Engineering Science of Biological Systems **Biotechnology and Bioengineering**

DOI 10.1002/bit.26244

Computational *de novo* Design of Antibodies binding to a Peptide with High Affinity[†]

Venkata Giridhar Poosarla[1,2‡], Tong Li[1‡], Boon Chong Goh[3], Klaus Schulten[3], Thomas K. Wood[1,2*], and Costas D. Maranas[1*]

[1]Department of Chemical Engineering and [2]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, 16802, USA. [3]Department of Physics and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA

*To whom correspondence should be addressed: costas@psu.edu, tuw14@psu.edu

[‡]These authors have contributed equally to this work

[†]This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: [10.1002/bit.26243]

Additional Supporting Information may be found in the online version of this article.

Received October 7, 2016; Revision Received December 23, 2016; Accepted January 5, 2017

ABSTRACT

Antibody drugs play a critical role in infectious diseases, cancer, autoimmune diseases, and inflammation. However, experimental methods for the generation of therapeutic antibodies such as using immunized mice or directed evolution remain time consuming and cannot target a specific antigen epitope**.** Here, we describe the application of a computational framework called OptMAVEn combined with molecular dynamics to *de novo* design antibodies. Our reference system is antibody 2D10, a single-chain antibody (scFv) that recognizes the dodecapeptide DVFYPYPYASGS, a peptide mimic of mannose-containing carbohydrates. Five *de novo* designed scFvs sharing less than 75% sequence similarity to all existing natural antibody sequences were generated using OptMAVEn and their binding to the dodecapeptide was experimentally characterized by biolayer interferometry and isothermal titration calorimetry. Among them, three scFvs show binding affinity to the dodecapeptide at the nM level. Critically, these *de novo* designed scFvs exhibit considerably diverse modeled binding modes with the dodecapeptide. The results demonstrate the potential of OptMAVEn for the de novo design of thermally and conformationally stable antibodies with high binding affinity to antigens and encourage the targeting of other antigen targets in the future.

Keywords: OptMAVEn; computational antibody design; *de novo* design; antibody structure prediction; single-chain antibodies; biolayer interferometry

This article is protected by copyright. All rights reserved

INTRODUCTION

Antibodies are protective proteins used by the immune system to recognize and neutralize foreign objects through interactions with a specific part of the target, called an antigen. The high specificity and binding affinity of antibodies with antigens enables them to be used as therapeutic agents for the treatment of different diseases (Nelson et al. 2010). Although these methods are effective, they remain time consuming, cannot target a specific epitope, and are unable to tolerate rapid changes of antigens(Sormanni et al. 2015). Hence, advances in antibody design technology and a deeper understanding of the action of therapeutic antibodies are required for improved therapeutic antibodies(Tiller and Tessier 2015). Computational methods for the *de novo* design of a fully human antibody against any specific antigen provide a route to resolve these issues(Li et al. 2014). This strategy, if validated, could offer a general route to therapeutic antibodies for many pathogens that have resisted traditional vaccine development, including highly antigenically-variable viruses such as HIV, influenza and Ebola virus.

To this end, we have developed a computational framework named Optimal Method for Antibody Variable region Engineering (OptMAVEn)(Li et al. 2014) for the *de novo* design of fully-human antibody variable domains to bind any specified antigen by assembling the six best-scored modular antibody parts (MAPs)(Pantazes and Maranas 2013). In particular, OptMAVEn is implemented as a three-step workflow. First, for a given antigen, an ensemble of possible antigen binding conformations is generated in a modeled antibody-binding site. Second, the top scored antigen conformations and antibody models assembled by the combinations of six modular antibody parts from the MAPs database are selected. Third, random mutations are introduced to the antibody models for improved binding affinity to the antigens. This idea is inspired by the natural evolution of an antibody *in vivo*, in that the gene of a "germline" antibody

variable domain is initially assembled by combining random variable (V), diversity (D), and joining (J) "germline" genes to create an antibody variable domain(Peled et al. 2008). We have created the MAPs database(Pantazes and Maranas 2013) approximately analogous to the naturally occurring human repertoire of V, D and J genes. To improve the description of antibody structures, the MAPs database utilizes CDR3 as a structural component instead of D region (heavy chain) and partial region of V region (light chain). The primary advantage of using MAPs to *de novo* assemble an antibody model is that the computational overhead is manageable as the need for *de novo* folding calculations is bypassed by assembling structural domains. In addition, compared to traditional fragment-assembly-based approaches(Simons et al. 1997) for *de novo* protein structure prediction, OptMAVEn efficiently samples both local and non-local contacts that are inherently present in the relatively large structural fragments. To generate antibodies of high quality, molecular dynamics (MD) simulation is incorporated in the present study into the design workflow of OptMAVEn to screen against unstably bound with antigen antibodies. Although OptMAVEn is capable of generating novel computational antibody models with numerous interactions with their target epitopes, the feasibility of this protocol has not been validated by experiments until now.

Using OptMAVEn, we *de novo* designed single-chain antibodies (scFvs) against the dodecapeptide antigen DVFYPYPYASGS. The dodecapeptide with its Tyr-Pro-Tyr motif mimics the carbohydrate methyl α-*D*-mannopyranoside, which has been studied extensively since it is the recognition target of lectin concanavalin A (con A) and mAb-2D10(Krishnan et al. 2007; Tapryal et al. 2013). Association with the two antigens (dodecapeptide and methyl α-*D*-mannopyranoside) occurs at different sites for con A. Because 2D10-Fab (fragment antigen-binding) could not be crystallized, a shorter recombinant scFv-2D10 was constructed,

refolded(Tapryal et al. 2010), and crystallized with both antigens which revealed that 2D10 has a rigid binding site that does not change upon antigen binding(Tapryal et al. 2013). We chose dodecapeptide antigen (DVFYPYPYASGS) to test the OptMAVEn designs because it is easily synthesized, has a crystal structure bound to the scFv-2D10 (PDB ID 4H0H)(Tapryal et al. 2013), and scFv-2D10 can be produced in *Escherichia coli* and activated by *in vitro* refolding(Tapryal et al. 2010).

We designed five *de novo* scFvs possessing distinct sequences compared to all existing natural antibody sequences that can bind with the dodecapeptide antigen using OptMAVEn. Since the stability and activity of a protein often depend not only on its static structure but also on its dynamic properties(Mou et al. 2015), an additional conformational sampling and stability evaluation using MD simulation was performed. The lack of sequence similarity between our designs and scFv-2D10 demonstrates that the OptMAVEn procedure broadly samples the vast sequence space and arrives at designs unbiased by the original antibody residue composition. All five scFvs were found to be actively folded and stable in solution, and three of the scFvs showed high binding affinity to the dodecapeptide antigen (nanomolar level) while exhibiting diverse binding modes to the antigen. The results demonstrate that OptMAVEn combined with MD simulation are a promising computational approach for the *de novo* design of thermally and conformationally stable antibodies with high binding affinities.

MATERIALS AND METHODS

Computational antibody generation. Figure 1 shows an overview of OptMAVEn to design the entire variable domain of an antibody targeting a specific epitope (Supplementary Text 1). Framework domains of OptMAVEn generated antibodies were aligned with existing antibodies (Supplementary Data 1) to identify allowed mutable residues.

Experimental procedures. A series of experimental steps involving cloning methods (fasta sequences of scFvs in Supplementary Data 2 and oligonucleotide sequences used for amplifying the scFvs in Supplementary Table I), expression, purification and refolding (Supplementary Table II) of scFvs, analysis of refolded scFvs by circular dichroism (CD), scFv-antigen binding affinity measurements by biolayer interferometry and isothermal titration calorimetry were described in detail in Supplementary Text 1.

RESULTS

Computational workflow with the incorporation of MD simulation. "Germline" antibody models with favorable dodecapeptide interactions were generated using OptMAVEn. Table I shows the diversity in the chosen antibody parts to recognize the manifold spatial poses of the antigen. Thereafter, all-atom MD simulations were performed to computationally assess the stability of the antibody binding site with the epitope. Additionally, for the antigens that bind stably with the antibody, the interaction at the antigen-antibody interface was refined during MD simulations. The root mean square deviations (rmsd) of the antigen with respect to the antibody were plotted in Fig 2a and listed in Supplementary Table III. Antigens with low binding affinity are seen to unbind from the antibody within 50 ns. To ensure the antigen-antibody interface is sufficiently refined, every complex was simulated for 100 ns. Eight out of the 31 designs (Fig. 2a-g) were stably bound throughout the 100 ns-long MD simulations (see Supplementary Text 2 and 3 for details).

***In silico* affinity maturation.** The assembled "germline" antibodies were expected to have only low or moderate binding affinity to the antigen. To further enhance antibody affinity we performed *in silico* affinity maturation to the MD-refined "germline" antibodies by randomly selecting mutations that are predicted to improve binding energy. Mutations were not limited to

residues within the CDR loops but allowed to arise along the entire variable fragment sequence. However, mutations in residues in the CDRs were allotted a 3-fold higher preference over those in the frameworks (FRs) to match expected mutational patterns observed in matured antibodies(Pantazes and Maranas 2013). Twenty-seven "germline" antibodies refined by MD were submitted for *in silico* affinity maturation. The complex and IEs were used to evaluate the stability of the complex and the binding strength between the antibody and antigen, respectively. Not surprisingly, we found that the best ten designs with the lowest IE (ranging from -550 to -360 kcal/mol) were from groups MD_SB (four) and MD_RE (six) while none were from group MD_PB (Supplementary Table IV). This is indicative of the difficulty in introducing favorable interactions to designs with antigens partially bound.

Designs from group MD_SB were preferred as the binding location and mode was fully retained after MD simulation(Kiss et al. 2010). We selected four designs from group MD_SB with the lowest complex and interaction energies (Table II) denoted as scFv-1 (120_6290_1), scFv-2 (120_15439_2), scFv-3 (140_10149_1) and scFv-4 (140_15899_5). In addition, we selected design scFv-5 (200_15222_5) that did not have one of the lowest complex and interaction energies but possessed the same binding mode as the selected design 200_15222_2. It is important to note that in the five selected designs the sequence distribution of mutations involved in affinity maturation is very diverse, with no clearly preferred amino acids/positions (Supplementary Fig. 1). The total number of accumulated mutations in mature form of scFv-1-4 is between 34% and 38% (Table II) compared to their "germline" versions. This is somewhat higher than what is seen in natural antibodies (typically 5 to 20% changes in somatic mutations and average ~27 mutations per antibody variable domain(Burkovitz et al. 2014)) but comparable to that of HIV-1-neutralizing antibodies (ranging from 15% to 44%(Corti and Lanzavecchia

2013)) which take several years to accumulate. Our results suggest that the mutation rate from computational affinity maturation is higher than the average mutation rate in nature but comparable to that of somatic mutations evolved in *vivo*. Furthermore, all five designed antibodies share less than 50% sequence similarity with the scFv-2D10 antibody (Fig. 3 and Table III) and less than 60% sequence similarity within the design cohort (Table III). We also compared the sequences of our five designs with non-redundant protein sequences (77,704,910 sequences) using BLAST(Altschul et al. 1990) and the nearest sequence to the designs identified in the protein database is CAA81438.1 (accession number). This sequence has only 74% identity with the heavy chain of scFv-5 (Supplementary Table V), which suggests that the *de novo* designs truly sample new antibody sequences not observed in existing antibody sequence space. The stated five designs (scFv-1 to scFv-5) were chosen for experimental characterization.

Purification of the *de novo*-designed scFvs. We produced six scFvs (five *de novo* designed scFvs and scFv-2D10) in *E. coli* as inclusion bodies and refolded them to validate the binding of *de novo* scFvs to antigen. Each scFv was built with a heavy chain (V_H) and light chain (V_L) of the variable regions, a hexa-histidine tag at the C-terminus, and a glycine-serine flexible linker $(Gly_4Ser_1)_3$ between the V_H and V_L regions (Fig. 4a), and was amplified with the respective primers (Fig. 4b). Note that the His tag was 42 Å away from the antigen-peptide active binding site so it is unlikely to affect antigen binding. All the scFvs formed inclusion bodies after induction with IPTG, and the inclusion bodies were washed in Triton X-100 detergent for removal of membrane debris(Palmer and Wingfield 2004), solubilized in 8M urea, and purified via FPLC using Ni-NTA affinity chromatography. The purified scFvs were refolded by gradually reducing the urea concentration (from 8 M to 0 M, Supplementary Table II). To accomplish removal of urea from 2 M to 0 M, arginine hydrochloride to suppresses protein aggregation and

oxidized glutathione for the formation of disulfide bonds was added to the refolding buffer(Umetsu et al. 2003). Yields for the six scFvs were verified by SDS-PAGE (Fig. 4c).

Secondary structure analysis of *de novo* designed scFvs. To confirm correct folding of scFvs, the secondary structures of the six purified scFvs were monitored by far-UV CD spectroscopy and compared to that of other scFvs(Blanco-Toribio et al. 2014; Glaven et al. 2012; Song et al. 2014). The far-UV CD spectra (190-240 nm) for the six scFvs (Fig. 4d) showed the scFvs were predominantly β-sheets, as expected for typical members of the immunoglobulin family, including scFvs (Supplementary Table VI)(Blanco-Toribio et al. 2014; Glaven et al. 2012; Gregoire et al. 1996; Lee et al. 2013; Song et al. 2014). The scFv-2D10 (PDB ID- 4H0H) crystal structure showed the presence of 4% helical and 47% beta sheet secondary structures(Tapryal et al. 2013), whereas the CD data for the scFv-2D10 in our hands retrieved 3.4%, 5%, and 9% helical and 40%, 38%, and 27% beta sheet structures from CDSSTR, CONTIN, and GOR4, respectively (**Supplementary Table 6**). With the recognition that CD spectral data is suitable primarily for the determination of gross secondary structure changes, the crystal structure information (percentage of helical and beta sheet content) for scFv-2D10 (PDB ID- 4H0H) is clearly similar to that obtained from CD data both in solution (CDSSTR and CONTIN) (**Supplementary Table 6**) as well as predicted (GOR4) for the *de novo* designed scFvs (scFv-1 to scFv-5). Therefore, the CD spectral data indicate that all of our scFvs were folded properly and that *de novo* designs did not cause significant structural distortions.

Binding of the scFvs to the dodecapeptide using biolayer interferometry. As a quantitative assessment of the scFv-antigen complex, the association constant (k_a), dissociation constant (k_d), and equilibrium dissociation constant (K_D) values were determined(Fischer et al. 2015; Lee et al. 2014; Prischi et al. 2010; Tang et al. 2014) by loading the refolded His_6 tag scFvs onto Ni-NTA

biosensors and titrating with the dodecapeptide antigen using the biolayer interferometry via Octet QK system. The scFvs (550 nM) showed binding to dodecapeptide antigen (1000 nM) (Fig. 5a). Because of the high concentrations of the antigenic peptide (1000 nM), a dissociation curve (k_d) was not observed for scFv-1, 2 and 4. From the initial sensogram (Fig. 5a), it was clear the antigen as well as antibody concentrations were non-optimal for obtaining the proper kinetic rates for each scFv tested against the antigen; hence, we optimized the antigen concentrations using a fixed scFv-2 concentration (550 nM) and varying antigen concentrations (1000, 300, 100, 30, 10, and 3 nM). Only the 100 nM and 30 nM concentrations of antigen displayed the K_D values of 2.55 nM and 21.5 nM, respectively. The antigen concentration of 30 nM was selected to preclude the after effects of addition of high antigen concentrations when deriving the kinetic values for each scFv in this study (Fig. 5b). For optimizing the antibody titers, different concentrations of scFv-2 (1110, 550, 460, 370, 280 and 180 nM) were tested with a constant antigen concentration of 30 nM. The K_D values were very close (~ 17.5 nM) for the variable scFv concentrations tested, which clearly indicates that changes in antigen concentration impact the kinetic data whereas varying scFv concentrations have little effect (Fig. 5c). Since we controlled the antigen concentrations (the dodecapeptide antigen was synthesized and loaded the same for each scFv), the system is robust for determining the kinetic values for the different scFvs. To rule out the non-specific binding with the biolayer interferometry, we also tested our system with various negative controls: (i) all the scFvs of our present system were titrated against a non-related peptide antigen, and (ii) a non-related scFv which was recombinantly cloned, expressed, purified, and refolded similar to the current system was titrated with the dodecapeptide antigen. Neither of the two sets of negative controls showed binding which confirms the *de novo* designs were specific towards their dodecapeptide targets (Fig. 5d-e). As a

positive control, we obtained a binding constant consistent with an independent lab (0.3 nM) for an unrelated DNA-binding protein. Using the optimized scFv (550 nM) and dodecapeptide antigen (30 nM) concentrations, the binding constant K_D was determined for each *de novo* designed scFv. Significant binding was observed between *de novo* designed scFv-1, 2, and 4 with the antigen along with the scFv-2D10 (Fig. 5f). The scFv-3 and scFv-5 did not exhibit any binding. Based on the K_D values for all the six scFvs tested, the best proved to be scFv-2D10 (3.7 nM) followed by scFv-1 (8.9 nM), scFv-2 (14.4 nM) and scFv-4 (23.8 nM) (Table IV).

Isothermal titration calorimetry (ITC) was performed to corroborate the biolayer interferometry binding results using the best binders from the Octet binding study, scFv-1 and scFv-2D10, which showed a positive response with biolayer interferometry. The scFv-1 along with scFv-2D10 displayed the ~1:1 binding and showed sigmoidal behavior, an indicative of exothermic binding (Supplementary Fig. 2). Because of the high affinity of association (K_D) (~ 1 nM), different presentation of the antigen (not conjugated to BSA), and change in buffer conditions from biolayer interferometry, it is difficult to use ITC to measure the equilibrium binding constant with precision; however, the enthalpy (ΔH), the entropy (ΔS) and stoichiometry (N) of the two scFvs that showed a positive shift with biolayer interferometry are indicative of binding to dodecapeptide antigen (Supplementary Table VII)(Kubala et al. 2010).

Diversity of antibody binding sites in successfully designed scFvs. As all three designs scFv-1, scFv-2 and scFv-4 exhibit low nanomolar to the dodecapeptide in the binding affinity range, their respective residue compositions were assessed to identify structural determinants favorable for binding. Figure 6a and 6b show the comparison of the amino acid compositions between the "germline" and mature antibodies in scFv-1, 2 and 4. Interestingly, Asp, Glu, Gly, Lys, and Arg occur more than any other amino acid (counts ≥ 4) in mature sequences, whereas Ala, Asn, Ile,

Leu, Thr and Tyr occur less frequently (counts ≤ 4). One of the most striking results is the dramatic decrease of aromatic residues especially tyrosine, which is consistent with the trend of *in vivo* somatic hypermutations of natural antibodies(Clark et al. 2006). Tyrosine, which can provide hydrogen bonding and substantial hydrophobic interaction, is frequently found in the paratope of "germline" antibodies to promote low-affinity binding to new antigens(Dalkas et al. 2014). Another apparent trend is that charged residues are favorable in high-affinity mature sequences. Charged-residue driven interaction could provide more specific binding than aromatic binding. Meanwhile, the numbers of aliphatic and polar residues are significantly decreased. More polar and charged residue occurrences contribute to the improvement of binding affinity and complex stability in the solvent. A comparison of mutations before and after affinity maturation demonstrates that the net overall effect is a migration from residue types that could provide nonspecific binding to new ones that generally provide specific binding.

The sequence diversity of the three scFvs alludes to the presence of diverse modes of antigen recognition. A detailed view of the modeled interactions between the three scFvs and dodecapeptide (Fig. 6 c-f) reveals the differences with the scFv-2D10 antibody. In the natural antibody-peptide complex, the phenyl rings of the CDR H2 residue Tyr55 and CDRH3 residue Tyr102 of the antibody form π–π stacking interactions with the phenyl rings of the antigen residues Tyr4 and Tyr6, respectively. In contrast, cation/amino–π interactions are found to be significantly more frequent in the designed scFvs. For example, the amino group of Lys113 in the H chain of scFv-1 forms cation-π interaction with the phenyl ring of the antigen residues Tyr6 and the amino group of Lys66 in the H chain of scFv-4 forms cation-π interactions with the phenyl ring of the antigen residues Tyr6. This finding is in agreement with previous studies demonstrating that cation–π interactions are an important stabilizing factor that is more

frequently found in antibody-antigen than in protein-protein interfaces(Dalkas et al. 2014). The analysis of the hydrogen bond pairs revealed a preference of the hydrogen bonds formed between charged residues in the designs with the antigen. For example, Glu59 in the H chain of scFv-1 forms a hydrogen bond with the hydroxyl group of the antigen Tyr8; Asp34 and Glu108 in the L chain of scFv-2 forms a hydrogen bond with the hydroxyl groups of the antigen Tyr6 and Tyr8, respectively; the CDRH3 residue Asp109 of scFv-4 forms a hydrogen bond with the main chain nitrogen of Phe3. Notably, mutations to charged residues during affinity maturation also contribute considerably to the heavy chain variable domain–light chain variable domain (V_H–V_L) association, which might stabilize the two domains and maintain the relative positions of the CDRs loops, which, in turn, can affect the antigen specificity(Chailyan et al. 2011). For example, the CDRH3 residue Lys114 (mutated from Ala) of scFv-1 form interaction with Tyr 55 of the L chain; the CDRL2 residue Arg56 (mutated from Glu) of scFv-2 interacts with Glu113 (mutated form Asn) of the heavy chain; the CDRL2 residue Arg56 (mutated from Ala) of scFv-4 interacts with Asp109 of the heavy chain.

DISCUSSION

Despite significant progress over the past few years, the success rate of computationally designed libraries of antibodies to bind difficult-to-target epitopes is still very low(Kuroda et al. 2012) mostly due to our limited understanding of the process of antibody affinity maturation *in vivo*. It is often unclear how "germline" antibodies are assembled in the first response to antigens and how the mutations selected during the affinity maturation process contribute to improving binding affinity or selectivity or stability. To overcome these challenges, we first developed an approach named OptCDR(Pantazes and Maranas 2010) that is the first formal computational workflow for optimally selecting the residue composition of the CDRs to enhance binding for a

given antigen. OptMAVEn(Li et al. 2014) extends this to *de novo* designing the entire variable region of antibodies by drawing structural parts from the MAPs database(Pantazes and Maranas 2013), inspired by the natural evolution of an antibody *in vivo*, where the gene of a "germline" antibody is initially assembled by V-(D)-J recombination. In the present study, we aimed to experimentally validate epitope-focused antibody design by designing antibodies against a small peptide.

Despite the challenges encountered in the *de novo* design of antibodies, antibody generation *in silico* targeting a 12-mer antigen was shown here to be successful. Five optimally binding antibodies in the format of scFv emerged as anti-peptide candidates for experimental characterization. Among them, three show experimentally-validated binding to the dodecapeptide in the low nanomolar range without any directed-evolution effort. To our knowledge, the three scFvs are the first experimental validated high-affinity antibodies completely *de novo* designed *in silico*.

The success rate (three out of five designs, 60%) in *de novo* antibody design using OptMAVEn coupled MD simulations is much higher than those of *de novo* protein binder design against the steroid digoxigenin (2 out of 17 designs, 12%)(Tinberg et al. 2013) and stem region of influenza hemagglutinin (2 out of 73 designs, 3%)(Fleishman et al. 2011). This quantitatively demonstrates the efficacy of OptMAVEn coupled MD simulations in capturing the critical structural features of antibodies key to tight binding to antigens. Due to the similarity of general structures of the antibodies, OptMAVEn mediated *de novo* antibody design could efficiently sample both local and non-local contacts that are inherently present in the antibody structures using relatively large structural fragments (V, D, and J modular parts) extracted from known crystal structures, which is important for the successful design. All-atom, explicit solvent MD

simulations have previously proved to be effective at discerning active from inactive computationally designed Kemp eliminases(Kiss et al. 2010; Privett et al. 2012) and aiding in the prediction of domain swapping of computationally designed engrailed homeodomain protein variants(Mou et al. 2015). Inspired by their success, we carried out MD simulations for all *de novo* designed "germline" antibodies with the dodecapeptide prior to in *silico* affinity maturation and experimental validation. Success using MD underlines its importance as a routine approach for involving dynamic properties of proteins in the selection. Encouragingly, all three of our designs exhibit high binding affinity, suggesting the potential of completely bypassing the laborious and time-consuming *in vitro* directed evolution and directly obtaining high affinity antibodies against a specific antigen.

Comparison of "germline" and mature sequences in the three successful designs show that a large number of mutations (>35%) in both framework and CDRs regions were introduced by *in silico* affinity maturation protocol. Although design scFv-5 correctly folded in solution, the failure of design scFv-5 (possessing only 22% mutations during computational affinity maturation) binding to the antigen might be due to the shortage of favorable mutations incorporated. Charged residues, especially Lys, are significantly preferred in our affinity matured designs. Analysis of the modeled interactions between the three successful scFv-dodecapeptide designs show that the CDR H3 loops of all the designed scFvs involve direct interactions with the antigen and highlight that the cation–π interactions and hydrogen bonds formed between them by charged residues are critical determinants for the high binding affinity.

Furthermore, the differences with our scFv designs and the naturally-occurring antibodies suggest that computational *de novo* designs generate a variety of possible antibody solutions binding to the same antigen alluding to a plethora of possible binding possibilities to achieve

high affinity. OptMAVEn's epitope-specificity implies that a cocktail of antibodies, each targeting different epitopes of the antigen, may be designed simultaneously. A cocktail of therapeutic monoclonal antibodies might be more tolerant toward antigen mutations and show better efficacy than single monoclonal antibody. For example, ZMAb, a combination of multiple-neutralizing monoclonal antibodies that recognize three different areas of the Ebola envelope GP protein have been shown to be an effective strategy to improve survival of Ebola-infected patients(Qiu et al. 2012).

In the present OptMAVEn guided antibody generation, the target epitope comprised eight contiguous amino acids. However, most B-cell epitopes(Haste Andersen et al. 2006) in nature consist of residues from different regions of the sequence and are discontinuous. The *de novo* antibody designs against discontinuous epitopes present additional challenges which will be tackled in future investigations using the OptMAVEn design strategy. The successful design for the linear epitope described herein, implies that the design methodology bears great promise for streamlining and greatly facilitating the development of high-affinity antibodies for a plethora antigens such as virus envelope proteins (e.g. HIV gp120(Pantophlet and Burton 2006)) and tumor-associated surface proteins (e.g. B-lymphocyte antigen CD20(Czuczman 2008)).

CONCLUSIONS

We demonstrate here the successful application of a computational framework called OptMAVEn combined with molecular dynamics to *de novo* design antibodies against specific antigen-peptide target using reference system i.e. scFv-2D10, a peptide mimic of mannose-containing carbohydrates. Among the five OptMAVEn *de novo* designed scFvs, three scFvs show nanomolar binding affinities to the dodecapeptide. These results reveal that OptMAVEn can efficiently *de novo* design thermally and conformationally stable antibodies with high

binding affinity to antigens and encourage the targeting of other antigen targets such as virus envelope proteins and tumor-associated surface proteins in the future.

Acknowledgements. This work was supported by the grants of National Science Foundation CBET 1133040 (T.K.W. and C.D.M.) and ACI-1524703 (B.C.G. and K.S.) and the National Institutes of Health grant 9P41GM104601 (B.C.G. and K.S.). The MD simulations in this research were performed on the Blue Waters supercomputers supported by NSF awards OCI-0725070 and ACI-1238993, the state of Illinois, and "Development of rapid diagnostics for Ebola" NSF award ACI-1524703. We thank Dr. Neela Yennawar and Julia Fecko of the Penn State Huck Institutes of the life sciences for assistance with the biolayer interferometry, isothermal titration calorimetry, and circular dichroism.

Author Contributions. V.G.P., T.L., T.K.W. and C.D.M., designed the research. V.G.P., T.L., T.K.W. and C.D.M. analyzed the data and wrote the paper. V.G.P. performed the biochemical work, T.L. performed the computational design work, and B.C.G. and K.S. performed molecular dynamics simulations. All authors discussed the results and commented on the manuscript.

Competing financial interests. The authors declare no competing financial interests.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215(3):403-10.

Blanco-Toribio A, Lacadena J, Nunez-Prado N, Alvarez-Cienfuegos A, Villate M, Compte M, Sanz L, Blanco FJ, Alvarez-Vallina L. 2014. Efficient production of single-chain fragment variable-based N-terminal trimerbodies in Pichia pastoris. Microb. Cell Fact. 13:116.

Burkovitz A, Sela-Culang I, Ofran Y. 2014. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. FEBS J. 281(1):306-19.

Chailyan A, Marcatili P, Tramontano A. 2011. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. FEBS J. 278(16):2858-66.

Clark LA, Ganesan S, Papp S, van Vlijmen HW. 2006. Trends in antibody sequence changes during the somatic hypermutation process. J. Immunol. 177(1):333-40.

Corti D, Lanzavecchia A. 2013. Broadly neutralizing antiviral antibodies. Annu. Rev. Immunol. 31:705-42.

Czuczman MS. 2008. The biology of B-cell CD20 antigen and its potential as a therapeutic target. Ann. Oncol. 19:257-257.

Dalkas GA, Teheux F, Kwasigroch JM, Rooman M. 2014. Cation-pi, amino-pi, pi-pi, and H-bond interactions stabilize antigen-antibody interfaces. Proteins 82(9):1734-46.

Fischer N, Elson G, Magistrelli G, Dheilly E, Fouque N, Laurendon A, Gueneau F, Ravn U, Depoisier JF, Moine V and others. 2015. Exploiting light chains for the scalable generation and platform purification of native human bispecific IgG. Nat. Commun. 6:6113.

Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D. 2011. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332(6031):816-21.

Glaven RH, Anderson GP, Zabetakis D, Liu JL, Long NC, Goldman ER. 2012. Linking Single Domain Antibodies that Recognize Different Epitopes on the Same Target. Biosensors 2(1):43-56.

Gregoire C, Malissen B, Mazza G. 1996. Characterization of T cell receptor single-chain Fv fragments secreted by myeloma cells. European Journal of Immunology 26(10):2410-2416.

Haste Andersen P, Nielsen M, Lund O. 2006. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 15(11):2558-67.

Kiss G, Rothlisberger D, Baker D, Houk KN. 2010. Evaluation and ranking of enzyme designs. Protein Sci. 19(9):1760-73.

Krishnan L, Lomash S, Raj BP, Kaur KJ, Salunke DM. 2007. Paratope plasticity in diverse modes facilitates molecular mimicry in antibody response. J. Immunol. 178(12):7923-31.

Kubala MH, Kovtun O, Alexandrov K, Collins BM. 2010. Structural and thermodynamic analysis of the GFP:GFP-nanobody complex. Protein Science 19(12):2389-2401.

Kuroda D, Shirai H, Jacobson MP, Nakamura H. 2012. Computer-aided antibody design. Protein Eng. Des. Sel. 25(10):507-21.

Lee J, Kim HJ, Roh J, Seo Y, Kim M, Jun HR, Pham CD, Kwon MH. 2013. Functional Consequences of Complementarity-determining Region Deactivation in a Multifunctional Anti-nucleic Acid Antibody. J. Biol. Chem. 288(50):35877-35885.

Lee PS, Ohshima N, Stanfield RL, Yu WL, Iba Y, Okuno Y, Kurosawa Y, Wilson IA. 2014. Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus. Nat. Commun. 5:3614.

Li T, Pantazes RJ, Maranas CD. 2014. OptMAVEn--a new framework for the de novo design of antibody variable region models targeting specific antigen epitopes. PLoS One 9(8):e105954.

Mou Y, Huang PS, Thomas LM, Mayo SL. 2015. Using Molecular Dynamics Simulations as an Aid in the Prediction of Domain Swapping of Computationally Designed Protein Variants. J. Mol. Biol. 427(16):2697-706.

Nelson AL, Dhimolea E, Reichert JM. 2010. Development trends for human monoclonal antibody therapeutics. Nat. Rev. Drug. Discov. 9(10):767-74.
Palmer I, Wingfield PT. 2004. Preparation and extraction of insoluble (inclusion-body) proteins from Escherichia coli. Curr. Protoc. Protein Sci. Chapter 6:Unit 6 3.
Pantazes RJ, Maranas CD. 2010. OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. Protein Eng. Des. Sel. 23(11):849-58.
Pantazes RJ, Maranas CD. 2013. MAPs: a database of modular antibody parts for predicting tertiary structures and designing affinity matured antibodies. BMC Bioinformatics 14(1):168.
Pantophlet R, Burton DR. 2006. GP120: target for neutralizing HIV-1 antibodies. Annu. Rev. Immunol. 24:739-69.
Peled JU, Kuang FL, Iglesias-Ussel MD, Roa S, Kalis SL, Goodman MF, Scharff MD. 2008. The biochemistry of somatic hypermutation. Annu. Rev. Immunol. 26:481-511.
Prischi F, Konarev PV, Iannuzzi C, Pastore C, Adinolfi S, Martin SR, Svergun DI, Pastore A. 2010. Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly. Nat. Commun. 1:95.
Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, Hilvert D, Houk KN, Mayo SL. 2012. Iterative approach to computational enzyme design. Proc. Natl. Acad. Sci. USA 109(10):3790-5.
Qiu X, Audet J, Wong G, Pillet S, Bello A, Cabral T, Strong JE, Plummer F, Corbett CR, Alimonti JB and others. 2012. Successful treatment of ebola virus-infected cynomolgus macaques with monoclonal antibodies. Sci. Transl. Med. 4(138):138ra81.
Simons KT, Kooperberg C, Huang E, Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J. Mol. Biol. 268(1):209-25.
Song HN, Jang JH, Kim YW, Kim DH, Park SG, Lee MK, Paek SH, Woo EJ. 2014. Refolded scFv Antibody Fragment against Myoglobin Shows Rapid Reaction Kinetics. Int. J. Mol. Sci. 15(12):23658-23671.
Sormanni P, Aprile FA, Vendruscolo M. 2015. Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins. Proc. Natl. Acad. Sci. USA 112(32):9902-9907.
Tang XC, Agnihothram SS, Jiao YJ, Stanhope J, Graham RL, Peterson EC, Avnir Y, Tallarico AS, Sheehan J, Zhu Q and others. 2014. Identification of human neutralizing antibodies against MERS-CoV and their role in virus adaptive evolution. Proc. Natl. Acad. Sci. USA 111(19):6863-6863.
Tapryal S, Gaur V, Kaur KJ, Salunke DM. 2013. Structural evaluation of a mimicry-recognizing paratope: plasticity in antigen-antibody interactions manifests in molecular mimicry. J. Immunol. 191(1):456-63.
Tapryal S, Krishnan L, Batra JK, Kaur KJ, Salunke DM. 2010. Cloning, expression and efficient refolding of carbohydrate-peptide mimicry recognizing single chain antibody 2D10. Protein Expr. Purif. 72(2):162-8.
Tiller KE, Tessier PM. 2015. Advances in Antibody Design. Annu. Rev. Biomed. Eng.
Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL and others. 2013. Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501(7466):212-6.
Umetsu M, Tsumoto K, Hara M, Ashish K, Goda S, Adschiri T, Kumagai I. 2003. How additives influence the refolding of immunoglobulin-folded proteins in a stepwise dialysis system - Spectroscopic evidence for highly efficient refolding of a single-chain FV fragment. J. Biol. Chem. 278(11):8979-8987.

FIGURE LEGENDS

Fig. 1. Revised OptMAVEn strategy for *de novo* antibody design. (a) Antigen dodecapeptide (PDB 4H0H) (b) Step 1: sample antigen positions in a predefined antibody binding site. The binding site is represented by a rectangular box that covers all mean epitope coordinates by analyzing 750 antibody-antigen structures which are superimposed onto a reference antibody structure whose coordinate center of CDRs attachment points was placed at the origin. (c) Step 2: assign best V, (D) and J MAPs antibody modular parts (Pantazes and Maranas 2013) to assemble the germline antibody models against the dodecapeptide. (d) Step 3: refine the antibody-peptide conformation using molecular dynamics. (e) Step 4: select the favorable mutations in the antibody to improve the stability and binding to the dodecapeptide.

Fig. 2. (a) Time evolution of the RMSD values during the 100 ns molecular dynamics of the 31 antibody designs complexed with the dodecapeptide. (b)-(f) Snapshots of representative designed scFvs complexed with the dodecapeptide during simulations. (b) Antigen stably bound. The scFv remains stably bound to the dodecapeptide in the same binding pose during 100ns of simulation with an accompanying rearrangement of a few residues (shown in sticks) surrounding the dodecapeptide. (c) Antigen relocation. The dodecapeptide relocates to a new binding pocket. (d) Antigen reorientation. The dodecapeptide remains in the same binding pocket but executes a significant orientation change. (e) Antigen partially bound. Parts of the dodecapeptide become unbound. (f) Antigen unbound. The dodecapeptide shifts out of the binding site. (g) Flowchart of the selections of "Germline" antibodies using MD for further in silico affinity maturation.

Fig. 3. Multiple sequence alignments of the five *de novo* designs and scFv-2D10 antibody. (a) H chain. (b) L chain. The alignments were performed with UGENE (Okonechnikov et al. 2012) and the Clustal X coloring scheme was used. FRs and CDRs regions are indicated on top of each alignment based on the IMGT numbering system.

Fig. 4. Construction of *de novo* scFv for protein expression in *E. coli*. (a) Schematic representation of scFv gene construction. (b) Agarose gel analysis of PCR products of all

the scFvs in this study indicating amplified gBlock fragments. *Lane M*, DNA marker (1 kb ladder); *Lane 1*, scFv-1; *Lane 2*, scFv-2; *Lane 3*, scFv-3; *Lane 4*, scFv-4; *Lane 5*, scFv-5; and *Lane N*, scFv-2D10. (c) Analysis of purified and refolded *de novo* designed scFvs on SDS-PAGE. *Lane M*, protein molecular weight markers; *Lane 1*, scFv-1; *Lane 2*, scFv-2; *Lane 3*, scFv-3; *Lane 4*, scFv-4; *Lane 5*, scFv-5; *Lane N*, scFv-2D10. (d) Far UV CD spectra of refolded scFvs characterized in this study. The far-UV CD spectra of scFv proteins were recorded in wavelength range from 190 to 250 nm (x-axis) and are expressed as CD[mDeg] value, which represents the ellipticity (y-axis). The predicted (GOR4) and actual (CDSSTR and CONTIN of CDPro software) analysis were in agreement in having more β-sheets than α-helical content.

Fig. 5. Binding curves of *de novo* designed scFvs with dodecapeptide used to determine K_D values *via* biolayer interferometry on the Octet QK System. All scFvs were His_6-tagged to load on Ni-NTA biosensors. (a) The sensogram shows the interaction of *de novo* designed scFvs (550 nM) binding to dodecapeptide (1000 nM) (b) Optimization of antigen concentrations, the sensogram shows the interaction of scFv-2 (550 nM) with different antigen concentrations (1000 nM, 300 nM, 100 nM, 30 nM, 10 nM, and 3 nM). (c) Optimization of scFv concentrations, the sensogram shows the interaction of antigen (30 nM) with different scFv-2 concentrations (1110 nM, 550 nM, 460 nM, 370 nM, 280 nM, and 180 nM). (d) Each scFv (550 nM) of the current system titrated against 30 nM of the non-related antigen, GCN4 (e) Non-related scFv (550 nM) designed for yeast transcription factor, GCN4 (YHLENEVARLKK-C-BSA) (Zahnd et al. 2004) titrated against the dodecapeptide antigen (30 nM) in the present study (e) With the optimized antibody and antigen concentrations, the sensogram shows the interaction of *de novo* designed and 2D10 scFvs (550 nM) binding to dodecapeptide antigen (30 nM).

Fig. 6. (a) Counts of amino acid mutation types before and after computational affinity maturation of the three successful designs: scFv-1, scFv-2, scFv-4. b. Amino acid propensities of the three successful designs. (c-e) Models of scFv-1, scFv-2 and scFv-4 bound to dodecapeptide. Hydrogen bonds and cation-π interactions are shown in black

and magenta dashed lines, respectively. (f) Structure of scFv-2D10 bound to dodecapeptide (PDB 4H0H).

Table I. Summary of 31 best-designed "germline" antibodies ranked by interaction energy or RMSDs.

Antibody[a]	H chain			L chain[b]			IE[c]	RMSD[d]
	V	CDR3	J	V	CDR3	J		
120_10148_1	7	343	3	2 (K)	25 (K)	2 (K)	-530	
120_10148_2	7	343	3	3 (K)	25 (K)	2 (K)	-530	
160_8161_1	13	368	5	45 (K)	10 (K)	1 (K)	-527	
140_10149_1	135	43	5	2 (K)	171 (K)	2 (K)	-502	
140_10149_2	136	43	5	3 (K)	171(K)	2 (K)	-502	
140_9977_4	116	126	5	32 (L)	14 (L)	1 (L)	-488	
160_8161_2	13	270	5	45 (K)	10 (K)	1 (K)	-483	
120_15439_1	39	124	1	9 (K)	19 (K)	3 (K)	-481	
160_10005_1	13	62	5	2 (K)	6 (K)	5 (K)	-480	
160_10005_2	13	62	5	3 (K)	6 (K)	5 (K)	-480	
120_13389_2	30	207	1	45 (K)	86 (K)	3 (K)	-478	
160_10175_1	13	43	5	9 (K)	85 (K)	2 (K)	-468	
160_10162_1	13	43	5	2 (K)	2 (K)	3 (K)	-467	
160_10162_2	13	43	5	3 (K)	2 (K)	3 (K)	-467	
120_15439_2	39	124	1	48 (K)	19 (K)	3 (K)	-467	
140_9976_1	120	126	5	31 (L)	14 (L)	1 (L)	-465	
140_9976_2	120	126	5	32 (L)	14 (L)	1 (L)	-464	
120_13389_2	30	207	1	45 (K)	145 (K)	3 (K)	-458	
120_6290_1	12	207	5	45 (K)	10 (K)	1 (K)	-456	
140_4402_1	12	257	5	31 (K)	33 (K)	1 (K)	-455	
200_15222_5	137	283	1	45 (K)	158 (K)	3 (K)		2.9
140_15899_5	59	43	3	16 (K)	81 (K)	4 (K)		3.0
120_13234_5	97	212	1	45 (K)	30 (K)	1 (K)		3.2
160_15235_1	87	283	1	45 (K)	25 (K)	1 (K)		3.4
160_14896_1	133	283	1	45 (K)	25 (K)	1 (K)		3.6
240_4115_2	135	30	5	44 (K)	6 (K)	5 (K)		3.6
140_16152_1	13	7	5	9 (K)	15 (K)	1 (K)		3.6
160_15235_2	77	283	1	45 (K)	25 (K)	1 (K)		3.7
140_13234_5	116	207	1	45 (K)	30 (K)	5 (K)		3.7
200_15729_2	20	283	4	48 (K)	3 (K)	5 (K)		3.9
200_15222_2	133	283	1	45 (K)	158 (K)	3 (K)		4.0

[a] Antibody is named based on the rotation angle and order during the antigen sampling in the antibody-binding site.

[b] The number of V, CDR3 and J modular parts in the MAPs database. The MAPs database is composed of 929 "parts" that can be assembled to create 2.3×10^{10} unique antibodies which is in fact more antibodies than can be assembled by the human immune system through rearrangement of the V, D, and J gene (Pantazes and Maranas 2013). K and L in the parenthesis represent KAMPA and LAMBDA light chains, respectively.

[c] MILP interaction energies. Unit in kcal/mol.

[d] The RMSD between the docked and best-positioned antigen conformations.

Table II. Summary of energies and mutations of the five *de novo* designed scFvs against the dodecapeptide.

Antibody	Stage[a]	Complex Energy[b]	IE[c]	Mutation count[d]			
				H chain		L chain	
				FR	CDR	FR	CDR
scFv-1	Before	-6585	-43	23	17	24	16
	After	-11939	-474	(10%)	(7%)	(10%)	(7%)
scFv-2	Before	-8046	-37	28	16	26	19
	After	-11629	-428	(12%)	(7%)	(11%)	(8%)
scFv-3	Before	-4712	-51	24	16	31	13
	After	-7168	-360	(11%)	(7%)	(14%)	(6%)
scFv-4	Before	-6602	-96	23	21	19	20
	After	-11110	-550	(10%)	(9.4%)	(9%)	(9%)
scFv-5	Before	-5708	-4.8	13	12	14	11
	After	-8704	-225	(6%)	(5%)	(6%)	(5%)
2D10[e]		-3868	-181	6 (3%)	5 (2%)	1 (1%)	0 (0%)

[a] Before or after computational affinity maturation.
[b] The entire complex energy. Unit in kcal/mol.
[c] The interaction energy between the antibody and antigen using CHARMM force field. Unit in kcal/mol.
[d] The number of mutations between the designed mature sequence and the "germline" sequence. The mutation frequency was calculated by (number of mutations)/(number of residues in scFv except for the linker).
[e] The mutation count for scFv-2D10 is based on the germline gene assignment in IMGT.

Table III. Pairwise sequence similarity of the designed scFv and 2D10. Lower left and upper right triangular parts are the sequence identities of H and L chains, respectively. The sequence alignments were performed by BLAST (Altschul et al. 1990).

	2D10	**scFv-1**	**scFv-2**	**scFv-3**	**scFv-4**	**scFv-5**
2D10	1	56%	49%	47%	41%	57%
scFv-1	43%	1	57%	52%	44%	54%
scFv-2	36%	49%	1	46%	47%	47%
scFv-3	38%	57%	49%	1	54%	45%
scFv-4	32%	39%	48%	44%	1	43%
scFv-5	46%	49%	46%	62%	41%	1

Table IV. Kinetic data for the binding of refolded *de novo* designed scFvs with the dodecapeptide obtained by Octet QK[a].

scFv	Molar Conc [nM]	k_{obs}[b] x 10^{-3} [s^{-1}]	Error in k_{obs} x 10^{-5}	k_d x 10^{-4} [s^{-1}]	Error in k_d x 10^{-5}	k_a[c] x 10^4 [$M^{-1}s^{-1}$]	Error in k_a x 10^{-5}	K_D[d] [nM]	Error in K_D [nM]
scFv-1	30	3.3	1.3	7.6	1.2	8.6	1.8	8.9	2.3
scFv-2	30	4.1	0.8	13.2	0.7	9.1	1.1	14.4	1.9
scFv-4	30	3.9	1.1	17.1	1.3	7.2	1.7	23.8	5.8
scFv-2D10	30	7.0	4.0	7.8	1.5	20.6	4.3	3.8	1.1

[a] Errors are from model fitting.
[b] k_{obs} is the observed rate constant that reflects the overall rate of the combined association and dissociation of the two binding partners.
[c] $k_a = (k_{obs} - k_d)$ / [Analyte].
[d] K_D represents the ratio of the association rate constant (k_a) to the dissociation rate constant (k_d).

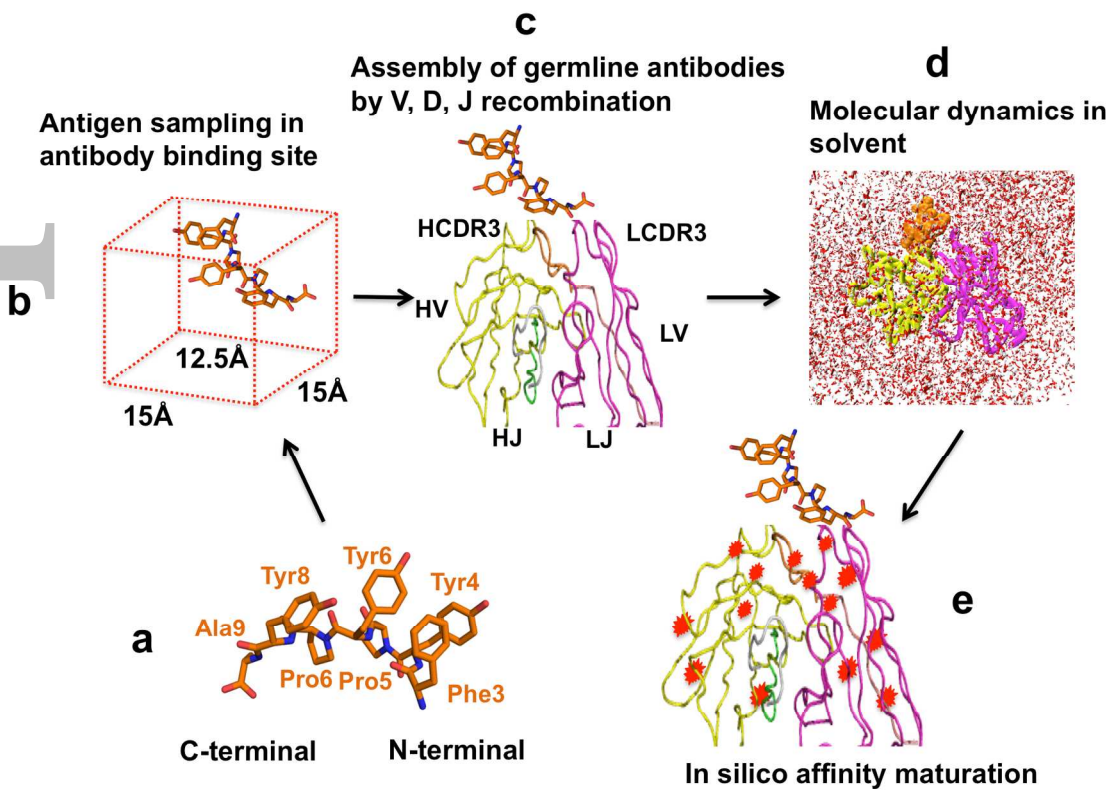


Figure 1

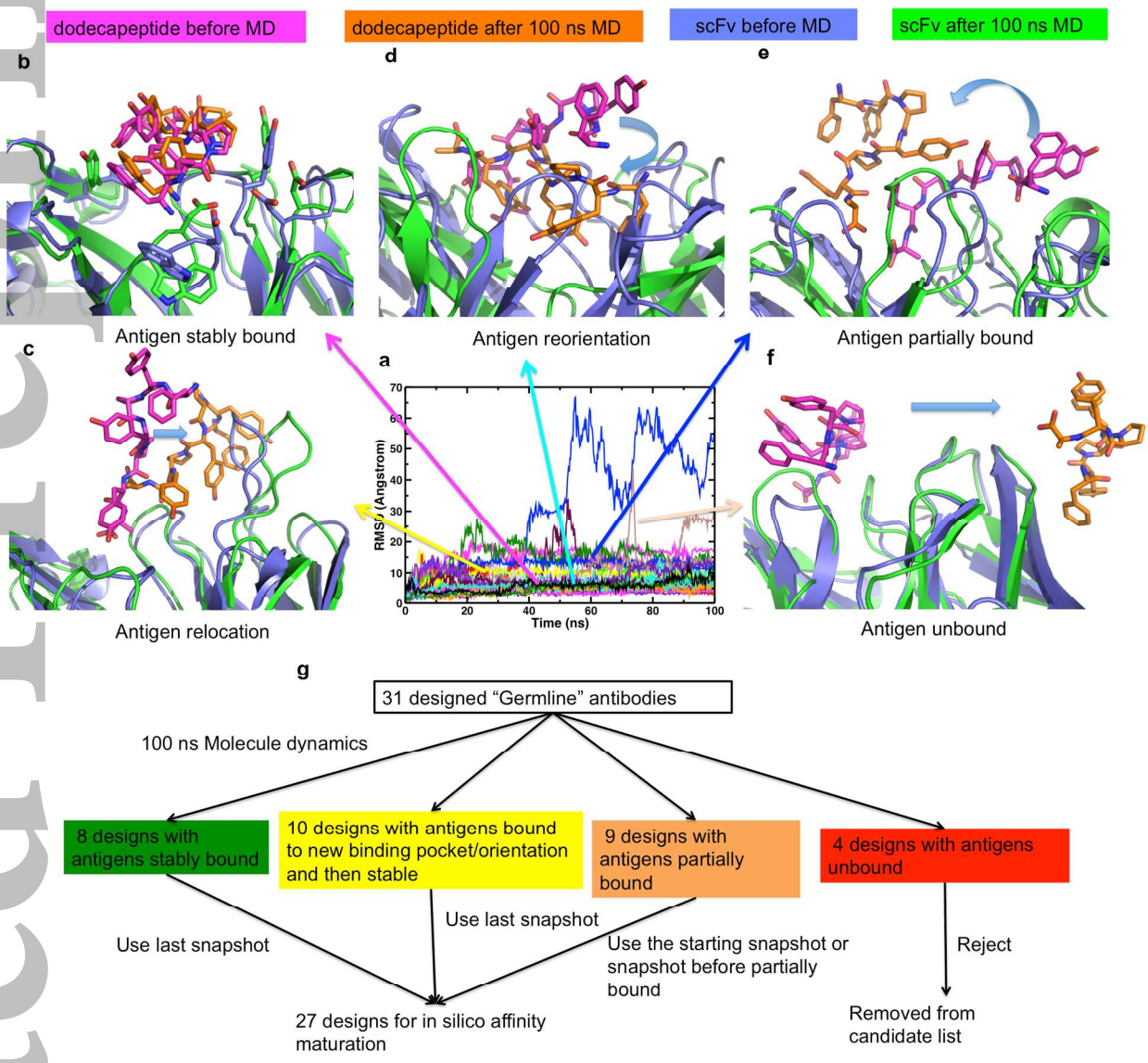


Figure 2

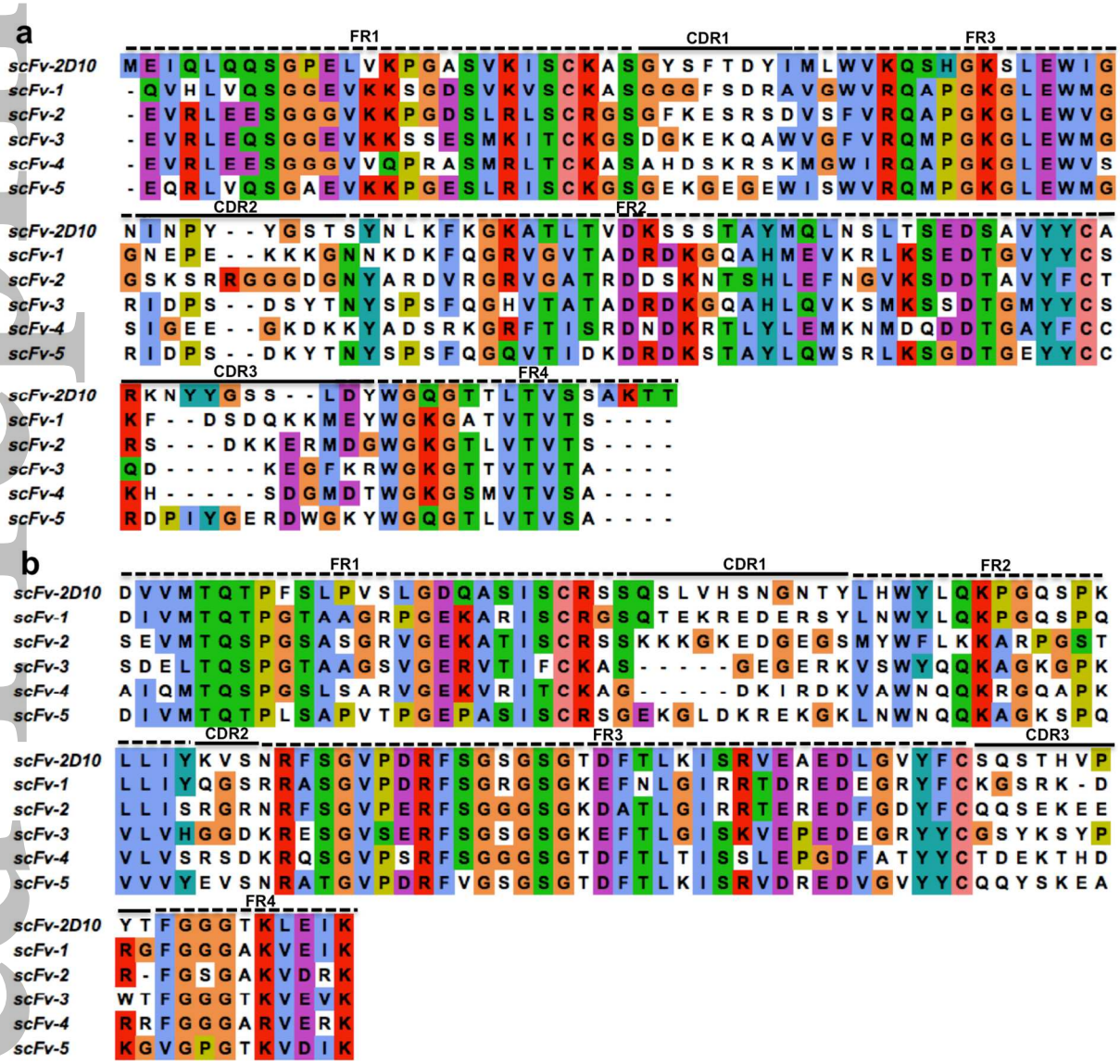
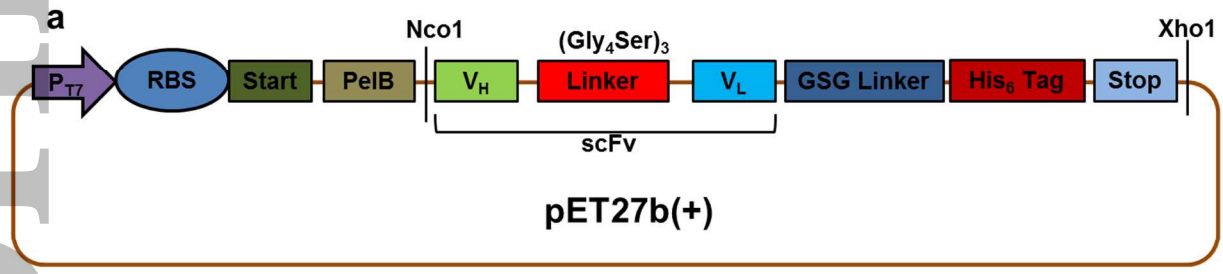
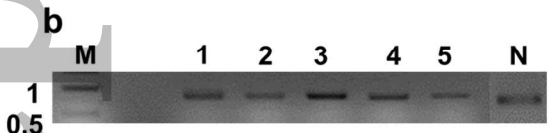
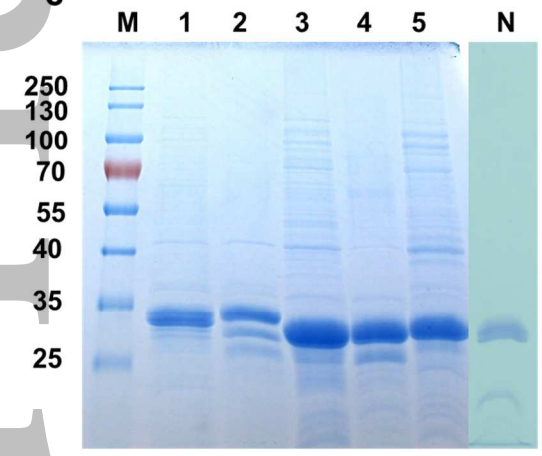
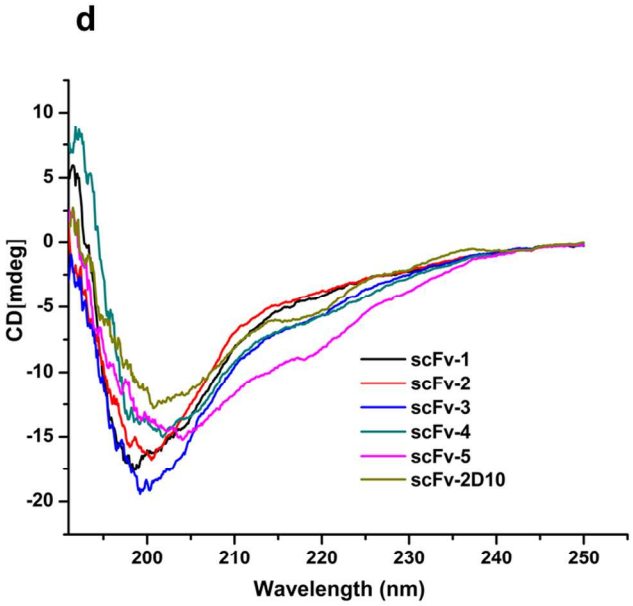


Figure 3









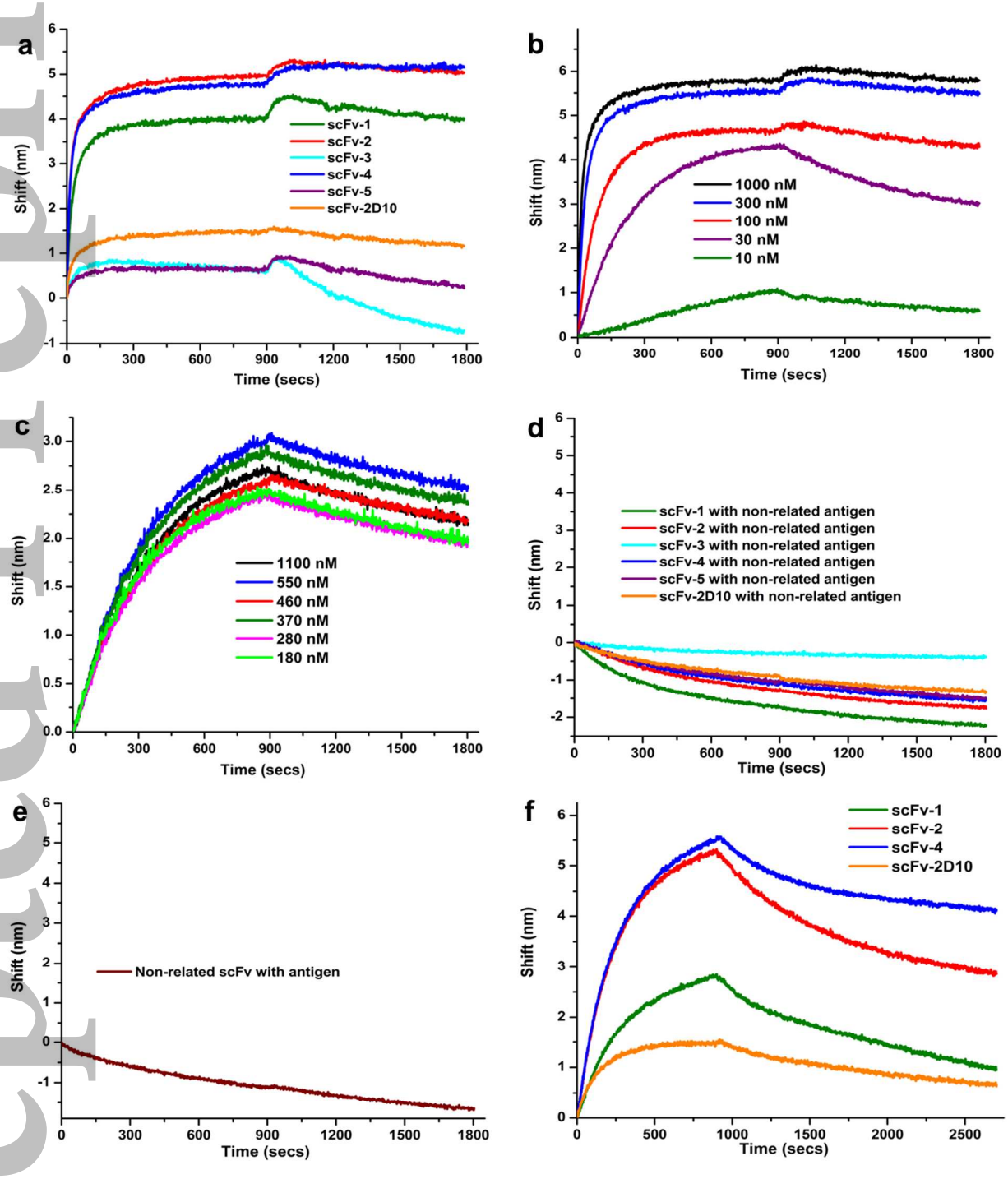


Figure 5

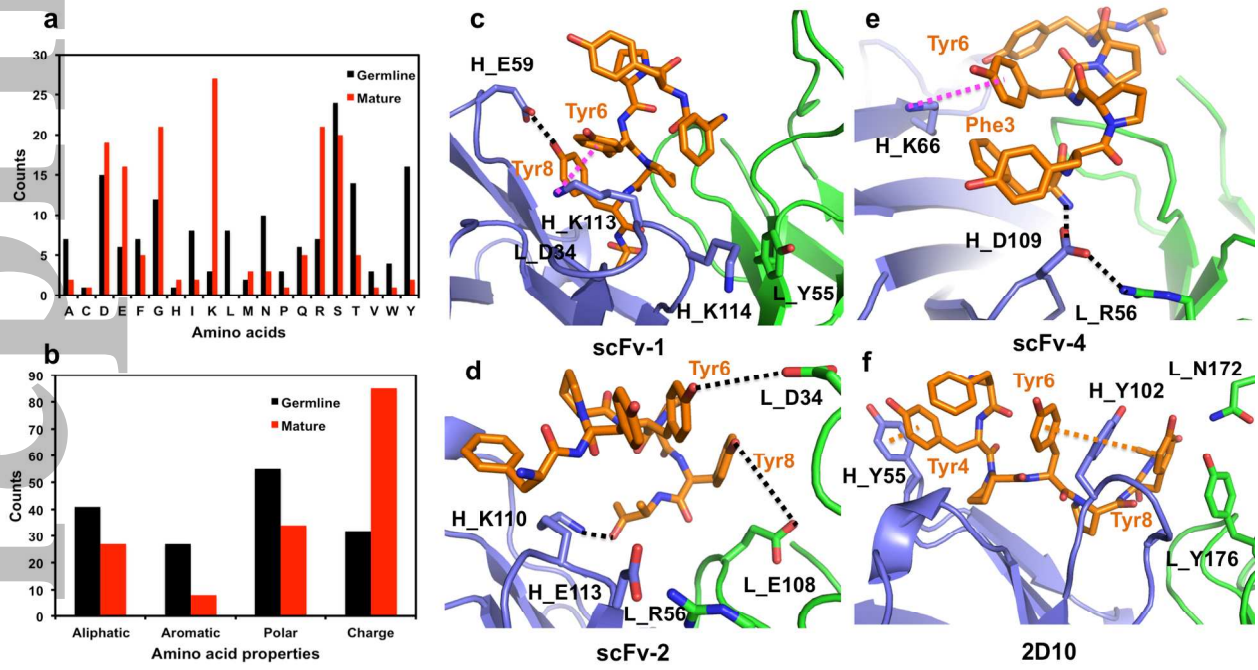


Figure 6