




REPORT



## Effective binding to protein antigens by antibodies from antibody libraries designed with enhanced protein recognition propensities

Jih-Wei Jian <sup>a,b,c,\*</sup>, Hong-Sen Chen<sup>a\*</sup>, Yi-Kai Chiu<sup>a</sup>, Hung-Pin Peng <sup>a</sup>, Chao-Ping Tung <sup>a</sup>, Ing-Chien Chen<sup>a</sup>, Chung-Ming Yu<sup>a</sup>, Yueh-Liang Tsou<sup>a</sup>, Wei-Ying Kuo <sup>a</sup>, Hung-Ju Hsu<sup>a</sup>, and An-Suei Yang <sup>a\*</sup>

<sup>a</sup>Genomics Research Center, Academia Sinica, Taipei, Taiwan; <sup>b</sup>Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan; <sup>c</sup>Bioinformatics Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei, Taiwan

### ABSTRACT

Antibodies provide immune protection by recognizing antigens of diverse chemical properties, but elucidating the amino acid sequence-function relationships underlying the specificity and affinity of antibody-antigen interactions remains challenging. We designed and constructed phage-displayed synthetic antibody libraries with enriched protein antigen-recognition propensities calculated with machine learning predictors, which indicated that the designed single-chain variable fragment variants were encoded with enhanced distributions of complementarity-determining region (CDR) hot spot residues with high protein antigen recognition propensities in comparison with those in the human antibody germline sequences. Antibodies derived directly from the synthetic antibody libraries, without affinity maturation cycles comparable to those in *in vivo* immune systems, bound to the corresponding protein antigen through diverse conformational or linear epitopes with specificity and affinity comparable to those of the affinity-matured antibodies from *in vivo* immune systems. The results indicated that more densely populated CDR hot spot residues were sustainable by the antibody structural frameworks and could be accompanied by enhanced functionalities in recognizing protein antigens. Our study results suggest that synthetic antibody libraries, which are not limited by the sequences found in antibodies in nature, could be designed with the guidance of the computational machine learning algorithms that are programmed to predict interaction propensities to molecules of diverse chemical properties, leading to antibodies with optimal characteristics pertinent to their medical applications.

### ARTICLE HISTORY

Received 7 September 2018  
Revised 5 November 2018  
Accepted 16 November 2018

### KEYWORDS

antibody engineering;  
synthetic antibody library;  
antibody-antigen affinity  
prediction; anti-HER2  
antibodies; affinity  
maturation; hot spot  
residues for antibody-  
protein interactions

## Introduction

One major biological system that defends individuals against diverse immunogens relies on rapidly developing high affinity immunogen-specific antibodies from the individual's naïve B cell receptor (BCR) repertoire through clonal selection and affinity maturation. A repertoire of naïve BCRs are encoded in a large number of B cells, each of which expresses a sequence-wise unique BCR through antibody gene segment recombination and segment junction diversification.<sup>1</sup> However, it has been estimated that more than half of all human BCRs expressed by early immature B cells, where the BCR sequences are the same or closely related to the germline sequences, are polyreactive; the BCR polyreactivity was defined as that the BCR bound to at least two structurally and chemically distinguished antigens, albeit with relatively low affinity compared with affinity-matured antibodies by orders of magnitude in dissociation constant.<sup>2–6</sup> The polyreactivity of the immature BCRs is likely to be a prerequisite condition bridging the gap between the humoral innate immunity and the adaptive immunity, allowing the adaptive immune system to rapidly develop humoral protection by matured antibodies through finite cycles of stochastic somatic hyper mutation (SHM)<sup>7</sup> and clonal selection<sup>1</sup> of the BCRs. In

light of the notion that the naïve antibody repertoires decoupled from the *in vivo* affinity maturation processes might not be productive sources for highly specific antibodies against diverse antigens, the goal of our study was to address the question of practical importance as to what sequences in the complementarity-determining regions (CDRs) of antibodies could confer high affinity and specificity on the antibodies against their cognate antigens, with the focus of engineering antibody-antigen interaction hot spot residues in the CDRs.

Antibodies provide immune protections by recognizing antigens with remarkable affinity and exquisite specificity, in large part through key antibody-antigen interface residues known as hot spot residues to reflect their binding energetics contributions. Surveys of general protein-protein interactions (PPIs) indicate that the hot spot residues in the interfaces contribute substantially to the energetics of the PPIs.<sup>8–10</sup> Antibodies recognize antigens with sub-nanomolar affinity through the aromatic residue-enriched CDRs,<sup>11–14</sup> which are populated with hot spot residues that contribute substantially to the antibody-antigen interaction energy.<sup>15–17</sup>

Study of the functionalities of the hot spot residues of antibodies in nature have helped establish an understanding

**CONTACT** An-Suei Yang  [yangas@gate.sinica.edu.tw](mailto:yangas@gate.sinica.edu.tw)  Genomics Research Center, Academia Sinica, 128 Academia Rd., Sec.2, Nankang Dist., Taipei 115, Taiwan

\*These authors contribute equally.

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2019 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

of antibody-protein recognitions,<sup>18</sup> but methods to engineer CDR hot spot residues to enhance antibody-antigen recognitions have not been established. It is not known if the engineered CDR hot spot residues are functionally sustainable in the antibody structure frameworks, and if more densely populated CDR hot spot residues above the hot spot residue distribution level in antibodies in nature could be accompanied by enhanced functionalities for the engineered antibodies to recognize antigens. Moreover, identifying hot spot residues on a vast number of antibodies is intractable with experimental methods. An alternative method feasible for large-scale hot spot residue evaluations is thus indispensable for designing and engineering hot spot residues on antibodies.

To address these questions, we constructed phage-displayed synthetic antibody libraries with single-chain variable fragment (scFv) variants encoded with densely enhanced CDR hot spot residues and tested the functionalities of these synthetic antibody libraries against protein antigens. The synthetic antibody scFv libraries were constructed with oligonucleotide-directed mutagenesis;<sup>19-24</sup> the enriched antigen-recognition determinants on the scFv variants were calculated with the ISMBLab package, which is a collection of machine learning predictors for quantitative antigen-recognition propensities on the antigen binding sites of the antibodies.<sup>18,24-29</sup> The antigen-recognition propensity computation indicated that the scFv variants of the designed synthetic antibody libraries were encoded with multiple folds of CDR hot spot residues with high protein antigen recognition propensities compared with those of the human antibody germline sequences. Selected anti-protein antibodies from the synthetic antibody libraries were highly specific against the corresponding protein antigens with sub-nanomolar affinity without *in vivo* affinity maturation, indicating that the antibodies encoded with enhanced population of CDR hot spot residues could bind to protein antigens with high specificity and affinity, bypassing the *in vivo* affinity maturation processes involving somatic hyper mutations and clonal selections.

As antibodies are becoming the most prominent class of protein therapeutics,<sup>30</sup> better understanding of the principles governing antibody affinity and specificity will facilitate in understanding humoral immunity and in developing novel antibody-based therapeutics. Computational and experimental results indicate that the antibody CDR sequences alternative to human germline sequences can be rationally directed to recognize antigens with high specificity and affinity. Molecular understanding of antibody specificity and affinity could lead to antibodies with optimal characteristics pertinent to their medical applications.

## Results

### **Synthetic antibody library constructs based on the prominent canonical CDR structures in the antibodies with known structure in protein data bank**

We designed and constructed three sets of antibody libraries based on the CDR sequence length configurations of the most prominent antibody structures found in the antibodies with known structure in Protein Data Bank (PDB). Antibody

variable domain fragment structures archived in PDB mostly bind to four classes of antigen: protein, peptide, carbohydrate, or small molecule/hapten. By removing redundant antibody structures with pairwise sequence identity above 95% and other criteria (Supplementary Methods), we attained four antibody-antigen complex structure datasets (Figure S1 A): Ab-PRO (281 antibody-protein complex structures), Ab-PEP (111 antibody-peptide complex structures), Ab-CARB (14 antibody-carbohydrate complex structures), and Ab-LIG (106 antibody-hapten complex structures).

The CDR H1-H2-L1-L2-L3 length configuration 13-10-11-8-9 (the definition of the CDRs follows the work by North et al.<sup>31</sup>) with the canonical structure configuration 1-2-2-1-1 as defined by Chothia and coworkers<sup>32,33</sup> is the most prominent antibody structure class in these antibody-antigen complex structures (Figure S1 B ~ E), as well as in human antibody repertoires and known antibody sequence databases (Figure S2 A ~ C). The peak of the CDR-H3 length distribution of these 13-10-11-8-9 antibodies of known structure is centered at 12 ~ 13 residues (Figure S2 D).

In addition to the 13-10-11-8-9 antibodies, the CDR length configuration 13-10-16-8-9 with the main canonical structure configuration 1-2-4-1-1 as defined by Chothia<sup>32,33</sup> is found most frequently in the antibody-peptide (Ab-PEP), antibody-carbohydrate (Ab-CARB), and antibody-hapten (Ab-LIG) complexes (Figure S1 A, C ~ E). Longer CDR-H3s (> 17 residues) do not seem compatible with this class of antibodies (Figure S2 D).

Based on these prominent antibody structure classes, three sets of synthetic antibody libraries with increasing structural complexity were designed and constructed: set(1) GH2-13 (one antibody library): 13-10-11-8-9 antibodies with 13-residue CDR-H3s (CDR sequence designs shown in Tables S1 and S2); set(2) GH2-6 ~ 20 (11 antibody libraries): 13-10-11-8-9 antibodies with 6 ~ 20-residue CDR-H3s, excluding 13-residue CDR-H3s (CDR sequence designs shown in Tables S1 and S3); set(3) GH3-6 ~ 13 (8 libraries): 13-10-16-8-9 and 13-10-17-8-9 antibodies with 6 ~ 13-residue CDR-H3s (CDR sequence designs shown in Tables S4 and S5). The CDRs of the scFv variants in the synthetic antibody libraries were designed to encode densely enhanced CDR hot spot residues, as analyzed in a later section (see the section: *Hot spot residues are substantially more densely distributed in the CDRs of the synthetic antibody library scFv variants*). One key feature of these synthetic antibody libraries is that each of the libraries contains only one or two CDR H1-H2-L1-L2-L3 length configurations with fixed CDR-H3 length, such that the functional contributions from the CDR lengths can be elucidated respectively.

These 20 synthetic antibody libraries were constructed and expressed as phage-displayed scFv libraries based on the framework of the VL/VH combination IGKV1-NL1/IGHV3-23 with the well-established method previously published<sup>19,21</sup> (see Methods and Supplementary Methods). The phage-displayed synthetic antibody library construction procedure has been validated with next-generation sequencing (NGS).<sup>19</sup> Each of the synthetic scFv libraries contained at least 10<sup>9</sup> well-folded scFv variants (Protein A/L binding to the IGKV1-NL1/IGHV3-

23 framework as indication for the structural stability of the scFv variants<sup>22,23</sup>) with sequence diversified in all the 6 CDRs (Tables S1~ S5).

### **Experimental validations of the phage-displayed synthetic antibody libraries**

The phage-displayed synthetic antibody libraries were first experimentally validated with scFvs isolated from the antibody libraries binding to 22 randomly selected protein antigens (Table S6). After this first general test of the synthetic antibody libraries, we further tested the synthetic antibody libraries on one antigen (human epidermal growth factor receptor 2 extracellular domain (HER2-ECD)) with more comprehensive analyses of the scFv variants and the IgGs reformatted from the selected scFv candidates, as described in a later section (see the section: *Antibodies from the synthetic antibody libraries bind to HER2-ECD through diverse epitopes with high affinity and specificity without affinity maturation*). The experimental procedures for the biopanning of the phage-displayed scFv libraries against protein antigens and the binding evaluation of the isolated soluble scFv molecules to the corresponding antigen have been standardized and published previously<sup>19,21</sup> (Supplementary Methods). To evaluate the general functionalities of the synthetic antibody libraries, we carried out the standard biopanning with the synthetic antibody libraries against randomly selected protein antigens and evaluated the binding of the resultant scFvs against the protein antigens. We did not enumerate all 20(libraries)× 22(antigens) possible library-antigen biopannings; rather, we carried out 118 randomly selected antibody discovery campaigns. Only 13 of 118 biopannings failed to generate any positive binders; 105 of 118 antibody discovery campaigns yielded positive scFv binders against the corresponding protein antigen with stable native scFv structure. The number of non-redundant scFv binders and the number of positive clones sequenced are shown for each of the antibody discovery campaigns in Table S6. The increases of the binding affinities of the polyclonal soluble scFvs of the output phage libraries from each round of the biopannings are also shown for each of the antibody discovery campaigns in Figure S3. Overall, as shown in Table S6, scFv binders from each of the synthetic antibody libraries were attained from the standard biopanning against at least one randomly selected protein antigens, and binding scFvs against each of the protein antigens were attainable from at least one randomly selected synthetic antibody library. Moreover, 18 of the 20 antibody libraries tested were proven to contain specific scFv binders against more than one randomly selected protein antigens; one antibody library (GH2-13) was tested on 20 randomly selected protein antigens and only one antigen failed to generate a specific scFv binder (Table S6). These experimental validations, in terms of generating multiple and specific scFv binders with stable native scFv structure against the corresponding randomly selected antigens, suggested that each of the GH synthetic antibody libraries contains well-folded scFv variants binding to randomly selected protein antigens in general.

### **Hot spot residues are substantially more densely distributed in the CDRs of the synthetic antibody library ScFV variants**

In order to compare the hot spot residue distributions on antibodies in nature versus those on the antibodies from the synthetic antibody libraries, we predicted hot spot residues on the scFv structures with the machine learning methods that we have previously developed for predicting the PPI propensities of protein surface atoms<sup>18</sup> (ISMBLab-PPI, see Methods and Supplementary Methods). A query scFv structure derived experimentally or computationally is the only required input for the prediction of the atomistic interaction propensities of the query antibody surface atoms to be involved in a combination site for a protein antigen. The output of the predictors for each of the query antibody surface atoms is normalized into a prediction confidence level (PCL) ranging from 0 to 1, which represents the atomistic propensity for the query antibody surface atom to interact with a protein antigen. Note that the atomistic propensities were calculated only for protein surface atoms; any solvent-inaccessible atom has zero propensity to interact with the protein antigen. A residue on the query antibody structure with maximal atomistic propensity  $\geq 0.45$  is predicted as a hot spot residue. The predictions are correlated with experimentally determined hot spot residues defined by the threshold of  $\Delta\Delta G \geq 1$  kcal/mol in alanine-scanning experiments with Matthews correlation coefficient of 0.43 and F1 score of 0.51.<sup>18</sup> While the alanine-scanning of hot spot residues is experimentally intractable for the large number of scFv variants from the synthetic antibody libraries and from antibodies in nature, the computational hot spot predictions provide an alternative for evaluating the hot spot residue distributions in the CDRs of the scFvs, albeit with uncertainty in prediction accuracy to an extent.

Although the CDR sequence length configurations of the synthetic antibody libraries resemble those of the prominent antibody structures (see above), the CDRs of the synthetic antibody libraries are much more densely enriched with hot spot residues than those of the antibodies with known structures in PDB. In Table S6, each scFv library had two sets of scFv sequences collected from the validation experiments described in the previous section: set(F) – scFv sequences known to fold properly (Protein A/L binding; numbers of non-redundant sequences as shown in the column labelled ProA/L in Table S6; total 1819 sequences), and set(FB) – scFv sequences known to fold and to bind to the corresponding protein antigen (binding to Protein A/L and the corresponding antigens; numbers of non-redundant sequences as shown in the columns labelled by the corresponding antigen names in Table S6; total 1563 sequences). In addition, for each of the scFv libraries, the scFv sequence set(D) contained 200 randomly selected theoretical scFv sequences based on the CDR designs (Tables S1~ S5). The 3D structures of these sets of scFvs (1819, 1563 and 4000 scFvs for set(F), set(FB) and set(D), respectively) were modeled with default FoldX (Supplementary Methods) and the CDR hot spot residue distributions were predicted with ISMBLab-PPI and compared with those of the Ab-PRO 13–10–11–8–9 antibody structures (90 entries) and the Ab-PRO 13–10–16/17–8–9

antibody structures (20 entries) (Figure 1). The CDR-L1~H2 in the scFvs of the set(D)s (blue box plots in Figure 1(a,c,e)) are designed with enhanced hot spot residues by several folds as compared with those in the corresponding antibody structures in the Ab-PRO dataset (green box plots in Figure 1(a,c,e)); P-value for the comparison of each pair of the blue box plot and green box plot is consistently less than  $10^{-5}$  (Figure 1(a,c,e)). The CDR-H3 hot spot residues increase with the CDR-H3 sequence length to average maximal 2 fold (blue box plots in Figure 1(b,d,f)) compared to those of the Ab-PRO antibodies (green box plots in Figure 1(b,d,f)); P-values for the comparisons of the pairs of the blue box plot and green box plot indicate that the synthetic antibody libraries contain similar or more hot spot distributions in the CDR-H3s in comparison with those in the CDR-H3s of the antibodies in nature (Figure 1(b,d,f)). These results indicate that the scFv variants in the synthetic antibody libraries were substantially enriched with CDR hot spot residues in comparison with those of the antibodies with known structures in PDB.

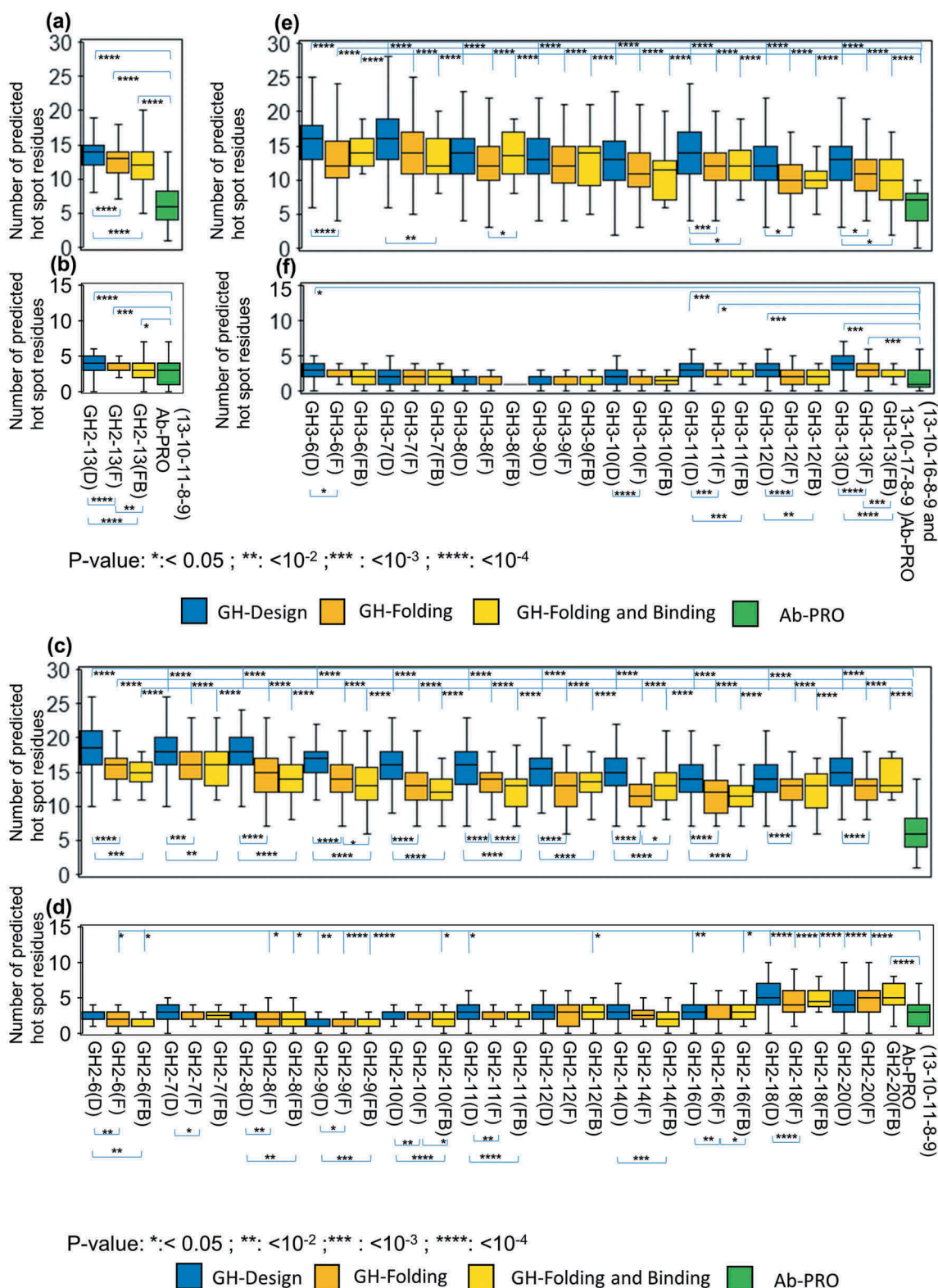
The CDR hot spot residues of the scFvs in set(F) (orange box plots in Figure 1) and set(FB) (yellow box plots in Figure 1) generally decrease in comparison with those in set(D) (blue box plots in Figure 1), with half of the individual decreasing trends statistically significant as defined by the P-value  $< 0.05$  shown in Figure 1, indicating that the folding and the binding requirements for the scFvs restrict the CDR hot spot residue distributions. The folding and binding sequence requirements are summarized in Figures S4~S9, where the sequence preference profiles for the well-folded scFvs (set(F), Figures S4, S6 and S8) and the folded scFvs binding to the corresponding antigens (set(FB), Figures S5, S7 and S9) are in parallel presented with the amino acid type distribution designs in each of the CDR positions of the synthetic antibody libraries. Indeed, as expected, the CDR sequence profiles of the scFvs in set(F) and set(FB) show prominent sequence features in the CDR sequence profiles of the scFvs (Figure S4~S9), especially in the scFvs of set(FB) (Figure S5, S7 and S9) likely due to the requirements of both folding of the scFv and binding of the CDRs to the corresponding protein antigens. The prominent sequence features in the CDR sequence profiles, which reflect the binding and/or folding requirements of the amino acid types at specific sequence positions, are likely to restrict the distributions of the CDR hot spot residues in these sequence positions, explaining the decreasing of the CDR hot spot residues in the scFvs of the set(F) and set(FB). Nevertheless, the functional synthetic scFvs that were folded and bound to protein antigens were frequently encoded with many more CDR hot spot residues (orange and yellow box plots in Figure 1) in comparison with those of the corresponding functional antibody structures in the Ab-PRO dataset (green box plots in Figure 1) with P-values less than  $10^{-5}$ , suggesting that the antibody variable domain CDRs have substantial structural tolerance for enhanced distributions of hot spot residues.

The CDR hot spot residue distributions on the scFv variants in the synthetic antibody libraries are different to an extent from those on human germline antibody variable domain sequences. Figure 2 shows the hot spot occurrence probabilities for each amino acid type at each of the 13-10-11-8-9 CDR positions in the scFvs of set(D) (Figure 2(b)),

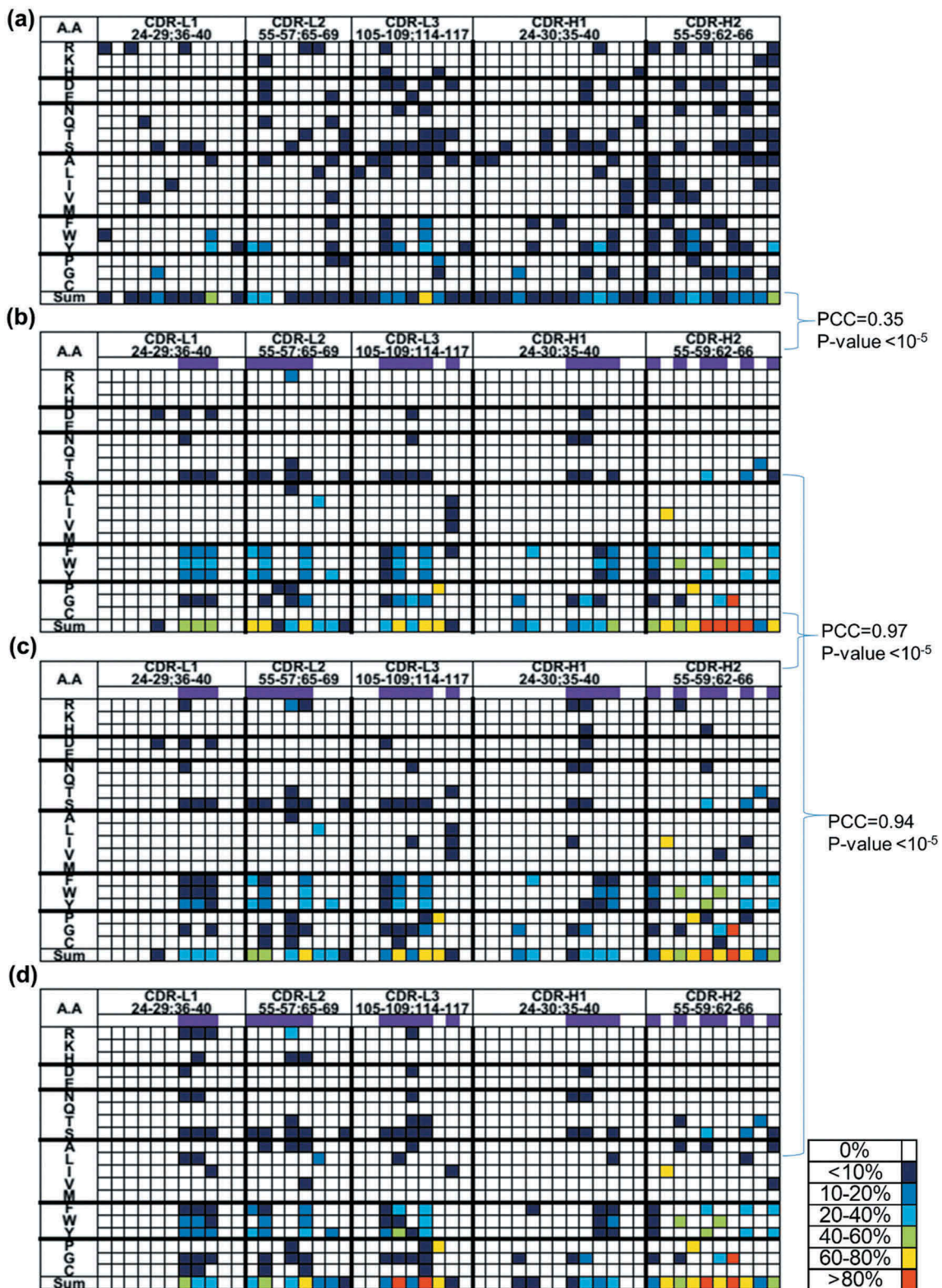
set(F) (Figure 2(c)) and set(FB) (Figure 2(d)) from the synthetic scFv libraries compared with those in the human germline antibody variable domain sequences of 13-10-11-8-9 CDR length configuration (Figure 2(a)). The scFvs from set(D), set(F) and set(FB) have highly similar distribution patterns (Pearson correlation coefficient  $> 0.9$ ) for the hot spot residue positions and amino acid types (Figure 2(b,c,d)), indicating that the folding and binding requirements of the functional scFvs did not severely restrict the general distributions of the designed hot spot residue positions and amino acid types; the synthetic scFvs have enhanced hot spot occurrence probabilities in the CDR with comparable position distributions as in human germline antibody variable domains (Figure 2(a)) of the same CDR length configuration. This conclusion is also valid for the 13-10-17-8-9 and 13-10-16-8-9 scFvs from the synthetic antibody libraries compared with the human germline antibody sequences of the same CDR length configuration (Figures 3 and 4, respectively). The spatial distributions of the hot spot occurrence probabilities on antibody 3D structures indicate that the designed scFvs have extensive paratopes for protein binding, comparable to those on the human germline antibody variable domains (Figure 5). The scFv variants of the synthetic antibody libraries are distinguishable from the human germline antibody variable domain sequences by the enhanced distributions of CDR hot spot residues for PPIs.

In addition to the antibody-protein interactions discussed above, antibody-peptide interactions also need to be considered in the designs of the scFv variants in the synthetic antibody libraries targeting protein antigens. This is because antibodies could recognize corresponding protein antigens through linear epitopes with specific antibody-peptide interactions. We developed and validated a machine learning predictor (ISMBLab-PEP, see Supplementary Methods) for CDR hot spot residues in antibody-peptide interactions based on the same algorithm and parameterization as in the ISMBLab-PPI predictors. The results equivalent to Figures 1~5 are shown in Figures S10~S14, which were calculated with the ISMBLab-PEP predictors instead of ISMBLab-PPI predictors.

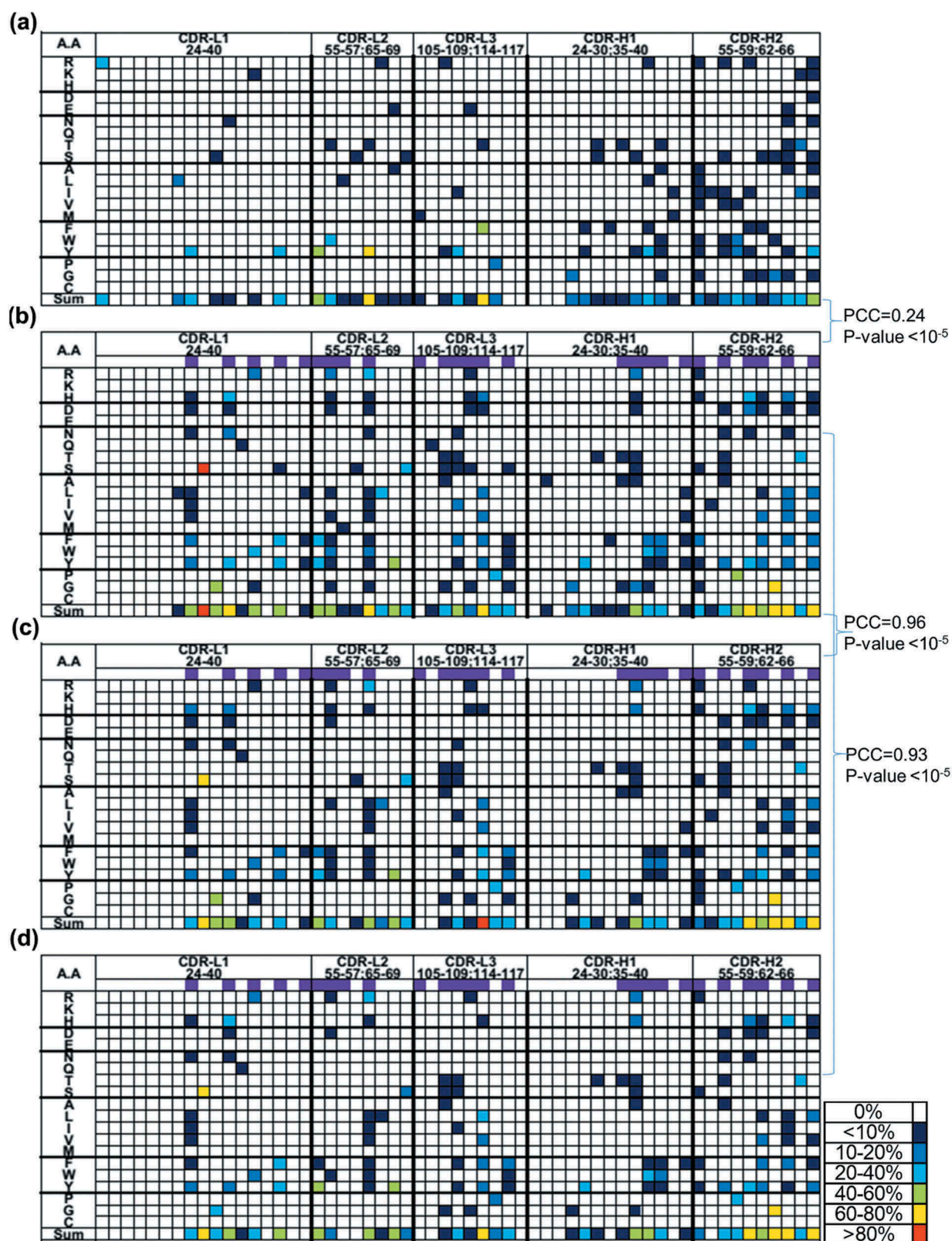
The general conclusions for antibody-protein interactions apply to antibody-peptide interactions as well, suggesting that the scFv variants of the synthetic antibody libraries are anticipated to be applicable to recognize both conformational (antibody-protein interactions) and linear (antibody-peptide interactions) epitopes on protein antigens. The distributions of the hot spot residues for both antibody-protein and antibody-peptide interactions are CDR position-dependent, mostly due to the dependence of the amino acid type distribution and the exposure level of the amino acid sidechain on its CDR position. The antibody-protein interaction hot spot residues are more abundant and are distributed in a more extensive surface area than the antibody-peptide interaction hot spot residues, in agreement with the general experimental observation that the peptide binding sites are smaller than the protein binding sites on antibodies. Overall, the amino acid type distributions of the predicted hot spot residues for both antibody-protein and antibody-peptide interactions are more prominent for the residues with aromatic sidechains, in agreement with the hot spot residues in PPIs.<sup>8,9</sup>



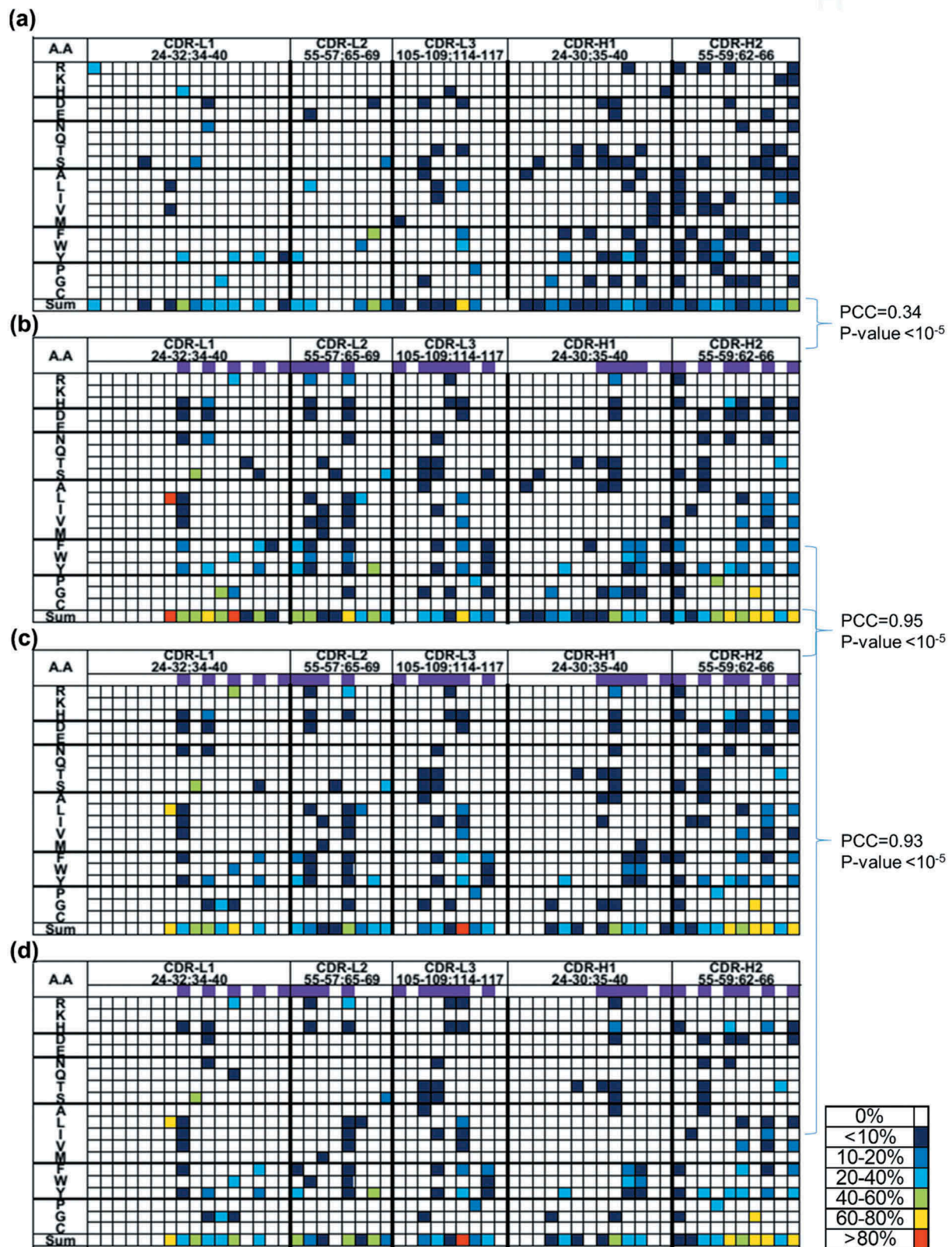
**Figure 1.** Side-by-side comparisons of numbers of hot spot residues for antibody-protein interactions on the CDRs of antibodies from the synthetic antibody libraries and antibodies in nature with the corresponding CDR length configurations. (a) The three scFv sets: set(D), set(F) and set(FB), are described in the main text. The hot spot residues were predicted for each of the modeled scFv structures with ISMBLab-PPI and PCL  $\geq 0.45$ . The distributions of the numbers of the hot spot residues in the CDR L1~H2 per scFv for the scFv sets, as indicated in the x-axis, are shown as the blue, orange and yellow box plots, respectively; the line in each of the box plots represents the median number of hot spots per scFv (the bottom and top of the box are the first and the third quartile) and the lower and upper bars show the minimum and maximum number of the distribution, respectively. These distributions are compared with the green box plot, which shows the distribution of numbers of hot spots on the CDR L1~H2 per scFv calculated with ISMBLab-PPI and confidence level  $\geq 0.45$  for the VL-VH variable domain structures in the Ab-PRO dataset with 13~10~11~8~9 CDR length configuration. The P-values between the scFv dataset and Ab-PRO dataset are shown above the box plots while the P-values between the scFv datasets are shown below the box plots. The P-values greater than 0.05 are not shown. (b) Distributions of hot spot residues on CDR-H3s are shown as the box plots. The descriptions are the same as in panel A. (c) The distributions of the numbers of the hot spot residues in the CDR L1~H2 per scFv for the scFv sets, as indicated in the x-axis, are shown as the blue, orange and yellow box plots. The descriptions are the same as in panel A. (d) The descriptions are the same as in panel B for the data sets shown in the x-axis. (e) The distributions of the numbers of the hot spot residues in the CDR L1~H2 per scFv for the scFv sets, as indicated in the x-axis, are shown as the blue, orange and yellow box plots. The descriptions are the same as in panel A. (f) The descriptions are the same as in panel B for the data sets shown in the x-axis.



**Figure 2.** Predicted antibody-protein interaction hot spot residue frequencies and amino acid types in each CDR-L1~H2 position of the 13–10–11–8–9 CDR length configuration for antibodies from the synthetic antibody libraries and human germline antibody variable domain sequences with the same CDR length configuration. (a) The residue positions, occurrence probabilities, and amino acid types of predicted hot spot residues in five CDRs (CDR-L1~H2) for human germline antibody variable domain sequences with 13–10–11–8–9 CDR length configuration calculated with ISMBlab-PPI (threshold 0.45) are shown as the heat map. Each of the 32 human antibody light chain germline V genes with 11–8–9 CDR length configuration were paired with 35 human antibody heavy chain germline V genes with 13–10 CDR length configuration (listed in order in Table S7; gene sequences obtained from the IMGT database) to construct antibody germline VL-VH variable domain structures with invariable light chain J (IGKJ2\*02) and heavy chain D-J (IGHD3-3\*02-IGHJ4\*01) germline gene sequences. 10 best VL-VH variable domain computational model structures for each pair of the germline V genes were derived by running Rosetta Antibody with default settings. These  $32 \times 35 \times 10$  model structures were then used as inputs for hot spot residue predictions with ISMBlab-PPI. A hot spot residue is defined as the residue with at least one atom having predicted confidence level  $\geq 0.45$ . The hot spot occurrence probability for each amino acid type in each position was determined by the number of predicted hot spot residues for the amino acid type over the total number of antibody cases in this position. The top row shows the IMGT positions for start and end residues of each CDR. The bottom color bar shows the color scheme of the probability heat maps. (b) Same as in panel A for the scFvs from set(D). The purple color in the second row highlights the CDR positions with variable amino acid type in the synthetic antibody libraries. (c) Same as in panel B for the scFvs from set(F). (d) Same as in panel B for the scFvs from set(FB).

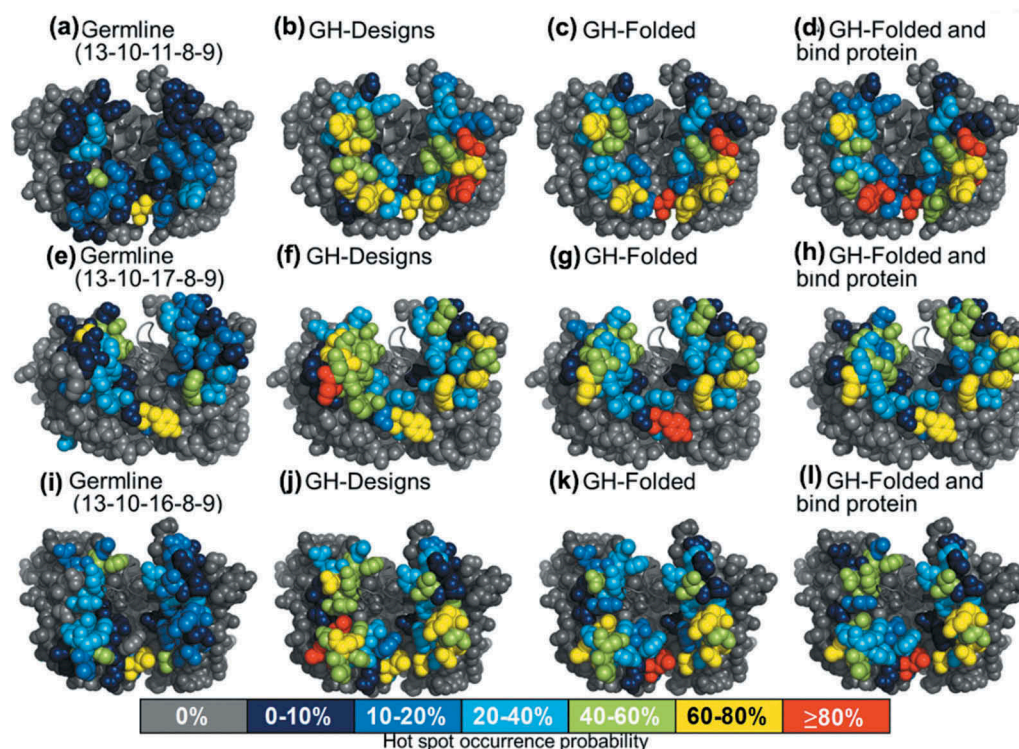


**Figure 3.** Predicted antibody-protein interaction hot spot residue frequencies and amino acid types in each CDR-L1~ H2 position of the 13-10-17-8-9 CDR length configuration for antibodies from the synthetic antibody libraries and human germline antibody variable domain sequences with the same CDR length configuration. (a) The residue positions, occurrence probabilities, and amino acid types of predicted hotspot residues in five CDRs (CDR-L1~ H2) for human germline antibody variable domain sequences with 13-10-17-8-9 CDR length configuration calculated with ISMLab-PPI (threshold 0.45) are shown as the heat map. Each of the 3 human antibody light chain germline V genes with 17-8-9 CDR length configuration were paired with 35 human antibody heavy chain germline V genes with 13-10 CDR length configuration (listed in order in Table S7; gene sequences obtained from the IMGT database) to construct antibody germline VL-VH variable domain structures with invariable light chain J (IGKJ2\*02) and heavy chain D-J (IGHD3-3\*02-IGHJ4\*01) germline gene sequences. 10 best VL-VH variable domain computational model structures for each pair of the germline V genes were derived by running Rosetta Antibody with default settings. These 3 × 35 × 10 model structures were used as inputs for hot spot residue predictions with ISMLab-PPI. The calculations and general descriptions of the heat maps are the same as in Figure 2. (b) Same as in panel A for the scFvs from set(D). The purple color in the second row highlights the CDR positions with variable amino acid type in the synthetic antibody libraries. (c) Same as in panel B for the scFvs from set(F). (d) Same as in panel B for the scFvs from set(FB).



**Figure 4.** Predicted antibody-protein interaction hot spot residue frequencies and amino acid types in each CDR-L1~H2 position of the 13-10-16-8-9 CDR length configuration for antibodies from the synthetic antibody libraries and human germline antibody variable domain sequences with the same CDR length configuration. (a) The residue positions, occurrence probabilities, and amino acid types of predicted hot spot residues in five CDRs (CDR-L1~H2) for human germline antibody variable domain sequences with 13-10-16-8-9 CDR length configuration calculated with ISMBLab-PPI (threshold 0.45) are shown as the heat map. Each of the 8 human antibody light chain germline V genes with 16-8-9 CDR length configuration were paired with 35 human antibody heavy chain germline V genes with 13-10 CDR length configuration (listed in order in Table S7; gene sequences obtained from the IMGT database) to construct antibody germline VL-VH variable domain structures with invariable light chain J (IGKJ2\*02) and heavy chain D-J (IGHD3-3\*02-IGHJ4\*01) germline gene sequences. 10 best VL-VH variable domain computational model structures for each pair of the germline V genes were derived by running Rosetta Antibody with default settings. These  $8 \times 35 \times 10$  model structures were used as inputs for hot spot residue predictions with ISMBLab-PPI. The calculations and general descriptions of the heat maps are the same as in Figure 2. (b) Same as in panel A for the scFvs from set(D). The purple color in the second row highlights the CDR positions with variable amino acid type in the synthetic antibody libraries. (c) Same as in panel B for the scFvs from set(F). (d) Same as in panel B for the scFvs from set(FB).





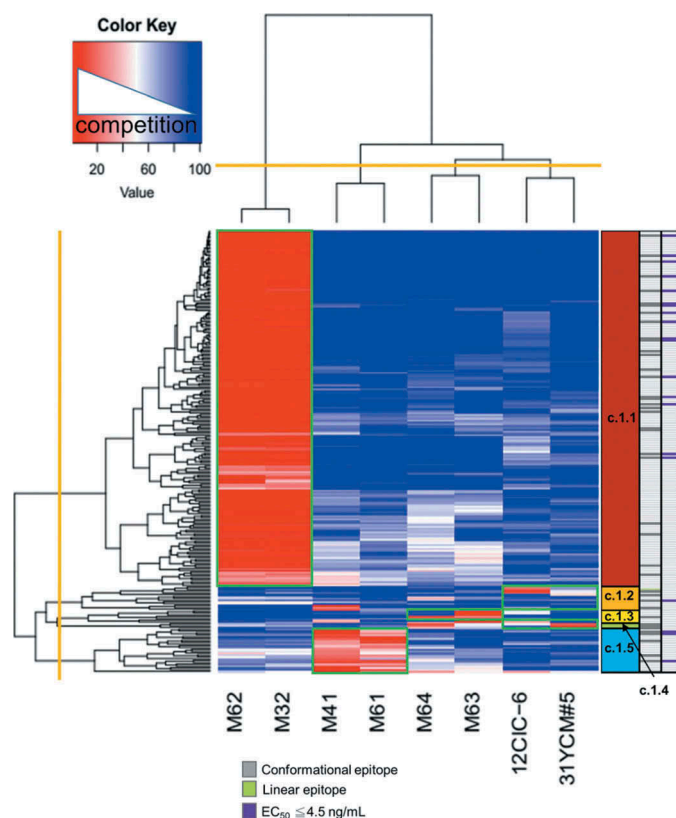
**Figure 5.** Antibody-protein interaction hot spot occurrence probabilities mapped on the surfaces of antibody variable domain model structures. (a) The hot spot occurrence probabilities in CDR-L1~H2 predicted with ISMBLab-PPI and confidence level  $\geq 0.45$  for the human germline antibody variable domain sequences of CDR length configuration 13-10-11-8-9 are mapped onto the model structure, where the sidechains of the CDR-H3 are not shown. The occurrence probabilities are reproduced from the heat map row labelled 'Sum' in Figure 2(a). The bottom color bar shows the interval of hot spot occurrence probability. (b)~(d) The hot spot occurrence probabilities in CDR-L1~H2 for the modeled scFv structures in set(D) of GH2-13 and GH2-6~20 (data from Figure 2(b)), set(F) of GH2-13 and GH2-6~20 (data from Figure 2(c)), and set(FB) of GH2-13 and GH2-6~20 (data from Figure 2(d)) are shown in panels (b)~(d), respectively. (e) Same as in panel (a) for CDR length configuration 13-10-17-8-9 with the hot spot occurrence probabilities reproduced from the heat map row labelled 'Sum' in Figure 3(a). (f)~(h) The hot spot occurrence probabilities in CDR-L1~H2 for the modeled scFv structures with 13-10-17-8-9 CDR length configuration in set(D) of GH3-6~13 (data from Figure 3(b)), set(F) of GH3-6~13 (data from Figure 3(c)), and set(FB) of GH3-6~13 (data from Figure 3(d)) are shown in panels (f)~(h), respectively. (i) Same as in panel (a) for CDR length configuration 13-10-16-8-9 with the hot spot occurrence probabilities reproduced from the heat map row labelled 'Sum' in Figure 4(a). (j)~(l) The hot spot occurrence probabilities in CDR-L1~H2 for the modeled scFv structures with 13-10-16-8-9 CDR length configuration in set(D) of GH3-6~13 (data from Figure 4(b)), set(F) of GH3-6~13 (data from Figure 4(c)), and set(FB) of GH3-6~13 (data from Figure 4(d)) are shown in panels (j)~(l), respectively.

### **Antibodies from the synthetic antibody libraries bind to HER2-ECD through diverse epitopes with high affinity and specificity without affinity maturation**

To scrutinize the antibody molecules from the synthetic antibody libraries, we focused on the HER2-ECD-binding scFvs from the synthetic antibody libraries to characterize the spectrum of epitopes on HER2-ECD interacting with the scFvs. Fourteen of the 20 synthetic antibody libraries yielded HER2-ECD-binding scFvs (Table S6). The epitopes of a total of 241 randomly selected non-redundant HER2-ECD-binding scFvs were grouped according to the competition binding pattern attained with ELISA to the antigen in the presence of each of the 8 competing IgGs known to bind to HER2-ECD at different epitopes (Figure 6) – 6 of the competing IgGs (M62, M32, M41, M61, M63 and M64) had been characterized in a previous study;<sup>21</sup> 2 additional competing IgGs (12CIC-6 and 31YCM#5) were added to cover additional epitopes in this study, so that each of the 241 HER2-ECD-binding scFvs competed with at least one of the 8 competing IgGs for HER2-ECD binding (Figure 6). The objective thresholds for epitope clustering should optimally separate the competition blocks colored in

blue with minimal number of clusters. In Figure 6, the thresholds (orange cutoff lines shown in the x- and y-axis) defining the clustering of 4 groups of the competing IgGs and 5 groups of the scFvs lead to a minimal number of blocks enclosed in green lines (c.1.1::M62/M32; c.1.2&c.1.4::12CIC-6/31YCM#5; c.1.3::M63/M64; c.1.5::M41/M61), which optimally separate the competition patterns with minimal epitope groups: the scFvs from the synthetic antibody libraries bind to HER2-ECD through at least 5 independent epitope groups.

The sequence lengths of the antibody CDRs could affect the epitope preference of the antibody. All the major epitope clusters were correspondingly distributed by the scFv variants from the three sets of synthetic antibody libraries, except for the c.1.3 epitope group, for which the scFv variants were mostly from the GH3-6~13 synthetic antibody library set (Figure 7(a)); the scFv variants in the c.1.3 epitope group were dominated by the CDR sequence length 13-10-17/16-8-9 with 9-residue CDR-H3 (Figure 7(b)), suggesting that the length of the CDR-L1 could play an important role in determining the epitope location on the antigen. Nevertheless, none of the epitope groups were dominated by a single CDR length configuration (Figure 7(b)). Strikingly, the CDR sequence length distributions in Figure 7(a) indicate that the



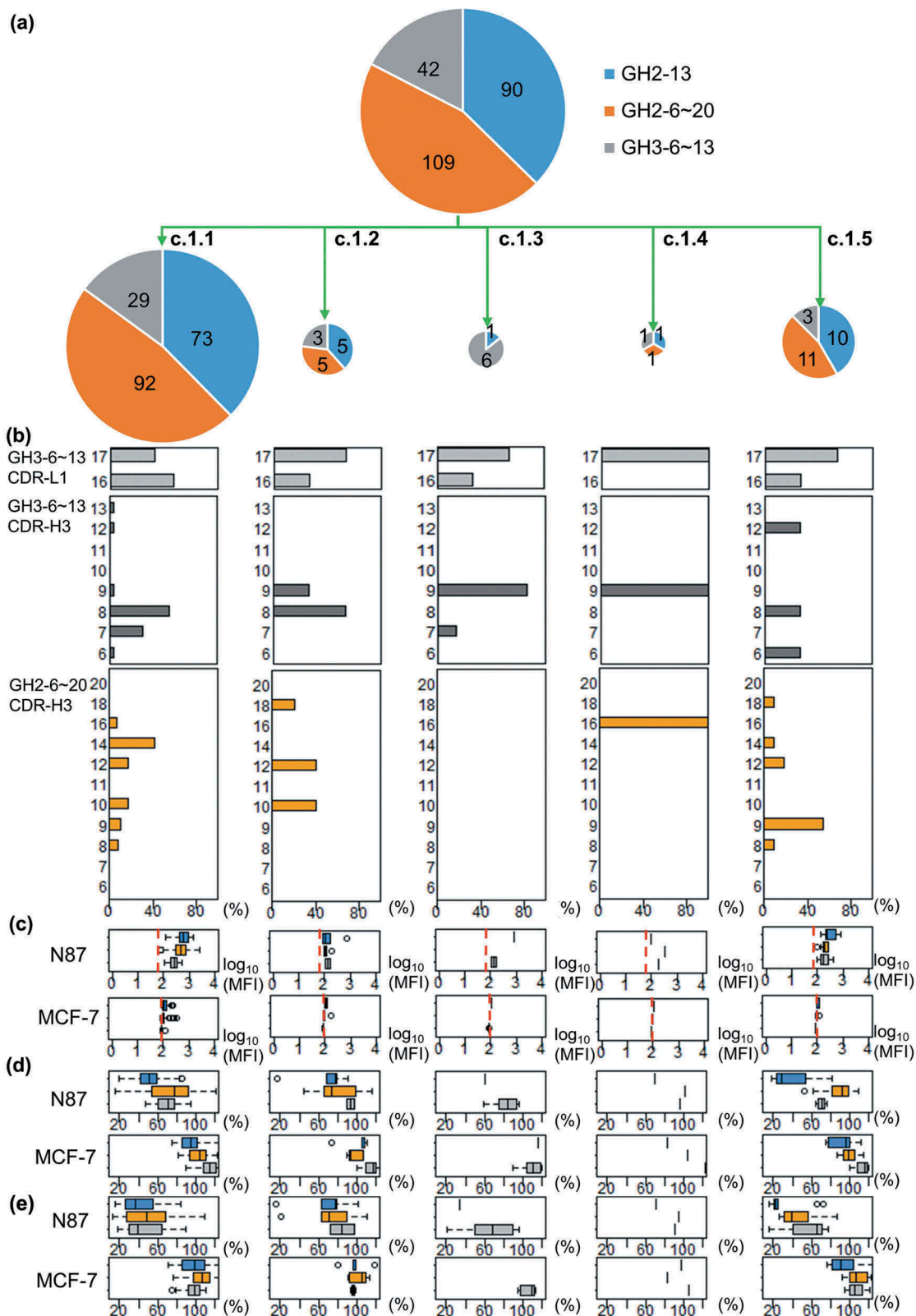
**Figure 6.** Clustering the epitope groups of the HER2-ECD-binding scFvs with competition ELISA patterns against 8 competing IgGs binding to HER2-ECD at different epitopes. The competition ELISA patterns are color-coded from blue to red with decreasing relative signal (the color scheme on the upper-left corner of the panel) resulting from the scFv (y-axis) binding to HER2-ECD immobilized in the ELISA well in the presence of the competing IgG (x-axis). The dendrogram and the heat map were calculated with the competition matrix using the gplots package of R software (Supplementary Method). The color bar next to the heat map shows the minimal epitope groups of the 241 HER-ECD-binding scFvs. Solid lines in the column next to the epitope group column indicate the scFvs reformatted into IgG1s and the purple-colored solid lines in the rightmost column indicate the IgG1s with affinity superior to that of trastuzumab (Table S8). Experimental details are described in Supplementary Method. See main text for discussion.

CDR-H3 length diversity of the GH2-6 ~ 20 synthetic antibody libraries did not enhance the preferences of the antibody-HER2-ECD interactions on the epitopes that had not already been used in the HER2-ECD-binding scFvs from the GH2-13 antibody library with fixed CDR-H3 length. This result is surprising because the structure of the CDR-H3, determined by the length of the loop structure, dictates the shape of the paratope surface and thus the tentative complementary binding location on the antigen; diversity of CDR-H3 length had been anticipated to result in more diverse epitopes. But this anticipation was not aligned with the results shown in Figure 7(a,b).

To test if the scFvs binding to recombinant HER2-ECD also recognize the HER2 receptor expressed on cell surfaces, we evaluated the binding of each of the 241 scFvs to cell surface HER2 with cell-based assays and the cytotoxicity of receptor-mediated endocytosis of the PE38-conjugated scFv binding to cell surface HER2. The scFv variants were cytotoxic to varying extent when non-covalently conjugated with *Pseudomonas* exotoxin AL1-PE38KDEL or AL2-PE38KDEL<sup>34</sup> as shown in Figure S15, the data of which are summarized in Figure 7(d,e), respectively, in box plots. These cytotoxicity data are shown in parallel with the mean fluorescence intensity (MFI) measurements of the RFP (red fluorescence protein)-labeled scFvs binding to HER2-ECD on cell surface

(Figure S16). The MFI data sets are summarized in Figure 7(c) in box plots. The binding affinities are not correlated with receptor-mediated cytotoxicities of the scFv-PE38 non-covalent immunoconjugates (Figure S16). To validate the specificity of the scFvs binding to HER2-ECD, we compared the cytotoxicity/MFI data from N87 cells known to express HER2 on the cell surface with the corresponding data from MCF-7 cells known to express negligible HER2 on the cell surface as negative control in Figure S16 and Figure 7(c,d,e). The negative control cytotoxicity/MFI data sets indicate no cross reaction of the 241 HER2-ECD-positive scFvs to the cells without HER2 expression (Figure S16), indicating specific binding of the scFvs to the cell surface HER2-ECD. These results indicate that the scFvs selected and screened from the synthetic antibody libraries binding to recombinant HER2-ECD immobilized in ELISA wells also specifically bound to HER2 on the cell surface.

Finally, we characterized human IgG1s reformatted from 43 scFvs randomly selected from the 241 HER2-ECD-binding scFvs. The epitope group, CDR sequences, production yield and affinity to HER2-ECD ( $EC_{50}$ ) for each of these IgG1s are summarized in Table S8. The expression and purification of these IgG1s were characterized with SDS-PAGE gels (Figure S17). The scFv variants from the synthetic antibody libraries were anticipated to bind to protein antigens through either



**Figure 7.** Distributions of the CDR length configurations and cell-surface HER2 interactions of the HER2-ECD-binding scFvs in the context of the minimal epitope groups. (a) The pie charts show the distributions of the 241 HER2-ECD-binding scFvs from the three sets of synthetic antibody libraries (GH2-13 in blue, GH2-6 ~ 20 in orange and GH3-6 ~ 13 in grey) in the 5 minimal epitope groups. (b) The length distributions of CDR-L1 and CDR-H3 of the scFvs from the GH3-6 ~ 13 and GH2-6 ~ 20 library sets are shown in the histograms for the scFvs from each epitope group. The x-axis shows the percentage of each CDR length in the corresponding epitope group. (c) Distributions of the mean fluorescence intensities (MFI in  $\log_{10}$  scale in the x-axes) for the scFvs (0.5 nM) from each epitope group complexed with AL1-RFP binding to cell surface HER2 on N87 cells measured by flow cytometry are shown with the box plots. Red dash line shows the negative control background intensity. In parallel, the MFI measurements for the same set of scFvs on the HER2-negative MCF-7 cells are also shown. (d) Distributions of cell viabilities (percentage of survival cells shown in the x-axes) for N87 (upper panels) and MCF-7 (lower panels) cells treated with 0.5 nM scFvs complexed with AL1-PE 38KDEL at the 1:1 molar ratio are shown with the box plots. (e) Distributions of cell viabilities (percentage of survival cells shown in the x-axes) for N87 (upper panel) and MCF-7 (lower panel) cells treated with 0.5 nM scFvs complexed with AL2-PE38KDEL at the 2:1 molar ratio. The line in each of the box plots represents the median value (the bottom and top of the box are the first and the third quartile) and the lower and upper bars show the minimum and maximum number of the distribution, respectively. The color scheme for panel (a) is used for panel (c), (d) and (e). Experimental details are shown in Supplementary Method.

conformational or linear epitopes (see above), but the epitopes discovered were nevertheless dominated by conformational ones: 42 of the 43 IgG1s bound to HER2-ECD through conformational epitopes (Figure S18); only one IgG1 (GH12CIC-6, see Table S8 and Figure S18) bound to HER2-ECD through a linear epitope. Although the CDR length configurations 13–10–16/17–8–9 were rarely found in antibody-protein complexes in PDB (Figure S1B), the antibodies from the synthetic antibody library set GH3-6 ~ 13 nevertheless bound to the protein antigen HER2-ECD with high affinity through diverse conformational epitopes (Table S8). The affinities of the IgG1s were compared with that of trastuzumab in  $EC_{50}$  measurements, with the binding versus IgG1 concentration curves presented in Figure S19: 19 (curves plotted in black in Figure S19) of the 43 IgG1s had higher affinity than that of trastuzumab in terms  $EC_{50}$  measurements on the basis of the statistics of the binding versus IgG1 concentration curves (Figure S19). The high affinity antibodies bound to the antigen through diverse epitopes (Figure 6 and Table S8). These results indicated that selected scFvs from the synthetic antibody libraries herein can be reformatted and expressed as IgG1 with affinity frequently superior to that of the affinity-matured antibodies without explicit affinity maturation processes.

## Discussion

The functionalities of the synthetic scFv libraries are attributed to the enhanced distributions of the CDR hot spot residues in the scFv variants. *In vivo* affinity maturation by stochastic SHM does not occur with specific preference to antibody residues in contact with the antigens,<sup>35,36</sup> and has been known to enhance the antibody-antigen binding through diverse mechanisms, such as by increasing nonpolar complementarity of the antigen-contact areas,<sup>37</sup> introducing subtle structural variations between the variable domain interface,<sup>38</sup> optimizing electrostatic interactions,<sup>38,39</sup> and reducing antigen-binding site flexibility.<sup>40–42</sup> Alternatively, we found that antibodies with enhanced protein-specific hot spot residue distributions on the CDRs could confer the antibodies with high affinity and specificity to protein antigens without affinity maturation. The rationales are that more hot spot residues in the antibody-protein antigen binding surfaces could enhance more energetically favorable antibody-antigen interactions and that combinations of more densely distributed hot spot residues could result in more diverse and effective antibody-protein recognition surfaces on the antibody CDRs, increasing the probability of forming antigen binding sites with high specificity due to better complementarity to cognate protein antigen's surfaces.<sup>24</sup> Experiments herein indicate that the antibody variable domains have substantial structural tolerance for enhanced hot spot density by multiple folds in comparison with those in the human germline antibody variable domain sequences; it seems that germline antibodies are under-equipped with hot spot residues in exchange for keeping options open for polyreactive recognition of diverse antigens. Many functional antibodies, for which the specificity and affinity are comparable or superior to the corresponding control antibody optimized through *in vivo* affinity

maturation, can be readily discovered from these artificially designed synthetic scFv libraries with enhanced hot spot residue distributions in the CDRs, suggesting an alternative to the *in vivo* affinity maturation procedure required for progeny antibodies from germline antibody lineages, which do not necessarily satisfy all the characteristics needed in antibody-based therapeutics.

Phage-displayed synthetic antibody libraries have been developed into technological platforms and successfully applied to antibody discoveries. The human Combinatorial Antibody Libraries (HuCAL) technology of MorphoSys is built on 49 variable domain frameworks, each of which consists of one VH domain from 7 representative human VH consensus sequences and one VL domain from 7 representative human VL consensus sequences.<sup>43,44</sup> Only the CDR3L and CDR3H sequences were diversified according to the knowledge-based designs derived from public domain antibody sequence and structural databases. The first HuCAL scFv library designs had a theoretical sequence space on the order of  $10^{18}$  ( $49 \times 10^7 \times 10^9$ ).<sup>44</sup> Only around  $10^{10}$  HuCAL variants have been expressed as phage-displayed libraries, mainly due to the upper limit in the bacterial host transformation of the phage display technology settings.<sup>44</sup> The subsequent HuCAL GOLD technology incorporated knowledge-based designs to diversify all six CDRs in the antibody fragment variable domain,<sup>43</sup> and the HuCAL PLATINUM technology improved the 49 framework sequences based on the human antibody germline sequences.<sup>45</sup> The magnitudes of the library complexity remain on the order of  $10^{10}$  for all libraries of the HuCAL series.<sup>45</sup>

Based on the notion that antibody paratopes are enriched with aromatic residues, noticeable upon solving the first structures of antibody-antigen complexes<sup>11</sup> and surveys of antibody structures and sequences thereafter,<sup>12–14</sup> the fundamental role of the tyrosyl side chains in antibody-antigen recognition has been demonstrated<sup>46–49</sup> with the functional antibodies selected and screened from the minimalist designs of antibody CDR libraries with only a small subset of amino acid types (Tyr, Ala, Asp, Ser)<sup>49</sup> or with binary code (Tyr and Ser).<sup>48</sup> Synthetic antibody libraries based on the observed distributions of amino acid type in antibodies have also been demonstrated to be productive tools in antibody discoveries.<sup>50</sup>

The antibody libraries developed in this study are different from the previously developed synthetic antibody libraries in that our antibody libraries have been constructed with antibody-antigen (proteins and peptides) interaction propensity predictions as guidance. We used the computational methodology to assess antibody library designs to ensure that the antibody variants in the synthetic antibody libraries have substantial antibody-antigen interaction propensities, rather than just to imitate antibody sequences in nature or limit to the minimalist designs for antibody library constructions. The experimental tests of the functionalities of the antibody libraries indicated that these antibody library designs can be constructed and expressed with the phage display system, and the selected antibodies are functional both in stability, affinity and specificity to the target antigen comparable

to the antibodies derived from natural immune systems. These results suggest that antibody libraries could be designed with specificity guidance by the computational machine learning algorithms that are programmed to predict interaction propensities to molecules of diverse chemical properties, such as carbohydrates and haptens. As such, a large diversity of antibody libraries, which are not limited by the sequences found in natural antibodies, could be envisaged of being developed specifically targeting antigens of diverse chemical properties. In the future, these antibody library designs could be combined with computational methods for human T-cell epitope predictions<sup>51</sup> to avoid CDR sequences of potential immunogenicity so as to enhance the applicability of the synthetic antibody libraries for developing antibody therapeutics. These directions could lead to antibodies with optimal characteristics pertinent to their medical applications.

## Materials and methods

### Computational methods

#### ISMBLab-PPI and ISMBLab-PEP predictors

The general principles and technical details of the ISMBLab-PPI/PEP predictors have been published previously.<sup>18,24-29</sup> Details of the computational methods are described in Supplementary Methods.

### Experimental methods

#### Phage-displayed synthetic ScFv library construction and quality control by NGS

The construction and characterization of the phage-displayed synthetic scFv libraries followed the same procedure, without modification, as described in previously published studies.<sup>19,21</sup> Detailed methods are described in Supplementary Methods. The CDR sequence designs are shown in Tables S1~ S5.

#### Experimental procedures associated with the molecular and cellular assays

All the experimental procedures and reagents are described in detail in the Supplementary Methods.

## Abbreviations

Ab-CARB	antibody-carbohydrate complex structures
Ab-LIG	antibody-hapten complex structures
Ab-PEP	antibody-peptide complex structures
Ab-PRO	antibody-protein complex structures
BCR	B cell receptor
CDR	complementarity determining regions
HER2-ECD	human epidermal growth factor receptor 2 – extracellular domain
ISMBLab	In Silico Molecular Biology Laboratory
PCL	prediction confidence level
PDB	protein data bank
PPI	protein-protein interaction
scFv	single-chain variable fragment
SHM	somatic hyper mutation

## Acknowledgments

This research was supported by Academia Sinica and Ministry of Science and Technology [MOST106-0210-01-15-02 and MOST107-0210-01-19-01][AS-SUMMIT-108] and by the Taiwan Protein Project [MOST105-0210-01-12-01 and MOST106-0210-01-15-04][AS-KPQ-105-TPP]. We would also like to thank the support from Program for Translational Innovation of Biopharmaceutical Development [AS-KPQ-106-TSPA].

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Funding

This work was supported by the Academia Sinica [MOST105-0210-01-12-01]; Academia Sinica [MOST106-0210-01-15-04]; Ministry of Science and Technology, Taiwan [MOST105-0210-01-13-01]; Ministry of Science and Technology, Taiwan [MOST104-0210-01-09-02]; Ministry of Science and Technology, Taiwan [MOST106-0210-01-15-02].

## ORCID

Jih-Wei Jian  <http://orcid.org/0000-0001-9870-0678>  
 Hung-Pin Peng  <http://orcid.org/0000-0002-9398-8120>  
 Chao-Ping Tung  <http://orcid.org/0000-0002-8699-3167>  
 Wei-Ying Kuo  <http://orcid.org/0000-0001-8006-3594>  
 An-Suei Yang  <http://orcid.org/0000-0002-4699-873X>

## References

- Rajewsky K. Clonal selection and learning in the antibody system. *Nature*. 1996;381:751–758. doi:10.1038/381751a0.
- Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science*. 2003;301:1374–1377. doi:10.1126/science.1086907.
- Panda S, Ding JL. Natural antibodies bridge innate and adaptive immunity. *J Immunol*. 2015;194:13–20. doi:10.4049/jimmunol.1400844.
- Gunti S, Notkins AL. Polyreactive antibodies: function and quantification. *J Infect Dis*. 2015;212(Suppl 1):S42–6. doi:10.1093/infdis/jiu512.
- Van Regenmortel MH. Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition. *J Mol Recognit*. 2014;27:627–639. doi:10.1002/jmr.2394.
- Dimitrov JD, Planchais C, Roumenina LT, Vassilev TL, Kaveri SV, Lacroix-Desmazes S. Antibody polyreactivity in health and disease: statu variabilis. *J Immunol*. 2013;191:993–999. doi:10.4049/jimmunol.1300880.
- Burkovitz A, Sela-Culang I, Ofra Y. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J*. 2014;281:306–319. doi:10.1111/febs.12597.
- Moreira IS, Fernandes PA, Ramos MJ. Hot spots—a review of the protein-protein interface determinant amino-acid residues. *Proteins*. 2007;68:803–812. doi:10.1002/prot.21396.
- Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. *J Mol Biol*. 1998;280:1–9. doi:10.1006/jmbi.1998.1843.
- Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. *Science*. 1995;267:383–386.
- Davies DR, Padlan EA, Sheriff S. Antibody-antigen complexes. *Annu Rev Biochem*. 1990;59:439–473. doi:10.1146/annurev.bi.59.070190.002255.
- Mian IS, Bradwell AR, Olson AJ. Structure, function and properties of antibody binding sites. *J Mol Biol*. 1991;217:133–151.

13. Kringleum JV, Nielsen M, Padkjaer SB, Lund O. Structural analysis of B-cell epitopes in antibody: protein complexes. *Mol Immunol.* 2013;53:24–34. doi:10.1016/j.molimm.2012.06.001.
14. Ramaraj T, Angel T, Dratz EA, Jesaitis AJ, Mumey B. Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim Biophys Acta.* 2012;1824:520–532. doi:10.1016/j.bbapap.2011.12.007.
15. Dall'Acqua W, Goldman ER, Eisenstein E, Mariuzza RA. A mutational analysis of the binding of two different proteins to the same antibody. *Biochemistry.* 1996;35:9667–9676. doi:10.1021/bi960819i.
16. Sundberg EJ, Mariuzza RA. Molecular recognition in antibody-antigen complexes. *Adv Protein Chem.* 2002;61:119–160.
17. Bostrom J, Yu SF, Kan D, Appleton BA, Lee CV, Billeci K, Man W, Peale F, Ross S, Wiesmann C, et al. Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. *Science.* 2009;323:1610–1614. doi:10.1126/science.1165480.
18. Peng HP, Lee KH, Jian JW, Yang AS. Origins of specificity and affinity in antibody-protein interactions. *Proc Natl Acad Sci USA.* 2014;111:E2656–65. doi:10.1073/pnas.1401131111.
19. Chen I-C, Chiu Y-K, Yu C-M, Lee C-C, Tung C-P, Tsou Y-L, Huang Y-J, Lin C-L, Chen H-S, Wang AH-J, et al. High throughput discovery of influenza virus neutralizing antibodies from phage-displayed synthetic antibody libraries. *Sci Rep.* 2017;7:14455. doi:10.1038/s41598-017-14823-w.
20. Tung CP, Chen IC, Yu CM, Peng HP, Jian JW, Ma SH, Lee Y-C, Jan J-T, Yang A-S. Discovering neutralizing antibodies targeting the stem epitope of H1N1 influenza hemagglutinin with synthetic phage-displayed antibody libraries. *Sci Rep.* 2015;5:15053. doi:10.1038/srep15053.
21. Chen HS, Hou SC, Jian JW, Goh KS, Shen ST, Lee YC, You J-J, Peng H-P, Kuo W-C, Chen S-T, et al. Predominant structural configuration of natural antibody repertoires enables potent antibody responses against protein antigens. *Sci Rep.* 2015;5:12411. doi:10.1038/srep12411.
22. Hsu HJ, Lee KH, Jian JW, Chang HJ, Yu CM, Lee YC, Chen I-C, Peng H-P, Wu CY, Huang Y-F, et al. Antibody variable domain interface and framework sequence requirements for stability and function by high-throughput experiments. *Structure.* 2014;22:22–34. doi:10.1016/j.str.2013.10.006.
23. Chang HJ, Jian JW, Hsu HJ, Lee YC, Chen HS, You JJ, Hou S-C, Shao C-Y, Chen Y-J, Chiu K-P, et al. Loop-sequence features and stability determinants in antibody variable domains by high-throughput experiments. *Structure.* 2014;22:9–21. doi:10.1016/j.str.2013.10.005.
24. Yu CM, Peng HP, Chen IC, Lee YC, Chen JB, Tsai KC, Chen C-T, Chang J-Y, Yang E-W, Hsu P-C, et al. Rationalization and design of the complementarity determining region sequences in an antibody-antigen recognition interface. *PLoS ONE.* 2012;7:e33340. doi:10.1371/journal.pone.0033340.
25. Jian JW, Elumalai P, Pitti T, Wu CY, Tsai KC, Chang JY, Peng H-P, Yang A-S, Zhang Y. Predicting ligand binding sites on protein surfaces by 3-dimensional probability density distributions of interacting atoms. *PLoS ONE.* 2016;11:e0160315. doi:10.1371/journal.pone.0160315.
26. Mahalingam R, Peng HP, Yang AS. Prediction of fatty acid-binding residues on protein surfaces with three-dimensional probability distributions of interacting atoms. *Biophys Chem.* 2014;192c:10–19. doi:10.1016/j.bpc.2014.05.002.
27. Mahalingam R, Peng HP, Yang AS. Prediction of FMN-binding residues with three-dimensional probability distributions of interacting atoms on protein surfaces. *J Theor Biol.* 2014;343:154–161. doi:10.1016/j.jtbi.2013.10.020.
28. Tsai KC, Jian JW, Yang EW, Hsu PC, Peng HP, Chen CT, Chen J-B, Chang J-Y, Hsu W-L, Yang A-S, et al. Prediction of carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. *PLoS ONE.* 2012;7:e40846. doi:10.1371/journal.pone.0040846.
29. Chen CT, Peng HP, Jian JW, Tsai KC, Chang JY, Yang EW, Chen J-B, Ho S-Y, Hsu W-L, Yang A-S, et al. Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. *PLoS ONE.* 2012;7:e37706. doi:10.1371/journal.pone.0037706.
30. Sliwkowski MX, Mellman I. Antibody therapeutics in cancer. *Science.* 2013;341:1192–1198. doi:10.1126/science.1241145.
31. North B, Lehmann A, Dunbrack RL Jr. A new clustering of antibody CDR loop conformations. *J Mol Biol.* 2011;406:228–256. doi:10.1016/j.jmb.2010.10.030.
32. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol.* 1997;273:927–948. doi:10.1006/jmbi.1997.1354.
33. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR. Conformations of immunoglobulin hypervariable regions. *Nature.* 1989;342:877–883. doi:10.1038/342877a0.
34. Hou SC, Chen HS, Lin HW, Chao WT, Chen YS, Fu CY, Yu C-M, Huang K-F, Wang AH-J, Yang A-S. High throughput cytotoxicity screening of anti-HER2 immunotoxins conjugated with antibody fragments from phage-displayed synthetic antibody libraries. *Sci Rep.* 2016;6:31878. doi:10.1038/srep31878.
35. Raghunathan G, Smart J, Williams J, Almagro JC. Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens. *J Mol Recognit.* 2012;25:103–113. doi:10.1002/jmr.2158.
36. Tomlinson IM, Walter G, Jones PT, Dear PH, Sonnhhammer EL, Winter G. The imprint of somatic hypermutation on the repertoire of human germline V genes. *J Mol Biol.* 1996;256:813–817. doi:10.1006/jmbi.1996.0127.
37. Li Y, Li H, Yang F, Smith-Gill SJ, Mariuzza RA. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nat Struct Biol.* 2003;10:482–488. doi:10.1038/nsb930.
38. Midelfort KS, Hernandez HH, Lippow SM, Tidor B, Drennan CL, Witttrup KD. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J Mol Biol.* 2004;343:685–701. doi:10.1016/j.jmb.2004.08.019.
39. Chong LT, Duan Y, Wang L, Massova I, Kollman PA. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proc Natl Acad Sci USA.* 1999;96:14330–14335.
40. Willis JR, Briney BS, DeLuca SL, Crowe JE Jr., Meiler J. Human germline antibody gene segments encode polyspecific antibodies. *PLoS Comput Biol.* 2013;9:e1003045. doi:10.1371/journal.pcbi.1003045.
41. Ag S, Xu H, Khan AR, O'Donnell T, Khurana S, King LR, Manischewitz J, Golding H, Suphaphiphat P, Carfi A, et al. Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody. *Proc Natl Acad Sci USA.* 2013;110:264–269. doi:10.1073/pnas.1218256109.
42. Zimmermann J, Romesberg FE, Brooks CL 3rd, Thorpe IF. Molecular description of flexibility in an antibody combining site. *J Phys Chem B.* 2010;114:7359–7370. doi:10.1021/jp906421v.
43. Rothe C, Urlinger S, Lohning C, Prassler J, Stark Y, Jager U, Hubner B, Bardroff M, Pradel I, Boss M, et al. The human combinatorial antibody library HuCAL GOLD combines diversification of all six CDRs according to the natural immune system with a novel display method for efficient selection of high-affinity antibodies. *J Mol Biol.* 2008;376:1182–1200. doi:10.1016/j.jmb.2007.12.018.
44. Knappik A, Ge L, Honegger A, Pack P, Fischer M, Wellnhofner G, Hoess A, Wölle J, Plückthun A, Virnekäs B. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol.* 2000;296:57–86. doi:10.1006/jmbi.1999.3444.

45. Ponsel D, Neugebauer J, Ladetzki-Baehs K, Tissot K. High affinity, developability and functional size: the holy grail of combinatorial antibody library generation. *Molecules*. 2011;16:3675–3700. doi:10.3390/molecules16053675.
46. Koide S, Sidhu SS. The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem Biol*. 2009;4:325–334. doi:10.1021/cb800314v.
47. Birtalan S, Zhang Y, Fellouse FA, Shao L, Schaefer G, Sidhu SS. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol*. 2008;377:1518–1528. doi:10.1016/j.jmb.2008.01.093.
48. Fellouse FA, Esaki K, Birtalan S, Raptis D, Cancasci VJ, Koide A, Jhurani P, Vasser M, Wiesmann C, Kossiakoff AA, et al. High-throughput generation of synthetic antibodies from highly functional minimalist phage-displayed libraries. *J Mol Biol*. 2007;373:924–940. doi:10.1016/j.jmb.2007.08.005.
49. Fellouse FA, Wiesmann C, Sidhu SS. Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition. *Proc Natl Acad Sci USA*. 2004;101:12467–12472. doi:10.1073/pnas.0401786101.
50. Sidhu SS, Li B, Chen Y, Fellouse FA, Eigenbrot C, Fuh G. Phage-displayed antibody libraries of synthetic heavy chain complementarity determining regions. *J Mol Biol*. 2004;338:299–310. doi:10.1016/j.jmb.2004.02.050.
51. Gfeller D, Bassani-Sternberg M. Predicting antigen presentation—what could we learn from a million peptides? *Front Immunol*. 2018;9. doi:10.3389/fimmu.2018.01716.