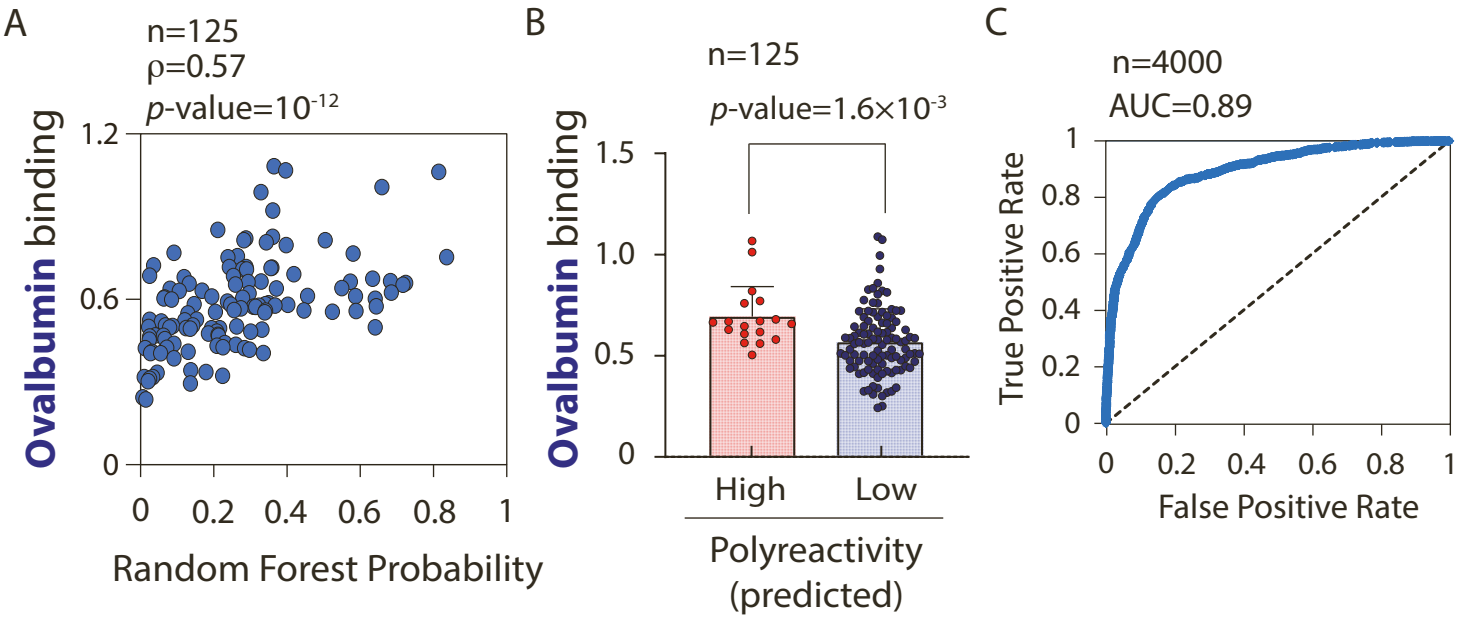


Figure S6. Evaluation of random forest model predictions using a set of emibetuzumab variants. Related to Figure 6. (A-B) Single-chain antibodies (scFabs, n=125 variants) were evaluated for their binding to ovalbumin and compared to model predictions in terms of the model (A) probability prediction or (B) classification. (C) ROC curve analysis for predicting the classification of 4000 emibetuzumab variants with either high or low polyreactivity. In (A), the statistical significance of Spearman correlation was calculated using a two-sided Student's t-test. In (B), the *p*-values were calculated using the Anderson-Darling test. The experimental data is from a previous publication [S2]. The data are the mean values of three biological replicates and the errors are the standard deviation.

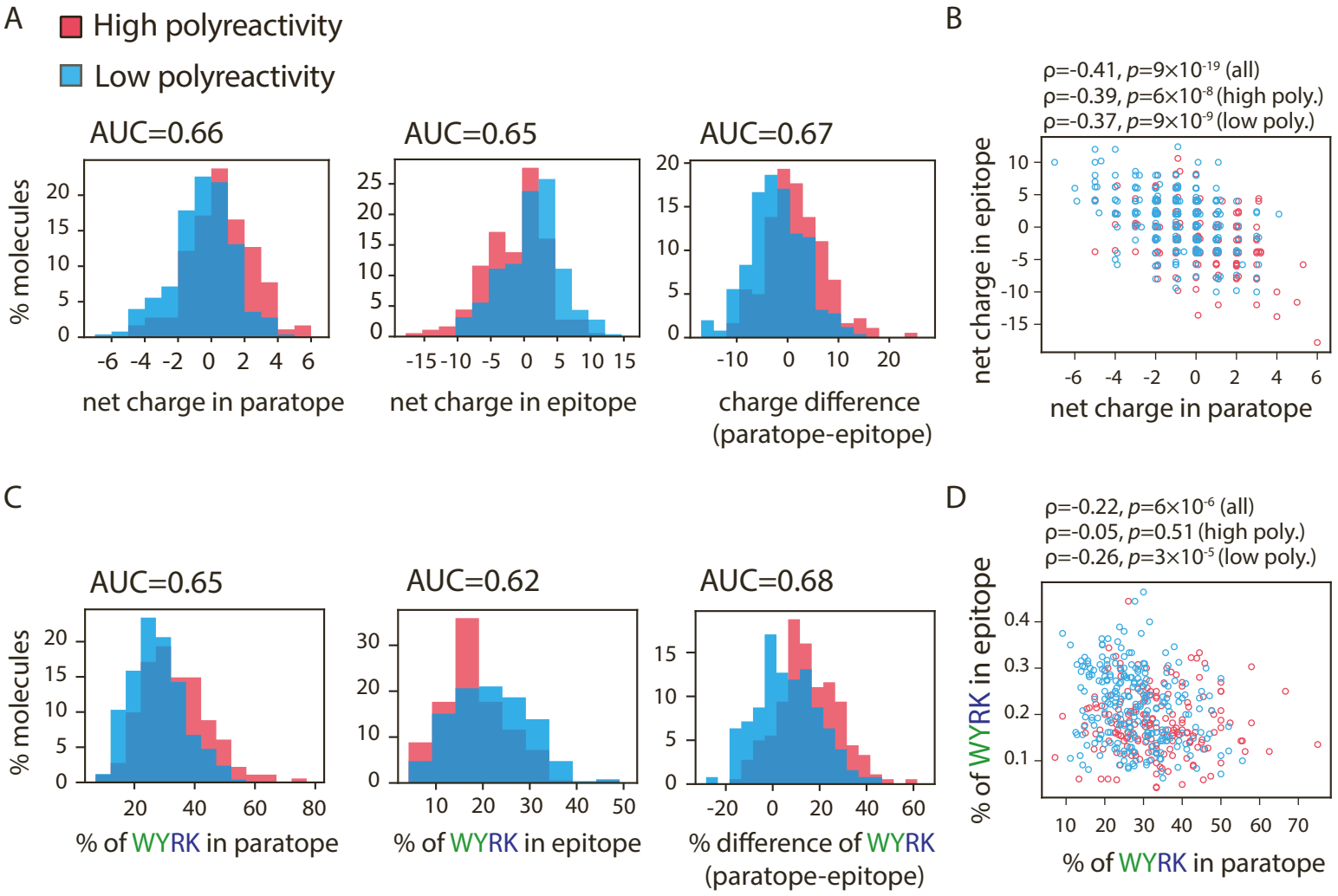


Figure S7. Distribution of molecular features linked to antibody polyreactivity in paratopes and epitopes. Related to Figure 5. (A) Net charge (pH 7.4) in antibody paratopes (left) and epitopes (center), as well as the charge difference for paratopes relative to epitopes (right) for antibodies with predicted high and low polyreactivity. (B) The relationship between net charge for antibody paratopes versus epitopes. (C) The normalized number of Trp, Tyr, Arg, and Lys residues in antibody paratopes (left) and epitopes (center), as well as the difference for paratopes relative to epitopes (right) for antibodies with high and low predicted polyreactivity. (D) The relationship between the number of Trp, Tyr, Arg, and Lys residues for antibody paratopes relative to epitopes. In (A-D), 468 antibody/antigen complexes from the PDB are used. In (C) and (D), the number of Trp, Tyr, Arg, and Lys residues is normalized by the length of the paratopes or epitopes.

Table S4. Summary of the most selective molecular features for differentiating between human antibodies with high and low polyreactivity. Related to Figure 3 and 4. The AUC values of the most significant sequence-based features are reported for libraries #1 and #2. Each feature was calculated for multiple antibody regions (Fv, V_H, V_L, CDR, HCDR and LCDR). The minus sign denotes negative correlations between the feature value and polyreactivity (i.e., increasing feature values reduce polyreactivity). Features with AUC values >0.8 for both libraries are bolded.

	Area under the Curve (AUC), Library #1/Library #2					
Feature	**Fv**	**V_H**	**V_L**	**CDR**	**HCDR**	**LCDR**
Net charge (pH 7.4)	**0.91/0.89**	**0.90/0.83**	0.75/0.78	**0.89/0.90**	**0.85/0.86**	0.76/0.62
WRK	**0.83/0.86**	**0.82/0.80**	0.66/0.73	0.81/0.79	**0.81/0.81**	0.59/0.58
IWR	**0.80/0.81**	0.83/0.77	0.60/0.69	0.73/0.80	0.74/0.79	0.57/0.62
WYRK	**0.86/0.82**	**0.85/0.80**	0.67/0.66	0.86/0.79	**0.86/0.81**	0.61/0.55
PYRK	**0.84/0.83**	0.84/0.78	0.67/0.70	0.81/0.74	0.81/0.76	0.61/0.54
IWRK	**0.81/0.88**	**0.84/0.84**	0.63/0.70	0.77/0.84	0.78/0.83	0.58/0.65
PWRK	**0.82/0.88**	**0.83/0.82**	0.63/0.76	0.77/0.78	0.77/0.79	0.58/0.59
WRHK	**0.81/0.84**	0.80/0.79	0.65/0.69	0.79/0.76	0.80/0.79	0.60/0.55
IYRK	**0.81/0.83**	0.82/0.79	0.64/0.70	**0.81/0.80**	0.80/0.78	0.62/0.63
FWRK	**0.80/0.80**	0.78/0.72	0.68/0.72	0.80/0.74	0.79/0.74	0.62/0.59
IPWR	**0.80/0.83**	0.82/0.78	0.59/0.72	0.65/0.77	0.65/0.75	0.55/0.63
FIPWY	0.82/0.76	0.84/0.73	0.61/0.63	0.64/0.72	0.64/0.71	0.54/0.57
IPWY	0.80/0.78	0.82/0.76	0.53/0.60	0.69/0.76	0.71/0.76	0.52/0.56
IWY	0.77/0.74	0.81/0.76	0.51/0.55	0.75/0.77	0.77/0.77	0.53/0.56
IPVWY	0.68/0.72	0.78/0.75	0.57 (-)/0.51	0.64/0.66	0.75/0.71	0.58 (-)/0.53 (-)
IW	0.71/0.77	0.77/0.75	0.50/0.63	0.56/0.75	0.60/0.74	0.53 (-)/0.61
W	0.74/0.73	0.77/0.76	0.51 (-)/0.52	0.74/0.73	0.77/0.76	0.53 (-)/0.52
Y	0.67/0.52	0.71/0.59	0.52/0.58 (-)	0.74/0.58	0.76/0.62	0.55/0.54 (-)
RK	0.76/0.82	0.72/0.73	0.67/0.76	0.78/0.73	0.76/0.72	0.62/0.60
D	0.76 (-)/0.72 (-)	0.67 (-)/0.74 (-)	0.67 (-)/0.74 (-)	0.76 (-)/0.80 (-)	0.63 (-)/0.73 (-)	0.73 (-)/0.68 (-)
NDE	0.79 (-)/0.70 (-)	0.69 (-)/0.65 (-)	0.73 (-)/0.63 (-)	0.75 (-)/0.64 (-)	0.62 (-)/0.55 (-)	0.73 (-)/0.63 (-)
QDE	0.77 (-)/0.73 (-)	0.78 (-)/0.65 (-)	0.61 (-)/0.68 (-)	0.78 (-)/0.69 (-)	0.77 (-)/0.63 (-)	0.63 (-)/0.63 (-)
NQDE	0.76 (-)/0.72 (-)	0.63 (-)/0.64 (-)	0.71 (-)/0.66 (-)	0.77 (-)/0.57 (-)	0.71 (-)/0.52 (-)	0.68 (-)/0.62 (-)
G	0.76 (-)/0.72 (-)	0.71 (-)/0.74 (-)	0.73 (-)/0.57 (-)	0.68 (-)/0.71 (-)	0.68 (-)/0.68 (-)	0.56 (-)/0.60 (-)

Table S5. Summary of the most selective molecular features based on individual CDRs for differentiating between human antibodies with different levels of polyreactivity. Related to Figure 4. The AUC values of the most significant sequence-based features are reported for library #1 and #2. The amino acid features are the number of the designated amino acids in each antibody region. The minus sign denotes negative correlations between the feature value and polyreactivity (i.e., increasing feature values reduce polyreactivity). Features with AUC values >0.7 for both libraries are underlined.

	Area under the Curve (AUC), Library #1/Library #2					
Feature	**HCDR1**	**HCDR2**	**HCDR3**	**LCDR1**	**LCDR2**	**LCDR3**
Net charge (pH 7.4)	0.57(-)/0.66	0.78/0.71	0.72/0.73	0.62/0.64	0.64/0.55	0.68/0.70
WRK	0.75/0.66	0.71/0.69	0.64/0.69	0.57/0.60	0.52/0.59(-)	0.50/0.62
IWR	0.69/0.72	0.61/0.71	0.61/0.61	0.54/0.66	0.53(-)/0.62(-)	0.54/0.55
WYRK	0.64/0.68	0.75/0.66	0.68/0.67	0.57/0.55	0.52/0.59(-)	0.60/0.61
PYRK	0.52(-)/0.60	0.75/0.75	0.63/0.57	0.60/0.52	0.57/0.59(-)	0.64/0.67
IWRK	0.68/0.73	0.66/0.75	0.61/0.62	0.53/0.67	0.52/0.59(-)	0.56/0.63
PWRK	0.74/0.65	0.70/0.81	0.60/0.62	0.62/0.60	0.57(-)/0.59(-)	0.57/0.67
WRHK	0.67/0.66	0.72/0.67	0.64/0.68	0.56/0.60	0.54/0.59(-)	0.53/0.57
IYRK	0.53(-)/0.66	0.72/0.73	0.63/0.57	0.55/0.61	0.52/0.59(-)	0.63/0.65
FWRK	0.57/0.55(-)	0.69/0.76	0.67/0.63	0.58/0.61	0.52/0.59(-)	0.57/0.62
IPWR	0.68/0.72	0.54/0.76	0.57/0.56	0.59/0.67	0.62(-)/0.60(-)	0.57/0.62
FIPWY	0.54(-)/0.62	0.61/0.78	0.62/0.50	0.55/0.59	0.62(-)/0.59(-)	0.57/0.54
IPWY	0.57/0.72	0.67/0.78	0.60/0.51	0.54/0.59	0.63(-)/0.59(-)	0.56/0.55
IWY	0.57/0.72	0.71/0.71	0.61/0.55	0.51/0.58	0.51/0.50	0.54/0.50
IPVWY	0.68/0.74	0.67/0.73	0.58/0.51(-)	0.53(-)/0.50	0.61(-)/0.56(-)	0.53(-)/0.55(-)
IW	0.67/0.71	0.55(-)/0.67	0.55/0.53	0.53(-)/0.64	0.51/0.50	0.53(-)/0.50
W	0.74/0.64	0.67/0.63	0.58/0.61	0.51/0.56	0.50/0.50	0.54(-)/0.54(-)
Y	0.55(-)/0.58	0.75/0.56	0.61/0.53	0.51/0.57(-)	0.50/0.50	0.56/0.52
RK	0.56/0.53	0.70/0.64	0.60/0.63	0.57/0.58	0.52/0.59(-)	0.59/0.66
D	0.62/0.62(-)	0.62(-)/0.59(-)	0.67(-)/0.65(-)	0.62(-)/0.64(-)	0.64(-)/0.59(-)	0.63(-)/0.60(-)
NDE	0.62/0.56(-)	0.65(-)/0.61	0.69(-)/0.66(-)	0.69(-)/0.53(-)	0.65(-)/0.66(-)	0.63(-)/0.58(-)
QDE	0.59/0.62(-)	0.78(-)/0.58	0.67(-)/0.64(-)	0.56(-)/0.59(-)	0.63(-)/0.60(-)	0.55(-)/0.55(-)
NQDE	0.61/0.56(-)	0.76(-)/0.70	0.67(-)/0.65(-)	0.66(-)/0.52(-)	0.64(-)/0.67(-)	0.57(-)/0.57(-)
G	0.56(-)/0.61	0.71(-)/0.77(-)	0.53(-)/0.53(-)	0.63(-)/0.53(-)	0.52/0.63(-)	0.59/0.55(-)

Table S6. Charge and hydrophobicity features in individual HCDRs that differentiate between various HCDR3 length of human antibodies with high and low polyreactivity. Related to Figure 4. The AUC values of significant sequence-based features are reported for libraries #1 and #2. Each feature was calculated for 3 HCDR regions (HCDR1, HCDR2, HCDR3). The minus sign denotes negative correlations between the feature value and polyreactivity (i.e., increasing feature values reduce polyreactivity). Features with AUC values >0.7 for both libraries are bolded, and those with AUC values >0.8 are also underlined.

HCDR3 length	**HCDR region**	**Area under the Curve (AUC), Library #1/Library #2**				
		Net charge (pH 7.4)	PYRK	PWRK	IYRK	IWY
All	HCDR1	0.57 (-)/0.66	0.52 (-)/0.60	0.74/0.65	0.53 (-)/0.66	0.57/0.72
	HCDR2	**0.78/0.71**	**0.75/0.75**	**0.70/0.81**	**0.72/0.73**	**0.71/0.71**
	HCDR3	**0.72/0.73**	0.63/0.57	0.60/0.62	0.63/0.57	0.61/0.55
5	HCDR1	0.67 (-)/0.55	0.57 (-)/0.51 (-)	0.69/0.64	0.66/0.52	0.83/0.61
	HCDR2	**0.99/0.74**	0.91/0.67	**0.89/0.73**	0.94/0.60	0.97/0.61
	HCDR3	0.56/0.75	0.96/0.59 (-)	0.91/0.52 (-)	0.93/0.55 (-)	0.66/0.59 (-)
8	HCDR1	0.57/0.64	0.58/0.60	0.57/0.63	0.60/0.61	0.57/0.64
	HCDR2	0.61/0.65	0.60/0.73	0.58/0.71	0.64/0.69	0.66/0.61
	HCDR3	**0.82/0.82**	0.73/0.69	0.68/0.67	0.66/0.70	0.60/0.59
10	HCDR1	0.60 (-)/0.65	0.53 (-)/0.604	0.72/0.714	0.50/0.670	0.60/0.719
	HCDR2	0.85/0.58	**0.78/0.71**	**0.78/0.76**	0.74/0.65	0.67/0.65
	HCDR3	0.55/0.70	0.55/0.54 (-)	0.67/0.72	0.56 (-)/0.58 (-)	0.55/0.53 (-)
12	HCDR1	0.64 (-)/0.68	0.65 (-)/0.57	0.86/0.65	0.75 (-)/0.60	0.60 (-)/0.67
	HCDR2	**0.89/0.74**	**0.90/0.77**	**0.87/0.84**	**0.88/0.74**	0.86/0.69
	HCDR3	**0.73/0.83**	**0.75/0.79**	0.70/0.67	0.65/0.78	0.52/0.65
15	HCDR1	0.77 (-)/0.69	0.68 (-)/0.60	0.83/0.58	0.75 (-)/0.65	0.60 (-)/0.67
	HCDR2	0.84/0.66	**0.89/0.75**	**0.84/0.78**	**0.84/0.79**	**0.75/0.80**
	HCDR3	**0.86/0.93**	0.75/0.65	0.73/0.62	**0.74/0.74**	0.62/0.70

Table S7. Summary of the AUC values of the significant sequence-based features in HCDR2 for common germline genes in library #1 and #2. Related to Figure 4. The minus sign denotes negative correlations between the feature value and polyreactivity (i.e., increasing feature values reduce polyreactivity). Features with AUC values >0.7 for both libraries are bolded, and those with AUC values >0.8 are also underlined.

	Area under the Curve (AUC), Library #1/Library #2				
Germline gene	Charge	PYRK	IWY	NQDE	G
VH1-69	0.72/0.65	0.52 (-)/0.54	0.68/0.63	0.54 (-)/0.54 (-)	0.52/0.62 (-)
VH1-46	0.69/0.54 (-)	0.54 (-)/0.62	0.56 (-)/0.68	0.74 (-)/0.74 (-)	0.52 (-)/0.64
VH1-2	**0.76/0.70**	0.75/0.51	0.54 (-)/0.52(-)	0.64 (-)/0.63 (-)	0.52/0.65
VH1-3	0.74/0.63	0.60/0.74 (-)	0.51(-)/0.55(-)	0.56/0.72	0.56/0.50(-)
VH1-18	0.50/0.64 (-)	0.70/0.60(-)	0.53 (-)/0.52(-)	0.52/0.77 (-)	0.54/0.52
VH1-8	0.54 (-)/0.51 (-)	0.63 (-)/0.51	0.55/0.53	0.66/0.53 (-)	0.60/0.53 (-)
VH1-24	0.56/0.62 (-)	0.66/0.60 (-)	0.50/0.51 (-)	0.51 (-)/0.61	0.71/0.61
VH3-30	**0.70/0.70**	0.61/0.70	0.57 (-)/0.69	0.64 (-)/0.63 (-)	0.54 (-)/0.58 (-)
VH3-30-3	0.66/0.60	0.83/0.66	0.62/0.68	0.79 (-)/0.58 (-)	0.51 (-)/0.52 (-)
VH3-33	0.91/0.51	0.92/0.67	0.92/0.71	0.91 (-)/0.60 (-)	0.56/0.51 (-)
VH3-48	0.65/0.52 (-)	0.88/0.61	0.80/0.69	**0.74 (-)/0.71 (-)**	0.72/0.61 (-)
VH3-49	0.51/0.51	0.56/0.77	0.62/0.75	0.61/0.50	0.53 (-)/0.5
VH3-53	0.93/0.68	**0.76/0.75**	**0.86/0.75**	0.70 (-)/0.65 (-)	0.64/0.53 (-)
VH3-64	0.99/0.61	0.78/0.51	0.69/0.52	0.79 (-)/0.68	0.84/0.65 (-)
VH3-7	**0.94/0.81**	**0.84/0.76**	0.53 (-)/0.63 (-)	0.81 (-)/0.52 (-)	0.75/0.58 (-)
VH3-72	0.78 (-)/0.60	0.89 (-)/0.65	0.54 (-)/0.52	0.52 (-)/0.80	0.57/0.61 (-)
VH3-74	**0.90/0.77**	0.83/0.63	0.53 (-)/0.58 (-)	0.75 (-)/0.58 (-)	0.70/0.54 (-)
VH3-15	**0.96/0.79**	0.92/0.56 (-)	0.72/0.53	0.80 (-)/0.66 (-)	**0.73/0.77**
VH3-23	0.57 (-)/0.53	0.69/0.62	0.83/0.51	0.58 (-)/0.58	0.58/0.60 (-)
VH3-11	**0.96/0.72**	0.72/0.55	0.54 (-)/0.74	0.59 (-)/0.81 (-)	0.50/0.55 (-)
VH3-9	0.67/0.53	0.65 (-)/0.52	0.52/0.50	0.52/0.51 (-)	0.87/0.52 (-)
VH4-4	0.62/0.64 (-)	0.51 (-)/0.64	0.59/0.83	**0.94 (-)/0.83 (-)**	0.69 (-)/0.53
VH4-61	0.64/0.63	0.67/0.59	0.62/0.51	0.56 (-)/0.60	0.62 (-)/0.51 (-)
VH4-39	0.98/0.51	0.90/0.65 (-)	0.94/0.68 (-)	0.81 (-)/0.54	0.56/0.51
VH5-51	0.55 (-)/0.55 (-)	0.52 (-)/0.52	0.75 (-)/0.51	0.69 (-)/0.52	0.59/0.52 (-)
VH5-10-1	0.5/0.55 (-)	0.93 (-)/0.55	0.57/0.53	0.93/0.51 (-)	0.5/0.5
VH6-1	0.65/0.68	0.58/0.67	0.59/0.55	0.67 (-)/0.56	0.50 (-)/0.51

Table S9. Summary of the molecular features enriched or depleted in antibody paratopes. Related to Figure 5. The AUC values of sequence-based features are reported for the dataset that contains 468 antibody/antigen complexes (antibody set #7). Each feature was separately calculated for V_H and V_L domains. The minus sign denotes depletion from the paratope. Features with AUC values >0.8 for both libraries are bolded. The median values of both paratope and non-paratope regions are also reported.

	V_H			V_L		
Feature	**Area under the Curve (AUC)**	**Median of paratopes**	**Median of non-paratopes**	**Area under the Curve (AUC)**	**Median of paratopes**	**Median of non-paratopes**
Net charge (pH 7.4)	0.75 (-)	0.0	1.1	0.70 (-)	0.0	1.0
WRK	0.52 (-)	11.1%	11.8%	0.51	9.1%	9.2%
IWR	0.61	14.3%	10.8%	0.50	10.0%	9.9%
WYRK	**0.86**	30.0%	16.5%	**0.87**	28.6%	13.4%
PYRK	**0.84**	27.3%	16.3%	0.71	27.3%	18.3%
IWRK	0.54	15.8%	14.7%	0.54 (-)	12.5%	14.3%
PWRK	0.52 (-)	14.3%	14.8%	0.62 (-)	11.1%	15.2%
WRHK	0.52	13.3%	12.2%	0.55	11.1%	9.4%
IYRK	**0.88**	28.6%	16.3%	0.79	27.3%	17.2%
FWRK	0.55	15.8%	14.4%	0.51	13.3%	13.1%
IPWR	0.58	15.8%	13.6%	0.65 (-)	12.5%	16.2%
FIPWY	**0.94**	31.8%	16.5%	0.76	30.0%	20.6%
IPWY	**0.92**	27.8%	13.5%	0.76	27.3%	16.7%
IWY	**0.93**	25.0%	10.6%	**0.87**	25.0%	10.4%
IPVWY	**0.81**	31.7%	22.7%	0.68	28.6%	21.8%
IW	0.59	7.7%	5.9%	0.55 (-)	0.0%	6.1%
W	0.50	3.5%	3.0%	0.67 (-)	0.0%	1.0%
Y	**0.97**	16.4%	4.7%	**0.85**	16.7%	4.1%
RK	0.59 (-)	7.1%	8.6%	0.57 (-)	0.0%	8.0%
D	0.63	5.9%	3.8%	0.65	0.0%	4.0%
NDE	0.72	15.0%	10.0%	0.63 (-)	15.4%	8.3%
QDE	0.60 (-)	10.5%	12.9%	0.56	10.0%	14.7%
NQDE	0.57	16.7%	15.2%	0.65 (-)	20.0%	15.8%
G	0.59 (-)	9.1%	11.1%	0.51	0.0%	9.5%

Table S10. Summary of the 43 molecular features used in developing the first-generation random forest model for predicting human antibody polyreactivity. Related to Table 1. The identities of 43 molecular features, including 31 SB features and 12 SB-SE features, are reported. Their respective antibody regions (Fv, V_H and CDRs), feature types (SB and SB-SE) and relative importance in the model predictions are also reported. The 12 SB-SE features used in this model were reported previously as chemical rules [S3].

#	Feature	Region	Feature type	Importance	Chemical Rule #
F1	net charge (pH 7.4)	Fv	SB	16.6%	NA
F2	net charge (pH 7.4)	V_H	SB	11.5%	NA
F3	WYRK	V_H	SB	8.9%	NA
F4	WRK	Fv	SB	7.2%	NA
F5	IYRK	V_H	SB	5.2%	NA
F6	FWRK	Fv	SB	5.1%	NA
F7	DEN	Fv	SB	4.4%	NA
F8	IPWY	V_H	SB	4.4%	NA
F9	IWY	V_H	SB	4.2%	NA
F10	DEQ	Fv	SB	2.4%	NA
F11	DENTVA	H23 L123	SB-SE	2.4%	#8
F12	DENQAILMPH	V_H	SB-SE	2.4%	#10
F13	DENQSLMWHR	V_H	SB-SE	2.3%	#11
F14	FIPWY	Fv	SB	2.0%	NA
F15	IWRK	V_H	SB	1.5%	NA
F16	IWRK	Fv	SB	1.4%	NA
F17	W	V_H	SB	1.4%	NA
F18	RHFWPYA	H13 L123	SB-SE	1.2%	#4
F19	DETIVWF	H123 L123	SB-SE	1.1%	#9
F20	WYRK	Fv	SB	1.1%	NA
F21	KMWYGQT	V_H	SB-SE	1.1%	#6
F22	G	V_H	SB	1.0%	NA
F23	IPVWY	V_H	SB	1.0%	NA
F24	RVPWYQE	H2 L13	SB-SE	1.0%	#3
F25	D	Fv	SB	0.9%	NA
F26	DENQ	Fv	SB	0.9%	NA
F27	RKHWIVMPYQ	H13 L2	SB-SE	0.8%	#1
F28	DENTYPMH	H2 L1	SB-SE	0.8%	#7
F29	RKHFWLPYQ	H13 L3	SB-SE	0.7%	#2
F30	RHPMLYWQ	H123 L23	SB-SE	0.7%	#5
F31	WRK	V_H	SB	0.7%	NA
F32	W	Fv	SB	0.6%	NA
F33	PWRK	V_H	SB	0.5%	NA
F34	RK	Fv	SB	0.4%	NA
F35	IW	V_H	SB	0.4%	NA
F36	IPWY	Fv	SB	0.3%	NA

F37	PYRK	Fv	SB	0.3%	NA
F38	G	Fv	SB	0.3%	NA
F39	DNTLMFP	Fv	SB-SE	0.3%	#12
F40	PWRK	Fv	SB	0.2%	NA
F41	IPWR	Fv	SB	0.2%	NA
F42	WRHK	Fv	SB	0.2%	NA
F43	IWR	Fv	SB	0.2%	NA

Table S11. Summary of the 18 molecular features used in developing the final random forest model for predicting human antibody polyreactivity. Related to Table 1. The identities of 18 molecular features, including 7 SB features and 11 SB-SE features, are reported. Their respective antibody regions (Fv, V_H and CDRs), feature types (SB and SB-SE) and relative importance in the model predictions are also reported. The 11 SB-SE features used in this model were reported previously as chemical rules [S3].

#	Feature	Region	Feature type	Importance	Chemical Rule #
F1	net charge (pH 7.4)	VH	SB	27.3%	NA
F2	net charge (pH 7.4)	Fv	SB	16.3%	NA
F3	WYRK	VH	SB	11.9%	NA
F4	IPWY	VH	SB	10.0%	NA
F5	FWRK	Fv	SB	5.6%	NA
F6	DEN	Fv	SB	4.8%	NA
F7	DENQAILMPH	VH	SB-SE	4.5%	#10
F8	IYRK	Fv	SB	4.1%	NA
F9	DENQSLMWHR	VH	SB-SE	4.0%	#11
F10	DETIVWF	H123 L123	SB-SE	2.5%	#9
F11	RVPWYQE	H2 L13	SB-SE	2.0%	#3
F12	RHFWPYA	H13 L123	SB-SE	1.8%	#4
F13	KMWYGQT	VH	SB-SE	1.6%	#6
F14	RKHFWYPLQ	H13 L3	SB-SE	1.0%	#2
F15	RKHWIVMPYQ	H13 L2	SB-SE	1.0%	#1
F16	DENTYPMH	H2 L1	SB-SE	0.9%	#7
F17	RHPMLYWQ	H123 L23	SB-SE	0.8%	#5
F18	DNTLMFP	Fv	SB-SE	0.0%	#12

Table S13. The AUC values of the molecular features linked to predicted antibody polyreactivity in paratopes and epitopes for 468 antibody/antigen complexes. Related to Figure 5. The difference for paratopes relative to epitopes for each feature was also calculated. The number of amino acids was normalized by the length of paratopes or epitopes. The minus sign denotes negative correlations between the feature value and polyreactivity (i.e., increasing feature values reduce polyreactivity).

Feature	Paratope	Epitope	Paratope-Epitope
Net charge (pH 7.4)	0.66	0.65	0.67
WRK	0.63	0.55 (-)	0.63
IWR	0.59	0.52	0.56
WYRK	0.65	0.62 (-)	0.68
PYRK	0.63	0.63 (-)	0.67
IWRK	0.60	0.51	0.56
PWRK	0.63	0.53 (-)	0.61
WRHK	0.65	0.55 (-)	0.64
IYRK	0.61	0.59 (-)	0.63
FWRK	0.61	0.61 (-)	0.64
IPWR	0.60	0.54	0.55
FIPWY	0.58	0.57 (-)	0.61
IPWY	0.59	0.52 (-)	0.59
IWY	0.57	0.55 (-)	0.59
IPVWY	0.56	0.52 (-)	0.57
IW	0.54	0.60	0.53 (-)
W	0.60	0.53	0.57
Y	0.54	0.60 (-)	0.61

REFERENCES

S1. Jain, T., Sun, T., Durand, S., Hall, A., Houston, N.R., Nett, J.H., Sharkey, B., Bobrowicz, B., Caffry, I., Yu, Y., et al. (2017). Biophysical properties of the clinical-stage antibody landscape. Proceedings of the National Academy of Sciences *114*, 944–949. https://doi.org/10.1073/pnas.1616408114.

S2. Makowski, E.K., Kinnunen, P.C., Huang, J., Wu, L., Smith, M.D., Wang, T., Desai, A.A., Streu, C.N., Zhang, Y., Zupancic, J.M., et al. (2022). Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. Nat Commun *13*, 3788. https://doi.org/10.1038/s41467-022-31457-3.

S3. Zhang, Y., Wu, L., Gupta, P., Desai, A.A., Smith, M.D., Rabia, L.A., Ludwig, S.D., and Tessier, P.M. (2020). Physicochemical Rules for Identifying Monoclonal Antibodies with Drug-like Specificity. Mol Pharm *17*, 2555–2569. https://doi.org/10.1021/acs.molpharmaceut.0c00257.

S4. Makowski, E.K., Wu, L., Desai, A.A., and Tessier, P.M. (2021). Highly sensitive detection of antibody nonspecific interactions using flow cytometry. mAbs *13*, 1951426. https://doi.org/10.1080/19420862.2021.1951426.