

1  
2       **An *in silico* method to assess antibody fragment polyreactivity**  
3

4       Edward P. Harvey<sup>1,\*</sup>, Jung-Eun Shin<sup>2,3,\*</sup>, Meredith A. Skiba<sup>1,\*</sup>, Genevieve R. Nemeth<sup>1</sup>,  
5       Joseph D. Hurley<sup>1</sup>, Alon Wellner<sup>4,5,6</sup>, Ada Y. Shaw<sup>2</sup>, Victor G. Miranda<sup>1</sup>, Joseph K. Min<sup>2</sup>,  
6       Chang C. Liu<sup>4,5,6</sup>, Debora S. Marks<sup>2,3,†</sup>, Andrew C. Kruse<sup>1,†</sup>  
7

8       <sup>1</sup>Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute,  
9       Harvard Medical School, Boston, MA 02115, USA.

10      <sup>2</sup>Department of Systems Biology, Harvard Medical School, Boston, MA 02215, USA  
11

12      <sup>3</sup>Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA  
13

14      <sup>4</sup>Department of Biomedical Engineering, University of California, Irvine, CA 92692, USA  
15

16      <sup>5</sup>Department of Chemistry, University of California, Irvine, CA 92697, USA  
17

18      <sup>6</sup>Department of Molecular Biology & Biochemistry, University of California, Irvine, CA  
19       92697, USA  
20

21  
22  
23  
24  
25      \*These authors, listed alphabetically, contributed equally  
26

27      †Correspondence and requests for materials should be sent to Debora S. Marks  
28      ([Debora.Marks@hms.harvard.edu](mailto:Debora.Marks@hms.harvard.edu)) or Andrew C. Kruse  
29      ([Andrew.kruse@hms.harvard.edu](mailto:Andrew.kruse@hms.harvard.edu))

30 **ABSTRACT**

31 Antibodies are essential biological research tools and important therapeutic agents, but  
32 some exhibit non-specific binding to off-target proteins and other biomolecules. Such  
33 polyreactive antibodies compromise screening pipelines, lead to incorrect and  
34 irreproducible experimental results, and are generally intractable for clinical development.  
35 We designed a set of experiments using a diverse naïve synthetic camelid antibody  
36 fragment ('nanobody') library to enable machine learning models to accurately assess  
37 polyreactivity from protein sequence ( $AUC > 0.8$ ). Moreover, our models provide  
38 quantitative scoring metrics that predict the effect of amino acid substitutions on  
39 polyreactivity. We experimentally tested our model's performance on three independent  
40 nanobody scaffolds, where over 90% of predicted substitutions successfully reduced  
41 polyreactivity. Importantly, the model allowed us to diminish the polyreactivity of an  
42 angiotensin II type I receptor antagonist nanobody, without compromising its  
43 pharmacological properties. We provide a companion web-server that offers a  
44 straightforward means of predicting polyreactivity and polyreactivity-reducing mutations  
45 for any given nanobody sequence.

46

47

48

49

50

51

52

53

## 54 INTRODUCTION

55 Due to their specificity and affinity, antibodies are an indispensable class of  
56 biomedical research tools as well as important therapeutics for the treatment of cancer,  
57 autoimmune, and infectious diseases. Current antibody discovery methods prioritize the  
58 generation of antibodies and antibody fragments with high target specificity. However,  
59 some antibodies that strongly bind one target interact with additional antigens with low-  
60 affinity. In clinical development, these non-specific or polyreactive antibodies show poor  
61 pharmacokinetics or other liabilities that limit clinical use<sup>1-3</sup>. Additionally, polyreactive  
62 antibodies encountered in the basic research setting cause misinterpretation of results,  
63 low reproducibility in routine experiments, and wasted time and money<sup>4</sup>. Thus, there have  
64 been several calls to standardize the quality and specificity of antibodies used in research  
65 settings similar to those in the clinic<sup>5,6</sup>.

66 Developing and improving methods to detect and quantify polyreactivity are  
67 essential for enhancing the quality of antibodies in both clinical development and basic  
68 research settings. Many experimental methods that evaluate polyreactivity<sup>7-14</sup> are low-  
69 throughput and require experimental screening with purified antibody. The degree of  
70 polyreactivity is highly method and reagent-dependent and is typically measured after  
71 antigen selection, making it difficult to prioritize the most promising clones. Understanding  
72 sequence features of polyreactive antibodies could provide an efficient avenue to  
73 quantitatively assess antibody polyreactivity without experimental effort. Previous  
74 computational methods<sup>15-22</sup> have revealed features of polyreactivity antibodies, such as  
75 J- and V-chain usage<sup>17</sup>, high isoelectric points in the complementarity determining regions  
76 (CDRs)<sup>16,18-25</sup>, longer CDR3s<sup>16,23</sup>, enrichment of arginine, glycine, valine, and tryptophan

77 containing motifs<sup>18</sup>, and glutamine residues<sup>23</sup>. Despite these extensive analyses the  
78 relative importance of many characteristics is disputed<sup>21</sup> and prediction software cannot  
79 quantitate polyreactivity<sup>17</sup>.

80 For broad utility, a computational method should accurately predict the *degree* of  
81 polyreactivity and compute candidate rescue mutations from the input of a user sequence  
82 alone. To achieve this goal, we designed experiments to learn features of high and low  
83 polyreactivity clones from a naïve synthetic yeast display library of heavy-chain only  
84 camelid antibody fragments (nanobodies)<sup>26,27</sup> through computational methods. Synthetic  
85 nanobodies provide an ideal reductionist system to probe polyreactivity in the context of  
86 a fixed framework without the influence of heavy and light chain pairing effects. These  
87 methods result in generalizable software that quantifies nanobody polyreactivity based  
88 on sequence alone and most importantly designs specific mutations to decrease  
89 polyreactivity.

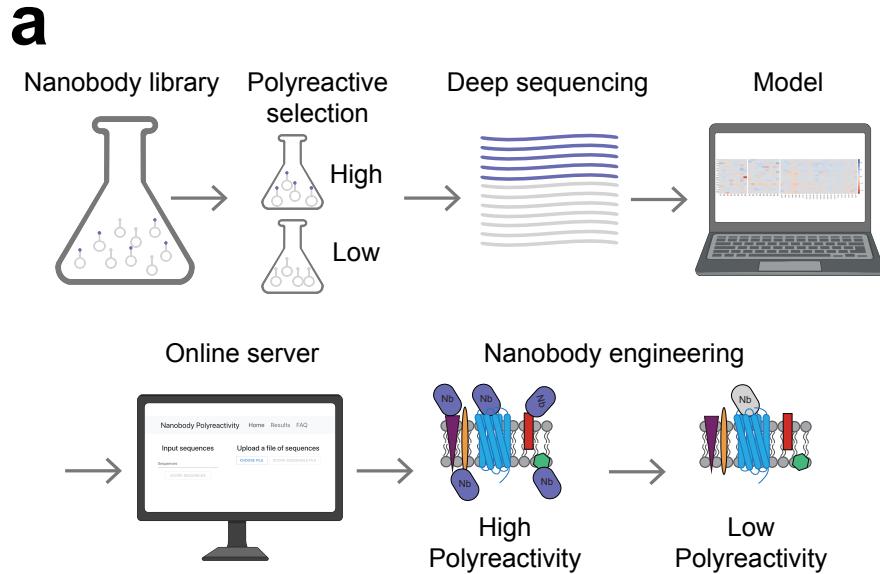
90 We successfully applied our software to three polyreactive nanobodies, including  
91 AT118i4h32, a nanobody antagonist of the angiotensin II type I receptor (AT1R)<sup>28</sup>, where  
92 we reduced polyreactivity without compromising binding affinity or target-specific  
93 pharmacology. This sequence-based approach may be a generally useful tool for  
94 prioritizing nanobody clones identified in selection experiments and improving  
95 nanobodies targeting diverse antigens. While nanobodies are gaining popularity as next  
96 generation biotherapeutics<sup>29</sup> that target antigen surfaces and tissue types not accessible  
97 to conventional antibodies, the approaches developed here are in principle fully  
98 applicable to conventional antibodies as well.

99

100 **RESULTS**

101 **Enriching naïve library for polyreactive clones**

102       Unlike previous analyses of antibody polyreactivity which relied on clinical  
103 candidates<sup>23-25</sup>, clones enriched for antigen binding<sup>17</sup>, or primarily focused on the  
104 contribution of V<sub>H</sub> CDR3 antibody polyreactivity<sup>18,21</sup>, we designed experiments to assess  
105 polyreactivity of clones from a naïve synthetic yeast display library through binding to  
106 detergent-solubilized *Spodoptera frugiperda* (Sf9) insect cell membranes (Figure 1)<sup>14</sup>.  
107 This mixed protein polyspecificity reagent (PSR) is compatible with sorting large pools of  
108 antigen naïve clones, allowing us to determine global contributions to polyreactivity in an  
109 unbiased manner. The yeast display library contains  $>2 \times 10^9$  unique nanobody clones that  
110 mimic a naïve llama immune repertoire in CDR sequence composition and CDR3 length  
111 and possesses moderate diversity in the CDR1 and CDR2 regions and extensive diversity  
112 in the CDR3 region. We used Magnetic-Activated Cell Sorting (MACS) to both enrich for  
113 polyreactive clones and deplete non-expressing clones from the library. Following MACS,  
114 distinct populations of clones with high and low polyreactivity were isolated by  
115 Fluorescence-Activated Cell Sorting (FACS) (Supplementary Figure 1A-B).

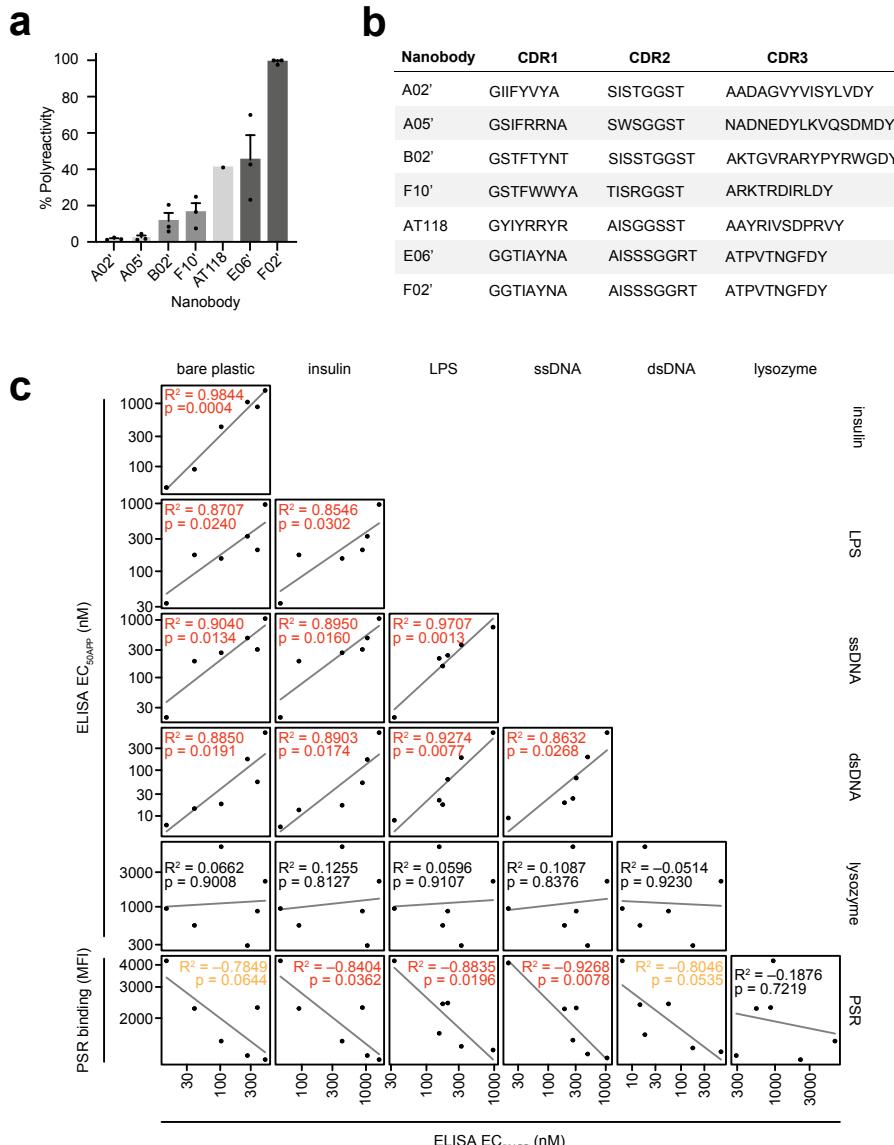


116  
117

118 **Figure 1. Development of computational tool to assess and mitigate polyreactivity.**  
119 Starting from a large, naïve synthetic nanobody library, pools of nanobodies with low and  
120 high polyreactivity were isolated. Machine learning models were trained on deep  
121 sequencing data from these pools to learn sequence features of low and high polyreactive  
122 nanobodies. These algorithms were incorporated into software that quantitatively predicts  
123 polyreactivity levels and recommends substitutions that reduce it.  
124

125 PSR reagent has not been used to assess nanobody polyreactivity, but is well  
126 validated against other measures of polyreactivity for conventional antibodies<sup>2,14,15</sup>. To  
127 validate PSR performance on nanobodies, we recombinantly expressed six nanobodies  
128 with varying levels of polyreactivity from our FACS sorted pools and assessed  
129 polyreactivity by conventional ELISA assays against lysozyme, double stranded DNA  
130 (dsDNA), single stranded DNA (ssDNA), insulin, lipopolysaccharide (LPS), and bare  
131 plastic (Figure 2, Supplementary Figure 2A-F). ELISA polyreactivity assays performed  
132 using different reagents correlated well with one another ( $r^2$  values between 0.789 and  
133 0.986,  $p < 0.05$ ) with the exception of lysozyme ( $r^2$  values between -0.109 and 0.045,  $p$ -  
134 values between 0.8127 and 0.9230), which did not correlate with the other reagents.

135 Furthermore, direct ELISA assays strongly correlated with insect cell PSR ( $r^2$  values  
136 between 0.7849 and 0.9268) except for lysozyme which exhibited a very weak correlation  
137 ( $r^2 = -0.1876$ ). The correlations between insulin, LPS, and ssDNA direct ELISA assays to  
138 insect cell PSR staining were highly significant ( $p < 0.05$ ), while bare plastic and dsDNA  
139 direct ELISA assays were modestly significant ( $p < 0.10$ ). Lysozyme direct ELISA assays  
140 did not significantly correlate with insect cell PSR staining ( $p = 0.7219$ ). We also observed  
141 that polyreactive clones had increased retention times in conventional size exclusion  
142 chromatography albeit not with statistical significance ( $r^2 = 0.7836$ ,  $p = 0.1168$ ),  
143 suggesting that nanobody polyreactivity may be detected during routine protein  
144 purification (Supplementary Figure 2G). Overall, the ELISA experiments support that the  
145 pools of nanobodies selected by PSR staining possess high and low levels of  
146 polyreactivity. Armed with this validation, we deep-sequenced the two FACS sorted pools  
147 and obtained 65,147 unique low polyreactivity sequences and 69,155 unique highly  
148 polyreactive sequences that contained 51,308 and 59,623 distinct CDR regions.



149

150

151 **Figure 2. Correlations between direct ELISA assays and insect cell polyspecificity**  
 152 **reagent (PSR) staining.**

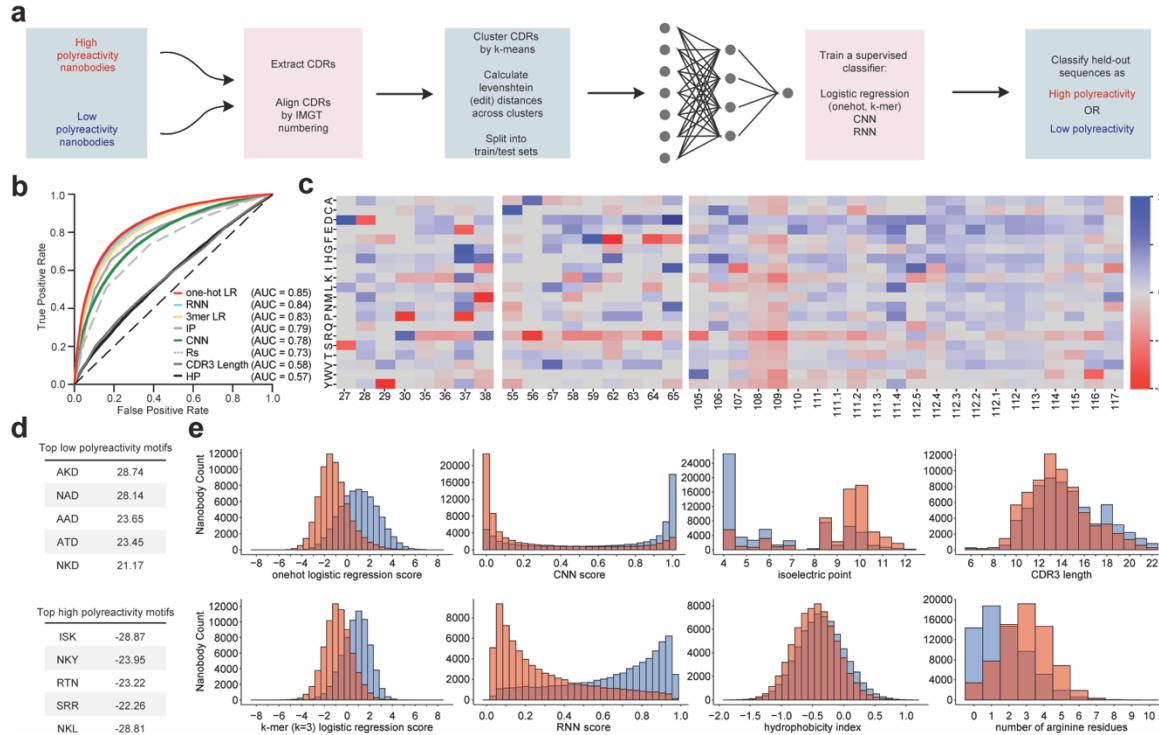
153 **a**, *Spodoptera frugiperda* (Sf9) insect cell PSR staining of single nanobodies isolated from  
 154 FACS sorts. Data are mean +/- SEM of three independent biological experiments  
 155 performed in technical triplicate. Polyreactivity levels are normalized with respect to the  
 156 highest value. **b**, CDR sequences of isolated nanobodies. **c**, Direct ELISA assays  
 157 measured the apparent EC<sub>50</sub> (EC<sub>50APP</sub>) of five index panel members and nanobody AT118  
 158 to the specified reagents. ELISA data are representative of two independent experiments,  
 159 each performed in technical triplicates.

160

161 **Development of computational method**

162 We developed computational models trained on the sequences from the FACS-  
163 sorted pools to classify nanobodies as possessing high or low polyreactivity. We  
164 constructed a suite of supervised, discriminative models that can separate high and low  
165 polyreactivity sequences (Figure 3A-B). These models include a logistic regression model  
166 of a one-hot embedding of the CDR sequences, a logistic regression model of a k-mer  
167 embedding (k=3) of the CDR sequences, a convolutional neural network (CNN), and a  
168 recurrent neural network (RNN). The one-hot logistic regression model learns weights for  
169 each amino acid type at each position in the CDR sequences that are most predictive of  
170 polyreactivity; the k-mer logistic regression learns weights for each motif (lengths 1, 2,  
171 and 3) that are most predictive of polyreactivity, irrespective of where they occur within a  
172 given CDR sequence. Convolutional neural networks use convolutional filters to learn  
173 spatial information (e.g., an amino acid and its neighboring residues) and are often used  
174 in image classification. Recurrent neural networks capture sequential information (e.g.  
175 the probability of a residue given the previous residues) and are frequently used in text  
176 and audio analysis. For the one-hot logistic regression and for the CNN, we align the CDR  
177 sequences using the IMGT numbering scheme with ANARCI<sup>30</sup>. The k-mer logistic  
178 regression and the RNN methods do not require aligned CDR sequences. In order to test  
179 the generalizability of our models, we clustered the nanobody sequences using k-means  
180 clustering to generate five clusters of sequences, which we used to build train and test  
181 splits. These splits and careful selection allowed us to avoid over-optimistic prediction  
182 accuracies that result from the test sets overlapping or close to the training sets<sup>31</sup>.  
183 Specifically, we ensured that all sequences in the test sets were more than 10 edit-

184 distance (Levenshtein distance) and possessed only ~75% sequence similarity in the  
 185 CDR sequences from each other (Figure 3A).  
 186



187  
 188  
 189  
 190  
 191  
 192 **Figure 3. Development of computational models to predict polyreactivity.**  
 193 Supervised models were trained on pools of high and low polyreactivity sequences. **a**,  
 194 Pipeline of computational model development, from raw NGS data to held-out predictions  
 195 with sequence clustering for rigorous validation. **b**, Comparison of supervised models  
 196 (one-hot and k-mer logistic regression, RNN, CNN) and biochemical properties such as  
 197 hydrophobicity, isoelectric point, CDR3 lengths, and number of arginine residues. **c**,  
 198 Trained parameters of a one-hot logistic regression model, showing which amino acids  
 199 at specific positions are most predictive of high polyreactivity and low polyreactivity (red  
 200 and blue, respectively). **d**, Polyreactivity scores of top motifs learned from a k-mer logistic  
 201 regression model most predictive of low and high polyreactivity (top and bottom,  
 202 respectively). **e**, Separation of high and low polyreactivity nanobodies by each of the  
 203 models and biochemical properties displayed in panel b.  
 204

205 The one-hot logistic regression, k-mer logistic regression, and RNN models  
206 performed well at classifying distant nanobody sequences as high or low polyreactivity,  
207 achieving 0.85, 0.83, and 0.84 Area Under Curve (AUC) respectively (Figure 3B).  
208 Whereas, the CNN (AUC=0.78, Figure 3B) achieved similar performance to metrics as  
209 described previously in literature, such as isoelectric point<sup>16,22-24</sup> and the number of  
210 arginine residues<sup>18,20,21,25</sup> (AUCs of 0.79 and 0.73 respectively, Figure 3B). Consistent  
211 with previous literature<sup>15,23</sup>, we found that hydrophobicity, as described by the  
212 hydrophobicity index, is not strongly predictive of polyreactivity (AUC of 0.57, Figure 3B).  
213 However, CDR3 length, which is a reported feature of polyreactive antibodies<sup>16,23</sup> is not  
214 highly predictive of nanobody polyreactivity (AUC of 0.58, Figure 3B). Score and  
215 measurement distributions of the nanobody sequences for each of these metrics,  
216 separated by labeled class are displayed in Figure 3E.

217 In addition to the models' robust performance, sequence features learned by the  
218 logistic regression methods are easily interpretable. A distinct advantage of the one-hot  
219 logistic regression model is its ability to produce a picture of amino acid contribution to  
220 polyreactivity at each position of nanobody CDR sequences (Figure 3C). In agreement  
221 with previous findings, we find that acidic residues in CDRs 2 and 3 are characteristic of  
222 low polyreactivity clones and the presence of arginine residues across all CDRs, and  
223 lysine, tryptophan, or tyrosine in CDR3 contribute to higher polyreactivity. Despite the  
224 overall enrichment of arginine and tryptophan polyreactive clones, the position specific  
225 analysis provided by the one-hot model indicates that low polyreactivity clones tolerate  
226 arginine in positions 30 and 38 of CDR1 and tryptophan in position 105 in CDR3.

227 Furthermore, the k-mer logistic regression model provides insight into sequence  
228 dependencies on the local level in high or low polyreactivity clones (Figure 3D). K-mer  
229 motifs containing negatively charged residues such as glutamate and aspartate are highly  
230 associated with low polyreactivity sequences, and positively charged residues such as  
231 arginine and lysine are predicted to contribute to polyreactivity, agreeing with the  
232 predictions of the one-hot logistic regression model. These motifs differ from previously  
233 reported polyreactive motifs, that were enriched in glycine and the hydrophobic amino  
234 acids valine and tryptophan<sup>18</sup>. However, these previously reported motifs were derived  
235 from a library where only CDR3 was diversified. We proceeded to use the one-hot and k-  
236 mer logistic regression models for further analysis based on of their accuracy and  
237 interpretability.

238

### 239 **Quantitative scoring of nanobody polyreactivity**

240 In order to test if our model could go beyond predicting binary classification labels  
241 and quantitatively score polyreactivity, we stained 48 nanobodies isolated from MACS and  
242 FACS pools with PSR to obtain an “index set” of sequenced clones with defined levels of  
243 polyreactivity (Figure 4A, Supplementary Table 1). Index panel nanobodies partitioned  
244 into three groups according to their level of polyreactivity: minimal polyreactivity (light  
245 gray), moderate polyreactivity (gray), and high polyreactivity (dark gray). To validate the  
246 rank order of the 48 nanobodies we measured the polyreactivity of index panel members  
247 using PSR reagent derived from solubilized HEK293 cell membranes. We found that  
248 insect cell and HEK293 derived PSR staining are highly correlated ( $r^2 = 0.895$ ,  $p <$   
249 0.0001), indicating that polyreactivity levels do not vary with PSR reagent type

250 (Supplementary Figure 3C). Furthermore, to confirm that the rank order was not skewed  
251 by PSR binding to unfolded nanobodies on the surface of yeast, the index set was stained  
252 with an anti- $V_{HH}$  antibody, which recognizes the folded nanobody framework region  
253 (Supplementary Figure 3A). Levels of anti- $V_{HH}$  antibody staining are not correlated to  
254 insect cell PSR staining ( $r^2 = 0.046$ ,  $p = 0.1446$ , Supplementary Figure 3B), indicating that  
255 unfolded clones do not confound our dataset.

256 Biophysical characteristics of clones in our index set were reflective of the learned  
257 features in our high and low polyreactivity pools. There is a modest correlation between  
258 PSR staining of the index set and nanobody isoelectric point ( $r^2 = 0.390$ ,  $p < 0.0001$ ,  
259 Supplementary Figure 3D). While nanobodies with low isoelectric points possess low  
260 polyreactivity, nanobodies with high pl values demonstrate a range of polyreactivity.  
261 Similarly, nanobody hydrophobicity index values are not correlated with polyreactivity ( $r^2$   
262 = 0.036,  $p = 0.195$ , Supplementary Figure 3E).

263 Of the 48 nanobodies, 4 were previously seen in our training set, so we did not  
264 include these in our quantitative tests. Each of the 44 remaining nanobodies had at least  
265 6 mutations from any single nanobody sequence in the training set; the median of the  
266 minimum edit distance (a proxy for the number of mutations) of each of these index set  
267 nanobodies to the training set was 10 edit distance (the maximum similarity to the training  
268 set was 75% sequence identity). The correlation between the quantitative model  
269 predictions and the experimental binding scores to PSR, are strong - about 85% of the  
270 maximum theoretical correlation (Spearman  $\rho_s$  of 0.77 and 0.79, for the one-hot and k-  
271 mer logistic regression models, respectively) (Figure 3B). For comparison, the Spearman  
272 correlations between the three independent biological replicate experiments were 0.87,

273 0.87, and 0.95. Thus, our models trained on sequence pools of high and low polyreactivity  
274 nanobody CDR sequences are highly accurate for both classification and regression  
275 tasks for clones with distinct sequences.

276

277 **Model performance at predicting polyreactivity of closely related sequences**

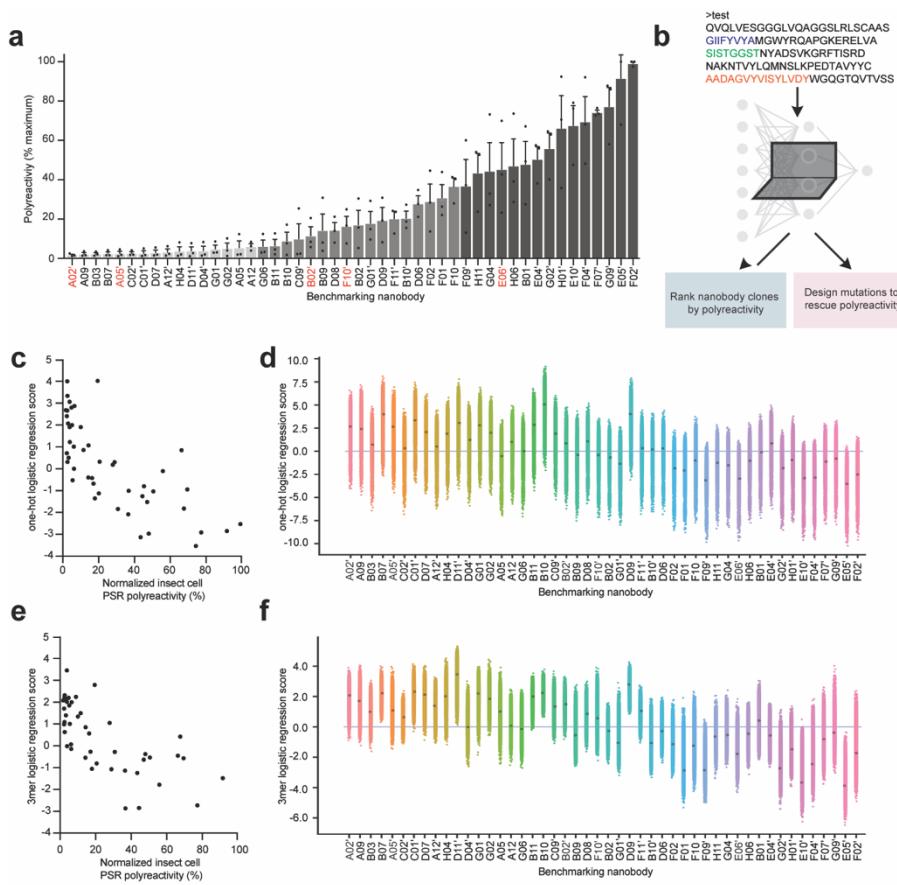
278 To determine if our computational model could accurately assess the influence of  
279 point mutations in single nanobody clones, we utilized the autonomous hypermutation  
280 yeast surface display (AHEAD) error-prone DNA replication system<sup>32</sup> to rapidly evolve the  
281 four most polyreactive clones from our index set (Nb E05', F02', G09', and F07') to have  
282 reduced binding to the PSR reagent. Over the course of four AHEAD cycles involving  
283 nanobody hypermutation and FACS sorting, global PSR staining of the evolved nanobody  
284 population decreased (Supplementary Figure 4). Deep sequencing analysis following the  
285 fourth FACS round revealed variation in the CDR regions of each of the four nanobodies.

286 A large proportion of the clones enriched by AHEAD are predicted to have reduced  
287 polyreactivity by both the one-hot and 3-mer logistic regression models. For the four  
288 clones, 97%, 67%, 69%, and 93% of the observed mutations are predicted to decrease  
289 polyreactivity by the one-hot logistic regression model, with similar decreases predicted  
290 by the k-mer logistic regression model (Supplementary Table 2). Furthermore, K31E<sup>36</sup>,  
291 A50T<sup>55</sup>, and R57P<sup>64</sup> substitutions that arose in nanobody E05' reflect the position specific  
292 analysis provided by the one-hot logistic regression model, where K, R, and A are  
293 characteristic of polyreactive nanobodies at positions 36, 55, and 64 and all three  
294 substitutions are characteristic of clones with reduced polyreactivity (Figure 3C). In a  
295 computational ranking of the polyreactivity of all 494 single amino acid substitutions using

296 the one-hot logistic regression model in the CDR regions of E05' found in our AHEAD  
297 experiment, from lowest to highest, R57P<sup>64</sup> ranked 28<sup>th</sup>, K31E<sup>36</sup> ranked 37<sup>th</sup>, and A50T<sup>55</sup>  
298 is 101<sup>st</sup>. Overall, the AHEAD-based directed evolution experiment produces clones that  
299 our computational models predict to have reduced polyreactivity suggesting that our  
300 models can accurately score the polyreactivity of closely related sequences.

301 With confidence in our models' performance on related clones, we employed our  
302 computational model to independently predict sequence substitutions to reduce  
303 polyreactivity of the highly polyreactive clone E10' and moderately polyreactive clone D06  
304 from our index set. We performed a comprehensive *in silico* single and double mutant  
305 scan, scored each sequence with both the one-hot logistic regression model and the k-  
306 mer logistic regression model (Figure 4B-D), and ranked all the possible single and  
307 double mutants, including insertions and deletions, surrounding the seed sequence. We  
308 sampled the substitutions most likely to reduce polyreactivity (with the exception of a  
309 substitution that would have introduced a cysteine that could disrupt disulfide bond  
310 formation) by selecting diverse mutations across residue types and positions that are  
311 contained within a single CDR and span each of the possible combinations of different  
312 CDR regions. Furthermore, if there was a mutation indicated to decrease polyreactivity  
313 by the k-mer logistic regression that scored similarly according to the one-hot logistic  
314 regression model, we selected the sequence with a higher k-mer logistic regression score  
315 to take into account local sequence dependencies. We selected the three top scoring  
316 single mutations for each of the CDR regions, the top scoring double mutants within a  
317 single CDR region, and the top scoring double mutants spanning two CDR regions where

318 at least one of the individual single mutations had not already been tested in a different  
 319 combination.



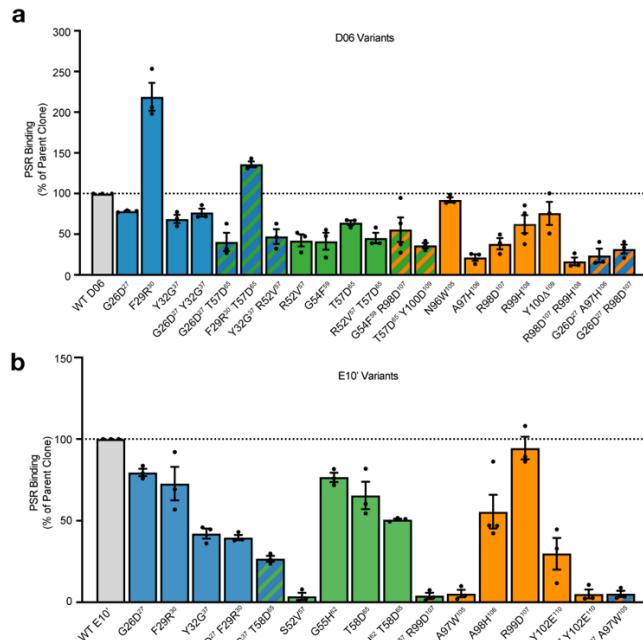
320  
 321 **Figure 4. Validation of computational model for quantitative predictions of**  
 322 **polyreactivity and design of rescue mutations.** **a**, Generation of an index panel of  
 323 **polyreactivity mutants by *Spodoptera frugiperda* (Sf9) insect cell membranes protein**  
 324 **polyspecificity reagent (PSR) staining of yeast displaying 48 unique nanobodies isolated**  
 325 **from MACS enrichment as well as non-reactive and polyreactive FACS pools. Data are**  
 326 **mean +/- SEM of three independent biological experiments performed in technical**  
 327 **triplicate. **b**, New nanobody sequence(s) can be input into a webserver, which will output**  
 328 **computational predictions of polyreactivity and biochemical properties of the sequence(s).**  
 329 **It is also possible to input a nanobody sequence to retrieve top scoring rescue mutations**  
 330 **predicted to decrease polyreactivity. **c, e**, The one-hot logistic regression model and k-**  
 331 **mer logistic regression model trained on the full NGS dataset from FACS sorts with PSR**  
 332 **binding were used to test quantitative predictions and rankings of the index set of clones**  
 333 **spanning a wide range of polyreactivity levels (as measured by PSR binding) (spearman**  
 334  $\rho_s$  **of 0.77 and 0.79, respectively). **d, f**, An *in silico* double mutation scan (spanning**  
 335 **substitutions, insertions, and deletions) was scored for predicted polyreactivity using both**  
 336 **the one-hot logistic regression model and k-mer logistic regression model. From these *in***  
 337 ***silico* double mutation scans, a diverse set (spanning each CDR and combinations of**

338 CDRs) of high scoring mutations predicted to have low polyreactivity were selected as  
339 rescue mutations for experimental testing from two parent clones, E10' and D06.

340  
341 For the moderately polyreactive D06 nanobody, 18 out of 21 variants that were  
342 computationally designed to decrease polyreactivity reduced levels of binding to insect  
343 cell PSR staining (Figure 5A). More stringently, 11 out of 21 mutations exhibited at least  
344 two-fold reductions in polyreactivity. Although substitutions in each of the CDR regions  
345 were able to lower polyreactivity, CDR3 appeared to drive polyreactivity as the most  
346 significant reductions in polyreactivity occurred from variations in the CDR3 region  
347 including A97H<sup>106</sup> and R98D<sup>107</sup> R99H<sup>108</sup>.

348 For the highly polyreactive E10' nanobody, 15 out of 16 computationally predicted  
349 single and double substitutions reduced binding to PSR reagent (Figure 5B). 9 out of the  
350 16 substitutions reduced polyreactivity by at least 50%, including mutations in each of the  
351 three CDR regions. Strikingly, the R99D<sup>107</sup> Y102E<sup>110</sup> clone, which was predicted to have  
352 the lowest polyreactivity value using the k-mer logistic regression model has very low  
353 polyreactivity by experimental PSR staining.

354

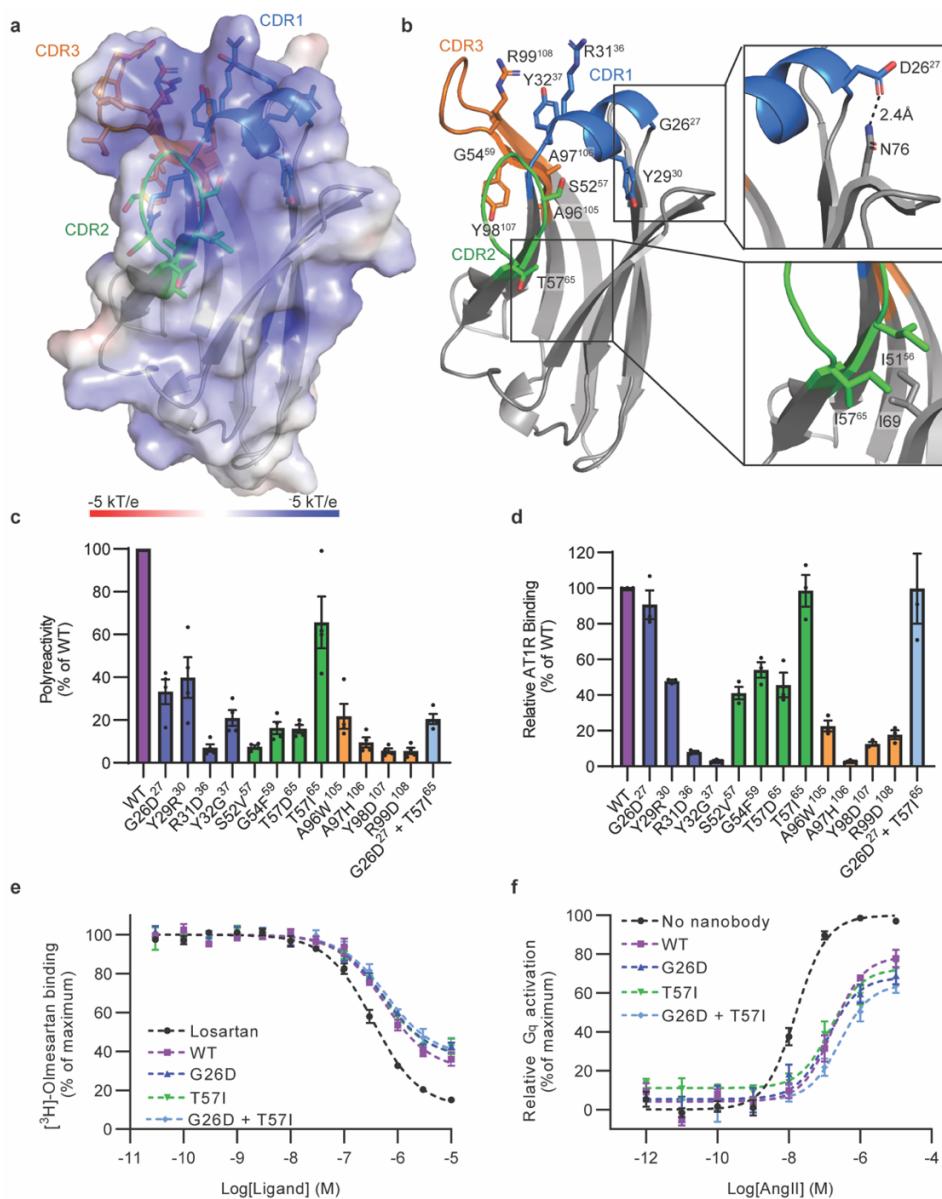


355  
356  
357 **Figure 5. In silico designed substitutions reduce nanobody polyreactivity. a,**  
358 Polyspecificity reagent (PSR) staining of yeast displaying D06 variants. For the  
359 moderately polyreactive D06 nanobody, 18 out of 21 variants that were computationally  
360 designed to decrease polyreactivity reduced levels of binding to insect cell PSR staining  
361 Data in a comprise the mean +/- SEM of at least three independent experiments, each  
362 performed in technical triplicate. **b**, PSR staining of yeast displaying E10' variants. For  
363 the highly polyreactive E10' nanobody, 15 out of 16 computationally predicted single and  
364 double substitutions reduced binding to PSR reagent. Data in b comprise the mean +/-  
365 SEM of at least three independent experiments, each performed in technical triplicate.  
366

### 367 Reducing polyreactivity of a functional clone

368 We next tested if our model could be employed to decrease the polyreactivity of  
369 nanobody clone that was independently selected for antigen specificity. AT118i4h32 is a  
370 nanobody antagonist for the angiotensin II type 1 receptor (AT1R), a G protein-coupled  
371 receptor (GPCR) that is a central regulator of blood pressure and renal function.  
372 AT118i4h32 directly competes with the binding of small molecule and peptide ligands to  
373 the AT1R and is active *in vivo*, reducing mouse blood pressure in a comparable degree  
374 to the clinically used angiotensin receptor blocker losartan<sup>28</sup>. Additionally, AT118i4h32

375 has been humanized with 11 amino acid substitutions to resemble a human V<sub>H</sub>3-23.  
376 Although pharmacologically intriguing, AT118i4h32 is highly polyreactive in the PSR  
377 assay and has a high pI value (9.6), which is characteristic of polyreactive antibodies.  
378 Furthermore, a crystal structure of AT118i4h32 displays large patches of positive charge  
379 on the protein surface (Figure 6a, Supplementary Table 3) and enrichment of both solvent  
380 exposed arginine and hydrophobic residues in the CDR regions (Figure



381

382 **Figure 6. Development of AT118i4h32 variants with reduced polyspecificity.** **a**,  
383 electrostatic surface of AT118i4h32. CDR1, CDR2, and CDR3 are colored blue, green,  
384 and orange. All positions substituted to produce variants of AT118i4h32 with reduced  
385 polyreactivity are shown in sticks with atomic coloring **b**, AT118i4h32 structure as colored  
386 in a. G26D<sup>27</sup> and T57I<sup>65</sup> substitutions are boxed. **c**, PSR staining of yeast displaying  
387 AT118i4h32 variants. All amino acid substitutions decrease polyreactivity. Data in c  
388 comprise the mean +/- SEM of four independent experiments, each performed in  
389 technical triplicate. **d**, binding of AT118i4h32 variants to HEK293 suspension cells  
390 expressing FLAG-AT1R. Cells were stained with AT118i4h32-V5-His variants,  
391 AlexaFlour-488 conjugated anti-FLAG, and AlexaFlour-647 conjugated anti-V5  
392 antibodies, then analyzed by flow cytometry. Data in d is the average of three independent  
393 experiments performed in technical triplicate, error bars are shown as SEM. **e**, radioligand  
394 competition binding of AT118i4h32 variants or the small molecule antagonist losartan and  
395 [<sup>3</sup>H]-olmesartan to AT1R in cell membranes. Like WT AT118i4h32, the G26D, T57I, and  
396 G26D+T57 variants compete with olmesartan for binding to the AT1R. Data in e is the  
397 average of three independent experiments performed in technical triplicate, error bars are  
398 shown as SEM. **f**, suppression of Gq-mediated inositol monophosphate production by  
399 AT118i4 in response to AngII stimulation. HEK293 suspension cells expressing FLAG-  
400 AT1R were treated with 5  $\mu$ M AT118i4h32 or no nanobody prior to AngII stimulation. Data  
401 in d is the average of three independent experiments performed in technical triplicate, error bars are shown as SEM.  $K_i$  values are reported in Supplementary Table 3.

403

404 We analyzed the sequence of AT118i4h32 and selected twelve single amino acid  
405 substitutions scattered throughout each CDR predicted to reduce polyreactivity based on  
406 the one-hot logistic regression model. AT118i4h32 variants were displayed on the surface  
407 of yeast and all showed reduced levels of PSR binding (Figure 6C). Neutralizing the highly  
408 basic patch composed of R30<sup>35</sup>, R31<sup>36</sup>, and R99<sup>108</sup> on the surface of AT118i4h32 (Figure  
409 6A) with R31D<sup>36</sup> and R99D<sup>108</sup> substitutions substantially reduces AT118i4h32  
410 polyreactivity. Notably, introduction of an additional arginine residue with the Y29R<sup>30</sup>  
411 substitution, which introduces a RRR sequence motif into CDR1, reduces polyreactivity,  
412 further demonstrating that arginine's contribution to polyreactivity is highly position  
413 dependent.

414 To assess the effects of these substitutions on antigen binding, AT118i4h32  
415 variants were recombinantly expressed in *E. coli* and purified to evaluate AT1R binding

416 by flow cytometry (Figure 6D). Two AT118i4h32 variants, G26D<sup>27</sup> and T57I<sup>65</sup>, retained at  
417 least 80% of wild-type binding levels to the AT1R. Combination of the G26D<sup>27</sup> and T57I<sup>65</sup>  
418 substitutions retained high levels of binding to the AT1R and yielded a clone with a modest  
419 decrease in PSR binding compared to the G26D<sup>27</sup> variant (Figure 6C), bringing the overall  
420 level of polyreactivity close to that of the clinically approved nanobody drug  
421 Cablivi/caplacizumab<sup>33</sup> (Supplementary Figure 5A). Additionally, the G26D<sup>27</sup>, T57I<sup>65</sup>  
422 variant has reduced polyreactivity compared to the wild-type nanobody as measured by  
423 ELISA assay (Supplementary Figure 5B-G). AT118i4h32 variants containing G26D<sup>27</sup> and  
424 T57I<sup>65</sup> maintain the ability to act as receptor antagonists, displacing small molecule  
425 orthosteric antagonists (Figure 6E) and suppressing receptor signaling upon angiotensin  
426 II (AngII) stimulation (Figure 6F).

427 To investigate how the G26D<sup>27</sup> T57I<sup>65</sup> substitutions alter AT118i4h32's structure  
428 and contribute to reduce polyreactivity, we crystallized AT118i4h32 G26D<sup>27</sup> T57I<sup>65</sup> and  
429 solved the structure at 1.6 Å resolution (Figure 6B, Supplementary Table 3). The T57I<sup>65</sup>  
430 substitution is located at the end of CDR2. I57<sup>65</sup> forms more favorable hydrophobic  
431 interactions with neighboring I51<sup>56</sup> and I65 side chains than T57<sup>65</sup>. In the case of  
432 AT118i4h32, maintaining this hydrophobic interaction is essential for antigen recognition,  
433 as the T57D<sup>65</sup> substitution diminished AT1R binding two-fold (Figure 6D). While the T57I<sup>65</sup>  
434 mildly decreases polyreactivity, AT118i4h32 variants containing the T57I<sup>65</sup> substitutions  
435 had slightly decreased thermal stability (Supplementary Table 4), indicating that changes  
436 in reduced polyreactivity are not necessarily correlated with thermal stability.

437 Residue D26<sup>27</sup>, found at the N-terminus of helical CDR1, forms a hydrogen bond  
438 with the side chain of framework residue N76 in all eight copies of the nanobody in the

439 crystal structure's asymmetric unit (Figure 6B). This hydrogen bond rigidifies the CDR1  
440 position and may reduce the flexibility of the nanobody's CDR regions. Additionally, the  
441 G26D substitution improves AT118i4h32's stability; we observed a five-fold increase in  
442 AT118i4h32 G26D<sup>27</sup> yield from *E. coli* and a two degree increase in melting temperature  
443 of the G26D<sup>27</sup> variant (Supplementary Table 4) over wild-type levels. Corresponding  
444 G26D<sup>27</sup> substitutions reduced the polyreactivity of nanobodies D06 and E10'. Despite  
445 occurring in just 0.05% of sequences from the naïve repertoire of seven llamas<sup>34</sup> (1.12  
446 million unique nanobody sequences), the D27 substitution may be both beneficial and  
447 tolerated in many sequence contexts and may broadly reduce polyreactivity by reducing  
448 the conformational flexibility of the CDR regions<sup>35</sup>.

449

#### 450 **Expansion of computational method**

451 Upon examination of corresponding substituted positions in D06, E10', and  
452 AT118i4h32 we observe some substitutions reduce polyreactivity in all clones, such as  
453 G26D<sup>27</sup>, whereas other mutations dramatically reduced polyreactivity of some  
454 nanobodies (i.e., E10' A97W<sup>105</sup> and AT118i4h32 A96W<sup>105</sup>) while having little to no effect  
455 in another clone (i.e., D06 N96W<sup>105</sup>). This suggests that *position dependency is critical*  
456 *for polyreactivity*, which may be more accurately captured with a larger data set.  
457 Therefore, we sought to improve our *in silico* method with expanded sequencing data.  
458 Through additional rounds of FACS selection, we collected 1,221,800 unique low  
459 polyreactivity clones and 1,058,842 unique high polyreactivity clones. We trained our  
460 suite of supervised classification models on this extended dataset and included analysis

461 of an extra position at the end of CDR2, which has some variability in the synthetic  
462 nanobody library, but was not included in the initial analysis.

463 To test classification accuracy, we clustered the sequences into 10 clusters using  
464 a k-means algorithm for train/test splits, and again limited our training dataset to  
465 sequences with at least 10 mutations as compared to any sequence in the test sets. We  
466 achieved comparable classification AUCs to the logistic regression and RNN models  
467 trained on the original FACS sorts (one-hot logistic regression: 0.83, 3-mer logistic  
468 regression: 0.83, RNN: 0.84) (Supplementary Figure 6A). The convolutional neural  
469 network model received a significant performance boost (CNN: 0.83 compared to  
470 previously 0.78 AUC) (Supplementary Figure 6A). For the higher throughput dataset, we  
471 see that the models that capture more complexities in sequences, such as the CNN and  
472 RNN, have higher accuracies, suggesting that there are meaningful dependencies in  
473 nanobody sequences that contribute to polyreactivity beyond site-specific amino acid  
474 contributions and/or 3-mer motifs and would allow us to make more accurate predictions  
475 to reduce polyreactivity for individual sequences. Furthermore, for each of these models  
476 we see an improved correlation (Spearman R) of polyreactivity scores with the index set  
477 measurements (one-hot logistic regression: 0.87, 3-mer logistic regression: 0.86, CNN:  
478 0.88, RNN: 0.88) (Supplementary Figure 6B-E). The majority of substitutions applied to  
479 clones D06, E10', and AT118i4h32 are still predicted to decrease polyreactivity across  
480 the four models trained on the deeper FACS sequencing experiments (37, 37, 41, and 23  
481 out of 45 mutations for one-hot logistic regression, k-mer logistic regression, CNN, and  
482 RNN respectively; for the RNN in particular, most mutations that were not predicted to

483 decrease polyreactivity had very small changes in predicted signal, Supplementary Table  
484 6).

485 As a resource to the field, we provide open-access use of our polyreactivity  
486 prediction software on our webpage (<http://18.224.60.30:3000/>). The webserver allows  
487 users to input a nanobody sequence(s) in FASTA format and outputs the aligned  
488 nanobody sequence with IMGT numbering using ANARCI<sup>30</sup>, along with biochemical  
489 properties of the sequence, including isoelectric point, hydrophobicity, CDR definitions  
490 (IMGT), CDR lengths, and computational predictions of polyreactivity scores using the  
491 one-hot logistic regression models that were trained for the design of rescue mutations.

492

## 493 **DISCUSSION**

494 Previous work has identified some biophysical characteristics of polyreactivity, but  
495 these studies have generally been performed on relatively small sets of antibody  
496 sequences without an explicit attempt to improve polyreactivity properties. Here, we  
497 designed and conducted high-throughput experiments to capture diverse clones that were  
498 not influenced by other selection pressures, facilitating an unbiased analysis of nanobody  
499 polyreactivity. Starting with a large naïve synthetic library mimicking the llama  
500 immunological repertoire, we isolated large pools of high and low polyreactivity nanobody  
501 clones based upon binding to the mixed-protein PSR reagent. Our models are over 80%  
502 accurate in discriminating between clones with high and low polyreactivity (Figure 3B),  
503 rank levels of polyreactivity with high fidelity (Figure 4), and reliably identify amino acid  
504 substitutions that reduce polyreactivity (Figures 5 and 6C).

505 Since our models were built upon experiments that were intentionally designed to  
506 interrogate sequence contributions to polyreactivity, they are highly accurate at  
507 measuring polyreactivity. In accordance with previous studies, our deep dive results  
508 suggest that arginine generally promotes nanobody polyreactivity while glutamate and  
509 aspartate usually decrease polyreactivity. However, we find amino acid contributions to  
510 polyreactivity are highly position dependent and more nuanced than broad  
511 generalizations suggest. This finding is in agreement with a recent independent study that  
512 analyzed polyreactivity of a subset of antibodies<sup>17</sup>. Furthermore, our computational  
513 models' ability to accurately quantify polyreactivity from sequence identity constitutes a  
514 large step forward as we can diagnose and engineer away polyreactivity of existing  
515 clones. More complex models including the CNN and RNN models also allowed us to  
516 evaluate dependencies of amino acids in different locations in nanobodies to  
517 polyreactivity. We find such dependencies contribute to polyreactivity indicating that both  
518 local and global characteristics of nanobodies influence their degree of polyreactivity.

519 We provide to the community an easy-to-use webserver that encapsulates our  
520 computational methods. These methods can guide antibody discovery campaigns at  
521 many points in the discovery pipeline. For instance, our software can be used to  
522 prospectively predict amino acid substitutions that will reduce polyreactivity of a single  
523 clone such as AT118i4h32. Moreover, the polyreactivity of a list of antigen binders can  
524 be ranked for clone prioritization during selection campaigns. We found that substitutions  
525 in each of the CDR regions of D06, E10', and AT118 reduce polyreactivity, suggesting  
526 that each CDR region contributes to polyreactivity. Therefore, if a certain CDR region is  
527 critical for antigen recognition, substitutions in alternative CDR regions can potentially

528 compensate in reducing polyreactivity. In addition, our success in reducing polyreactivity  
529 of AT118i4h32, where the humanized framework region differs from clones in the training  
530 set, indicates that our methods are applicable to nanobodies from a range of sources.  
531 Although outside the scope of this manuscript, similar approaches can be applied to  
532 conventional antibodies, adding in the three light-chain CDRs and germline gene choice  
533 as additional factors for polyreactivity prediction and optimization.

534

### 535 **Statistical Methods**

536 Prism software (Graphpad) was used to analyze data and perform error calculations. Data  
537 are expressed as arithmetic / geometric mean  $\pm$  SEM or arithmetic / geometric mean  $\pm$   
538 SD.

539

### 540 **Data Code Availability Statement**

541 The code for scoring new sequences for polyreactivity, designing rescue mutations,  
542 training polyreactivity models, and calculating biochemical properties of a sequence can  
543 be found on github: <https://github.com/debbiemarkslab/nanobody-polyreactivity>, and the  
544 webserver is available here: (<http://18.224.60.30:3000/>). Coordinates and structure  
545 factors for the AT118i4h32 structures are deposited in the Protein Data Bank under  
546 accession codes 7T83 and 7T84.

547

### 548 **ACKNOWLEDGMENTS**

549 This work was funded by a Merck Postdoctoral Fellowship from the Helen Hay Whitney  
550 Foundation to M.A.S.; NIH training grant 5T32GM132089-03 to V.G.M; NIH TR01 grant

551 1R01CA260415 to C.C.L, D.S.M., and A.C.K; 5R21HD101596 to A.C.K.; the Moore  
552 Inventor Fellowship to C.C.L. We thank Dr. Laura Wingler and Dr. Dean Staus for  
553 providing AT118i4h32 for crystallization experiments and Dr. Marie Bao for critical reading  
554 of the manuscript. We thank the staff at Advanced Photon Source GM/CA beamlines for  
555 support of X-ray data collection. GM/CA@APS is funded by the National Cancer Institute  
556 (ACB-12002) and the National Institute of General Medical Sciences (AGM-12006,  
557 P30GM138396). The Eiger 16M detector at GM/CA-XSD was funded by NIH grant S10  
558 OD012289. Portions of this research was conducted at the Advanced Photon Source, a  
559 U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE  
560 Office of Science by Argonne National Laboratory under Contract No. DE-AC02-  
561 06CH11357. We thank SBGrid Consortium for structural biology software support. D.S.F.  
562 experiments were carried out at the Center for Macromolecular Interactions in the  
563 Department of Biological Chemistry and Molecular Pharmacology at Harvard Medical  
564 School with support from Dr. Kelly Arnett.

565

## 566 **AUTHOR CONTRIBUTIONS**

567 M.A.S., E.P.H., J.S., D.S.M., A.C.K designed research. M.A.S. and E.P.H. performed  
568 MACS and FACS selections. E.P.H., M.A.S., and G.R.N. analyzed nanobody  
569 polyreactivity. J.S. and A.Y.S. designed computational algorithm. A.W. performed  
570 AHEAD experiments under the supervision of C.C.L. J.S. and E.P.H. analyzed AHEAD  
571 evolution experiments. G.R.N., J.H., E.P.H. and M.A.S purified nanobody variants. E.P.H.  
572 and J.H. performed nanobody size exclusion chromatography, E.P.H., J.H., and V.M.  
573 developed and ran ELISA assays. E.P.H. and J.H. performed and analyzed anti-

574 nanobody antibody staining experiments. J.K.M. and A.Y.S. designed webserver. M.A.S.  
575 generated PSR reagent, performed mammalian cell binding, thermal stability, radioligand  
576 binding, and AT1R signaling assays. M.A.S. and G.R.N. determined the crystal structures  
577 of AT118i4h32. M.A.S., E.P.H., and J.S. wrote the manuscript with input from all authors.  
578

579 **COMPETING INTERESTS STATEMENT**

580 C.C.L is a co-founder of K2 Biotechnologies Inc., which applies continuous evolution  
581 technologies to antibody engineering. D.S.M. is an advisor for Dyno Therapeutics, Octant,  
582 Jura Bio, Tectonic Therapeutic and Genetech, and is a co-founder of Seismic  
583 Therapeutic. A.C.K. is a co-founder and consultant for Tectonic Therapeutic and Seismic  
584 Therapeutic and for the Institute for Protein Innovation, a non-profit research institute.

585 **REFERENCES**

586

587 1 Sigounas, G., Harindranath, N., Donadel, G. & Notkins, A. L. Half-life of  
588 polyreactive antibodies. *J Clin Immunol* **14**, 134-140, doi:10.1007/BF01541346  
589 (1994).

590 2 Kelly, R. L. *et al.* High throughput cross-interaction measures for human IgG1  
591 antibodies correlate with clearance rates in mice. *MAbs* **7**, 770-777,  
592 doi:10.1080/19420862.2015.1043503 (2015).

593 3 Cunningham, O., Scott, M., Zhou, Z. S. & Finlay, W. J. J. Polyreactivity and  
594 polyspecificity in therapeutic antibody development: risk factors for failure in  
595 preclinical and clinical development campaigns. *MAbs* **13**, 1999195,  
596 doi:10.1080/19420862.2021.1999195 (2021).

597 4 Berglund, L. *et al.* A genecentric Human Protein Atlas for expression profiles  
598 based on antibodies. *Mol Cell Proteomics* **7**, 2019-2027,  
599 doi:10.1074/mcp.R800013-MCP200 (2008).

600 5 Baker, M. Reproducibility crisis: Blame it on the antibodies. *Nature* **521**, 274-276,  
601 doi:10.1038/521274a (2015).

602 6 Bradbury, A. & Pluckthun, A. Reproducibility: Standardize antibodies used in  
603 research. *Nature* **518**, 27-29, doi:10.1038/518027a (2015).

604 7 Frese, K., Eisenmann, M., Ostendorp, R., Brocks, B. & Pabst, S. An automated  
605 immunoassay for early specificity profiling of antibodies. *MAbs* **5**, 279-287,  
606 doi:10.4161/mabs.23539 (2013).

607 8 Wardemann, H. *et al.* Predominant autoantibody production by early human B  
608 cell precursors. *Science* **301**, 1374-1377, doi:10.1126/science.1086907 (2003).

609 9 Mouquet, H. *et al.* Polyreactivity increases the apparent affinity of anti-HIV  
610 antibodies by heteroligation. *Nature* **467**, 591-595, doi:10.1038/nature09385  
611 (2010).

612 10 Lueking, A. *et al.* A nonredundant human protein chip for antibody screening and  
613 serum profiling. *Mol Cell Proteomics* **2**, 1342-1349, doi:10.1074/mcp.T300001-  
614 MCP200 (2003).

615 11 Kelly, R. L. *et al.* Chaperone proteins as single component reagents to assess  
616 antibody nonspecificity. *MAbs* **9**, 1036-1040,  
617 doi:10.1080/19420862.2017.1356529 (2017).

618 12 Hotzel, I. *et al.* A strategy for risk mitigation of antibodies with fast clearance.  
619 *MAbs* **4**, 753-760, doi:10.4161/mabs.22189 (2012).

620 13 Jacobs, S. A., Wu, S. J., Feng, Y., Bethea, D. & O'Neil, K. T. Cross-interaction  
621 chromatography: a rapid method to identify highly soluble monoclonal antibody  
622 candidates. *Pharm Res* **27**, 65-71, doi:10.1007/s11095-009-0007-z (2010).

623 14 Xu, Y. *et al.* Addressing polyspecificity of antibodies selected from an in vitro  
624 yeast presentation system: a FACS-based, high-throughput selection and  
625 analytical tool. *Protein Eng Des Sel* **26**, 663-670, doi:10.1093/protein/gzt047  
626 (2013).

627 15 Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape.  
628 *Proc Natl Acad Sci U S A* **114**, 944-949, doi:10.1073/pnas.1616408114 (2017).

629 16 Shehata, L. *et al.* Affinity Maturation Enhances Antibody Specificity but  
630 Compromises Conformational Stability. *Cell Rep* **28**, 3300-3308 e3304,  
631 doi:10.1016/j.celrep.2019.08.056 (2019).

632 17 Boughter, C. T. *et al.* Biochemical patterns of antibody polyreactivity revealed  
633 through a bioinformatics-based analysis of CDR loops. *Elife* **9**,  
634 doi:10.7554/elife.61393 (2020).

635 18 Kelly, R. L., Le, D., Zhao, J. & Wittrup, K. D. Reduction of Nonspecificity Motifs in  
636 Synthetic Antibody Libraries. *J Mol Biol* **430**, 119-130,  
637 doi:10.1016/j.jmb.2017.11.008 (2018).

638 19 Kelly, R. L., Zhao, J., Le, D. & Wittrup, K. D. Nonspecificity in a nonimmune  
639 human scFv repertoire. *MAbs* **9**, 1029-1035,  
640 doi:10.1080/19420862.2017.1356528 (2017).

641 20 Tiller, K. E. *et al.* Arginine mutations in antibody complementarity-determining  
642 regions display context-dependent affinity/specificity trade-offs. *J Biol Chem* **292**,  
643 16638-16652, doi:10.1074/jbc.M117.783837 (2017).

644 21 Birtalan, S. *et al.* The intrinsic contributions of tyrosine, serine, glycine and  
645 arginine to the affinity and specificity of antibodies. *J Mol Biol* **377**, 1518-1528,  
646 doi:10.1016/j.jmb.2008.01.093 (2008).

647 22 Bumbaca Yadav, D. *et al.* Evaluating the Use of Antibody Variable Region (Fv)  
648 Charge as a Risk Assessment Tool for Predicting Typical Cynomolgus Monkey  
649 Pharmacokinetics. *J Biol Chem* **290**, 29732-29741, doi:10.1074/jbc.M115.692434  
650 (2015).

651 23 Lecerf, M., Kanyavuz, A., Lacroix-Desmazes, S. & Dimitrov, J. D. Sequence  
652 features of variable region determining physicochemical properties and  
653 polyreactivity of therapeutic antibodies. *Mol Immunol* **112**, 338-346,  
654 doi:10.1016/j.molimm.2019.06.012 (2019).

655 24 Rabia, L. A., Zhang, Y., Ludwig, S. D., Julian, M. C. & Tessier, P. M. Net charge  
656 of antibody complementarity-determining regions is a key predictor of specificity.  
657 *Protein Eng Des Sel* **31**, 409-418, doi:10.1093/protein/gzz002 (2018).

658 25 Zhang, Y. *et al.* Physicochemical Rules for Identifying Monoclonal Antibodies  
659 with Drug-like Specificity. *Mol Pharm* **17**, 2555-2569,  
660 doi:10.1021/acs.molpharmaceut.0c00257 (2020).

661 26 McMahon, C. *et al.* Yeast surface display platform for rapid discovery of  
662 conformationally selective nanobodies. *Nat Struct Mol Biol* **25**, 289-296,  
663 doi:10.1038/s41594-018-0028-6 (2018).

664 27 Schoof, M. *et al.* An ultrapotent synthetic nanobody neutralizes SARS-CoV-2 by  
665 stabilizing inactive Spike. *Science* **370**, 1473-1479, doi:10.1126/science.abe3255  
666 (2020).

667 28 McMahon, C. *et al.* Synthetic nanobodies as angiotensin receptor blockers. *Proc  
668 Natl Acad Sci U S A* **117**, 20284-20291, doi:10.1073/pnas.2009029117 (2020).

669 29 Jovcevska, I. & Muyldermans, S. The Therapeutic Potential of Nanobodies.  
670 *BioDrugs* **34**, 11-26, doi:10.1007/s40259-019-00392-z (2020).

671 30 Dunbar, J. & Deane, C. M. ANARCl: antigen receptor numbering and receptor  
672 classification. *Bioinformatics* **32**, 298-300, doi:10.1093/bioinformatics/btv552  
673 (2016).

674 31 Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of  
675 applying machine learning in genomics. *Nat Rev Genet*, doi:10.1038/s41576-  
676 021-00434-9 (2021).  
677 32 Wellner, A. *et al.* Rapid generation of potent antibodies by autonomous  
678 hypermutation in yeast. *Nat Chem Biol*, doi:10.1038/s41589-021-00832-4 (2021).  
679 33 Scully, M. *et al.* Caplacizumab Treatment for Acquired Thrombotic  
680 Thrombocytopenic Purpura. *N Engl J Med* **380**, 335-346,  
681 doi:10.1056/NEJMoa1806311 (2019).  
682 34 McCoy, L. E. *et al.* Molecular evolution of broadly neutralizing Llama antibodies  
683 to the CD4-binding site of HIV-1. *PLoS Pathog* **10**, e1004552,  
684 doi:10.1371/journal.ppat.1004552 (2014).  
685 35 Prigent, J. *et al.* Conformational Plasticity in Broadly Neutralizing HIV-1  
686 Antibodies Triggers Polyreactivity. *Cell Rep* **23**, 2568-2581,  
687 doi:10.1016/j.celrep.2018.04.101 (2018).  
688