EPFL

Thèse n°7545

Functional de novo Protein Design for Targeted Vaccines and Synthetic Biology Applications

Présentée le 17 janvier 2020

à la Faculté des sciences et techniques de l'ingénieur
Laboratoire de conception de protéines et d'immuno-ingénierie
Programme doctoral en biotechnologie et génie biologique

pour l'obtention du grade de Docteur ès Sciences

par

Che YANG

Acceptée sur proposition du jury

Prof. M. Dal Peraro, président du jury
Prof. B. E. Ferreira De Sousa Correia, directeur de thèse
Prof. O. Hartley, rapporteur
Prof. I. André, rapporteur
Prof. O. Hantschel, rapporteur

École polytechnique fédérale de Lausanne

2020

Acknowledgements

For a long time, I have truly admired for an old saying: “The ultimate purpose in one's life is to help each other to fulfill their own mission”. Therefore, if the work presented in this thesis contributes to the progress of scientific research, it has been built on countless supports and encouragement of many extraordinary people.

First and utmost important, I would like to express my sincere gratitude to my advisor and lifelong mentor, Prof. Bruno Correia, for providing me an opportunity to conduct the research with him and his incredibly enthusiastic support throughout the last four years. You led me into the field of computational biology, and always being there to give me your intellectual suggestions. Your unlimited positive energy and a strong will are absolutely unique, which made me grow mentally stronger than what I can’t imagine. I would specifically like to thank you for always reminding me to concentrate on dealing with my defects and firmly improving it when I was at a loss to know how to move on. I am incredibly grateful for this experience and for the many things both personally and professionally that I learned from you, the way you thought and the action you took. Truly, I could not have imagined having a better advisor and mentor for my Ph.D. study rather than you. Thank you!

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Oliver Hartley, Prof. Ingemar André, Prof. Oliver Hantschel and Prof. Matteo Dal Peraro, for their insightful comments and encouragement, but also for the inspired question which incited me to widen my research from various perspectives. It was a great pleasure and an amazing experience for me to present my work and sharing ideas with you.

Next, I would like to thank Fabian Sesterhenn, who took the challenge together with me to be the pioneers in this newly founded lab, and worked closely throughout the last four years with me. At least half of the works presented in my thesis were attributed to his efforts. We harmoniously worked together to moving forwards our project and had lots of fun on the journey. I always appreciate your talent in science, making you have an extraordinary passion for it and think the related problem in a constructive way. I consider myself very fortunate for having a chance to learn, grow and mature with you, and eventually, I believe both of us will succeed in the field we want to explore. Personally, we become a lifelong friend to share everything and every moment we experienced, and I want to express my deep appreciation for you (which I definitely owe you) that you are always standing beside me throughout all my ups and downs, happy and not so happy moments. I credit you as one of the most crucial elements for my own growth. I would like to thank you very, very much.

Special thanks are also given to Dr. Jaume Bonet and Dr. Pablo Gainza, postdocs as two pillars in our laboratory, for their insightful advice and valuable discussions regarding both science and life decision. Both of you are an exceptionally talented scientist and set up the highest standard of being a senior in the scientific field. Not only is a bellwether in our team but more importantly, you were always extremely supportive of in leading me through the process of Ph.D. Also, I would like to thank you for carefully checking my thesis and for helping me to prepare my Ph.D. private defense. Thank you for being a great friend, for the time we spent at lunchtime and after-work-drink.

All the great idea requires the experimental proof, with a special mention to Stéphane Rosset, Patricia Corthésy, Sandrine Georgeon, Mélanie Villard and Katyayanee Neopane, the very talented and diligent

technicians of the lab, your help made all the designed proteins real. It was fantastic to have the opportunity to work with you. Thank you very much for the work you have done.

During the last four years, I have a wonderful time surrounded by fascinating people in LPDI. I would like to thank Andreas Scheck for helping me with enormous questions regarding RosettaScript and also sharing with many interesting things happening in your band. Thank Sailan Shui as my great chatting companion to discuss many interesting sciences and also providing many great suggestions for my project. Thank all the members in LPDI that your guidance and accompany along the way are the warmest comfort and power for me.

Special mention goes to those helpful members in protein production and structure core facility, Dr. Pojer, Dr. Hacker, Dr. Lau, Dr. Abriata, and Dr. Lopez in Prof. Dal Peraro's group, they taught me the principle of many cutting-edge techniques and how to use them to facilitate the structural determination of our designed proteins. Besides, their encouragement and help made me feel confident to fulfill my desire and to overcome every difficulty I encountered.

As a foreigner studying abroad, friends with the same root always provided me moral and emotional support along the way. Especially thank my bosom friend Kuang-Yu Yang as my big brother for mentally accompanying me during the first and second years of Ph.D. period, and sharing your wise opinions with pure enthusiasm in life when I faced difficulties. Thank Hsiang-Chu Wang for comprehensive care of my daily life and feeding me with your spectacular food. Thanks to Yuan-Peng, Shang-Jung, Henry, Kun-Han, Chin-Lin, Chun-Lam, and so many others for your friendships and supports.

Finally, the family is the strongest backup for me, and I am very grateful for my parents, Yu-Hui and Yao-Sung for their unlimited love and continued support. Your understanding and your love encouraged me to pursue my dream and what I like even I need to be separated ten thousand kilometers away from you. Your firm and kindhearted personality have affected me to be steadfast and never bend to difficulty. Also, I thank my brother, Sheng, that you take care of mom and dad during the time I am absent while I know you are also struggling with your heavy workload. For all of you, thank you for showing me the attitude of life and the responsibilities coming with love.

Last, I would like to thank the most significant person in my life – Min Hsin, my loving wife. We have gone through every stage of life and weathered every up and down, happiness and sadness, and whenever I turned back, that is you always standing behind me and giving me the courage to sail through the storm. Without you, I have already staggered on the way and lost the direction in the wind, thank you for being my best half and making me complete as a man. I am looking forward to our next adventure and continue…

In the end, with my honor, this thesis is dedicated to all who care about me, who help me, thank you.

Lausanne, 1 January 2020 Che Yang

Abstract

Evolution has created, selected and evolved large repertoires of proteins that operate in various biological systems. Nowadays biotechnological needs are coming up orders of magnitude faster than proteins naturally evolve. The emergence of *de novo* protein design accelerates the process of creating new proteins to explore a virtually infinite number of protein folds and sequences that can potentially be empowered with functionality. So far, the majority of successful *de novo* proteins were ideal folds abundant in regular secondary structures to achieve the goals of structural accuracy and thermostability of proteins. Albeit invaluable, these studies have limited the generalization of this approach for designing "non-ideal" features in the protein. In fact, often molecular function of a protein involves irregular structural segments, including flexible loops and cavities to interact with their biological counterparts. Thus, there remains a significant need for research in this field for installing and stabilizing structurally complex segments in designed proteins, and advance is needed in order to build and understand the structure–function relationship in the upcoming challenges in biological applications.

As part of my thesis, I described two computational protocols to approach the design of proteins carrying irregular functional motifs from different perspectives; although the conceptual ideas varied, both methods eventually converged to construct the scaffolds for a function of choice while simultaneously optimizing protein stability. The first method, Rosetta FunFolDes, aimed to maximize the backbone flexibility of the existing protein structure to effectively stabilize the grafted structural motif. The second method, on the other hand, systematically built *de novo* protein topologies centered at a given functional motif to customize the structural features required for motif stabilization.

A particular field that will benefit from these computational approaches is the engineering of the epitope-focused immunogens for vaccine development. Previously, epitope immunogens have proved to be effective in eliciting potent neutralizing antibodies for an intractable pathogen: respiratory syncytial virus (RSV). The development of those immunogens relies on the identification of the neutralizing antibodies and structural characteristics of their epitopes, followed by the transplantation of the epitope from its viral context into a heterologous scaffold for epitope stabilization. However, due to the lack of a generalized tool, the translation of complex molecular details into the design of effective immunogens is lagging far behind the available antibody–epitope information.

Here, we borrowed the structural information of complex neutralization epitopes isolated from RSV and applied computational methods in order to engineer *de novo* proteins carrying irregular and discontinuous epitopes. Noticeably, we demonstrated that the structures of designed scaffolds closely resemble the design models, with sub-angstrom structural accuracy on the region of the epitopes. *In vivo*, we showed that cocktail formulations of three *de novo* designed immunogens presenting distinct epitopes induced RSV neutralizing antibodies in naïve nonhuman primates. Moreover, when pre-existing immunity is established, the designed immunogens are able to act as boosting immunogens to redirect the response onto the defined specificity. Also, we further functionalized the epitope-scaffolds as a BRET-based diagnostic biosensor to profile antibody responses in the serum, providing a valuable tool with which to quantify vaccination result with single epitope resolution.

Beyond immunogen design, the use of our computational methods was applicable to assembling a protein topology bi-functionalized presenting two distinct binding motifs, which we applied as a non-natural inducer to control receptor-mediated signaling, and modulation of transgene expression in synthetic cells.

Altogether, our advance in RSV-based studies demonstrates a substantial step for the utilization of epitope immunogens based on computationally designed proteins and may contribute to developing immunogens for other pathogens. The new methodological pipeline enables the design of a *de novo* protein with complex functional sites, which is a general blueprint for protein design to unravel new rules to create previously unimagined functional proteins.

Keywords: *De novo* protein design, protein design, protein engineering, Rosetta, epitope-focused immunogens, reverse vaccinology, vaccine, respiratory syncytial virus (RSV)

Graphical abstract

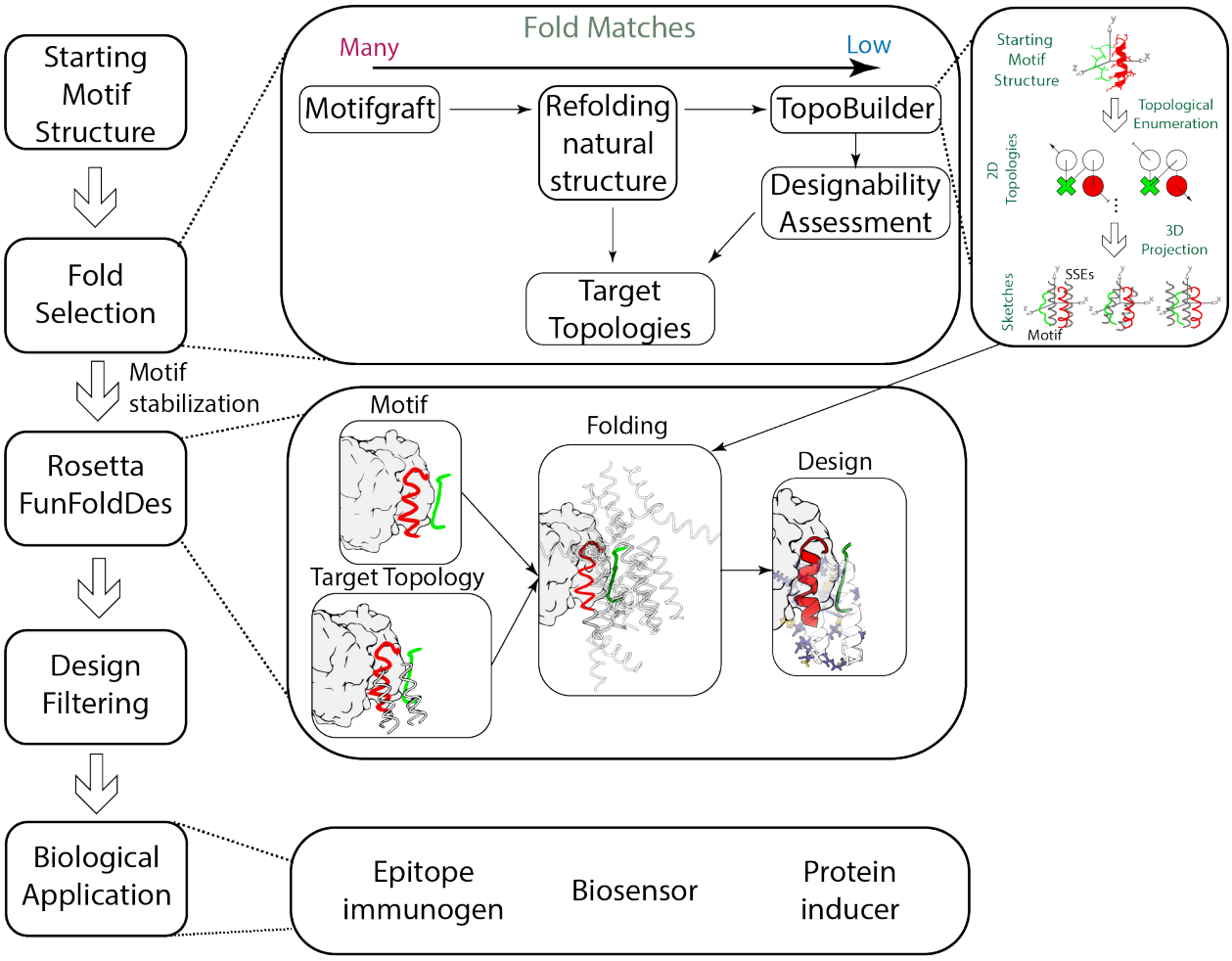


Figure 0.1. Pipeline for repurposing functional motifs for *de novo* protein design.

This thesis describes two computational protocols to leverage designable protein topology for repurposing functional motifs onto heterologous scaffolds, and applies the designed proteins to vaccine and biomedical developments.

Résumé

Le mécanisme évolutif a progressivement sélectionné et développé 10 000 protéines uniques, jouant un rôle dans diverses fonctions biologiques chez l'être humain; De nos jours, nos besoins évoluent plus rapidement que les protéines naturelles. L'émergence de la conception de protéines *de novo* accélère le processus d'exploration d'un nombre pratiquement infini de replis protéiques pouvant potentiellement être dotés de fonctionnalités. Jusqu'à présent, la majorité des succès de nouvelles protéines exigent la conception de plis idéaux abondants dans les structures secondaires régulières pour obtenir des protéines structurellement précises et thermostables, limitant ainsi la généralisation de cette approche pour la conception de caractéristiques «non idéales» dans la protéine. En effet, la fonction moléculaire d'une protéine implique généralement les segments structurels irréguliers, y compris les boucles flexibles et les cavités, pour interagir avec leurs homologues biologiques. Il reste donc un besoin important de recherche en ce qui concerne l'installation et la stabilisation de segments structurellement complexes pour la protéine conçue, ainsi que des progrès sont nécessaires pour établir et comprendre la relation structure-fonction dans les défis à venir des applications biologiques.

Dans le cadre de ma thèse, nous avons décrit deux protocoles informatiques permettant d'aborder la conception de motifs fonctionnels irréguliers porteurs de protéines selon différentes perspectives; bien que les idées conceptuelles aient varié, les deux méthodes ont finalement convergé pour construire les échafaudages protéique selon une fonction de choix tout en optimisant simultanément la stabilité de la protéine. La première méthode, RosettaFunFolDes, visait à maximiser la flexibilité de l'épine dorsale de la structure protéique existante pour stabiliser efficacement le motif structural greffé. La seconde méthode, par contre, construisait systématiquement les topologies de protéines *de novo* centrées sur un motif fonctionnel donné pour personnaliser le contrôle exquis des caractéristiques structurelles requises pour la stabilisation du motif.

L'ingénierie d'immunogènes à base d'épitopes pour le développement de vaccins est un domaine qui bénéficiera de ces approches informatiques. Auparavant, les immunogènes d'épitopes se sont révélés efficaces pour contrôler les spécificités des anticorps fins et pour susciter de puissants anticorps neutralisants pour un certain nombre d'agents pathogènes intraitables. Le développement de ces immunogènes repose sur l'identification des anticorps neutralisants et des caractéristiques structurelles de leurs épitopes, puis sur le transfert de la région de contact hors de son contexte viral dans l'échafaudage protéique hétérologue pour la stabilisation des épitopes. Toutefois, en raison d'absence d'outil généralisé, la traduction de détails moléculaires complexes dans la conception d'immunogènes efficaces prend beaucoup de retard par rapport aux informations disponibles sur les anticorps et les épitopes.

Ici, nous avons emprunté les informations structurelles des épitopes de neutralisation complexes isolés du VRS et des méthodes de calcul appliquées afin de mettre au point des protéines *de novo* portant des épitopes irréguliers et discontinus. De manière remarquable, nous avons démontré que les structures des échafaudages conçus ressemblent beaucoup aux modèles de conception, avec une précision structurelle inférieure à l'angström sur la région des épitopes complexes. *In vivo*, nous avons montré des formulations de cocktails de trois immunogènes conçus *de novo* et présentant des épitopes distincts induisant des anticorps neutralisant le VRS chez des primates naïfs non humains. En outre, lorsque l'immunité préexistante est établie, les immunogènes conçus sont capables d'agir en tant qu'immunogènes amplificateurs pour rediriger la réponse sur l'immunité définie. En outre, nous avons en outre fonctionnalisé l'immunogène d'épitope en tant que biocapteur de diagnostic basé sur le BRET pour établir le profil des réponses d'anticorps dans le sérum, ce qui constitue un outil précieux pour quantifier le résultat de la vaccination avec une résolution à un seul épitope.

Au-delà de cette conception immunogène, l'utilisation de nos méthodes informatiques s'appliquait à l'assemblage d'une topologie protéique bi-fonctionnalisée avec deux motifs de liaison distincts, que nous avons appliquée en tant qu'inducteur non naturel pour contrôler la signalisation médiée par le récepteur et la modulation de l'expression transgénique dans des cellules artificielles.

Globalement, notre avancée dans les études basées sur le VRS démontre une étape importante dans l'utilisation des immunogènes d'épitopes à base de protéines conçues par ordinateur et peut contribuer à développer l'immunogène précis pour d'autres pathogènes. En outre, le nouveau pipeline méthodologique permet de concevoir une protéine *de novo* avec des sites fonctionnels complexes, ce qui constituera un modèle polyvalent à utiliser dans le domaine de la conception de protéines afin de définir de nouvelles règles pour créer une protéine fonctionnellement complexe auparavant inimaginable.

Mots - clés : Conception de protéines *de novo* , conception de protéines, ingénierie des protéines, Rosetta, immunogènes focalisés sur les épitopes, vaccinologie inverse, vaccin, virus respiratoire syncytial (VRS)

Table of Contents

List of Figures

List of Supplementary Figures

List of Tables

List of Abbreviations

BnAb	Broadly neutralizing antibody
BRET	Bioluminescence resonance energy transfer
CPD	Computational protein design
DNA	Deoxyribonucleic acid
DOF	Degree of freedom
E.*coli*	Escherichia coli
FunFolDes	Functional folding and design algorithm
GEMS	Generalized extracellular molecule sensor
GMEC	Global minimum energy conformation
HA	Hemagglutinin
HIV	Human immunodeficiency virus
hMPV	Human metapneumovirus
mAb	Monoclonal antibody
MC	Monte Carlo-based algorithms
MRES-CoV	Middle East respiratory syndrome coronavirus
nAb	Neutralizing antibody
NHP	Non-human primate
NMR	Nuclear magnetic resonance
PDB	Protein data bank
RMSD	Root mean square deviation
RNA	Ribonucleic acid
RSV	Respiratory syncytial virus
RSVF	Respiratory syncytial virus fusion protein
SPR	Surface plasmon resonance
SSE	Secondary structure element
TriVax	Trivalent cocktail vaccine

Chapter 1 Introduction

In the drama of life on a molecular scale, proteins are where the action starts. --- Arthur M. Lesk, 2001

1.1 Overview

As a genetic transcript from DNA, proteins encode the essential molecules for supporting living organisms spanning a variety of cellular functions that include enzymatic activities, molecular scaffolding, mechanical matrix and immune responses among others (*1*). Proteins are synthesized according to the DNA sequence of the corresponding gene as a polypeptide chain covalently linked between amino acids through an amide bond. Naturally, there are 20 enantiomeric L-form amino acids used to build protein polymers in cells, and each amino acid composed of an asymmetric carbon, Cα linked to an amine group (NH_2), a carboxyl group (COOH), a hydrogen H and a side chain determining the physicochemical properties of the amino acid.

Based on Taylor's classification (*2*), amino acids can be categorized as three categories: charged, polar, hydrophobic. Under physiological conditions, a nascent polypeptide chain spontaneously collapses into a folded structure by local or long-range interactions between residues and eventually adopts the lowest energy state. In 1954, Anfinsen through a seminal experiment of demonstrated using ribonucleaseA that all the information necessary for folding a protein into a three-dimensional structure is solely encoded in its amino-acid sequence (primary structure) (*3*). Theoretically, all the bonds and angles in a polypeptide chain are rotatable, two rotations along the protein backbone (a repeating unit of (NH)Cα(CO)) are described respectively by the dihedral angles Ø and Ψ (Fig. 1.1); whereas, the rotatable dihedral angles in the side-chains are denoted by χ_1, χ_2. During the translation process, the growing polypeptide chain forms local substructures by adopting dihedral angles that fold into optimal conformations to maximize the molecular interactions, especially forming hydrogen bonds between backbone atoms. The possible angles of Ø and Ψ are constrained by steric hindrance across the peptide plane, the most frequent conformations for Ø and Ψ were well described by the Ramachandran distributions (*4*).

Based on the hydrogen-bond patterns and dihedral angles along the backbone of the polypeptide chain, two regular secondary structures emerge, α-helix and the β-sheet, which were described by Pauling and Corey in 1951 and the rest of local structures without a distinct pattern are often referred to as a loop or turn (*5*). The α-helix is a repetitive and relatively compact local structure forming a regular pattern of hydrogen bonds involving the NH group of n^{th} residue with the CO group of $(n+4)^{th}$ residue. Due to the configuration of the L-amino acids, the α-helix adopts a right-hand direction with a single complete turn occurring every 3.6 residues. On the other hand, the β-sheet is stabilized by local and non-local hydrogen bonds between distant residues along the primary sequence and generally categorize into two distinct types, according to the relative orientation of the strands: parallel and antiparallel β-sheets. Loop and turn are less regular local structures that bridge regular secondary structures, usually having a high level of structural flexibility.

The spatial arrangement of local secondary structures leads to the formation of specific three-dimensional structures of the protein, which is referred as tertiary structure. Tertiary structure is stabilized by the balance between enthalpy and entropy mainly relying on the stabilization of non-local interactions like disulfide bridges between spatially dispersed cysteine; burial of hydrophobic residues into protein core; salt bridges and others. Eventually, the tertiary structure adopted in the native conformation of the protein under physiological conditions and enables its biological function. In the cellular

environment, proteins are rarely isolated and rather organized into even higher order structure by the assembly of other counterparts into quaternary structure that will ultimately perform biological function. Within quaternary structure, each of these protein chains is called as monomer (or subunit) and multimerize as multi-subunit complexes through non-covalent interactions or sometimes disulfide bridges, for instance hemoglobin is a classic example of quaternary structure consisting of 4 subunits: two pairs of α and β subunits (Fig. 1.1).

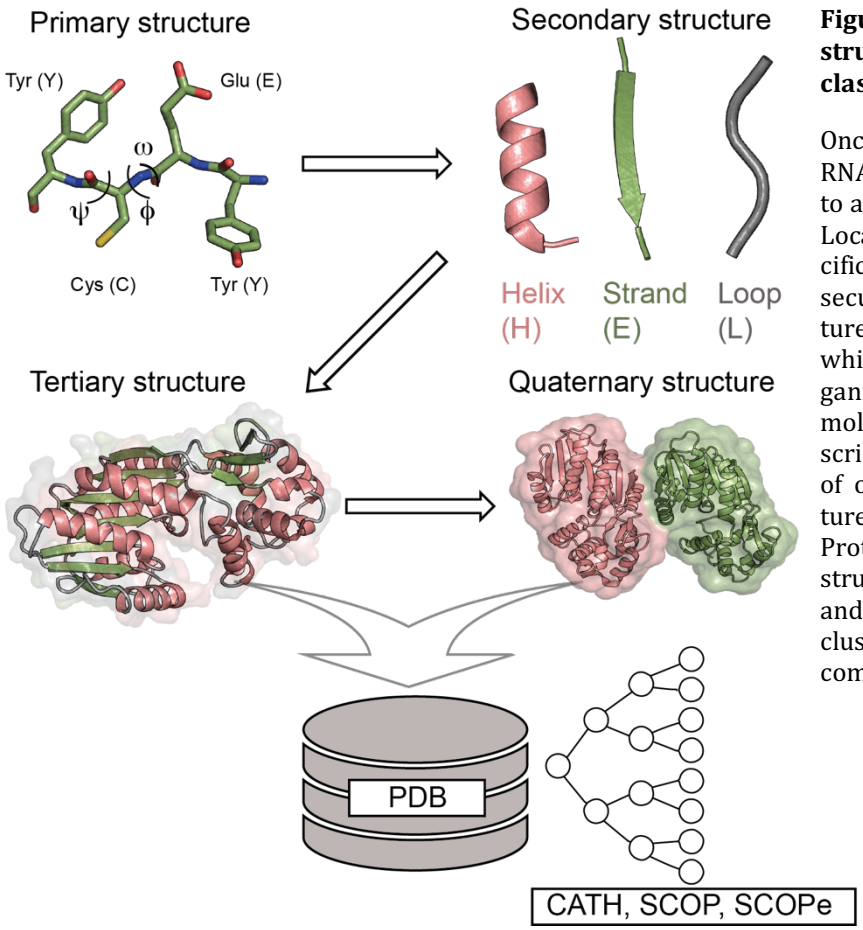


Figure 1.1. Orders of protein structure and protein structure classification databases

Once a protein is translated from the RNA transcript, the polypeptide starts to adopt the local secondary structure. Local structures account for the specific conformation of segments of consecutive amino acids. Tertiary structure is the three-dimensional structure which entails the complete spatial organization of the local structures of the molecule. Quaternary structure describes the organization of the subunits of oligomeric proteins. Protein structures are generally deposited into the Protein Data Bank (PDB). Based on the structural features of the protein, CATH and SCOP databases are used to further cluster protein structures according to common folds or topologies.

As briefly mentioned before, protein function is intrinsically encoded in three-dimensional structure of a protein; therefore, the precise understanding of protein structure at the atomic scale remains of the utmost importance for the broad scientific community in order to decipher the roles and functions of biomolecules. Advances in the field of structural biology including X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy provide the means to determine protein structures at the atomic level. The number of deposited structures in the Protein Data Bank (PDB) (*6*) has grown exponentially since the late 1990s (Fig. 1.2).

In the midst of the sequencing era, the gap between the protein sequence universe and known protein structures is tremendously enlarged, therefore, determining the structure of each sequence by experimental approaches are simply not feasible (*7*). To address this gap, computational approaches for protein structure modeling have become increasingly popular. Two computational approaches are commonly used to model protein structure from primary sequence.

The first approach is sequence-based methods which take a sequence and identify similar structures within solved structures and eventually build a three-dimension model of the query structure. These

approaches are usually referred as knowledge-based methods and two broad groups of methods are associated with this approach – comparative modeling and fold recognition (*8*).

- Knowledge-based methods

  - Comparative or homology modeling: These methods rely on the search for structures that share sequence "similarities" with the query sequences. The essence of comparative modeling is to build the three-dimensional structure of target protein from one or more structurally solved proteins with high sequence similarity. The modeling is generally done in three stages: 1) the selection of the most biologically relevant reference structure, 2) the alignment of the target protein to selected templates, 3) the construction of the actual model based on conserved regions. Two commonly used methods to deal with the sequence alignment between target and template sequence are PSI-BLAST or chains of Hidden Markov (HMM, Hidden Markov Models) and several popular algorithms performing comparative modeling by homology, like SWISS-MODEL, MODELLER, ROSETTA and others. A useful rule of thumb is that once the sequence similarity exceeds 50 %, modeled structure can achieve a reasonable RMSD of 2-3 Å compared to the template structure (*9, 10*).

  - Fold recognition modeling: fold recognition is frequently used when the homology modeling fails. This method relies on using an algorithm to identify the possible protein folds imprinted inside the predicted sequence and find a synonym out of estimated protein fold database (1,000-10,000 different folds). In general, this method is less sensitive to the sequence variability and leading to the failure of structure prediction (*11*).

A second type of approaches, however, seeks to build a three-dimensional protein structure directly from the primary sequence based on the physical principles that govern atomic interactions. This approach is usually referred as *ab initio* protein structure prediction, which will be described in a following section. Basically, *ab initio* methods construct physical models of the sequence and perform an exhaustive search of the conformational space to identify the lowest energy conformation, which will potentially be the native structure of protein.

1.2 *Ab initio* protein structure prediction

A successful *ab initio* prediction method based on the possible folding models to search the right path for the protein folding. Although the fundamental difference between folding problem and prediction problem is correlative, the protein prediction only focuses on the accuracy of final three-dimensional structure. The prevailing folding models are nucleation, hydrophobic collapse, framework model, and energy landscapes. A unified model centered on the energy landscape illustrates the folding trajectory as a globally funneled shape and any folding events can occur during the process toward the native conformation of protein. To generate the informative protein model on the basis of folding mechanism mentioned above, protein folding prediction is often referred as a multi-objective optimization problem and need to consider mainly three tasks in creating an *ab initio* model (*8*).

- Low-resolution representation of the protein structure.
  The energy landscape of protein folding has many local minima and maxima that preventing the general method such as gradient-based optimization methods to find the global minimum. A

common strategy for dealing with these difficulties is to reduce the degrees of freedom available in a conformational search algorithm. The conventional approaches using the lattice (amino acid is divided to polar and non-polar and simplified as bead) and off-lattice (limited degrees of freedom of side-chain and bond lengths) model to represent subtle geometry of amino acids. One of off-lattice model, coarse-grained models is broadly used to represent full-atom structures. A frequently used description represents explicitly the backbone atoms and the side-chains as spheres typically referred to as centroids. This simplified representation allows for significant computational speedups to accelerating the conformational search process, but often fail to capture the level of complexity needed to distinguish native from non-native conformations. Altogether, a low-resolution representation of protein structure allows performing extensive conformational searches in a manageable time frame (*9*).

- Energy function to model the interactions during the sampling process.
  The scoring function must compute efficiently and accurately the energy of each sampled conformation to guide the conformational search. Currently, the most common scoring functions lie between two extremes, with a mixture of physical force derived from classical mechanics and empirical potential derived from statistics of protein structures. Quantum mechanics (QM) accurately describes the physical principle including Newton, Coulomb interaction for each atom but it requires massive computational costs to even simulate a system with a small number of atoms. Another approach is to decompose the exact energy as an approximation of multiple energy potentials such as the statistical potential or pairwise energy potential (*12*). Since this type of energy function can be pre-computed, sampling the same pair of rotamer during the search process does not require recalculating. Therefore, selection of an appropriate energy function is dependent on the request for the structure prediction, and finding a correct balance between simulation time and accuracy is another part required to be optimized. But eventually optimal energy function is applicable to model a behavior already observed and is able to describe the physicochemical mechanisms of the behavior remain unexplored.

- Efficient energy optimization method to search an energy landscape
  The ultimate goal of the conformational searching is to identify a set of local minima including the Global Minimum Energy Conformation (GMEC) in the landscape. The algorithms used for sampling the conformational space also need a compromise between speed of execution and completeness of exploration. In addition, the choice of the method is strongly dependent on the representation of the conformational space and the energy function employed, for example some methods require discrete representation of the amino acids, and are restricted to approximated energy functions decomposable into a sum of energies of pairs.

  Broadly, search methods are classified into deterministic and stochastic methods. Deterministic method, also known as semi-exhaustive algorithms systematically converge on a unique solution for a given set of parameters. Two algorithms are predominantly used for deterministic methods: the average field and the Dead-End Elimination. On the other hand, the stochastic or semi-random algorithms randomly sample a large space of possible conformations for one cycle of optimization but without a guarantee to obtain a global minimum of the folding funnel in a single heuristic cycle. Thus, two independent simulations may generate different conformations located in different local minima (*8*).

Example of deterministic methods:

- The average field method substitutes all the pairwise interactions between rotamers to a single mean interaction. To compute a mean interaction, the Boltzmann distribution is used to weight the rotameric interaction with their respective probabilities.

- Dead-End Elimination is an algorithm that eliminates rotamers that are absent in the GMEC. For a given rotamer at residue i_1, if its energy contribution is always inferior by using an alternative i_2 in the same conformation, i_1 is eliminated. The iterative elimination continues until no dead-ending rotamers is found for each residue at a given conformation (*13*).

Example of stochastic methods:

- Monte Carlo-based algorithms (MC): The general MC process starts as following: First, a subset of rotamers is randomly selected for each residue and a rotamer that contributes to the enhancement of favorable energy state is repeatedly selected during heuristic cycles until the energy no longer is optimized. In order to cross an energy barrier to explore other local minima than searching inside a single one, rotamers increasing the total energy is accepted with a Boltzmann probability. Besides, simulated annealing metaheuristic (MCSA) usually used to couple with MC process by introducing the temperature variable into the simulation by heating and cooling steps provided by the Metropolis guidelines.

- Genetic Algorithm (GA): These algorithms simulate natural selection processes on protein structural searches by executing three evolutionary processes: mutation, selection, recombination. Briefly, GA first generates a population of P (parent generation) with a set of rotamers and P is continuously mutated according to a given probability distribution. A succeeded set of optimal rotamer is kept and randomized recombination for the next generation (P*). The energy in each generation is always compared with the random rotameric set S. The procedure is repeated until the entire population reach an equilibrium (*14*).

Recently, co-evolutionary information derived from homologous sequences provides evolutionary correlation between pairs of residues that preserved for functional fitness and implies the structural proximity of residue pair. By using co-evolutionary correlation together with other structural properties such as the solvent accessibility and secondary structural propensity of each residue, a searching space of *ab initio* prediction is constrained towards a favorable protein conformation. While in some cases the predictions have proven very accurate, structure prediction algorithms are in many cases not sufficiently accurate to predict a protein's 3D structure reliably and with high accuracy. Thus, experimental strategies to solve protein structures have thus far remained indispensable. By mutual integration of computational and experimental methods, it is possible to extend the determination of larger and more complex proteins that has been out of reach for any of the current approaches.

1.3 Protein structural space

The PDB has more than 150,000 protein structures deposited (*6*) (Fig. 1.2), and like the DNA gene banks, the PDB contains many homologous structures with very similar sequences. However, the theoretical protein space for the proteins with an average length of 200 amino-acids can generate 20^{200} sequences,

which largely outnumbers the estimated protein repertoire from all the known organism which is around 10^{10}-10^{12}. In fact, evolution only explores a small fraction of the protein sequence by incrementally accumulating mutations within the existent protein clusters rather than the emergence of full new protein clusters (*15*). Indeed, structural classification of the proteins determined experimentally so far, one can find that nature repetitively reused a subset of protein structures during the evolutionary mechanism. Because the selection pressure often selects the protein structure containing the function for an adaptation. As a result, different selected sequences can eventually adopt the same structure to execute same function; some sequence space probably hasn't been explored because the function encoded wasn't driven by natural selection. Indeed, nature leverages estimated 1,000 to 10,000 basic structural units (Fig. 1.2): fold or domain to perform the necessary functions (*11*). For instance, the enzymes, pyruvate decarboxylase (PDC) and benzoylformate decarboxylase (BFD), share a nearly identical protein fold but with only 21 % sequence identity, indicating that the protein fold is more evolutionarily conserved than protein sequence.

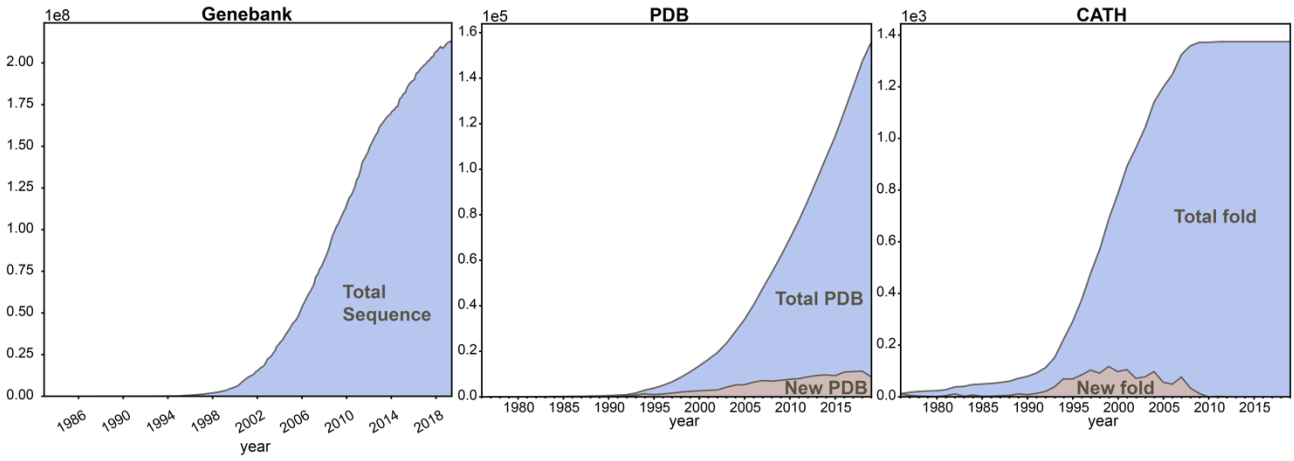


Figure 1.2. Growth of protein sequence, solved structure, and clusters of folds over time.

The data were extracted from GenBank (total sequence of all organisms), the PDB, and CATH databases, respectively. The area colored in light blue represents the total amount of data deposited in each database, and the brown region shows the number of novel structures or folds.

Folds are defined by the arrangement of secondary structures in space and represent the basic building unit in protein structure (*16*). To systematically organize the existing folds in nature, protein structural classification databases have categorized structural trends in proteins and evolutionary relationships. SCOP2 (Structural Classification Of Proteins) (*17*) and CATH (Class, Architecture, Topology, Homologous superfamily) (*18*) are the two classifications most widely used and both databases organized the folds in a hierarchical order to highlight the phylogenetic relationships between proteins. SCOP2 classifies folds into four hierarchical levels. 1) Protein types: proteins are sorted into four types, soluble, membrane, fibrous and intrinsically disordered according to the sequence and structural features of protein. 2) Evolutionary events: this category is used to reorder the annotation of various structural rearrangements due to the recombination of protein fragments. These two levels are innovative classifications originally not being used in SCOP. 3) Structural classes: Classes bring together folds with a similar composition in secondary structures of protein. Four main classes are classically identified: all-alpha (consisting mostly of α-helix), all-beta (majority β-structure), mixed α and β (denoted as α/β and α+β, consisting of different compositions of α- and β-structure) and small protein (the protein with a small portion of secondary structures). 4) Protein relationships: this category is composed of three subcategories-

Structural, evolutionary and other relationships. Evolutionary relationships describe the evolutionary levels of protein, it is defined by the sequence identity of protein into four levels: superfamily, family, protein and species. Structural relationships, on the other hand, strictly categorizes the protein on the basis of global structural features of protein domains (the composition of secondary structures in the domain, and architecture of domains) to avoid the sequence homologies were classified into the same fold.

CATH organizes the domains according to their similarities of structure and sequence. Five levels are defined 1) Class: the class characterizes composition in secondary structures and their arrangement, 2) Architecture: the architecture describes more precisely the arrangement of the secondary structures without taking connectivity into account, 3) Topology: the topology or fold categories the general structure of the domain and connectivity between secondary structure elements, 4) Homologous superfamily: the category classifies the domains sharing a common ancestor or sequence similarity, finally, 5) Family: the last category groups domains with significant sequence similarities.

Based on the Darwinian theory, natural evolution has fine-tuned all the properties of proteins for millions of years and reshapes the protein repertoire in a direction to fulfill the current biological need (*19*). Although nature has developed countless biomacromolecules, as described above, to perform diverse functional tasks for sustaining living organisms, the currently existing proteins may not be the best scaffold for specific applications in the following age. To explore the dark matter of the protein universe, two approaches have been used to expand the protein sequence space: 1) Directed evolution, and 2) Computational protein design (*7*).

Directed evolution

Directed evolution is a widely used experimental strategy that simulates the natural process of evolution to select protein variants with certain phenotypes in a short timescale. The main objective of this strategy is to generate and identify protein variants with the desired characteristics and the population is artificially shaped by purifying selection towards mutants with the properties of interest. This approach requires two essential steps: I) construction of DNA libraries composed by diverse genotypes; II) selection of proteins encoded a chosen phenotype (*20*). The rapid development of DNA synthesis over the past two decades has led to an increase of possibilities for the generation of DNA libraries with different levels of control over the location and identity of the inserted mutations. This DNA synthesis advances confer important advantages over the error-prone polymerase chain reaction (epPCR) generated libraries which produce random mutations in DNA. Where one can have a much more precise control over composition and avoid combinatorial explosion that will impede screening the full library. Second, to screen the library of sequences, there are several systems to present/express the protein variants *in vitro* for functional screening. For example, yeast and phage display methods (*21*) were widely used to incorporate the DNA libraries encoding variant libraries and display them on the surface of yeast cells or bacteriophages for the functional screening.

The applications of directed evolution are limited to protein engineering problems in which the desired phenotype can be easily screened. Still, many successes resulting from screening diverse libraries demonstrate the power of natural evolutionary processes in the optimization of binding affinity for a target molecule, the catalytic activity and stability of natural proteins (*20*). Recent advances in implementing machine-learning guided library construction will enlarge the scope of directed evolution in creating a new function or improving one or more pre-existing properties of the proteins (*22*).

Computational protein design (CPD): protein redesign and de novo protein design

Computational protein design was first proposed as the inverse protein folding problem, in which to find the compatible sequences to fold into a given protein conformation (*23*). Basically, algorithms of computational protein design perform two main tasks similar to those described for protein structure prediction: 1) extensive sampling of the conformational space of conformational degrees of freedom for both backbone and side-chains; 2) scoring and evaluating the energy of different sequence combinations for a given conformation. Unlike protein structure prediction, where the query sequence and the search are centered in exploring the conformational space available to a particular sequence; whereas, protein design requires to sample the energy landscape of numerous sequences, which increased the time necessary to sample and score many of the possible sequence-structure combinations (Fig. 1.3).

Depending on the degrees of freedom of the polypeptide chain, protein design approaches can be split into two major categories: protein redesign and *de novo* protein design. Generally, protein redesign uses an existing protein structure as a starting template and improve a certain property of template structure by changing a subset of residues. Thus, the overall protein structure is generally conserved after redesign. The conformational search is often restricted keeping the backbone fixed and searching for amino-acids and side chain conformations of native residues to form the optimal interactions with introduced mutations. The pioneering work in protein redesign focuses on the design of protein cores to improve its stability and develop powerful biosynthetic catalysts that can operate under difficult conditions (*24*). Dantas *et al*. redesigned the hydrophobic core of nine globular proteins and demonstrated that thermostability of several redesigned proteins was increased relative to the native proteins (*25*). Redesign approaches have also been applied to the engineering the tasks related to molecular recognition and biological functions, for example adopting natural protein interface on existing protein scaffolds for metal binding, therapeutic markers binding and others (*26-29*). And other successful attempts tried to engineer the catalytic mechanism of enzyme (*30, 31*). These successes in protein redesign involved grafting of protein segments into existing protein structures; whereas, an alternative approach is to customized design of *de novo* backbones to accommodate interface patterns.

Due to the topological constraints imposed by the architecture of existing structures, the exploration of the conformational space is limited by the initially chosen scaffold. To expand beyond natural protein folds, *de novo* protein design was proposed to generate novel proteins from scratch based on the first principles and searching for new sequences to fold into the defined protein conformation. As discussed in the previous section, nature only samples the structural space that benefits fitness at the level of organisms; thus it is unlikely that nature explores more diversity than it is necessary to sustain life. Therefore, *de novo* protein design aims to explore the "protein universe" to uncover protein structures only based on their feasibility of physical principle without a drive for fitness optimization (*7*). One frequently used approach in CPD is utilizing custom libraries of fragments derived from known structures to assemble the topology of target structure (*32*). By doing so, the sequence-dependent local interactions inherited from natural fragments pre-define the local conformations; thus the search is reduced to sampling nonlocal interactions.

Efforts in *de novo* protein design began in 1979 by Gutte *et al*., who started an attempt to design a protein using a simple rule derived from secondary structural preferences of different amino acids and applied this rule to model a continuous ββα motif of ribonuclease A (*33*). After this initial success, in 1981 Drexler stated that to predict the conformation of all the natural proteins may be challenging, but

exploring one sequence to satisfy a given conformation of protein is a practical task (*34*). Later in 1983, Carl Pabo first described that protein design was the inverse folding problem, where we search the set of primary sequences that adopt a targeted protein conformation (*23*). The following efforts used the empirical principles observed from the structural patterns of natural SSEs to design α-helical or β-sheet, like the attempts done by Lau *et al*., and Moser *et al*. (*35, 36*). In another important breakthrough in the late 1980s, DeGrado *et al*. used a graphical representation to pattern a coiled-coil tetramer and rationally specified the defined amino acids for each position in the helical bundle (*37-39*). Though many attempts in designing the *de novo* proteins from 1980 to 1990, none of those designed proteins were crystallized to reveal the design's accuracy at the atomic level. A landmark achievement in 1997, Dahiyat and Mayo demonstrated the first example of computationally designed sequence folding into a target zinc finger structure (named as pda8d) that matched with NMR solution structure to a resolution less than RMSD of 1.4 Å (*40*). Following with this success, more attention was put on developing the energy potential to decompose the energy into several scoring terms. The first crystal structure confirmed design came out shortly after the attempt of Harbury *et al*. (*41*) that using a flexible backbone design incorporated with Crick's parametric equation to design a super-helical coil (dimer, trimer and tetramer) and the crystal structure matched perfectly with the design at a remarkable RMSD of 0.2 Å . The next most noticeable success in protein design involved in creating a novel protein fold Top7 by Kuhlman and Baker (*42*). This 93-residue protein adopted a novel globular α/β fold that was never reported before. Due to the great improvements in computational power and structural biology during that period, the protein design community started to incorporate fundamental rules and structural features learned from the protein folding mechanisms and known protein structures. Comprehensive studies were presented by Koga *et al*. and Lin *et al*. to describe general rules for designing ideal structures and building the correct loop segments linking SSEs in a geometrically correct manner (*43, 44*). Based on these principles and strategies, a broad variety of *de novo* proteins were designed, for example tunable helical coiled coils (*45, 46*), β-structure with bulged strands encoded (*47*), TIM barrels (*48*), jelly-roll fold (*49*) and others (*50-52*).

Protein design has matured to a level where it can make an impact on many aspects of biological applications. However, most of the design successes mainly focused on creating thermostable proteins, rather than bringing back intrinsic essence of natural proteins - biological function. In order to consistently design functional proteins, efforts in the community should focus on improved representations of models, especially in the following areas: 1) evaluating protein scaffolds for specific design goals (designability assessment); 2) handling protein flexibility to portray the dynamics of protein structure; 3) developing energy functions that allow rapid high-throughput to explore the large sequence and conformational spaces, but still precise enough to correctly model the properties of the protein and distinguish the best candidate sequences; 4) developing objective energy functions to favor the physicochemical properties required for biological functions during the design process.

1.4 The Rosetta protein modeling suite

There are several programs for rational computational protein design, such as Rosetta (*15*), OSPREY (*53*), or Orbit (*54*). The Rosetta software traces its origins to David Baker's work in structure prediction (*55*), but has grown to become the dominant protein design program developed collaboratively by more than 40 independent research groups. Beyond protein design, Rosetta performs a multitude of structural bioinformatics tasks such as structure prediction, small molecule docking, or binding prediction (*56*).

Essential to Rosetta are its underlying high-resolution and low-resolution modes, and the energy functions used in each of these modes. Specifically, in Rosetta's folding simulations (which as we will see are essential for protein design), two different levels of atomic resolution are applied in the optimization phase. A low-resolution energy function is used in coarse-grain exploration of the conformational landscape to rapidly search the position of side chains, represented not with a full-atom description, but as summarized *centroids*. This low-resolution mode allows Rosetta to quickly explore structural conformations and it is biased to favor a reduced subset of force field potentials, such as solvation energy, electrostatics, a measure of compactness (derived from a conformation's radius of gyration) and a measure of steric hindrance between centroids. While Rosetta's low-resolution mode is useful for fast exploration of the conformational space, a full-atom high-resolution mode is used on optimized conformations to accurately score each one. The high-resolution energy function implemented in Rosetta is composed of around 20 terms and decomposed as pairwise interactions, meaning that the total energy of a given conformational state can be represented by the sum of the energies of atom/residue pairs (*57*). Thus, the total energy is a linear summation of each energy term multiplied with different weights to represent the enthalpic potential of a modeled conformation. The energy terms in the full-atom energy function include Lennard-Jones potential to approximate the Van der Waals force, the Lazaridus-Karplus implicit solvation term, the Ramachandran and rotameric term empirically derived from databases to evaluate the preference of both backbone and side chain for certain torsional constraints, the hydrogen bonding term, and the coulombic electrostatic potential. Improving Rosetta's energy function is an active field of research focused on both accurate and faster calculations (*57, 58*).

Rosetta model proteins in the following way. First, to represent the backbone conformation of proteins, Rosetta uses a torsional space representation: conformations are specified as a list of torsional angles (Φ, Ψ, ω). In parallel, to evaluate the energy of each conformation it is also necessary to obtain a Cartesian coordinate system, which is generated for all heavy atoms in the protein backbone. Finally, to represent amino acid side-chains in proteins, Rosetta uses two different representations according to the low-resolution or high-resolution mode: centroid description, where the side chain is represented by its center of mass, or full-atom side chain description, where a side chain is modeled in discrete rotameric conformations (*59*) (Fig. 1.3).

Rosetta performs three tasks that are critical for *de novo* topology design. The first task is amino acid sequence design (*42, 60*), also known as the protein design problem, in which given a protein backbone conformation, Rosetta searches an optimized sequence for the input backbone. Second, Rosetta performs local backbone perturbations on the input conformation to expand the number of sequences that may adopt the target fold. Under this approach, random angle perturbation perturbs selected backbone angles (phi, psi) to modify the dihedral angles along the entire protein backbone, and usually the random angle perturbation is coupled with gradient descent method to gradually sample the conformation towards local minimum.

The third Rosetta task critical for *de novo* topology design is *ab initio* protein structure prediction, a method originally developed for the protein folding problem. In the context of protein design, *ab initio* can perform larger conformational sampling of backbones that adopt the target topology, but differ significantly from the original backbone structure. At the beginning of the simulation, the starting point is a fully unfolded protein chain and usually a large number of independent simulations are performed in parallel. *Ab initio* prediction takes place in several stages: 1) Fragments of three- and nine-residues that can describe the conformational space of a defined topology are selected (*32*). The fragments are extracted from the library, which was built from high resolution structures in the Protein Data Bank. The subset of fragments used in the simulation is chosen according to both sequence and local structural

similarity of the target topology against the database. 2) nine-residue fragments are then assembled using a Monte Carlo method combined with a simulated annealing algorithm; At each step of simulated annealing, a semi-randomly selected nine-residue fragment is inserted to replace the torsional angles of one selected window. Once the insertion causes the decrease of energy, this move is accepted and the fragment assembly continuously propagates to next window in the query sequence. 3) structures are then refinement according to three-residue fragment: a refinement of each decoy is performed by inserting the predicted 3-residue fragments to replace the local conformation with the same procedure done for the insertion of 9-residue fragments. 4) the centroid, low-resolution representation of side-chains is replaced with a set of discrete rotamer for each residue position, 5) random backbone perturbation operated by Monte Carlo-plus-minimization strategy to accurately sample the local residue and side-chain conformation, and the last 6) evaluation of the final decoy compared with the target structure. Since the *ab initio* process is stochastic, repeated runs are needed to converge the simulated conformation. Therefore, at the end of this process, all the simulated decoys are usually clustered based on RMSD value to differentiate the structural diversity (*61*).

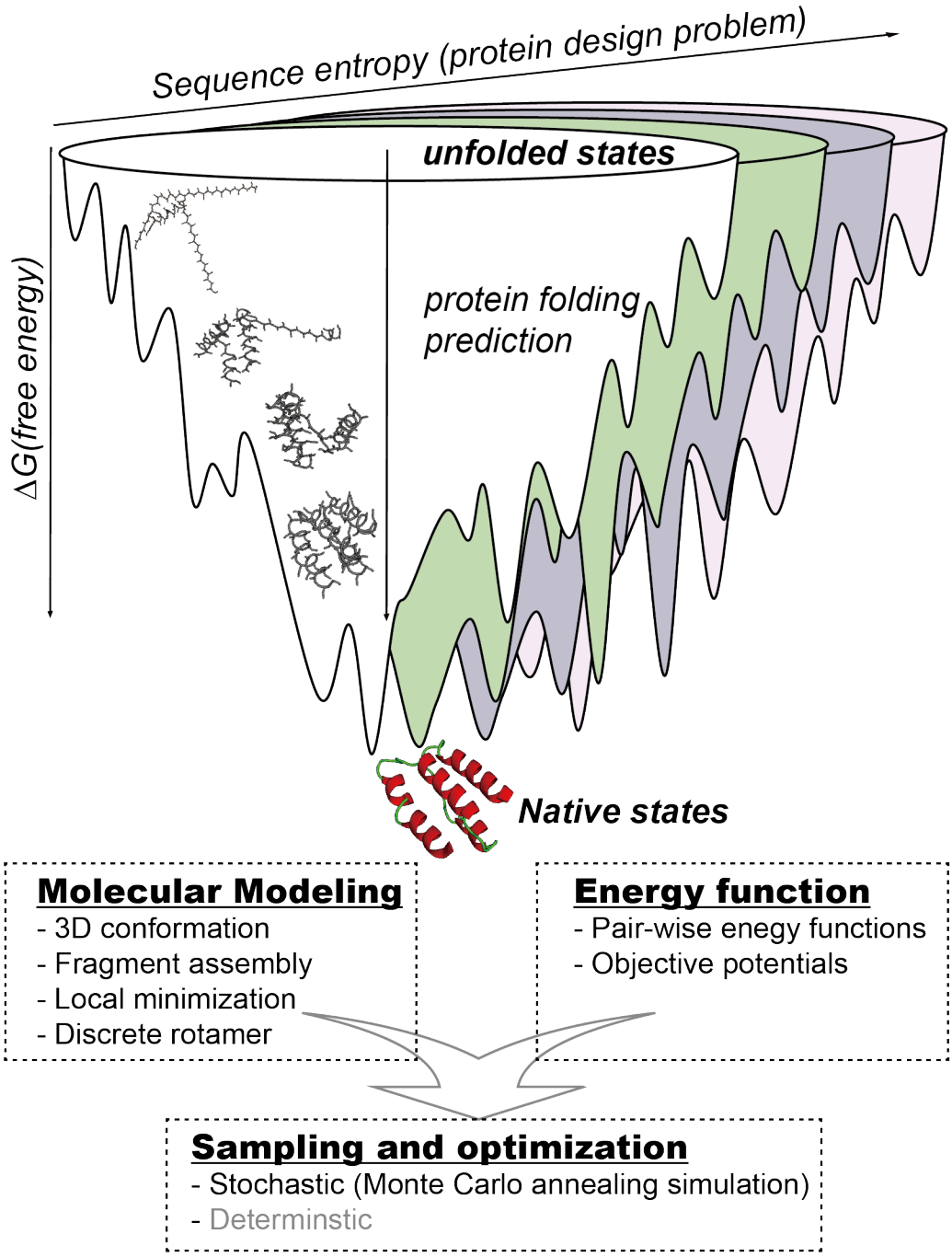


Figure 1.3 A generic process for protein design

The folding pathway is represented by the folding funnel. For a given sequence, the protein folding prediction is used to search an energy landscape (shown in white), in order to find a global energy minimum which is represented as a native conformation. The protein design aims to search through the sequence space (shown in green, purple and pink) to find the optimal sequences folding into a given native structure. This figure shows the general protein design process, using the Rosetta algorithm as an example. A given 3D structure first defines the search space. Energy functions are then used to score each sampling conformation. In Rosetta, the energy function is decomposable into pairwise energy potentials, and comprises two types of energy function, applied in two different levels of the searching process: coarse-grained and full atoms. The conformational searching is performed by fragment assembly at a low-resolution scale to quickly sample the shape of the protein. This is followed by local backbone perturbations (local minimization) to refine the full-atom structure. The scheme of the optimization in Rosetta uses stochastic sampling by Monte Carlo annealing simulation to perform a GMEC search. The protein structure demonstrated in the figure is the designed protein 5cwj39 used in this thesis.

1.5 Functional protein design

A fundamental feature of protein function is their ability to specifically recognize other biomolecules (e.g., lipid, protein, nucleic acids, sugar, ions, natural chemicals, etc.), because subsequent functions (such as catalysis) depend on first establishing an interaction (*62*). These interactions occur through regions in the protein surface called binding sites, in which amino acids are organized in a specific network to form non-covalent interactions with their ligand. In natural proteins, the binding site is also referred to as the *first shell* of the protein and it is highly susceptible to mutations (*63, 64*): changing the amino acid type of these residues often results in lost function. In general, the physical interactions occurring through the first shell trigger a conformational change in the protein to enable the subsequent action. For example, the structure of an enzyme can undergo a conformational change upon binding to the substrate (to a transition state), which consequently enable catalysis. In other contexts, biomolecule recognition itself is sufficient to operate biological function, such as when a natural antagonist pairs with a receptor, or antibodies bind to antigens and block, for example, binding of a viral particle to its receptor. While protein function is intimately linked to its three-dimensional shape, a modification of the 3D structure, even if minimal, can result in a decrease or loss of protein activity (*65, 66*). Therefore, accurately predicting the sequence and structural features of a protein binding site is a critical step required to design functional proteins.

Most of the work performed to date in the field of *de novo* protein design has focused on engineering highly thermostable proteins (*42-44, 48, 67, 68*). These successes show that current empirical models used in computational protein design can generate new proteins from scratch. However, the majority of *de novo* proteins thus far have lacked any biological function. State-of-the-art *de novo* design protocols first define a target protein topology, and then compute an optimized sequence (according to the protein design energy function) for this topology (*43, 69*). These protocols, however, are unaware of the function that will be conferred on the designed protein: the target topology explicitly defined by the designer may not be the optimal topology for a desired function. Therefore, functional design of proteins using existing methods is limited to redesigning existing protein backbones, which may or may not be able to accommodate the desired function (*70*). Despite several successes in design of functional protein, the procedure still requires iterative experimental optimizations and selection to obtain the final functionally compatible designs (*71*), and the experimentally optimized structures often deviate from the originally intended designs. So far, creating proteins with novel molecular or even biological functions from scratch is far from a solved problem.

There are a few exceptions in the field of functional protein design which are worth discussing. These include the design of a *de novo* catalytic triad for a retro-aldol reaction (*30*), for Diels-Alder reaction, for the Kemp elimination reaction (*31*), and others (*72, 73*). In all these cases, the binding module was inserted onto a pre-existing protein scaffold to perform catalysis. However, those successes relied on transplanting a binding motif to a pre-existing scaffold already arranged with a geometry nearly identical to the required one. This technique is known as *functional site transplantation* (Fig. 1.4). Functional site transplantation consists of extracting (or designing from scratch) the binding site or functional motif out from its natural context and transplanting the site into another protein topology. Designing the novel functional protein with this route, dense clusters of residues (usually referred as hot spots) (*74*) inside the binding motif are geometrically tight and provide an important amount of the binding energy to ensure the optimal interaction to its binding counterpart (*75*). Moreover, thanks to this technique, the entire rigid-body orientation of the protein complex is guided by the functional motif to retain the physicochemical property and shape complementarity required for binding, avoiding a systematic exploration of the docked protein conformation space. Once a functional site has been transplanted, the

design exercise consists of optimizing the amino acid identity of the residues surrounding the functional site for binding affinity to the ligand and for stability of the transplanted site. Functional site transplantation is, by far, the most successful technique in *de novo* design of functional proteins.

One specific strategy to perform functional site transplantation is motif grafting (*76*) (Fig. 1.4), which consist of three steps: 1) construction of a library of protein structures on which the functional site will be "engrafted". The ideal database will be as large as possible, but processed to remove homologous structures. 2) Extraction of the binding motif from the native context and search within the pre-generated database for proteins containing the structurally similar region to the isolated motif. 3) Adjustment of local conformation to allow the proper stabilization of the grafted motif and sequence design of surrounding residues for affinity to the target and stability of the engrafted motif. If the isolated motif is well matched to the putative protein, only the side-chains of the motif need to be transplanted, which is referred as "side chain grafting". In contrast, "backbone-grafting" is applied to reposition the entire conformation of motif into the query region of the scaffold and this procedure usually requires careful inspection to filter the decoy with suboptimal insertion which may impair the integrity of protein structure. Motif grafting has been used to design large numbers of novel functional proteins (*26, 29, 77-84*), ranging from ligand binders, protein inhibitors to the *in vivo* therapeutic reagent as immunogen, antiviral protein, neutralizers for chemical toxin and others.

While side chain grafting has been shown to be highly successful in designing functional proteins, cases of successful backbone grafting are much rarer. In this thesis, I hypothesize that this is due to backbone grafting requires the search of a much broader backbone conformational space of proteins. I propose that by improving the sampling of backbone conformations, one can better adapt pre-existing protein topologies to accommodate desired functions. Computational protein design typically relies on a constrained (reduced) subset of possible backbone conformations approximations to constrain the astronomical DOFs a protein backbone contains. This reduces the simplifies the problem enough so that it is tractable to solve with modern computers. However, conformational constraints limit the natural freedom of the protein backbone, thus potentially restricting the search for sequences needed for protein functionality. In fact, many experiments have shown that rigid backbone conformations fail to adapt to some residues or side-chain conformations relevant to binding and therefore a large portion of the sequence space will be discarded during the design process; however, minor adjustments of the backbone angles can help resolve these issues (*85*). In addition, conformational changes in the protein backbone are also critical for a protein to recognize other molecules (*86*). Therefore, the adjustability of protein backbone must be integrated in the design process to improve the accuracy of CPD predictions and extend the use of CPD methods to different design goals especially the design of proteins with functionality.

Despite the difficulty and complexity to integrate protein backbone flexibility in the design process, several attempts have tackled backbone flexibility since the early days of computational protein design. Mayo and his colleagues (*40*) used the characteristics of secondary structural elements to encode backbone flexibility into the design of a helical tetramer, which is only practical if the given backbone parametric descriptions are available for a topology of interest. Later, Desjarlais and Handel (*87*) used random dihedral values derived from the PDB to replace the torsional angles of their designed protein, and uncovered an effective mutation not found by a fixed backbone approach. After those pioneering works, several studies started to systematically implant the tolerance to sample the backbone flexibility during the simulation and proven that allowing the perturbations on the protein backbone, even the subtle variation applied on backbone conformation (less than 2 Å), is capable of enlarging the sequence space (*88-94*). Protein structure is believed to be robust to mutations (*95*); therefore, improvements in

refining protein conformation by local backbone perturbation allow protein design programs to recapitulate naturally occurring sequences.

In practice, successful grafting of functional motifs is applicable to a very small fraction of them, typically linear helical segments (*28, 78, 81, 96*), with only few cases showing the success to graft irregular motifs (*80*). Paradoxically, most functional sites found in proteins, such as binding and catalysis, are usually composed of one or multiple non-ideal elements that are arranged in a tertiary structural space (i.e., multiple discontinuous segments) and supported by a defined topological relationship to maintain a structural feature (e.g., irregular helix, bulged strand, cavity inside the protein core). Therefore, grafting such complex functional sites to an existing protein with close local structural similarity for every discontinuous segment of motifs becomes infeasible, and the difficulty is scaled with the complexity and number of motif segments. Moreover, the backbone of the chosen scaffold on which to graft the motif is rigidly constrained to maintain a global structure and preserve the protein stability inherited from the parental scaffold. Extensively sampling the backbone conformation space using existing approaches may fail to find a suitable conformation to tolerate the trade-off between non-ideal local structure and stability (*97, 98*).

Previously, Correia *et al*. (*99*) demonstrated an alternative method Rosetta Fold From Loops (FFL) to expand the applicability of protein grafting by overcoming the dependency of local structural similarity required for motif transfer (Fig. 1.4). The FFL protocol leverages the advantage of Rosetta's fragment assembly machinery, which can sample a large conformational landscape within a given topology, to fold a protein scaffold together with the motif of interest from the scratch. Correia *et al*. used this folding-based grafting approach to design an immunogen that presents one neutralization epitope of respiratory syncytial virus (RSV), and the crystal structure of designed protein showed the grafted motif is highly mimic as native conformation. Moreover, immunization of this computationally designed immunogen elicited the functionally relevant antibodies in the non-human primate model, which is the first study to reveal the concept of computational epitope scaffolding for vaccine development. Recently, we advanced this algorithm named Rosetta Functional Folding and Design (FunFolDes) to incorporate additional folding constraints by directing the folding process in the presence of binding target to enrich the conformer presenting the correct rigid body orientation can bind the targets (the detailed process will be explained in chapter 2).

Above mentioned strategy still, though sampling different freedom of backbone flexibilities, are still highly dependent on using pre-existing structure, either natural or *de novo* protein, as starting scaffold and repurposing the functional site on it. Since the chosen scaffold defines topological constraints of designed protein, in which impose a structurally searching limit to accurately stabilize the functional motif or even misplace functional residues at their suboptimal position and further impairing the designed function. Thus, giving the protein topology is predefined, modifying those structures beyond their topological definition seems unrealistic and may undermine the protein stability.

Contrary to the previous two strategies, both termed as "motif-driven design," recent studies have shown the success in incorporating the functional sites at the beginning of the design stage and tailor the *de novo* protein backbone to stabilize the motif. This alternative approach, referred to as "motif-centric design," is not generalizable in the field yet (Fig. 1.4); however, the central principle is to define the virtual protein topology centering around the functional motif and designing *de novo* segments, which can be any secondary structural types, to fit the context of motif. Conceptually, this approach consists of building the custom-made structure for the given motif, aiming to design the structure-to-functionality relationship in one design step, rather than post-incorporation of the functional residues

into a stable protein scaffold. Examples of successful attempts for this task include a *de novo* four-helix bundle in a favorable orientation to bind the porphyrin cofactor (*100*), and stabilize the catalytic triad (*101*). In addition, a recent study (*51*) reported that introducing the binding residues for fluorophore at the early design strategy led to design a *de novo* β-barrel protein, holding a specific conformation to support the functional module. Finally, a recent study conducted by Silva and colleagues (*102*) designed a *de novo* interleukin 2 (IL-2) mimetic, which preserved the substantial interface prerequisite for engaging the binding, and specifically manipulated the IL-2 signaling pathway to eradicate the tumor cell *in vivo*. Although, these "motif-centric design" strategies were successful, more evidence is required to prove its applicability in the future. These studies are an example that designing functional protein through this route can lead to biologically functional proteins with superior thermostability.

Overall under the trend in the field to computationally design functional proteins, further development is needed to strike a balance between physicochemical feature or characteristics and functions. A conservative approach like grafting inherits structural constraints on the initial scaffold, and disfavors the transplantation of a structurally distinct motif. At the same time, a conservative approach almost guarantees the fundamental stability and foldability of protein. On the other extreme, to design a functional protein from scratch with a motif-centric approach such as the one described, the tailor-made topology goes beyond the known structures to freely sample the topological space needed to stabilize the functional motif and have full control over physicochemical and structural features. However, due to inaccuracies in design methods, these *de novo* designed topologies and sequences likely need to be screened by experimental assays to, eventually, generate the folded and functional protein.

In my thesis, we take these approaches further to globally sample the space of protein topologies with a fine control of the spatial relationship between *de novo* backbone and functional motif and directing the protein folding under the functional constraint toward searching the relevant conformation. We employ both FunFolDes and motif-centric design approaches to functionally transplant a viral epitope in a *de novo* protein. We show that the computationally designed protein can elicit protective antibodies in animal models against the original virus. In addition, we use motif-centric design approaches to build novel protein segments around two binding motifs, performing a double epitope transplantation to a *de novo* protein, the first of its kind, to our knowledge. We show its ability to bivalently bind to their targets in an engineered cell system to regulate the cellular activity, and we show that, potentially, it can be used to program the synthetic cell for therapeutic applications. In the following section, the challenge for the vaccine design field is introduced, which the protein design work of my thesis aims to address.

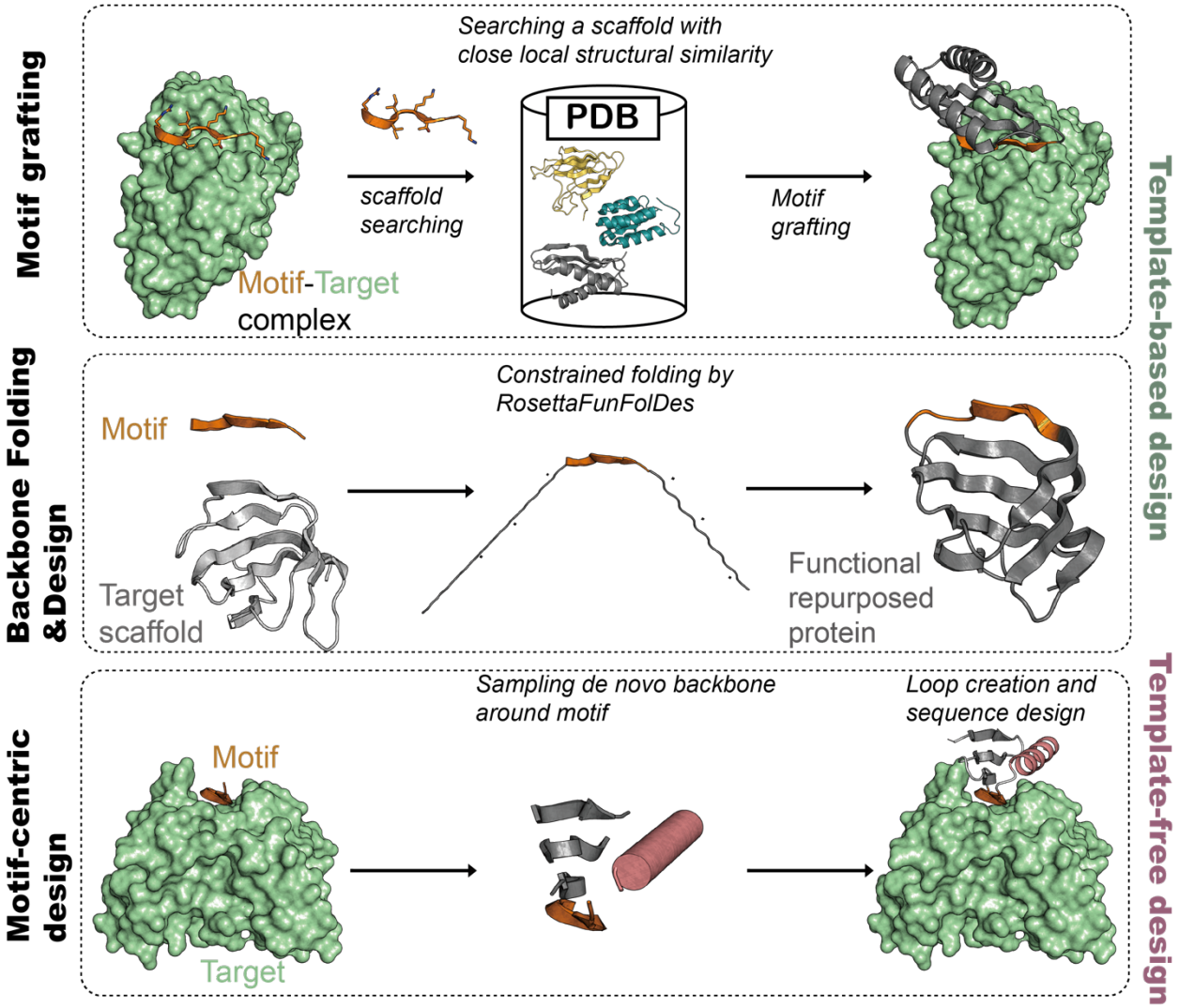


Figure 1.4. Common strategies for transplantation of functional motif.

Based on the template dependency, the current strategy to repurpose the functional motif into the protein scaffold can be categorized into two approaches: **template-based** and **template-free** approach. 1) Motif-grafting: A functional motif is transplanted onto a pre-existing protein structure that shows local structural similarity to the motif of interest. 2) Backbone folding and design: A chosen topology for a given functional motif folds from the extended polypeptide chain to sample a larger backbone flexibility to accommodate the chosen motif structure. 3) Motif-centric design: A *de novo* protein backbone is custom-made for a motif of interest by centering all the design steps around the motif. Chapters 3 and 4 describe each approach in more detail.

1.6 Reverse vaccinology

The primary correlate of protection of successful vaccines is the induction of neutralizing antibodies (*103-105*). Neutralizing antibodies target important functional sites of viral proteins, thus inhibiting the cellular infection or promoting viral clearance through Fc mediated effector functions (*106*). However, conventional vaccines based on attenuated or inactivated virus-based formulations, have failed to elicit neutralizing antibodies against several pathogens. Examples include human immunodeficiency virus 1 (HIV-1), various paramyxoviruses such as respiratory syncytial virus (RSV) or human metapneumovirus (hMPV), as well as a protective vaccine against multiple influenza strains. The main obstacle for efficacious vaccine development is the capability of some pathogens to shield their vulnerable

regions (*107*), resulting in the induction of non-neutralizing antibodies targeting irrelevant sites, which are thus unable to eradicate the pathogen. An illustrative example is influenza, where the majority of antibodies elicited during natural infection or seasonal vaccination target against the mutable head domain, whereas antibodies targeting the sequence-conserved stem region are subdominant (*108, 109*).

To overcome this hurdle, reverse vaccinology 2.0 has been proposed to rationally design immunogens that direct the immune response towards to the vulnerable sites of pathogen (*110*). This renewed reverse vaccinology emerged as a structure-based approach to develop vaccines by leveraging structural information of neutralization epitopes in complex with their target antibody to design novel immunogens that re-elicit an epitope-focused antibody response (Fig. 1.5). Thus, the approach starts with isolation of neutralizing antibodies from protected individuals, and structural characterization of such antibodies in complex with their targeted antigen is used to inform design of novel immunogens to elicit protective antibody responses. The last step is to spotlight the essential epitopes of antigen via protein engineering or rational protein design to stabilize the native conformation of epitope structure, and eventually presents designed immunogen back to the immune system.

During the last decade, numerous neutralizing monoclonal antibodies have been isolated and structurally characterized in complex with viral proteins for pathogens such as HIV (*111-113*), RSV (*114-119*), and influenza (*120-126*). This period has coincided with an advance in other nobly high-throughput technology, including a single cell sequence and isolation of huge panels of antibody sequence targeting virus to potentiate vaccine design (*127*). Together, these efforts have yielded templates for the structure-based design of novel immunogens, aiming to re-elicit defined antibody responses *in vivo*.

So far, there are several proof-of-concepts using this approach to elicit neutralizing antibodies by an engineered immunogen targeting to the defined immunity (*128*). They can be categorized into four different structure-based design approaches: 1) Masking immunodominant sites, 2) conformational stabilization of viral protein, 3) germline targeting, and 4) epitope scaffolding. Each strategy has been reviewed by Sesterhenn *et al*., and is briefly summarized in the following section (*129*).

Masking immunodominant sites

To avoid antibody responses against unwanted epitopes, a simple but elegant strategy is to mask those epitopes through: 1) deletion and mutation to shield the immunodominant epitopes (*130-133*) or 2) masking through hyper-glycosylation (*109, 134, 135*). Several attempts for influenza have shown that vaccination with the stem region of the hemagglutinin mounted a broadly cross-reactive antibody response to protect the immunized animals, indicating that antibodies elicited by HA stem can confer protection against the diverse group of influenza strains, which is rarely achieved by using the full construct of HA protein. A similar outcome was also achieved by using a HA variant with an hyperglycosylated head region of HA, focusing the antibody response toward the stem region. These strategies also led the successes in silencing the immunodominant site for different pathogens such as RSV and HIV, and further demonstrating that artificial introduction of glycosylation, mutation, and deletion on the dominant region potentially detoured the immune response toward subdominant epitopes.

Conformational stabilization of viral protein

Enveloped viruses, such as HIV-1, Influenza, Respiratory Syncytial Virus (RSV), Metapneumovirus (MPV) and others infect host cells through fusion of the host cellular and viral membrane (*136*). This infection process was specifically mediated by class I viral fusion proteins, in which these structurally complex proteins exist in a metastable prefusion conformation, and undergo structural conversion into a stable postfusion conformation during the infection (*137*). Many studies have proved that blocking this structural arrangement of fusion protein ultimately prohibits the viral infection from entering the tissue cell. Thus, eliciting antibodies that target prefusion-specific epitopes has become an important goal for vaccine development, particularly for RSV, but also for MPV and others. For RSV, it has been shown that a prefusion-stabilized conformation of the viral fusion protein elicits superior neutralizing antibody responses than its postfusion conformation (*138-140*).

Based on this observation, many groups have successfully stabilized the fusion protein in a neutralization-sensitive, prefusion conformation for HIV Env glycoprotein (*141-143*), MPV (*144*) and MERS-CoV virus (*145*) by using the rational protein design or *in vitro* evolution to identify the structural features that favor the prefusion conformation, for example, introduction of the disulfide bridge, replacement of the residue in the flexible loop, and increase of the protein core compactness.

Germline targeting immunogen

This approach is frequently used in the HIV vaccine research field. A major hurdle for eliciting broadly neutralizing antibodies against HIV-1 is that those antibodies are heavily somatically mutated, and often the inferred germline precursors of these antibodies lack detectable affinity for the viral protein. Thus, the germline targeting approach was proposed to use a rationally designed immunogen to activate the specific B-cell lineage, which encoded the unmutated precursors of the bnAbs. This work was pioneered by William Schief's group, which engineered the outer domain (*146*) of the viral gp120 protein to bind the germline precursors of VRC01 (glVRC01) with low affinity (*147-149*). Through iterative sequence optimizations, they evolved the immunogen towards binding a panel of VRC01-class antibodies and the immunogen enabled to trigger VRC01-like precursors in transgenic mice models. Several follow-up studies using the same approach to engineer Env trimers demonstrated the capability to engage germline precursors of antibodies of other specificities (*150, 151*), e.g., PGT121-like antibodies.

Epitope-focused scaffolding

The epitope scaffolding approach was proposed to use the co-crystal structure of a nAb with its targeting epitope on the viral protein and transferred this binding conformation of the epitope into a non-viral protein scaffold, yielding an epitope-focused immunogen. By doing so, the engineered immunogen presents the epitope in the correct conformation relevant for antibody binding, but the epitope becomes detached from its native context for a visible presentation to the immune system.

For a few cases, the nAb only requires the linear epitope to engage the specific binding. Thus, simply using the peptide form of the linear sequence with the conjugation of the protein carrier can be sufficient to present the epitope to the B cell. Several studies have shown that by linking the fusion peptide (FP) of HIV Env trimer to the nano-carrier enhanced the immunogenicity of the fusion peptide and the immunization study also revealed that FP-particle elicits HIV neutralizing antibodies with noticeable breadth (*152*). A similar effect was also demonstrated in the RSV research field (*153, 154*).

An alternative approach is using computational protein design to transfer the epitope to the heterogeneous scaffold, and, as the epitope structure is stabilized by the rest of the protein scaffold, the epitope conformation is maintained. Beyond linear epitopes, many nAbs target epitopes consisting of multiple segments (discontinuous epitopes). Thus, only by using the computational approach it is possible to stabilize such complex epitopes and preserving the native conformation of the epitope. Several groups have used a computational design approach to successfully generate epitope scaffolds that retain the structure of the epitope for, including 4E10 (*78*) and 2F5 epitope (*79, 155*) of gp41 for HIV, and the Mota epitope of the RSV fusion protein (*99, 156*). Impressively, immunization with an RSV epitope scaffold has shown the capability to elicit, albeit low, neutralization activity across 40 % of non-human-primate model (NHP). Antibodies isolated from the sera specifically bound to the presented epitope with high affinity (*99*). Altogether, the epitope scaffolding approach demonstrated high potential to emerge as an alternative vaccine concept, aiming to elicit precisely controlled nAb responses.

Overall, reverse vaccinology holds great promise for designing immunogens that elicit antibody responses focused on defined neutralization sensitive epitopes (*128, 157*). However, a single strategy is unlikely to be applicable to design efficient immunogens for all pathogens. As a consequence, only a combination of different approaches may overcome the specific challenges associated with each pathogen, for instance using a single epitope scaffold as an immunogen elicits the defined antibodies, but the overall immunogenicity has remained inferior compared to the virus-based immunogens. Thus, future effort is needed to find a suitable combination of different approaches that balance specificity and immunogenicity to advance reverse vaccinology into the next era.

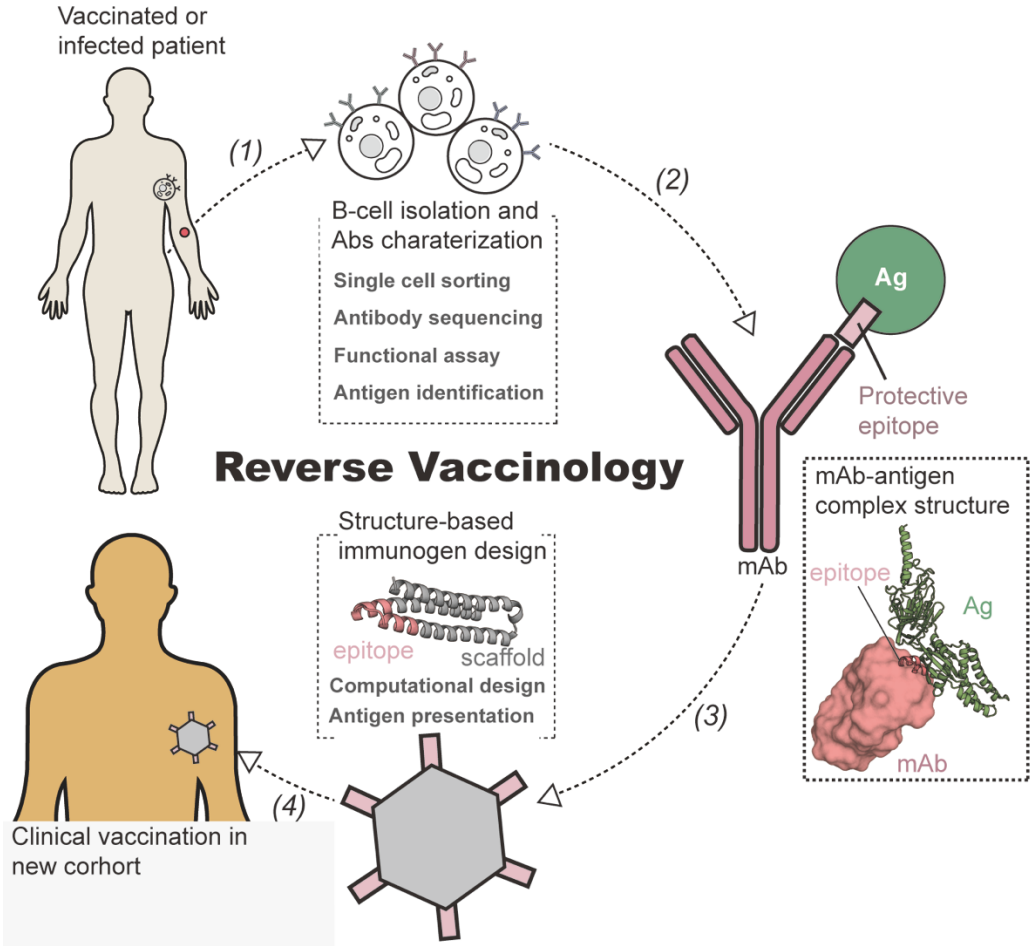


Figure 1.5. The concept of reverse vaccinology.

Step 1: Tracing the B-cell lineage and analyzing the antibody repertoires from infected patients or an immunized cohort that shows a neutralizing serum response against the pathogen. This is followed by isolating potent neutralizing antibodies. Step 2: Characterizing the structural interaction of the antibody with the targeted site (epitope). Step 3: Designing a structure-based immunogen or scaffold to accommodate the epitope in the antibody-bound conformation. Step 4: Re-eliciting the immune response specific to the presented epitope(s). This figure is inspired by Rappuoli et al., 2016 (*110*).

1.7 Respiratory syncytial virus

According to the statistics published by the World Health Organization, lower respiratory tract infection was the second leading cause of premature deaths worldwide between 2000 and 2013. Among these viruses, RSV infections account for 3.4 million hospitalizations each year in children younger than six months and are responsible for 3 to 9 % of the cases of early death in children under five, which amounts to approximately 55,000 to 200,000 deaths annually (*146*). However, as briefly mentioned in the previous section, RSV is still one of the pathogens without a licensed vaccine and with resistance to traditional vaccine development efforts. The only clinically approved treatment for infected patients is an antibody-based prophylaxis with the anti-RSV-F antibody palivizumab (*158, 159*), which is effective in reducing hospitalization of young children at high risk of RSV infection. However, the cost of Synagis manipulation prevents its widespread use, urging the need for the development of an effective vaccine.

Basic virology

RSV is an enveloped *Mononegavirale* belonging to the *Pneumoviridae* family and the genus *Orthopneumovirus* (*160*). The genome of the virus (15.2 kb) is in the form of a linear RNA single strand and has a negative polarity. It comprises ten genes, coding for eleven proteins. The nine structural proteins are: N; P; M; SH; G; F; M2-1; M2-2 and L, and the two nonstructural proteins are nonstructural protein 1 (NS1) and nonstructural protein 2 (NS2) (Fig. 1.6). The G, F and SH proteins make up the viral envelope, while M forms the matrix. There are four viral nucleocapsid forming proteins (N, P, L and M2-1). The virions are spherical (about 200 nm diameter) or filamentous (2-8 μm). RSV primarily targets the ciliated epithelial cells that will produce new viral particles, whose filamentous form favors the spread from cell to cell. The attachment of viral particles to cells is mediated by the glycoprotein of envelope G (RSVG). The glycoprotein F (RSVF) then allows the fusion of the viral and cellular membranes. RSVF also mediates fusion of the membranes of infected cells and those adjacent to the syncytial formation (*161*).

Vaccine trial for RSV

RSV has resisted vaccine development for more than 60 years. During the 1960s, a formalin-inactivated viral isolate was tested in children, leading to a disastrous outcome. Not only did this vaccine fail to protect, but instead resulted in many infants experiencing severe disease enhancement upon seasonal infection resulting in the death of two immunized infants due to vaccine-enhanced disease (*162*). While the reasons for the failure of the FI-RSV vaccine are not fully clear, recent studies have suggested that the presence of non-neutralizing antibodies, directed against structurally altered neutralization epitopes, has played a major role in the disease enhancement (*163-165*). This unfortunate outcome set a large safety constraint for the development of an RSV vaccine. Recently, a new subunit vaccine candidate based on the prefusion-stabilized RSV fusion protein has reported promising results from a phase I trial in adults, providing the first clinical proof-of-concept for structure-based immunogen design (*140, 166*).

RSV surface glycoprotein

In the past five years, our understanding of the structure and function of the RSV surface glycoprotein has progressed considerably and this is a key factor resulting in a recently encouraging clinical outcome (*167*). There are two proteins covering the surface of RSV, G and F protein, which execute the function of recognition of host cells and invasion of host cells, respectively. Among the different RSV subgroups, the surface F glycoprotein is structurally conserved, with sequence identities higher than 90 % across different strains; whereas, the G-glycoprotein is the most variable structural protein of RSV and has no sequence homology with other paramyxovirus attachment proteins. Both F and G glycoproteins are the main antigenic targets recognized by neutralizing antibodies induced by infection; however, further analysis of sera has shown that most of the antibody neutralizing activity is directed against the F protein (*140*).

RSVG is a single-pass type II integral membrane protein possessing molecular motifs similar to the chemokine CX3CR1-Ligand that allows it to bind to the CX3CR1 receptor. The ectodomain of G protein consists of two larger mucin-like repeats flanking a central conserved domain and a heparin-bind domain. The sequence of the conserved region is strictly conserved with an arrangement forming a CX3C

motif. This molecular interaction between the CX3C receptor pair facilitates viral infection and induces leukocyte recruitment *in vitro*. Similarly, the virus integrates into the epithelial cells using the interaction of its viral G protein with glycosaminoglycans (GAGs) (*115, 168*). Although anti-G protein neutralizing antibodies are sparse in the protected serum, recent studies have shown several potent neutralizing antibodies specifically targeting to G protein (*115*).

RSVF protein is a class I fusion protein and is expressed as the single-chain precursor F_0 decorated with 5 – 6 N-linked glycan, followed by intracellular furin protease cleavage to become fusion competent. The mature RSVF protein comprises two subunits F_1 and F_2, which are covalently linked through two disulfide bonds to yield a heterodimeric protomer forming a spheroidal shape trimer with the molecular weight of 180 KDa at the surface of the virus. The fusion peptide is located at the end of the F1 subunit and buried inside the core of the protein. After cleavage and trimerization, the RSVF adopts the prefusion conformation. This prefusion conformation, however, is metastable and easily rearranges into the compact postfusion conformation, mediating the viral entry into the host cell. This refolding process is irreversible, and the post F is extremely stable with a melting temperature higher than 90 °C (*161*).

Based on the surface-exposed region of RSVF protein and the secondary structure elements, the F protein is distinguished into six antigenic patches: site 0; I; II; III; IV and V (Fig 1.6). Among these patches, site 0, I, III and V undergo drastic structural rearrangement during the viral invasion. Therefore, the three antigenic sites that are exclusively present in the prefusion conformation of RSVF are known as prefusion-specific epitopes. In contrast, both antigenic sites II and IV maintain their structural conformation in both pre-and postfusion form ("conformation independent epitopes"). On the other hand, the site I is structurally rearranged during the pre-to-post transition, and the isolated antibodies targeting this site are specific for the postfusion state ("postfusion specific").

Recent study has proved that a large proportion of neutralizing antibodies specifically recognize prefusion-specific epitopes (*139*). Besides, a detailed analysis of the failure of the FI-RSV trial revealed that the majority of fusion proteins at the viral surface were spontaneously transformed into the postfusion conformation due to the formulation preparation (*164*). One of the major reasons for the high potency of prefusion-specific antibodies is that they prevent the structural transition to the postfusion state, which is essential for host cell infection (*114, 117, 119*). Among the most potent prefusion-specific neutralizing antibodies, D25 and 5C4 are well characterized structurally with prefusion RSVF. The neutralizing antibodies were categorized as conformational-independent Nab (*116, 118, 169*), for example palivizumab (site II targeting antibody) and 101F (site IV) are known to block the fusion of the membranes due to binding to either the prefusion form or in the intermediate states and causing steric effects that hinder fusion. In contrast, antibodies solely bound to postfusion conformation, such as the majority of the site I-specific antibodies, are generally non-neutralizing (*170*).

Due to structural and sequence conservation, antibodies targeting to site III and IV are found to broadly neutralize other viruses in the *Pneumoviridae* family, like MPV, another major cause of lower respiratory tract infection in the elderly and children (*114, 171, 172*). Several recent studies have isolated many human mAbs that can cross-react to both RSVF and hMPVF at site III and IV, highlighting those conserved regions as valuable sites for developing vaccines that afford broad protection against multiple viruses (*173, 174*).

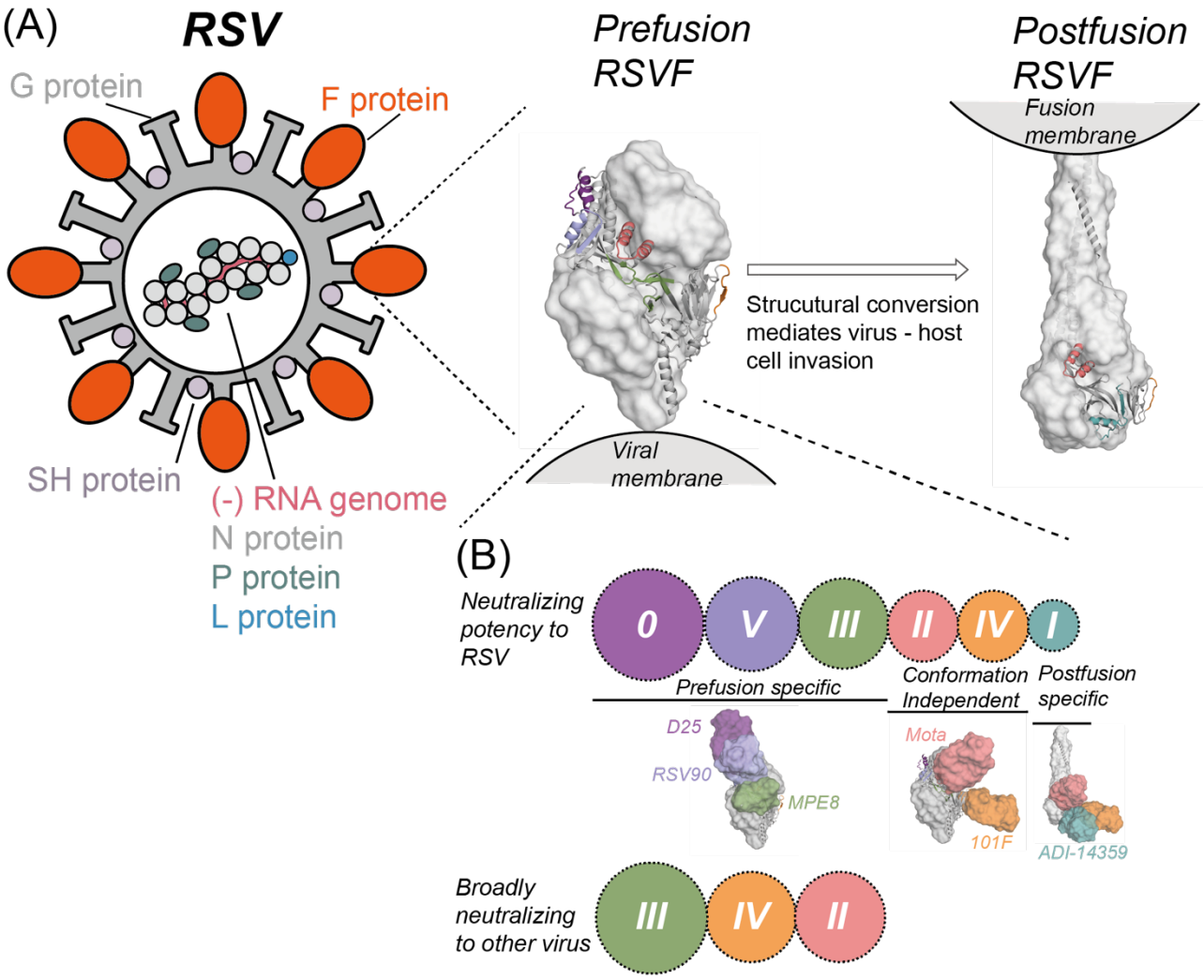


Figure 1.6. RSV viral structures and neutralizing epitope (site) on pre- and post-fusion of RSV.

A) RSV is the RNA virus and comprises eleven proteins. Among these proteins, the fusion protein is a crucial viral surface protein evolving in membrane fusion with the host cell. During the fusion process, RSVF undergoes a drastic conformational transition from pre-fusion form (left, diamond-like shape) to post-fusion conformation (right, core-like shape). B) Qualitative ranking of the neutralization potency of six antigenic sites on RSVF. The neutralization epitopes of RSVF are colored in both conformational states, named antigenic site 0-V. Several neutralizing antibodies that recognize the site III, site IV and site II of RSVF also show a neutralizing effect across the virus in the same phylogenetic family. The radius of the circle represents the potency of the antibodies targeting the given sites. This figure is inspired by Graham et al., 2019 (*128*).

Outlines

My Ph.D. thesis will essentially have two main objectives: *de novo* design of protein presenting complex epitopes, and the induction of nAbs using synthetic immunogens *in vivo*. To be specific, my research aimed to develop and validate new methodologies for the design of functional *de novo* proteins, applied to different biological questions as outlined below. As an important aspect of my work, the experimental and functional characterizations served to optimize the computational design process.

Objective I: Validate the computational methods for modeling extensive backbone movement of protein to present structurally distinct functional motifs

To date, the field of computational protein design has been mostly focused on the design of protein structures that are thermostable and structurally accurate, but these were often deprived of biochemical function. One of the reasons for this is the lack of a designable protein backbone suitable to accommodating functional modules. To address this specific aim, chapter 2 and chapter 4 present two design algorithms that allow tailoring of the optimal protein scaffolds for transplanting naturally occurring functional motifs. In chapter 2, we improved the previously published algorithm by Correia *et al*. in 2014, which was invented to couple the protein folding process with the sequence space exploration in a single simulation, thus sampling larger backbone movement in order to properly stabilize the inserted functional motif. In our improved version, named Rosetta FunFolDes, we embedded several technical improvements allowing the design of complex functional proteins presented in chapters 3 and 4. The main improvements included: 1) capability of inserting multi-segment functional sites into one topology, which leads to design of the bi-functional *de novo* protein, in chapter 4, for modulating the heterodimerization of synthetic receptors and 2) incorporation of the binding target as a functional constraint in the simulation. We then presented *in silico* benchmarks to demonstrate the improvement of predictive binding energy under the simulation, imposing binding constraints and also experimentally showing two new proteins that carry the antibody-binding motif specifically bound to target antibody.

Beyond repurposing existing protein scaffolds, chapters 3 and 4 present a novel algorithm to build novel functional proteins from scratch (TopoBuilder). Its aim is to systematically generate protein topologies by assembling the ideal protein secondary structural elements around the motif of interest and ultimately designing the protein with native-like local features. This approach will potentially expand the protein structural space to provide more topologies available to design the functionality, and reduce the dependency, of using natural protein as design template.

Objective II: *De novo* design of epitope-focused immunogens to present structurally irregular neutralization epitopes

So far, the development in protein design methodology has enabled the design of diverse *de novo* proteins with control of structural features that fold into a prescribed three-dimensional structure. The design of a *de novo* protein with a functional motif installed, however, has been proved to be far more challenging. For most of the successful attempts, the functional motifs that have been transplanted are restricted to commonly structural segments, such as the regular, linear helical segment. However, the

majority of functional motifs found in natural proteins is comprised of multiple irregular segments that are discontinuously arranged in the tertiary structure of the proteins. Towards this specific aim, we demonstrated two design strategies in chapter 3 for transplanting binding motifs with unprecedented structural complexity (irregular β-strand and discontinuous segments) into the *de novo* proteins. To improve the success of designing functional proteins, chapters 3 and 4 also showcase that a fine control of subtle topological orientation during the design may obviate the iterative experimental optimizations for functional improvement.

Objective III: Eliciting RSV neutralizing antibodies with cocktails of designed immunogens

Traditional vaccine development efforts are facing a roadblock, failing to induce protective immunity against several pathogens. This failure is generally attributed to the fact that such vaccine approaches do not elicit antibody responses focused towards sites where the pathogen is vulnerable, allowing it to escape the host immune response. The main effort in this field throughout the last few decades has been the design of immunogens that spotlight neutralization epitopes for efficient recognition through the immune system. Using the computational methods that we developed in chapters 2 and 4, we have engineered *de novo* proteins that present RSV neutralization epitopes outside of their native environment (RSVF). Using a cocktail of three epitope-focused immunogens, we show the consistent elicitation of neutralizing antibodies in mice and non-human primates targeting the presented antigenic sites. Then, chapter 3 also shows that, a cocktail formulation of multiple epitope-focused immunogens is able to reshape serum antibody specificities under conditions of preexisting immunity in non-human primates.

Objective IV: Generic approach to design functional *de novo* proteins

In chapter 4, my thesis describes a "bottom-up" functional protein design strategy, aiming at centering the design process around the functional module. This generalized methodology allows the modularly-assembly of protein secondary structural elements to accommodate the functional binding motifs extracted from natural proteins. By this means, we were able to customize the fine topology depending on the functional motif's constraints to stabilize functional sites in *de novo* proteins. Using the "bottom-up" design approach, we designed the *de novo* protein to stabilize RSV epitopes and applied those designed proteins as synthetic immunogens (chapters 2 and 3) and as biosensors to quantitatively profile epitope-specific antibody responses directly from the patient's sera (chapter 4). Besides this, we were able to assemble the protein topology simultaneously accommodating two binding motifs assembling the shared characteristic of a natural protein having more than one binding motif in order to interplay with other molecules. Using this *de novo* bi-functional protein, we successfully triggered the dimerization of synthetic receptors in the cell, allowing manipulating the gene expression *in vitro* (chapter 4). Altogether, we presented a versatile strategy for the *de novo* design of customized proteins carrying structurally complex binding motifs, applicable to a wide range of functional protein design challenges.

Chapter 2 Rosetta FunFolDes – a general framework for the computational design of functional proteins

This chapter is published in Plos Computational biology in 2018 (10.1371/journal.pcbi.1006623).

Authors and affiliations:

Jaume Bonet[1,2¶], Sarah Wehrle[1,2¶], Karen Schriever[1,2¶], **Che Yang**[1,2¶], Anne Billet[1,2], Fabian Sesterhenn[1,2], Andreas Scheck[1,2], Freyr Sverrisson[1,2], Barbora Veselkova[3,4], Sabrina Vollers[1,2], Roxanne Lourman[1,2], Mélanie Villard[1,2], Stéphane Rosset[1,2], Thomas Krey[3,4], Bruno E. Correia[1,2*].

[1]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland. [3]Institute of Virology, Hannover Medical School, Hannover, Germany. [4]German Center for Infection Research (DZIF), Hannover, Germany. *Corresponding author: bruno.correia@epfl.ch ¶These authors contributed equally to this work.

Contributions:

J.B. coded the algorithm described. A.S coded the StructFragMover. K.S., A.B., A.S. and F.S. performed computational design simulations. C.Y., S.W., K.S., A.B., F.S., S.V., R. L., M. V. and S.R. contributed to experimental characterization of the designed proteins. J.B. and B.E.C. designed the study and wrote manuscript.

2.1 Abstract

The robust computational design of functional proteins has the potential to deeply impact translational research and broaden our understanding of the determinants of protein function and stability. The low success rates of computational design protocols and the extensive *in vitro* optimization often required, highlight the challenge of designing proteins that perform essential biochemical functions, such as binding or catalysis.

One of the most simplistic approaches for the design of function is to adopt functional motifs in naturally occurring proteins and transplant them to computationally designed proteins. The structural complexity of the functional motif largely determines how readily one can find host protein structures that are "designable", meaning that are likely to present the functional motif in the desired conformation. One promising route to enhance the "designability" of protein structures is to allow backbone flexibility. Here, we present a computational approach that couples conformational folding with sequence design to embed functional motifs into heterologous proteins - Rosetta Functional Folding and Design (FunFolDes).

We performed extensive computational benchmarks, where we observed that the enforcement of functional requirements resulted in designs distant from the global energetic minimum of the protein. An observation consistent with several experimental studies that have revealed function-stability tradeoffs.

To test the design capabilities of FunFolDes we transplanted two viral epitopes into distant structural templates including one *de novo* "functionless" fold, which represent two typical challenges where the designability problem arises. The designed proteins were experimentally characterized showing high

binding affinities to monoclonal antibodies, making them valuable candidates for vaccine design endeavors.

Overall, we present an accessible strategy to repurpose old protein folds for new functions. This may lead to important improvements on the computational design of proteins, with structurally complex functional sites, that can perform elaborate biochemical functions related to binding and catalysis.

2.2 Author Summary

The ability to use computational tools to manipulate the structure and function of proteins has the potential to impact many facets of fundamental and translational science. Due to our limited understanding of the principles that govern protein function and structure, the computational design of functional proteins remains challenging. We developed a computational protocol (Rosetta FunFolDes) to facilitate the insertion of functional motifs into heterologous proteins. We performed extensive in silico benchmarks, and found that when the design of function is required the global energy minima may not be the optimal solution, in line with previously reported experimental studies.

Further, we used FunFolDes to design two novel functional proteins, displaying two viral epitopes that can be of interest for vaccine development. The designed proteins were experimentally characterized, showing that functionalization was successfully achieved. These results highlight the capability of FunFolDes to address common challenges on the design of functional proteins. In particular, the reduced structural compatibility between functional sites and host scaffolds, effectively enabling the repurposing of old protein folds for new functions. Overall, FunFolDes provides new means to accomplish the challenging task of functionalizing computationally designed proteins.

2.3 Introduction

Proteins are one of the main functional building blocks of the cell. The ability to create novel proteins outside of the natural realm has opened the path towards innovative achievements, such as new pathways (*175*), cellular functions (*176*), and therapeutic leads (*177-179*). Computational protein design is the rational and structure-based approach to solve the inverse folding problem, i.e. the search for the best putative sequence capable of fitting and stabilizing a protein's three-dimensional conformation (*180*). As such, a great deal of effort has been placed into understanding the rules of protein folding and stability (*181, 182*) and its relation to the appropriate sequence space (*183*).

Computational protein design approaches focus on exploring two interconnected landscapes related to sampling of the conformational and sequence spaces. Fixed backbone approaches use static protein backbone conformations, which greatly constrain the sequence space explored by the computational algorithm (*183*). Following the same principles of naturally occurring homologs, which often exhibit confined structural diversity, flexible backbone approaches enhance the sequence diversity, adding the challenge of identifying energetically favorable sequence variants that are correctly coupled to structural perturbations (*184*).

Another variation for computational design approaches i*s de novo* design, in which protein backbones are assembled *in silico*, followed by sequence optimization to fold into a pre-defined three-dimensional conformation without being constrained by previous sequence information (*185-187*). This approach

tests our understanding of the rules governing the structure of different protein folds. The failures and successes of this approach confirm and correct the principles used for the protein design process (*181, 182*).

One of the main aims of computational protein design is the rational design of functional proteins capable of carrying existing or novel functions into new structural contexts (*188*). Broadly, there are three main approaches for the design of functional proteins: redesigning of pre-existing functions, grafting of functional sites onto heterologous proteins, and designing of novel functions not found in the protein repertoire. The redesign of a pre-existing function to alter its catalytic activity (*189*) or improve its binding target recognition (*190*) can be considered the most conservative approach. It is typically accomplished by point mutations around the functional area of interest and tends to have little impact on global structure of the designed protein. On the other extreme, the design of fully novel functions has most noticeably been achieved by applying chemical principles that tested our fundamental knowledge of enzyme catalysis (*191, 192*).

Between these two approaches resides protein grafting. This method aims to repurpose natural folds as carriers for exogenous known functions. It relies on the strong structure-function relationship present in proteins, to endow an heterologous protein with an exogenous function by means of transferring a structural motif that performs such function (*177-179, 193-196*).

At the biochemical level, grafting approaches have been used to design high binding affinity protein-protein interactions, by stabilizing binding motifs removing the entropic cost of binding (e.g., flexible peptides) (*195*), and also by extending the binding interfaces to allow for additional energetically favorable interactions. The extended interfaces also provide opportunities to tune the specificity of the designed proteins (*195*). On the practical side, some of the most notable applications of protein grafting thus far, have been the design of novel viral inhibitors (*195, 197*) and epitope-focused immunogens for vaccine design (*177-179*). Following this strategy one can easily imagine applications to functionalize protein-based biomaterials (*27*) or to design novel biosensors (*198*). The importance of robust grafting approaches to functionalize heterologous proteins is related to the fact that the proteins that naturally perform these functions, may lack the best biochemical properties in terms of size, affinity, solubility, immunogenicity and other application specific factors.

Thus far, the most successful grafting approaches are highly dependent on structural similarity between the functional motif and the insertion region in the protein scaffold. When the functional motif and the insertion region are identical in backbone conformation, the motif transfer can be performed by side chain grafting, i.e. mutating the target residues into those of the functional motif (*177, 179*). In much more challenging scenarios, full backbone grafting may be used in conjunction with directed evolution (*193*). Nevertheless, motif transfer is limited between very similar structural regions, which greatly constrains the subset of putative scaffolds that can be used for this purpose. The lack of compatibility between the putative scaffolds and the functional sites has been referred to as a “designability” problem (*199, 200*), which refers to the likelihood of a protein backbone to host and stabilize a structural motif. The designability problem becomes more obvious as the structural complexity of the functional motif grows, drastically limiting the types of functional motifs that can be transferred. Previously, we have demonstrated the possibility of expanding protein grafting to scaffolds with segments that have low structural similarity. To accomplish that task, we developed the prototype protocol Rosetta Fold From Loops (*159*) (*178, 195*).

The distinctive feature of our protocol is the coupling of the folding and design stages to bias the sampling towards structural conformations and sequences that stabilize the grafted functional motif. In the past, FFL was used to obtain designs that were functional (synthetic immunogens (*178*) and protein-based inhibitors (*195*)) and where the experimentally determined crystal structures closely resembled the computational models. However, the structures of the functional sites were structurally very close to the insertion segments of the hosting scaffolds. The architecture of FFL was intrinsically limited in the types of constraints available and the grafting of linear, single segment functional motifs.

Here, we present a complete re-implementation of FFL with enhanced functionalities, simplified user interface and complete integration with other Rosetta protocols. We have called this new, more generalist protocol Rosetta Functional Folding and Design (FunFolDes). We benchmarked FunFolDes extensively, unveiling important technical details to better exploit and expand the capabilities of the protocol. Furthermore, we challenged FunFolDes with two design tasks of transplanting viral epitopes to heterologous scaffolds, and by doing so probe the applicability of the protocol. The design tasks were centered on using distant structural templates as hosting scaffolds, and functionalizing a *de novo* designed protein, - FunFolDes succeeded in both challenges. These results are encouraging and provide a solid basis for the broad applicability of FunFolDes as a strategy for the robust computational design of functionalized proteins.

2.4 Results

The original prototype of the Rosetta Fold From Loops (*159*) protocol was successfully used to transplant the structural motif of the Respiratory Syncytial Virus protein F (RSVF) site II neutralizing epitope into a protein scaffold in the context of a vaccine design application (*178*).

FFL enabled the insertion and conformational stabilization of the structural motif into a defined protein topology by using Rosetta's fragment insertion machinery to fold an extended polypeptide chain to adopt the desired topology (*201*) which was then sequence designed. Information from the scaffold structure was used to guide the folding, ensuring an overall similar topology while allowing for the conformational changes needed to stabilize the inserted structural motif.

The final implementation of FunFolDes is schematically represented in **Fig 2.1**, and fully described in Materials and Methods. Our upgrades to FFL focused on three main aims: I) improve the applicability of the system to handle more complex structural motifs (i.e. multiple discontinuous backbone segments); II) enhance the design of functional proteins by including binding partners in the simulations; III) increase the control over each stage of the simulation improving the usability for non-experts. These three aims were achieved through the implementation of five core technical improvements described below.

Insertion of multi-segment functional sites. Most functional sites in proteins typically entail, at the structural level, multiple discontinuous segments, as is the case for protein-protein interfaces, enzyme active-sites, and others (*202, 203*). FunFolDes handles functional sites with any number of discontinuous segments, ensuring the native orientations of each of the segments. These new features enhance the types of structural motifs that can be handled by FunFolDes, widening the applicability of the computational protocol.

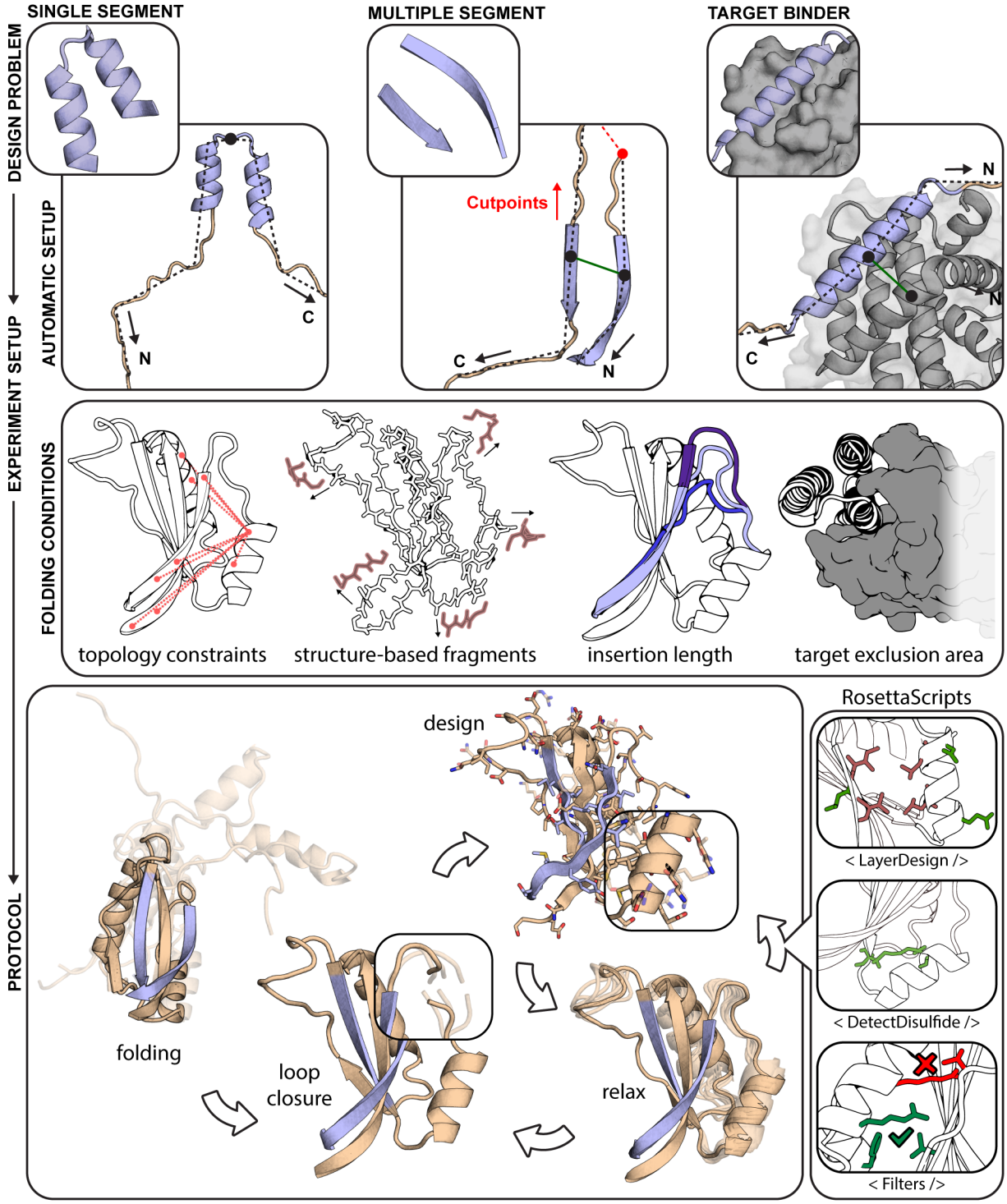


Figure 2.1. Rosetta FunFolDes - method overview.

FunFolDes was devised to tackle a wide range of functional protein design problems, combining a higher user control of the simulation parameters whilst simultaneously lowering the level of expertise required. FunFolDes is able to transfer single- and multi-segment motifs (light blue) together with the target partner (*204*) by exploiting Rosetta's FoldTree framework (top row). A wider range of information can be extracted from the template (wheat) to shift the final conformation towards a more productive design space (middle row), including targeted distance constraints, generation of structure-based fragments, motif insertion in sites with different residue length and presence of the binding target to bias the folding stage. The bottom row showcases the most typical application of the FunFolDes protocol. Implementation in RosettaScripts allows to tailor FunFolDes behavior. A seamless

integration with other protocols and complex selection logics can be added to address the different needs in each design task.

Structural folding and sequence design in the presence of a binding target. Many of the functional roles of proteins in cells require physical interaction with other proteins, nucleic acids, or metabolites (*205*). The inclusion of the binder has two main advantages: I) explicit representation of functional constraints to bias the designed protein towards a functional sequence space, resolving putative clashes derived from the template scaffold; II) facilitate the design of new additional contact residues (outside of the motif) that may afford enhanced affinity and/or specificity.

Region-specific structural constraints. FunFolDes can collect from full-template to region-specific constraints, allowing greater levels of flexibility in areas of the scaffold that can be critical for function (e.g., segments close to the interface of a target protein). The type of distance constrains used in the protocol are soft constrains with score penalties if the defined standard deviation is exceeded in the upper and lower bounds. Furthermore, FunFolDes is no longer limited to atom-pair distance constraints (*206*) and can incorporate other types of kinematic constraints, such as angle and dihedral constraints (*207*), which have been used extensively to design beta-rich topologies (*182*).

On-the-fly fragment picking. Classically, fragment libraries are generated through sequence-based predictions of secondary structure and dihedral angles that rely on external computational methods (*208*). We leveraged internal functionalities in Rosetta so that FunFolDes can assemble fragment sets on-the-fly. Using this feature, we can assemble fragment sets based on the structure of the input scaffold. Sequence-based fragments remain an option, however this feature removes the need for secondary applications, boosting the usability of FunFolDes. Lastly, the on-the-fly fragment picking enables the development of protocols with mutable fragment sets along the procedure.

Compatibility with other Rosetta modules. Finally, FunFolDes is compatible with Rosetta's modular xml-interface - Rosetta Scripts (RS) (*209*). Enabling customization of the FunFolDes protocol and, more importantly, cross-talk with other protocols and filters available through the RS interface.

We devised two benchmark scenarios to test the performance of FunFolDes. One of these aimed to capture conformational changes in small protein domains caused by sequence insertions or deletions, and the second scenario assessed protocol performance to fold and design a binder in the presence of the binding target.

Capturing conformational and sequence changes in small protein domains

Typical protein design benchmarks are assembled by stripping native side chains from known protein structures and evaluating the sequence recovery of the design algorithm (*183*). The main design aim of FunFolDes is to insert structural motifs into protein folds while allowing flexibility across the overall structure. This conformational freedom allows the full protein scaffold to adapt and stabilize the functional motif's conformation. This is a main distinctive point from other approaches to design functional proteins that rely on a mostly rigid scaffold (*176, 177, 185, 193, 203, 210*). For many modeling problems, such as protein structure prediction, protein-protein and protein-ligand docking, and protein design, standardized benchmark datasets are available (*211*) or easily accessible. Devising a benchmark for

designed proteins with propagating conformational changes across the structure is challenging, as we are assessing both structural accuracy as well as sequence recovery of the protocol.

To address this problem, we analyzed structural domains found repeatedly in natural proteins and clustered them according to their definition in the CATH database (*212*). As a result, we selected a set of 14 benchmark targets labeled T01 through T14 (**Fig 2.2A**). A detailed description on the construction of the benchmark can be found in the Materials and Methods section.

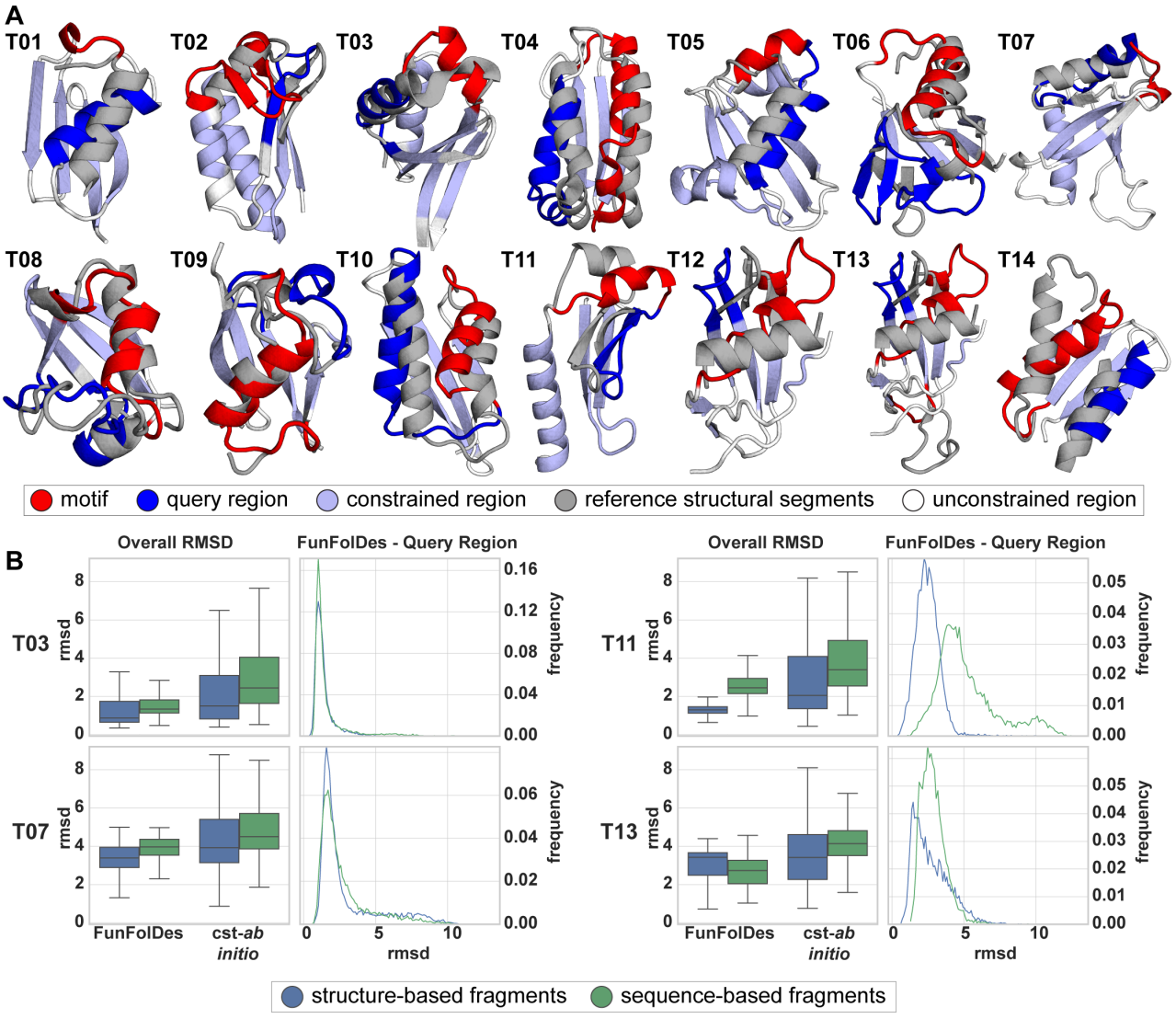


Figure 2.2. Benchmark test set to evaluate FunFolDes structural sampling.

A) Structural representation of the 14 targets used in the benchmark. In each target is highlighted the motif (red) and query (blue) regions, and the positions from which distance constraints were generated (light blue). Conformations of the motif and query regions, as found in the template structures, are shown superimposed in light grey. B) Full structure RMSD (Overall RMSD) and local RMSD for the query region (FunFolDes – Query Region) is presented for four targets (full dataset presented in **Fig. S2.1**). Overall RMSD compares results for the two simulation modes (FunFolDes Vs. constrained–*ab initio (cst-ab initio*)) and the two fragment generation methods (structure (blue) Vs. sequence-based fragments (*63*)) against their original target. FunFolDes more frequently samples RMSDs closer to the conformation of the target structure. Generally, structure-based fragments contribute to lower mean overall RMSDs. The FunFolDes – Query Region RMSD distributions show that the two fragment sets do not have a major importance in the structural recovery of the query region.

Briefly, for the benchmark we selected proteins with less than 100 residues, where each test case was composed of two proteins of the same CATH domain cluster. One of the proteins is the template, and serves as a structural representative of the CATH domain. The second protein, dubbed target, contains structural insertions or deletions (motif region), to which a structural change in a different segment of the protein could be attributed (query region). The motif and query regions for all the targets are shown in **Fig 2.2A** and quantified by the percentage of overall secondary structure in **Fig. S2.1A**. To a great extent, these structural changes due to natural sequence insertions and deletions are analogous to those occurring in the design scenarios for which FunFolDes was conceived.

Using FunFolDes, we folded and designed the target proteins while maintaining the motif segment structurally fixed, mimicking a structural motif insertion. Distance constraints between residues were extracted from the template in the regions of shared structural elements of the template and the target, and were used to guide the folding simulations.

To check whether FunFolDes enhances sequence and structural sampling, we compared the simulations to constrained *ab initio* (cst-*ab initio*) simulations (*207*). As Rosetta conformational sampling is highly dependent upon the fragment set (*213*), in this benchmark we also tested the influence of structure- and sequence-based fragments. The performance of the two protocols was analyzed regarding global and local recovery of both structure and sequence.

Structural recovery was assessed through two main metrics: (a) global RMSD of the full decoys against the target and (b) local RMSD of the query region. When evaluating the distributions for global RMSD in the designed ensembles, FunFolDes outperformed cst-*ab initio* by consistently producing populations of decoys with lower mean (RMSDs mostly found below 5 Å), a result observed in all 14 targets (**Fig 2.2B**, **Fig. S2.1B**). This result is especially reassuring considering that FunFolDes simulations contain more structural information of the target topology than the cst-*ab initio* simulations.

The local RMSDs of the query unconstrained regions presented less clear results across the benchmark (**Fig. S2.1B**). In 13 targets, FunFolDes outperformed cst-*ab initio*, showing lower mean RMSDs but in some targets with minor differences.

When comparing fragment sets (structure- vs sequence-based), both achieved similar mean RMSDs in the decoy populations; nonetheless, the structure-based fragments more often reached the lowest RMSDs for overall and query RMSDs (**Fig 2.2B**, **Fig. S2.1**). This is consistent with what would be expected from the structural information content within each fragment set. When paired with the technical simplicity of use, time-saving and enhanced sampling of the desired topology, the structure-based fragments are an added value for FunFolDes.

We also quantified sequence recovery, both in terms of sequence identity and similarity according to the BLOSUM62 matrix (*214*) (**Fig 2.3A**). In all targets, FunFolDes showed superior recoveries than cst-*ab initio*, and at the levels of other design protocols using Rosetta (*184*) (**Fig 2.3A**). This type of metrics has been shown to be highly dependent on the exact backbone conformation used as input (*183, 184*). Given that FunFolDes is exploring larger conformational spaces, as a proxy for the quality of the sequences generated, we used the target's Hidden Markov Models (HMM) (*215*) and quantified the designed sequences that were identified as part of the target's CATH superfamily according to its HMM definition (**Fig 2.3B**). FunFolDes decoy populations systematically outperformed those from cst-*ab initio* (**Fig 2.3B**). The performance of the two fragment sets shows no significant differences.

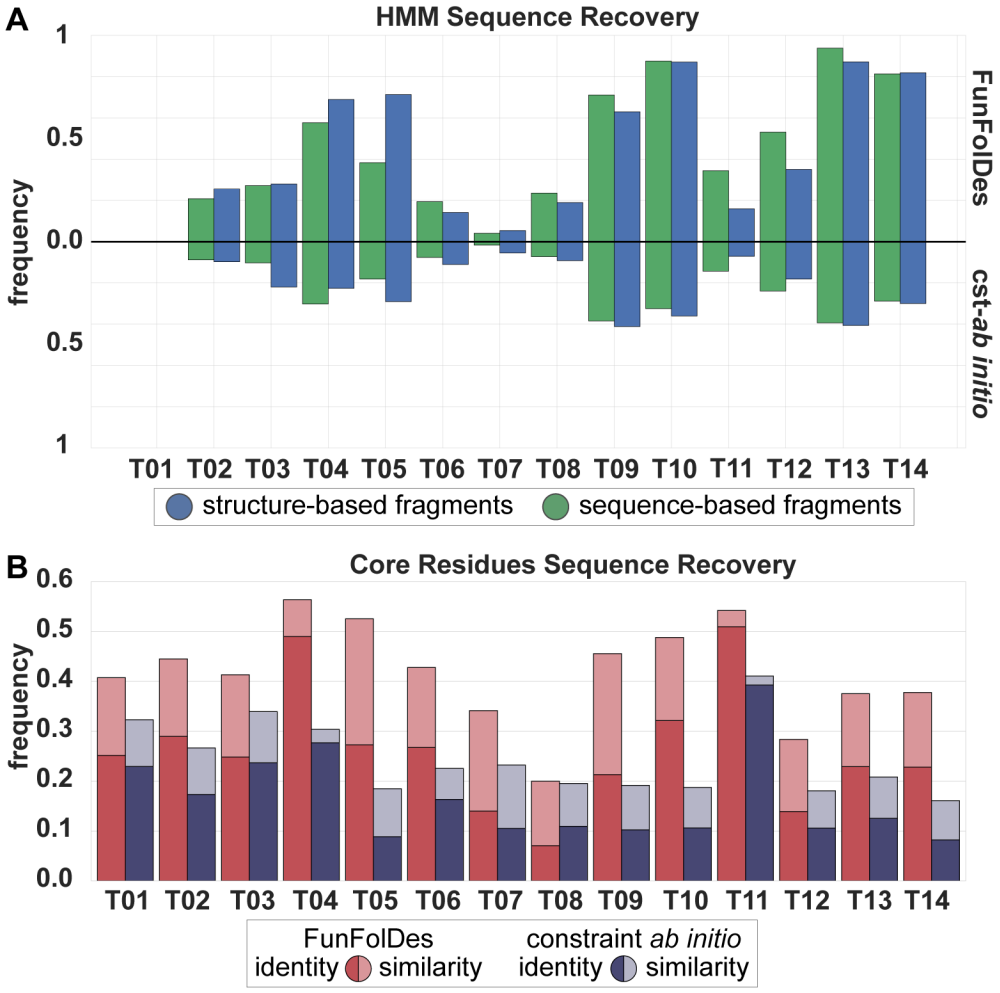


Figure 2.3. Assessment of FunFolDes sequence sampling quality.

A) HMM Sequence Recovery measures the percentage of decoys generated that can be assigned to the original HMM from the CATH superfamily. FunFolDes consistently outperforms cst–*ab initio*, in agreement with the structural recovery metrics. B) Core Residues Sequence Recovery shows the sequence recovery between the core residues of the designs set and the target. Recovery is measured in terms of sequence identity and sequence similarity (as assigned through BLOSUM62). Core sequence identity and similarity was assessed over the structure-based fragment set. According to this metric FunFolDes outperforms cst-*ab initio* in every instance, reaching for some populations, levels of conservation similar to those found in more restrained flexible-backbone design approaches (*184*).

In summary, this benchmark highlights the ability of FunFolDes to generate close-to-native scaffold proteins to stabilize inserted structural motifs. FunFolDes aims to refit protein scaffolds towards the structural requirements of a functional motif. It is thus critical, to explore within certain topological boundaries, structural variations around the original templates. This benchmark points to several variables in the protocol that resulted in enhanced structural and sequence sampling.

Target-biased folding and design of protein binders

The computational design of proteins that can bind with high affinity and specificity to targets of interest remains a largely unsolved problem (*216*). Within the FunFolDes conceptual approach of coupling folding with sequence design, we sought to add the structure of the binding target (**Fig 2.1**) to attempt to bias sampling towards functional structural and sequence spaces.

Previously, we used FFL to design a new binder (BINDI) to BHRF1 (**Fig 2.4A**), an Epstein-Barr virus protein with anti-apoptotic properties directly linked to the tumorigenic activity of EBV (*195*). FFL designs bound to BHRF1 with a dissociation constant (K_D) of 58-60 nM, and after affinity maturation reached a K_D of 220±50 pM. BINDI was designed in the absence of the target and then docked to BHRF1 through the known interaction motif. A striking observation from the overall approach was that the FFL stage was highly inefficient, generating a large fraction of backbone conformations incompatible with the binding mode of the complex.

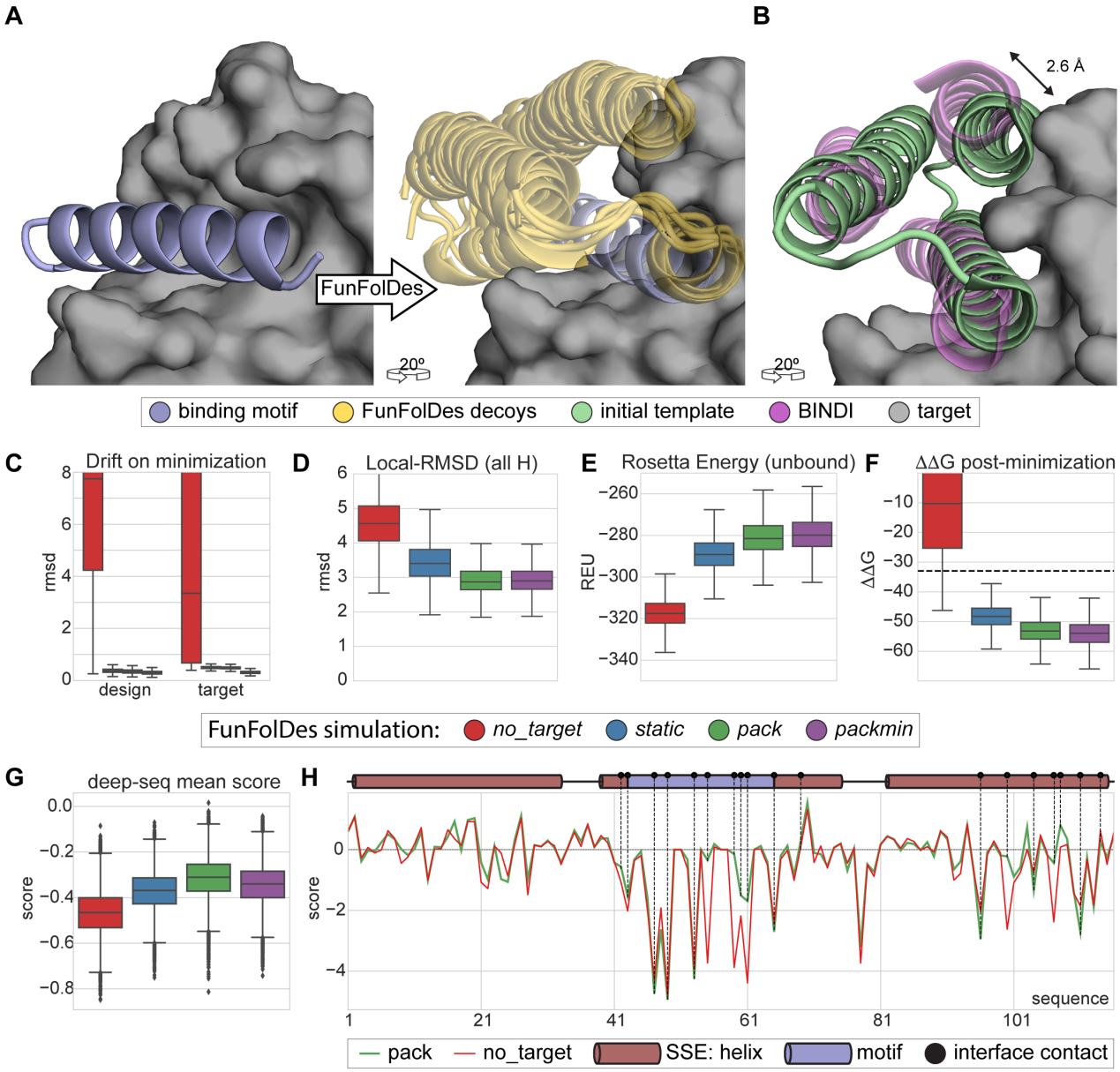


Figure 2.4. Target-biased design of a protein binder and performance assessment based on saturation mutagenesis.

A) Depiction of the initial design task, a single-segment binding motif (BIM-BH3) shown in light blue cartoons, with its target (BHRF1) shown in gray surface, is used by FunFolDes to generate an ensemble of designs compatible with the binding mode shown in light orange cartoons. B) Conformational difference between the initial template (PDB ID: 3LHP), shown in light brown and the previously designed binder (BINDI), shown in violet cartoons, helix 3 requires a subtle but necessary shift (2.6 Å) to avoid steric clashes with the target. C-G) Scoring metrics for design populations according to the simulation mode: *no_target* - FunFolDes was used without the target protein; *static* - target present no flexibility allowed; *pack* - target allowed to repack the side-chains; *packmin* – side chain repacking plus minimization and backbone minimization were allowed for the target. The target flexibility was allowed during the relax-design cycles of FunFolDes. C) Structural drift observed for design and target binder measured as the RMSD between pre- and post-minimization conformations. D) Structural recovery of the conformation observed in the BINDI-BHRF1 assessed over the 3 helical segments of the bundle. E) Rosetta energy for the designs in the unbound state generated by different simulation modes. F) Interaction energy (ΔΔG) between the designs and the target. G) Deep-sequencing score distribution for each design population, computed as the mean score of each sequence after applying a position score matrix based on the deep-sequencing data. The *pack* population slightly outperforms the other simulation modes. H) Per-residue scoring comparison of the *no_target* and the *pack* populations according to the deep-sequencing data. Although the behavior is overall similar, *pack* outperforms *no_target* in multiple positions, several of which are highlighted (black dots) as interfacial contacts or second shell residues close to the binding site.

To test whether the presence of the target could improve structural and sequence sampling, we leveraged the structural and sequence information available for the BINDI-BHRF1 system and benchmarked FunFolDes for this design problem. As described by Procko and colleagues, when comparing the topological template provided to FFL and the BINDI crystal structure, the last helix of the bundle (helix 3) was shifted relative to the template ensuring structural compatibility between BINDI and BHRF1 (**Fig 2.4B**). We used this case study to assess the capabilities of FunFolDes to sample closer conformations to those observed in the BINDI-BHRF1 crystal structure. In addition, we used the saturation mutagenesis data generated for BINDI (*195*) to evaluate the sequence space sampled by FunFolDes.

A detailed description of this benchmark can be found in the Materials and Methods section. Briefly, we performed four different FunFolDes simulations: I) binding target absent (*no_target*); II) binding target present with no conformational freedom (*static*); III) binding target present with side chain repacking (*21*); IV) binding target present with side chain repacking plus minimization and backbone minimization (*packmin*).

no_target simulations generated a low number of conformations compatible with the target (<10% of the total generated designs) (**Fig. S2.2A**). Upon global minimization more than 60% (**Fig. S2.2A**) of the decoys were compatible with the binding target, at the cost of considerable structural drifts for both binder (mean RMSD 3.3 Å) and target (mean RMSD 7.7 Å) (**Fig 2.4C**). These structural drifts reflect the energy optimization requirements by the relaxation algorithms but are deemed biologically irrelevant due to the profound structural reconfigurations. In contrast, simulations performed in the presence of the target clearly biased the sampling to more productive conformational spaces. RMSD drifts upon minimization were less than 1 Å for both designs and binding target (**Fig 2.4C**).

Global structural alignments of the designs fail to emphasize the differences of the helical arrangements (**Fig. S2.2B**). Thus, we aligned all the designs on the conserved binding motif (**Fig 2.4A**) and measured the RMSD over the three helices that compose the fold. FunFolDes simulations in the presence of the target sampled a mean RMSD of 3 Å (lowest ≈ 2 Å) compared to the BINDI structure (**Fig 2.4D**), with the closest designs at approximately 2 Å, while the *no_target* simulation showed a mean RMSD of 4.5 Å (lowest ≈ 2.5 Å). While we acknowledge that these structural differences are modest, the data suggests that they can be important to sample conformations and sequences competent for binding.

We also analyzed Rosetta energy distributions of designs in the unbound state for the different simulations. We observed noticeable differences for the designs generated in the absence (*no_target*) and the presence of the binding target, -320 and -280 Rosetta Energy Units (REUs), respectively (**Fig 2.4E**). This difference is significant, particularly for a small protein (116 residues). We also observed considerable differences for the binding energies (ΔΔG) of the *no_target* and the bound simulations with mean ΔΔGs of -10 and -50 REUs, respectively (**Fig 2.4E**).

The energy metrics provide interesting insights regarding the design of functional proteins. Although the sequence and structure optimization for the designs in the absence of the target reached lower energies, these designs are structurally incompatible with the binding target and, even after refinement, their functional potential (as assessed by the ΔΔG) is not nearly as favorable as those performed in the presence of the binding target (**Fig 2.4F**). These data suggest that, in many cases, to optimize function it may be necessary to sacrifice the overall computed energy of the protein, a common proxy to the experimental thermodynamic stability of the protein (*217*). The existence of stability-function tradeoffs has been the subject of many experimental studies (*218, 219*), however, it remains a much less explored strategy in computational design, where it may also be necessary to design proteins with lower stability to ensure that the functional requirements can be accommodated. This observation provides a compelling argument to perform biased simulations in the presence of the binding target, which can be broadly defined as a "functional constraint".

To evaluate sequence sampling quality, we compared the computationally designed sequences to a saturation mutagenesis dataset available for BINDI (*195*).

The details of the dataset and scoring scheme can be found in the methods and S2 Figure. Briefly, point mutations beneficial to the binding affinity to BHRF1 have a positive score, deleterious mutants a negative score, and neutral score 0. Such a scoring scheme, will yield a score of 0 for the BINDI sequence.

Designs performed in the presence of the binding target obtained higher mean scores as compared to the *no_target* designs (**Fig 2.4G**). The *pack* simulation, showed the highest distribution mean, having one design scoring better than the BINDI sequence. In some key positions at the protein-protein interface, the *pack* designs clearly outperformed those generated by the *no_target* simulation, when quantified by a per-position score (**Fig 2.4H**); meaning that amino-acids productive for binding interactions were sampled more often. This benchmark provides an example of the benefits of using a "functional constraint" (binding target) to improve the quality of the sequences obtained by computational design.

Overall, the BINDI benchmark provided important insights regarding the best FunFolDes protocol to improve the design of functional proteins.

Repurposing a naturally occurring fold for a new function

To further test FunFolDes's design capabilities, we sought to transplant a contiguous viral epitope that is recognized by a monoclonal antibody with high affinity (**Fig 2.5A**). For this design, we used the RSVF site II epitope (PDB ID: 3IXT (*220*)) as the functional motif. This epitope adopts a helix-loop-helix conformation recognized by the antibody motavizumab (mota) (*220*). In previous work we have designed proteins with this epitope, but started from a structurally similar template, where the RMSD between the epitope and the scaffold segment was approximately 1 Å over the helical residues. Here, we sought to challenge FunFolDes by using a distant structural template where the local RMSDs of the epitope and

the segment onto which it was transplanted were higher than 2 Å. We used MASTER (*221*) to perform the structural search (detailed description in Materials and Methods) and selected as template scaffold the structure of the A6 protein of the Antennal Chemosensory system from the moth *Mamestra brassicae* (PDB ID: 1KX8 (*222*))(**Fig 2.5A**). The backbone RMSD between the conformation of the epitope and the insertion region in 1kx8 is 2.37 Å (**Fig 2.5B**).

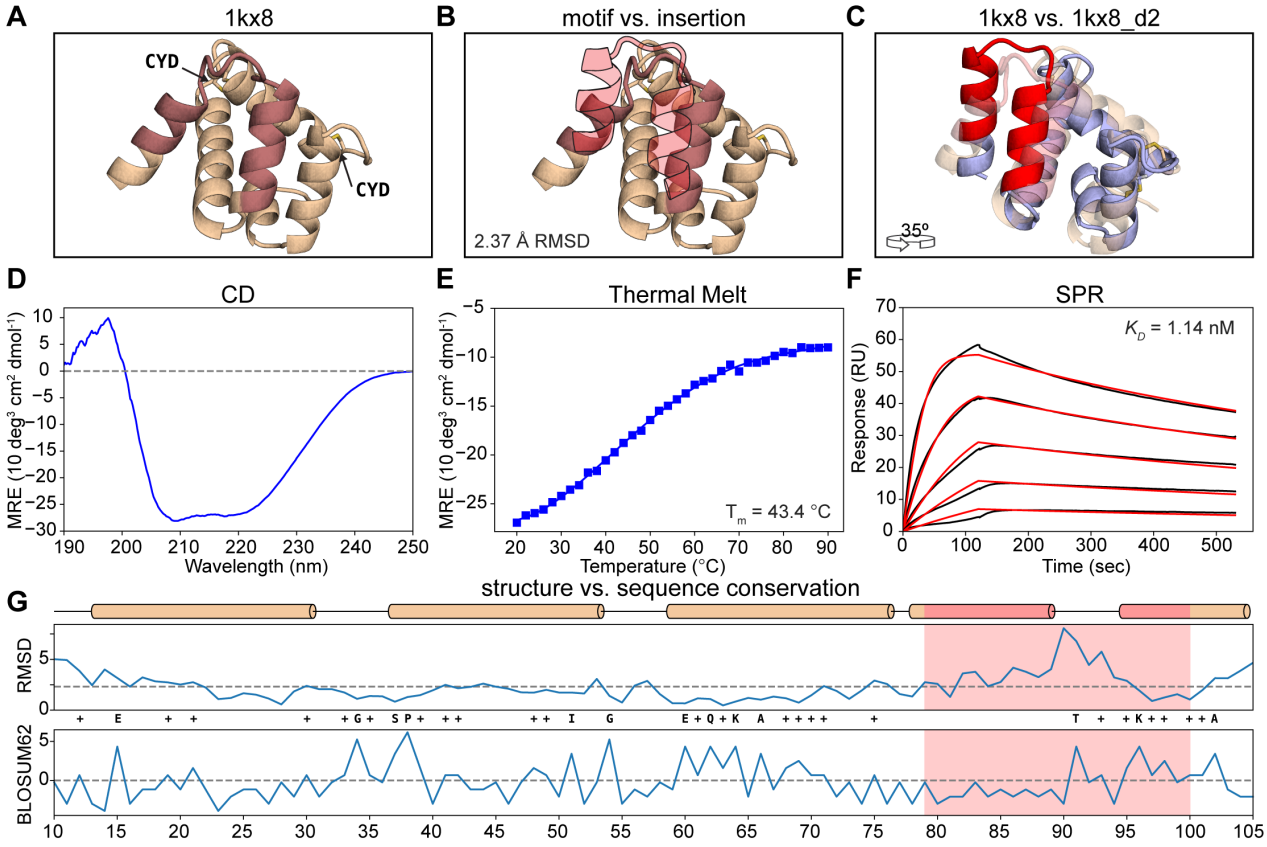


Figure 2.5. Functional design of a distant structural template.

A) Structural representation of 1kx8. The insertion region is colored in light red and the two disulfide bonds are labeled (CYD). B) Structural comparison between the insertion region of 1kx8 and the site II epitope (light red-filled silhouette). The local RMSD between the two segments is 2.37 Å. C) Superposition between 1kx8_d2 design model (blue with red motif) and the 1kx8 template (wheat and light red insertion site). Multiple conformational shifts are required throughout the structure to accommodate the site II epitope. D) CD spectrum of 1kx8_d2 showing a typical alpha-helical pattern with the ellipticity minima at 208 nm and 220 nm. E) 1kx8_d2 shows a melting temperature (*223*) of 43.4 °C. F) Binding affinity determined by SPR. 1kx8_d2 shows a K_D of 1.14 nM. Experimental sensorgrams are shown in black and the fitted curves in red. G) Per-position evaluation of structural (top) and sequence (*110*) divergence between the design model 1kx8_d2 and the starting template 1kx8. The largest structural differences are observed in the epitope insertion region, the overall difference of the two structures is 2.25 Å (dashed line). The sequence was evaluated using the BLOSUM62 score matrix, yielding a total of 13.5% identity and 38.5% similarity. The epitope region is colored in light red. Identical positions between the 1kx8_d2 and 1kx8 are labeled with the residue one letter code, while positively scored changes are labeled with plus (+).

The A6 protein is involved in chemical communication and has been shown to bind to fatty-acid molecules with hydrophobic alkyl chains composed of 12-18 carbons. Two prominent features are noticeable in the structure: two disulfide bonds (**Fig 2.5A**) and a considerable void volume in the protein core (**Fig.**

S2.3), thought to be the binding site for fatty acids. These features emphasize that the initial design template is likely not a very stable protein.

In the design process we performed two stages of FunFolDes simulations to obtain a proper insertion of the motif in the topology (**Fig 2.5C**). A detailed description of the workflow and metrics used for selection (**Fig. S2.3**) can be found in the Materials and Methods. A striking feature of our designs, when compared to the starting template, is that they had a much lower void volume, showing that FunFolDes generated structures and sequences that yielded well packed structures (**Fig. S2.3**).

We started by testing experimentally seven designs. Those that expressed in bacteria were further characterized using size exclusion chromatography coupled to a multi-angle light scatter (SEC-MALS) to determine the solution oligomerization state. To assess their folding and thermal stability (*223*) we used Circular Dichroism (CD) spectroscopy, and finally to assess their functional properties we used surface plasmon resonance (SPR) to determine binding dissociation constants (K_Ds) to the mota antibody. Out of the seven designs, six were purified and characterized further. The majority of the designs were monomers in solution and showed CD spectra typical of helical proteins. Regarding, thermal stabilities we obtained designs that were not very stable and did not unfold cooperatively (1kx8_02), however we also obtained very stable designs that did not fully unfold under high temperatures (1kx8_07) (**Fig. S2.4**).

The determined binding affinities to mota ranged from 34 to 208 nM, which was an encouraging result. Nevertheless, compared to the peptide epitope (K_D = 20 nM) and other designs previously published (K_D = 20 pM) (*178*), there was room for improvement. Therefore, we generated a second round of designs to attempt to improve stability and binding affinities.

Driven by the observation that the native fold has two disulfide bonds, in the second round, we tested eight designed variants with different disulfide bonds and, if necessary, additional mutations to accommodate them. The disulfide bonded positions were selected according to the spatial orientation of residues in the designed models, with most of the disulfide bonds being placed at distal locations from the epitope (>20 Å). All eight designs were soluble after purification and two were monomeric: 1kx8_d2 and 1kx8_3_d1, showing CD spectra typical of helical proteins (**Fig 2.5D**) with melting temperatures (T_ms) of 43 and 48°C (**Fig 2.5E**), respectively. Remarkably, 1kx8_d2 showed a K_D of 1.14 nM (**Fig 2.5F**), an improvement of approximately 30-fold compared to the best variants of the first round. 1kx8_d2 binds to mota with approximately 20-fold higher affinity than the peptide-epitope ($K_D \approx$ 20 nM), and 50-fold lower compared to previously designed scaffolds (K_D = 20 pM) (*178*). This difference in binding is likely reflective of how challenging it can be to accomplish the repurposing of protein structures with distant structural similarity.

Post-design analyses were performed to compare the sequence and structure of the best design model with the initial template. The global RMSD between the two structures is 2.25 Å. Much of the structural variability arises from the inserted motif, while the surrounding segments adopt a configuration similar to the original template scaffold. The sequence identity of 1kx8_d2 as compared to the native protein is approximately 13%. The sequence conservation per-position (**Fig 2.5G**) was evaluated through the BLOSUM62 matrix, where positive scores are attributed to the original residue or favorable substitutions and negative if unfavorable. Overall, 38.5% of the residues in 1kx8_d2 scored positively, and 61.4% of the residues had a score equal or lower than 0. This is particularly interesting, from the perspective that several residues, unfavorable according to BLOSUM62, yielded well folded and functional proteins. To further substantiate our experimental results, we performed structure prediction simulations of the designed sequences, where we observed that 1kx8_d2 presents a higher folding propensity than the WT

protein (**Fig. S2.5A**). To evaluate if the predicted models presented the correct epitope conformation, we performed docking simulations and observed that they obtained lower binding energies than the native peptide-antibody complex, within similar RMSD fluctuations (**Fig. S2.5A**).

The successful design of this protein is a relevant demonstration of the broad usability of FunFolDes and the overall strategy of designing functional proteins by coupled folding and design to incorporate functional motifs in unrelated protein scaffolds.

Functionalization of a functionless fold

Advances in computational design methodologies have achieved remarkable results in the design of *de novo* protein sequences and structures (*181, 182, 185*). However, the majority of the designed proteins are "functionless" and were designed to test the performance of computational algorithms for structural accuracy. Here, we sought to use one of the hallmark proteins from *de novo* design efforts – TOP7 (*187*) (**Fig 2.6A**) – and functionalize it using FunFolDes. The functional site selected to insert into TOP7 was a different viral epitope from RSVF, site IV, which is recognized by the 101F antibody (*224*). When bound to the 101F antibody, site IV adopts a β-strand-like conformation (**Fig 2.6B**), which in terms of secondary structure content is compatible with one of the edge strands of the TOP7 topology (**Fig 2.6C**). Despite the secondary structure similarity, the RMSD of the site IV backbone in comparison with TOP7 is 2.1 Å over 7 residues, and the antibody orientation in this particular alignment reveals steric clashes with TOP7. Therefore, this design challenge is yet another prototypical application for FunFolDes, and we followed two distinct design routes: I) a conservative approach where we fixed the amino-acid identities of roughly half of the core of TOP7 and allowed mutations mostly on the contacting shell of the epitope insertion site; and II) a sequence unconstrained design where all the positions of the scaffold were allowed to mutate. We attempted five designs for recombinant expression in *E. coli* and two (TOP7_full and TOP7_partial) were selected for further biochemical and biophysical characterization, one from each of the two design strategies mentioned above. According to SEC-MALS, both behaved as monomers in solution, with TOP7_partial showing higher aggregation propensity. Both TOP7_full and TOP7_partial (**Fig. S2.6**) were folded according to CD measurements. TOP7_full showed a CD spectrum (**Fig 2.6D**) very similar to that of native TOP7 (*187*). We observed that TOP7_full was much less stable than the original TOP7 (**Fig 2.6E**)(T_m = 54.5 °C). To quantify the functional component of TOP7_full, we determined a K_D of 24.2 nM with 101F (**Fig 2.6F**), within the range measured for the native viral protein RSVF (3.6 nM) (*224*). Importantly, the K_D for TOP7_full is 2400 fold lower than that of the peptide-epitope (58.4 μM) (*224*), suggesting that productive conformational stabilization and/or extra contacts to the scaffold were successfully designed.

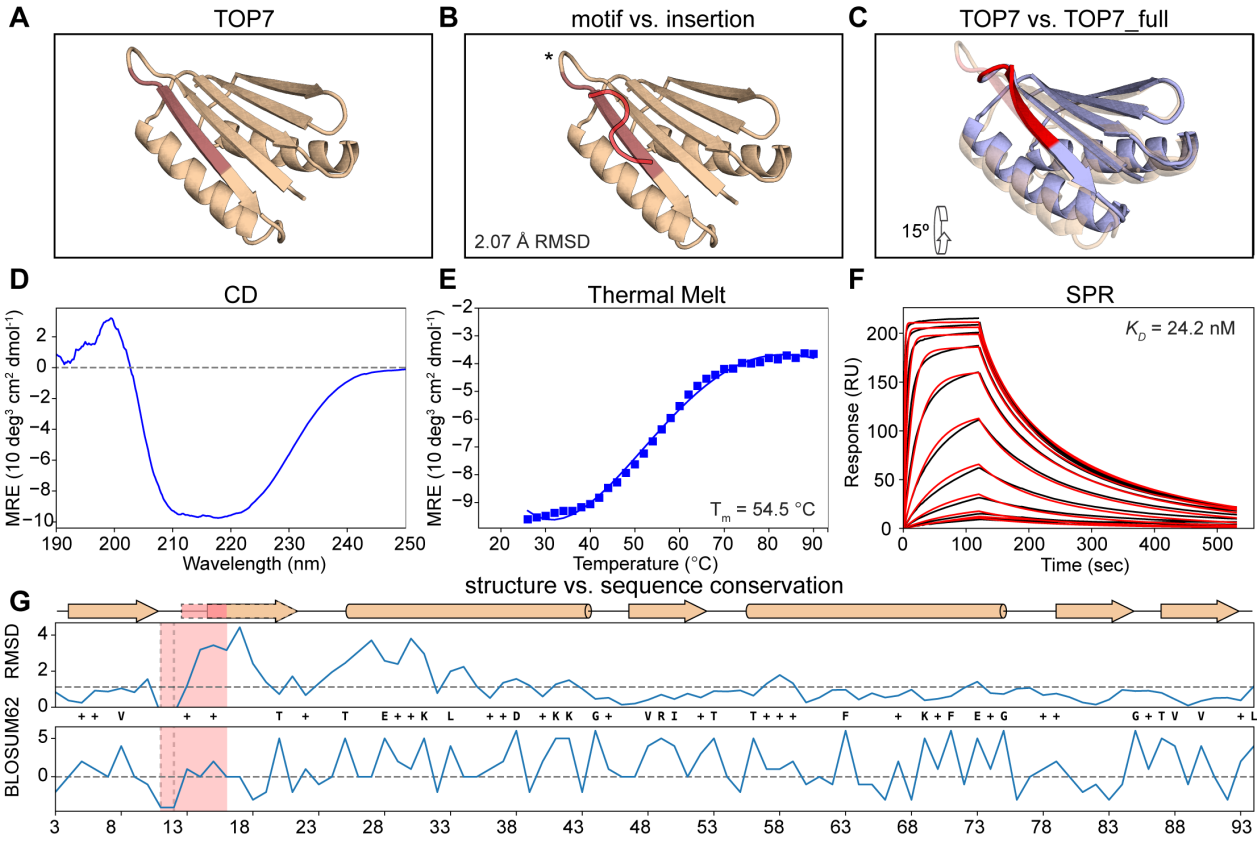


Figure 2.6. Functionalization of the functionless *de novo* fold TOP7.

A) Structure of TOP7 with the insertion region highlighted in light red. B) Structural comparison between 101F and TOP7's insertion region shows a 2.1 Å RMSD. C) TOP7_full model (in blue and red for the motif) superimposed over the TOP7 crystal structure. 101F's insertion is structurally compensated mostly by the first pairing beta strand and a shift of the first alpha helix. D) CD spectrum shows a broad ellipticity signal between 210 nm and 222 nm as a representative of mixed alpha and beta secondary structures. E) The T_m for TOP7_full was 54.5 °C. F) Binding affinity determined by SPR. TOP7_full shows a K_D of 24.2 nM. Experimental sensorgrams are shown in black and the fitted curves in red. G) Per-position evaluation of structural (top) and sequence (*110*) divergence between the design model TOP7_full and the starting template TOP7. The largest structural differences are observed in the region downstream of the site IV epitope, the overall difference of the two structures is 1.5 Å (dashed horizontal line). The connecting loop between the strand that holds the epitope and the adjacent strand was also shortened to obtain a tighter connection between the 2 strands (dashed vertical region). Sequence divergence is evaluated by applying the BLOSUM62 score matrix to the sequences, yielding a total of 27.7% identity and 52.2% similarity. The epitope region is colored in light red. Identical positions between the TOP7_full and TOP7 are displayed as their residue types while positively scored changes according to BLOSUM62 are labeled with a plus (+).

Per-residue structural similarity and sequence recovery were evaluated for TOP7_full against TOP7 (**Fig 2.6G**). Most conformational changes occur on the site IV insertion region and displacement of the neighboring alpha-helix, with the overall backbone RMSD being 1.5 Å.

Remarkably, the sequence identity of the most aggressive design (TOP7_full) is only 28 %, and using the BLOSUM62 based scoring system, we observe that most of the TOP7_full residues were actually favorable, obtaining positive scores. This low conservation is especially relevant considering that intensive studies on TOP7 have revealed the importance of beta-sheet conservation in order to keep its foldability (*196, 225, 226*). Sequence folding prediction experiments showed that TOP7_full has a similar folding

propensity to TOP7 and docking simulations also show lower binding energies as compared to the native peptide-antibody complex, reinforcing the experimental results obtained (**Fig. S2.5b**),

In summary, our results show that FunFolDes repurposed a functionless protein by folding and designing its structure to harbor a functional site, which in this case was a viral epitope. Previously, these computationally designed proteins with embedded viral epitopes were dubbed epitope-scaffolds and showed their medical applicability as immunogens that elicited viral neutralizing antibodies (*178*).

2.5 Discussion and Conclusions

The robust computational design of proteins that bear a biochemical function remains an important challenge for current methodologies. The ability to consistently repurpose old folds for new functions or the *de novo* design of functional proteins could bring new insights into the determinants necessary to encode function into proteins (e.g., dynamics, stability, etc.), as well as, important advances in translational applications (e.g., biotechnology, biomedical, biomaterials, etc.).

Here, we present Rosetta FunFolDes, that was conceived to embed functional motifs into protein topologies. This protocol allows for a global retrofitting of the overall protein topology to favorably host the functional motif and enhance the designability of the starting structural templates. FunFolDes has evolved to incorporate two types of constraints to guide the design process: topological and functional. The former entails the fragments to assemble the protein structure and sets of spatial constraints that bias the folding trajectories towards a desired topology; and the latter are the structure of the functional motif and the binding target.

Our methodological approach fills the gap between conservative grafting approaches where the structure of the host scaffold is mostly fixed (Rosetta Epigraft and Motifgraft (*193, 227*)) and the full *de novo* assembly of non-predefined protein topologies bearing functional motifs (*228, 229*). FunFolDes lies in between, by affording considerable structural flexibility to the host scaffold within the boundaries of its topology. In our view, FunFolDes is the most appropriate tool in situations where the structural mimicry of the functional motif is distant from the receiving scaffold's site and overall conformational adaptations are necessary to design viable protein structures and sequences.

We have extensively benchmarked FunFolDes, leveraging natural structural and sequence variation of proteins within the same fold, as well as deep mutational scanning data for the computationally designed protein BINDI (*195*). In our first benchmark, we observed that FunFolDes biases the sampling towards improved structural and sequence spaces. Improved sampling may contribute to solve some of the major limitations in protein design, related to "junk" sampling, where many designs are not physically realistic, exhibiting flaws according to general principles of protein structure. Importantly, higher quality sampling will likely contribute to improve the success rate of designs that are tested experimentally. The BINDI benchmark allowed us to test FunFolDes in a system with extensive experimental data, which included both sequences and structures. Perhaps the most interesting observation was that designs that were theoretically within a sequence/structure space productive for binding, were far from the energetic minimum accessible to the protein fold in the absence of the binding target. This observation resembles the stability-function tradeoffs that have been reported from *in vitro* evolution experimental studies (*218, 219*). The large majority of the design algorithms are energy "greedy" and the sequence/structure searches are performed with the central objective of finding the global minimum of the energetic landscape. By introducing functional constrains into the simulations, FunFolDes presents

an alternative way of designing functional molecules and skew the searches towards off-minima regions of the global landscape. We anticipate that such finding will be more relevant for protein scaffolds that need to undergo a considerable structural adaptation to perform the desired function. If confirmed that this finding is generalized across multiple design problems, it could be an important contribution for the field of computational protein design.

Furthermore, we used FunFolDes to tackle two design challenges and functionalized two proteins with two distinct viral epitopes generating synthetic proteins that could have important translational applications in the field of vaccine development. In previous applications, FFL always used three-helix bundles as design templates, here we diversified the template folds and used an all-helical protein that is not a bundle (1kx8) and a mixed alpha-beta protein (TOP7), clearly demonstrating the applicability to other folds. For the 1kx8 design series, we evaluated the capability of using distant structural templates as starting topologies as a demonstration of how to functionally repurpose many naturally occurring protein structures available. We obtained stable proteins that were recognized by an anti-RSV antibody with high affinity, showing that in this case, we successfully repurposed a distant structural template for a different function, a task for which other computational approaches (*76*) would have limited applicability. We see this result as an exciting step forward towards using the wealth of the natural structural repertoire for the design of novel functional proteins.

In a last effort, we functionalized a "functionless" fold, based on one of the first *de novo* designed proteins – TOP7. For us, this challenge has important implications to understand the design determinants and biochemical consequences of inserting a functional motif into a protein that was mainly optimized for thermodynamic stability. We were successful in functionalizing TOP7 differently than previous published efforts. Previously, TOP7 was mostly used as a carrier protein with functional motifs fused onto loop regions or side chains grafted in the helical regions (*196, 225, 226*), while our functional motif was embedded in the beta-sheet region. Exciting advances in the area of *de novo* protein design are also yielding many new proteins (*185-187*), which could then be functionalized with FunFolDes, highlighting the usefulness of this approach. Interestingly, we observed that the functionalized version of TOP7 showed a dramatic decrease in thermodynamic stability as compared to the parent protein. While this observation can be the result of many different factors, it is compelling to interpret it as the "price of function", meaning that to harbor function, TOP7 was penalized in terms of stability, which would be consistent with our findings in the BINDI benchmark and the experimental studies on stability-function tradeoffs.

Recently, there have also been several *de novo* proteins designed for functional purposes (*230*); however, these efforts were limited to linear motifs that carried the functions, and the functionalization was mainly accomplished by side chain grafting (*177, 179*), relying on screening a much larger number of designed proteins.

In the light of all the technical improvements, FunFolDes has matured to become a valuable resource for the robust functionalization of proteins using computational design. Here, we presented a number of important findings provided by the detailed benchmarks performed and used the protocol to functionalize proteins in design tasks that are representative of common challenges faced by the broad scientific community when using computational design approaches.

2.6 Materials and Methods

Computational protocol description

Rosetta Functional Folding and Design (FunFolDes) is a general approach for grafting functional motifs into protein scaffolds. Its main purpose is to provide an accessible tool to tackle specifically those cases in which structural similarity between the functional motif and the insertion region is low, thus expanding the pool of structural templates that can be considered useful scaffolds. This objective is achieved by folding the scaffold after motif insertion while keeping the structural motif static. This process allows the scaffold's conformation to change and properly adapt to the three-dimensional restrictions enforced by the functional motif. The pipeline of the protocol (summarized in **Fig 2.1**) proceeds as follows:

I) *Selection of the functional motif.* A single or multi-segment motif must be selected and provided as an input. In the most common mode of the protocol, dihedral angles, side chain identities and conformations are kept fixed throughout the whole protocol. Conserved sequence length between the motif and the insertion region is not required.

II) *Selection of the protein scaffold*. Searches for starting protein scaffolds can be achieved, but are not limited to, RMSD similarity matches to the Protein Data Bank (PDB) (*231*). The ability of FunFolDes to adapt the scaffold to the needs of the motif widens the structural space of what can be considered a suitable template. Thus, this step requires human intervention and is performed outside of the main protocol.

III) *Generation of fragment databases*. The usage of fragments lies at the core of many Rosetta protocols, particularly those that perform large explorations of the conformational space required for structure prediction and design. The most standard way of assembling fragment sets is to generate sequence-based fragments using the *FragmentPicker* application (*213*). While sequence-based fragments are critical in structure prediction problems, FunFolDes designs have a higher dependency from the structural content of the template rather than its sequence. Thus, we implemented the *StructFragmentMover*, a mover that performs on-the-fly fragment picking based on secondary structure, dihedral angles and solvent accessibility, calculated from the template's structural information. The typical three- and nine residue-long fragment sets are generated from the global fragment database included in the Rosetta tools release.

IV) *Generation of constraints.* Residue-pair distance and backbone dihedral angle constraints can be extracted from the protein scaffold to guide the folding process. These constraints may include the full-length protein or focus in specific segments while allowing a wider flexibility in other regions. Although not required, the use of constraints greatly increases the quality of the sampling. The protocol can be also made aware of other constraint types (such as cartesian constraints) by properly modifying the score functions applied to the *ab initio* stage (*232*).

V) *Construction of the extended pose*. The extended structure is composed of all the segments of the target motif maintaining their native backbone conformation and internal rigid body orientation. The scaffold residues are linearly attached to previously defined insertion points. In multi-segment motif scenarios, the construct will present a chain break between each of the motif-composing segments. This also allows for the segments to be placed into the design in non-consecutive sequence order. Details on how these chain breaks are created can be found in **S2.1Text**. Once the extended pose is assembled, it

is represented at the centroid level (all side-chain atoms in a single virtual atom) to reduce the computational cost of the simulation.

VI) *Folding the extended pose.* Fragment insertion is performed to accomplish the folding stage. Kinematics of the pose are controlled through the FoldTree (*233*), a system to control the propagation of the torsion angles applied to a structure. The procedure on how the FoldTree is build and exploited to maintain the appropriate position between different segments of the functional motif is detailed in **S2.1Text**. By default, the folding stage is allowed 10 trials to generate a decoy bellow a user defined RMSD threshold. In case the threshold is not reached, this trajectory is skipped and no design will be output.

VII) *Inclusion of the binding target*. If a binding target (protein, nucleic acid or small molecule ligand) is provided, a new FoldTree node is added to the closest residue between the first motif segment and each binding element. Similarly to the multi-segment kinematics, this ensures that the rigid-body orientation between the motif and its target is maintained. FunFolDes can handle simulations with both multi-segment and binding targets simultaneously.

VIII) *Folding post-processing*. Folding trajectories are considered successful if they generate structures under a user-defined RMSD threshold of the starting scaffold. In case of a multi-segment motif, a preliminary loop closure will be executed to generate a continuous polypeptide chain, and the kinematic setup maintained to avoid segment displacement during the design step. After the centroid folding stage, the full-atom information pose is recovered. All the steps necessary to perform the setup of the extended pose (kinematic setup, folding, post-processing) are carried out by a newly implemented mover called *NubInitioMover*.

IX) *Protein design and conformational relaxation*. The folded structure is subjected to iterative cycles of sequence design (*234*) and structural relaxation (*235*) in which the sequence search is coupled with confined conformational sampling (*236*). A MoveMap is defined to control backbone dihedrals and side chain conformations of the motif segments and the binding target while allowing for backbone and side-chain sampling of the movable residues (**S2.1Text**). TaskOperations are used to avoid undesired mutations in the functional motif.

X) *Loop closure*. If multi-segment motifs are used, a final loop closure step is required in order to obtain a polypeptide chain without breaks. The *NubInitioLoopClosureMover* performs this last step using the Cyclic Coordinate Descend (CCD) protocol (*233*), while ensuring that the original conformation and rigid-body orientation of the motifs is maintained. After the closure of each cut-point, a final round of fixed backbone design is performed on the residues of the cut-points and surroundings.

XI) *Selection, scoring and ranking*. Finally, the decoys are ranked and selected according to Rosetta energy, structural metrics (core packing, buried unsatisfied polar atoms, etc.) (*237*), sequence-based predictions such as secondary structure propensity (*238*) and folding propensity (*232*) or any other metrics accessible through RosettaScripts (RS). The *in* silico benchmarks and the design assessments in this work, we used the rstoolbox (*239*) to produce the statistical analysis and select the best-ranked decoys.

The pipeline components described here represent the most standardized version of the FunFolDes protocol. By means of its integration in RS, different stages can be added, removed or modified to tailor the protocol to the specific needs of the design problem at hand.

Capturing conformational and sequence changes in small protein domains

To test the ability of FunFolDes to recover the required conformational changes to stabilize a given structural motif, we created a benchmark of 14 target cases of proteins with less than 100 residues, named T01 to T14. Each target case was composed of two structures of the same CATH superfamily (*212*). One of the structures was representative of the shared structural features of the CATH family; we called this structure the reference. The second protein within each target case can present two types of structural variations with respect to the reference: I) an insertion or deletion (indel) region and II) a conformational change. Direct structural contacts between these two regions make it so that the indel region is likely the cause for the conformational change. We called this second structure the target (**Fig. 2.2**, **Table 2.1**).

ID	CATH	#	reference	target	motif range
T01	CATH.3.40.140.10	1	1pgxA	2pw9C	69-73
T02	CATH.3.30.310.50	1	3i3wA	4bjuA	464-486
T03	CATH.3.30.70.980	1	1lfpA	1mw7A	140-150
T04	CATH.3.30.70.100	1	1rjjA	1lq9A	19-45
T05	CATH.3.10.20.30	1	2q5wD	2pkoA	49-64
T06	CATH.2.30.29.30	1	1c1yB	1h4rA	39-59
T07	CATH.3.10.20.90	1	2bkfA	2al6B	115-119
T08	CATH.3.10.20.90	1	1wj4a	1wiaA	181-200
T09	CATH.3.10.20.90	1	3ny5B	3phxB	100-121
T10	CATH.3.10.20.310	1	2x8xX	2qdfA	103-121
T11	CATH.3.10.320.10	1	4p5mA	2bc4C	56-66
T12	CATH.2.40.40.20	1	1cr5B	2pjhB	119-142
T13	CATH.2.40.40.20	2	1cr5B	2pjhB	119-142, 168-173
T14	CATH.3.30.110.40	1	1jdqA	3lvjC	14-37

Table 2.1. Targets included in the conformational and sequence recovery benchmark.

For each of the benchmark targets is indicated the CATH superfamily and representatives used in the simulations. (#) indicates the number of segments in the target protein that are considered motif. Motif range indicates the residues considered motif according to the PDB numbering.

For each template protein we generated approximately 10000 decoys with FunFolDes by folding the target with the following conditions: 1) the indel region was considered as the motif, meaning that its structural conformation was kept fixed and no mutations allowed; 2) residue-pair distance constraints were derived from the secondary structure elements conserved between reference and the target (constrained region); 3) the region of the protein which showed the largest structural variations (query region) was constraint-free throughout the simulation.

FunFolDes simulations were compared with constrained *ab initio (*cst-*ab initio)* simulations, the key difference being that the cst-*ab initio* simulations allowed for backbone flexibility in the motif region. The comparison between both approaches provides insights on the effects of a static segment in the folding trajectory of the polypeptide chain. In both scenarios a threshold was set after the folding stage where only decoys that had less than 5 Å RMSD from the template were carried to the design stage.

The importance of the input fragments was assessed in our benchmark. Both protocols were tested with sequence-based fragments from *FragmentPicker* and structure-based fragments generated on-the-fly by FunFolDes. Comparison between the two types of fragments provides insight into how to utilize FunFolDes in the most productive manner.

Structural recovery was evaluated by RMSD with the target structure. Global RMSD, understood as the minimum possible RMSD given the most optimal structural alignment, was used to assess the overall structural recovery of each decoy population. Local RMSD, was evaluated for the unconstrained (query) region and the motif by aligning each decoy to the template through the constrained segments (excluding the motif). This metric aimed to capture the specific conformational changes required to accommodate the motif into the structure (**Fig. 2.2B**, **Fig. S2.1B**).

Sequence recovery was evaluated through two different criteria, sequence associated statistics and Hidden Markov Model (HMM) (*215*). For the sequence associated statistics, we quantified sequence identity and similarity according to BLOSUM62 for the core residues of each protein, as defined by Rosetta's *LayerSelector* (*181*). Motif residues, that were not allowed to mutate, were excluded from the statistics. In the second criteria, position specific scoring matrices with inter-position dependency known as Hidden Markov Model (HMM) were used to evaluate fold specific sequence signatures. In this case, the closest HMM to the template structure provided by CATH was used to query the decoys and identify those that matched the HMM under two conditions: I) an e-value under 10 and II) a sequence coverage over 50%. Although these conditions are wide, they were within the ranges found between members of CATH superfamilies with high structural and sequence variability like the ones used in the benchmark.

Target-biased design of protein binders

To assess the performance of FunFolDes in the presence of a binding target we recreated the design of BINDI as a binder for BHRF1 (*195*), the BHRF1 binding motif from the BIM-BH3 protein (PDB ID:2WH6 (*240*)) was inserted into a previously described 3-helix bundle scaffold (PDB ID:3LHP (*177*)).

Four different design simulations were performed, one without the binder (*no_target*) and three in the presence of the binder (*static*, *pack* and *packmin*). The difference between the last three relates to how the binding target was handled. In the *static* simulations the binding target was kept fixed and no conformational movement in the side chains was allowed throughout the protocol. In the *pack* simulations the side chains of the binding target were repacked during the binder design stage. Finally, in the *packmin* simulations the binding target side-chains were allowed to repack and both side-chains and backbone were subjected to minimization. In all cases, the two terminal residues on each termini of the binding motif were allowed backbone movement to optimize the insertion in the 3-helix bundle scaffold. For each of these simulations, approximately 20000 decoys were generated.

For the *no_target* simulations the FunFolDes designs were docked to BHRF1 using the inserted motif as guide to assess their complementarity and interface metrics. In all the simulations, a final round of global minimization was performed, where both proteins of the complex were allowed backbone flexibility. During this minimization, the rigid-body orientation between the design and target was kept fixed. The final ΔΔG of the complexes was measured after the minimization step to enable comparisons between the *no_target* decoys and the remaining simulation modes. Structural changes related to this minimization step were evaluated as the global RMSD between each structure before and after the process, this measure is referred to as RMSD drift.

Structural evaluation includes global RMSD against the BINDI crystal structure (PDB ID: 4OYD (*195*)) as well as local RMSDs against regions of interest in BINDI. In the Local-RMSD calculations the structures were aligned through the inserted motif, as its conformation and orientation relative to BINDI were kept fixed throughout all simulations. The local RMSD analysis was performed over all the helical segments contained in the structures (all H), which provided a measurement of the structural shifts on the secondary structure regions of the designs.

To evaluate the sequence recovery we leveraged BINDI's saturation mutagenesis data analyzed by deep sequencing performed by Procko *et al*. (*195*). The experimental fitness of each mutation was summarized in a score matrix where a score was assigned to each amino-acid substitution for the 116 positions of the protein (**Fig. S2.2C**). In summary, point mutations that improved BINDI's binding to BHRF1 are assigned positive scores while deleterious mutations present negative values. These scores were computed based on experimental data where the relative populations of each mutant were compared between a positive population of cells displaying the designs (binders) and negative populations (mutants that display but don't bind), these experiments have been described in detail elsewhere (*195*). Upon normalization by the BINDI sequence score, a position sequence specific matrix (PSSM) was created. Like the original data, this matrix also assigns a positive score to each point mutation if it resulted in an improved binding for the design. This normalization provides a score of 0 for the BINDI sequence, which is useful as a reference score.

Repurposing naturally occurring folds for a new functions

To experimentally validate the capabilities of FunFolDes and insert functional sites in structurally distant templates, we grafted the 11 residues from the site II epitope from the Respiratory Syncytial Virus (RSV) protein F (PDB ID:3IXT (*220*)), residues 256 to 276 in chain P (NSELLSLINDMPITNDQKKLMSN), into heterologous scaffolds. This is a continuous, single segment, helix-loop-helix conformation epitope. The main objective was to challenge the capabilities of FunFolDes to reshape the structure of the scaffold to the requirements of the functional motif. We searched for insertion segments with RMSDs towards the site II structure higher than 2 Å.

The structural searches were performed using MASTER (*221*) where we used the full-length site II segment as a query against a subset of 17539 protein structures from the PDB, composed of 30% non-redundant sequences included in the MASTER distribution. The RMSD between the query and segments on the scaffolds were assessed using backbone C_αs. All matches with $RMSD_{C\alpha} < 5.5$ Å relative to site II were further filtered by protein size, where only proteins between 50 and 100 residues were kept. These scaffolds were then ranked regarding antibody-binding compatibility, where each match was realigned to the antibody–epitope complex and steric clashes between all glycine versions of the scaffold and antibody were quantified using Rosetta. All matching scaffolds with ΔΔG values above 100 REU were discarded under the assumption that their compatibility with the antibody binding mode was too low. The remaining scaffolds were visually inspected and PDB ID: 1kx8 (*222*) ($RMSD_{C\alpha}$ = 2.37 Å) was selected for design with FunFolDes. The twenty-one residues from the site II epitope (motif) as present in 3IXT were grafted into a same sized segment (residues 79-100) of 1kx8 using the *NubInitioMover*. Up to three residues in each insertion region of the motif were allowed backbone flexibility in order to model proper conformational transitions in the insertion points. Atom pair constraints with a standard deviation of 3 Å were defined for all template residues, leaving the motif segment free of constraints. The generous standard deviation was defined to allow for necessary conformational changes to retrofit the motif

within the topology. The total allowed deviation from the template was limited to 5 Å to ensure the retrieval of the same topology. In this design series we used sequence-based fragments generated with the 1kx8 native sequence. Three cycles of design/relax were performed on the template residues with the *FastDesignMover*.

A first generation of 12500 designs was ranked according to Rosetta energy. From the top 50 decoys, only one presented the motif without distortions on the edges derived from the allowed terminal flexibility. This decoy was used as template on the second generation of FunFolDes to enhance the sampling of properly folded conformations, with the same input conditions as before.

In the second generation, the top 50 decoys according to Rosetta energy were further optimized through additional cycles of design/relax. The final designs were again selected using a composite filter based on Rosetta energy (top 50), buried unsatisfied polar atoms (<15), cavity volume (< 75 $Å^3$) and we obtained a final set of 15 candidates from which we prioritized 6 upon the inspection of the computational models. In addition, we also quantified the secondary structure prediction using PSIPRED (*238*), all the tested designs had more 65% (ranging from 65% to 92%) of the residues with correct secondary structure prediction. The final designs were manually optimized, this process entailed the removal of designed hydrophobic residues in solvent exposed positions, in this designs series we performed between 2 and 4 mutations obtaining 7 designs from the previous 6. After the initial characterization, designs with added disulfide bridges were generated to improve protein stability and affinity (**Fig. S2.3**, **Fig. S2.4**). To do so, we use the Rosetta *DisulfidizeMover*, which screened the designed models for pairs of residues with favourable three-dimensional orientations to host disulfide bonds. Upon the placement of the disulfide bond, the neighbouring residues within 10 Å of the disulfide, were designed to optimize the residue interactions and improve the packing of the designed region.

Functionalization of a functionless fold

In a second effort to test the design capabilities of FunFolDes we sought to insert a functional motif in one of the first *de novo* designed proteins – TOP7 (PDB ID: 1QYS (*187*))

Six residues from the complex between the antibody 101F and the peptide-epitope, corresponding to residues 429-434 in chain P (RGIIKT) on the full-length RSV F protein (*224*), were grafted into the edge strand of the TOP7 backbone using FunFolDes. The choice between epitope and hosting scaffold was made based on the secondary structure adopted by the epitope and the structural compatibility of TOP7, the $RMSD_{C\alpha}$ between the epitope an the insertion segment was 2.07 Å.

To ensure that the majority of the β-strand secondary structure was maintained throughout the grafting protocol, the epitope motif was extended by one residue and a designed 4-residue β-strand (KVTV) pairing with the backbone of the C-terminal epitope residues was co-grafted as a discontinuous segment into the adjacent strand of the TOP7 backbone. With this strategy we circumvented a Rosetta sampling limitation, where often times extensive sets of constraints to achieve backbone hydrogen bonds on beta-strands are necessary (*182*). After defining the motif consisting of the epitope plus the pairing strand and the sites of insertion on the TOP7 scaffold, FunFolDes was used to graft the motif.

Backbone flexibility was allowed for the terminal residues of the functional motif and a β-turn connection between the two strands was modelled during the folding process (*NubInitioMover*). During the folding process, the 101F antibody was added to the simulation in order to limit the explored

conformational space productive for binding. Finally, the *NubInitioLoopClosureMover* was applied to ensure that a proper polypeptide chain was modelled and no chain-breaks remained, a total of 800 centroid models were generated after this stage. Next, we applied an RMSD filter to select scaffolds with similar topology to TOP7 (< 1.5 Å) and a hydrogen bond long-range backbone score (HB_LR term) to favour the selection of proteins with proper beta-sheet pairing. The top 100 models according the HB_LR score and <1.5 Å to TOP7, were then subjected to an iterative sequence-design relax protocol, alternating fixed backbone side-chain design and backbone relaxation using the *FastDesignMover*. Two different design strategies were pursued: I) partial design - amino acid identities of the C-terminal half of the protein (residues 45 through 92) were retained from TOP7 while allowing repacking of the side chains and backbone relaxation; II) full-design - the full sequence space in all residues of the structure (with the exception of the 101F epitope) was explored. No backbone or side chain movements were allowed in the 6-residue epitope segment whereas the adjacently paired β-strand was allowed to both mutate and relax. Tight Cα atom-pair distance constraints (standard deviation of 0.5 Å) were used to restrain movements of the entire sheet throughout the structural relaxation iterations.

From the 100 designs generated, only those that passed a structural filter requiring 80% beta-sheet secondary structure composition after backbone relaxation were selected for further analysis.

The 93 designs passing this filter were evaluated with a composite filter based on REU score (Top 50), hydrogen-bond long-range backbone interactions (< -113) and core packing (> 0.7). The selected designs were finally submitted to human-guided optimisation to correct for hydrophobic residues that were designed in solvent exposed positions (1-3) and shortening of the connecting loop between the two inserted strands using the Rosetta Remodel application (*241*).

Interestingly, in an attempt to reproduce the same grafting exercise with *MotifGraftMover* (*76*), this resulted in non-resolvable chain breaks when trying to graft either the two segment-motif or the epitope alone into the TOP7 scaffold.

Protein Expression and Purification

DNA sequences of the designs were purchased from Twist Bioscience. For bacterial expression the DNA fragments were cloned via Gibson cloning into a pET21b vector containing a C-terminal His-tag and transformed into *E. coli* BL21(DE3). Expression was conducted in Terrific Broth supplemented with ampicillin (100 µg/ml). Cultures were inoculated at an OD_{600} of 0.1 from an overnight culture and incubated at 37°C with a shaking speed of 220 rpm. After reaching OD_{600} of 0.7, expression was induced by the addition of 1 mM IPTG and cells were further incubated for 4-5h at 37°C. Cells were harvested by centrifugation and pellets were resuspended in lysis buffer (50 mM TRIS, pH 7.5, 500 mM NaCl, 5% Glycerol, 1 mg/ml lysozyme, 1 mM PMSF, 1 µg/ml DNase). Resuspended cells were sonicated and clarified by centrifugation. Ni-NTA purification of sterile-filtered (0.22 µm) supernatant was performed using a 1 ml His-Trap™ FF column on an ÄKTA pure system (GE healthcare). Bound proteins were eluted using an imidazole concentration of 300 mM. Concentrated proteins were further purified by size exclusion chromatography on a Superdex™ 75 300/10 GL or a Hiload 16/600 Superdex™ 75 pg column (GE Healthcare) using PBS buffer (pH 7.4) as mobile phase.

For IgG expression, heavy and light chain DNA sequences were cloned separately into pFUSE-CHIg-hG1 (InvivoGen) mammalian expression vectors. Expression plasmids were co-transfected into HEK293-F cells in FreeStyle™ medium (Gibco™) using polyethylenimine (Polysciences) transfection.

Supernatants were harvested after 1 week by centrifugation and purified using a 5 ml HiTrap™ Protein A HP column (GE Healthcare). Elution of bound proteins was accomplished using a 0.1 M glycine buffer (pH 2.7) and eluents were immediately neutralized by the addition of 1 M TRIS ethylamine (pH 9). The eluted IgGs were further purified by size exclusion chromatography on a Superdex™ 200 10/300 GL column (GE Healthcare) in PBS buffer (pH 7.4). Protein concentrations were determined by measuring the absorbance at 280 nm using the sequence calculated extinction coefficient on a Nanodrop (Thermo Scientific).

Circular Dichroism (CD)

Far-UV circular dichroism spectra of designed scaffolds were collected between a wavelength of 190 nm to 250 nm on a Jasco J-815 CD spectrometer in a 1 mm path-length quartz cuvette. Proteins were dissolved in PBS buffer (pH 7.4) at concentrations between 20 µM and 40 µM. Wavelength spectra were averaged from two scans with a scanning speed of 20 nm min^{-1} and a response time of 0.125 sec. The thermal denaturation curves were collected by measuring the change in ellipticity at 220 nm from 20 to 95 °C with 2 or 5 °C increments.

Size-exclusion Chromatography combined with Multi-Angle Light-Scattering (SEC-MALS)

Multi-angle light scattering was used to assess the monodispersity and molecular weight of the proteins. Samples containing between 50 -100 µg of protein in PBS buffer (pH 7.4) were injected into a Superdex™ 75 300/10 GL column (GE Healthcare) using an HPLC system (Ultimate 3000, Thermo Scientific) at a flow rate of 0.5 ml min^{-1} coupled in-line to a multi-angle light scattering device (miniDAWN TREOS, Wyatt). Static light-scattering signal was recorded from three different scattering angles. The scatter data were analyzed by ASTRA software (version 6.1, Wyatt)

Surface Plasmon Resonance (SPR)

To determine the dissociation constants of the designs to the mota or 101F antibodies, surface plasmon resonance was used. Experiments were performed on a Biacore 8K at room temperature with HBS-EP+ running buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 3mM EDTA, 0.005% v/v Surfactant P20) (GE Healthcare). Approximately 1200 response units of mota or 101F antibody were immobilized via amine coupling on the methylcarboxyl dextran surface of a CM5 chip (GE Healthcare). Varying protein concentrations were injected over the surface at a flow rate of 30 µl/min with a contact time of 120 sec and a following dissociation period of 400 sec. Following each injection cycle, ligand regeneration was performed using 3 M $MgCl_2$ (GE Healthcare). Data analysis was performed using 1:1 Langmuir binding kinetic fits within the Biacore evaluation software (GE Healthcare).

2.7 Supporting information

S2.1 Text: FoldTree and MoveMap

FoldTree: Basics

In Rosetta, the FoldTree (*233*) is a graph representation of the connectivity of a structure that controls its kinematics through the propagation of changes to the torsion angles applied to the structure. An extensive explanation of the FoldTree can be found in the official Rosetta documentation:

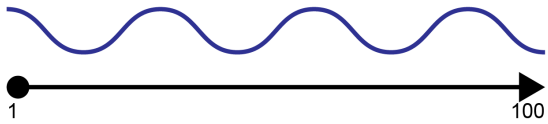
https://www.rosettacommons.org/demos/latest/tutorials/fold_tree/fold_tree.

For the purposes of this work, the following properties are important to highlight:

1. A FoldTree has to cover the totality of the structure. This includes all protein chains and, if applicable, small-molecules.
2. FoldTrees are not cyclic and start at a given node dubbed as **root**.
3. A FoldTree contains two main types of connectivity: **peptide edges** (indicated with the value -1), that represent peptide connections through which the angle torsions are propagated, and **jumps** (indicated with consecutive positive values), that represent non polymeric connections between different points of the fold tree and that translate in the 3D space to the maintenance of defined rigid body orientations between the segments they connect.
4. The ends on a continuous stretch of peptide edges are labeled as **cutpoints**. These could represent either **chain ends** (like N- and C-terminal) or **chain breaks**, discontinuities inside a single chain that avoid the transfer of torsion changes.

Thus, a typical FoldTree for a 100-residue protein would by default look like this:

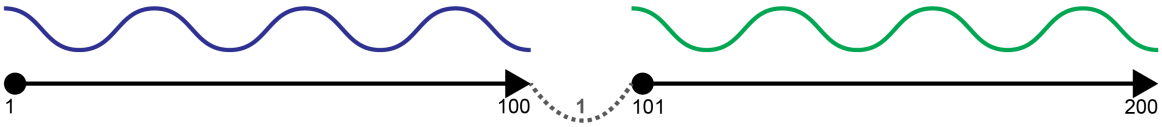
```
FOLD_TREE EDGE 1 100 -1
```



Default FoldTree. By default a FoldTree (*242*) will transfer the torsion effects from the first to the last residue of a given protein structure (blue).

The FoldTree for two bound 100-residue proteins would be as follows:

```
FOLD_TREE EDGE 1 100 -1  EDGE 100 101 1  EDGE 101 200 -1
```



FoldTree for two bound proteins. When two proteins (blue and green) form a complex, a jump between the two chains is defined, maintaining the rigid body orientation between the two proteina and creating a FoldTree to describe the full complex.

MoveMap: Basics

The MoveMap (*233*) definition controls the degrees of freedom of each individual residue. The degrees of freedom can be controlled at two levels backbone (**BB**: True/False), that defines whether or not backbone angles (ϕ, ψ) can be altered; side-chain (**CHI**: True/False) that defines the ability of the side chain angles (χ_n) to be changed.

Thus, a flexible backbone relaxation with full mobility on a 100-residue protein will be defined as:

```
RESIDUE 1 100 BBCHI
```

While a fixed backbone with only side chain repacking would be described as:

```
RESIDUE 1 100 CHI
```

In combination with the FoldTree, there is a third variant definition in the MoveMap: the **JUMP** (True/False), defines the ability for the residues at both sides of a FoldTree jump to change their relative positions.

For example, given our previous two-protein FoldTree:

```
FOLD_TREE EDGE 1 100 -1  EDGE 100 101 1  EDGE 101 200 -1
```

Applying structure relaxation with a MoveMap such as

```
RESIDUE 1 100 BBCHI

RESIDUE 101 200 BBCHI

JUMP 1 YES
```

will allow for the proteins to change their rigid body orientation to each other, while

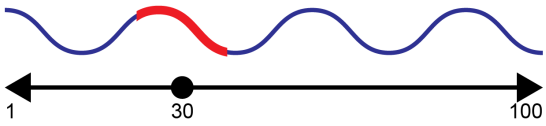
```
RESIDUE 1 100 BBCHI

RESIDUE 101 200 BBCHI

JUMP 1 NO
```

will keep that orientation fixed.

FunFolDes: Single-segment motif

The most straightforward application of FunFolDes is the grafting of a single-segment motif. Given that we want to insert an 11-residue motif between positions 25 to 35 in a 100-residue protein. The final FoldTree will look like this:

```
FOLD_TREE EDGE 30 1 -1  EDGE 30 100 -1
```



Single-segment motif FunFolDes design. The root of the FoldTree is placed in the middle of the motif (red) and expands towards both ends of the protein.

While the default MoveMap will be:

```
RESIDUE 1 24 BBCHI

RESIDUE 25 35 NO

RESIDUE 36 100 BBCHI
```

In this case, the root of the FoldTree is placed in the middle of the grafted segment and expands towards both N- and C-terminus of the protein. Combined with the MoveMap restrictions, this setup allows for the segments at each side of the motif to refold while the motif stays static.

FunFolDes: Multi-segment motif

A second level of complexity is the grafting of multi-segment functional motifs, i.e., of non-contiguous structural fragments.

In this scenario, a continuous tree in which torsion changes are propagated in the region between the two insertions would result in changes regarding the rigid body orientation between the individual structural motifs, thus interfering in the proper mimicry of the full functional motif. Therefore, a **cut-point** has to be introduced in this segment to avoid this propagation of torsional changes. As this has to happen between each pair of inserted segments, the number of chain breaks in the FoldTree is:

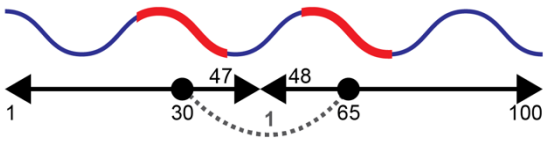
$$count_{chainbreaks} = count_{segments} - 1 \qquad \text{Eq1}$$

To select the appropriate region where to generate the cutpoint, FunFolDes uses the secondary structure assignment provided by Rosetta's internal implementation of DSSP (*243*) to locate a loop region between the two motif segments. If found, the midpoint of the loop is selected as the cutpoint. When no loop is found, the cutpoint is set in the middle residue between the two segments.

Once the cutpoints have been chosen, a FoldTree root is placed in the middle of each grafted motif segment and expanded in both directions towards their flanking cutpoints (here N- and C-terminus are considered as cutpoints, although they are not chain breaks). Finally, jumps span from the root of the first segment (in sequential order) towards the rest of the roots of the other segments, creating a fully connected FoldTree.

As an example, assuming that we want to graft two segments into a 100-residue protein, one between residues 25 and 35 and another between residues 60 and 70. Following the previous explanation, and premising a loop region in residues 45-50, the FoldTree is defined as:

```
FOLD_TREE EDGE 30 1 -1  EDGE 30 47 -1  EDGE 30 65 1 EDGE 65 48 -1 EDGE 65 100 -1
```



Multi-segment motif FunFolDes design. Two bi-directional FoldTrees are rooted in their dependent segment moti (red)f. The roots are joined by a jump in order to form a complete FoldTree and maintain the rigid body orientation of the motif segments.

In order to keep different segments in a FunFolDes run fixed within themselves and with respect to each other, the MoveMap needs to be set so that the internal conformation of the motif and the jumps cannot be altered:

```
RESIDUE 1 24 BBCHI

RESIDUE 25 35 NO

RESIDUE 36 59 BBCHI

RESIDUE 60 70 NO

RESIDUE 71 100 BBCHI

JUMP 1 NO
```

FunFolDes: Target binder

First of all, it is important to highlight that a target binder (if available) has to be provided through the same PDB file (*231*) as the functional motif, to maintain the exact rigid body orientation between the motif and the target binder.

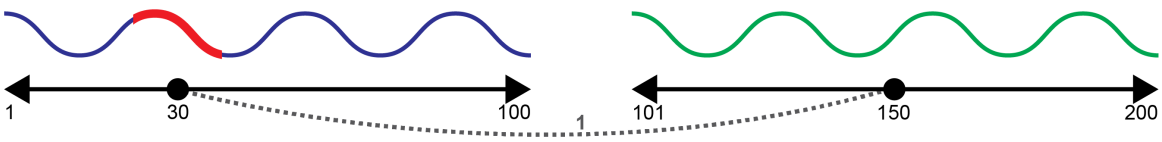
From the point of view of the FoldTree, adding one or multiple binders (e.g., the two protein chains of a target antibody) is easier than inserting multiple segments. Simply put, for each binder chain, the closest residue to the root of the FoldTree (located in the first segment of the motif) is identified and set up as root of an individual FoldTree that spans in both directions of the binder chain. Finally, a jump is set up between the old and the new root to unify the two FoldTrees into one.

Thus, if we add a 100-residue partner to our single-segment motif design (assuming the closest residue to the motif is residue 50 of the binding partner), the FoldTree will transform from

```
FOLD_TREE EDGE 30 1 -1  EDGE 30 100 -1
```

to

```
FOLD_TREE EDGE 30 1 -1  EDGE 30 100 -1  EDGE 30 150 1  EDGE 150 101 -1  EDGE 150 200 -1
```



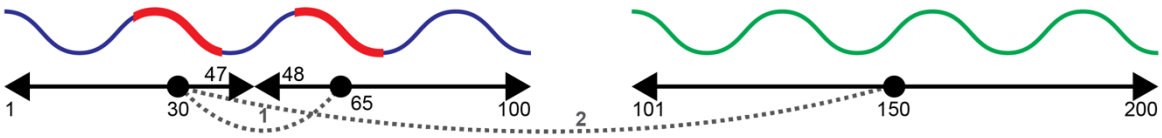
Single-segment motif design with binder. The two proteins are attached through a jump defined between the tree's root in the motif and the closest residue (in space) from the target binder.

Repeating the same exercise on our two-segment motif design will change the original FoldTree

```
FOLD_TREE EDGE 30 1 -1  EDGE 30 47 -1  EDGE 30 65 1 EDGE 65 48 -1 EDGE 65 100 -1
```

to

```
FOLD_TREE EDGE 30 1 -1  EDGE 30 47 -1  EDGE 30 65 1 EDGE 65 48 -1 EDGE 65 100 -1

EDGE 30 150 2  EDGE 150 101 -1 EDGE 150 200 -1
```



Multi-segment motif design with binder. The FoldTree construction follows the same logic as before, adding one more jump to keep the target binder in place with respect to the functional motif.

As before, setting the jumps of the MoveMap to static is key to ensure that motif and target binder are correctly positioned with respect to each other. Furthermore, it is necessary to prevent the target binder(s) from moving at all during this process. Because of that, the MoveMap will look like

```
RESIDUE 1 24 BBCHI

RESIDUE 25 35 NO

RESIDUE 36 100 BBCHI

RESIDUE 101 200 NO

JUMP 1 NO
```

for the single-segment motif with binder and

```
RESIDUE 1 24 BBCHI

RESIDUE 25 35 NO

RESIDUE 36 59 BBCHI

RESIDUE 60 70 NO

RESIDUE 71 100 BBCHI
```

```
RESIDUE 101 200 NO

JUMP 1 NO

JUMP 2 NO
```

for the multi-segment motif with binder.

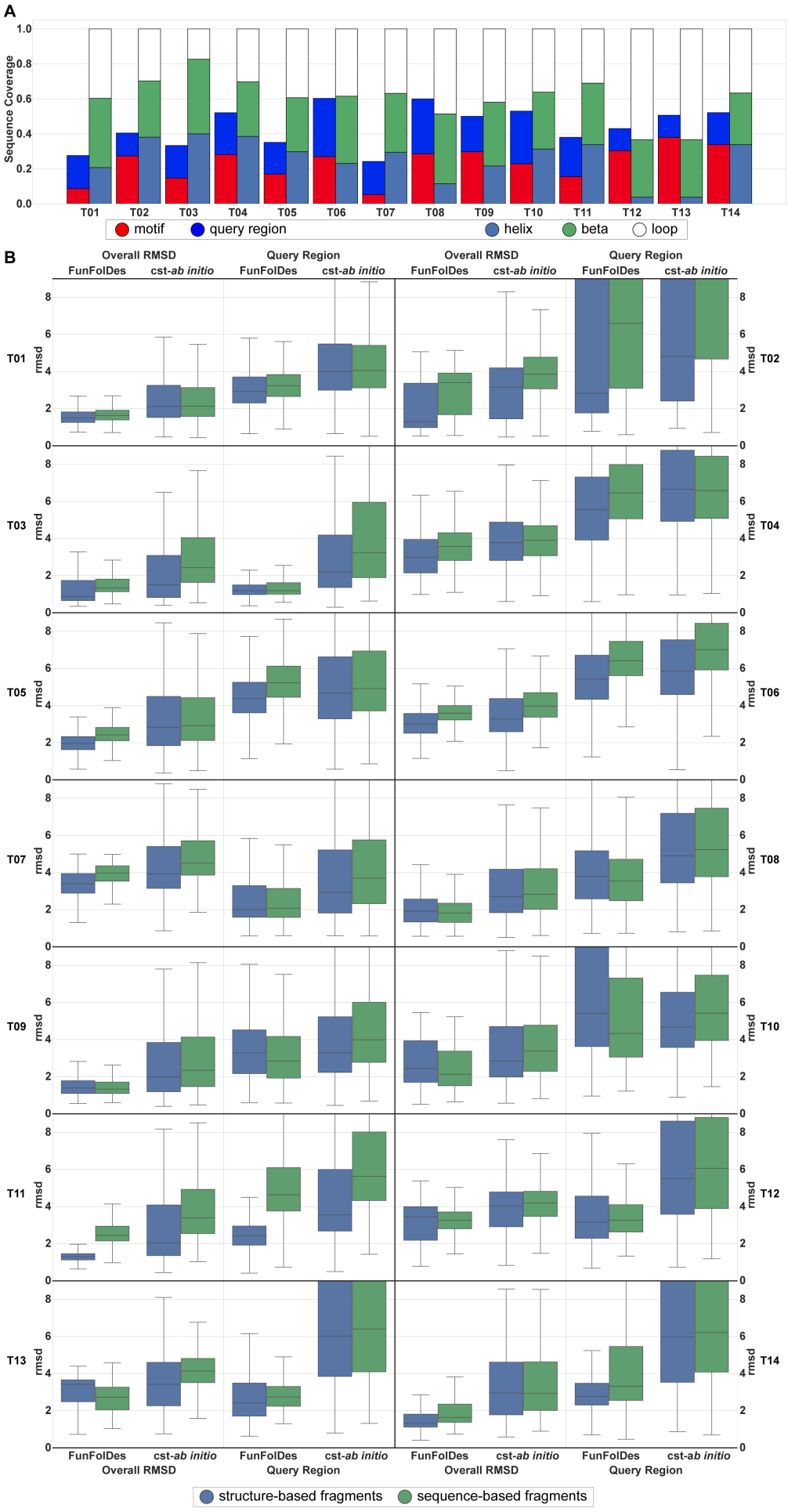

A
Sequence Coverage
1.0
0.8
0.6
0.4
0.2
0.0
T01 T02 T03 T04 T05 T06 T07 T08 T09 T10 T11 T12 T13 T14
motif
query region
helix
beta
loop
B
Overall RMSD
Query Region
FunFolDes
cst-ab initio
rmsd
T01 T02 T03 T04 T05 T06 T07 T08 T09 T10 T11 T12 T13 T14
structure-based fragments
sequence-based fragments

Figure S 2.1. Structural composition and overall results of the benchmark targets.

A) Percentage of secondary structure type, motif and query region in the overall structures. B) Full structure RMSD (Overall RMSD) and local RMSD for the query region (Query Region) between the decoy populations and their respective targets. FunFolDes tends to outperform cst–*ab initio* in all scenarios and the structure-based fragments yield decoy population with lower mean RMSDs, albeit with small differences relative to the sequence-based fragments.

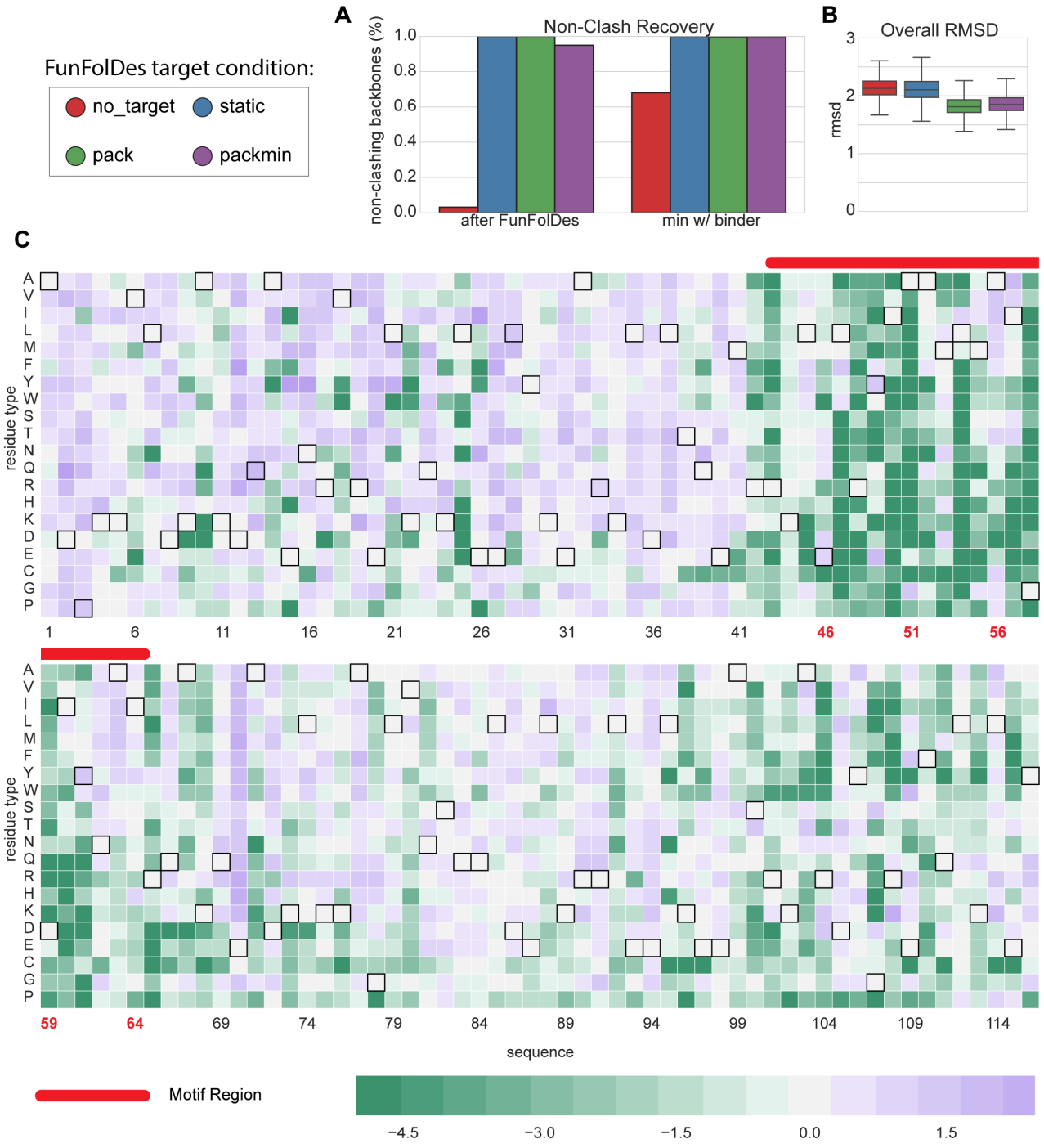


Figure S 2.2. Target-biased folding and design: structural features of the modeled designs and saturation mutagenesis data used for sequence recovery benchmark.

A) Quantification of the percentage of decoys compatible with a design-target binding conformation for the different simulation modes. The simulations performed without the target yield a very low percentage of binding

compatible conformations. After minimization, this percentage increases with significant structural drifts. B) The initial template is a 3-helix bundle structure, the slight shift needed to adopt a binding-compatible conformation produces only a small global RMSD. C) Graphical representation of the deep-sequencing data as a position-specific score matrix. Black borders highlight the native BINDI residue type for each position. Mutations for which no data were obtained, likely reflect that these protein variants were unable to fold or display at the surface of yeast and were assigned the lowest score of -5.

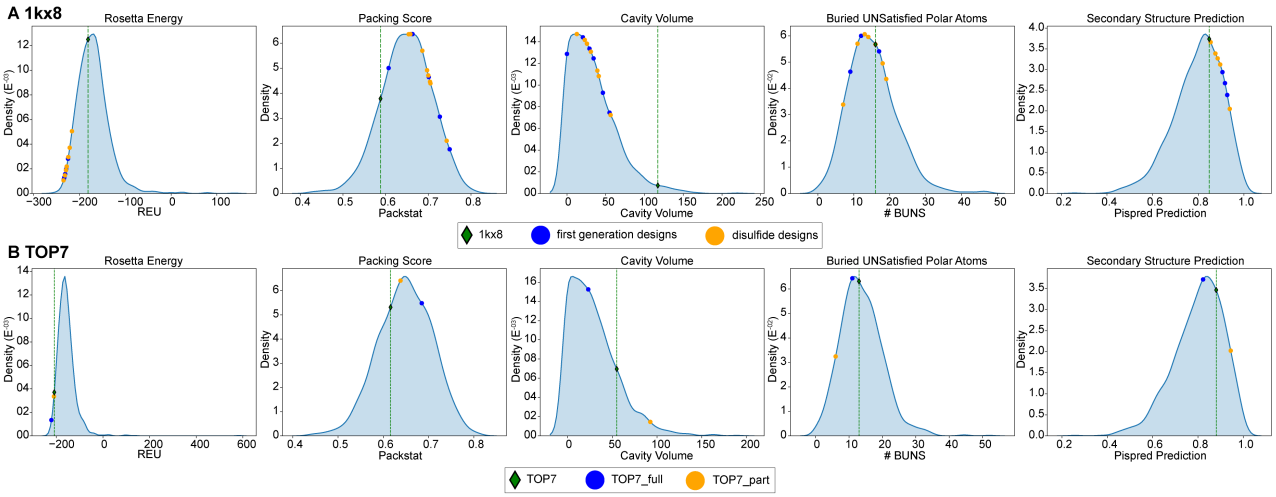
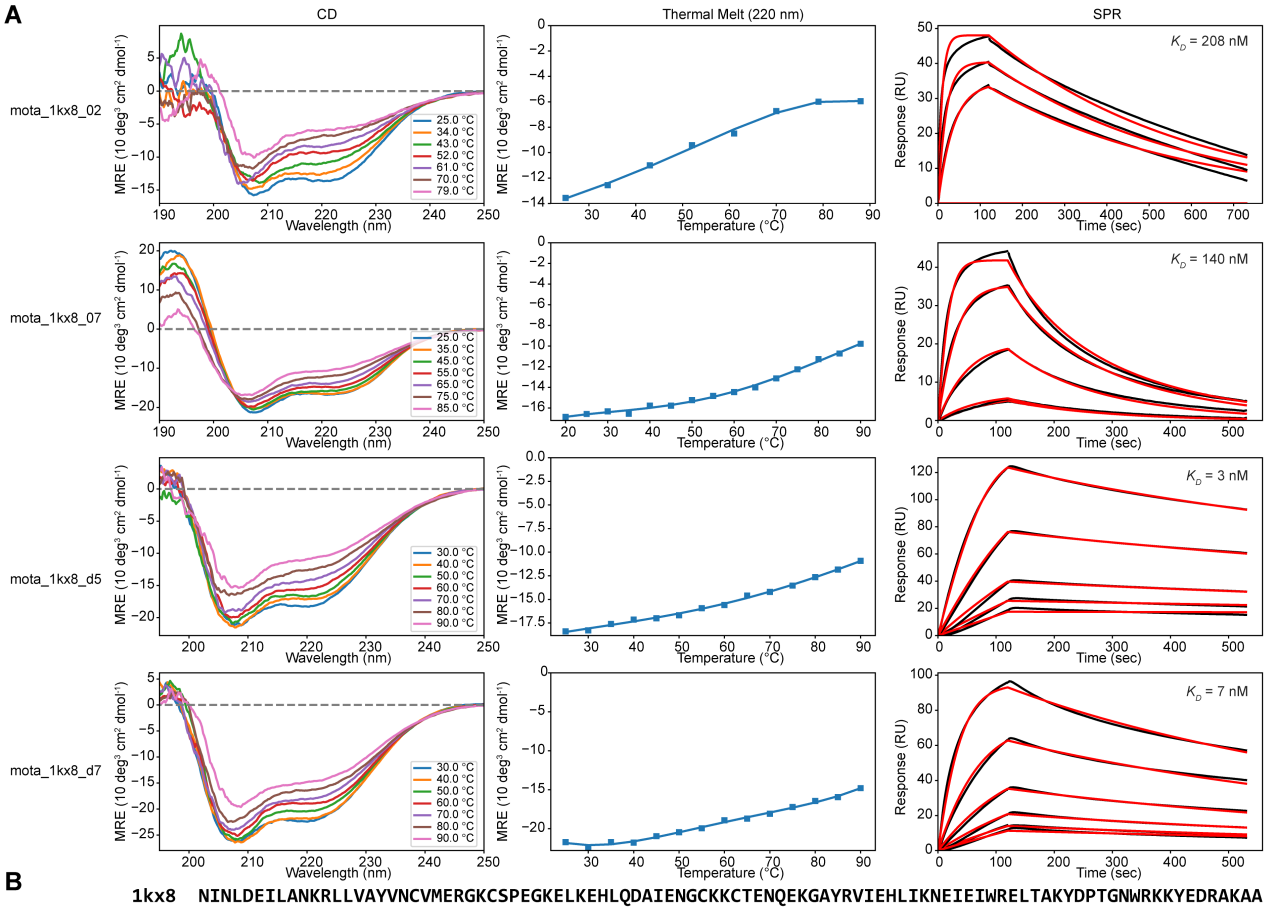


Figure S 2.3. Structural and sequence evaluation of the computational designs.

Assessment of structural and sequence features: Rosetta Energy, packing score (packstat) (*237*), cavity volume, Buried UNSatisfied polar atoms and secondary structure prediction (PSIPRED) for the template and the computational designs. Each native template (green diamond and vertical dashed line) and design (yellow and blue circles) are compared against a set of non-redundant minimized structures of similar size (± 15 residues). A) Due to its natural function, 1kx8 presents of a large cavity to bind its hydrophobic ligands. As such, the structure presents generally low scores as compared to computationally designed proteins. B) Distributions of the structural and sequence features of natural proteins and the TOP7 series of designs.



B

```
1kx8        NINLDEILANKRLLVAYVNCVMERGKCSPEGKELKEHLQDAIENGCKKCTENQEKGAYRVIEHLIKNEIEIWRELTAKYDPTGNWRKKYEDRAKAA
mota_1kx8_01 TKDMSQILKSEEAWERIIKEFEGRGQGTEDVRELARIARIWLKKGGKDLDDEERRRAKQFFAWVARQRSSEALSWINDAPITNDAKKKYSNLLAKF
mota_1kx8_02 SQSTLEVLKDPKALKEIQKQLNNQGQSSDEAKRLADIFENYIFGGGTGADEDQQKEAQKIMKWFWEFSPSELLSWINDAPITNDAKKEFSNQAANF
mota_1kx8_03 SQSTLEVLKDPKALKEIQKQLNNQGQSSDEAKRLADIFENYIFGGGTGADEDQQKEAQKIMKWFQEYSPSELLSWINDAPITNDAKKEFSNQAANF
mota_1kx8_04 SSGLELAMKGKEEAKKVKAAFEGKSEGSDDMKELAKALQEFLKNGSKGIDEKRERMARAFLAWLARQNPSEILSWINDMPITNDAKKKLSNEVAKW
mota_1kx8_05 TNLKKEIQRGQKAEDELRNAIENKGERSDEAQEIAEIFKKYFVLGGKGLSEERKQIARELMIYFARKRPSELLSWINDWPITNDAKKEYSNEAAKN
mota_1kx8_06 TEGAEAIAKDPEAWKKLKKEFEGQGQGSQDMKEIARLLKIWIVKGGKDLDDEERKQAEKIAAAVARGWGSELLSWINDAPITNDAKKKWSNQAAKL
mota_1kx8_07 PETSADVMRDPEATERLKAAIQGKGTASEDLKRLARALQEWWANGGSGLGEDEQKAAENILRWFVKQSKSELLSLINDAPITNDKKKEASNRAAQG
mota_1kx8_03_d1 SQSTLEVLKDPKALKEIQKQLNNQGQSSDEAKRLADIFENYIFGGGTGADEDQQKEAFKIMCWFQEYSPSELLSWINDAPITNDAKKEFSNQAANC
mota_1kx8_d1 TLPCDELNKSEEQVERAIKEFEGRGQGTEDARELARAARIWLKKGGKDLNDDERECAKRFFAFVARQRSSEALSWINDAPITNDAKKKYSNLLAKF
mota_1kx8_d2 KKSPLEWAQDPKCGEEMRRFLNNQGQASPDCEKAARIFENYIFGGGTGADEDQQKEAQKIMKWFWEFSPSELLSWINDAPITNDAKKEFSNQAANF
mota_1kx8_d3 RKSPLEWLNDPKCLEEIQKWLNNQGTSSDECRRVAEIFENYIFGGGTGADEDQQKEAQKIMKWFQEYSPSELLSWINDAPITNDAKKEFSNQAANF
mota_1kx8_d4 GDPAECVMKNPECGKEAKAAFEGKSEAHDEFKELAKALQEFLKNGSKGIDEKRERMARRALAWLARQNPSEILSWINDMPITNDAKKKLSNEVAKW
mota_1kx8_d5 TNACQEAQRGQKVEDEVRNAIENKGERSDEAQEVAEILKKYFVLGGKGLSEERKRCARELMIYFARKRPSELLSWINDWPITNDAKKEYSNEAAKN
mota_1kx8_d6 TEGWEAIAKDPEAWKKFKKACEGQGQASQDMEEACRQVRIWIVKGGKDLDDEERKQAEKAAAALARGWGSELLSWINDAPITNDAKKKWSNQAAKL
mota_1kx8_d7 TCRPDDVMRNPEATERLKAAIQGKGTALECLKRLARALQEWWANGGSGLGEDEQKAAENILRWFVKQSKSELLSLINDAPITNDKKKEASNRAAQG
```

Figure S 2.4. Examples of experimental characterization performed for other variants on the 1kx8 design series.

A) CD wavelength spectra (left column), thermal denaturations (middle column) and SPR binding assays with the mota antibody (right column) were performed. B) Global sequence alignment of the wild-type protein 1kx8 and the computationally designed sequences. Red positions highlight the site II epitope insertion. Green positions highlight the cysteines performing the disulfide bridges. The two positions that consistently kept the original residue type of 1kx8 are highlighted in bold.

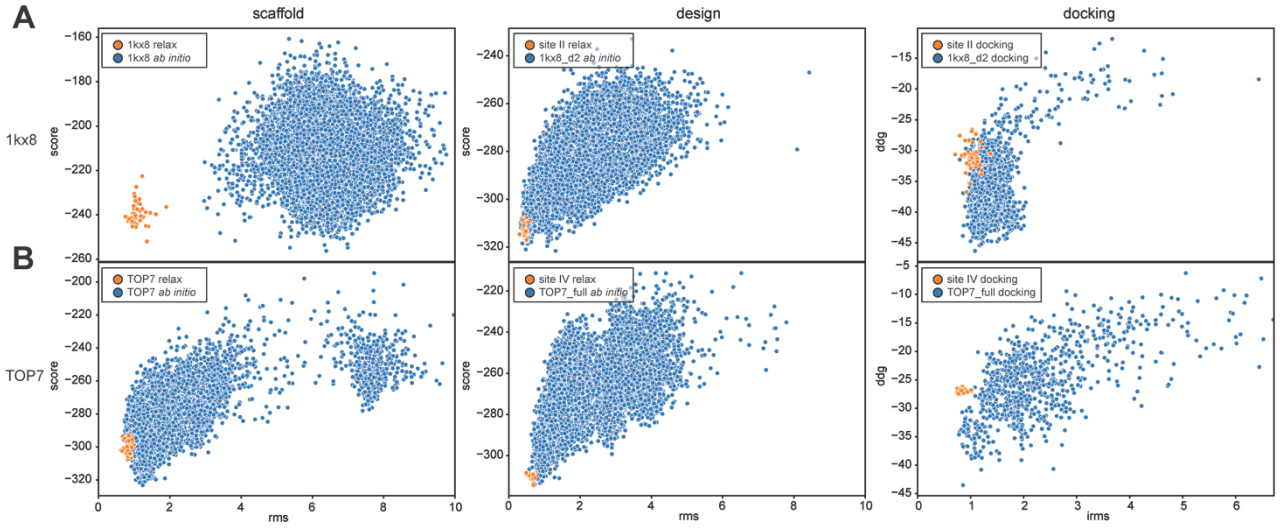


Figure S 2.5. In silico assessment of 1kx8_d2 and TOP7_full computational designs.

A) *Ab initio* folding simulations for wild-type 1kx8 (left) and design 1kx8_d2 (center), ensembles generated by relaxing the starting structures are shown in orange. The inability of 1kx8 to form a proper folding funnel could be explained by the big internal cavity of the protein due to its fatty-acid binding pocket. Docking-minimization simulations (right) performed with the top50 scoring *ab initio* decoys. The docking simulations reveal a similar binding configuration between the peptide motif and the full design, and ΔΔGs are within a similar range to those of the native peptide antibody complex. B) Same simulations as described in A) for wild-type TOP7 (left) and TOP7_full (center). We observe energetically favorable folding funnels for both wildtype and design. The docking simulations showed that the complex between the design and the antibody is formed in a similar structural configuration to the peptide-antibody complex achieving a similar range of ΔΔGs.

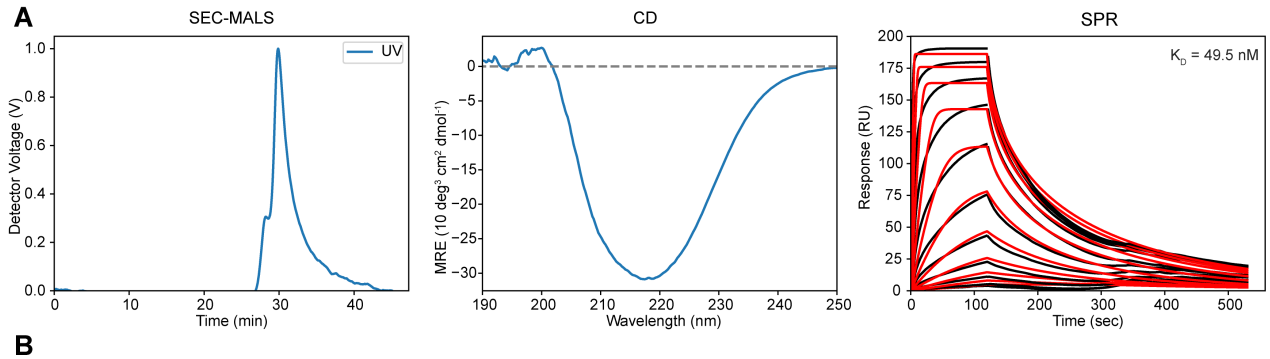


Figure S 2.6. Experimental characterization of TOP7_variants.

A) Experimental characterization for the TOP7_partial design: SEC-MALS elution profile (left column); CD wavelength scan spectrum; SPR binding assays with the 101F antibody (right column). The TOP7_partial CD spectrum is notably different from WT TOP7 and the TOP7_full design. B) Global sequence alignment of the wild-type protein TOP7 and the computationally designed sequences. Red positions highlight the site IV epitope insertion.

Chapter 3 Trivalent cocktail of *de novo* designed immunogens enables the robust induction and focusing of functional antibodies *in vivo*

Chapter 3 is based on an article that is currently under preparation. A preprint is revealed and available in bio-Rxiv (doi: https://doi.org/10.1101/685867).

Authors

Sesterhenn F[1,2*], Yang C[1,2*,] Cramer JT[3], Bonet J[1,2], Wen X[4], Abriata LA[1,2], Kucharska I[5,6], Chiang CI[7], Wang YM[7], Castoro G[3], Vollers SS[1,2], Galloux M[8], Rosset S[1,2], Corthésy P[1,2], Georgeon S[1,2], Villard M[1,2], Descamps D[8], Delgado T[9], Rameix-Welti MA[10], Más V[9], Ervin S[11], Eléouët JF[8], Riffault S[8], Bates JT[12], Ju-lien JP[5,6], Li YX[7], Jardetzky T[4], Krey T[3,13] & Correia BE[1,2].

*These authors contributed equally.

Affiliations:

[1]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. [2]Swiss Insti-tute of Bioinformatics (SIB), Lausanne CH-1015, Switzerland. [3]Institute of Virology, Hannover Medical School, Germany. [4]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA; [5]Program in Molecular Medicine, Hospital for Sick Children Research Institute, Toronto, ON, M5G 0A4, Canada. [6]Departments of Biochemistry and Immunology, University of Toronto, Toronto, ON M5S 1A8, Canada. [7]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850, USA. [8]Unité de Virologie et Immunologie Moléculaires (UR892), INRA, Université Paris-Saclay, 78352, Jouy-en-Josas, France. [9]Centro Nacional de Microbiología, Instituto de Salud Carlos III, Madrid, Spain. [10]UMR1173, INSERM, Université de Versailles St. Quentin, 78180 Montigny le Bretonneux, France. [11]Wake Forest Baptist Medical Center, Winston Salem NC 27157, USA. [12]University of Missis-sippi Medical Center, Mississippi 39216, USA. [13]German Center for Infection Research (DZIF), Hannover, Germany.

Author contributions:

FS, CY and BEC conceived the work and designed the experiments. FS and CY performed computational design and experimental characterization. JTC, GC, TK, XW and TJ solved X-ray structures. JB developed the TopoBuilder protocol. LAA performed NMR characterization and solved NMR structure. IK and JPJ performed and analysed samples by electron microscopy. CIC, YW, SSV, MG, SR, PC, SG, MV and MAR performed experiments and analysed data. JTB contributed to the design and planning of animal studies. FS, CY and BEC wrote the manuscript, with input from all authors.

3.1 Abstract

De novo protein design has been increasingly successful in expanding beyond nature's sequence and structural space. However, most *de novo* designed proteins lack biological function, in part due to the structural complexity required for functional purposes. An important domain where protein design has raised expectations was on the induction of precise antibody responses that may lead to improved vaccines. Here, we showcase two computational design approaches to stabilize irregular and discontinuous binding motifs in *de novo* designed immunogens and tested them for the induction of respiratory

syncytial virus neutralizing antibodies *in vivo*. The designs mimic the native conformations of the neutralization epitopes with sub-angstrom accuracy. *In vivo*, cocktail formulations of the immunogens induce robust neutralizing serum responses targeting three epitopes, and refocus pre-existing antibody responses towards bona fide neutralization epitopes. Our work provides a blueprint for epitope-centric vaccine design for pathogens that have frustrated traditional vaccine development efforts, and a general methodological pipeline to create novel proteins with functional sites within tailored protein topologies.

3.2 Introduction

Efforts to design novel proteins from first principles have revealed a variety of rules to control the structural features of *de novo* proteins (*181, 182, 244, 245*). However, the *de novo* design of functional proteins has been far more challenging (*7*). A commonly used strategy to design *de novo* functional proteins is to transplant the binding motifs found in existing protein structures to pre-existing or *de novo* templates. In nearly all cases, the binding motifs transplanted were commonly found in existing protein structures, such as linear helical segments, allowing the grafting of such motifs without extensive backbone adjustments (*195, 230, 246*). Most protein functional sites, however, are not contained within a single, regular segment in protein structures but arise from the three-dimensional arrangement of several, often irregular, structural elements that are supported by defined topological features of the overall structure (*247-249*). As such, it is of utmost importance for the field to develop computational approaches to endow *de novo* designed proteins with irregular and multi-segment complex structural motifs that can perform the desired functions.

Functional protein design has raised expectations in the domain of the immune response modulation; in particular, on the induction of neutralizing antibodies (nAbs) *in vivo* (*178*). Inducing nAbs targeting defined epitopes remains an overarching challenge for vaccine development (*250*). Our increasing structural understanding of many nAb-antigen interactions has provided templates for the rational design of immunogens for respiratory syncytial virus (RSV), influenza, HIV, dengue and others (*251-253*). Despite this extensive structural knowledge, these and other pathogens are still lacking efficacious vaccines, highlighting the need for next-generation vaccines to efficiently guide antibody responses towards key neutralization epitopes in both naïve and pre-exposed immune systems. The elicitation of antibody responses with defined epitope specificities has been an enduring challenge for immunogens derived from modified viral proteins (*254*).

Recently, Correia and colleagues (*178*) have shown that computationally designed immunogens could elicit epitope-specific responses. The RSVF antigenic site II, a linear helix-turn-helix motif, was transplanted onto a heterologous protein scaffold, which elicited nAbs in non-human primates (NHPs) after repeated boosting immunizations. Despite being a proof-of-principle for the induction of functional antibodies using a computationally designed immunogen, several major caveats emerged; namely, the lack of applicability of the computational approach to structurally complex epitopes, and the inconsistent neutralization titers observed in the immunogenicity studies.

To address these limitations, here we designed epitope-focused immunogens mimicking irregular and discontinuous RSV neutralization epitopes (site 0 (*255*) and IV (*224*), Fig 3.1) and showcase two computational design methodologies that enable the presentation of these structurally challenging motifs in *de novo* designed proteins. *In vivo*, cocktail formulations including an optimized site II immunogen (*256*) yielded consistent neutralization levels above the protective threshold directed against all three epitopes. The design strategies presented provide a blueprint to engineer proteins stabilizing irregular

and discontinuous binding sites, applicable to vaccine design for pathogens that require fine control over the antibody specificities induced, and more generally for the design of *de novo* proteins displaying complex functional motifs.

3.3 Results

De novo design of immunogens with structurally complex epitopes

Designing proteins with structurally complex functional sites has remained a largely unmet challenge in the field of computational protein design (*7*). We sought to design accurate mimetics of RSV neutralization epitopes, which have been structurally well characterized, and evaluate their functionality in immunization studies. We chose antigenic sites 0 and IV (Fig 3.1), which are both targeted by potent nAbs, and are structurally distinct from functional motifs that have previously been handled by computational protein design algorithms. The antigenic site 0 presents a structurally complex and discontinuous epitope consisting of a kinked 17-residue alpha helix and a disordered loop of 7 residues, targeted by nAbs D25 and 5C4 (*257, 258*), while site IV presents an irregular 6-residue bulged beta strands and is targeted by nAb 101F (*224*).

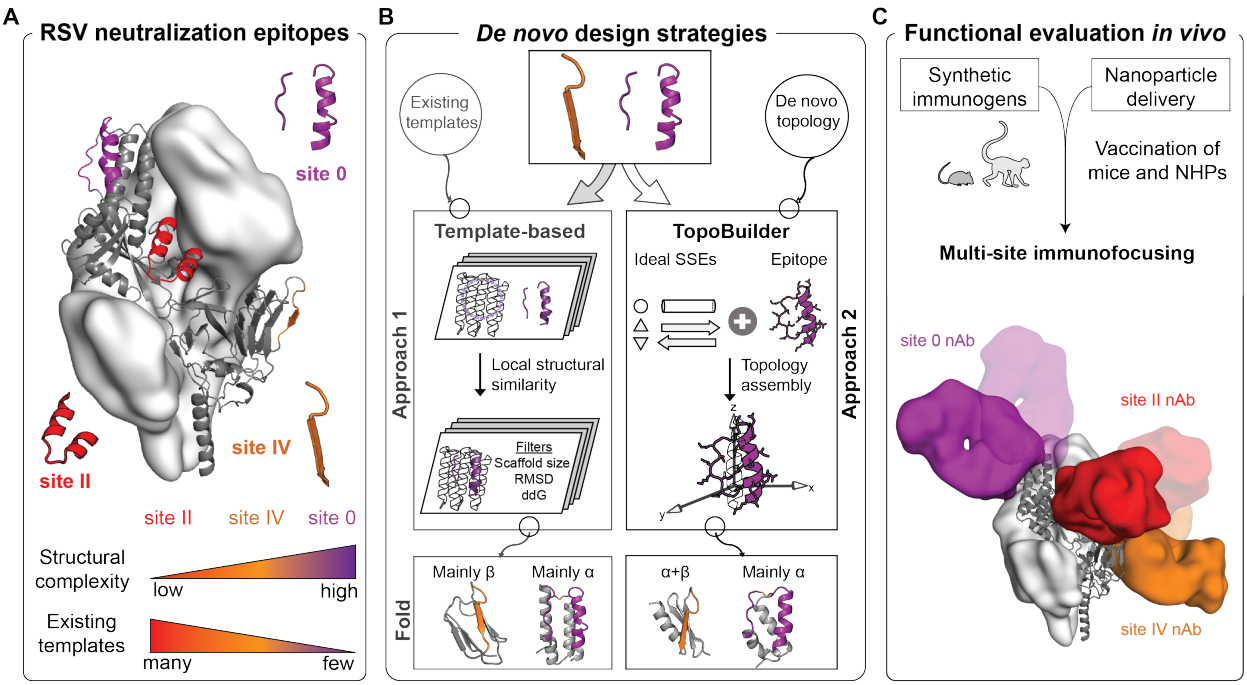


Figure 3.1. Conceptual overview of the computational design of immunogens to elicit RSV neutralizing antibodies focused on three distal epitopes.

(A) Prefusion RSVF structure (PDB 4JHW) with sites 0, II and IV highlighted. An immunogen for site II was previously reported by Correia *et al.* (B) Computational protein design strategies. Approach 1: Design templates were identified in the PDB based on loose structural similarity to site 0/IV, followed by *in silico* folding and design, and sequence optimization through directed evolution. Approach 2: A motif-centric *de novo* design approach was developed ("TopoBuilder") to tailor the protein topology to the motif's structural constraints. Bottom: Computational models of designed immunogens using different approaches. (C) Cocktail formulations of three synthetic immunogen nanoparticles to elicit nAbs focused on three non-overlapping epitopes.

The computational design of proteins mimicking structural motifs has previously been performed by first identifying compatible protein scaffolds, either from naturally occurring structures or built *de novo*, which then serve as design templates to graft the motif (*26, 193, 195, 230, 259*). Given the structural complexity of sites 0 and IV, this approach did not provide any promising matches, even using loose structural criteria.

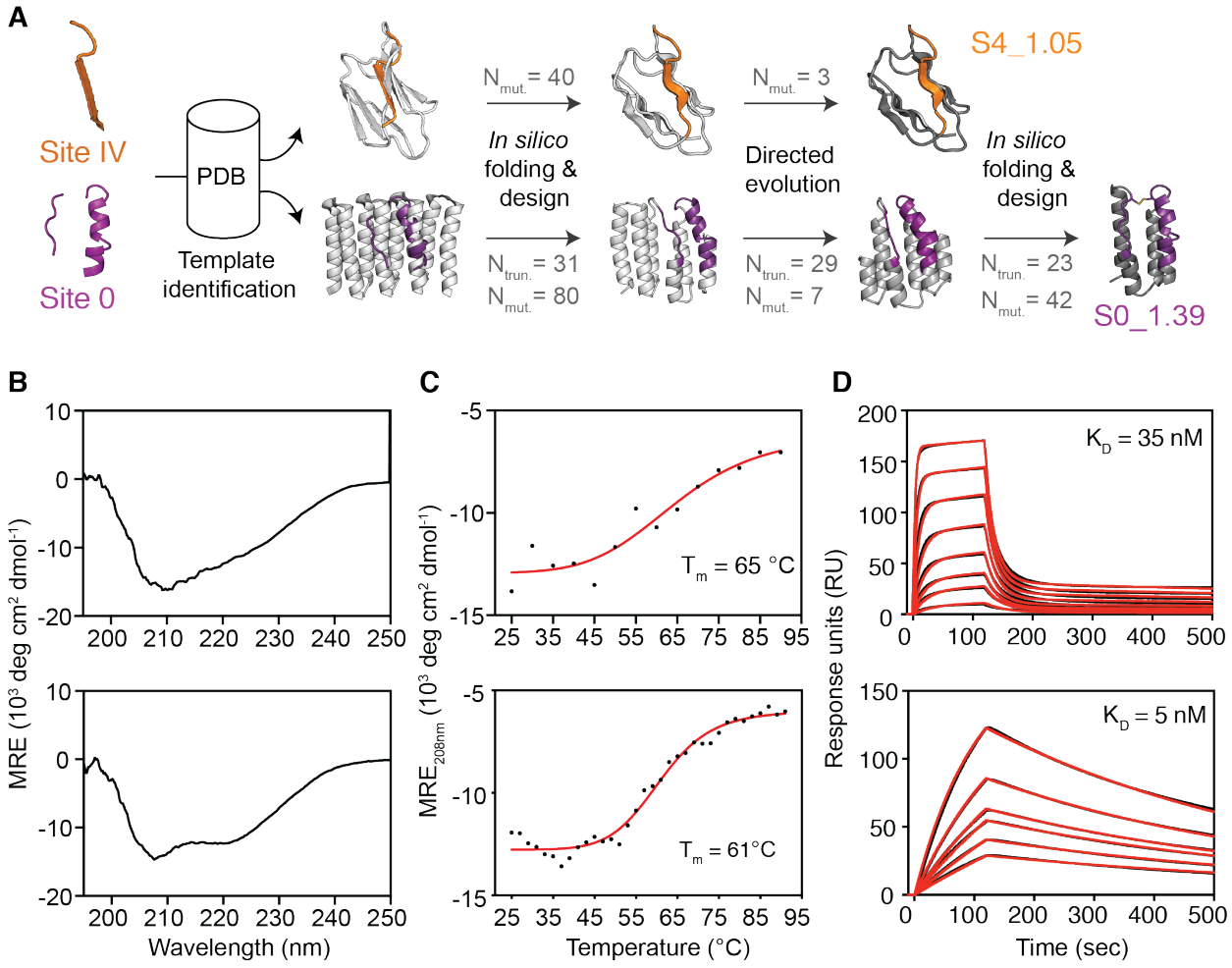


Figure 3.2. Templated computational design and biophysical characterization of synthetic immunogens.

(A) Protein design strategy - templates with structural similarity to sites IV and 0 were identified by native domain excision or loose structural matching, followed by in silico folding, design and directed evolution. An additional in silico folding and design step was necessary to install site 0 on a truncated template sequence revealed by directed evolution. Computational models of intermediates and final designs (S4_1.5 and S0_1.39) are shown, and the number of mutations (Nmut) and truncated residues (Ntrun) are indicated for each step. (B) CD spectra measured at 20 °C of S4_1.5 (top) and S0_1.39 (*110*), are in agreement with the expected secondary structure content of the design model. (C) Thermal melting curves measured by CD at 208 nm in presence of 5 mM TCEP reducing agent. (D) Binding affinity measured by SPR against target antibodies 101F (top) and D25 (*110*). Sensorgrams are shown in black and fits in red lines. CD - Circular dichroism, T_m - melting temperature, SPR - Surface plasmon resonance.

Thus, for site IV, we noticed that a small structural domain that resembles an immunoglobulin fold containing the epitope could be excised from the prefusion RSVF (preRSVF) structure. We hypothesized this would be a conservative approach to maintain its native, distorted epitope structure (Fig S3.1). We optimized the sequence for stability and epitope mimicry using Rosetta FunFolDes (*260*) (Fig 3.2a), and

our best computational design (S4_1.1) bound with a K_D > 85 μM to the 101F target antibody. To improve binding affinity, we performed deep mutational scanning followed by next-generation sequencing, as previously described (*261*) (Fig S3.1). We tested combinations of enriched positions in recombinantly expressed proteins for antibody binding, obtaining a double mutant (S4_1.5) that bound with a K_D of 35 nM to the 101F target antibody, showed a circular dichroism (CD) spectrum of a folded protein, and was thermostable up to 65 °C (Fig 3.2b-d and Fig S3.2).

The discontinuous structure of site 0 was not amenable for domain excision and stabilization. Thus, we searched for template structures that mimicked the helical segment of the epitope, and simultaneously allowed grafting the loop segment, ultimately selecting a designed helical repeat protein as a template (PDB 5cwj) (Fig 3.2a and Fig S3.3) (*262*). In order to avoid steric clashes with the target antibody D25, we truncated the N-terminal 31 residues of the 5cwj template, and performed *in silico* folding and design simulations to sample local and global changes on the scaffold to allow the presentation of the site 0 epitope (Fig 3.2a). Out of 9 sequences tested, the best design (S0_1.1) bound with a K_D of 1.4 μM to the D25 antibody (Fig S3.4), which is four orders of magnitude lower than the affinity of D25 to preRSVF (*263*). Following multiple rounds of directed evolution using yeast display, we found an enriched sequence that was C-terminally truncated by 29 residues, and showed a ~5-fold increased affinity towards D25 (Fig S3.3-S3.4). We used the truncated structure as a new template for *in silico* folding and design. Ultimately, this multi-stage process yielded S0_1.39, a design further truncated by additional 23 residues, which bound with 5 nM to D25 (Fig 3.2d). S0_1.39 was also recognized by the 5C4 antibody (Fig S3.4), which has been shown to engage site 0 in a different orientation compared to D25 (*258*), with an affinity of 5 nM, identical to that of the 5C4-preRSVF interaction (*263*).

The primary goals for the designs were achieved in terms of the stabilization of irregular and complex binding motifs in a conformation relevant for antibody binding, however, the overall strategy presented important limitations with respect to its general utility. Despite the large number of structures available to serve as design templates, the fraction of those that are practically useful for the design of functional proteins becomes increasingly limited with the structural complexity of the motif. As described above, suboptimal design templates require extensive backbone flexibility on the design process and multiple rounds of directed evolution until a sequence with high-affinity binding is identified. Additionally, the starting topology determines the overall shape of the designed protein, which may be suboptimal for the accurate stabilization of the motif, and may impose unwanted tertiary steric constraints that interfere with the designed function. In particular, for immunogen design it is desired to preserve native-like accessibility of the epitope in the context of the designed immunogen, thereby maximizing the induction of antibodies that can cross-react with the native antigen presented by the pathogen. An illustrative example on how a template-based design approach can fail to fulfill these criteria is the comparison between the quaternary environment of the site 0 epitope in preRSVF and S0_1.39 showing that this topology does not mimic such environment, albeit allowing the binding of several monoclonal antibodies (Fig S3.5).

To overcome these limitations, we developed a template-free design protocol - the TopoBuilder - that generates tailor-made topologies to stabilize complex functional motifs. Within the TopoBuilder, we parametrically sample the placement of idealized secondary structure elements which are then connected by loop segments, to assemble topologies that can stabilize the desired conformation of the structural motif. Next, these topologies are diversified to enhance structural and sequence diversity with a folding and design stage using Rosetta FunFolDes (see Fig S3.6 and methods for details). For this approach, we defined two new design objectives which were unmet by our previous template-based designs: 1) building stable *de novo* topologies that stabilize the epitope, while mimicking their native

quaternary environment; 2) fine-tuning the topology's secondary structure arrangements to maximize the fold stability and optimize epitope presentation for high affinity antibody binding.

To present antigenic site IV, we designed a fold composed of a β-sheet with 4 antiparallel strands and one helix (Fig 3.3a), referred to as S4_2 fold. Within the S4_2 topology, we generated three structural variants (S4_2_bb1-3), by parametrically sampling three distinct helical secondary structural elements, varying both orientations and lengths to maximize the packing interaction against the β-sheet. Sequences generated from two structural variants (S4_2_bb2 and S4_2_bb3) showed a strong propensity to recover the designed structures in Rosetta *ab initio* simulations (Fig 3.3a and Fig S3.7).

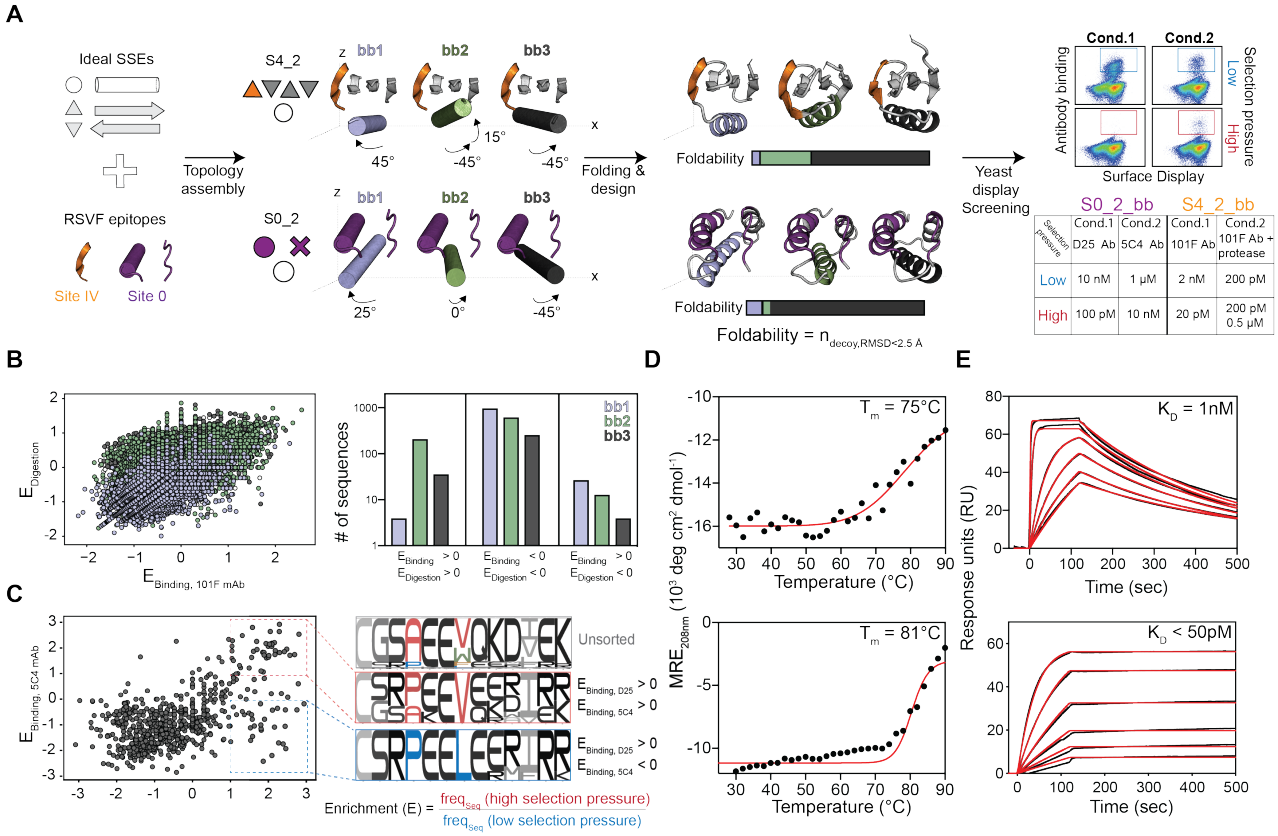


Figure 3.3. Motif-centric *de novo* design of epitope-focused immunogens.

(A) Ideal secondary structure elements (SSEs) are assembled around RSVF epitopes, sampling different orientations within the same topology, followed by a single round of *in silico* folding and design. Rosetta *ab initio* simulations are performed for designs of each topology to assess its propensity to fold into the designed structures, returning a foldability score. Selected designs are then displayed on yeast surface and sorted under two different selection pressures for subsequent deep sequencing. (B) All three designed topological variants were screened for high affinity binding and resistance to chymotrypsin to select stably folded proteins. Enrichment analysis revealed a strong preference for one of the designed helix orientations (S4_2_bb2, green) to resist protease digestion and bind with high affinity to 101F. (C) Enrichment analysis of sorted populations under high and low selective pressures. Sequences highly enriched for both D25 and 5C4 binding show convergent sequence features in critical core positions of the site 0 scaffold. (D) Thermal melting curves measured by CD for best designs (S4_2.45 (top) and S0_2.126 (*110*)) showing high thermostability. (E) Dissociation constants (K_D) of S4_2.45 to 101F (top) and S0_2.126 to D25 (*110*) antibodies measured by SPR.

We screened a defined set of computationally designed sequences using yeast display and applied two selective pressures – binding to 101F and resistance to the nonspecific protease chymotrypsin, an

effective method to digest partially unfolded proteins (*230, 264, 265*). Deep sequencing of populations sorted under different conditions revealed that S4_2_bb2-based designs were strongly enriched under stringent selection conditions for folding and 101F binding, showing that subtle topological differences in the design template can have a substantial impact on function and stability. We expressed 15 S4_2_bb2 design variants and successfully purified and biochemically characterized 14. The designs showed mixed alpha/beta CD spectra and bound to 101F with affinities ranging from 1 nM to 200 nM (Fig S3.8). The best variant, S4_2.45 (K_D = 1 nM), was well folded and thermostable according to CD and NMR with a T_m of 75 °C (Fig 3.3d and Fig S3.9).

Similarly, we built a minimal *de novo* topology to present the tertiary structure of the site 0 epitope. The choice for this topology was motivated by the fact that site 0, in its native preRSVF environment, is accessible for antibody binding from diverse angles (*258*), in contrast to the S0_39 natural template which topologically constrained site 0 accessibility (Fig S3.5). By building a template *de novo,* we attempted to respect site 0's native quaternary constraints, while stabilizing both irregular epitope segments with high accuracy.

We explored the topological space within the shape constraints of preRSVF and built three different helical orientations (S0_2_bb1-3) that support both epitope segments. Evaluation of the designed sequences with Rosetta *ab initio* showed that only sequences generated based on one topology (S0_2_bb3) presented a funnel-shaped energy landscape (Fig S3.10). A set of computationally designed sequences based on S0_2_bb3 was screened in yeast under the selective pressure of two site 0-specific antibodies (D25 and 5C4) to ensure the presentation of the native epitope's conformation. Deep sequencing of the double-enriched clones and subsequent sequence analysis revealed that a valine at position 28 is critical to retain a cavity formed between the two epitope segments, ensuring binding to both antibodies (Fig 3.3b). We selected five sequences, differing from 3 to 21 mutations, for further biochemical characterization (Fig S3.11). The design with best solution behaviour (S0_02.126) showed a CD spectrum of a mostly helical protein, with extremely high thermostability even under reducing conditions (T_m = 81 °C, Fig 3.3d) and a well-dispersed HSQC NMR spectrum (Fig S3.9). Strikingly, S0_2.126 bound with ~50 pM affinity to D25, similar to that of the preRSVF-D25 interaction (~150 pM), and with a K_D = 4 nM to 5C4 (Fig 3.3e and Fig S3.12).

Overall, the properties of the designs generated by topological assembly with the TopoBuilder showed improved binding affinities and thermal stabilities as compared to those using available structural templates. To investigate whether this design and screening procedure yielded scaffolds that better mimicked the viral epitope presented, or rather revealed sequences with a highly optimized interface towards the antibodies used during the selection, we determined the affinities of S4_2.45 and S0_2.126 against a panel of human site-specific antibodies. Compared to the first-generation designs, S4_2.45 and S0_2.126 showed large affinity improvements across the antibody panels, exhibiting a geometric mean affinity closely resembling that of the antibodies to preRSVF (Fig S3.12). In the light of such results, we concluded that the topologically designed immunogens were superior mimetics of sites IV and 0 relative to the template-based designs.

De novo designed topologies adopt the predicted structures with high accuracy

To evaluate the structural accuracy of the computational design approach, we solved the crystal structure of S4_2.45 in complex with 101F at 2.6 Å resolution. The structure closely matched our design model, with a full-atom RMSD of 1.5 Å (Fig 3.4a). The epitope was mimicked with an RMSD of 0.135 Å,

and retained all essential interactions with 101F (Fig 3.4d,e). Importantly, the structural data confirmed that we presented an irregular beta strand, a common motif found in many protein-protein interactions (*266*), in a fully *de novo* designed protein with sub-angstrom accuracy.

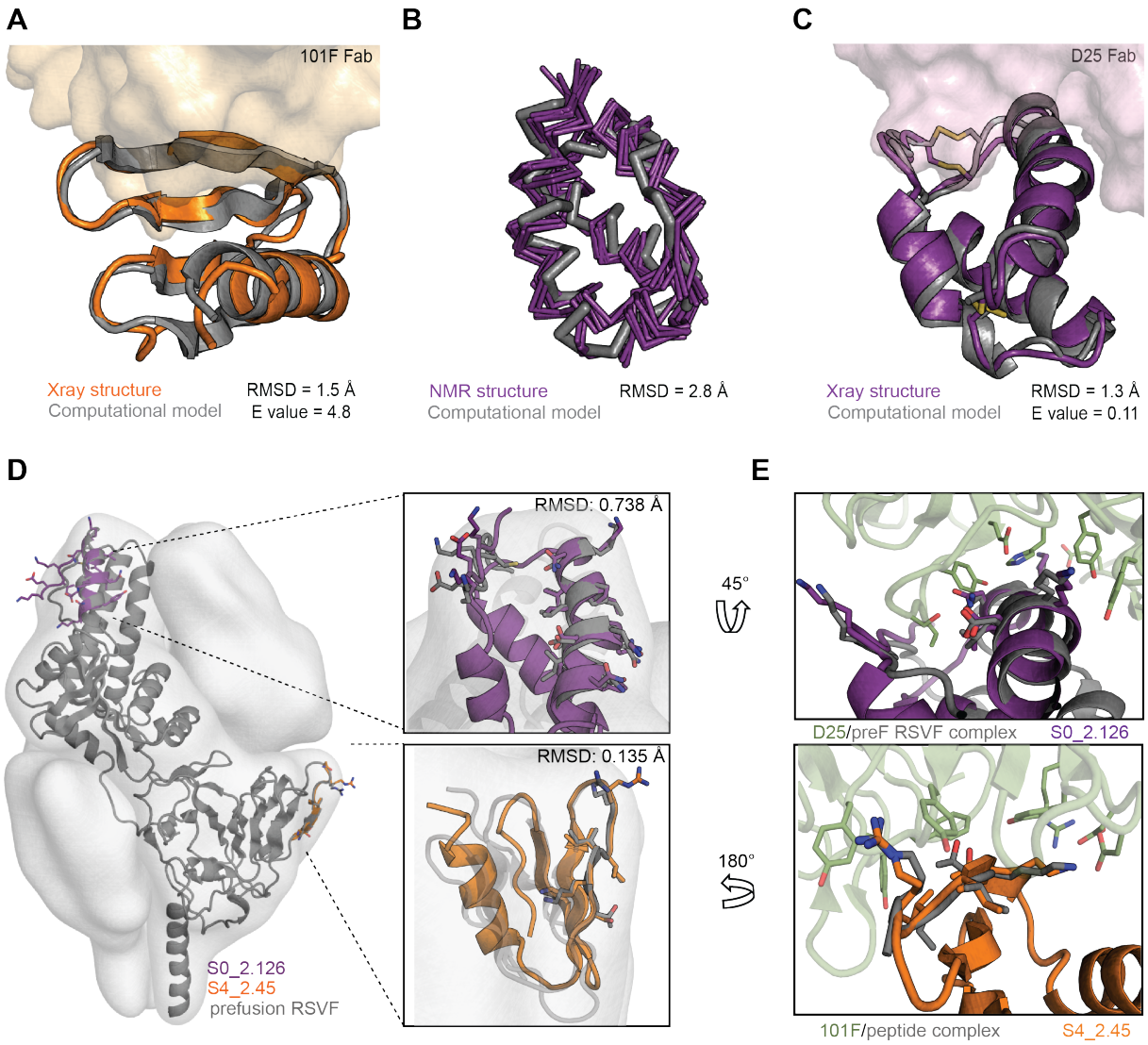


Figure 3.4. Structural characterization of *de novo* designed immunogens.

(A) Crystal structure of S4_2.45 (orange) bound to 101F Fab closely matches the design model (grey, RMSD = 1.5 Å). (B) NMR structural ensemble of S0_2.126 (purple) superimposed to the computational model (*204*). The NMR structure is in agreement with the design model (backbone RMSD of 2.8 Å). (C) Crystal structure of S0_2.126 (purple) bound to D25 Fab closely resembles the design model (grey, RMSD = 1.3 Å). (D) Superposition of the preRSVF sites 0/IV and designed immunogens show sub-angstrom mimicry of the epitopes. Designed scaffolds are compatible with the shape constraints of preRSVF (surface representation). (E) Close-up view of the interfacial side-chain interactions between D25 (top) and 101F (*110*) with designed immunogens as compared to the starting epitope structures.

Next, we solved an unbound structure of S0_2.126 by NMR, confirming the accuracy of the designed fold with a backbone RMSD between the average structure and the model of 2.8 Å (Fig 3.4b). Additionally, we solved a crystal structure of S0_2.126 bound to D25 at a resolution of 3.0 Å. The structure showed

an overall RMSD of 1.5 Å to the design model, and an RMSD of 0.9 Å over the discontinuous epitope compared to preRSVF (Fig 3.4c-e). To the best of our knowledge, this is the first computationally *de novo* designed protein that presents a two-segment, structurally irregular, binding motif with atomic-level accuracy. In comparison with native proteins, S0_2.126 showed exceptionally low packing due to a large core cavity (Fig S3.13), but retained a very high thermal stability. The core cavity was essential for antibody binding and highlights the potential of *de novo* approaches to design small proteins hosting structurally challenging motifs and preserving cavities required for function (*182*). Notably, due to the level of control and precision of the TopoBuilder, both designed antigens respected the shape constraints of the respective epitope in their native environment within preRSVF, a structural feature that may be important for the improved elicitation of functional antibodies (Fig S3.5).

Cocktails of designed immunogens elicit neutralizing Antibodies in vivo

Lastly, we sought to evaluate the designed antigens for their ability to elicit focused nAb responses *in vivo*. Our rationale for combining site 0, II and IV immunogens in a cocktail formulation is that all three sites are non-overlapping, as verified by electron microscopy analysis (Fig S3.14), and thus could induce a more potent antibody response *in vivo*. To increase immunogenicity, each immunogen was multimerized on self-assembling protein nanoparticles. We chose the RSV nucleoprotein (RSVN), a self-assembling ring-like structure of 10-11 subunits, previously shown to be an effective carrier for the site II immunogen (*256*), and formulated a trivalent immunogen cocktail containing equimolar amounts of S0_1.39, S4_1.05 and S2_1.2 immunogen nanoparticles ("Trivax1", Fig S3.15). The fusion of S0_2.126 and S4_2.45 to RSVN yielded poorly soluble nanoparticles, prompting us to use ferritin particles for multimerization, with a 50% occupancy (~12 copies), creating a second cocktail comprising S2_1.2 in RSVN and the remaining immunogens in ferritin ("Trivax2", Fig S3.16).

In mice, Trivax1 elicited low levels of RSVF cross-reactive antibodies, and sera did not show RSV neutralizing activity in most animals (Fig S3.17). In contrast, Trivax2 induced robust levels of RSVF cross-reactive serum levels, and the response was balanced against all three epitopes (Fig 3.5a,b). Strikingly, Trivax2 immunization yielded RSV neutralizing activity above the protective threshold in 6/10 mice (Fig 3.5c). These results show that vaccine candidates composed of *de novo* designed proteins mimicking viral neutralization epitopes can induce robust antibody responses *in vivo*, targeting multiple specificities. This is an important finding given that mice have been a traditionally difficult model to induce neutralizing antibodies with scaffold-based immunogens (*178, 259*).

In parallel, we sought to test the potential of a trivalent immunogen cocktail in NHPs. The previously designed site II immunogen showed promise in NHPs, but the induced neutralizing titers were low and inconsistent across animals, requiring up to five immunizations to elicit neutralizing antibodies in 2/4 animals (*178*). We immunized seven RSV-naïve NHPs with Trivax1, as detailed in Fig 3.5d. In contrast to mice, NHPs developed robust levels of RSVF cross-reactive serum titer in all animals (Fig 3.5e), and antibodies induced were directed against all three epitopes (Fig 3.5f). Strikingly, we found that 6/7 NHPs showed RSV neutralizing serum levels above the protective threshold after a single boosting immunization (median IC_{50} = 312) (Fig 3.5g). Neutralization titers were maximal at day 84 (median IC_{50} = 408), fourfold above the protective threshold (*263*), and measurements were confirmed by an independent laboratory (Fig S3.18).

Immunization studies in naïve animals are essential to test the capability of the designed immunogens to induce functional antibodies. However, an overarching challenge for vaccine development to target

pathogens such as RSV, influenza, dengue and others is to focus or reshape pre-existing immunity of broad specificity on defined neutralizing epitopes that can confer long-lasting protection (*256, 267, 268*). To mimic a serum response of broad specificity towards RSV, we immunized 13 NHPs with preRSVF. All animals developed strong preRSVF-specific titers and cross-reactivity with all the epitope-focused immunogens, indicating that epitope-specific antibodies were primed and recognize the designed immunogens (Fig S3.19). Group 2 (6 animals) subsequently served as a control group to follow the dynamics of epitope-specific antibodies over time, and group 3 (7 animals) was boosted three times with Trivax1 (Fig 3.5d). PreRSVF-specific antibody and neutralization titers maximized at day 28 and were maintained up to day 119 in both groups (Fig 3.5h,i). Analysis of the site-specific antibody levels showed that site 0, II and IV responses were dynamic in the control group, with site II dropping from 37% to 13% and site 0 from 17% to 4% at day 28 and 91, respectively (Fig 3.5j). In contrast, site IV specific responses increased from 13% to 43% over the same time span. Although Trivax1 boosting immunizations did not significantly change the magnitude of the preRSVF-specific serum response, they reshaped the serum specificities in primed animals. Site II specific titers were 6.5-fold higher (day 91) compared to the non-boosted control group (84% vs 13%, p = 0.02, Mann-Whitney), and unlike the rapid drop of site 0-specific antibodies in the non-boosted group, these antibodies were maintained upon Trivax1 boosting (25% vs 4%, p=0.02, Mann-Whitney) (Fig 3.5j). In contrast, site IV specific responses increased to similar levels in both groups, 43% and 40% in group 2 and 3, respectively. Strikingly, upon depletion of site 0, II and IV specific antibodies from pooled sera, we observed a 60% drop in neutralizing activity in group 3 as compared to only a 7% drop in the non-boosted control group, indicating that Trivax1 boosting reshaped a serum response of broad specificity towards a more focused response that predominantly relies on site 0, II and IV-specific antibodies for RSV neutralization (Fig 3.5k).

Altogether, the design strategies utilized, yielded antigens presenting structurally complex neutralization epitopes that induce neutralizing antibodies upon cocktail formulation, providing a strong rationale for including multiple, ideally non-overlapping epitopes in an epitope-focused vaccination strategy. While the first-generation immunogens were inferior according to biophysical parameters and failed to induce neutralization in mice, they were successful under two different immunological scenarios in NHPs, and we show that a second generation can now induce neutralizing antibodies in mice. This is an important step to optimize and test different nanoparticles, formulations and delivery routes in a small animal model, and we foresee that the second-generation immunogens will prove superior in inducing neutralizing serum responses in NHPs.

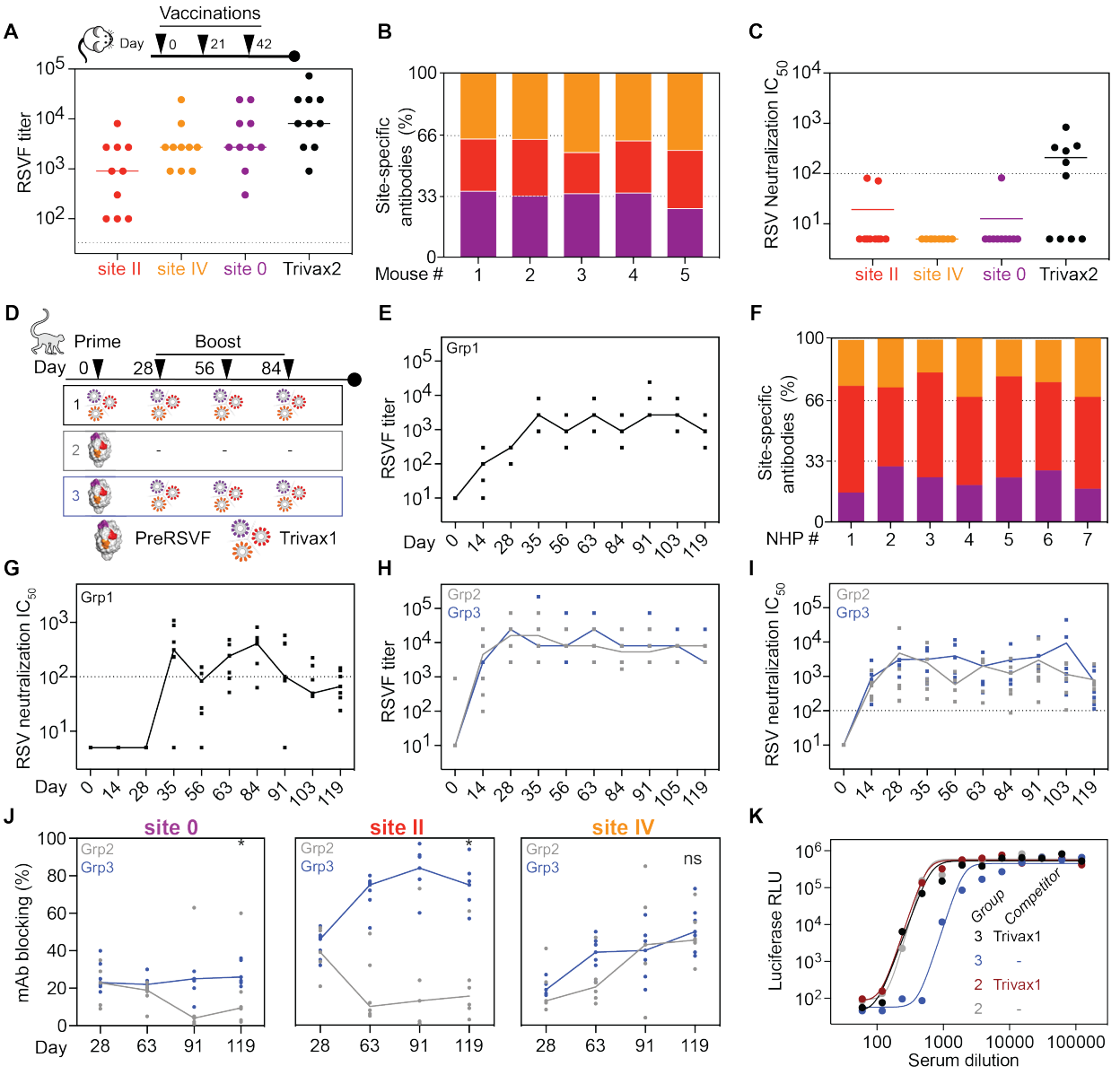


Figure 3.5. Synthetic immunogens elicit neutralizing serum responses in mice and NHPs, and focus pre-existing immunity on sites 0 and II.

(A-C) Trivax2 immunization study in mice. (A) PreRSVF cross-reactive serum levels following three immunizations with single immunogens or Trivax2 cocktail (day 56). (B) Serum specificity shown for 5 representative mice immunized with Trivax2, as measured by an SPR competition assay with D25, Motavizumab and 101F IgGs as competitors, exhibiting an equally balanced response towards all sites. (C) RSV neutralization titer of mice at day 56, immunized with Trivax2 components individually and as cocktail. Dotted line (IC_{50} = 100) indicates protective threshold as defined by protective level of Palivizumab. (D-K) Trivax1 immunization study in NHPs. (D) NHP immunization scheme. (E) PreRSVF cross-reactive serum levels for group 1. (F) Serum antibodies target all three antigenic sites in all 7 animals as measured by an SPR competition assay. (G) RSV neutralization titers of group 1. (H) PreRSVF titer in group 2 (*204*) and 3 (blue). (I) RSV neutralization titer of group 2 and 3. (J) Site-specific antibody levels measured by SPR competition assay. Site 0 and site II-specific titers were significantly higher in group 3 compared to 2 following Trivax1 boosting ($p < 0.05$, Mann-Whitney U test). (K) RSV neutralization curves upon depletion of day 91 sera with site 0, II, IV-specific scaffolds. 60% of the neutralizing activity is competed in group 3, whereas no significant decrease is observed in the control group 2.

3.4 Discussion and conclusions

Our work showcased two computational protein design strategies to design immunogens which present structurally complex epitopes with atomic accuracy, and validated their functionality to elicit nAb responses in cocktail formulations both in mice and NHPs. We have shown that through computational design of pre-existing templates with full backbone flexibility, irregular and discontinuous epitopes were successfully stabilized in heterologous scaffolds. However, this design strategy required extensive *in vitro* evolution optimization and the resulting scaffolds remained suboptimal regarding their biochemical and biophysical properties. Moreover, the lack of precise topological control of the designed proteins is a major limitation for the design of functional proteins that require defined topological similarity in addition to local mimicry of the transplanted site. For instance, the design template of the site 0 immunogen did not mimic the quaternary environment of the epitope of interest, which may have contributed to the low levels of functional antibodies induced in mice.

To overcome these limitations, we developed the TopoBuilder, a motif-centric design approach that tailors a protein fold directly to the functional site of interest. Compared to previously employed functional *de novo* design protocols, in which a stable scaffold topology was constructed first and endowed with binding motifs in a second step (*230*), our method has significant advantages for structurally complex motifs. First, it tailors the topology to the structural requirements of the functional motif from the start of the design process, rather than through the adaptation (and often destabilization) of a stable protein to accommodate the functional site. Second, the topological assembly and fine-tuning allowed to select for optimal backbone orientations and sequences that stably folded and bound with high affinity in a single screening round, without further optimization through directed evolution, as often necessary in computational protein design efforts (*102, 195, 197, 230*). Together, our approach enabled the computational design of *de novo* proteins presenting irregular and discontinuous structural motifs that are typically required to endow proteins with diverse biochemical functions (e.g., binding or catalysis), thus providing a new means for the *de novo* design of functional proteins.

As to the functional aspect of our design work, we showed *in vivo* that these immunogens consistently elicited neutralizing serum levels in mice and NHPs as cocktail formulations. The elicitation of focused neutralizing antibody responses by vaccination remains the central goal for vaccines against pathogens that have frustrated conventional vaccine development efforts (*250*). Using RSV as a model system, we have shown that cocktails of computationally designed antigens can robustly elicit neutralizing serum levels in naïve animals. These neutralization levels were much superior to any previous report on epitope-focused immunogens (*178*) and provide a strong rationale for an epitope-focused vaccination strategy involving multiple, non-overlapping epitopes. Also, their capability to dramatically reshape the nature of non-naïve repertoires in NHPs addresses an important challenge for many next-generation vaccines to target pathogens for which efficacious vaccines are needed (*256, 268, 269*). An important pathogen from this category is influenza, where the challenge is to overcome immunodominance hierarchies (*270*), which have been established during repeated natural infections, and that favour strain-specific antibody specificities, rather than cross-protecting nAbs found in the hemagglutinin stem region (*271*). The ability to selectively boost subdominant nAbs targeting defined, broadly protective epitopes that are surrounded by strain-specific epitopes could overcome long-standing challenges in vaccine development, given that cross-neutralizing antibodies can persist for years once elicited (*272*). A tantalizing future application for epitope-focused immunogens could marry this technology with engineered components of the immune system, which could be used to stimulate antibody production of adoptively transferred, engineered B-cells that express monoclonal therapeutic antibodies *in vivo* (*273*).

Altogether, this study provides a blueprint for the design of epitope-focused vaccines against pathogens that have eluded traditional vaccine development approaches. Beyond immunogens, our approach to design *de novo* proteins presenting complex binding sites will be broadly applicable to engineer novel functional proteins with defined structural properties.

3.5 Methods

Computational design of template-based epitope-focused immunogens

All code used for the computational design and analysis is available through a public github repository: https://github.com/lpdi-epfl/trivalent_cocktail. It contains the TopoBuilder source code, RosettaScripts used for the design, analysis scripts and detailed information on how designs were selected. Tables with the amino acid sequences of the experimentally characterized sequences are available in the supplementary material of a publicly available preprint of the manuscript, found under: https://doi.org/10.1101/685867.

Site 0

The structural segments entailing the antigenic site 0 were extracted from the prefusion stabilized RSVF Ds-Cav1 crystal structure, bound to the antibody D25 (PDB ID: 4JHW) (*255*). The epitope consists of two segments: a kinked helical segment (residues 196-212) and a 7-residue loop (residues 63-69).

The MASTER software (*221*) was used to perform structural searches over the Protein Data Bank (PDB, from August 2018), containing 141,920 protein structures, to select template scaffolds with local structural similarities to the site 0 motif. A first search with a Cα RMSD threshold below 2.5 Å did not produce any usable structural matches both in terms of local mimicry as well as global topology features. A second search was performed, where extra structural elements that support the epitope in its native environment were included as part of the query motif to bias the search towards matches that favoured motif-compatible topologies rather than those with close local similarities. The extra structural elements included were the two buried helices that directly contact the site 0 in the preRSVF structure (4JHW residues 70-88 and 212-229). The search yielded initially 7,600 matches under 5 Å of backbone RMSD, which were subsequently filtered for proteins with a length between 50 and 160 residues, high secondary structure content, as well as for accessibility of the epitope for antibody binding. Remaining matches were manually inspected to select template-scaffolds suitable to present the native conformation of antigenic site 0. Subsequently, we selected a computationally designed, highly stable, helical repeat protein (*262*) consisting of 8 regular helices (PDB ID: 5CWJ) with an RMSD of 4.4 Å to the query (2.82 Å for site 0 segments only). To avoid steric clashes with the D25 antibody, we truncated the 5CWJ template structure at the N-terminus by 31 residues, resulting in a structural topology composed of 7 helices.

Using Rosetta FunFolDes (*260*) the truncated 5CWJ topology was folded and designed to stabilized the grafted site 0 epitope recognized by D25. We generated 25,000 designs and selected the top 300 by Rosetta energy score (RE), designed backbones that presented obvious flaws, such as low packing scores, distorted secondary structural elements and buried unsatisfied atoms were discarded. From the top 300 designs, 3 were retained for follow-up iterative cycles of structural relaxation and design using Rosetta FastDesign (*234*), generating a total of 100 designed sequences.

The best 9 designs by Rosetta energy score were recombinantly expressed in *E. coli* 2 designed sequences derived from the same backbone, were successfully expressed and purified. The best variant was named S0_1.1, and subjected to experimental optimization using yeast surface display (Fig S3.3-S3.4). In one of the libraries, we found a truncated sequence (S0_1.17) enriched for expression and binding, which served as template for a second round of computational design (Fig S3.3-S3.4). We performed 25,000 folding and design simulations using Rosetta FunFolDes (*260*). The best 300 decoys by total Rosetta energy score were extracted, and relaxed using the Rosetta Relax application (*274*). We computed the mean total RE, and selected designs that showed a lower energy score than the mean of the design population (RE = -155.2), RMSD drift of the epitope after relaxing of less than 0.7 Å, and a cavity volume <60 $Å^3$. We selected one of the best 5 scoring decoys, truncated the C-terminal 29 and N-terminal 23 residues which did not contribute to epitope stabilization, and introduced a disulfide bond between residue 1 and 43. Four sequences were experimentally tested (S0_1.37-40). The best variant according to binding, S0_1.39, bound with 5 nM affinity to antibody D25, and, importantly, also gained binding to the 5C4 antibody (K_D = 5 nM).

Site IV

When the design simulations were carried out, there was no structure available of the full RSVF protein in complex with a site IV-specific nAb, nevertheless a peptide epitope of this site recognized by the 101F nAb had been previously reported (PDB ID: 3O41) (*224*).

The crystallized peptide-epitope corresponds to the residues 429-434 of the RSVF protein. Structurally the 101F-bound peptide-epitope adopts a bulged strand and several studies suggest that 101F recognition extends beyond the linear β-strand, contacting other residues located in antigenic site IV (*275*). Despite the apparent structural simplicity of the epitope, structural searches for designable scaffolds failed to yield promising starting templates. However, we noticed that the antigenic site IV of RSVF is self-contained within an individual domain that could potentially be excised and designed as a soluble folded protein. To maximize these contacts, we first truncated the seemingly self-contained region from RSVF pre-fusion structure (PDB ID: 4JHW, residue: 402-459) forming a β-sandwich and containing site IV. We used Rosetta FastDesign to optimize the core positions of this minimal topology, obtaining our initial design: S4_wt. However, S4_wt did not show a funnel-shaped energy landscape in Rosetta *ab initio* simulations, and we were unable to obtain expression in E.*coli*.

In an attempt to improve the conformation and stabilization of S4_wt, we used Rosetta FunFolDes to fold and design this topology, while keeping the conformation of the site IV epitope fixed. Out of 25,000 simulations, the top 1 % decoys according to RE score and overall RMSD were selected for manual inspection, and 12 designed sequences were selected for recombinant expression in E.*coli*.

TopoBuilder - Motif-centric de novo design

Given the limited availability of suitable starting templates to host structurally complex motifs such as site 0 and site IV, we developed a template-free design protocol, which we named TopoBuilder (see Fig S3.6). In contrast to adapting an existing topology to accommodate the epitope, the design goal is to build protein scaffolds around the epitope from scratch, using idealized secondary structures (beta strands and alpha helices). The length, orientation and 3D-positioning are defined by the user for each secondary structure with respect to the epitope, which is extracted from its native environment. The topologies built were designed to meet the following criteria: (1) Small, globular proteins with a high

contact order between secondary structures and the epitope, to allow for stable folding and accurate stabilization of the epitope in its native conformation; (*276*) Context mimicry, i.e. respecting shape constraints of the epitope in its native context (Fig S3.5). For assembling the topology, the default distances between alpha helices was set to 11 Å and for adjacent beta-strands was 5 Å. For each discontinuous structural sketch, a connectivity between the secondary structural elements was defined and loop lengths were selected to connect the secondary structure elements with the minimal number of residues that can cover a given distance, while maintaining proper backbone geometries.

For site 0, the short helix of S0_1.39 preceding the epitope loop segment was kept, and a third helix was placed on the backside of the epitope to: (1) provide a core to the protein and (*276*) allow for the proper connectivity between the secondary structures. A total of three different orientations (45°, 0° and -45° degrees to the plane formed by site 0) where tested for the designed supporting alpha helix (Fig. 3.3).

In the case of site IV, the known binding region to 101F (residues 428F-434F) was extracted from prefusion RSVF (PDB 4JWH). Three antiparallel beta strands, pairing with the epitope, plus an alpha helix on the buried side, were assembled around the 101F epitope. Three different configurations (45°, (-45°,0°,10°) and -45° degrees with respect to the β-sheet) were sampled parametrically for the alpha helix (Fig. 3.3 and Fig S3.7).

The structural sketches were used to generate Cα distance constraints to guide Rosetta FunFolDes (*260*) folding trajectories. Around 25,000 trajectories were generated for each sketch. The newly generated backbones were further subjected to layer-based FastDesign (*234*), meaning that each amino acid position was assigned a layer (combining “core”, “boundary”, “surface” and “sheet” or “helix”) on the basis of its exposure and secondary structure type, that dictated the allowed amino acid types at that position.

After iterative cycles of sequence design, unconstraint FastRelax (*235*) (i.e. sidechain repacking and backbone minimization) was applied over the designs to evaluate their conformational stability of the epitope region. After each relax cycle, structural changes of the epitope region were evaluated (epitope RMSD drift). Designs with epitope RMSD drifts higher than 1.2 Å were discarded. Designs were also ranked and selected according hydrophobic core packing (packstat score), with a cutoff of 0.5 for site 0 and 0.6 for the site IV design series, and a cavity volume of < 50 $Å^3$. Between 1,000 and 10,000 of the designed sequences were generated from this computational protocol. We evaluated sequence profiles for the designs, and encoded the critical positions combinatorially by assembling overlapping oligos. Upon PCR assembly, libraries were transformed in yeast and screened for antibody binding and stability as assessed by protease digestion assays (*230, 264, 265*).

Immunization studies

Mouse immunizations

All animal experiments were approved by the Veterinary Authority of the Canton of Vaud (Switzerland) according to Swiss regulations of animal welfare (animal protocol number 3074). Female Balb/c mice (6-week old) were purchased from Janvier labs.

Immunogens were thawed on ice, mixed with equal volumes of adjuvant (2% Alhydrogel, Invivogen or Sigma Adjuvant System, Sigma) and incubated for 30 minutes. Mice were injected subcutaneously with 100 µl vaccine formulation, containing in total 10 µg of immunogen (equimolar ratios of each

immunogen for Trivax immunizations). Immunizations were performed on day 0, 21 and 42. 100-200 µl blood were drawn on day 0, 14 and 35. Mice were euthanized at day 56 and blood was taken by cardiac puncture.

NHP immunizations

Twenty-one african green monkeys (AGM, 3-4 years) were divided into three experimental groups with at least two animals of each sex. AGMs were pre-screened as seronegative against prefusion RSVF (preRSVF) by ELISA. Vaccines were prepared 1 hour before injection, containing 50 µg preRSVF or 300 µg Trivax1 in 0.5 ml PBS, mixed with 0.5 ml alum adjuvant (Alhydrogel, Invivogen) for each animal. AGMs were immunized intramuscularly at day 0, 28, 56, and 84. Blood was drawn at days 14, 28, 35, 56, 63, 84, 91, 105 and 119.

Serum analysis

Serum fractionation

Monomeric Trivax1 immunogens (S2_1, S0_1.39 and S4_1.5) were used to deplete the site 0, II and IV specific antibodies in immunized sera. HisPurTM Ni-NTA resin slurry (Thermo Scientific) was washed with PBS containing 10 mM imidazole. Approximately 1 mg of each immunogen was immobilized on Ni-NTA resin, followed by two wash steps to remove unbound scaffold. 60 µl of sera pooled from all animals within the same group were diluted to a final volume of 600 µl in wash buffer, and incubated overnight at 4 °C with 500 µl Ni-NTA resin slurry. As control, the same amount of sera was incubated with Ni-NTA resin that did not contain scaffolds. Resin was pelleted down at 13,000 rpm for 5 minutes, and the supernatant (depleted sera) was collected and then used for neutralization assays.

RSV neutralization assay

The RSV neutralization assay was performed as described previously (*256*). Briefly, Hep2 cells were seeded in Corning 96-well tissue culture plates (Sigma) at a density of 40,000 cells/well in 100 µl of Minimum Essential Medium (MEM, Gibco) supplemented with 10% FBS (Gibco), L-glutamine 2 mM (Gibco) and penicillin-streptomycin (Gibco), and grown overnight at 37 °C with 5% CO2. Serial dilutions of heat-inactivated sera were prepared in MEM without phenol red (M0, Life Technologies, supplemented with 2 mM L-glutamine and penicillin/streptomycin) and were incubated with 800 pfu/well (final MOI 0.01) RSV-Luc (A2 strain carrying a luciferase gene) for 1 hour at 37 °C. Serum-virus mixture was added to Hep2 cell layer, and incubated for 48 hours. Cells were lysed in lysis buffer supplemented with 1 µg/ml luciferin (Sigma) and 2 mM ATP (Sigma), and luminescence signal was read on a Tecan Infinite 500 plate reader. The neutralization curve was plotted and fitted using the GraphPad variable slope fitting model, weighted by $1/Y^2$.

Dissection of serum antibody specificities by SPR

To quantify the epitope-specific antibody responses in bulk serum from immunized animals, we performed an SPR competition assay with the monoclonal antibodies (D25, Motavizumab and 101F) as described previously (*256*). Briefly, approximately 400 RU of prefusion RSVF were immobilized on a CM5 chip via amine coupling, and serum diluted 1:10 in running buffer was injected to measure the total response ($RU_{non\text{-}blocked\ surface}$). After chip regeneration using 50 mM NaOH, the site 0/II/IV epitopes were blocked by injecting saturating amounts of either D25, Motavizumab, or 101F IgG, and serum was

injected again to quantify residual response ($RU_{\text{blocked surface}}$). The delta serum response (ΔSR) was calculated as follows:

$$\Delta\text{SR} = \text{RU}_{\text{(non-)blocked surface}} - \text{RU}_{\text{Baseline}}$$

Percent blocking was calculated for each site as:

$$\%\ \text{blocking} = \left(1 - \left(\frac{\Delta\text{SR}_{\text{blocked surface}}}{\Delta\text{SR}_{\text{non-blocked surface}}}\right)\right) * 100$$

Competition ELISA

Prior to incubation with a coated antigen plate, sera were serially diluted in the presence of 100 μg/ml competitor antigen and incubated overnight at 4°C. ELISA curves of a positive control, Motavizumab, are shown Fig S2.10. Curves were plotted using GraphPad Prism, and the area under the curve (*197*) was calculated for the specific (NRM) and control (RSVN) competitor. % competition was calculated using the following formula (*277*):

$$\%\ \text{competition} = \left(1 - \left(\frac{AUC(specific\ competitor\ (NRM))}{AUC(control\ competitor\ (NR))}\right)\right) * 100$$

In vitro evolution

Yeast surface display

Libraries of linear DNA fragments encoding variants of the designed proteins were transformed together with linearized pCTcon2 vector (Addgene #41843) based on the protocol previously described by Chao and colleagues (*278*). Transformation procedures generally yielded ~10^7 transformants. The transformed cells were passaged twice in SDCAA medium before induction. To induce cell surface expression, cells were centrifuged at 7,000 r.p.m. for 1 min, washed with induction media (SGCAA) and resuspended in 100 ml SGCAA with a cell density of 1 x 10^7 cells/ml SGCAA. Cells were grown overnight at 30 °C in SGCAA medium. Induced cells were washed in cold wash buffer (PBS + 0.05% BSA) and labelled with various concentration of target IgG or Fab (101F, D25, and 5C4) at 4°C. After one hour of incubation, cells were washed twice with wash buffer and then incubated with FITC-conjugated anti-cMyc antibody and PE-conjugated anti-human Fc (BioLegend, #342303) or PE-conjugated anti-Fab (Thermo Scientific, #MA1-10377) for an additional 30 min. Cells were washed and sorted using a SONY SH800 flow cytometer in "ultra-purity" mode. The sorted cells were recovered in SDCAA medium, and grown for 1-2 days at 30 °C.

In order to select stably folded proteins, we washed the induced cells with TBS buffer (20 mM Tris, 100 mM NaCl, pH 8.0) three times and resuspended in 0.5 ml of TBS buffer containing 1 μM of chymotrypsin. After incubating five-minutes at 30°C, the reaction was quenched by adding 1 ml of wash buffer, followed by five wash steps. Cells were then labelled with primary and secondary antibodies as described above.

Site saturation mutagenesis library (*279*)

A SSM library was assembled by overhang PCR, in which 11 selected positions surrounding the epitope in the S4_1.1 design model were allowed to mutate to all 20 amino acids, with one mutation allowed at a time. Each of the 11 libraries was assembled by primers (Table S1) containing the degenerate codon "NNK" at the selected position. All 11 libraries were pooled, and transformed into EBY-100 yeast strain with a transformation efficiency of $1x10^6$ transformants.

Combinatorial library

Combinatorial sequence libraries were constructed by assembling multiple overlapping primers (Table S2) containing degenerate codons at selected positions for combinatorial sampling of hydrophobic amino acids in the protein core. The theoretical diversity was between $1x10^6$ and $5x10^6$. Primers were mixed (10 μM each), and assembled in a PCR reaction (55 °C annealing for 30 sec, 72 °C extension time for 1 min, 25 cycles). To amplify full-length assembled products, a second PCR reaction was performed, with forward and reverse primers specific for the full-length product. The PCR product was desalted, and transformed into EBY-100 yeast strain with a transformation efficiency of at least $1x10^7$ transformants (*278*).

Next-generation sequencing of design pools

After sorting, yeast cells were grown overnight, pelleted and plasmid DNA was extracted using Zymoprep Yeast Plasmid Miniprep II (Zymo Research) following the manufacturer's instructions. The coding sequence of the designed variants was amplified using vector-specific primer pairs, Illumina sequencing adapters were attached using overhang PCR, and PCR products were desalted (Qiaquick PCR purification kit, Qiagen). Next generation sequencing was performed using an Illumina MiSeq 2x150bp paired end sequencing (300 cycles), yielding between 0.45-0.58 million reads/sample.

For bioinformatic analysis, sequences were translated in the correct reading frame, and enrichment values were computed for each sequence. We defined the enrichment value E as follows:

$$E_{Seq} = \frac{count_{Seq}(high\ selective\ pressure)}{count_{Seq}(low\ selective\ pressure)}$$

The high selective pressure corresponds to low labelling concentration of the respective target antibodies (100 pM D25, 10 nM 5C4 or 20 pM 101F, as shown in Fig. 3.3), or a higher concentration of chymotrypsin protease (0.5 μM). The low selective pressure corresponds to a high labelling concentration with antibodies (10 nM D25, 1 μM 5C4 or 2 nM 101F), or no protease digestion, as indicated in Fig. 3.3. Only sequences that had at least one count in both sorting conditions were included in the analysis.

Protein expression and purification

Designed scaffolds

All genes of designed proteins were purchased as DNA fragments from Twist Bioscience, and cloned via Gibson assembly into either pET11b or pET21b bacterial expression vectors. Plasmids were transformed into E.*coli* BL21 (DE3) (Merck) and grown overnight in LB media. For protein expression, precultures were diluted 1:100 and grown at 37 °C until the OD_{600} reached 0.6, followed by the addition of 1 mM IPTG to induce expression. Cultures were harvested after 12-16 hours at 22 °C. Pellets were resuspended in lysis buffer (50 mM Tris, pH 7.5, 500 mM NaCl, 5% Glycerol, 1 mg/ml lysozyme, 1 mM

PMSF, 1 µg/ml DNase) and sonicated on ice for a total of 12 minutes, in intervals of 15 seconds sonication followed by 45 seconds pause. Lysates were clarified by centrifugation (20,000 rpm, 20 minutes) and purified via Ni-NTA affinity chromatography followed by size exclusion on a HiLoad 16/600 Superdex 75 column (GE Healthcare) in PBS buffer.

Ferritin-based immunogens

The gene encoding Helicobacter pylori ferritin (GenBank ID: QAB33511.1) was cloned into the pHLsec vector for mammalian expression, with an N-terminal 6x His Tag. The sequence of the designed immunogens (S0_2.126 and S4_2.45) were cloned upstream of the ferritin gene, spaced by a GGGGS linker. Ferritin particulate immunogens were produced by co-transfecting a 1:1 stochiometric ratio of "naked" ferritin and immunogen-ferritin in HEK-293F cells, as previously described for other immunogen-nanoparticle fusion constructs (*280*). The supernatant was collected 7-days post transfection and purified via Ni-NTA affinity chromatography and size exclusion on a Superose 6 increase 10/300 GL column (GE).

NRM

The full-length N gene (sequence derived from the human RSV strain Long, ATCC VR-26; GenBank accession number AY911262.1) was PCR amplified using the Phusion DNA polymerase (Thermo Scientific) and cloned into pET28a+ at NcoI-XhoI sites to obtain the pET-N plasmid. The sequence of FFLM was then PCR amplified and cloned into pET-N at NcoI site to the pET-NRM plasmid. *E. coli* BL21 (DE3) bacteria were co-transformed with pGEX-PCT (*281*) and pET-FFLM-N plasmids and grown in LB medium containing ampicillin (100 µg/ml) and kanamycin (50 µg/ml). The same volume of LB medium was then added, and protein expression was induced by the addition of 0.33 mM IPTG to the medium. Bacteria were incubated for 15 h at 28°C and then harvested by centrifugation. For protein purification, bacterial pellets were resuspended in lysis buffer (50 mM Tris-HCl pH 7.8, 60 mM NaCl, 1 mM EDTA, 2 mM dithiothreitol, 0.2% Triton X-100, 1 mg/ml lysozyme) supplemented with a complete protease inhibitor cocktail (Roche), incubated for one hour on ice, and disrupted by sonication. The soluble fraction was collected by centrifugation at 4 °C for 30 min at 10,000 x g. Glutathione-Sepharose 4B beads (GE Healthcare) were added to clarify supernatants and incubated at 4°C for 15h. The beads were then washed one time in lysis buffer and two times in 20 mM Tris pH 8.5, 150 mM NaCl. To isolate NRM, beads containing bound complex were incubated with thrombine for 16 h at 20 °C. After cleavage of the GST tag, the supernantant was loaded onto a Sephacryl S-200 HR 16/30 column (GE Healthcare) and eluted in 20 mM Tris-HCl, 150 mM NaCl, pH 8.5.

Antibody variable fragments (Fabs)

For Fab expression, heavy and light chain DNA sequences were purchased from Twist Biosciences and cloned separately into the pHLSec mammalian expression vector (Addgene, #99845) using AgeI and XhoI restriction sites. Expression plasmids were pre-mixed in a 1:1 stoichiometric ratio, co-transfected into HEK293-F cells and cultured in FreeStyle™ medium (Gibco, #12338018). Supernatants were harvested after one week by centrifugation and purified using a kappa-select column (GE Healthcare). Elution of bound proteins was conducted using 0.1 M glycine buffer (pH 2.7) and eluates were immediately neutralized by the addition of 1 M Tris ethylamine (pH 9), followed by buffer exchange to PBS pH 7.4.

Respiratory Syncytial Virus Fusion protein (prefusion RSVF)

Protein sequence of prefusion RSVF corresponds to the sc9-10 DS-Cav1 A149C Y458C S46G E92D S215P K465Q variant designed by Joyce *et al*. (*282*), which we refer to as RSVF DS2. RSVF DS2 was codon optimized for mammalian expression and cloned into the pHCMV-1 vector together with two C-terminal Strep-Tag II and one 8x His tag. Plasmids were transfected in HEK293-F cells and cultured in FreeStyle™ medium. Supernatants were harvested one week after transfection and purified via Ni-NTA affinity chromatography. Bound protein was eluted using buffer containing 10 mM Tris, 500 mM NaCl and 300 mM Imidazole (pH 7.5), and eluate was further purified on a StrepTrap HP affinity column (GE Healthcare). Bound protein was eluted in 10mM Tris, 150 mM NaCl and 20 mM Desthiobiotin (Sigma), pH 8, and size excluded in PBS, pH 7.4, on a Superdex 200 Increase 10/300 GL column (GE Healthcare) to obtain trimeric RSVF.

Negative-stain transmission electron microscopy

Sample preparation

RSVN and Ferritin- based nanoparticles were diluted to a concentration of 0.015 mg/ml. The samples were absorbed on carbon-coated copper grid (EMS, Hatfield, PA, United States) for 3 mins, washed with deionized water and stained with freshly prepared 0.75 % uranyl formate.

Data acquisition

The samples were viewed under an F20 electron microscope (Thermo Fisher) operated at 200 kV. Digital images were collected using a direct detector camera Falcon III (Thermo Fisher) with the set-up of 4098 X 4098 pixels. The homogeneity and coverage of staining samples on the grid was first visualized at low magnification mode before automatic data collection. Automatic data collection was performed using EPU software (Thermo Fisher) at a nominal magnification of 50,000X, corresponding to pixel size of 2 Å, and defocus range from -1 μm to -2 μm.

Image processing

CTFFIND4 program (*283*) was used to estimate the contrast transfer function for each collected image. Around 1000 particles were manually selected using the installed package XMIPP within SCIPION framework (*284*). Manually picked particles were served as input for XMIPP auto-picking utility, resulting in at least 10,000 particles. Selected particles were extracted with the box size of 100 pixels and subjected for three rounds of reference-free 2D classification without CTF correction using RELION-3.0 Beta suite (*285*).

RSVF-Fabs complex formation and negative stain EM

20 μg of RSVF trimer was incubated overnight at 4°C with 80 μg of Fabs (Motavizumab, D25 or 101F). For complex formation with all three monoclonal Fabs, 80 μg of each Fab was used. Complexes were purified on a Superose 6 Increase 10/300 column using an Äkta Pure system (GE Healthcare) in TBS buffer. The main fraction containing the complex was directly used for negative stain EM.

Purified complexes of RSVF and Fabs were deposited at approximately 0.02 mg/ml onto carbon-coated copper grids and stained with 2% uranyl formate. Images were collected with a field-emission FEI Tecnai F20 electron microscope operating at 200 kV. Images were acquired with an Orius charge-coupled device (CCD) camera (Gatan Inc.) at a calibrated magnification of ×34,483, resulting in a pixel size of

2.71 Å. For the complexes of RSVF with a single Fab, approximately 2,000 particles were manually selected with Cryosparc2 (*286*). Two rounds of 2D classification of particle images were performed with 20 classes allowed. For the complexes of RSVF with D25, Motavizumab and 101F Fabs, approximately 330,000 particles were picked using Relion 3.0 (*285*) and subsequently imported to Cryosparc2 for two rounds of 2D classification with 50 classes allowed.

Biophysical characterization of designed proteins

SEC-MALS

Size exclusion chromatography with an online multi-angle light scattering (MALS) device (miniDAWN TREOS, Wyatt) was used to determine the oligomeric state and molecular weight for the protein in solution. Purified proteins were concentrated to 1 mg/ml in PBS (pH 7.4), and 100 µl of sample was injected into a Superdex 75 300/10 GL column (GE Healthcare) with a flow rate of 0.5 ml/min, and UV_{280} and light scattering signals were recorded. Molecular weight was determined using the ASTRA software (version 6.1, Wyatt).

Circular Dichroism

Far-UV circular dichroism spectra were measured using a Jasco-815 spectrometer in a 1 mm path-length cuvette. The protein samples were prepared in 10 mM sodium phosphate buffer at a protein concentration of 30 µM. Wavelengths between 190 nm and 250 nm were recorded with a scanning speed of 20 nm min^{-1} and a response time of 0.125 sec. All spectra were averaged 2 times and corrected for buffer absorption. Temperature ramping melts were performed from 25 to 90 °C with an increment of 2 °C/min in presence or absence of 2.5 mM TCEP reducing agent. Thermal denaturation curves were plotted by the change of ellipticity at the global curve minimum to calculate the melting temperature (*223*).

Determining binding affinities by Surface plasmon resonance (SPR)

SPR measurements were performed on a Biacore 8K (GE Healthcare) with HBS-EP+ as running buffer (10 mM HEPES pH 7.4, 150 mM NaCl, 3 mM EDTA, 0.005% v/v Surfactant P20, GE Healthcare). Ligands were immobilized on a CM5 chip (GE Healthcare # 29104988) via amine coupling. Approximately 2000 response units (RU) of IgG were immobilized, and designed monomeric proteins were injected as analyte in two-fold serial dilutions. The flow rate was 30 µl/min for a contact time of 120 seconds followed by 400 seconds dissociation time. After each injection, surface was regenerated using 3 M magnesium chloride (101F as immobilized ligand) or 0.1 M Glycine at pH 4.0 (Motavizumab and D25 IgG as an immobilized ligand). Data were fitted using 1:1 Langmuir binding model within the Biacore 8K analysis software (GE Healthcare #29310604).

Structural characterization by NMR and X-ray crystallography

NMR

Protein samples for NMR were prepared in 10 mM sodium phosphate buffer, 50 mM sodium chloride at pH 7.4 with the protein concentration of 500 µM. All NMR experiments were carried out in a 18.8T (800 MHz proton Larmor frequency) Bruker spectrometer equipped with a CPTC $^{1}H,^{13}C,^{15}N$ 5 mm cryoprobe

and an Avance III console. Experiments for backbone resonance assignment consisted in standard triple resonance spectra HNCA, HN(CO)CA, HNCO, HN(CO)CA, CBCA(CO)NH and HNCACB acquired on a 0.5 mM sample doubly labelled with ^{13}C and ^{15}N (*287*). Sidechain assignments were obtained from HCCH-TOCSY experiments acquired on the same sample plus HNHA, NOESY-^{15}N-HSQC and TOCSY-^{15}N-HSQC acquired on a ^{15}N-labeled sample. The NOESY-^{15}N-HSQC was used together with a 2D NOESY collected on an unlabelled sample for structure calculations. Spectra for backbone assignments were acquired with 40 increments in the ^{15}N dimension and 128 increments in the ^{13}C dimension, and processed with 128 and 256 points by using linear prediction. HCCH-TOCSY were recorded with 64-128 increments in the ^{13}C dimensions and processed with twice the number of points. ^{15}N-resolved NOESY and TOCSY spectra were acquired with 64 increments in ^{15}N dimension and 128 in the indirect ^{1}H dimension, and processed with twice the number of points. ^{1}H-^{1}H 2D-NOESY and 2D TOCSY spectra were acquired with 256 increments in the indirect dimension, processed with 512 points. Mixing times for NOESY spectra were 100 ms and TOCSY spin locks were 60 ms. Heteronuclear ^{1}H-^{15}N NOE was measured with 128 ^{15}N increments processed with 256 points, using 64 scans and a saturation time of 6 seconds. All samples were prepared in 20 mM phosphate buffer pH 7, with 10% $^{2}H_2O$ and 0.2% sodium azide to prevent sample degradation.

All spectra were acquired and processed with Bruker's TopSpin 3.0 (acquisition with standard pulse programs) and analyzed manually with the program CARA (http://cara.nmr.ch/doku.php/home) to obtain backbone and sidechain resonance assignments. Peak picking and assignment of NOESY spectra (a ^{15}N-resolved NOESY and a 2D NOESY) were performed automatically with the program UNIO-ATNOS/CANDID (*288, 289*) coupled to Cyana 2.1 (*290*), using standard settings in both programs. The run was complemented with dihedral angles derived from chemical shifts with Talos-n (*291*). The solution NMR structure of S0_2.126 has been deposited in the Protein Data Bank under accession code 6S28.

NMR restraints	
Total NOEs from Unio [a]	306
Intraresidual	124
Interresidual	182
Sequential (i – j = 1)	112
Medium-range (1 < i – j < 5)	47
Long-range (i – j ≥ 5)	23
Dihedral Angles from Talos-n [b]	88
ϕ	43
ψ	45
Structural statistics	
Violations [c]	
Distance restraints (Å)	0.0254 ± 0.009
Dihedral angle constraints (°)	6.8 ± 0.12
Ramachandran plot (all residues/ordered residues) [d]	
Most favored (%)	84.7 / 95.8
Additionally allowed (%)	14.3 / 4.5
Generously allowed (%)	0.98 / 0.1
Disallowed (%)	0 / 0
Average pairwise RMSD (Å) [e]	
Heavy	3.3 / 1.8
Backbone	2.8 / 1.2
Structure Quality Factors (raw score/z-score) [e]	
Procheck G-factor (phi/psi)	0.15 / 0.9
Procheck G-factor (all)	-0.48 / -2.84
[a] From UNIO-ATNOS/CANDID's last cycle (cycle 7)	

[b] Obtained from chemical shifts with Talos-N server
[c] From Cyana in Unio's last cycle
[d] All residues from Cyana un Unio's last cycle; ordered residues (5-22,26-57) from the Protein Structure Validation Suite at http://psvs-1_5-dev.nesg.org/results/testbc/OUTPUT.html
[e] From the Protein Structure Validation Suite

Table S 3.1. Refinement statistics of the S0_2.126 NMR structure.

Co-crystallization of complex D25 Fab with S0_2.126

After overnight incubation at 4°C, the S0_2.126/D25 Fab complex was purified by size exclusion chromatography using a Superdex200 26 600 (GE Healthcare) equilibrated in 10 mM Tris pH 8, 100 mM NaCl and subsequently concentrated to ~10 mg/ml (Amicon Ultra-15, MWCO 3,000). Crystals were grown at 291K using the sitting-drop vapor-diffusion method in drops containing 1 µl purified protein mixed with 1 µl reservoir solution containing 10% PEG 8000, 100 mM HEPES pH 7.5, and 200 mM calcium acetate. For cryo protection, crystals were briefly swished through mother liquor containing 20% ethylene glycol.

Data collection and structural determination of the S0_2.126/D25 Fab complex

Diffraction data was recorded at ESRF beamline ID30B. Data integration was performed by XDS (*292*) and a high-resolution cut at I/σ=1 was applied. The dataset contained a strong off-origin peak in the Patterson function (88% height rel. to origin) corresponding to a pseudo translational symmetry of ½, 0, ½. The structure was determined by the molecular replacement method using PHASER (*293*) using the D25 structure (*257*) (PDB ID 4JHW) as a search model. Manual model building was performed using Coot (*294*), and automated refinement in Phenix (*295*). After several rounds of automated refinement and manual building, paired refinement (*296*) determined the resolution cut-off for final refinement (Table S3.2).

D25 S0_2.126	
Wavelength	0.9763
Resolution range	49.09-3.0 (3.107-3.0)
Space group	P 21 21 21
Unit cell	126.3 127.0 156.1 90 90 90
Total reflections	700184 (72248)
Unique reflections	50740 (5000)
Multiplicity	13.8 (14.4)
Completeness (%)	98.76 (99.22)
Mean I/sigma(I)	12.63 (2.00)
Wilson B-factor	74.78
R-merge	0.1622 (1.484)
R-meas	0.1684 (1.538)
R-pim	0.04506 (0.4019)
CC1/2	0.999 (0.893)
CC*	1 (0.971)
Reflections used in refinement	50284 (4971)
Reflections used for R-free	2519 (249)

R-work	0.2699 (0.3677)
R-free	0.2936 (0.3972)
CC(work)	0.949 (0.817)
CC(free)	0.958 (0.793)
Number of non-hydrogen atoms	14453
macromolecules	14452
Protein residues	1921
RMS(bonds)	0.004
RMS(angles)	1.02
Ramachandran favored (%)	94.45
Ramachandran allowed (%)	5.07
Ramachandran outliers (%)	0.48
Rotamer outliers (%)	0.00
Clashscore	7.35
Average B-factor	97.74
macromolecules	97.74
solvent	59.33
Number of TLS groups	12

Table S 3.2. X-ray data collection and refinement statistics of S0_2.126 crystal structure

Co-crystallization of complex 101F Fab with S4_2.45

The complex of S4_2.45 with the F101 Fab was prepared by mixing two proteins in 2:1 molar ratio for 1 hour at 4 °C, followed by size exclusion chromatography using a Superdex-75 column. Complexes of S4_2.45 with the 101F Fab were verified by SDS–PAGE. Complexes were subsequently concentrated to 6–8 mg/ml. Crystals were grown using hanging drops vapor-diffusion method at 20 °C. The S4_2.45/101F protein complex was mixed with equal volume of a well solution containing 0.2 M Magnesium acetate, 0.1 M Sodium cacodylate pH 6.5, 20 %(w/v) PEG 8000. Native crystals were transferred to a cryoprotectant solution of 0.2 M Magnesium acetate, 0.1 M Sodium cacodylate pH 6.5, 20 % (w/v) PEG 8000 and 15% glycerol, followed by flash-cooling in liquid nitrogen.

Data collection and structural determination of the S4_2.45/101F Fab complex

Diffraction data were collected at SSRL facility, BL9-2 beamline at the SLAC National Accelerator Laboratory. The crystals belonged to space group P3221. The diffraction data were initially processed to 2.6 Å with X-ray Detector Software (XDS) (Table S3.3).

Molecular replacement searches were conducted with the program PHENIX PHASER using 101F Fab model (PDB ID: 3O41) and S4_2.45/101F Fab computational model generated from superimposing epitope region of S4_2.45 with the peptide-bound structure (PDB ID: 3O41), and yielded clear molecular replacement solutions. Initial refinement provided a R_{free} of 42.43% and R_{work} of 32.25% and a complex structure was refined using Phenix Refine, followed by manual rebuilding with the program COOT. The final refinement statistics, native data and phasing statistics are summarized in Table S3.3.

101F S4_2.45	
Wavelength	0.98

Resolution range	38.49 - 2.6 (2.693 - 2.6)
Space group	P 32 2 1
Unit cell	148.224 148.224 45.046 90 90 120
Total reflections	113069 (7302)
Unique reflections	17464 (1567)
Multiplicity	6.5 (4.7)
Completeness (%)	98.57 (89.58)
Mean I/sigma(I)	17.03 (1.66)
Wilson B-factor	56.09
R-merge	0.06712 (0.8361)
R-meas	0.07282 (0.9424)
R-pim	0.02776 (0.4231)
CC1/2	0.999 (0.635)
CC*	1 (0.881)
Reflections used in refinement	17455 (1565)
Reflections used for R-free	1748 (166)
R-work	0.2298 (0.3682)
R-free	0.2736 (0.3503)
CC(work)	0.462 (0.203)
CC(free)	0.353 (0.190)
Number of non-hydrogen atoms	3794
macromolecules	3686
solvent	108
Protein residues	485
RMS(bonds)	0.010
RMS(angles)	1.46
Ramachandran favored (%)	93.53
Ramachandran allowed (%)	5.64
Ramachandran outliers (%)	0.84
Rotamer outliers (%)	0.96
Clashscore	2.19
Average B-factor	38.90
macromolecules	38.37
solvent	56.78
Number of TLS groups	3

Table S 3.3. X-ray data collection and refinement statistics of S4_2.45 crystal structure.

3.6 Supplementary information

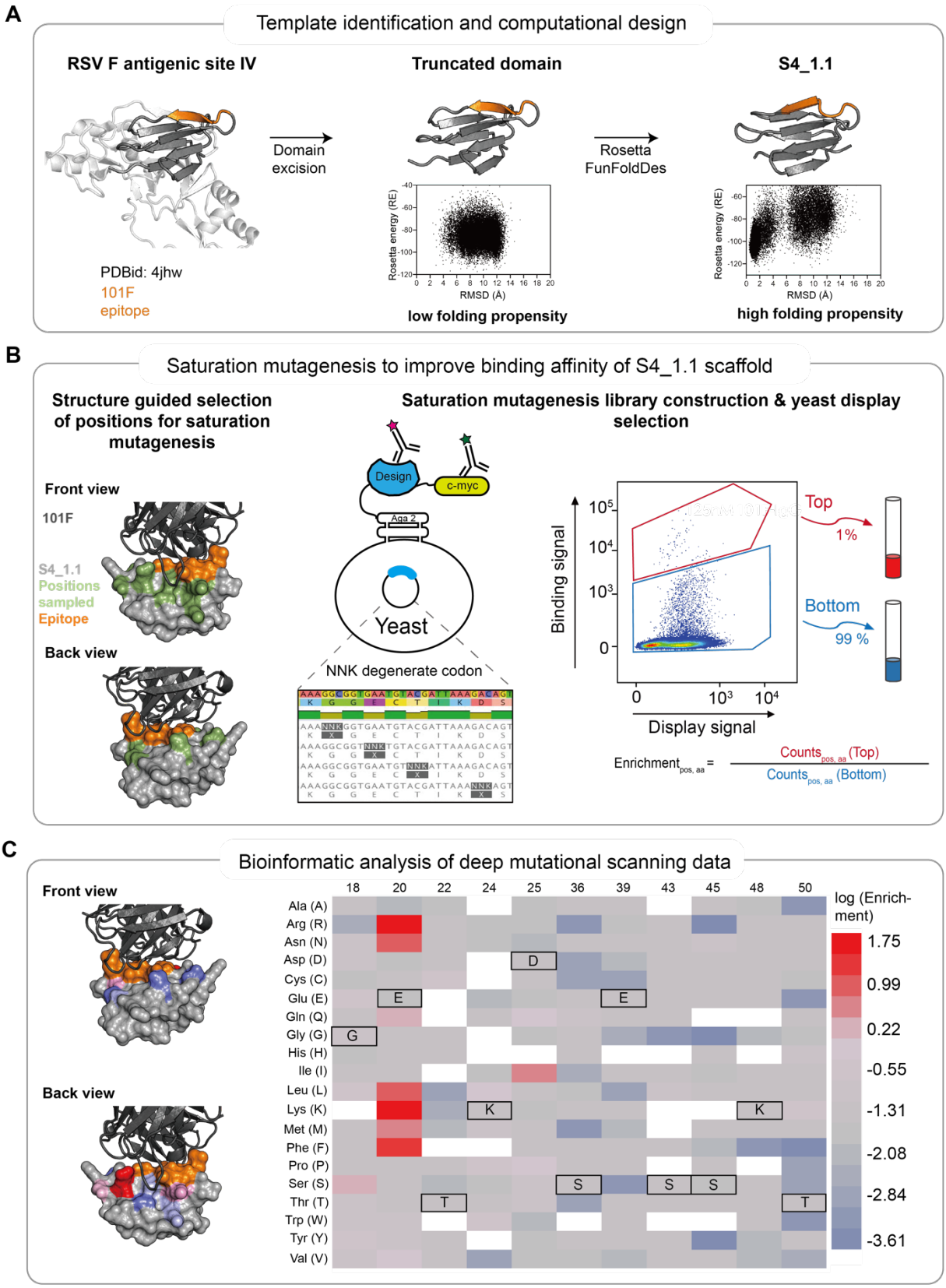


Figure S 3.1. Computational design and experimental optimization of S4_1 design series.

A) Template identification and computational design of S4_1.1. RSVF antigenic site IV is located in a small contained domain of preRSVF. This excised domain failed to show a folding funnel in Rosetta *ab initio* predictions, and failed to express recombinantly in E.*coli*. Using the excised domain as a template, we folded and sequence-designed

this topology using Rosetta FunFolDes, yielding design S4_1.1 which showed a strong funnel-shape energy landscape in *ab initio* folding simulation. B) Experimental optimization of S4_1.1 through saturation mutagenesis. A saturation mutagenesis library was constructed using overhang PCR for 11 positions proximal to the site IV epitope (orange), allowing one position at a time to mutate to any of the 20 amino acids, encoded by the degenerate codon "NNK". The library (size 11 positions x 32 codons = 352) was transformed in yeast, and designs were displayed on the cell surface. The selection was done by labeling the cells with 125 nM of 101F antibody. The top 1 % of clones binding with high affinity to 101F antibody were then sorted, as well as the bottom 99 % as shown. Following next-generation sequencing of the two populations, the enrichment values were computed for each sequence variant, corresponding to the relative abundance of each variant in the top versus bottom gate. C) Bioinformatic analysis of deep mutational scanning data. The log(enrichment) is shown as heatmap (right) for each sequence variant, and mapped to the structure (left). White indicates missing data. Position 20 showed the highest enrichment for arginine and lysine, together with other less pronounced enrichments seen for other positions.

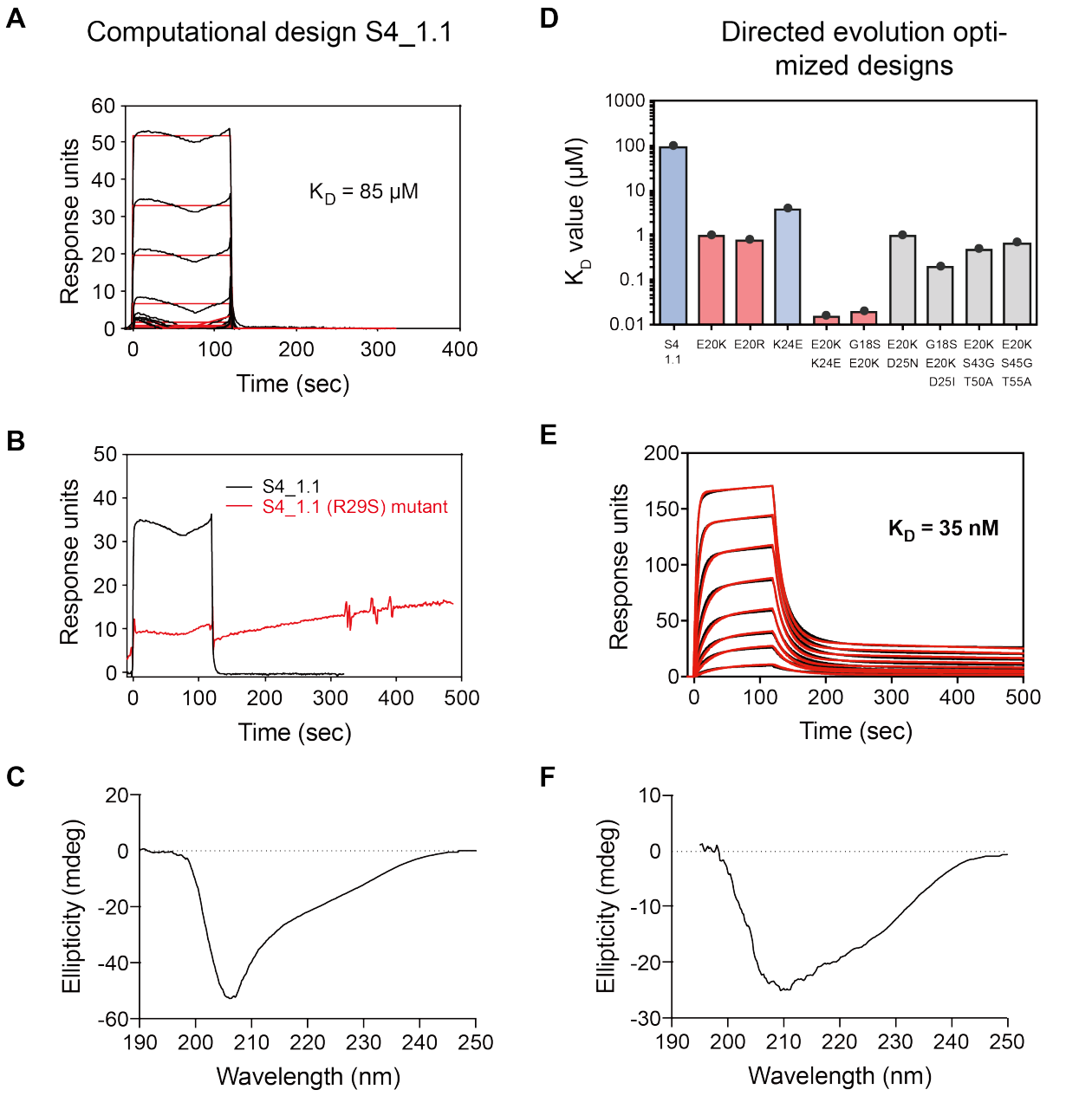


Figure S 3.2. Experimental characterization of S4_1 design series.

A) Surface plasmon resonance measurement for the initial computational design S4_1.1 against 101F antibody revealed a dissociation constant of > 85 μM. B) Despite low affinity, an R29S mutant revealed that binding was specific to the epitope of interest. C) Circular dichroism spectrum of S4_1.1 at 20 ºC. D) Dissociation constants (K_D) for single and combined mutations of S4_1.1 that were identified in the deep mutational scanning screen. E20K/K24E double mutant (named S4_1.5) showed a binding affinity of 35 nM. E) SPR sensorgram of S4_1.5 against 101F. F) Circular dichroism spectrum of S4_1.5 at 20 ºC.

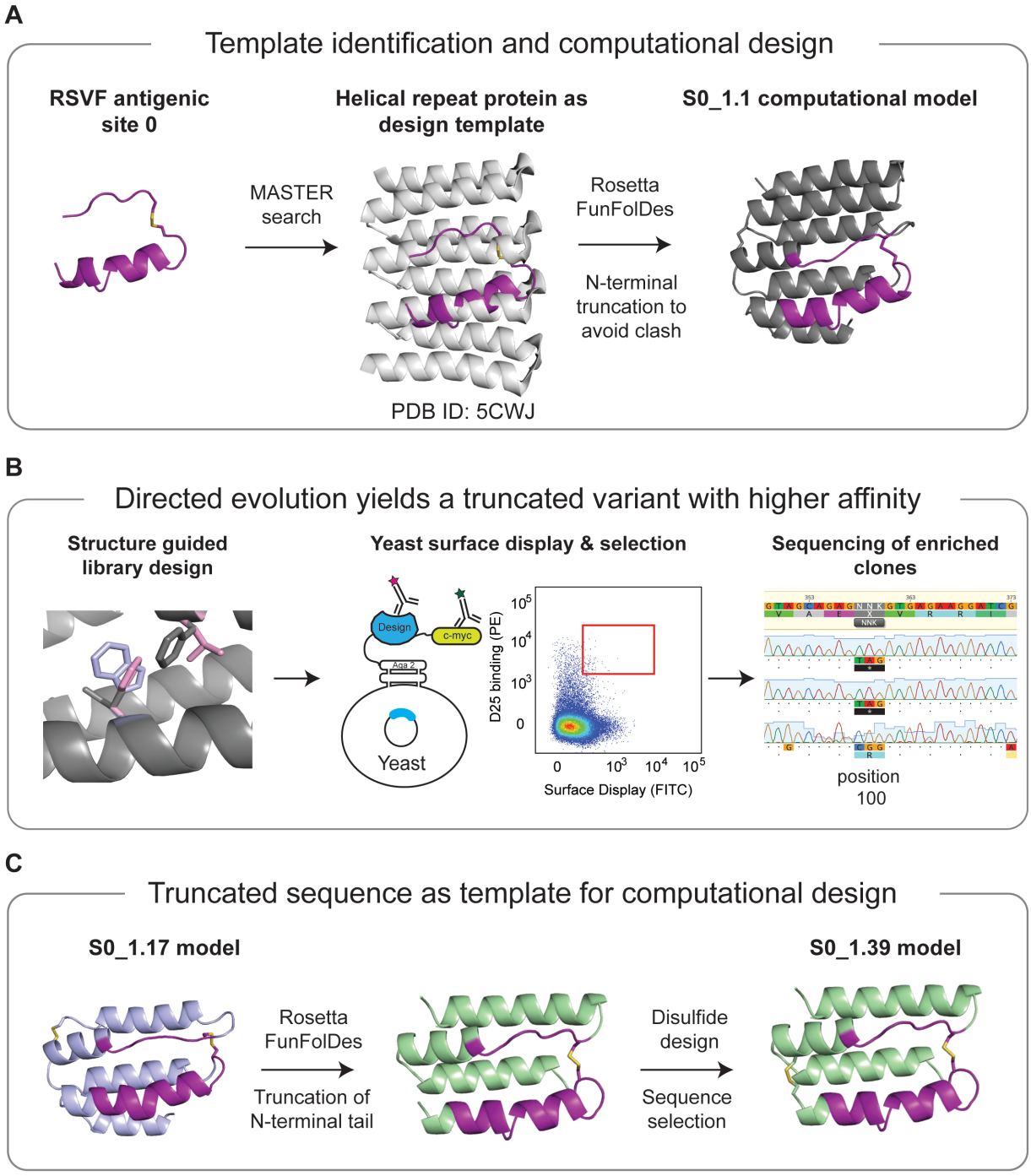


Figure S 3.3. Computational design and experimental optimization of S0_1 design series.

A) Template identification and design. Using MASTER, we identified a designed helical repeat protein (PDB ID: 5CWJ) to serve as a design template to present and stabilize antigenic site 0 (see methods for details). The N-terminal 29 residues were truncated to avoid clashing with the D25 antibody, and Rosetta FunFolDes was used to design S0_1.1. See methods for details on the design process. B) Based on S0_1.1, a combinatorial sequence library

was constructed and screened using yeast surface display. After three consecutive sorts of high-affinity binding clones, individual colonies were sequenced. Position 100 was frequently found to be mutated to a stop codon, leading to a truncated variant with increased expression yield, and a ~5-fold improved binding affinity to D25 (Fig S3.4). C) A model of the truncated variant served as a template for a second round of in silico folding and design. We truncated the template further by the N-terminal 14 residues, and introduced a disulfide bond between residues 1 and 43, leading to S0_1.39. See methods for full details on the design selection process.

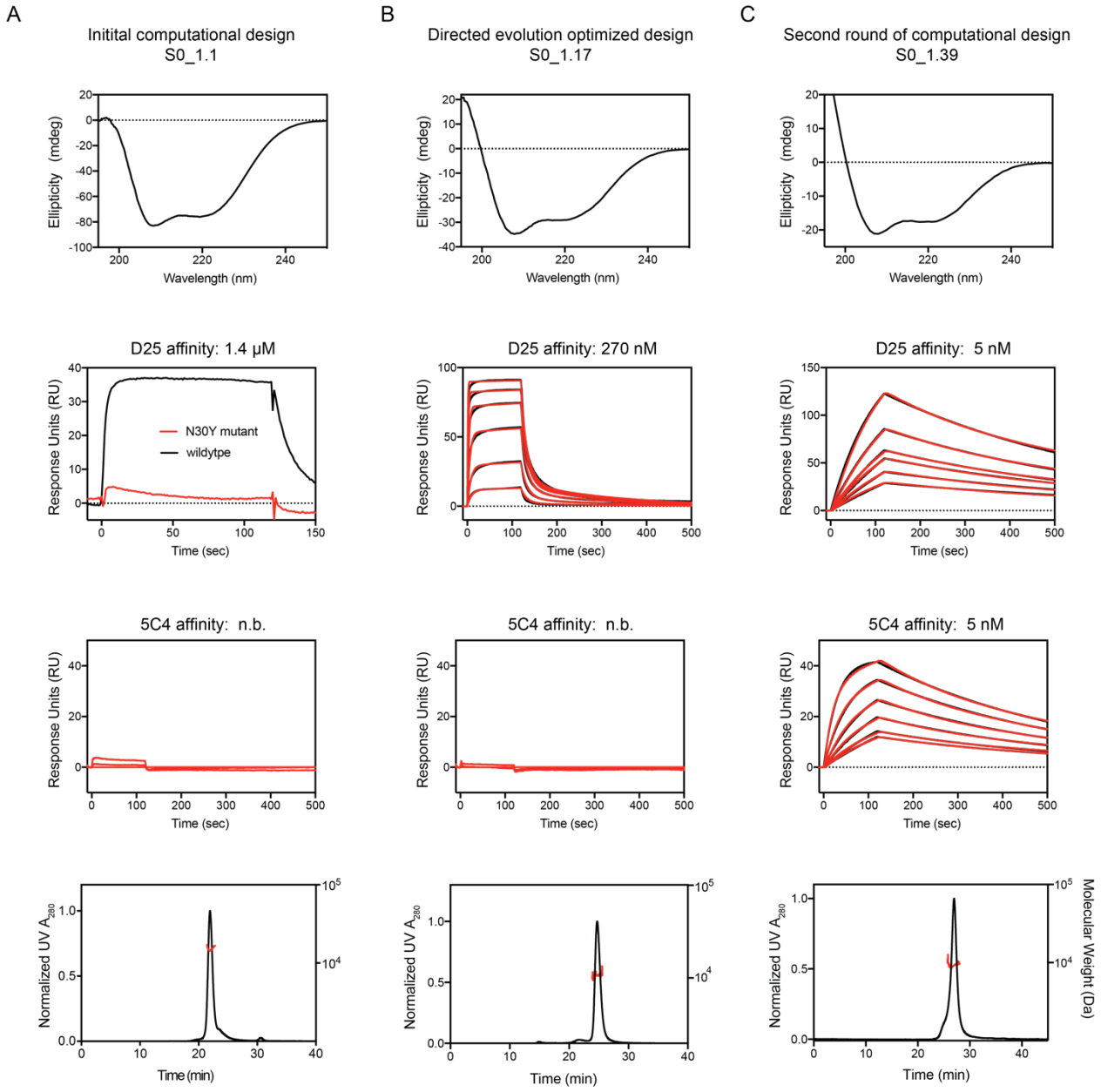


Figure S 3.4. Biophysical characterization of the S0_1 design series.

Top: Circular dichroism spectra at 20 ºC. Middle: Surface plasmon resonance measurements against D25 and 5C4. Bottom: Multi-angle light scattering coupled to size exclusion chromatography. A) S0_1.1 bound with a KD of 1.4 µM to D25 and no detectable binding to 5C4. To verify that the binding interaction was specific to the epitope we generated a knockout mutant (N30Y) and observed that the binding interaction was absent. B) S0_1.17 showed a KD of 270 nM to D25 and no binding to 5C4. C) SPR sensorgrams of S0_1.39 binding to D25 and 5C4 antibodies. D25 or 5C4 IgG was immobilized as ligand on the sensor chip surface, and S0_1.39 was flown as analyte. All designs shown CD spectra typical of helical proteins and behaved as monomers in the solution (Top and bottom rows).

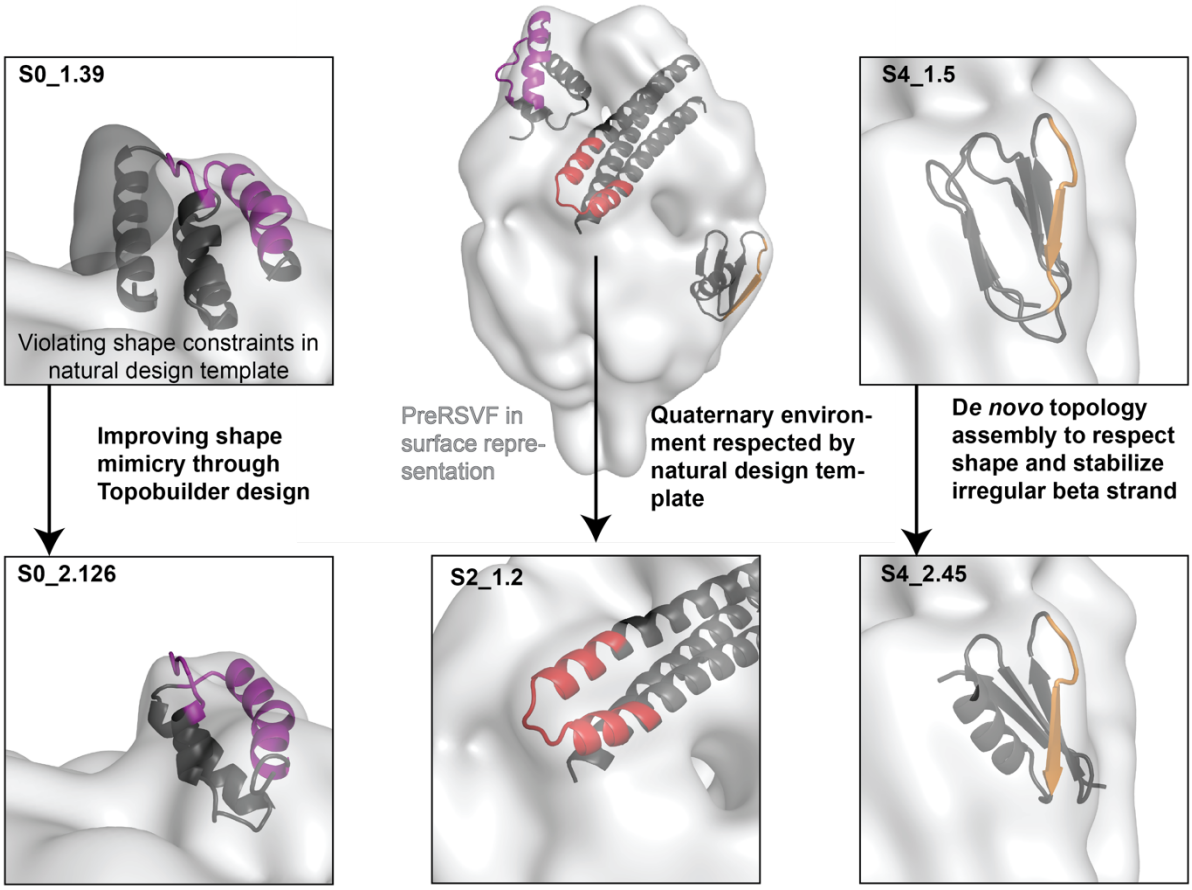


Figure S 3.5. Shape mimicry of computationally designed immunogens compared to prefusion RSVF.

Prefusion RSVF is shown in surface representation (light grey), with designed immunogens superimposed. Close-up views are shown for template-based designs (S0_1.39 and S4_1.5, top row). While site 0 is freely accessible for antibody binding in preRSVF, the C-terminal helix of S0_1.39 constrains its accessibility (dark grey surface). Through defined back-bone assembly using TopoBuilder, S0_2.126 was designed, mimicking the native quaternary environment of site 0 (bot-tom left). RSVF antigenic site II, which is a structurally simple helix-turn-helix motif frequently found in natural proteins, was previously designed based on a design template that respects the quaternary constraints of site II in its native environment (S2_1.2, bottom middle). For site IV, a topology was assembled (S4_2.45) that respects the shape constraints while improving the stabilization of the irregular, bulged beta strand compared to the S4_1.5 design (right).

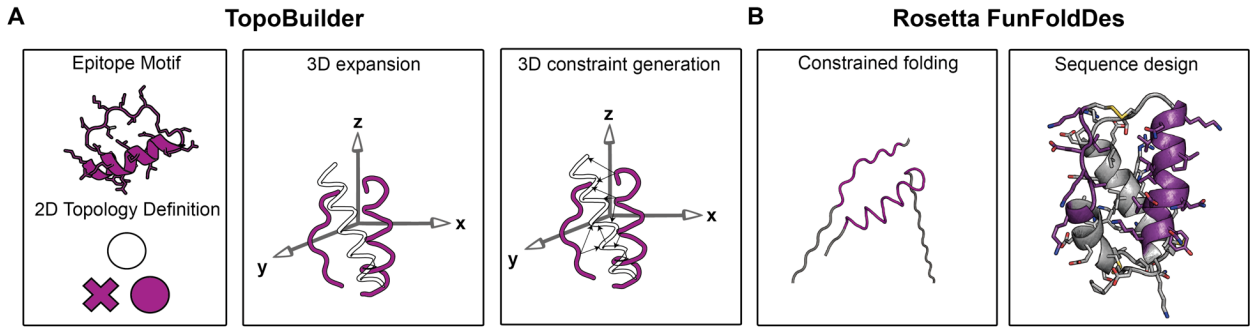


Figure S 3.6. TopoBuilder design strategy.

A) The motif of interest is extracted from its native environment, and a 2D form is generated that allows connecting the discontinuous epitope segments. The 2D form is then expanded to the 3D space, applying user defined rotations and translations along x,y and z coordinates. From the 3D sketch, Cα constraints are generated to guide the folding process. B) Rosetta FunFolDes is used to fold the idealized 3D sketch (using fragment insertions of sizes 3 and 9), and to build connecting loops between the secondary structures. A sequence that stabilizes the folded pose is designed in a last step using Rosetta FastDesign. Further details on the design process, the TopoBuilder code and scripts used for folding and design are available in the online repository.

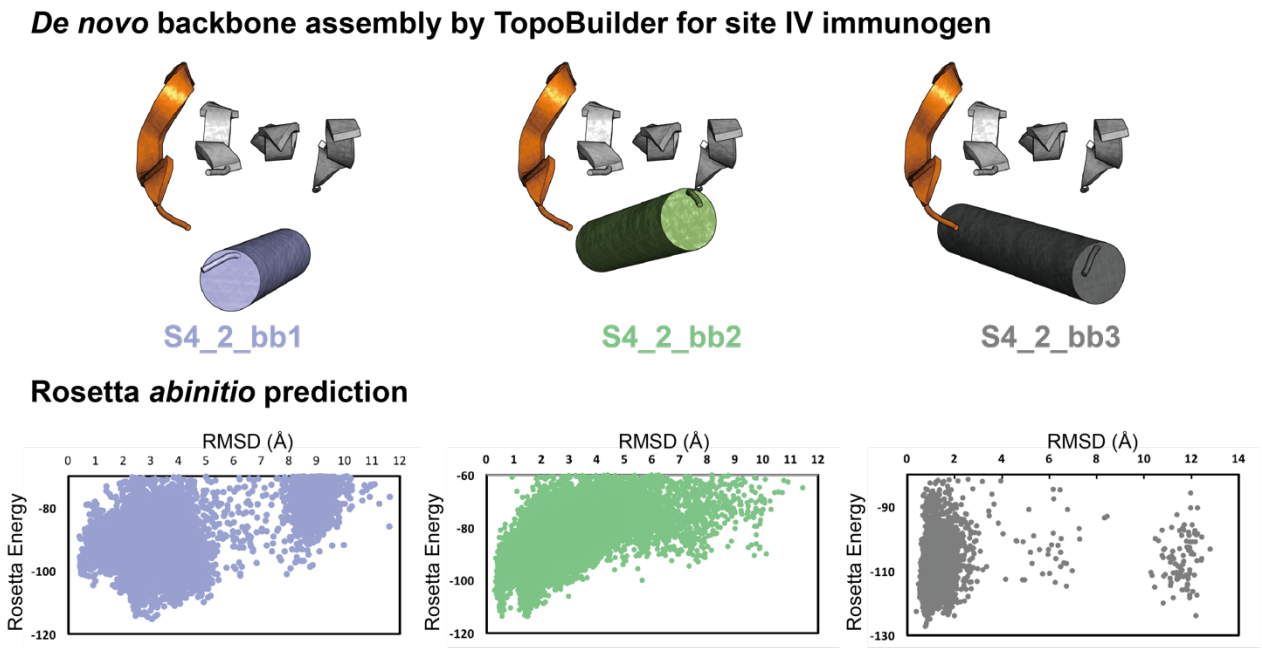


Figure S 3.7. *De novo* backbone assembly for site IV immunogen.

The site IV epitope was stabilized with three antiparallel beta strands built *de novo*, and a helix packing in various orientations against this beta sheet (bb1-bb3). Each backbone was simulated in Rosetta *ab initio* simulations for its ability to fold into a low energy state that is close to the design model, indicating that S4_2_bb2 and bb3 have a stronger tendency to converge into the designed fold.

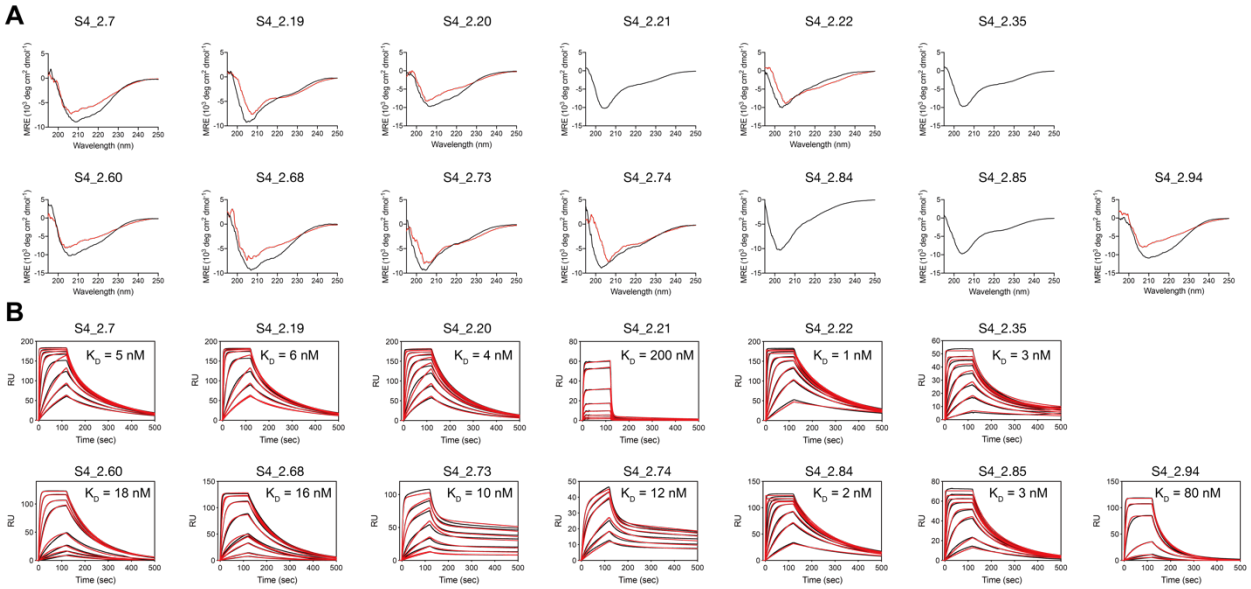


Figure S 3.8. Biophysical characterization of S4_2 design series.

A) Circular dichroism spectra for 13 designs of the S4_2 design series that were enriched for protease resistance and binding to 101F in the yeast display selection assay. Black: spectrum at 20 ºC. Red: spectrum at 90 ºC. B) SPR sensorgrams for binding to 101F for the same designs. 101F IgG was immobilized on the sensor chip surface, and the designs were flown as analyte.

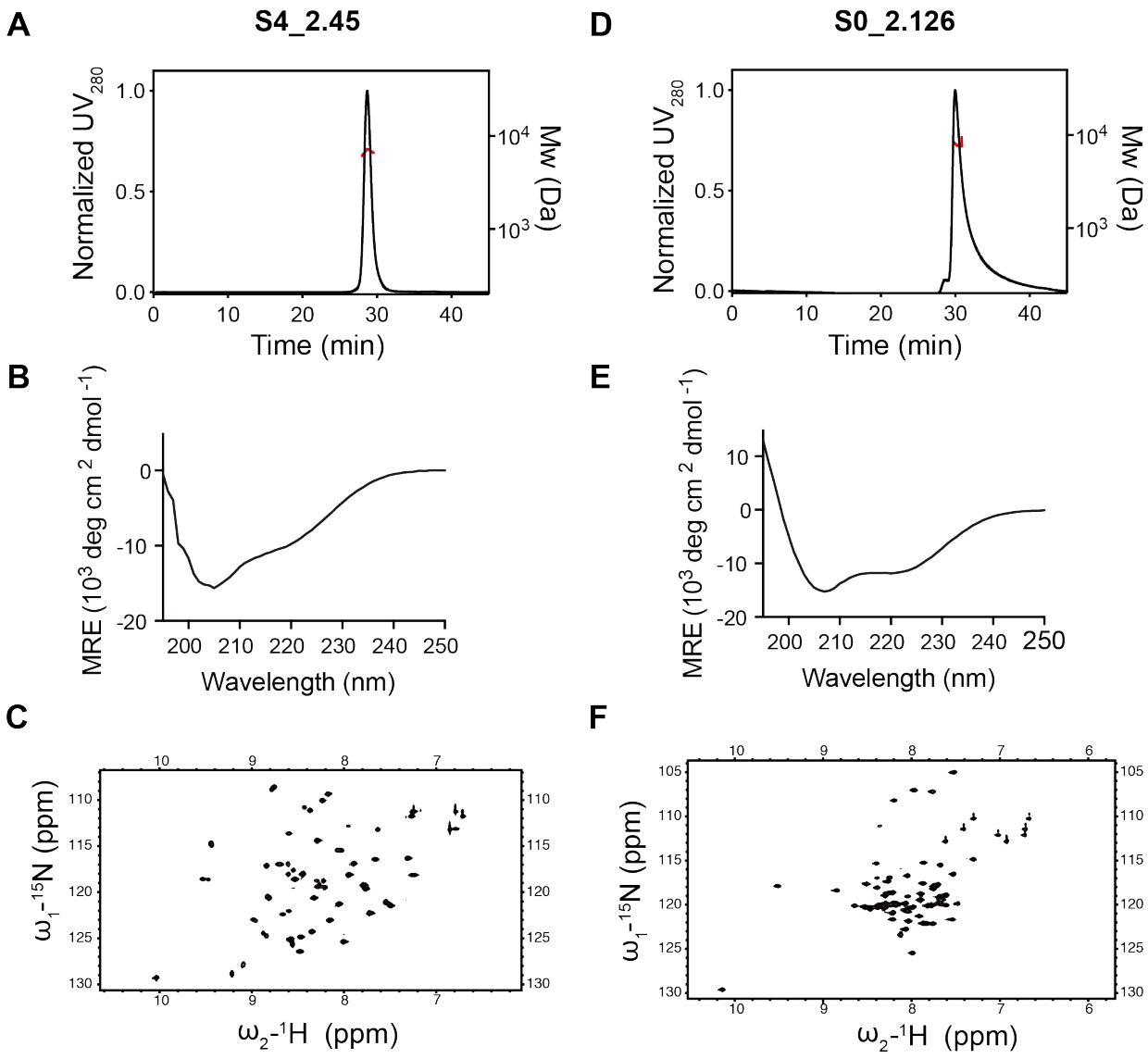


Figure S 3.9. Biophysical characterization of S4_2.45 and S0_2.126.

A,D: S4_2.45 (A) and S0_2.126 (D) are monomeric in solution as shown by SEC-MALS profile. B,E: Circular dichroism spectra at 25 ºC. C,F: 2D NMR of 15N HSQC spectra for S4_2.45 (C) and S0_2.126 (F) are well dispersed, confirming that the designs are well folded in solution. Mw: Molecular Weight.

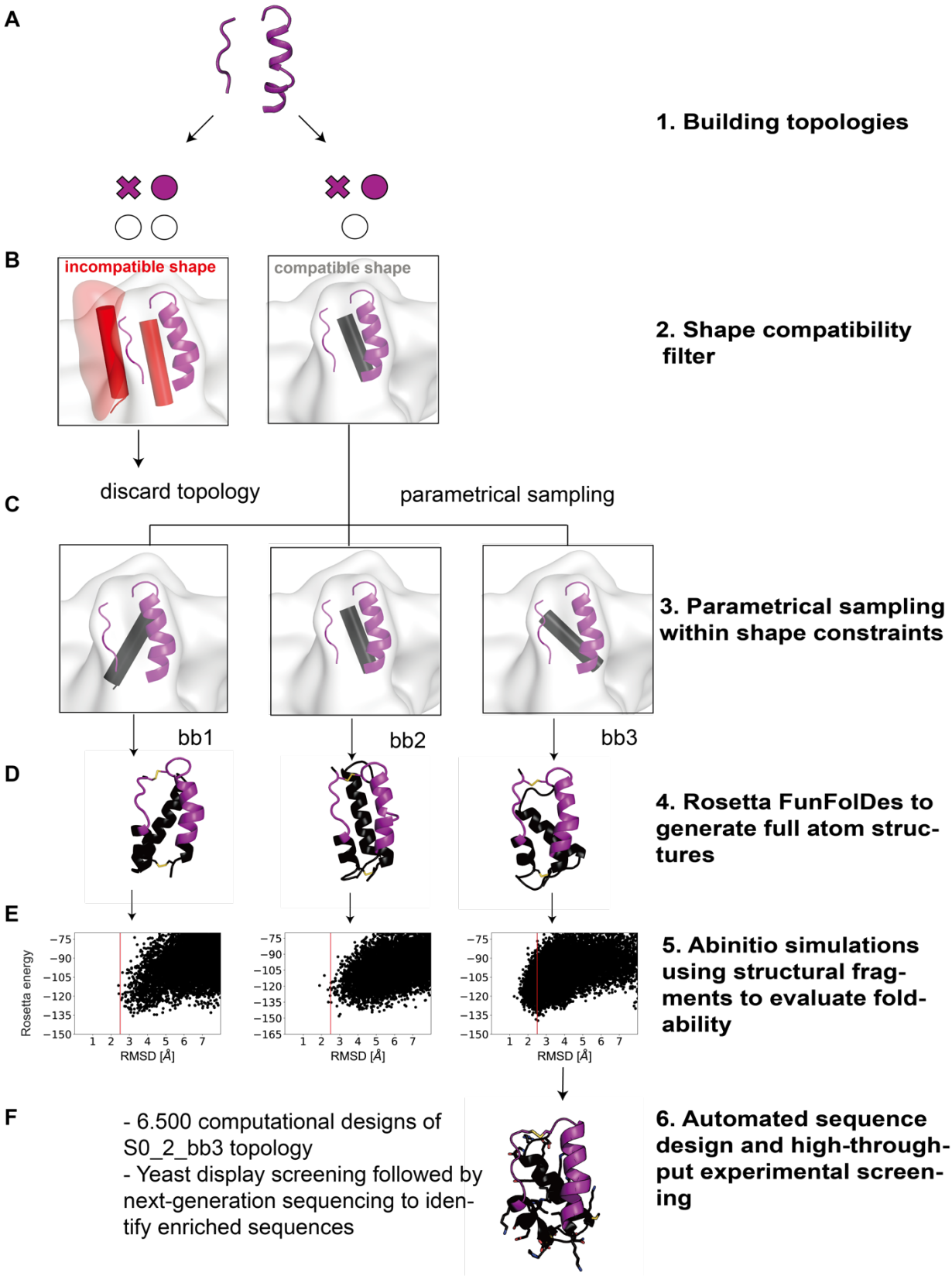


Figure S 3.10. ***De novo*** **topology assembly to stabilize site 0 using TopoBuilder.**

A) Topologies that allow to connect the discontinuous site 0 motif are built as 2D form, and then translated into a 3D sketch. B) Generated sketches are evaluated for their compatibility with the shape constraints of prefusion RSVF. A compatible shape (right) respects the quaternary constraints of the motif in its native environment, whereas an incompatible shape (left) is one that sterically constrains the motif's accessibility in a non-native manner (red surface). C) Once a compatible topology is identified, the secondary structure is parametrically sampled (3 different helical orientations, S0_2_bb1-bb3). D) Three customized helical orientations were assembled (S0_2_bb1-bb3) to support site 0 epitope, and evaluated for their ability to fold into the designed topology in Rosetta *ab initio* simulations (E). S0_2_bb3 showed a funnel-shaped energy landscape, and was selected for subsequent automated sequence design and experimental screening (F).

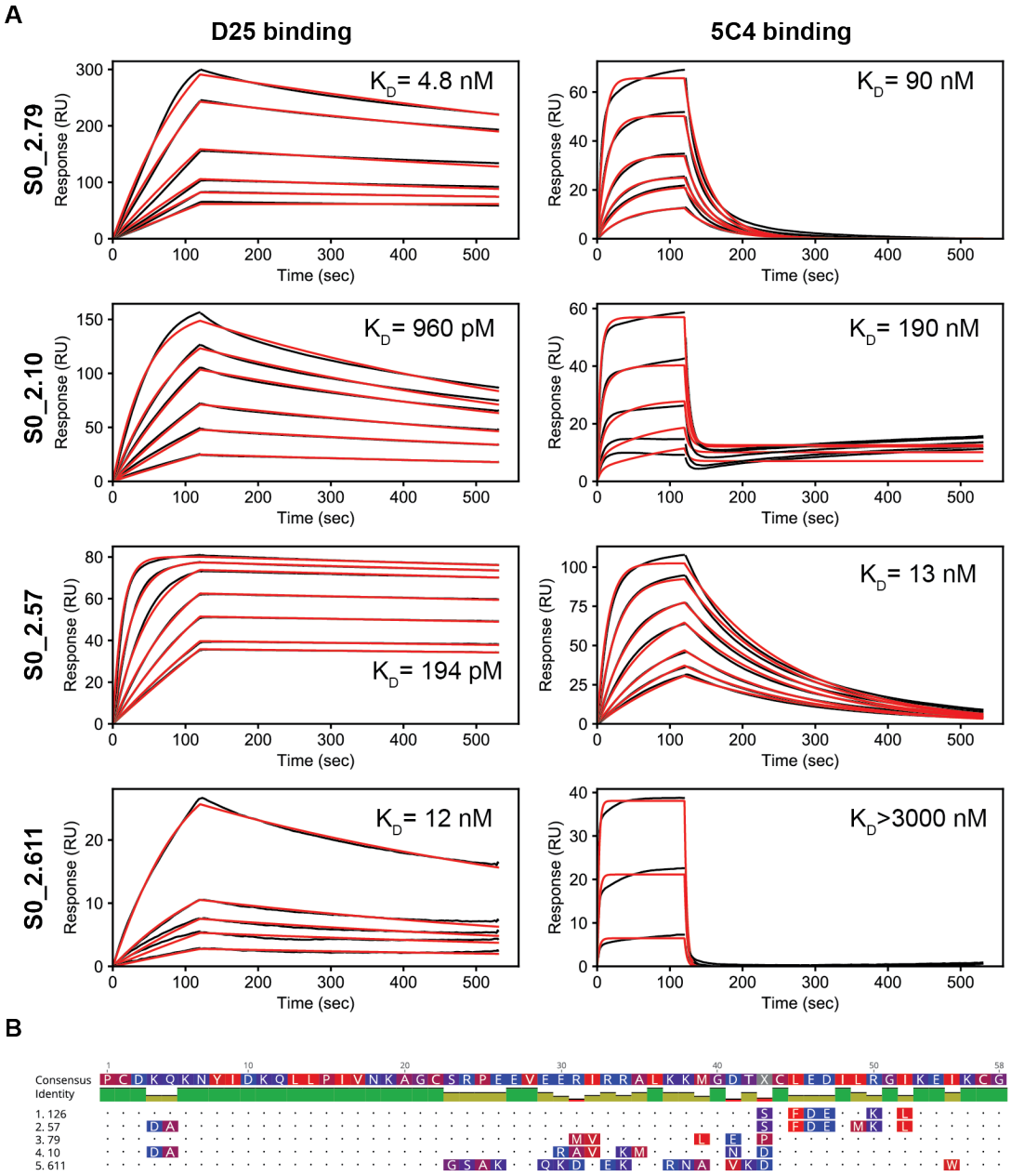


Figure S 3.11. Biophysical characterization of S0_2 design series.

(A) Binding affinity measurement for D25 and 5C4 binding of *de novo* site 0 scaffolds. Shown are the SPR sensorgrams of enriched designs that were successfully expressed and purified after the yeast display selection. D25 or

5C4 IgG were immobilized on the sensor chip surface, and scaffolds were injected as analyte. Dissociation constants shown are kinetic fits using a 1:1 Langmuir model. (B) Sequence alignment of experimentally characterized sequences, in comparison to S0_2.126. The closest sequence homolog to S0_2.126 is S0_2.57, differing in 3 amino acids.

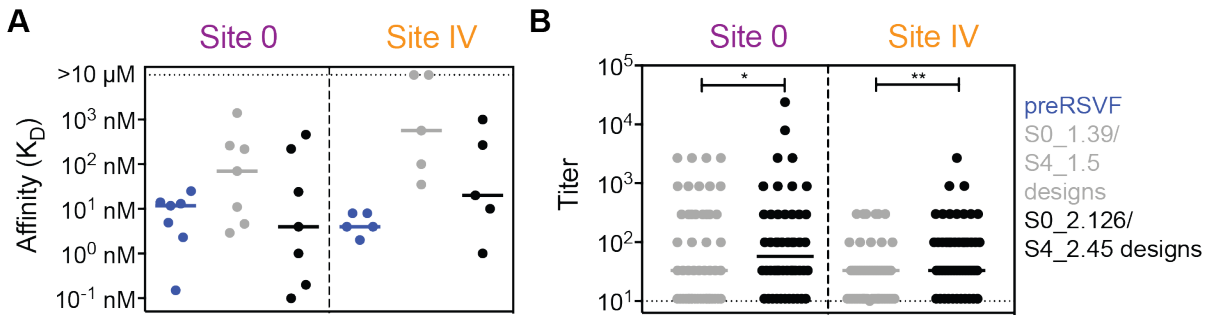


Figure S 3.12. Binding affinity of designed immunogens towards panels of site-specific, human neutralizing antibodies and human sera.

A) Binding affinity (K_D, determined by SPR flowing Fabs as analyte) of S0_1.39 (*204*) and S0_2.126 (*242*) towards a diverse panel of site-specific neutralizing antibodies, in comparison to prefusion RSVF (blue). Antibodies shown for site 0 are 5C4, D25 (*257*), ADI-14496, ADI-18916, ADI-15602, ADI-18900 and ADI-19009 (*297*). For site IV, the binding affinity was tested against 101F (*224*), ADI-15600 (*297*), 17E10, 6F18 and 2N6 (*298*), comparing S4_1.5 (*204*) and S4_2.45 (*242*) to prefusion RSVF. The higher binding affinity of the second-generation designs (S0_2.126 and S4_2.45) compared to the first generation and to prefusion RSVF indicates a greatly improved, near-native epitope mimicry of the respective antigenic sites in the designed immunogens. B) ELISA reactivity of designed immunogens with sera obtained from 50 healthy human adults that were seropositive for prefusion RSVF. Both S0_2.126 and S4_2.45 showed significantly increased reactivity compared to the first-generation designs, confirming an improved epitope-mimicry on the serum level (* $p< 0.05$ and ** $p< 0.01$, Wilcoxon test). Data are representative from one out of two independent experiments.

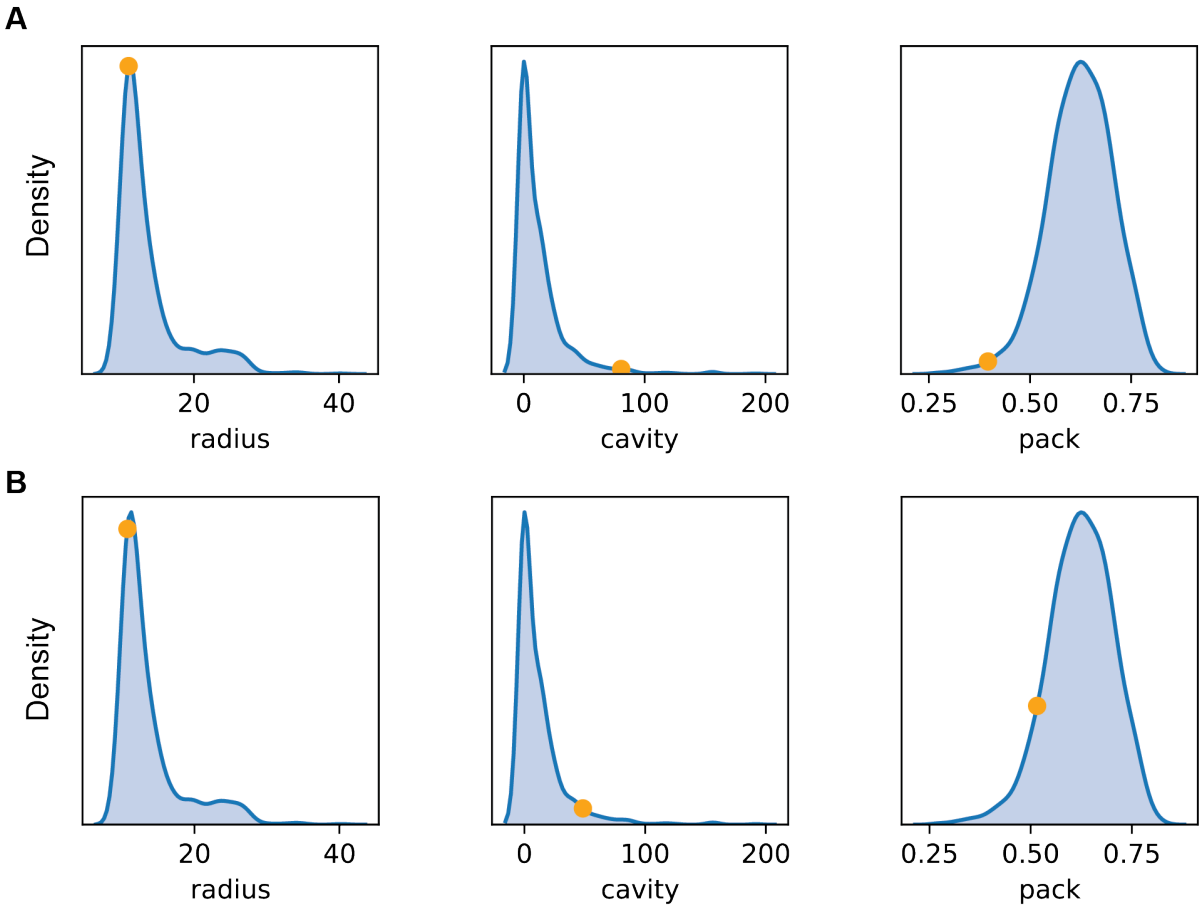


Figure S 3.13. Comparison of S0_2.126 Rosetta scores against natural proteins of similar size.

Protein structures within the same size as S0_2.126 (57 +/- 5 residues) were downloaded from the CATH database and filtered by 70 % sequence homology, yielding a representative database of natural proteins with similar size as S0_2.126 (n = 1,013 structures). Proteins were then minimized and scored by Rosetta to compute their radius of gyration, intra-protein cavities (cavity) and core packing (packstat). Plotted is the distribution for these score terms in 1,013 natural proteins (blue density plot), and the same scores for S0_2.126 are shown in orange. The NMR structure of S0_2.126 is shown in (A), the computational model of S0_2.126 is shown in (B), indicating that, despite similar radius of gyration, S0_2.126 shows a substantial cavity volume as well as a very low core packing compared to natural proteins of similar size. CATH database and scores were pre-calculated, loaded and visualized using the rstoolbox python library (*299*).

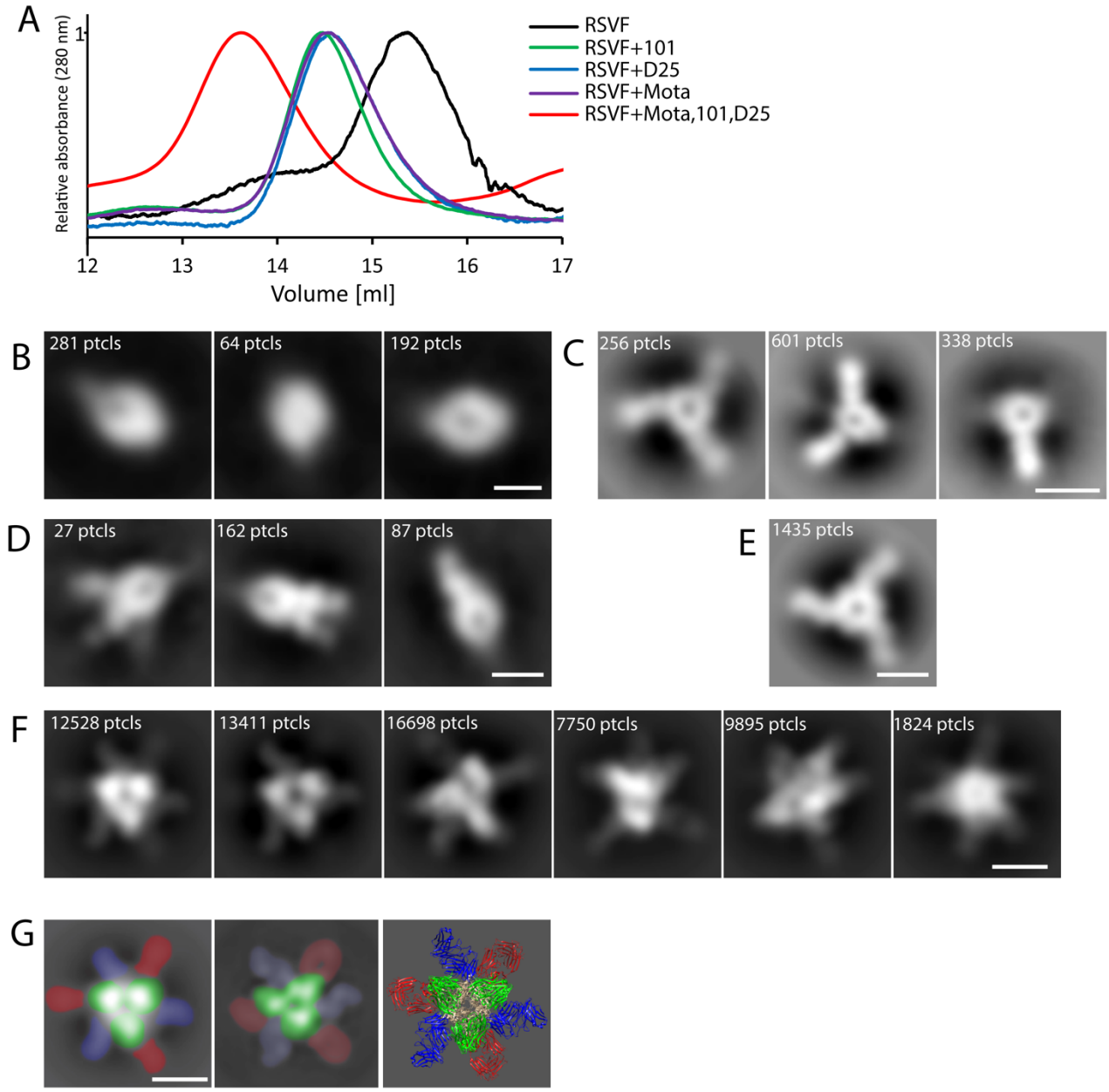


Figure S 3.14. Electron microscopy analysis of site-specific antibodies in complex with RSVF trimer.

(A) Superposed size-exclusion profiles of unliganded RSVF (black line) and RSVF in complex with 101F (green line), D25 (blue line), Mota (purple line) and all three (101F, D25, Mota - red line) Fabs. (B-F) Representative reference-free 2D class averages of the unliganded RSVF trimer (B) and RSVF in complex with 101F (C), D25 (D), Mota (E) or all three (101F, D25, Mota (F)) Fabs. Fully saturated RSVF trimers bound by Fabs are observed, as well as sub-stoichiometric classes. (G) Left panel: reference-free 2D class average of RSVF trimer with three copies of 101F, D25 and Mota Fabs visibly bound. The predicted structure of RSVF in complex with 101F, D25 and Mota was used to simulate 2D class averages in Cryosparc2, and simulated 2D class average with all three types of Fabs is shown in the middle panel. Right panel: predicted structure of RSVF trimer with bound 101F, D25 and Mota Fabs based on the existing structures of RSVF with individual Fabs (PDB ID 4JHW, 3QWO and 3O45). Fabs are colored as follows: red - 101F; blue - Mota; green - D25. Scale bar - 100 Å.

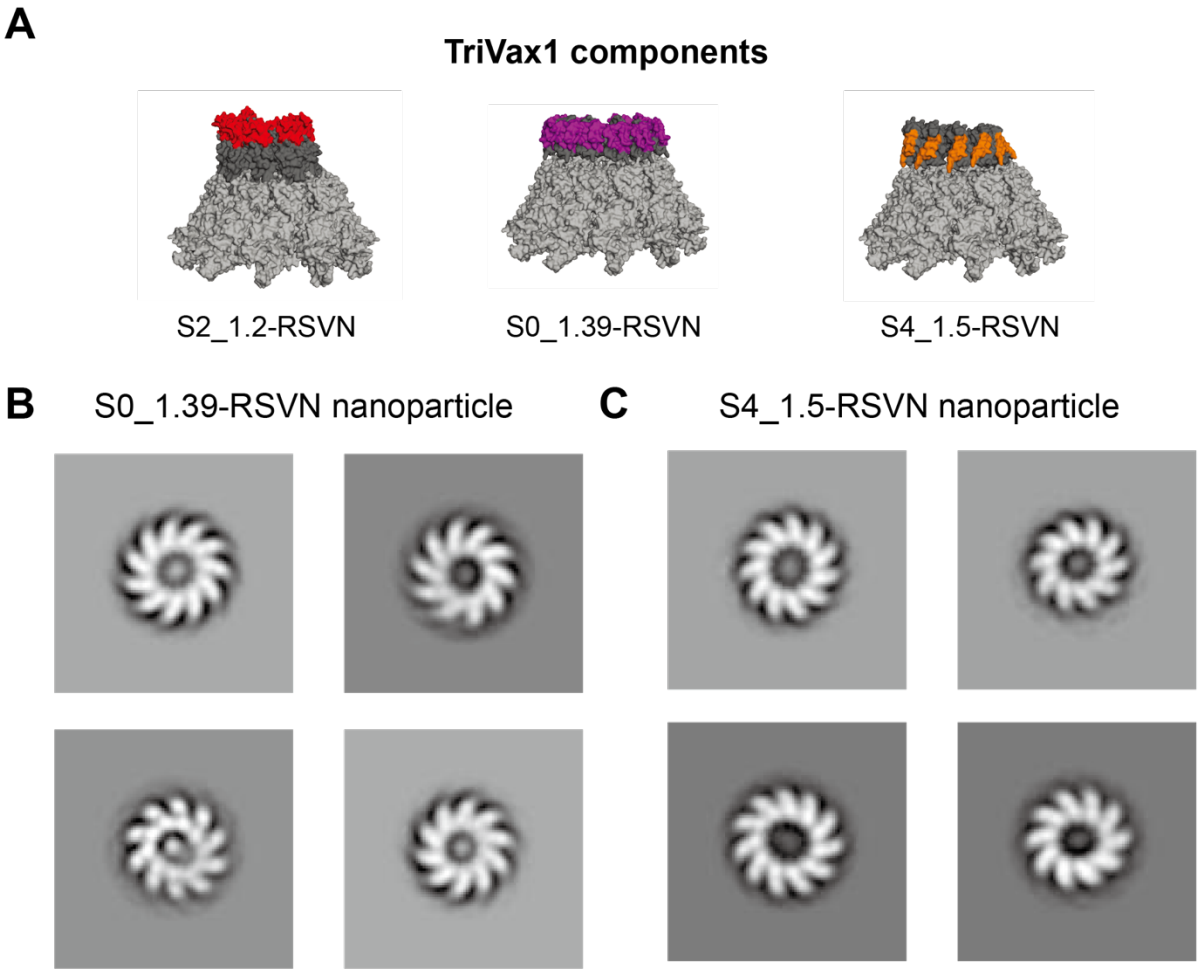


Figure S 3.15. Composition and EM analysis of Trivax1 RSVN nanoparticles.

A) Trivax1 contains equimolar amounts of site II, 0 and IV epitope focused immunogens fused to the self-assembling RSVN nanoparticle with a ring-like structure (n = 10-11 subunits). The site II-RSVN nanoparticle has been described previously (*256*). Shown are the computational models for the nanoparticles-immunogen fusion proteins. B,C) Negative stain electron microscopy for S0_1.39-RSVN and S4_1.5-RSVN nanoparticles confirms that the ring-like structure is maintained upon fusion of the designed immunogens.

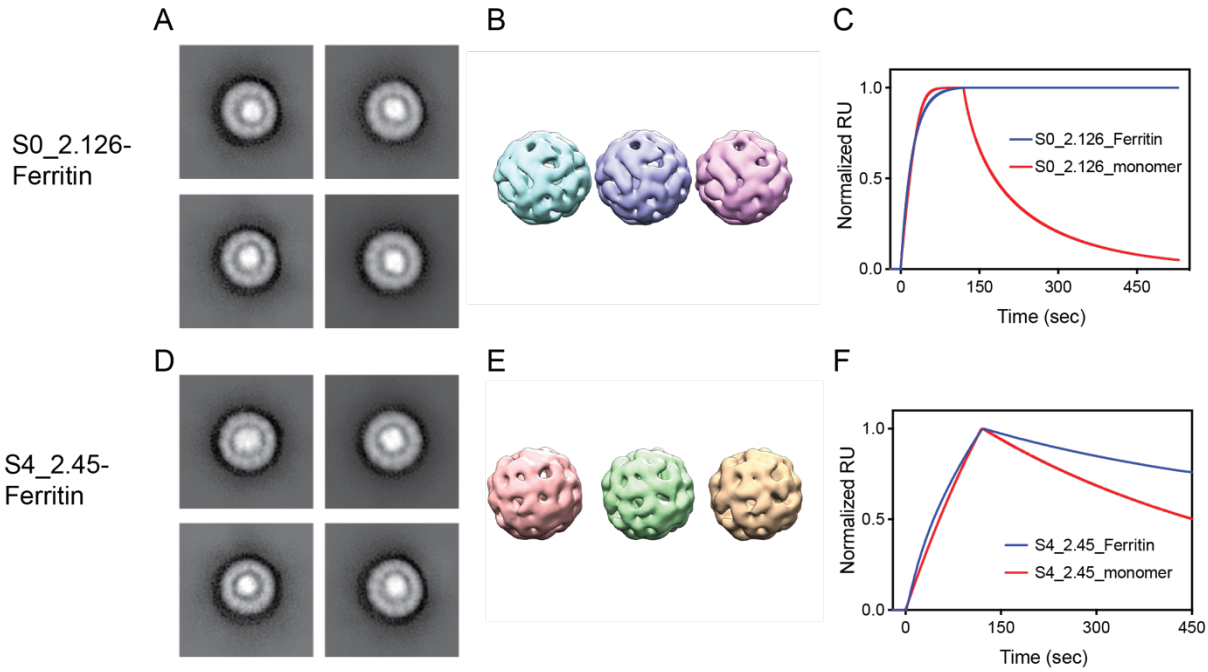


Figure S 3.16. EM analysis of Trivax2 ferritin nanoparticles.

A,B,D,E) Negative stain electron microscopy (A,D) and 3D reconstruction (B,E) for S0_2.126 and S4_2.45 fused to ferritin nanoparticles. C) Binding affinity of S0_2.126 nanoparticle (blue) to 5C4 antibody in comparison to S0_2.126 monomer (red), showing that S0_2.126 has been successfully multimerized and antibody binding sites are accessible. F) Binding of S4_2.45 to 101F antibody when multimerized on ferritin nanoparticle (blue) compared to monomeric S4_2.45 (red), indicating that the scaffold is multimerized and the epitope is accessible for antibody binding.

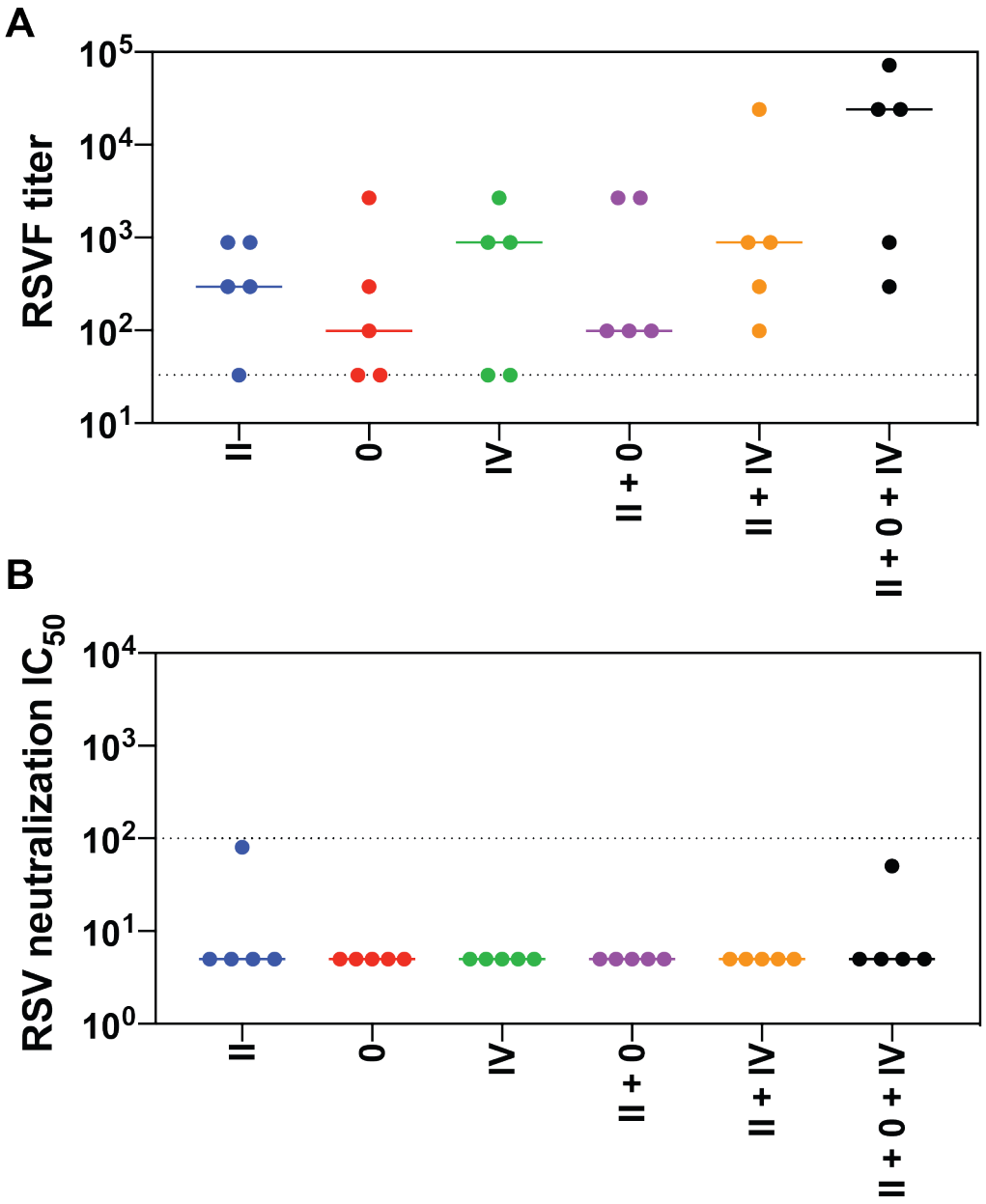


Figure S 3.17. Mouse immunization studies with Trivax1.

A) RSVF cross-reactivity of epitope-focused immunogens formulated individually, as a cocktail of two, and three (Trivax1). B) RSV neutralizing serum titer of mice immunized with designed immunogens and combinations scenarios.

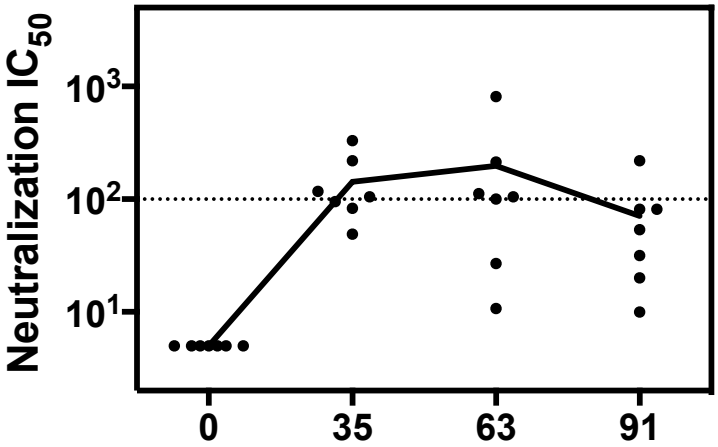


Figure S 3.18. NHP neutralization titer measured by an independent laboratory.

Sera from indicated time points were tested for RSV neutralization by an independent laboratory in a different RSV neutralization assay, using a Vero-118 cell line and a GFP readout.

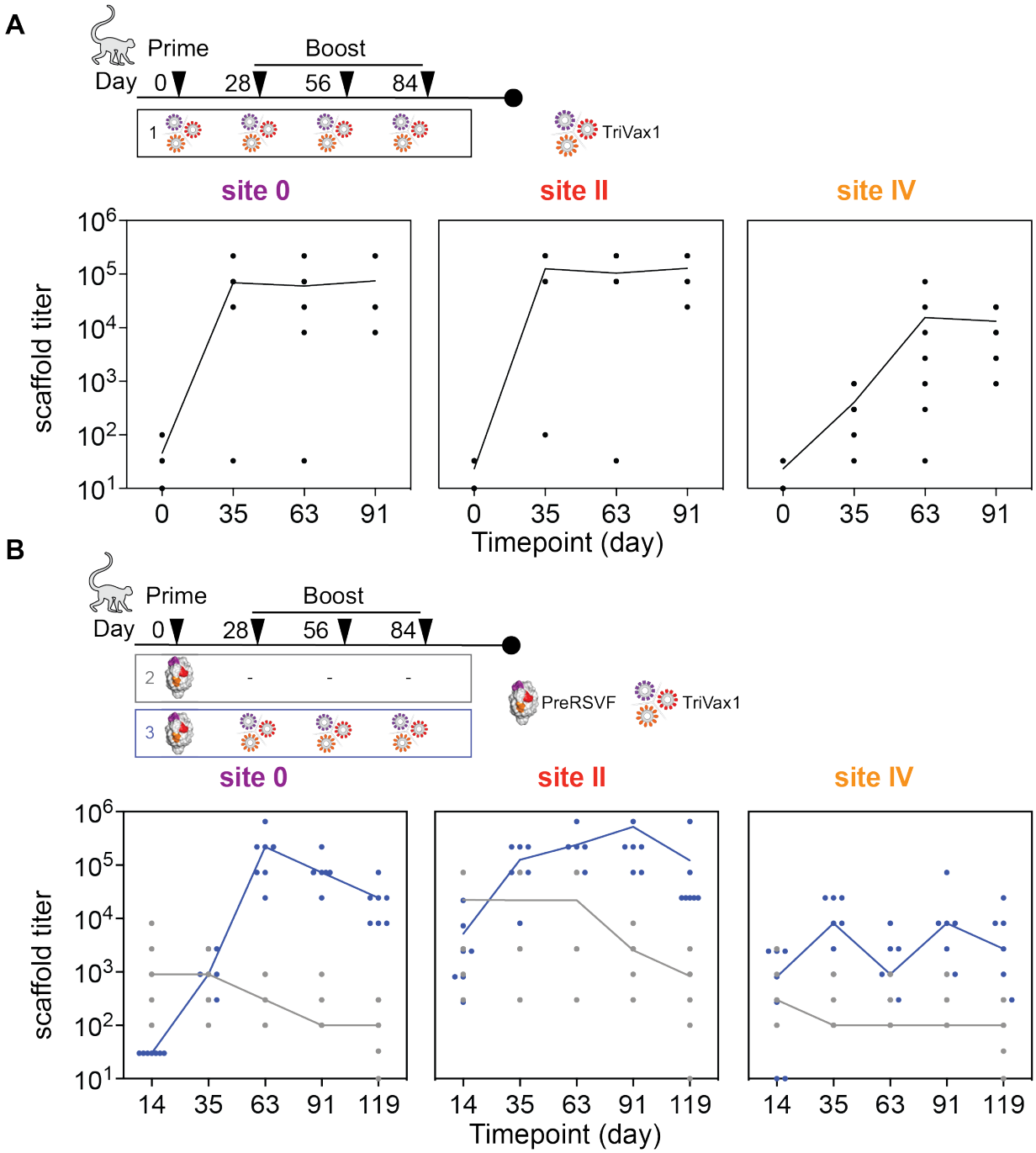


Figure S 3.19. NHP serum reactivity with designed immunogens.

A) ELISA titer of NHP group 1 (immunized with Trivax1) measured at different time points. All animals responded to Trivax1 immunogens at day 91, with site IV immunogen reactivity lower compared to site 0 and site II reactivity. B) ELISA titer of NHP group 2 (grey, RSVF prime) and 3 (blue, RSVF prime, Trivax1 boost). Following the priming immunization, all animals developed detectable cross-reactivity with the designed immunogens, indicating that the designed scaffolds recognized relevant antibodies primed by RSVF.

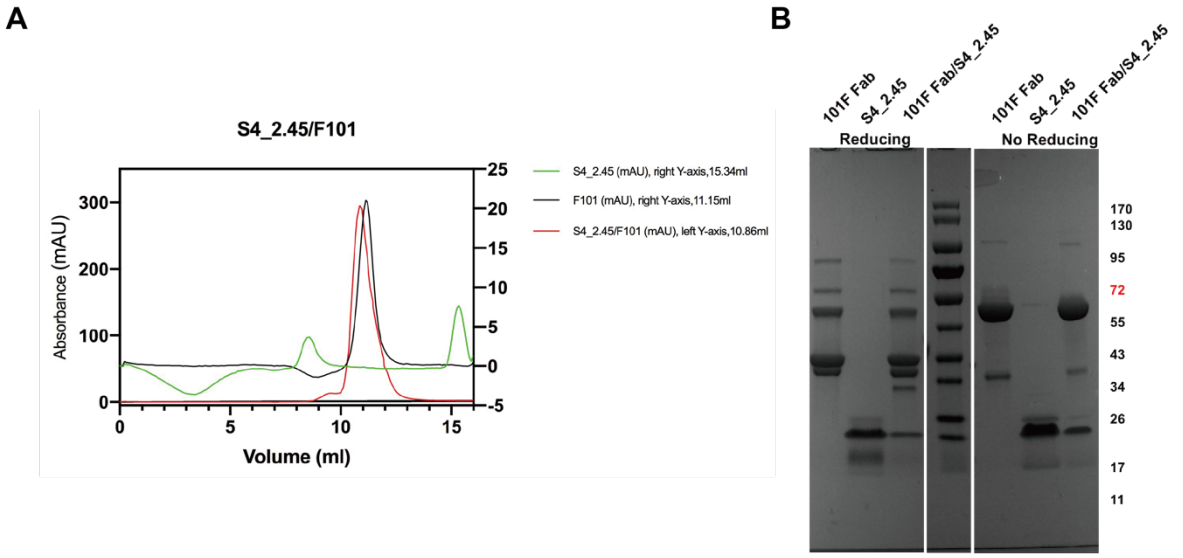


Figure S 3.20. Purification of protein complex of S4_2.45/101F Fab.

A) The size exclusion chromatograph of individual components (Fab and scaffold) and the protein complex. B) SDS gel of purified protein complex shows the collected fractions contain both scaffold and Fab.

Chapter 4 A bottom-up approach for *de novo* design of functional proteins

This chapter presents ongoing work towards a more general framework for the functional *de novo* protein design for carrying functional binding sites. A manuscript based on this work is currently in preparation.

Authors:

Yang C[1,2*], Sesterhenn F[1,2*], Aalen E[3], Scheller L[4], Bonet J[1,2], Cramer JT[5], Wen X[6], Abriata LA[1,2], Rosset S[1,2], Georgeon S[1,2], Jardetzky T[6], Krey T[5,7], Fussenegger M[4], Merkx M[3], Correia BE[1,2].

*These authors contributed equally.

Affiliations:

[1]Institute of Bioengineering, École Polytechnique Fédérale de Lausanne, Lausanne CH-1015, Switzerland. [2]Swiss Institute of Bioinformatics (SIB), Lausanne CH-1015, Switzerland. [3]Laboratory of Chemical Biology and Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands. [4]Department of Biosystems Science and Engineering, ETH Zurich, Basel CH-4058, Switzerland. [5]Institute of Virology, Hannover Medical School, Germany. [6]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305, USA; [7]German Center for Infection Research (DZIF), Hannover, Germany.

Author contributions:

CY, FS and BEC conceived the work and designed the experiments. FS and CY performed computational design and experimental characterization. CY solved X-ray structures. EA and MM performed biosensor characterization and analyzed serum sample by novel biosensor. LS and MF engineered the synthetic cell line, performed the cell assay and analyzed the result. JB developed the TopoBuilder protocol. LAA performed NMR characterization and solved NMR structure. JTC, TK, XW and TJ help with structural characterization. SR, and SG performed experiments and analysed data. FS, CY and BEC wrote the manuscript, with input from all authors.

4.1 Introduction

De novo protein design has emerged as a powerful approach to delineate rules of protein folding, and translate them into the design of novel proteins with defined structures (*300, 301*). Several successes have been reported for the *de novo* design of diverse protein folds. The primary design goal, for the vast majority of these studies, has been focused on structural accuracy of computationally generated models relative to experimentally determined (*46, 181, 302, 303*). In contrast, the design of *de novo* proteins with encoded biochemical function is lagging far behind. Nonetheless, successes reported illustrate the potential of *de novo* design to transform multiple arenas of biology and biotechnology, including the design of vaccine candidates (*178, 256, 304*), protein-based lead drugs (*102*), antivirals (*230*), pH-responsive carriers (*305*) and others (*176, 306, 307*).

Molecular recognition is central to protein function, the ability to specifically interact with other molecules, including proteins, small molecules, nucleic acids, metals and others determines many of the

protein's biological roles (*308*). Albeit not the only determinant of protein function, a fundamental requirement for *de novo* design of functional proteins is to endow *de novo* proteins with structural motifs that mediate binding interactions. Beyond the difficulty of designing structurally accurate *de novo* proteins, functional design requires an exceptional level of accuracy, which is further complicated by the high structural complexity of many binding sites relevant in biology.

A widely used approach to design functional proteins is to transplant functional sites from one protein to another, generally referred to as protein grafting (*76*). An essential prerequisite for protein grafting is the availability of template structures with enough local structural similarity to faithfully present the transplanted functional site. With few exceptions (*193*), grafting has been largely limited to single, regular secondary structures that are frequently found in natural proteins. However, the vast majority of functional sites is not contained in single, regular helical segments, but rather composed of multiple and often irregular structural segments, that are stabilized by the overall protein structure (*247-249*).

Beyond nature's repertoire, *de novo* protein design allows to create novel structural templates deprived of any evolutionary baggage. In a subsequent step, *de novo* proteins can be functionalized with structural motifs, as recently shown for mini-proteins endowed with influenza stem binding motifs (*230, 301*). We refer to this two-step approach, consisting of first building a stable, functionless scaffold first, which subsequently serves as a template for grafting, as "top-down" approach. More generally, the "top-down" approach also includes design strategies that repurpose protein structures from the natural repertoires for other functions (*28, 29, 82, 84*). There are several important limitations of a "top-down" approach for functional protein design: first, most *de novo* proteins are built with a high content of regular secondary structures, high-contact order and minimal loops (*181, 309*). While these proteins generally are highly thermodynamically stable, *de novo* proteins designed in a "function-agnostic" fashion are unlikely to have enough local structural similarity to an irregular, multi-segment, functional motif that would make them amenable to grafting approaches. Second, the chosen protein fold largely defines the topological constraints of the designed protein, beyond local structural similarity, the overall topology must be compatible with the designed function, as shown for *de novo* designed receptor agonists and epitope mimetics (*102, 304*).

In contrast to this "top-down" approach, a "bottom-up" (Fig. 4.1A) strategy could consider the local structural and global topological requirements of the functional motif, irrespective of its complexity, and build supporting secondary structure elements (SSEs) to stabilize the functional site in its native conformation (*310*). A few studies have shown such a function-centric design strategy (*102, 176, 311, 312*), but a general approach that allows to systematically build *de novo* proteins with embedded functional motifs, controlled connectivity and precise control over the spatial positioning of each motif and secondary structure element is thus far lacking.

Here, we describe a "bottom-up" approach to design functional *de novo* proteins by centering the design process on the functional motif. SSEs are assembled around the functional motif of interest, followed by parametric sampling of the SSEs to refine the topological features of the structure and a folding-design stage to finely sample the sequence-structure space (*86*).

We demonstrate the power of this approach by designing various protein folds (all-alpha, all-alpha with crossover, alpha-beta), to accommodate irregular and discontinuous binding motifs. These binding motifs are found in two proteins of the respiratory syncytial virus (RSV): the fusion protein (RSVF) (*116*) and glycoprotein (RSVG) (*115*); and are recognized by well characterized human monoclonal antibodies. On the functional side, we showcase two completely distinct applications of *de novo* designed proteins

(Fig. 4.1B). First, we show their utility as building blocks for BRET-based biosensors to detect and quantify epitope-specific antibodies, potentially useful to profile antibody responses with single epitope resolution. Additionally, we assembled a protein topology bi-functionalized with two distinct binding motifs, which we use as a non-natural inducer to control receptor-mediated signaling to and modulation of transgene expression in synthetic cells. Altogether, we present a versatile strategy for the design of functionalized *de novo* proteins carrying structurally complex binding motifs, that is applicable to a wide range of functional protein design challenges.

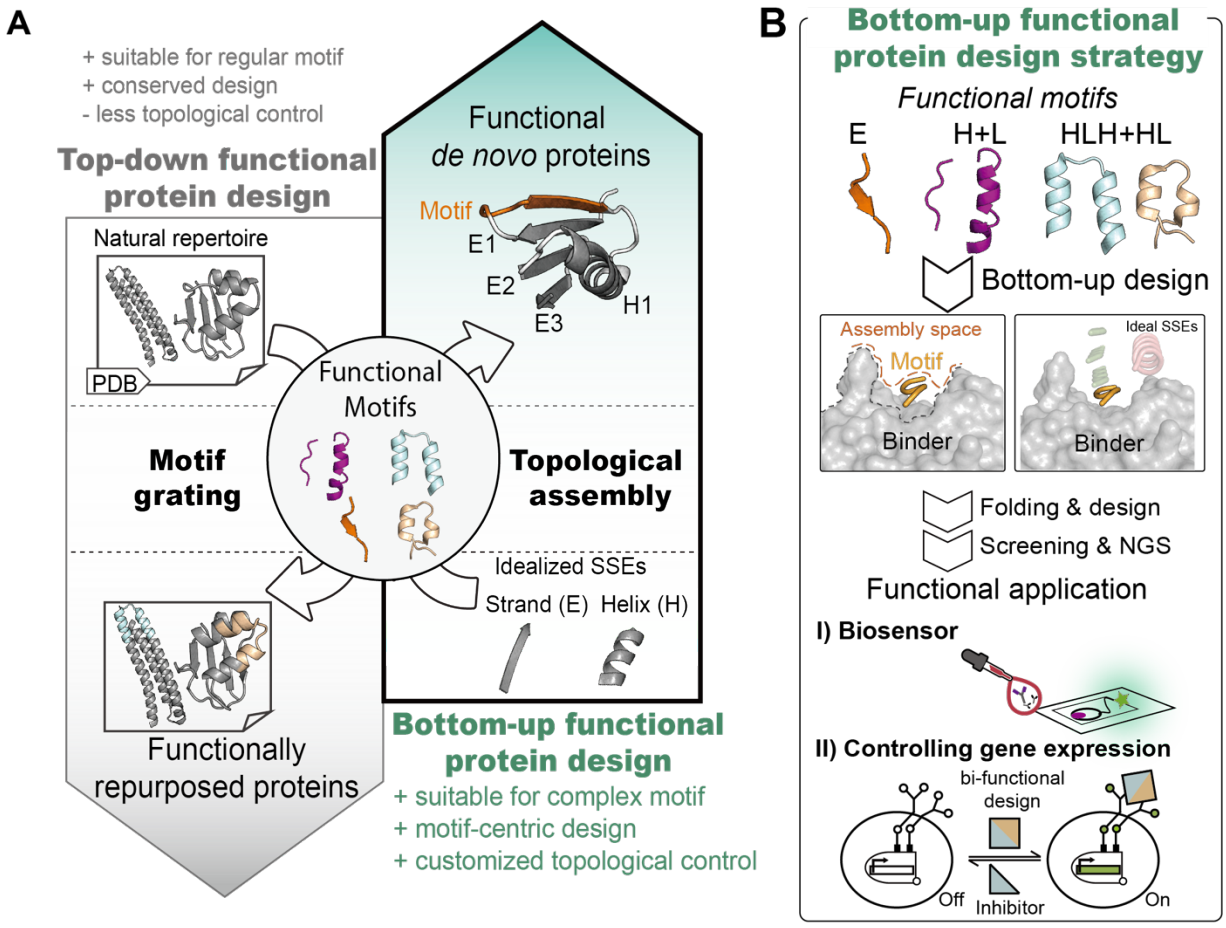


Figure 4.1. A bottom-up design strategy for the design of functional proteins.

A) Two conceptual frameworks for designing functional proteins. Top-down functional design strategies have been previously reported, in which functional motifs were grafted onto pre-existing templates found in the natural protein repertoire. In contrast, the presented bottom-up strategy allows the *de novo* assembly of stabilizing secondary structures around a given functional motif, yielding fully *de novo* functional proteins. B) Bottom-up *de novo* design of functional proteins for several structural motifs and biological applications. *De novo* proteins containing binding motifs of different secondary structure types and structural complexity were built using a bottom-up *de novo* design approach. Idealized secondary structural elements (SSEs) were assembled around the binding motifs, respecting spatial constraints imposed by the binder. To identify lead candidates for functional characterization, designed topologies were screened using yeast display, followed by next-generation sequencing (NGS). As functional applications, the designed proteins were used to create biosensors for the detection of epitope-specific antibodies and as signaling triggering molecules of synthetic cell-surface receptors that can control gene expression.

4.2 Results

Bottom-up functional de novo protein design

While numerous studies have reported the *de novo* design of proteins with high structural accuracy, the design of functional *de novo* proteins remains more challenging (*70*). With few exceptions (*101, 102, 313*), the majority of functional protein design efforts has employed a "top-down" approach (Fig. 4.1a), in which functional sites were installed on existing (*80, 99*) or previously designed protein scaffolds (*84*).

We have previously explored the bottom-up conceptual approach, namely the TopoBuilder protocol, for the *de novo* design of epitope-focused immunogens (*314*). Here, we enlarge the protein structural space explored, as well as, the scope of the applications, by introducing a more generalizable technical framework for function-centric, bottom-up *de novo* design of functional proteins (Fig 4.2).

As functional elements, we selected four structurally well-characterized antibody binding motifs of RSVF and RSVG proteins, differing in secondary structure content and structural complexity (Fig. 4.1 and Fig. 4.2): I) RSVF site IV, a linear, irregular beta strand (*169*); II) RSVF site 0, a discontinuous epitope consisting of a kinked alpha helix and a disordered loop (*117*); III) RSVF site II, a linear helix-turn-helix motif (*116*); IV) RSVG 2D10 epitope, a linear helix-loop segment that is constrained by two disulfide bonds (*315*).

The TopoBuilder presents three major assets for functional protein design, as shown in Fig. 4.2. First, it enables the modular assembly of user-defined protein topologies with respect to the functional motif's structural requirements. To do so, idealized SSEs with defined length are positioned one by one to support the motif, enabling the systematic building and exploration of multiple topologies accommodating the same motif. Leveraging this feature, we constructed three different topologies to present the site IV epitope: 3E2H (three-stranded sheet and two helices), 4E1H (four-stranded sheet and one helix) and 4E2H (four-stranded sheet and two helices). Second, it allows full control over the topology's connectivity, as shown by the design of two connectivities for a 3H1L topology (three helices with a crossover loop) assembled to host the discontinuous site 0 epitope. Third, we showcase the utility of the TopoBuilder to assemble a bi-functional protein and parametrically sample the placement of secondary structures and motifs. In one layer of the protein fold, the 4H topology presents the site II epitope, whereas the second helical layer contains the 2D10 epitope. To ensure optimal presentation of both functional motifs, we sampled two configurations (4H_01, 4H_02), where the first layer is tilted relative to the second layer sampling different degrees of structural compactness of the protein core (Fig. 4.2).

Each topology was initially defined in a two-dimensional form representation (*2*) and subsequently projected in the three-dimensional space, with respect to the coordinates of the functional motif extracted from its native context (RSVF or RSVG). Secondary structures were placed with a distance of 10-11 Å between the center of mass of alpha helices, and 4.5 Å between adjacent beta strands. From these idealized 3D sketches, pairwise distance constraints were derived to guide constrained *ab initio* folding coupled to sequence design using Rosetta FunFolDes (*86*).

Between 10,000 - 20,0000 designs were generated for each topology, and filtered for decoys with native-like features such as well-packed hydrophobic cores, and sequences that presented funnel-shaped energy landscapes in Rosetta *ab initio* simulations. Critical hydrophobic core positions based the design simulations were selected, and encoded combinatorially in DNA libraries for subsequent high-throughput yeast display screening (10^5 - 10^6 variants per topology).

Each library was sorted under double selective pressure: site IV designs - binding to 101F antibody, and residual binding under a pre-treatment of nonspecific protease chymotrypsin; site 0 and bifunctional designs - binding to two different antibodies (D25 and 5C4). Sorted populations were bulk-sequenced using next-generation sequencing, revealing sequence profiles and structural determinants for stability and accurate motif presentation of the different design series (Fig. 4.2B and Fig S4.1-S4.5).

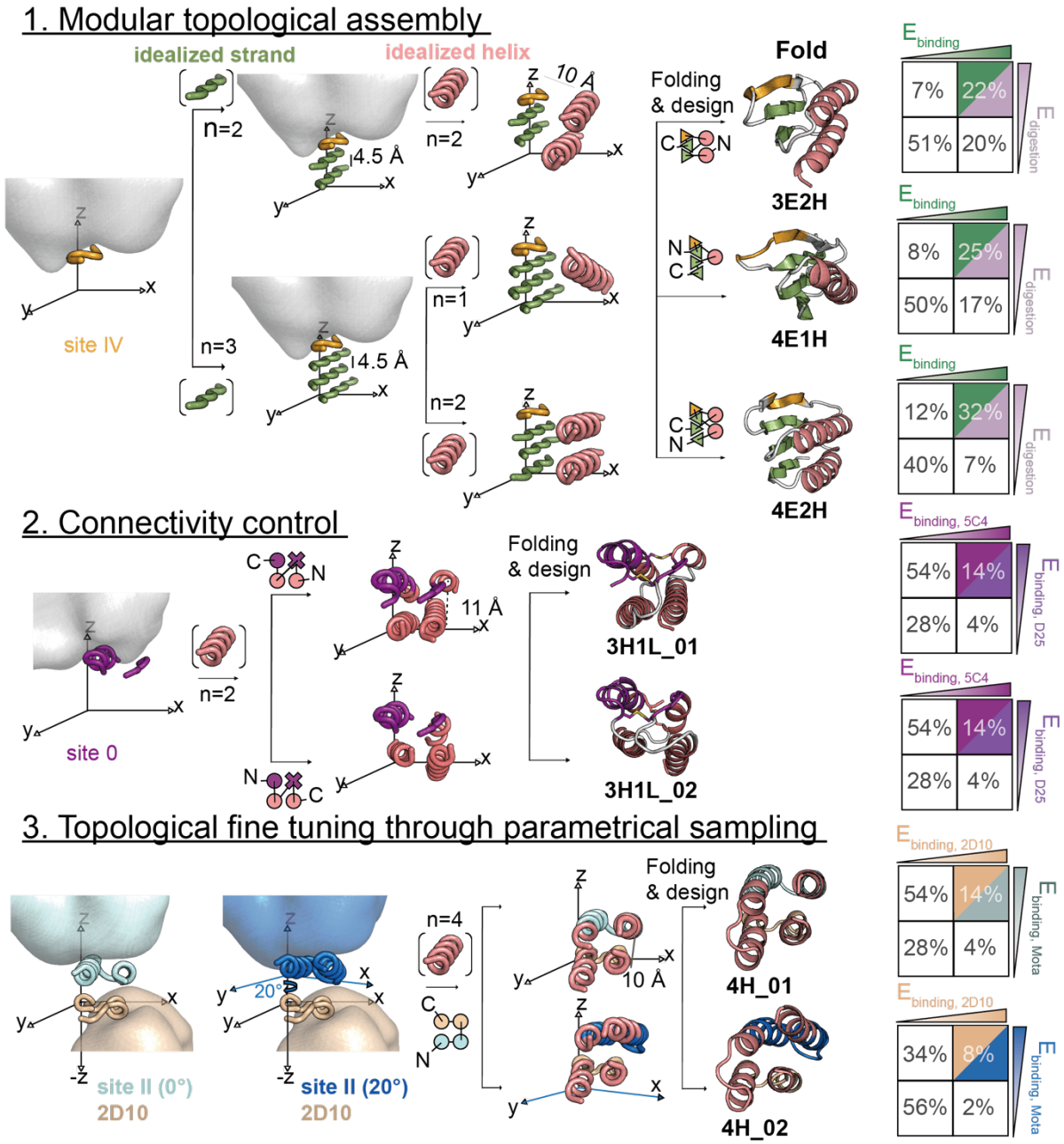


Figure 4.2. Tailoring distinct protein folds for four different binding motifs.

A) Protein backbone assembly and building stages of each defined topology. The TopoBuilder software enables a modular topological assembly of virtually any protein fold that can be described in layers, allowing to create both beta strands and alpha helices of defined length and spatial positioning. Three different folds were built to present the site IV epitope (3E2H, 4E1H, 4E2H). The topological connectivity of the folds is controllable, as illustrated by

the design of two 3H1L topologies to stabilize site 0. With all its features, the TopoBuilder enables the design of bi-functionalized proteins, carrying both the site II and the 2D10 epitopes, and fine-tune the spatial positioning of the motifs with respect to each other. E: beta strand, H: alpha helix, L: loop structure. B) Bulk enrichment analysis of sequences sorted under selective pressures for different topologies. X, Y-axis represents two selection conditions with an arrow gradient showing the magnitude of enrichment, with the upper right quadrant showing the percentage of clones that were enriched for both selective pressures. The raw data showing the distribution of each sequence is shown in Supplementary figures S4.1-S4.5.

Designs are well folded and bind with high affinity to target antibodies

For each topology, 10-20 designs that showed strong enrichment under double selection pressure in the yeast screen were selected for recombinant expression in E.*coli* and biophysical characterization.

All three topologies designed to accommodate site IV yielded well-folded and stable proteins that bound the 101F antibody with dissociation constants (K_D) ranging from 20 to 180 nM (Fig S4.6-S4.8). The best designs from each design series, 3E2H_37, 4E1H_95 and 4E2H_210, showed melting temperatures (T_ms) of 74 °C, 58 °C and 84 °C, respectively (Fig. 4.3).

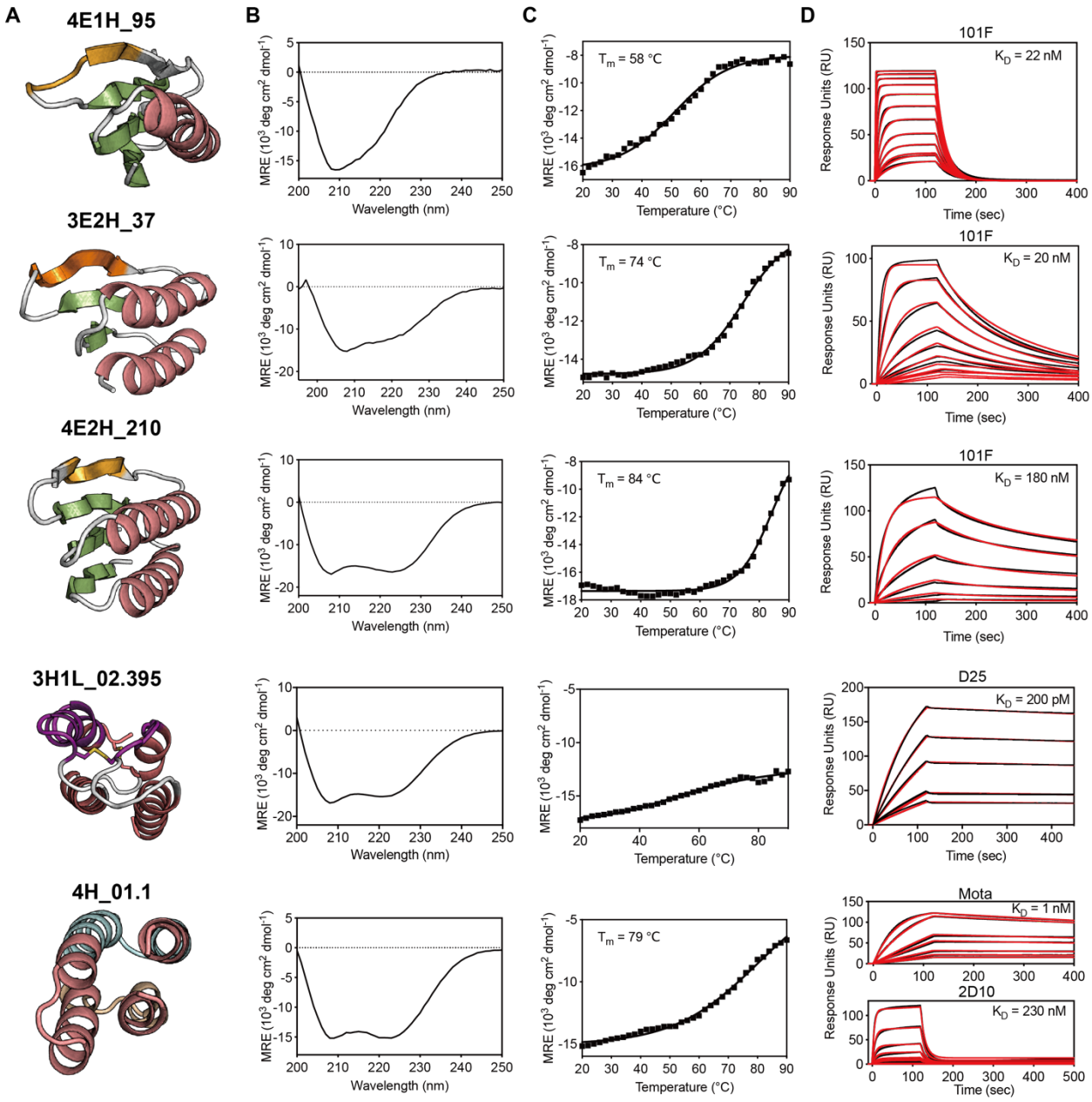


Figure 4.3. Biophysical characterization of lead variants from each topology.

A) Computational models of lead designs. B) Designs adopt folded conformations with the expected predominant secondary structure as assessed by circular dichroism spectroscopy (CD). C) Designed proteins are thermally stable as measured by CD. D) Binding affinity to target antibodies determined by SPR. Sensorgrams are shown in black, fitted curves in red. Binding of 4E1H_95 and 4E2H_210 was measured against immobilized 101F IgG, 3H1L_02.395 against D25 IgG, and 4H_01.1 against both Motavizumab (Mota) and 2D10 IgG.

In terms of connectivity exploration within the TopoBuilder, both variants of the 3H1L topology, presenting the discontinuous site 0 epitope, were recombinantly expressed and bound to their target antibody D25. However, all sequences tested from the 3H1L_01 design series were dimeric in solution, and failed to bind to the 5C4 antibody (Fig. S4.8). The 5C4 antibody was shown to engage site 0 from a different angle compared to D25 (*316*), and binding to both antibodies can thus serve as a probe for the conformational integrity of the discontinuous site 0. In contrast, the best design of the 3H1L_02 series

(3H1L_02.395) showed approximately 1000-fold higher binding affinity to D25 (K_D ~ 200 pM) (Fig. 4.3) than the 3H1L_01 designs, and bound to 5C4 with an affinity of 25 nM (Fig. S4.9). Both binding affinities closely match the reference binding affinity of the respective antibodies binding to prefusion RSVF (K_D = 150 pM), indicating that this discontinuous motif was mimicked accurately in a *de novo* designed protein. Also, 3H1L_02.395 was monomeric in solution, thermostable with a T_m of 59°C, and showed a well-dispersed HSQC NMR spectrum (Fig. 4.3 and Fig. S4.9).

Towards endowing computationally designed proteins with multiple functional sites, we assembled two antibody binding motifs (HLH/HL) in a single topology on a back-to-back orientation and sampled two distinct sheer angles, 0° (4H_01) and 30° (4H_02) (Fig. 4.2). We next screened a library based on the designed sequences of the topological variants using a dual-binding selection criterion to ensure that both motifs were presented in their native conformation. Deep sequencing of populations sorted under stringent selection conditions showed that the 4H_01 design series were highly enriched. These results further confirm our *in silico* analysis, where the 4H_01 design series showed a lower structural drift than the 4H_02 series (Fig. S4.10). We expressed 20 variants of the 4H_01 design series, out of which 14 were soluble upon purification and were then characterized biochemically. Overall, the designs showed CD spectra typical of alpha-helical proteins and bound to both target antibodies (Mota and 2D10) with K_Ds ranging from 0.5 to 10 nM for Mota and 200 to 400 nM for 2D10 (Fig. S4.10). The best variant, 4H_01.01, was well folded and thermostable according to CD and NMR, showing a T_m of 75 °C (Fig. 4.3). 4H_01.01 bound with a K_D of 0.5 nM to Mota and 200 nM to 2D10. Overall, we showed that the subtle topological tuning sampled by the TopoBuilder had an important impact on the stability and function of the designed proteins.

Experimentally determined structures confirm the accuracy of the designs

We next solved a crystal structure of 4E1H.95 in complex with the target antibody (101F) at 2.9 Å resolution. The structure is in close agreement with the computationally designed model, with a full-atom RMSD of 1.8 Å (Fig. 4.4A). The designed backbone hydrogen bonding network was accurately recapitulated in the crystal structure compared to the computational model, and the overall paring of the β-sheet was accurately recapitulated. The β-bulged structure of the site IV epitope also retained the designed hydrogen bond network supported by a contiguous strand and mimicked with sub-angstrom accuracy the conformation of the antibody-bound peptide epitope (Fig. 4.4B) (*169*).

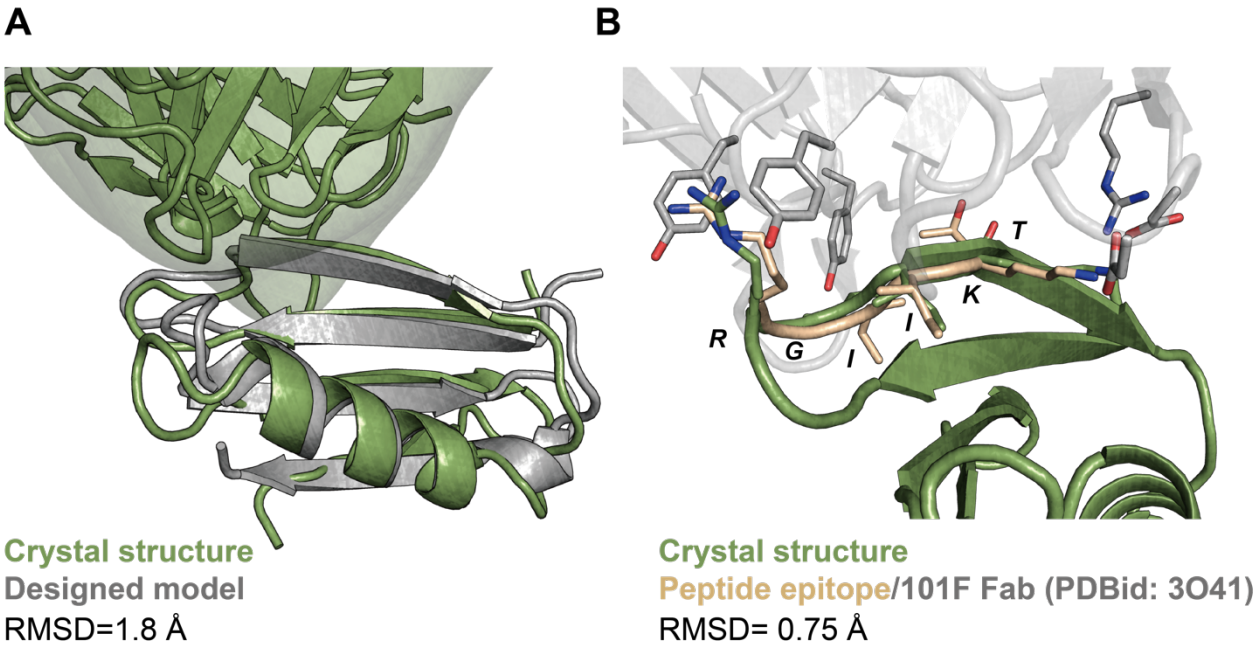


Figure 4.4. Crystal structure of 4E1H.95 design is in close agreement with the computational model.

A) A crystal structure of 4E1H_95 in complex with 101F Fab (green surface) was solved at 2.9 Å resolution, in close agreement with the design model (RMSD 1.8 Å). B) Superposition of 4E1H_95 crystal structure with the antibody-bound site IV peptide epitope (PDB ID: 3O41), shows accurate mimicry of the desired conformation with an RSMD of 0.9 Å.

Biosensors containing de novo designed modules detect epitope-specific antibodies

In recent years, numerous high-throughput antibody isolation campaigns have enlightened the molecular basis of humoral responses, yielding large repertoires of monoclonal antibodies from human and animal.

These efforts have revealed correlates of antibody specificities with the degree of protection afforded against several pathogens upon vaccination or natural infection (*103, 104, 317*). However, assays for the detection and quantification of antibody responses with epitope-level resolution in bulk sera are laborious, and often require complex liquid handling steps. With the growing structural understanding of pathogen's proteins and their neutralization epitopes, assays to dissect epitope-specific responses in bulk sera are urgently needed. These assays will contribute to the improvement of our understanding of natural and vaccine-induced immunity with single epitope resolution, and provide the grounds to design better vaccines and therapeutics.

We used the neutralizing epitopes identified in RSVF as a test case and hypothesized that a biosensor based on the designed site IV and site 0 scaffolds would be an ideal tool to detect and quantify epitope-specific antibodies in bulk serum. We designed a biosensor based on the recently developed Lumabs platform, as shown in Fig. 4.5A (*318*). Briefly, the sensor is based on bioluminescence resonance energy transfer (BRET) where in their "closed" conformation, BRET occurs between the nanoluciferase (Nluc) donor and the mNeonGreen (mNG) acceptor. In presence of antibodies specific for the presented epitope, the sensor adopts an open state, as measured by a decrease in BRET ratio (Fig. 4.5B).

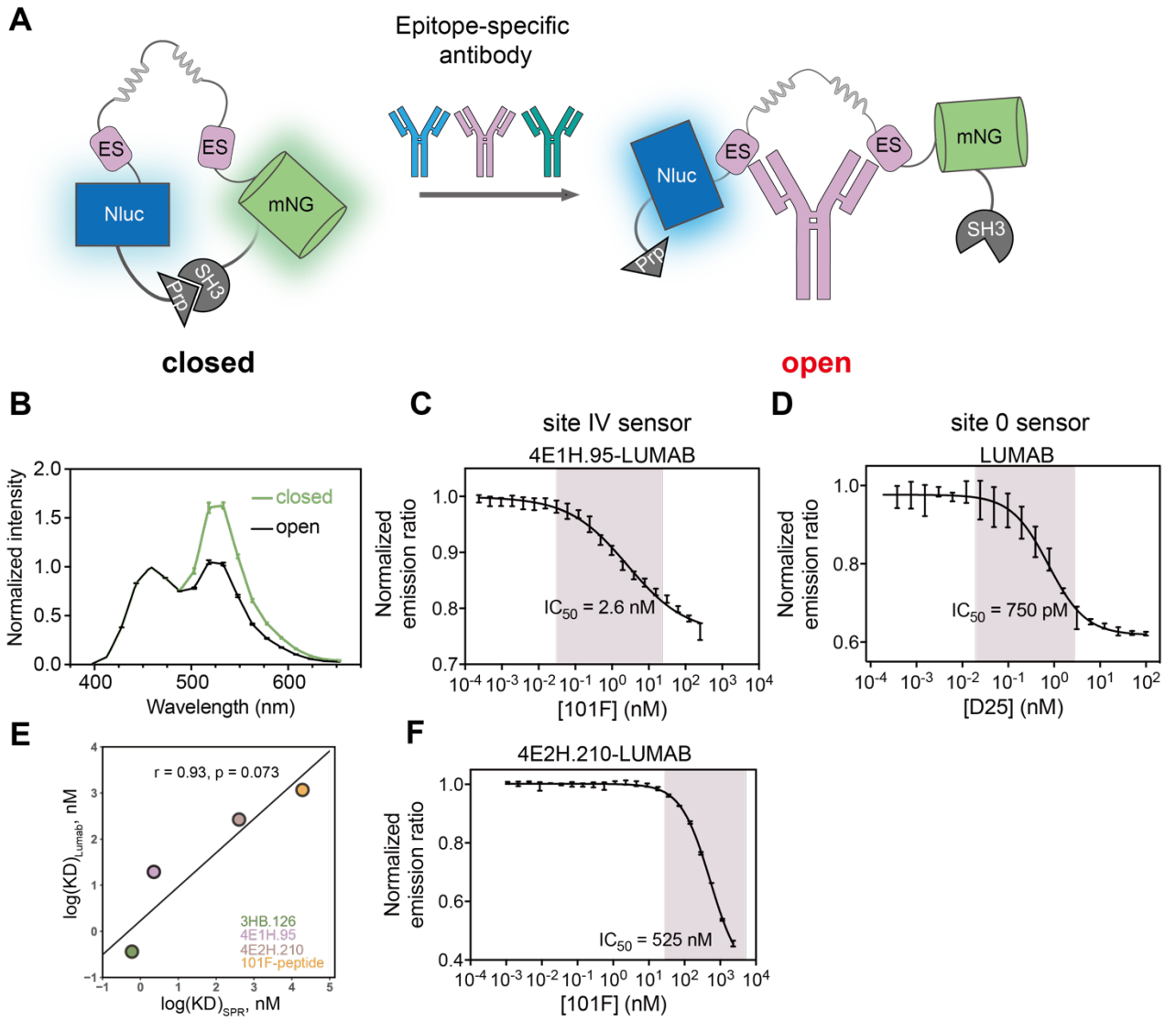


Figure 4.5. Antibody biosensors based on *de novo* designed proteins for the detection of site-specific responses.

A) Schematic representation of the Lumabs sensor platform. NanoLuc luciferase (Nluc) and mNeonGreen (mNG) are held in close proximity by two helper domains (SH3 domain and a proline-rich-peptide, Prp), allowing efficient BRET between Nluc and mNG. Antibodies binding to the designed epitope-scaffolds (ES) disrupt the weak Prp-SH3 interaction, opening the sensor and decreasing BRET efficiency. The fluorescence signals were normalized with the emission intensity of Nluc. B) Luminescence spectrum of the sensor in a closed (*63*) and open (*242*) conformation. C) Characterization of Site IV sensor based on the 4E1H.95 scaffold upon titration of 101F IgG. Plotted is the ratio between 518 and 458 nm emission upon titration of 101F antibody. The IC_{50} of the 4E1H.95_Lumab was 2.6 nM, which is in line with the determined binding affinity of 4E1H.95 to 101F (K_D = 20 nM). D) Characterization of the site 0 sensor based on S0_2.126 scaffold upon titration of D25 IgG. The IC_{50} using D25 mAb was 750 pM, also in line with the high affinity of S0_2.126 for D25 (~50 pM), allowing detection of site 0 antibodies of less than 0.1 µg/ml. E) Correlation of determined K_D derived from LUMAB sensor and SPR. F) Characterization of Site IV sensor based on the 4E2H.210 design, showing different sensitivity for target antibody, B-D, F) Data points are plotted as mean ± SEM.

As a proof-of-concept, we developed sensors for two important RSVF neutralization epitopes (site IV and site 0). The site IV sensor is based on the 4E1H.95 design, and the site 0 sensor is based on the S0_2.126 scaffold presented previously.

Both sensors were robustly expressed in E.*coli*, and showed a concentration-dependent decrease in BRET signal upon the addition of 101F or D25 mAbs (Fig 4.5C, D). The sensors were not responsive to an irrelevant IgG. Both D25 and 101F mAbs were detected at sub-nanomolar concentration (<0.1 μg/ml). To assess the importance of presenting the epitopes stabilized in their native conformation in the designed proteins we tested a Lumab sensor containing only the site IV peptide epitope, this sensor showed a remarkably lower performance with approximately 500-fold drop in sensitivity. This observation shows that the correct stabilization of the epitope through *de novo* design means was critical to have highly functional sensors.

Together, the sensors presented will enable the study of serum responses beyond bulk titers, allowing to dissect the level of epitope-specific antibodies elicited by viral infection and different vaccination strategies, potentially becoming a versatile tool for immune monitoring.

Bi-functionalized de novo designs regulate the activity of synthetic cell-surface receptors

Many fundamental biological processes at the cellular and organismal levels involve signaling pathways that are triggered by the interaction of soluble factors with cell-surface receptors. Within the realm of synthetic biology, considerable progress has been made towards the design of synthetic receptors, including chimeric antigenic receptors (*319*), the SynNotch (*320*) and others (*321*). These receptors control intracellular pathways and endow mammalian cells with novel sensor-effector responses. However, they still rely on natural ligands to modulate their activity, potentially raising concerns due to the lack of orthogonality with endogenous ligands.

With our bifunctional *de novo* design (4H_01), we sought to create a synthetic effector-receptor system that modulates the heterodimerization of synthetic receptor creating an orthogonal signaling pathway to control gene expression (Fig. 4.6A). To do so, we used the recently developed GEMS synthetic receptor platform developed by Scheller and colleagues (*322*), which we modified by fusing scFvs (single-chain variable fragments) of Mota and 2D10 mAbs to the extracellular domains of EpoR (erythropoietin receptor). We used the FGFR1 (fibroblast growth factor receptor 1) intracellular domains for triggering the endogenous MAPK signaling and express a reporter gene as output (Fig. 4.6A). The signaling promoted by the heterodimerization of the receptor is quantified by the expression of the reporter protein SEAP (human placental secreted alkaline phosphatase) (Fig. 4.6A).

As predicted by the structural model of the design, we confirmed biochemically that 4H_01 bound simultaneously to both mAbs (Fig. 4.6B). Negative-stain electron microscopy confirmed this finding, and through single-particle reconstruction of the ternary complex we observed that their binding mode was in close agreement with our computational model (Fig. 4.6C).

We then introduced the synthetic GEMS receptors ($scfv_{Mota}$-EpoR-FGFR1 and $scfv_{2D10}$-EpoR-FGFR1) into HEK cells and measured SEAP expression upon titration of 4H_01. We observed transgene expression when the engineered cells expressed both extracellular scFv-fused receptors and the response curve revealed a typical dose-dependent activation with a high sensitivity (EC_{50} = 214 pM) (Fig. 4.6D). Furthermore, gene expression was absent when cells were only transfected with a single scFv-fused receptor (Fig. 4.6D).

In native systems, cellular signaling is often the result of a delicate balance between agonists and antagonists, allowing to finely tune cellular behavior. Towards the engineering of a synthetic sensor-effector system controlled by computationally designed proteins, we used a previously designed epitope-scaffold (FFLM) containing only the site II epitope to work as an antagonist. FFLM binds to the mota mAb with high affinity (K_D= 20 pM) as previously reported by Sesterhenn and colleagues (*323*). We tested the response of the synthetic receptors to mixtures of agonist and antagonist designs in different molar ratios (Fig. 4.6A), and observed an antagonist dose-dependent response with increasing amount of inhibitor (Fig. 4.6E). The effective EC_{50} was shifted correspondingly from 214 pM (in the absence of an inhibitor) to 459 pM (100 pM of inhibitor), 2 nM (1 nM of inhibitor) and 34 nM (10 nM of inhibitor).

Altogether, we show a proof-of-concept for *de novo* design proteins to control the activity of synthetic receptors. The bi-functionalization of the designs was only made possible by the bottom-up approach described as we could not find a suitable design template in the natural protein repertoire. We also demonstrate the fine-tuning of the receptor's activity by using a computationally designed antagonist, which enables a more precise signaling regulation opening possibilities to encode complex regulatory mechanisms (e.g., feedback loops). This novel synthetic system based on non-natural ligand-receptor pair could be useful to dynamically control gene expression by tuning the agonist/antagonist ratios and should provide a general versatile platform for precision-guided cell-based therapeutics.

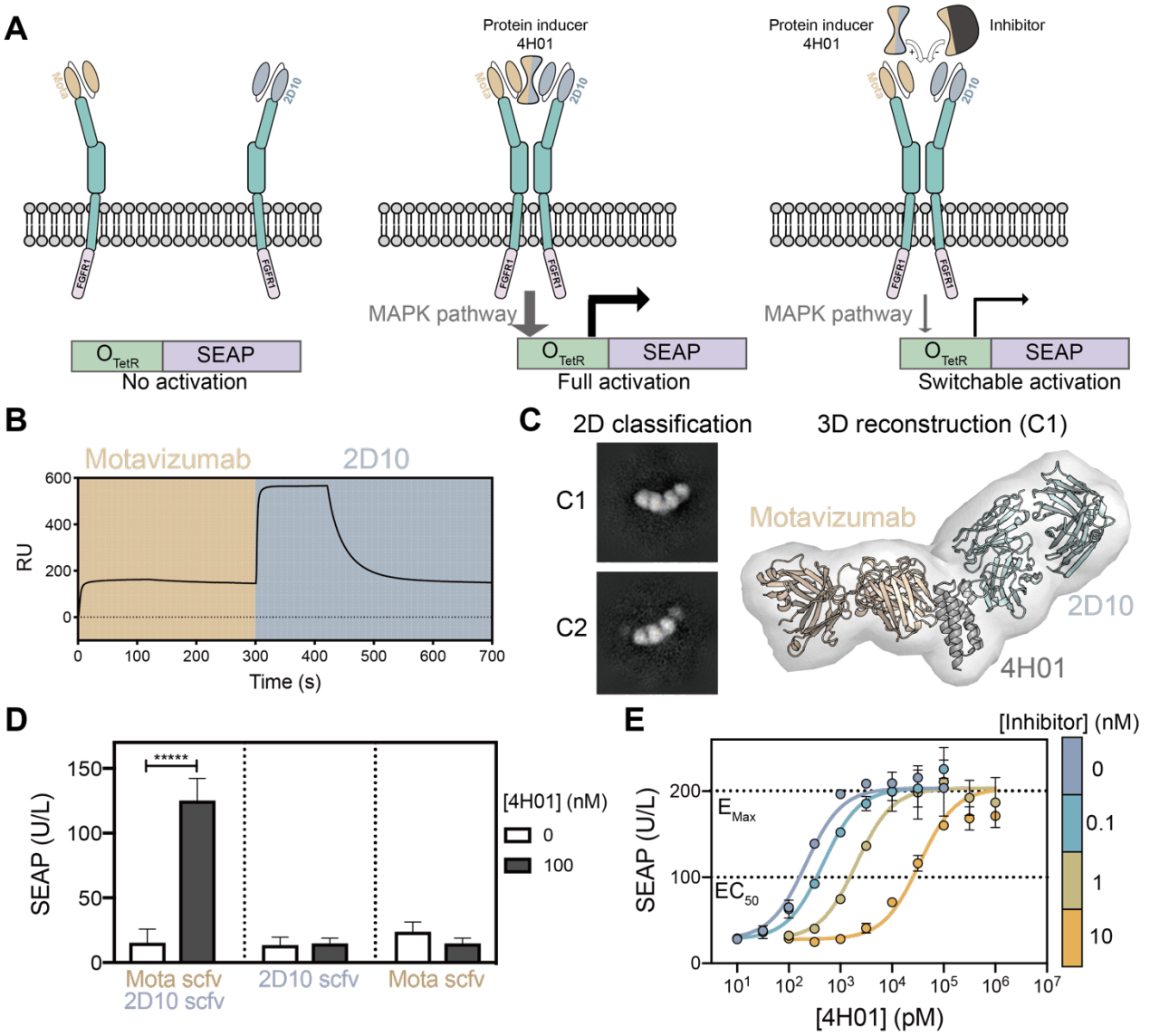


Figure 4.6. Bi-functional *de novo* design controls the activity of synthetic receptors.

A) Schematic representation of the receptor's architecture and the regulation of the signaling activity monitored through the expression of the reporter gene (SEAP). Two scFvs were used as ectodomains of the synthetic receptors which hetero-dimerize in the presence of 4H_01, resulting in activation of the MAPK signaling pathway that regulates the expression of a reporter gene. In presence of a computationally designed high-affinity antagonist carrying the Mota-epitope only, the heterodimerization is disrupted. B) SPR binding response of 4H_01 to target antibodies. 4H_01 is able to bind simultaneously to Mota and 2D10, as shown by the 2D10 binding signal in presence of saturating amounts of Mota. C) EM images of two 2D average classes and 3D reconstruction of the 4H_01 design complexed with two mAbs. The complexed model of 4H_01 with the two mAbs (shown in cartoon representation) is in close agreement with the EM 3D reconstruction of the complex (shown in gray surface). D) Responsiveness of dual and single receptor expressing cells. 15 hours after transfection, cells were treated with 100 nM of 4H_01. The dose-response rate was quantified in terms of SEAP expression after 24 hours. E) Inhibitory effect of a high-affinity antagonist. Transfected cells were treated with different molar ratios of 4H_01 and FFLM for 24 hours and SEAP production was quantified. The EC_{50} was determined by fitting the SEAP response with the dose-response equation provided by Prism (goodness of fit greater than $R^2 > 0.95$). Data shown in D) and E) was the result of three independent replicates (n = 3).

4.3 Discussion and Conclusions

The design of novel functional proteins is a fundamental test for our understanding of protein structure and function, has it has shown potential in multiple arenas of basic biology and biotechnology. However, the majority of studies where function was the main objective, existing proteins were repurposed and used as templates to transplant functional sites (*76*). This "top-down" approach faces inherent limitations for functional motifs with high structural complexity (e.g., irregular conformations or multi-segment), specifically, when the number of designable templates available in the natural protein repertoire is sparse. To overcome this critical bottleneck, we present a "bottom-up" approach for the *de novo* design of functional proteins.

In the described motif-centric *de novo* design approach, we account for the structural and sequence constraints of the functional motifs, serving as a "functional seed" to build tailored scaffolds that stabilize their native conformation. While previous reports have shown a conceptually similar "bottom-up" building and design of functional proteins, only one topology per functional site was reported, and topologies were structurally limited to helical bundles (*102, 176, 311, 312, 324*).

Thus, the design strategy presented here has significant advantages with regard to its general utility for the design of *de novo* functional proteins, allowing to systematically explore a wide variety of possible topologies tailored to a given functional motif, control their connectivity, and fine-tune the secondary structure arrangement through parametric sampling. To validate the methodology, we have attempted to design six different topologies, four of which have yielded well-folded, stable proteins. The functional sites were presented accurately in the *de novo* proteins, as shown by high affinity binding to their target antibodies. Importantly, a crystal structure of one of the designs confirmed the atomic level accuracy of the design model.

Regarding the functional significance of our work, one of the most attractive domains for *de novo* design is to create proteins with functional properties that are out of reach for natural proteins. Herein, we have shown the utility of the motif-centric design strategy to design functional proteins for two important applications that are largely out of reach for natural proteins.

First, the designed RSV epitope scaffolds were functional when embedded in a biosensor, serving as tools to detect and quantify epitope-specific antibodies. With our rapidly growing understanding of protective antiviral antibody responses, monitoring epitope-specific antibody responses on serum level is of utmost importance to assess the efficacy of novel vaccine candidates and decipher immune responses upon natural infection. Thus, the presented biosensors may become valuable diagnostic tools to dissect serum responses beyond bulk serum titers.

Second, we have proven the utility of the TopoBuilder beyond single functional sites, for the *de novo* design of proteins accommodating multiple binding sites. Interestingly, we foresee that upcoming protein design challenges (e.g., protein-protein interactions, ligand binding, enzymes) will benefit from this ability of encoding more functional components in small protein domains. As in nature, carrying multiple functional sites is a common feature in proteins that mediate common biological functions (*325, 326*). However, endowing *de novo* proteins with more than one (irregular) binding site has remained largely unachieved.

Leveraging the ability of the TopoBuilder to assemble secondary structures around multiple given functional motifs, we have shown the design of a helical topology with two binding sites in a defined spatial orientation, with the predicted antibody mode closely resembling a negative stain EM reconstruction.

An attractive application domain for *de novo* designed proteins is in synthetic biology; here we use a bi-functionalized *de novo* design to mediate the heterodimerization of synthetic receptors to trigger orthogonally controlled signaling pathways.

As a proof-of-principle, we have shown a relevant biological function of the *de novo* ligand to control a synthetic signaling pathway, resulting in transgene expression. Moreover, we showed that a pair of computationally designed proteins could function as agonists/antagonists to regulate signaling strengths. On the structural side of membrane-receptor activation, it is important to note that the precise spatial configuration of ligand-induced receptor dimerization is a critical determinant for efficient signaling (*327*), the TopoBuilder should prove as a versatile tool to design topologically controlled ligands to tune cellular function as demonstrated in our work.

Altogether, the presented motif-centric design strategy is a step forward for the design of functional proteins. While herein the functions were limited to antibody binding motifs, the approach should be widely applicable to any known functional motif, and ultimately, the ability to precisely control topological and physicochemical features will enable the design of novel functions absent in natural proteins.

4.4 Supplementary information

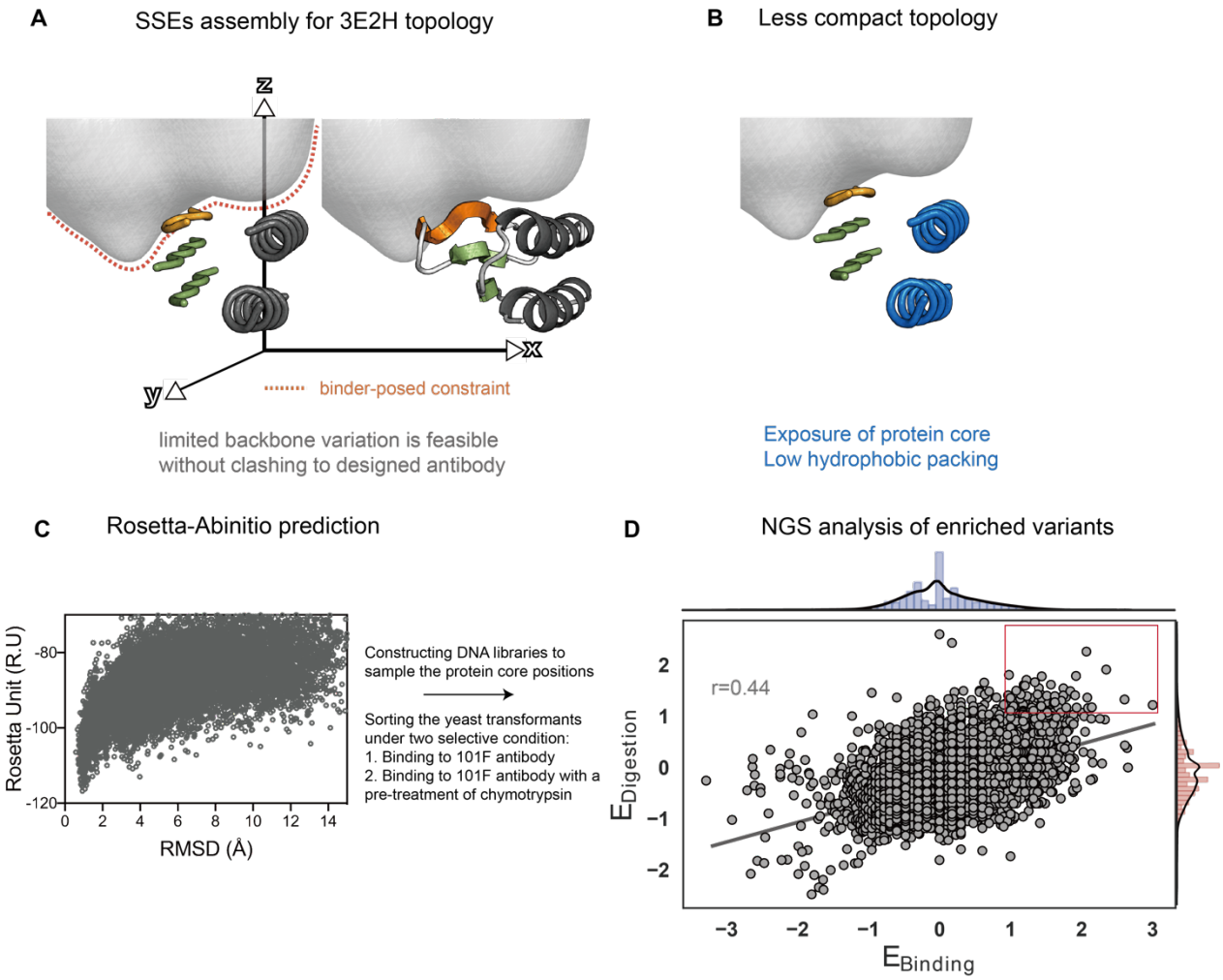


Figure S 4.1. Topological assembly of 3E2H topology for stabilization of 101F strand motif.

A) The binding constraint imposed by binder limited the sampling space for assembling 3E2H topology, in which only the few orientations were allowed to be sampled. B) The arrangement of SSEs potentially has the low core compactness. C) The design of the best energy was predicted by forward folding prediction by Rosetta-*Abinitio* showing the designed sequence is predicted to fold as computational model. Based on this sequence, we further construct the library combinatorial sampling the core position of defined sequence and screen the constructs by yeast surface display. D) Enrichment analysis of the sorted populations, each dot represents a unique sequence. The enrichment was computed for binding to 101F (x-axis) and the residual binding after the random protease digestion (y-axis). Sequences that were strongly enriched locates in the red square.

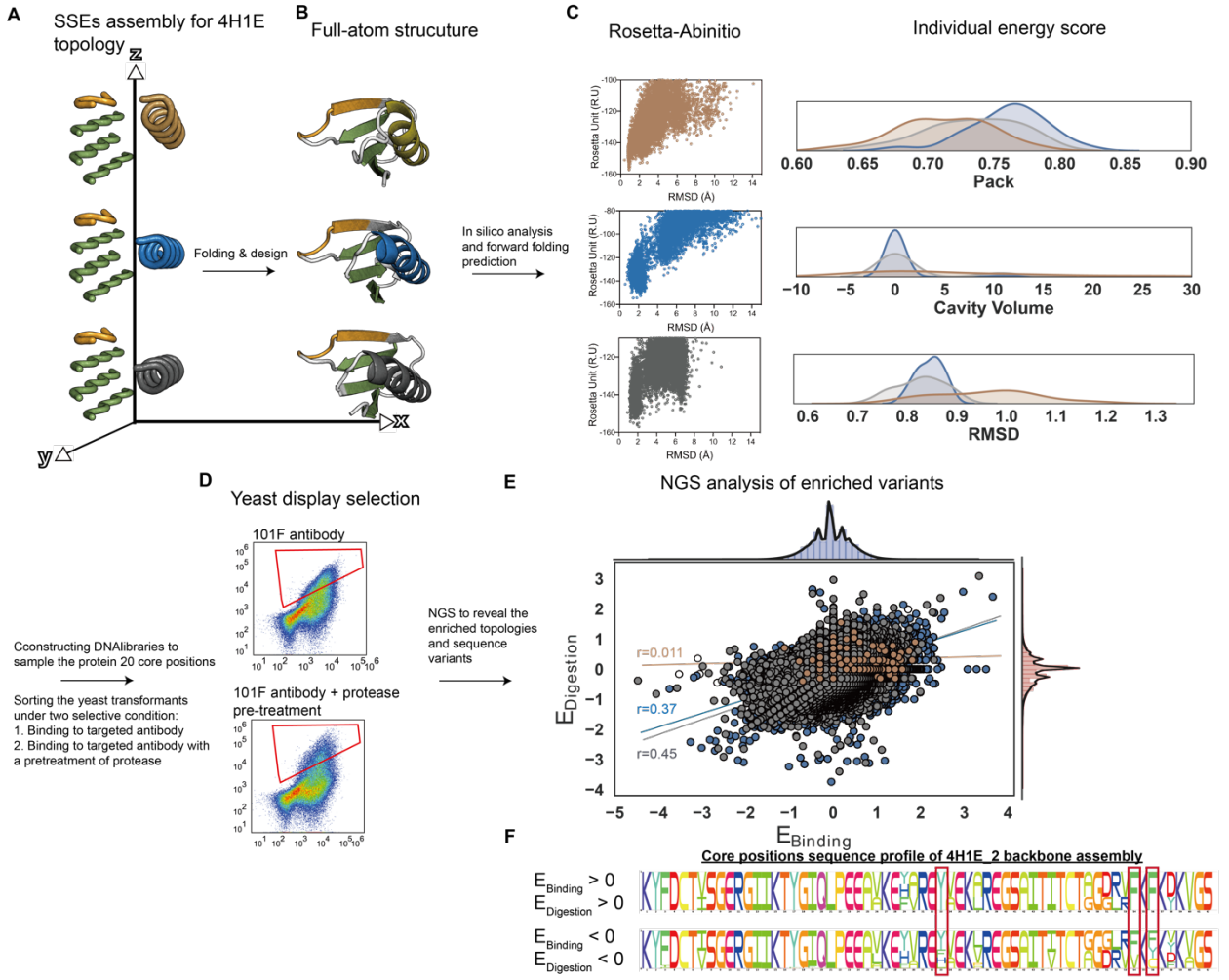


Figure S 4.2. Modular assembly of 4E1H topology for presentation of 101F strand motif.

A) Three different configurations of the 4E1H topology were built. The helical layer was varied to tilt by 30°, 10°, 0° relative to the β-sheet layer stabilizing 101F motif to generate 4E1H_1 (*159*), 4E1H_2 (blue) and 4E1H_3 (*242*) respectively. B) The full-atom structure of each backbone after the folding and sequence design performed by Rosetta-FunFoldDes. C) The designs of the best energy in each backbone variation were predicted by forward folding prediction by Rosetta-*Abinitio*. Besides, each computational energy term was plotted to represent the distribution of decoy in the energy state. Based on this sequence, we further construct the libraries for each backbone variation that combinatorial sampling the core position of defined sequence and pooled all the libraries together to screen the constructs with selected features by yeast surface display. D) Designs from three templates were screened in a single yeast library, and clones with high affinity binding to 101F and highly resistant to protease were sorted and sequenced. E) Enrichment analysis of the sorted populations, each dot represents a unique sequence. The enrichment was computed for binding to 101F (x-axis) and the residual binding after the random protease digestion (y-axis). Only the two backbone orientations showed the linear correlation with two selection dimensions. F) Sequence profile of the enriched population and non-enriched population in 4E1H_1 design series.

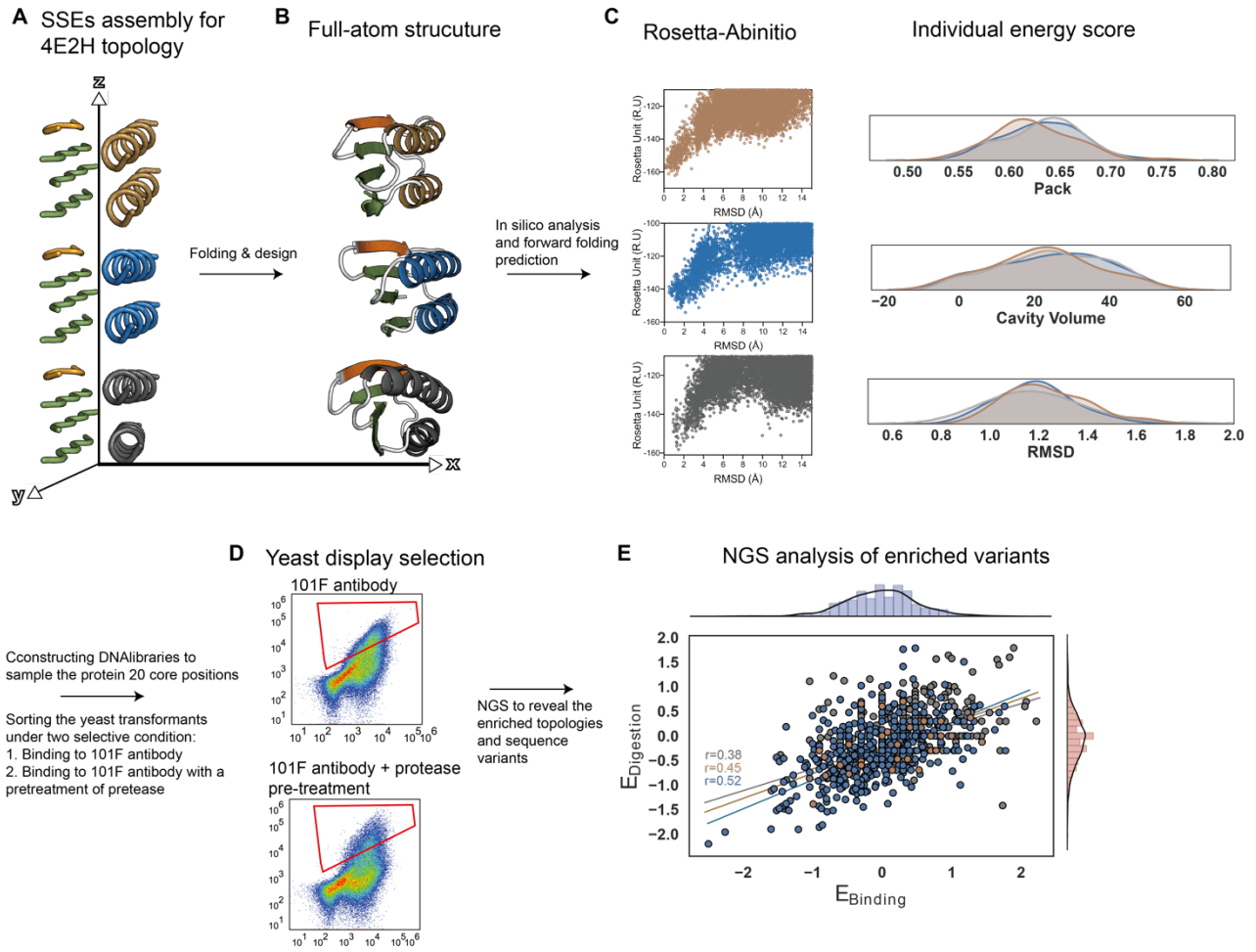


Figure S 4.3. SSEs assembly of 4E2H topology for stabilization of 101F binding motif.

A) Three different configurations of the 4E2H topology were built. The helical layer was varied to tilt by -15°, 0°, 15° relative to the β-sheet layer stabilizing 101F motif to generate 4E2H_1 (*159*), 4E2H_2 (blue) and 4E2H_3 (*242*) respectively. B) The full-atom structure of each backbone after the folding and sequence design performed by Rosetta-FunFoldDes. C) The designs of the best energy in each backbone variation were predicted by forward folding prediction by Rosetta-*Abinitio*. Besides, each computational energy term was plotted to represent the distribution of decoy in the energy state. Based on this sequence, we further construct the libraries for each backbone variation that combinatorial sampling the core position of defined sequence and pooled all the libraries together to screen the constructs with selected features by yeast surface display. D) Designs from three templates were screened in a single yeast library, and clones with high affinity binding to 101F and highly resistant to protease were sorted and sequenced. E) Enrichment analysis of the sorted populations, each dot represents a unique sequence. The enrichment was computed for binding to 101F (x-axis) and the residual binding after the random protease digestion (y-axis). All three backbone orientations demonstrated the linear correlation with two selection dimensions, indicating the higher binding affinity clone potentially obtaining the feature of high stability.

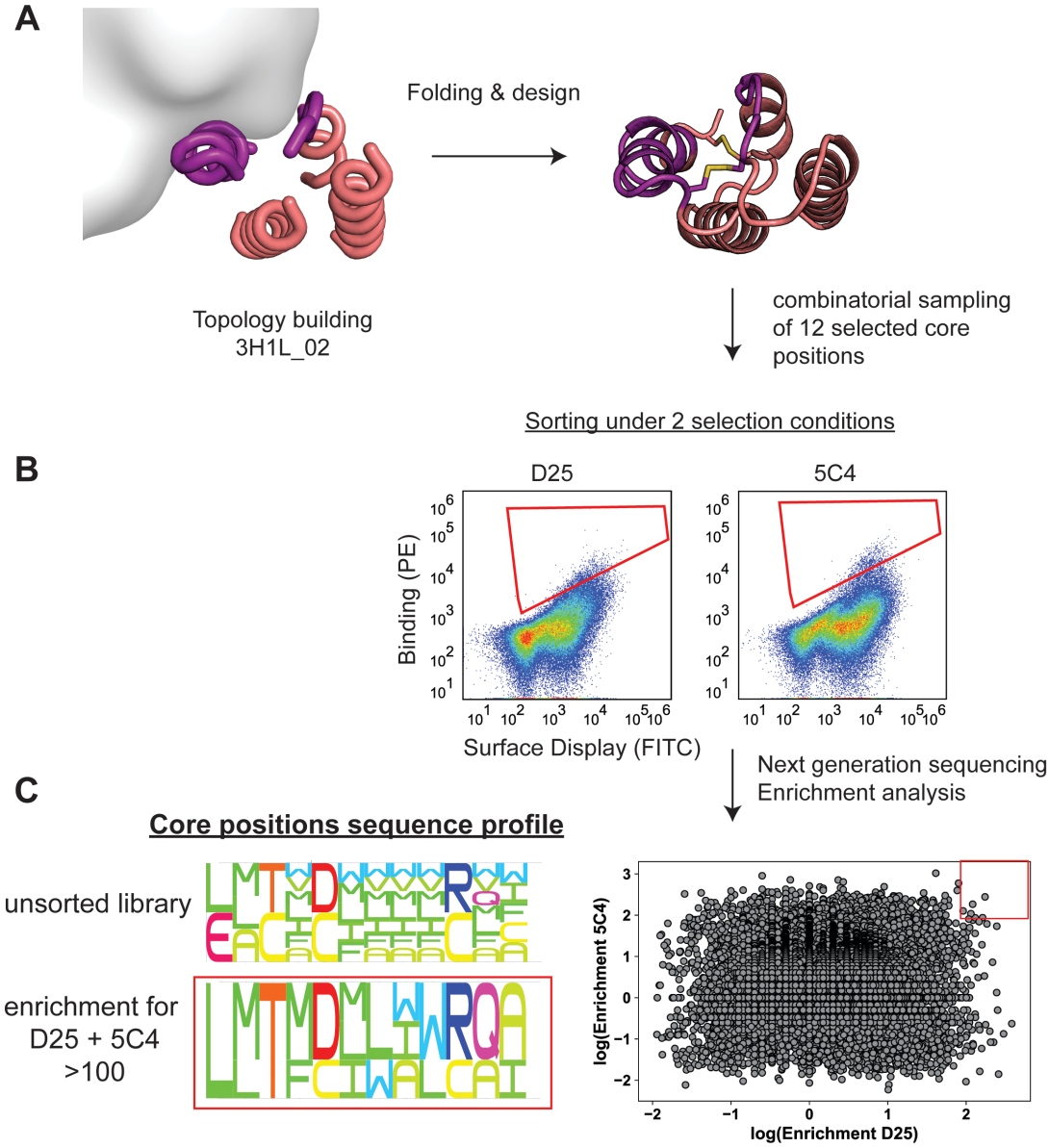


Figure S 4.4. Design and high-throughput screening for 3H1L_02 topology.

A) The topology was assembled using the TopoBuilder, followed by folding and design using Rosetta FunFolDes. The site 0 epitope is shown in purple, the secondary structure elements that were built by the TopoBuilder are colored salmon. From the ensemble of designed sequences, 12 critical core positions were selected and encoded in a combinatorial sequence library (theoretical size ~10^7). B) The library was transformed into yeast, and clones with high-affinity binding to D25 and 5C4 were sorted (red sorting gate). C) Right: The sorted populations were sequenced using next-generation sequencing, and enrichment scores were computed for each sequence binding to 5C4 and D25. Sequences that were strongly enriched for D25 and 5C4 (top right corner, red) were selected for recombinant expression and biophysical characterization. Left: Sequence profile of the 12 selected core positions in the unsorted library, and of sequences that were >100 fold enriched for D25 and 5C4 binding.

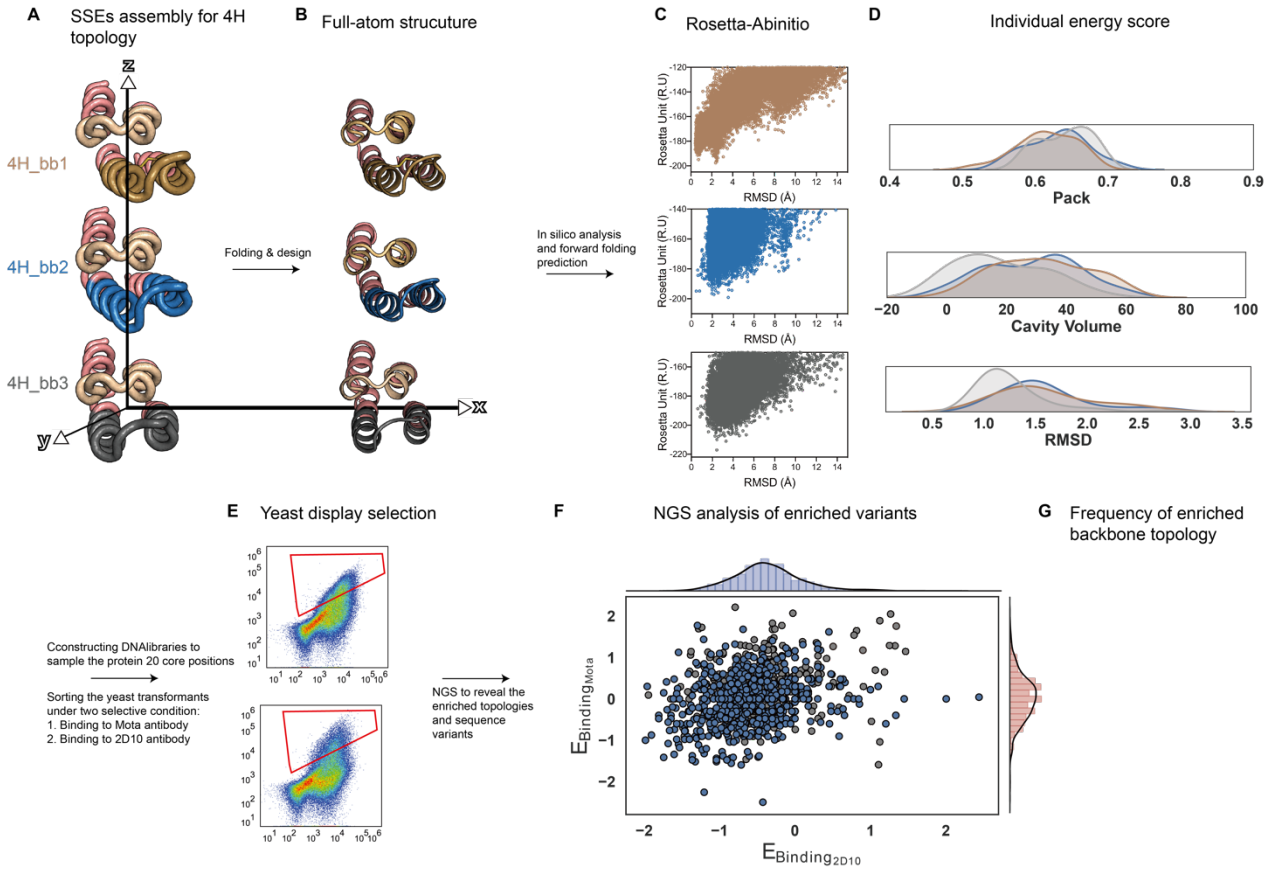


Figure S 4.5. Topological tuning of the 4H topology for optimal presentation of both binding motifs.

A) Two different configurations of the 4H topology were built. For 4H_02, the helical layer containing the site II epitope was tilted by 20° relative to the helical layer presenting the 2D10 binding motif. B) Designs from both templates were screened in a single yeast library, and clones with high affinity binding to Motavizumab or 2D10 were sorted and sequenced. C) Enrichment analysis of the sorted populations, each dot represents a unique sequence. Sequences from the 4H_01 design series are colored grey, 4H_02-based designs blue. For each sequence, an enrichment score (E) was computed based on its frequency under high selective pressure versus low selective pressure (see chapter 3 methods for details). The enrichment was computed for binding to Motavizumab and to 2D10. Sequences that are strongly enriched for binding to both antibodies (red square) uniquely derive from the 4H_01 design series (indicated in grey).

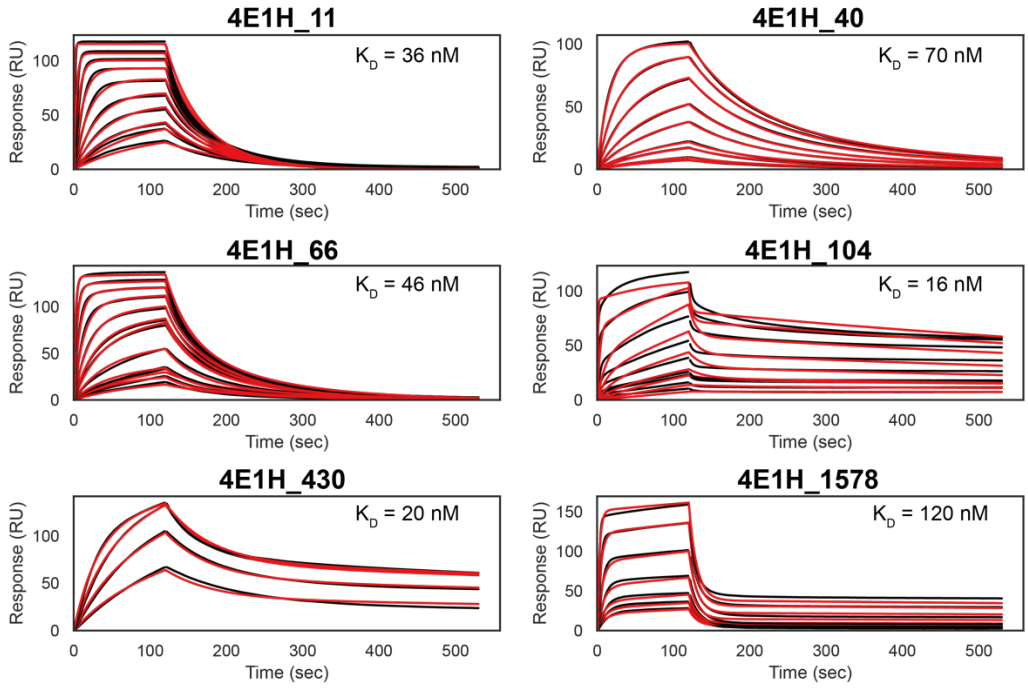


Figure S 4.6. Biophysical characterization of 4E1H design series.

The binding affinity of selected 4E1H designs was measured by SPR and revealed the binding affinity around 20-120 nM.

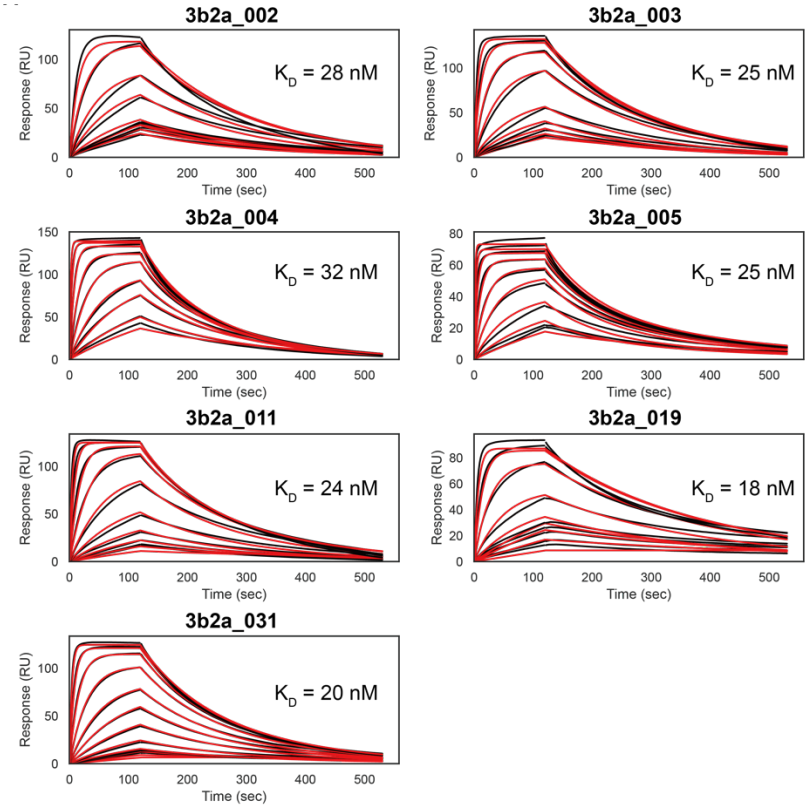


Figure S 4.7. Affinity measurement and solution oligomeric state of 3E2H design series.

The binding affinity was measured by immobilization of 3E2H on the sensor chip to detect the 101F Fab binding.

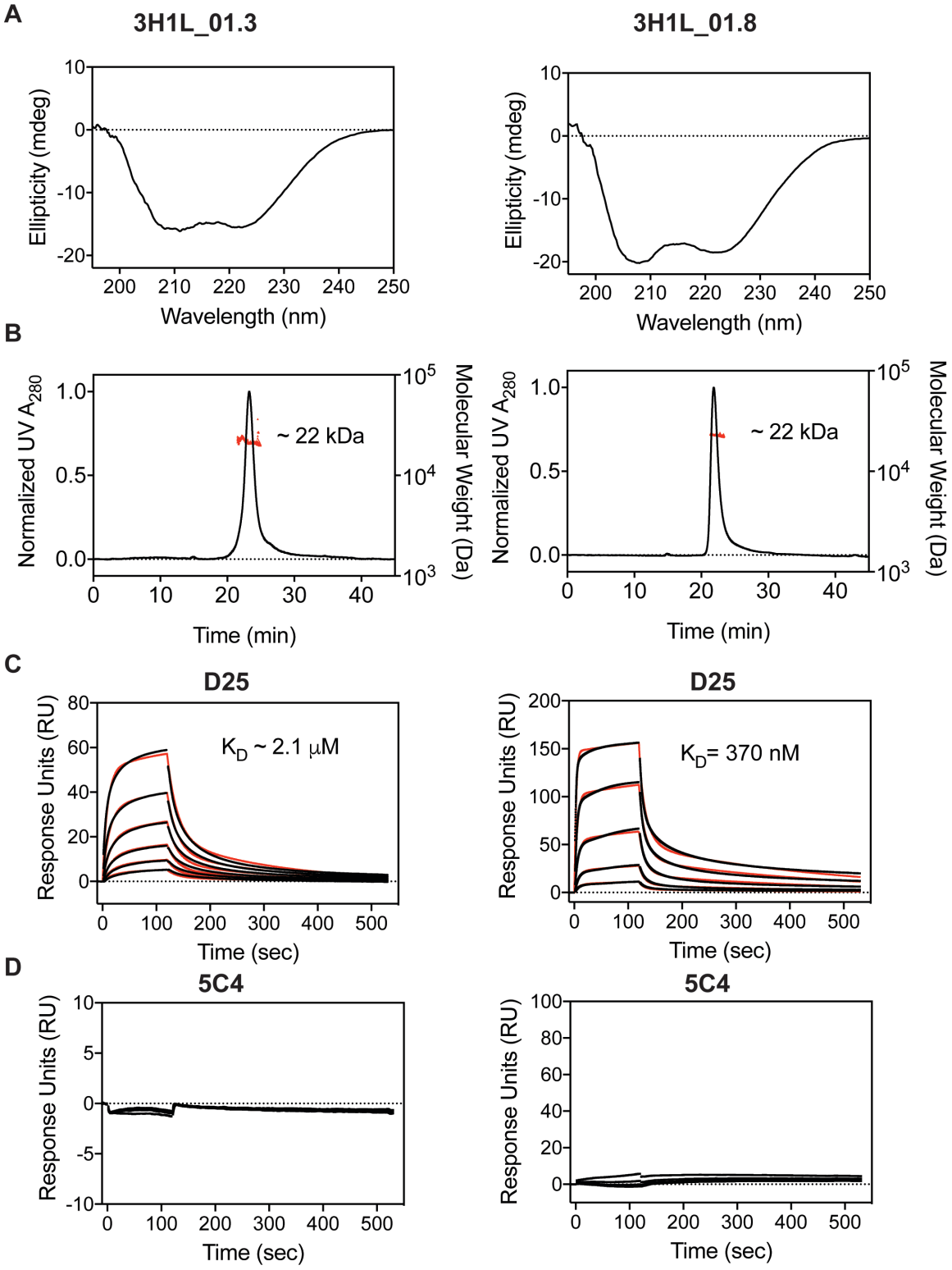


Figure S 4.8. Biophysical characterization of best 3H1L_01 variants.

Following yeast display screening, 10 sequences were selected for recombinant expression and biophysical characterization. Data are shown for the two best sequences (3H1L_01.3 and 3H1L_01.8) according to their D25 binding affinity. A) Circular dichroism spectra at 25 °C are in agreement with a helical protein. B) Size-exclusion chromatography with on-line multi-angle light scattering (SEC-MALS) shows that both designs are dimeric in solution. The expected molecular weight of monomeric 3H1L_01 designs is ~11 kDa. C,D) Measurement of binding affinity

to D25 and 5C4 by SPR. 3H1L_01 designs were immobilized on the sensor chip surface, and D25 or 5C4 Fab were injected as analyte in various concentrations. Kinetic dissociation constants are indicated, with no detectable binding to 5C4.

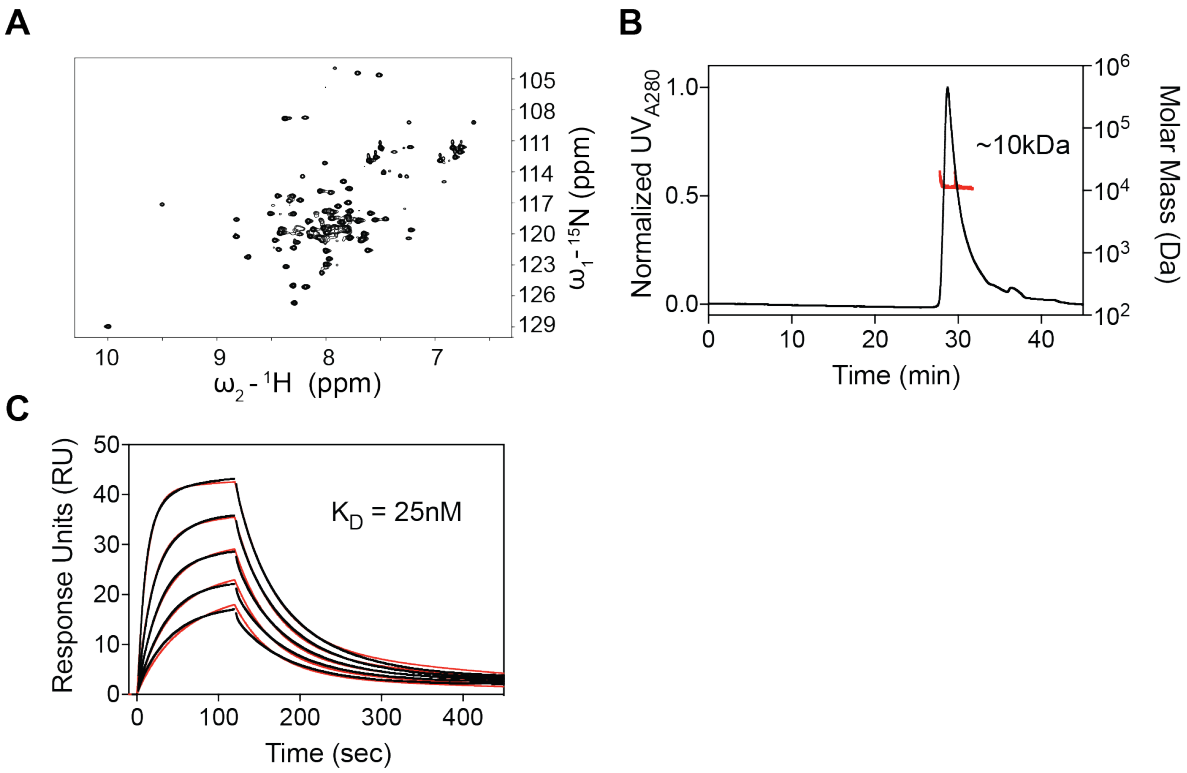


Figure S 4.9. Extended biophysical characterization of 3H1L_02.395.

A) The 2D NMR ^{15}N HSQC spectrum for 3H1L_02.395 is well dispersed, confirming that it is well folded in solution. B) 3H1L_02.395 is monomeric in solution, with the determined molecular weight of 10 kDa closely matching its theoretical molecular weight of 10.5 kDa. C) 3H1L_02.395 binds with a K_D of 25 nM to 5C4, as determined by SPR.

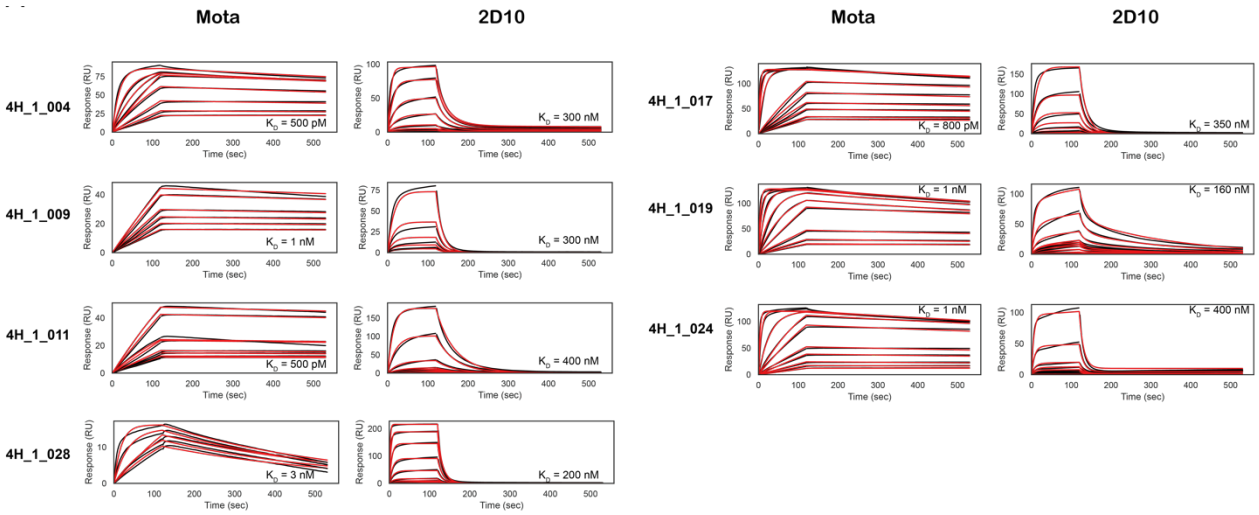


Figure S 4.10. Biophysical characterization of 4H_1 design series.

The affinity measurement by SPR for all the selected designs for 4H_1 series against two different targeting antibodies (Mota and 2D10).

Chapter 5 Conclusions and Perspectives

Protein engineering techniques have rapidly improved over the last decades, and engineered diverse proteins have made an impact on both clinical and industrial applications. The chapters presented here describe several new computational algorithms and conceptual strategies to aid in protein engineering, especially from the angle of using protein design to create functional proteins. Our results provide valuable insights to improve the generalization of design protocols that can be utilized to stabilize functional motifs for vaccine development and synthetic biology. The following sections explain in detail the significant progress achieved throughout my graduate studies.

5.1 *De novo* protein scaffolds enable the stability of structurally complex functional motifs

Generally, robust approaches for the design of functional proteins rely on natural protein scaffolds that can optimally fit a functional motif, usually referred to as protein grafting. Despite the successes in repurposing single regular and helical motifs for novel biological functions, one should notice this well-used approach is intrinsically limited due to two aspects: 1) the finite number of protein folds and conformations explored by nature with enough local structural similarity to fit the structure of the functional sites; 2) related to the first statement, the inability of transferring complex functional motifs into heterologous protein structures. The transplantation of a structurally complex motif onto a heterologous scaffold requires extraordinary precision of the design process. Specifically, due to the need to stabilize all the motif segments within the overall protein structure. The lack of compatible protein backbones to achieve the motif stabilization is often referred as the lack of "designability" of the structural templates. As described in the Aims chapter, we overcame these challenges via: I) the increased structural control over existing protein backbone degrees of freedom; II) the construction of function-compatible backbones from scratch. These two modalities correspond to two conceptually distinct design approaches for functional protein design: **top-down** versus **bottom-up**.

We advanced the algorithm FunFolDes based on a previous prototype (*99*) to handle the simultaneous insertion of multiple backbone segments (e.g., within the same functional motif or for different functional motifs) into the target scaffold. This possibility allowed us to design proteins with multiple functional sites as often observed in Nature. Using functional motifs as fixed nuclei to bias the sequence folding and design around it, the final topology allows a proper accommodation and structural stabilization of each motif segment. Moreover, we further implemented the possibility to consider a "binder constraint" to couple the folding process with the desired binding function in the conformational sampling process. In essence, the FunFolDes algorithm takes a protein topology with a low level of local structural similarity to the inserted motifs, and by extracting distance and dihedral constraints from the starting scaffold, it folds a similar topology around the fixed motif and binder. We computationally benchmarked the design of the new binder to BHRF1, named BINDI, in the presence of the target protein (BHRF1). With FunFolDes and the presence of the target protein, we improved the conformation and sequences sampling to favor the design of beneficial residues outside the contact region. Moreover, one helical structure of BINDI was shifted away from the target to ensure the structural compatibility between the motif and target, which experimentally was shown that the motion of the helix was crucial for the binding event.

Secondly, we showed that the FunFolDes algorithm is applicable when using structural templates with distant local RMSDs of the matching regions to epitope structures. We transplanted three structurally distinct neutralizing epitopes of RSVF, representing different levels of structural complexity, site II (helix-loop-helix), site IV (irregular strand), and site 0 (discontinuous epitope with kink helix and loop) into a protein template with unprecedentedly distant local structural similarity (2.8 Å for site 0, 2.3 Å for site II, and 2.2 Å for site IV), and the designed proteins were well folded with high thermostability. Although we could not solve crystal structures for the designed proteins for atomic-level assessment of the accuracy of the design, they performed the designed molecular function to bind the target antibodies, indicating the epitope structure was correctly mimicked for the binding interaction. After iterative rounds of sequence optimizations by the FunFolDes algorithm and directed evolution, we achieved high affinity designs for sites 0 and IV to present as epitope-focused immunogens, thus proving their functionality as precision immunogens in vivo.

Although the FunFolDes algorithm enlarges the protein plasticity to tolerate the insertion of distant structural motifs, improving the use of a top-down approach to design functional proteins; ultimately designable structural templates become scarce for the engineering of high-complexity motifs. The requirements for designable scaffolds are at two levels: 1) local similarity of the scaffold with the functional motif; 2) compatibility of the overall topology. Often it is complex to find design templates that fulfill both criteria.

To address this problem, we therefore describe an alternative design workflow to allow the assembly of *de novo* secondary structures around a given functional constraint, concomitantly achieving the design of function and structure of the *de novo* proteins. This approach, a "bottom-up" strategy, has been conceptually employed in previously reported design successes (*101, 102, 328*). Our bottom-up methodology, however, generalized the design concept for the creation of designable backbones, including but not limited to all-helical topologies. In addition, it provides a novel structural assembly stage by parametrically positioning ideal secondary structural elements and functional motifs in the intended protein topology. The process is initiated using a 2D description of the protein followed by a projection into the 3D space, the spatial constraints are then easily extracted from the putative fold for guidance of constrained folding-design simulations (FunFoldDes). The backbone construction step implemented in our algorithm to some extent may resemble a natural process for the recombination of structural elements on a scale of secondary structures to give rise to a novel functional topology, in which nature usually performs at a higher level to recombine independent folding domains to construct alternative tertiary and quaternary structures. A previously proposed method—SEWING (Structure Extension With Native-substructure Graphs) (*229*)—was proposed to computationally mimic the natural process of assembling protein folds to create new protein structures. Sewing was used to successfully build diverse helical structures not yet observed in natural repertoires.

Moreover, compared to other reported methods, our motif-centric strategy has the advantage of exploring the space of designable structures for a given functional motif. It does so by its generalizable modular assembly of SSEs per user's definition followed by the coupled structure-sequence design sampling. Other strategy by using a blueprint definition starts with a pre-defined topology (*42, 43*), followed by consecutively building residue-by-residue for the creation of an entire structure, and the functional fitness of the target structure is only evaluated once the topology is built. In other words, such strategies limit the robust construction of designable backbones that accommodate non-ideal functional motifs, and the design space for functional purposes is ultimately restricted by the imagination of the protein designer. In contrast, our algorithm allows a coarse-grained exploration of the topological space through parametric arrangements of SSEs with the embedded functional motifs to optimize the context for

functionality. This exploration of the topological space is guided by two complementary evaluations: I) SSEs pair correlative position; II) realistic SSEs connectivity. Without a strict definition in the arrangement of secondary structures for a target fold, it provides a significant potential for unbiased design of an optimal structure accomplishing a defined function. A recent success in using the citizen science software Foldit enables the non-expert creativity to flourish inside a knowledge-dependent arena to generate protein structures obeying the physical principle, leading to a broad exploration of designable proteins (*52*).

We used this approach to design eight distinct topologies suited with irregular and discontinuous antibody binding motifs as a demonstration for its broad applicability. First, we found that subtle variations of the backbone orientation within the same topology can lead to a change in fold stability and the presentation of the site IV neutralization epitope, ultimately affecting the binding affinity of a linear epitope. Remarkably, six out of eight design series were successfully expressed in E. *coli*, and the thermostability of the designs generated by this approach is much superior to previously reported template-based designs. This result highlights that the incorporation of the functional motif in *de novo* design yields the enhancement of protein stability without compromise of functionality. Structurally, three different designed topologies were confirmed at the atomic level showing a high accuracy of the design model. More importantly, the functional motifs were accurately stabilized in the *de novo* proteins, as shown by high-affinity binding to their target antibodies. For a detailed characterization of motif presentation, we've shown S0_2.126 and S4_2.45 bound to broad panels of neutralizing antibodies that engage the same epitopes, indicating that the designed protein also presented an overall molecular context around the functional motif similar to that of the native protein. Notably, we showed that the motif-centric designs upon formulated as a cocktail immunogen yielded neutralizing serum responses in a murine model, which has been a historically difficult model to induce neutralizing antibodies via scaffold-based design approaches (*314*).

Overall, tailoring the scaffold topology directly to the functional motif allowed for the selection of optimal backbones and sequences to achieve stable folds with high binding affinity. Importantly, these designs were directly identified in a single screening round a single library, bypassing further optimization through directed evolution. Also, it provides a way for the rigorous exploration of a structural and sequence space required for the stabilization of a functional motif to enlarges our capabilities to engineer novel functional proteins in the next era.

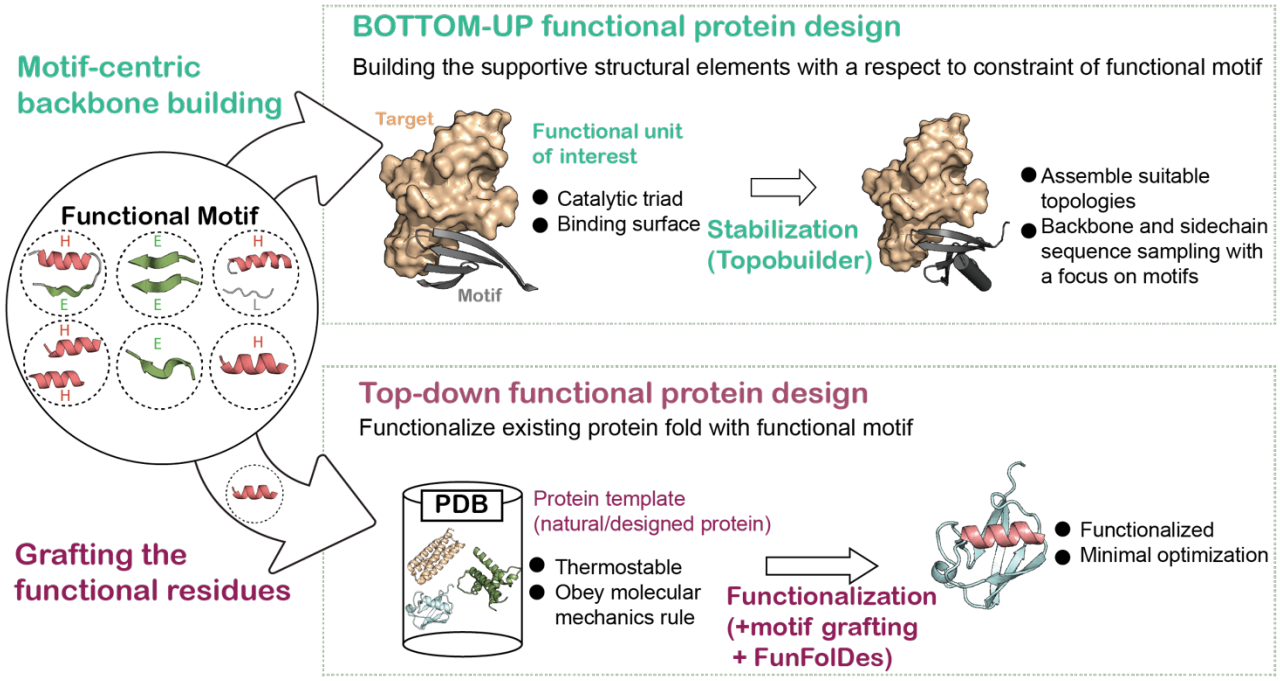


Figure 5.1. Strategies for designing functional *de novo* proteins: Top-down versus bottom-up approaches.

A top-down approach describes a design process that involves grafting the functional motif onto a pre-existing protein template found in the natural protein repertoire. A bottom-up approach allows the tailoring of a *de novo* protein around a given motif of interest. The whole process of backbone building, folding and designing is centered around embedded functional motifs. In Chapter 3, we used both approaches to design the epitope immunogens carrying irregular motifs and demonstrated that the designed proteins generated by the bottom-up approach generally had superior physicochemical and functional features than those generated by the top-down approach. Throughout Chapter 4, we further generalized the bottom-up approach by designing eight distinct protein topologies to incorporate functional motif(s) at the beginning of the design stage.

5.2 Synthetic epitope immunogens consistently elicit neutralizing antibodies *in vivo* and direct the antibody response towards defined antigenic sites in naïve and non-naïve immunity

Recent advances in high-throughput B-cell technologies have yielded comprehensive panels of nAbs for RSV, HIV, influenza and other pathogens. Subsequently, the structural characterization of a subset of those nAbs with their target antigen has elucidated the atomic details of antibody-epitope interaction. Together, those studies providing the crucial molecular understanding have fueled the idea of rational engineered immunogens that present such pathogen's sites for the development of structure-based vaccines (*129*). However, the translation of this structural information into the design of immunogens that elicit targeted antibodies has proven challenging so far, as described in the introduction. Previously, proof-of-principle work using epitope-based immunogens has been shown to hold great promise to elicit precisely controlled neutralizing antibodies in NHPs for RSV (*99*), as well as, to reshape pre-existing immunity towards the increased recognition of subdominant neutralization epitopes (*323*). Since using single epitope-based vaccinations is unlikely to generate the complete antibody response as using the full antigen, and may also exert a selective pressure which the virus may escape easily by mutational drift. Our studies have advanced this strategy using computational design of multiple epitope-focused immunogens presenting distal neutralization epitopes of RSVF and formulated them on the nanoparticle

as a trivalent cocktail vaccine (TriVax). Using Trivax, we unraveled unique features of epitope-focused immunogens for immunization in different immunological backgrounds (*314*).

Naïve immunological background

One of the main goals of vaccine development is to protect naïve individuals (e.g., infants) from infection later in life. Thus, vaccination studies in naïve animal models are an important first step to validate novel immunogens. In Chapter 3, we tested our cocktail formulation of designed immunogens in the NHP model, showing for the first time the ability of cocktail epitope scaffolds to elicit antibody responses on three distinct neutralizing epitopes. Specifically, the TriVax1 cocktail consistently elicited neutralizing serum activity above the protective threshold in six out of seven NHPs after just two immunizations, which is much superior compared to a previous study with a single epitope-focused immunogen where five immunizations were needed to elicit similar titers in half of the immunized animals. While the RSV serum neutralization level of the trivalent cocktail is an order of magnitude lower than prefusion RSVF, these results still provide a proof-of-concept that combining immunogens carrying distinct non-overlapping antigenic sites can prime functional antibodies targeting multiple sites, while yielding a substantially improved immunological outcome than focusing on a single site. Compared to potentially hundreds of epitopes on the viral protein, TriVax1 is only formulated to target three defined epitopes; the obtained titers are remarkable, and further optimization is required to enhance the efficacy and magnitude.

As discussed in Chapter 3, the second generation of Trivax, Trivax2, was developed using our novel motif-centric algorithm TopoBuilder. TopoBuilder allowed us to encode the global and local structural-similarity features of the target epitopes as a constraint to accurately design the epitope scaffold immunogen. Our second-generation immunogens substantially improve the structural mimicry of the epitope, resulting in an increase of binding affinity to multiple site-specific nAbs, compared to Trivax1. In Trivax2 the biophysical stability of immunogens is significantly improved, a feature that has been shown to strongly correlate with the elicitation of functional antibody responses in the DsCav2 version of prefusion RSVF (*166*). All of those improved factors eventually led to Trivax2 eliciting neutralizing serum levels in mice models, which has been reported as a challenging model to induce RSV neutralizing activity through epitope immunogens (*99, 323*). Although we are missing a comparable result in an NHP study for Trivax2, the success in producing the effect in the mouse model represents a substantial step forward for a vaccination development based on a cocktail of epitope-scaffold immunogens.

We proposed an approach that uses cocktail immunogens to prime functional antibodies against predefined pathogenic epitopes. We believe this approach may have broad implications for pathogens that display limited conserved epitopes. For example, one main challenge for influenza vaccine development is that often the immune response is biased towards the first strain encountered by the naive immune system (*329, 330*), or that responses are often directed against immunodominant, strain-specific epitopes rather than broadly neutralizing epitopes found both in the stem and head region of HA. In the case of Influenza, our epitope-focused immunogen strategy could be used to prime the immune system against highly conserved epitopes and prepare the effective antibody repertoire, while subsequent boosts with native-like viral proteins could guide the affinity maturation of B-cells towards generating bnAbs. The applicability of this approach has recently shown promise in the HIV field, where an HIV fusion peptide fused with a nanocarrier was used to prime for fusion-peptide-specific antibodies, followed by boosting immunizations with prefusion-stabilized envelope trimers of HIV, thus resulting in

broader neutralization across clades of HIV strains (*152*). It is likely that, since HIV is a hyper-mutable pathogen, a single epitope is unlikely to provide long-term protection. Our approach that focuses the antibody responses on multiple epitopes by scaffold-based immunogens provides an encouraging route for future HIV research.

Altogether, from the perspective of translational medicine, our trivalent immunogen cocktail may still far from eliciting a complete antibody response (*167*). It may require further optimization in adjuvant formulation, antigenic presentation and delivery route to improve the efficacy of the magnitude of synthetic epitope immunogens. Despite this, this is the first time that epitope-focused immunogens have been developed for three potent neutralization epitopes of the same pathogen, and our *in vivo* studies showed the potential of epitope immunogens to emerge as a promising alternative for vaccine development.

Pre-established immunological background

An important consideration for vaccine development against a seasonal virus, for example influenza, dengue, or RSV, is that the major target populations (adults and the elderly) have encountered multiple infections during their lifetime, and each of these infections likely left a pre-established immunological background in the host. Commonly, such recurrent natural infections create a B-cell immunodominance hierarchy (*108, 330*), where B-cell specific for non-neutralizing or strain-specific epitopes dominate the memory B-cell response, leading to an inefficacy for subsequent vaccinations. In response to this immunodominance hierarchy, we devoted significant efforts to overcome imprinted immunodominance hierarchies. In Chapter 3 we used our cocktail of epitope-immunogens as a boosting reagent to redirect the broad antibody response established by pre-existing immunity to the defined and focused specificities, especially site 0 and site II, in NHPs. While this change in antibody specificities did not enhance the overall serum neutralization activity, neutralization activity was highly dependent on the boosted antigenic sites, accounting for around 60 % of total reactivity. In contrast, the non-boosted control group lost site II specific antibodies over time, and had barely detectable site 0 specificity at the end of the study. Therefore, our work highlighted the potential of Trivalent cocktail as a precision immunogen to reshape the pre-existing antibody responses towards defined epitopes *in vivo*, which is usually unachievable by using a full viral protein.

Given that subdominance is a common immunological phenotype for many neutralization epitopes (*331*), boosting defined antibody specificities *in vivo* remains an outstanding challenge in vaccinology. This is especially true for influenza, where the antibody response predominantly targets variable and strain-specific regions (head domain of HA) and rarely targets broadly protective, conserved epitopes (stem region of HA). Even with seasonal vaccination or repetitive infections as a boosting mechanism, the immune response only sporadically elicits serum response against the conserved region. Thus, reshaping B cell repertories towards conserved regions across all the strains may increase the chances of developing a universal flu vaccine. Our *in vivo* data are an extraordinary finding that enables the development of precision vaccines to steer established immunity towards inducing high levels of antibodies targeting specific neutralization epitopes, which represents a highly relevant scenario for vaccine development against RSV, influenza, and dengue.

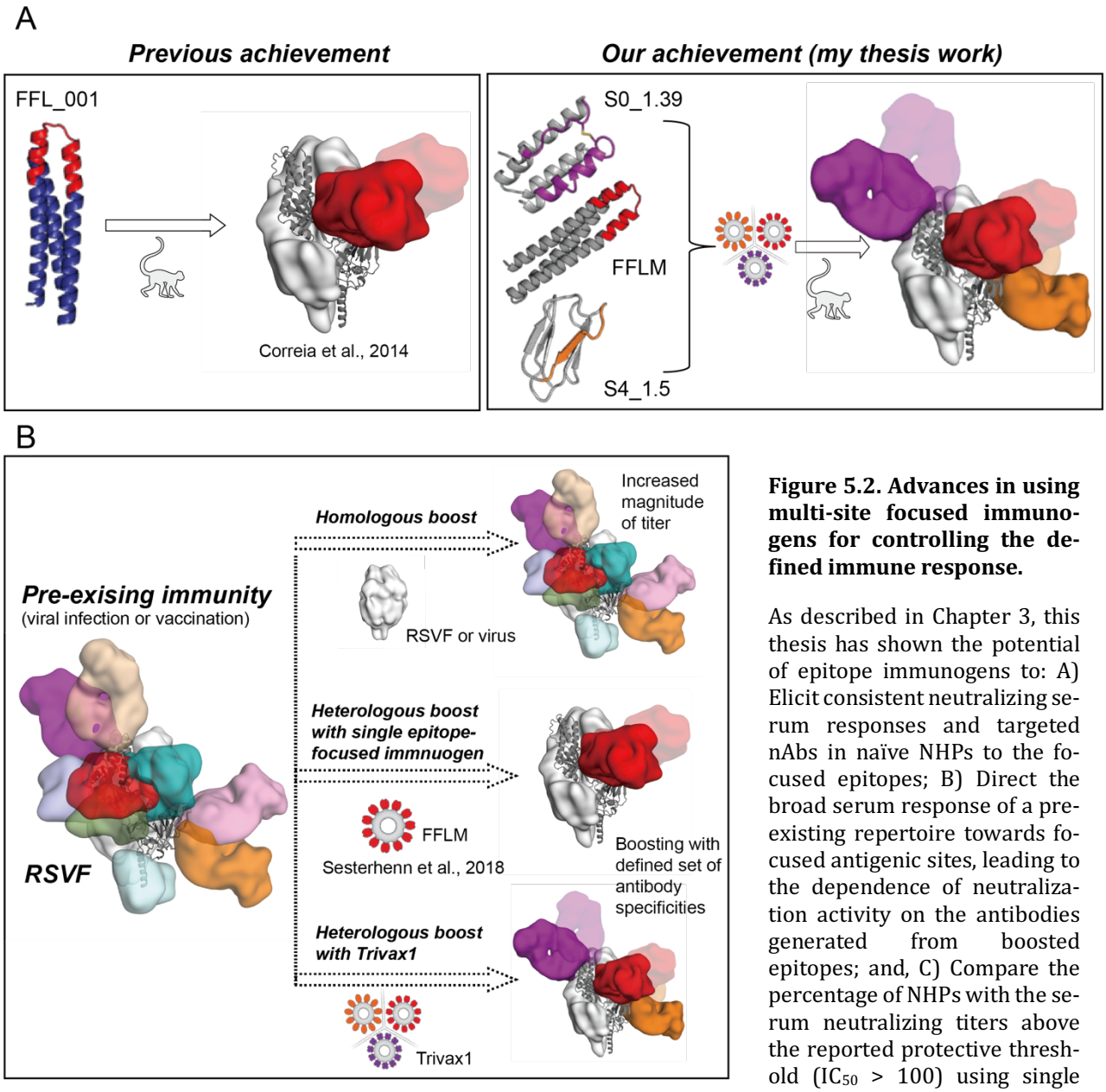


Figure 5.2. Advances in using multi-site focused immunogens for controlling the defined immune response.

As described in Chapter 3, this thesis has shown the potential of epitope immunogens to: A) Elicit consistent neutralizing serum responses and targeted nAbs in naïve NHPs to the focused epitopes; B) Direct the broad serum response of a pre-existing repertoire towards focused antigenic sites, leading to the dependence of neutralization activity on the antibodies generated from boosted epitopes; and, C) Compare the percentage of NHPs with the serum neutralizing titers above the reported protective threshold (IC_{50} > 100) using single epitope immunogens and TriVax.

5.3 Accurate *de novo* design of bi-functional proteins for the fine regulation of synthetic receptors

Advances in synthetic biology over the past several decades have accelerated the ability to engineer existing organisms and enable the modification or creation of biological systems not found in nature. One rapidly growing field in synthetic biology is programming a designed system to treat or diagnose individual patients in a custom-tailored manner, which is also a central dogma of precision medicine (*332*). Considerable progress in creating an orthogonally synthetic receptor including SynNotch, and Chimeric antigen receptors (CAR) to program intracellular responses and rewire gene circuits has significantly improved the treatment of cancer and opened a door for the implementation of synthetic receptors in biological systems (*321*). Those novel systems, however, still rely on using natural ligands to modulate the activation of receptors (*333*), and pose a limitation to precisely control the cellular activity especially in the presence of endogenous ligands. Besides, to achieve a precisely controlled designed system, a synthetic receptor is required to be highly plastic in order to respond to the various strengths of an activator according to different physiological conditions.

Thus, the next generation of synthetic cells were designed to be dynamic and orthogonally regulated by a novel inducer (*332, 334*). Several mechanisms exist in nature to control extracellularly the activation of receptors, including the level of receptor dimerization (*335*), light-induced conformational change (*336*), and others (*337*). Given that the precise spatial configuration of ligand-induced receptor dimerization is a critical determinant for efficient signaling (*338*), recently, designed protein was used to geometrically control the receptor dimerization of EpoR, leading to distinct magnitudes of gene activation (*339*). While, in a real scenario, the designed system needs to integrate multiple extracellular inputs in different doses and trigger a corresponding signaling strength (*340*). Programming the dynamics of a receptor activation in the presence of both agonist and antagonist is a milestone that has remained largely unachieved.

In my thesis, we leverage the ability of the TopoBuilder to assemble secondary structures around multiple functional motifs for designing *de novo* proteins, accommodating two binding sites for synthetic receptors. By using this method, we imposed subtle topological variations between the two binding motifs to maximize the proper functionality for both motifs. Subsequent experimental screening revealed that only one sub-topology was favorable to engage the simultaneous binding to two target counterparts, inferring the value of sampling the backbone orientation at the building stage. The final design, 4H01, shows a helical topology with two binding sites in a defined spatial orientation. This high-functional content encoded in our *de novo* designed proteins is yet another breakthrough for a *de novo* designed protein, where largely the focus has been on empowering the designed proteins with one sole functional motif (*29, 51, 84, 102*). Our computationally designed proteins are thus a step closer to natural proteins that play multiple functional roles in the finely tuned cellular environments.

In the context of synthetic biology, we showcase our *de novo* bi-functional protein in promoting the regulation of the activation of synthetic receptor, as a non-natural inducer of the hetero-dimerization of Generalized Extracellular Molecule Sensor (GEMS) (*322*) and ultimately manipulating signaling *in situ*. Importantly, the GEMS-engineered cells display clear OFF-state baseline and robust ON-state triggered by our bi-functional *de novo* design. The GEMS-engineered cells perceive various ranges of agonistic and antagonistic signals, and output a response reflective of the strength of the stimuli. This advance demonstrates the possibility for coupling synthetic receptors with *de novo* designed proteins, which will likely

be orthogonal to other endogenous protein components, to control the magnitudes of signaling and consequent gene expression.

As a proof-of-principle, we have shown a relevant biological function of the *de novo* ligand to control a synthetic signaling pathway, resulting in transgene expression. Moreover, the designed pair of computationally designed proteins could function as agonists/antagonists to regulate signaling strengths for highly demanding cellular regulation tasks. As such, this *de novo* ligand-receptor pair may be used as a plug-in molecular device for precise productions of transgenes in a therapeutic dose for other cell-based biotechnological applications, which is strongly required in diseases such as diabetes, depression and others (*341-345*).

Projecting beyond the scope we have covered in my thesis, our design pipeline can be a roadmap to design proteins with harmonious structure-function features and applied to design the functionality requiring multi-binding motifs to execute more complex biological activities, such as transcription factors and ribosome binding components. Altogether, we foresee that upcoming protein design challenges (e.g., protein-protein interactions, ligand binding, enzymes) will benefit from this ability of encoding more functional components in small protein domains. As in nature, carrying multiple functional sites is a common feature in proteins that mediate common biological functions. Learning from Nature and to benefit Nature, we believe *de novo* protein design can be expected to tackle intractable biological challenges in the coming age.

References

1. W. Liebermeister *et al.*, Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8488-8493 (2014).
2. W. R. Taylor, A 'periodic table' for protein structures. *Nature* **416**, 657-660 (2002).
3. C. B. Anfinsen, H. A. Scheraga, Experimental and theoretical aspects of protein folding. *Adv Protein Chem* **29**, 205-300 (1975).
4. G. N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95-99 (1963).
5. L. Pauling, R. B. Corey, H. R. Branson, The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37**, 205-211 (1951).
6. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).
7. P. S. Huang, S. E. Boyken, D. Baker, The coming of age of de novo protein design. *Nature* **537**, 320-327 (2016).
8. R. Bonneau, D. Baker, Ab initio protein structure prediction: Progress and prospects. *Annu. Rev. Biophys. Biomolec. Struct.* **30**, 173-189 (2001).
9. D. Baker, A. Sali, Protein structure prediction and structural genomics. *Science* **294**, 93-96 (2001).
10. E. T. Maggio, K. Ramnarayan, Recent developments in computational proteomics. *Drug Discov Today* **6**, 996-1004 (2001).
11. R. Kolodny, L. Pereyaslavets, A. O. Samson, M. Levitt, On the universe of protein folds. *Annu Rev Biophys* **42**, 559-582 (2013).
12. M. Y. Shen, A. Sali, Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507-2524 (2006).
13. J. Desmet, M. Demaeyer, B. Hazes, I. Lasters, The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542 (1992).
14. T. Dandekar, P. Argos, Folding the main-chain of small proteins with the genetic algorithm. *J. Mol. Biol.* **236**, 844-861 (1994).
15. B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 10383-10388 (2000).
16. A. D. Moore, A. K. Bjorklund, D. Ekrnan, E. Bornberg-Bauer, A. Elofsson, Arrangements in the modular evolution of proteins. *Trends Biochem.Sci.* **33**, 444-451 (2008).
17. A. Andreeva, D. Howorth, C. Chothia, E. Kulesha, A. G. Murzin, SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* **42**, D310-D314 (2014).
18. I. Sillitoe *et al.*, New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Res.* **41**, D490-D498 (2013).
19. L. Holm, C. Sander, Mapping the protein universe. *Science* **273**, 595-602 (1996).
20. P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866-876 (2009).
21. M. S. Packer, D. R. Liu, Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379-394 (2015).
22. K. K. Yang, Z. Wu, F. H. Arnold, Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687-694 (2019).
23. C. Pabo, Molecular technology - designing proteins and peptides. *Nature* **301**, 200-200 (1983).
24. W. A. Lim, R. T. Sauer, Alternative packing arrangements in the hydrophobic core of lambda-repressor. *Nature* **339**, 31-36 (1989).
25. G. Dantas, B. Kuhlman, D. Callender, M. Wong, D. Baker, A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449-460 (2003).
26. S. J. Fleishman *et al.*, Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science* **332**, 816-821 (2011).
27. N. P. King *et al.*, Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science* **336**, 1171-1174 (2012).
28. E. Procko *et al.*, A Computationally Designed Inhibitor of an Epstein-Barr Viral Bcl-2 Protein Induces Apoptosis in Infected Cells. *Cell* **157**, 1644-1656 (2014).
29. S. Berger *et al.*, Computationally designed high specificity inhibitors delineate the roles of BCL2 family proteins in cancer. *eLife* **5**, 31 (2016).
30. L. Jiang *et al.*, De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391 (2008).
31. D. Rothlisberger *et al.*, Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-U194 (2008).

32. K. T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225 (1997).
33. B. Gutte, M. Daumigen, E. Wittschieber, Design, synthesis and characterization of a 34-residue polypeptide that interacts with nucleic-acids. *Nature* **281**, 650-655 (1979).
34. K. E. Drexler, Molecular engineering - an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 5275-5278 (1981).
35. R. Moser, R. M. Thomas, B. Gutte, An artificial crystalline ddt-binding polypeptide. *FEBS Lett.* **157**, 247-251 (1983).
36. S. Y. M. Lau, A. K. Taneja, R. S. Hodges, Synthesis of a model protein of defined secondary and quaternary structure - effect of chain-length on the stabilization and formation of 2-stranded alpha-helical coiled-coils. *J. Biol. Chem.* **259**, 3253-3261 (1984).
37. S. P. Ho, W. F. Degrado, Design of a 4-helix bundle protein - synthesis of peptides which self-associate into a helical protein. *J. Am. Chem. Soc.* **109**, 6751-6758 (1987).
38. W. F. Degrado, Design of peptides and proteins. *Adv.Protein Chem.* **39**, 51-124 (1988).
39. W. F. Degrado, Z. R. Wasserman, J. D. Lear, Protein design, a minimalist approach. *Science* **243**, 622-628 (1989).
40. B. I. Dahiyat, S. L. Mayo, De novo protein design: Fully automated sequence selection. *Science* **278**, 82-87 (1997).
41. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, High-resolution protein design with backbone freedom. *Science* **282**, 1462-1467 (1998).
42. B. Kuhlman *et al.*, Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368 (2003).
43. N. Koga *et al.*, Principles for designing ideal protein structures. *Nature* **491**, 222-227 (2012).
44. Y. R. Lin *et al.*, Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5478-E5485 (2015).
45. J. M. Fletcher *et al.*, Self-assembling cages from coiled-coil peptide modules. *Science* **340**, 595-599 (2013).
46. A. R. Thomson *et al.*, Computational design of water-soluble alpha-helical barrels. *Science* **346**, 485-488 (2014).
47. E. Marcos *et al.*, Principles for designing proteins with cavities formed by curved beta sheets. *Science* **355**, 201-206 (2017).
48. P. S. Huang *et al.*, De novo design of a fourfold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12**, 29-+ (2016).
49. E. Marcos *et al.*, De novo design of a non-local beta-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **25**, 1028-+ (2018).
50. S. E. Boyken *et al.*, De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680-687 (2016).
51. J. Y. Dou *et al.*, De novo design of a fluorescence-activating beta-barrel. *Nature* **561**, 485-+ (2018).
52. B. Koepnick *et al.*, De novo protein design by citizen scientists. *Nature* **570**, 390-+ (2019).
53. P. Gainza *et al.*, in *Methods in Protein Design,* A. E. Keating, Ed. (Elsevier Academic Press Inc, San Diego, 2013), vol. 523, pp. 87-107.
54. B. I. Dahiyat, S. L. Mayo, Protein design automation. *Protein Sci.* **5**, 895-903 (1996).
55. B. Kuhlman, P. Bradley, Advances in protein structure prediction and design. *Nat Rev Mol Cell Biol* **20**, 681-697 (2019).
56. K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, J. Meiler, Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You. *Biochemistry* **49**, 2987-2998 (2010).
57. R. F. Alford *et al.*, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031-3048 (2017).
58. Z. X. Li, Y. D. Yang, J. Zhan, L. Dai, Y. Q. Zhou, in *Annual Review of Biophysics, Vol 42,* K. A. Dill, Ed. (Annual Reviews, Palo Alto, 2013), vol. 42, pp. 315-335.
59. R. L. Dunbrack, Rotamer libraries in the 21(st) century. *Curr. Opin. Struct. Biol.* **12**, 431-440 (2002).
60. P. Gainza, H. M. Nisonoff, B. R. Donald, Algorithms for protein design. *Curr. Opin. Struct. Biol.* **39**, 16-26 (2016).
61. P. Bradley, K. M. S. Misura, D. Baker, Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868-1871 (2005).
62. S. H. Gellman, Introduction: Molecular recognition. *Chem. Rev.* **97**, 1231-1232 (1997).
63. M. Babor, H. M. Greenblatt, M. Edelman, V. Sobolev, Flexibility of metal binding sites in proteins on a database scale. *Proteins* **59**, 221-230 (2005).

64. T. Dudev, Y. L. Lin, M. Dudev, C. Lim, First-second shell interactions in metal binding sites in proteins: A PDB survey and DFT/CDM calculations. *J. Am. Chem. Soc.* **125**, 3168-3180 (2003).
65. E. M. Meiering, L. Serrano, A. R. Fersht, Effect of active-site residues in barnase on activity and stability. *J. Mol. Biol.* **225**, 585-589 (1992).
66. R. A. Studer, B. H. Dessailly, C. A. Orengo, Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem. J.* **449**, 581-594 (2013).
67. G. Bhardwaj *et al.*, Accurate de novo design of hyperstable constrained peptides. *Nature* **538**, 329-+ (2016).
68. G. J. Rocklin *et al.*, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168-174 (2017).
69. P. S. Huang *et al.*, RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS One* **6**, 8 (2011).
70. W. M. Dawson, G. G. Rhys, D. N. Woolfson, Towards functional de novo designed proteins. *Curr Opin Chem Biol* **52**, 102-111 (2019).
71. H. K. Privett *et al.*, Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3790-3795 (2012).
72. J. Kaplan, W. F. DeGrado, De novo design of catalytic proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 11566-11570 (2004).
73. F. Richter *et al.*, Computational Design of Catalytic Dyads and Oxyanion Holes for Ester Hydrolysis. *J. Am. Chem. Soc.* **134**, 16197-16206 (2012).
74. A. A. Bogan, K. S. Thorn, Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1-9 (1998).
75. S. J. Fleishman *et al.*, Hotspot-Centric De Novo Design of Protein Binders. *J. Mol. Biol.* **413**, 1047-1062 (2011).
76. D. A. Silva, B. E. Correia, E. Procko, Motif-Driven Design of Protein-Protein Interfaces. *Methods Mol Biol* **1414**, 285-304 (2016).
77. S. Liu *et al.*, Nonnatural protein-protein interaction-pair design by key residues grafting. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5330-5335 (2007).
78. B. E. Correia *et al.*, Computational Design of Epitope-Scaffolds Allows Induction of Antibodies Specific for a Poorly Immunogenic HIV Vaccine Epitope. *Structure* **18**, 1116-1126 (2010).
79. G. Ofek *et al.*, Elicitation of structure-specific antibodies by epitope scaffolds. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 17880-17887 (2010).
80. M. L. Azoitei *et al.*, Computation-Guided Backbone Grafting of a Discontinuous Motif onto a Protein Scaffold. *Science* **334**, 373-376 (2011).
81. M. L. Azoitei *et al.*, Computational Design of High-Affinity Epitope Scaffolds by Backbone Grafting of a Linear Epitope. *J. Mol. Biol.* **415**, 175-192 (2012).
82. E. Procko *et al.*, Computational Design of a Protein-Based Enzyme Inhibitor. *J. Mol. Biol.* **425**, 3563-3575 (2013).
83. U. Scheib, S. Shanmugaratnam, J. A. Farias-Rico, B. Hocker, Change in protein-ligand specificity through binding pocket grafting. *J. Struct. Biol.* **185**, 186-192 (2014).
84. A. Chevalier *et al.*, Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74-+ (2017).
85. C. L. Kingsford, B. Chazelle, M. Singh, Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* **21**, 1028-1036 (2005).
86. J. Bonet *et al.*, Rosetta FunFolDes - A general framework for the computational design of functional proteins. *PLoS Comput. Biol.* **14**, 30 (2018).
87. J. R. Desjarlais, T. M. Handel, Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**, 305-318 (1999).
88. P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber, P. S. Kim, High-resolution protein design with backbone freedom. *Science* **282**, 1462-1467 (1998).
89. G. Dantas *et al.*, High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design. *J. Mol. Biol.* **366**, 1209-1221 (2007).
90. I. Georgiev, B. R. Donald, Dead-end elimination with backbone flexibility. *Bioinformatics* **23**, I185-I194 (2007).
91. J. R. Apgar, S. Hahn, G. Grigoryan, A. E. Keating, Cluster Expansion Models for Flexible-Backbone Protein Energetics. *J. Comput. Chem.* **30**, 2402-2413 (2009).
92. J. J. Havranek, D. Baker, Motif-directed flexible backbone design of functional interactions. *Protein Sci.* **18**, 1293-1305 (2009).
93. D. J. Mandell, T. Kortemme, Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.* **20**, 420-428 (2009).

94. P. M. Murphy, J. M. Bolduc, J. L. Gallaher, B. L. Stoddard, D. Baker, Alteration of enzyme specificity by computational loop remodeling and design. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9215-9220 (2009).
95. J. D. Bloom *et al.*, Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biol.* **5**, 20 (2007).
96. B. E. Correia *et al.*, Computational Protein Design Using Flexible Backbone Remodeling and Resurfacing: Case Studies in Structure-Based Antigen Design. *J. Mol. Biol.* **405**, 284-297 (2011).
97. B. M. Beadle, B. K. Shoichet, Structural bases of stability-function tradeoffs in enzymes. *J. Mol. Biol.* **321**, 285-296 (2002).
98. N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **4**, 7 (2008).
99. B. E. Correia *et al.*, Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201-206 (2014).
100. G. M. Bender *et al.*, De novo design of a single-chain diphenylporphyrin metalloprotein. *J Am Chem Soc* **129**, 10732-10740 (2007).
101. A. J. Burton, A. R. Thomson, W. M. Dawson, R. L. Brady, D. N. Woolfson, Installing hydrolytic activity into a completely de novo protein framework. *Nat. Chem.* **8**, 837-844 (2016).
102. D. A. Silva *et al.*, De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186-+ (2019).
103. S. A. Plotkin, Correlates of vaccine-induced immunity. *Clin. Infect. Dis.* **47**, 401-409 (2008).
104. S. A. Plotkin, Correlates of Protection Induced by Vaccination. *Clin. Vaccine Immunol.* **17**, 1055-1065 (2010).
105. D. R. Burton, Antibodies, viruses and vaccines. *Nat. Rev. Immunol.* **2**, 706-713 (2002).
106. D. J. DiLillo, P. Palese, P. C. Wilson, J. V. Ravetch, Broadly neutralizing anti-influenza antibodies require Fc receptor engagement for in vivo protection. *J. Clin. Invest.* **126**, 605-610 (2016).
107. A. B. Ward, I. A. Wilson, The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunol. Rev.* **275**, 21-32 (2017).
108. D. Angeletti *et al.*, Defining B cell immunodominance to viruses. *Nat. Immunol.* **18**, 456-+ (2017).
109. D. Eggink, P. H. Goff, P. Palese, Guiding the Immune Response against Influenza Virus Hemagglutinin toward the Conserved Stalk Domain by Hyperglycosylation of the Globular Head Domain. *J. Virol.* **88**, 699-704 (2014).
110. R. Rappuoli, M. J. Bottomley, U. D'Oro, O. Finco, E. De Gregorio, Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *J. Exp. Med.* **213**, 469-481 (2016).
111. R. M. F. Cardoso *et al.*, Broadly neutralizing anti-HIV antibody 4E10 recognizes a helical conformation of a highly conserved fusion-associated motif in gp41. *Immunity* **22**, 163-173 (2005).
112. J. P. Julien *et al.*, Broadly Neutralizing Antibody PGT121 Allosterically Modulates CD4 Binding via Recognition of the HIV-1 gp120 V3 Base and Multiple Surrounding Glycans. *PLoS Pathog.* **9**, 15 (2013).
113. G. Ofek *et al.*, Structure and mechanistic analysis of the anti-human immunodeficiency virus type 1 antibody 2F5 in complex with its gp41 epitope. *J. Virol.* **78**, 10724-10737 (2004).
114. D. Corti *et al.*, Cross-neutralization of four paramyxoviruses by a human monoclonal antibody. *Nature* **501**, 439-+ (2013).
115. S. O. Fedechkin, N. L. George, J. T. Wolff, L. M. Kauvar, R. M. DuBois, Structures of respiratory syncytial virus G antigen bound to broadly neutralizing antibodies. *Sci. Immunol.* **3**, 7 (2018).
116. J. S. McLellan *et al.*, Structural basis of respiratory syncytial virus neutralization by motavizumab. *Nat. Struct. Mol. Biol.* **17**, 248-250 (2010).
117. J. S. McLellan *et al.*, Structure of RSV Fusion Glycoprotein Trimer Bound to a Prefusion-Specific Neutralizing Antibody. *Science* **340**, 1113-1117 (2013).
118. J. S. McLellan, Y. P. Yang, B. S. Graham, P. D. Kwong, Structure of Respiratory Syncytial Virus Fusion Glycoprotein in the Postfusion Conformation Reveals Preservation of Neutralizing Epitopes. *J. Virol.* **85**, 7788-7796 (2011).
119. J. J. Mousa, N. Kose, P. Matta, P. Gilchuk, J. E. Crowe, A novel pre-fusion conformation-specific neutralizing epitope on the respiratory syncytial virus fusion protein. *Nat. Microbiol* **2**, 8 (2017).
120. K. Tharakaraman, V. Subramanian, D. Cain, V. Sasisekharan, R. Sasisekharan, Broadly Neutralizing Influenza Hemagglutinin Stem-Specific Antibody CR8020 Targets Residues that Are Prone to Escape due to Host Selection Pressure. *Cell Host Microbe* **15**, 644-651 (2014).
121. J. R. R. Whittle *et al.*, Broadly neutralizing human antibody that recognizes the receptor-binding pocket of influenza virus hemagglutinin. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 14216-14221 (2011).
122. P. S. Lee, I. A. Wilson, in *Influenza Pathogenesis and Control - Vol Ii,* M. B. A. Oldstone, R. W. Compans, Eds. (Springer-Verlag Berlin, Berlin, 2015), vol. 386, pp. 323-341.
123. D. Corti *et al.*, A Neutralizing Antibody Selected from Plasma Cells That Binds to Group 1 and Group 2 Influenza A Hemagglutinins. *Science* **333**, 850-856 (2011).

124. C. Dreyfus *et al.*, Highly Conserved Protective Epitopes on Influenza B Viruses. *Science* **337**, 1343-1348 (2012).
125. D. C. Ekiert *et al.*, Antibody Recognition of a Highly Conserved Influenza Virus Epitope. *Science* **324**, 246-251 (2009).
126. D. C. Ekiert *et al.*, A Highly Conserved Neutralizing Epitope on Group 2 Influenza A Viruses. *Science* **333**, 843-850 (2011).
127. S. Friedensohn, T. A. Khan, S. T. Reddy, Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires. *Trends Biotechnol.* **35**, 203-214 (2017).
128. B. S. Graham, M. S. A. Gilman, J. S. McLellan, in *Annual Review of Medicine, Vol 70,* M. E. Klotman, Ed. (Annual Reviews, Palo Alto, 2019), vol. 70, pp. 91-104.
129. F. Sesterhenn, J. Bonet, B. E. Correia, Structure-based immunogen design - leading the way to the new age of precision vaccines. *Curr. Opin. Struct. Biol.* **51**, 163-169 (2018).
130. V. V. A. Mallajosyula *et al.*, Influenza hemagglutinin stem-fragment immunogen elicits broadly neutralizing antibodies and confers heterologous protection. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2514-E2523 (2014).
131. H. M. Yassine *et al.*, Hemagglutinin-stem nanoparticles generate heterosubtypic influenza protection. *Nat. Med.* **21**, 1065-+ (2015).
132. J. C. Boyington *et al.*, Structure-Based Design of Head-Only Fusion Glycoprotein Immunogens for Respiratory Syncytial Virus. *PLoS One* **11**, 21 (2016).
133. A. Impagliazzo *et al.*, A stable trimeric influenza hemagglutinin stem as a broadly protective immunogen. *Science* **349**, 1301-1306 (2015).
134. R. P. Ringe *et al.*, Reducing V3 Antigenicity and Immunogenicity on Soluble, Native-Like HIV-1 Env SOSIP Trimers. *J. Virol.* **91**, 17 (2017).
135. D. W. Kulp *et al.*, Structure-based design of native-like HIV-1 envelope trimers to silence non-neutralizing epitopes and eliminate CD4 binding. *Nat. Commun.* **8**, 14 (2017).
136. S. C. Harrison, Viral membrane fusion. *Nat. Struct. Mol. Biol.* **15**, 690-698 (2008).
137. R. A. Lamb, T. S. Jardetzky, Structural basis of viral invasion: lessons from paramyxovirus F. *Curr. Opin. Struct. Biol.* **17**, 427-436 (2007).
138. M. Magro *et al.*, Neutralizing antibodies against the preactive form of respiratory syncytial virus fusion protein offer unique possibilities for clinical intervention. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3089-3094 (2012).
139. J. O. Ngwuta *et al.*, Prefusion F-specific antibodies determine the magnitude of RSV neutralizing activity in human sera. *Sci. Transl. Med.* **7**, 9 (2015).
140. T. J. Ruckwardt, K. M. Morabito, B. S. Graham, Immunological Lessons from Respiratory Syncytial Virus Vaccine Development. *Immunity* **51**, 429-442 (2019).
141. S. W. de Taeye *et al.*, Stabilization of the gp120 V3 loop through hydrophobic interactions reduces the immunodominant V3-directed non-neutralizing response to HIV-1 envelope trimers. *J. Biol. Chem.* **293**, 1688-1701 (2018).
142. R. W. Sanders *et al.*, Stabilization of the soluble, cleaved, trimeric form of the envelope glycoprotein complex of human immunodeficiency virus type 1. *J. Virol.* **76**, 8875-8889 (2002).
143. R. W. Sanders, J. P. Moore, Native-like Env trimers as a platform for HIV-1 vaccine design. *Immunol. Rev.* **275**, 161-182 (2017).
144. M. B. Battles *et al.*, Structure and immunogenicity of pre-fusion-stabilized human metapneumovirus F glycoprotein. *Nat. Commun.* **8**, 11 (2017).
145. J. Pallesen *et al.*, Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7348-E7357 (2017).
146. H. Nair *et al.*, Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet* **375**, 1545-1555 (2010).
147. J. Jardine *et al.*, Rational HIV Immunogen Design to Target Specific Germline B Cell Receptors. *Science* **340**, 711-716 (2013).
148. J. G. Jardine *et al.*, HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogene. *Science* **351**, 1458-1463 (2016).
149. J. G. Jardine *et al.*, Priming a broadly neutralizing antibody response to HIV-1 using a germline-targeting immunogen. *Science* **349**, 156-161 (2015).
150. M. Medina-Ramirez *et al.*, Design and crystal structure of a native-like HIV-1 envelope trimer that engages multiple broadly neutralizing antibody precursors in vivo. *J. Exp. Med.* **214**, 2573-2590 (2017).
151. J. M. Steichen *et al.*, HIV Vaccine Design to Target Germline Precursors of Glycan-Dependent Broadly Neutralizing Antibodies. *Immunity* **45**, 483-496 (2016).
152. K. Xu *et al.*, Epitope-based vaccine design yields fusion peptide-directed antibodies that neutralize diverse strains of HIV-1. *Nat. Med.* **24**, 857-+ (2018).

153. N. Jaberolansar *et al.*, Induction of high titred, non-neutralising antibodies by self-adjuvanting peptide epitopes derived from the respiratory syncytial virus fusion protein. *Sci Rep* **7**, 11 (2017).
154. P. L. Herve *et al.*, RSV N-nanorings fused to palivizumab-targeted neutralizing epitope as a nanoparticle RSV vaccine. *Nanomed.-Nanotechnol. Biol. Med.* **13**, 411-420 (2017).
155. J. Guenaga *et al.*, Heterologous Epitope-Scaffold Prime:Boosting Immuno-Focuses B Cell Responses to the HIV-1 gp41 2F5 Neutralization Determinant. *PLoS One* **6**, 12 (2011).
156. J. S. McLellan *et al.*, Design and Characterization of Epitope-Scaffold Immunogens That Present the Motavizumab Epitope from Respiratory Syncytial Virus. *J. Mol. Biol.* **409**, 853-866 (2011).
157. D. W. Kulp, W. R. Schief, Advances in structure-based vaccine design. *Curr. Opin. Virol.* **3**, 322-331 (2013).
158. T. F. Feltes *et al.*, Palivizumab prophylaxis reduces hospitalization due to respiratory syncytial virus in young children with hemodynamically significant congenital heart disease. *J. Pediatr.* **143**, 532-540 (2003).
159. D. Null *et al.*, Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants. *Pediatrics* **102**, 531-537 (1998).
160. B. Rima *et al.*, ICTV Virus Taxonomy Profile: Pneumoviridae. *J. Gen. Virol.* **98**, 2912-2913 (2017).
161. M. B. Battles, J. S. McLellan, Respiratory syncytial virus entry and how to block it. *Nat. Rev. Microbiol.* **17**, 233-245 (2019).
162. J. Chin, R. L. Magoffin, L. A. Shearer, J. H. Schieble, E. H. Lennette, Field evaluation of a respiratory syncytial virus vaccine and a trivalent parainfluenza virus vaccine in a pediatric population. *Am. J. Epidemiol.* **89**, 449-+ (1969).
163. B. R. Murphy, E. E. Walsh, Formalin-inactivated respiratory syncytial virus-vaccine induces antibodies to the fusion glycoprotein that are deficient in fusion-inhibiting activity. *J. Clin. Microbiol.* **26**, 1595-1597 (1988).
164. I. Widjaja *et al.*, Characterization of Epitope-Specific Anti-Respiratory Syncytial Virus (Anti-RSV) Antibody Responses after Natural Infection and after Vaccination with Formalin-Inactivated RSV. *J. Virol.* **90**, 5965-5977 (2016).
165. J. Loebbermann, L. Durant, H. Thornton, C. Johansson, P. J. Openshaw, Defective immunoregulation in RSV vaccine-augmented viral lung disease restored by selective chemoattraction of regulatory T cells. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2987-2992 (2013).
166. J. S. McLellan *et al.*, Structure-Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus. *Science* **342**, 592-598 (2013).
167. M. C. Crank *et al.*, A proof of concept for structure-based vaccine design targeting RSV in humans. *Science* **365**, 505-+ (2019).
168. H. G. Jones *et al.*, Structural basis for recognition of the central conserved region of RSV G by neutralizing human antibodies. *PLoS Pathog.* **14**, 23 (2018).
169. J. S. McLellan *et al.*, Structure of a Major Antigenic Site on the Respiratory Syncytial Virus Fusion Glycoprotein in Complex with Neutralizing Antibody 101F. *J. Virol.* **84**, 12236-12244 (2010).
170. E. Goodwin *et al.*, Infants Infected with Respiratory Syncytial Virus Generate Potent Neutralizing Antibodies that Lack Somatic Hypermutation. *Immunity* **48**, 339-+ (2018).
171. V. Mas *et al.*, Engineering, Structure and Immunogenicity of the Human Metapneumovirus F Protein in the Postfusion Conformation. *PLoS Pathog.* **12**, 21 (2016).
172. X. W. Wen *et al.*, Structural basis for antibody cross-neutralization of respiratory syncytial virus and human metapneumovirus. *Nat. Microbiol* **2**, 7 (2017).
173. J. J. Mousa *et al.*, Human antibody recognition of antigenic site IV on Pneumovirus fusion proteins. *PLoS Pathog.* **14**, 19 (2018).
174. A. M. Tang *et al.*, A potent broadly neutralizing human RSV antibody targets conserved site IV of the fusion glycoprotein. *Nat. Commun.* **10**, 13 (2019).
175. L. L. Cross *et al.*, Towards designer organelles by subverting the peroxisomal import pathway. *Nat Commun* **8**, 454 (2017).
176. N. H. Joh *et al.*, De novo design of a transmembrane Zn(2)(+)-transporting four-helix bundle. *Science* **346**, 1520-1524 (2014).
177. B. E. Correia *et al.*, Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope. *Structure* **18**, 1116-1126 (2010).
178. B. E. Correia *et al.*, Proof of principle for epitope-focused vaccine design. *Nature* **507**, 201-206 (2014).
179. M. R. Kulkarni *et al.*, Structural and biophysical analysis of sero-specific immune responses using epitope grafted Dengue ED3 mutants. *Biochim Biophys Acta* **1854**, 1438-1443 (2015).
180. I. Coluzza, Computational protein design: a review. *J Phys Condens Matter* **29**, 143001 (2017).
181. N. Koga *et al.*, Principles for designing ideal protein structures. *Nature* **491**, 222-227 (2012).

182. E. Marcos *et al.*, Principles for designing proteins with cavities formed by curved beta sheets. *Science* **355**, 201-206 (2017).
183. B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci U S A* **97**, 10383-10388 (2000).
184. G. S. Murphy *et al.*, Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic core. *Structure* **20**, 1086-1096 (2012).
185. R. B. Hill, D. P. Raleigh, A. Lombardi, W. F. DeGrado, De novo design of helical bundles as models for understanding protein folding and function. *Acc Chem Res* **33**, 745-754 (2000).
186. D. N. Woolfson *et al.*, De novo protein design: how do we expand into the universe of possible protein structures? *Curr Opin Struct Biol* **33**, 16-26 (2015).
187. B. Kuhlman *et al.*, Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368 (2003).
188. A. G. Street, S. L. Mayo, Computational protein design. *Structure* **7**, R105-109 (1999).
189. F. Yu *et al.*, Protein design: toward functional metalloenzymes. *Chem Rev* **114**, 3495-3578 (2014).
190. G. Guntas, C. Purbeck, B. Kuhlman, Engineering a protein-protein interface using a computationally designed library. *Proc Natl Acad Sci U S A* **107**, 19296-19301 (2010).
191. L. Jiang *et al.*, De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391 (2008).
192. H. Kries, R. Blomberg, D. Hilvert, De novo enzymes by computational design. *Curr Opin Chem Biol* **17**, 221-228 (2013).
193. M. L. Azoitei *et al.*, Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold. *Science* **334**, 373-376 (2011).
194. B. E. Correia *et al.*, Computational protein design using flexible backbone remodeling and resurfacing: case studies in structure-based antigen design. *J Mol Biol* **405**, 284-297 (2011).
195. E. Procko *et al.*, A computationally designed inhibitor of an Epstein-Barr viral Bcl-2 protein induces apoptosis in infected cells. *Cell* **157**, 1644-1656 (2014).
196. I. F. T. Viana *et al.*, De novo design of immunoreactive conformation-specific HIV-1 epitopes based on Top7 scaffold. *Rsc Adv* **3**, 11790-11800 (2013).
197. E. M. Strauch *et al.*, Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. *Nat Biotechnol* **35**, 667-671 (2017).
198. E. M. Strauch, S. J. Fleishman, D. Baker, Computational design of a pH-sensitive IgG binding protein. *Proc Natl Acad Sci U S A* **111**, 675-680 (2014).
199. J. W. Chin, A. Schepartz, Design and Evolution of a Miniature Bcl-2 Binding Protein. *Angew Chem Int Ed Engl* **40**, 3806-3809 (2001).
200. H. Domingues, D. Cregut, W. Sebald, H. Oschkinat, L. Serrano, Rational design of a GCN4-derived mimetic of interleukin-4. *Nat Struct Biol* **6**, 652-656 (1999).
201. C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, Protein structure prediction using Rosetta. *Methods Enzymol* **383**, 66-93 (2004).
202. R. Aragues, A. Sali, J. Bonet, M. A. Marti-Renom, B. Oliva, Characterization of protein hubs by inferring interacting motifs from protein interactions. *PLoS Comput Biol* **3**, 1761-1771 (2007).
203. F. Richter, A. Leaver-Fay, S. D. Khare, S. Bjelic, D. Baker, De novo enzyme design using Rosetta3. *PLoS One* **6**, e19230 (2011).
204. J. A. Marie-France de1 Guercio, Ralph T. Kubo,Thomas Arrhenius, Ajesh Maewal, Ettore Appellat, Stephen L. HoffmanS, Trevor Jonest, Danila Valmori , Kazuyasu Sakaguchit, Howard M. Grey and Alessandro Sette, Potent immunogenic short linear peptide constructs composed of B cell epitopes and Pan DR T Helper Epitopes (PADRE) for antibody responses in vivo. *Vaccine* **15**, 441-448 (1997).
205. J. Garcia-Garcia *et al.*, Networks of ProteinProtein Interactions: From Uncertainty to Molecular Details. *Mol Inform* **31**, 342-362 (2012).
206. C. A. Rohl, D. Baker, De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. *J Am Chem Soc* **124**, 2723-2729 (2002).
207. P. M. Bowers, C. E. Strauss, D. Baker, De novo protein structure determination using sparse NMR data. *J Biomol NMR* **18**, 311-318 (2000).
208. D. Gront, D. W. Kulp, R. M. Vernon, C. E. Strauss, D. Baker, Generalized fragment picking in Rosetta: design, protocols and applications. *PLoS One* **6**, e23294 (2011).
209. S. J. Fleishman *et al.*, RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PLoS One* **6**, e20161 (2011).
210. J. A. Fallas *et al.*, Computational design of self-assembling cyclic protein homo-oligomers. *Nat Chem* **9**, 353-360 (2017).
211. T. Vreven *et al.*, Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J Mol Biol* **427**, 3031-3041 (2015).

212. N. L. Dawson *et al.*, CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res* **45**, D289-D295 (2017).
213. D. E. Kim, B. Blum, P. Bradley, D. Baker, Sampling bottlenecks in de novo protein structure prediction. *J Mol Biol* **393**, 249-260 (2009).
214. S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
215. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
216. G. Schreiber, S. J. Fleishman, Computational design of protein-protein interactions. *Curr Opin Struct Biol* **23**, 903-910 (2013).
217. E. H. Kellogg, A. Leaver-Fay, D. Baker, Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830-838 (2011).
218. J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **103**, 5869-5874 (2006).
219. N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, How protein stability and new functions trade off. *PLoS Comput Biol* **4**, e1000002 (2008).
220. J. S. McLellan *et al.*, Structural basis of respiratory syncytial virus neutralization by motavizumab. *Nat Struct Mol Biol* **17**, 248-250 (2010).
221. J. Zhou, G. Grigoryan, Rapid search for tertiary fragments reveals protein sequence-structure relationships. *Protein Sci* **24**, 508-524 (2015).
222. A. Lartigue *et al.*, X-ray structure and ligand binding study of a moth chemosensory protein. *J Biol Chem* **277**, 32094-32098 (2002).
223. M. G. Joyce *et al.*, Iterative structure-based improvement of a fusion-glycoprotein vaccine against RSV. *Nature Structural & Molecular Biology* **23**, 811-820 (2016).
224. J. S. McLellan *et al.*, Structure of a major antigenic site on the respiratory syncytial virus fusion glycoprotein in complex with neutralizing antibody 101F. *J Virol* **84**, 12236-12244 (2010).
225. C. B. Boschek *et al.*, Engineering an ultra-stable affinity reagent based on Top7. *Protein Eng Des Sel* **22**, 325-332 (2009).
226. T. A. Soares, C. B. Boschek, D. Apiyo, C. Baird, T. P. Straatsma, Molecular basis of the structural stability of a Top7-based scaffold at extreme pH and temperature conditions. *J Mol Graph Model* **28**, 755-765 (2010).
227. M. L. Azoitei *et al.*, Computational design of high-affinity epitope scaffolds by backbone grafting of a linear epitope. *J Mol Biol* **415**, 175-192 (2012).
228. S. L. Guffy, F. D. Teets, M. I. Langlois, B. Kuhlman, Protocols for Requirement-Driven Protein Design in the Rosetta Modeling Program. *J Chem Inf Model* **58**, 895-901 (2018).
229. T. M. Jacobs *et al.*, Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687-690 (2016).
230. A. Chevalier *et al.*, Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74-79 (2017).
231. P. W. Rose *et al.*, The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* **45**, D271-D281 (2017).
232. K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* **Suppl 3**, 171-176 (1999).
233. C. Wang, P. Bradley, D. Baker, Protein-protein docking with backbone flexibility. *J Mol Biol* **373**, 503-519 (2007).
234. X. Hu, H. Wang, H. Ke, B. Kuhlman, High-resolution design of a protein loop. *Proc Natl Acad Sci U S A* **104**, 17668-17673 (2007).
235. M. D. Tyka *et al.*, Alternate states of proteins revealed by detailed energy landscape mapping. *J Mol Biol* **405**, 607-618 (2011).
236. B. Kuhlman, D. Baker, Exploring folding free energy landscapes using computational protein design. *Curr Opin Struct Biol* **14**, 89-95 (2004).
237. R. F. Alford *et al.*, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J Chem Theory Comput* **13**, 3031-3048 (2017).
238. D. T. Jones, Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202 (1999).
239. J. Bonet, Z. Harteveld, F. Sesterhenn, A. Scheck, B. E. Correia, rstoolbox: management and analysis of computationally designed structural ensembles. *bioRxiv*, (2018).
240. M. Kvansakul *et al.*, Structural basis for apoptosis inhibition by Epstein-Barr virus BHRF1. *PLoS Pathog* **6**, e1001236 (2010).
241. P. S. Huang *et al.*, RosettaRemodel: a generalized framework for flexible backbone protein design. *PLoS One* **6**, e24109 (2011).

242. R. Rappuoli, C. W. Mandl, S. Black, E. De Gregorio, Vaccines for the twenty-first century society. *Nat. Rev. Immunol.* **11**, 865-872 (2011).
243. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637 (1983).
244. P. S. Huang *et al.*, De novo design of a fourfold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* **12**, 29-34 (2016).
245. M. Mravic *et al.*, Packing of apolar side chains enables accurate design of highly stable membrane proteins. *Science* **363**, 1418-1423 (2019).
246. S. Berger *et al.*, Computationally designed high specificity inhibitors delineate the roles of BCL2 family proteins in cancer. *Elife* **5**, (2016).
247. S. Jones, J. M. Thornton, Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13-20 (1996).
248. N. D. Rubinstein *et al.*, Computational characterization of B-cell epitopes. *Mol Immunol* **45**, 3477-3489 (2008).
249. G. L. Holliday, J. D. Fischer, J. B. Mitchell, J. M. Thornton, Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *FEBS J* **278**, 3835-3845 (2011).
250. R. Rappuoli, M. J. Bottomley, U. D'Oro, O. Finco, E. De Gregorio, Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *J Exp Med* **213**, 469-481 (2016).
251. J. S. McLellan, Neutralizing epitopes on the respiratory syncytial virus fusion glycoprotein. *Curr Opin Virol* **11**, 70-75 (2015).
252. N. S. Laursen, I. A. Wilson, Broadly neutralizing antibodies against influenza viruses. *Antiviral Res* **98**, 476-483 (2013).
253. D. Sok, D. R. Burton, Recent progress in broadly neutralizing antibodies to HIV. *Nat Immunol* **19**, 1179-1188 (2018).
254. F. Sesterhenn, J. Bonet, B. E. Correia, Structure-based immunogen design-leading the way to the new age of precision vaccines. *Curr Opin Struct Biol* **51**, 163-169 (2018).
255. J. S. McLellan *et al.*, Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody. *Science* **340**, 1113-1117 (2013).
256. F. Sesterhenn *et al.*, Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen. *PLoS Biol* **17**, e3000164 (2019).
257. M. C. Jason S. McLellan, M. Gordon Joyce, Mallika Sastry,, Y. Y. Guillaume B. E. Stewart-Jones, Baoshan Zhang, Lei Chen, A. Z. Sanjay Srivatsan, Tongqing Zhou, Kevin W. Graepel, Azad Kumar, Syed Moin, Jeffrey C. Boyington, Gwo-Yu Chuang, Cinque Soto, Ulrich Baxa, Arjen Q. Bakker, Hergen Spits, Tim Beaumont, Zizheng Zheng, Ningshao Xia, Sung-Youl Ko, S. R. John-Paul Todd, Barney S. Graham, Peter D. Kwong, Structure-Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus. *Science* **342**, (2013).
258. D. Tian *et al.*, Structural basis of respiratory syncytial virus subtype-dependent neutralization by an antibody targeting the fusion glycoprotein. *Nat Commun* **8**, 1877 (2017).
259. J. S. McLellan *et al.*, Design and characterization of epitope-scaffold immunogens that present the motavizumab epitope from respiratory syncytial virus. *J Mol Biol* **409**, 853-866 (2011).
260. J. Bonet *et al.*, Rosetta FunFolDes - A general framework for the computational design of functional proteins. *PLoS Comput Biol* **14**, e1006623 (2018).
261. T. A. Whitehead *et al.*, Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat Biotechnol* **30**, 543-548 (2012).
262. T. J. Brunette *et al.*, Exploring the repeat protein universe through computational protein design. *Nature* **528**, 580-584 (2015).
263. J. S. McLellan *et al.*, Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. *Science* **342**, 592-598 (2013).
264. P. Kristensen, G. Winter, Proteolytic selection for protein folding using filamentous bacteriophages. *Fold Des* **3**, 321-328 (1998).
265. M. D. Finucane, M. Tuna, J. H. Lees, D. N. Woolfson, Core-directed protein design. I. An experimental method for selecting stable proteins from combinatorial libraries. *Biochemistry* **38**, 11604-11612 (1999).
266. A. M. Watkins, P. S. Arora, Anatomy of beta-strands at protein-protein interfaces. *ACS Chem Biol* **9**, 1747-1754 (2014).
267. S. F. Andrews *et al.*, Immune history profoundly affects broadly protective B cell responses to influenza. *Sci Transl Med* **7**, 316ra192 (2015).
268. G. Barba-Spaeth *et al.*, Structural basis of potent Zika-dengue virus antibody cross-neutralization. *Nature* **536**, 48-+ (2016).

269. S. F. Andrews *et al.*, High preexisting serological antibody levels correlate with diversification of the influenza vaccine response. *J Virol* **89**, 3308-3317 (2015).
270. D. Angeletti *et al.*, Defining B cell immunodominance to viruses. *Nat Immunol* **18**, 456-463 (2017).
271. D. Corti *et al.*, A neutralizing antibody selected from plasma cells that binds to group 1 and group 2 influenza A hemagglutinins. *Science* **333**, 850-856 (2011).
272. J. Lee *et al.*, Persistent Antibody Clonotypes Dominate the Serum Response to Influenza over Multiple Years and Repeated Vaccinations. *Cell Host Microbe* **25**, 367-376 e365 (2019).
273. H. F. Moffett *et al.*, B cells engineered to express pathogen-specific antibodies protect against infection. *Sci Immunol* **4**, (2019).
274. P. Conway, M. D. Tyka, F. DiMaio, D. E. Konerding, D. Baker, Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci* **23**, 47-55 (2014).
275. V. Mas *et al.*, Engineering, Structure and Immunogenicity of the Human Metapneumovirus F Protein in the Postfusion Conformation. *PLoS Pathog* **12**, e1005859 (2016).
276. M. C. Joan O. Ngwuta, 1* Kayvon Modjarrad,1,2* M. Gordon Joyce,1* Masaru Kanekiyo,1* Azad Kumar,1 Hadi M. Yassine,1 Syed M. Moin,1 April M. Killikelly,1 Gwo-Yu Chuang,1 Aliaksandr Druz,1 Ivelin S. Georgiev,1 Emily J. Rundlet,1 Mallika Sastry,1 Guillaume B. E. Stewart-Jones,1, B. Z. Yongping Yang, 1 Martha C. Nason,1 Cristina Capella,3 Mark E. Peeples,3, J. S. M. Julie E. Ledgerwood, 1,4 Peter D. Kwong,1 Barney S. Graham1†, Prefusion F–specific antibodies determine the magnitude of RSV neutralizing activity in human sera. *ScienceTranslational Medicine* **7**, (2015).
277. D. W. Kulp *et al.*, Structure-based design of native-like HIV-1 envelope trimers to silence non-neutralizing epitopes and eliminate CD4 binding. *Nat Commun* **8**, 1655 (2017).
278. G. Chao *et al.*, Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* **1**, 755-768 (2006).
279. S. D. Khare *et al.*, Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat. Chem. Biol.* **8**, 294-300 (2012).
280. B. Briney *et al.*, Tailored Immunogens Direct Affinity Maturation toward HIV Neutralizing Antibodies. *Cell* **166**, 1459-1470 e1411 (2016).
281. N. Castagne *et al.*, Biochemical characterization of the respiratory syncytial virus P-P and P-N protein complexes and localization of the P protein oligomerization domain. *J Gen Virol* **85**, 1643-1653 (2004).
282. M. G. Joyce *et al.*, Iterative structure-based improvement of a fusion-glycoprotein vaccine against RSV. *Nat Struct Mol Biol* **23**, 811-820 (2016).
283. A. Rohou, N. Grigorieff, CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**, 216-221 (2015).
284. J. M. de la Rosa-Trevin *et al.*, Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J Struct Biol* **195**, 93-99 (2016).
285. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530 (2012).
286. A. Punjani, J. L. Rubinstein, D. J. Fleet, M. A. Brubaker, cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**, 290-296 (2017).
287. M. Sattler, J. Schleucher, C. Griesinger, Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog Nucl Mag Res Sp* **34**, 93-158 (1999).
288. T. Herrmann, P. Guntert, K. Wuthrich, Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular Nmr* **24**, 171-189 (2002).
289. T. Herrmann, P. Guntert, K. Wuthrich, Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology* **319**, 209-227 (2002).
290. D. Gottstein, D. K. Kirchner, P. Guntert, Simultaneous single-structure and bundle representation of protein NMR structures in torsion angle space. *J Biomol NMR* **52**, 351-364 (2012).
291. Y. Shen, A. Bax, Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of Biomolecular Nmr* **56**, 227-241 (2013).
292. W. Kabsch, Xds. *Acta Crystallogr D* **66**, 125-132 (2010).
293. A. J. Mccoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
294. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D* **66**, 486-501 (2010).
295. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D* **66**, 213-221 (2010).

296. P. A. Karplus, K. Diederichs, Linking crystallographic model and data quality. *Science* **336**, 1030-1033 (2012).
297. M. S. Gilman *et al.*, Rapid profiling of RSV antibody repertoires from the memory B cells of naturally infected adult donors. *Sci Immunol* **1**, (2016).
298. J. J. Mousa *et al.*, Human antibody recognition of antigenic site IV on Pneumovirus fusion proteins. *PLoS Pathog* **14**, e1006837 (2018).
299. J. Bonet, Z. Harteveld, F. Sesterhenn, A. Scheck, B. E. Correia, rstoolbox - a Python library for large-scale analysis of computational protein design data and structural bioinformatics. *BMC Bioinformatics* **20**, 240 (2019).
300. D. Baker, What has de novo protein design taught us about protein folding and biophysics? *Protein Sci* **28**, 678-683 (2019).
301. G. J. Rocklin *et al.*, Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168-175 (2017).
302. E. Marcos *et al.*, De novo design of a non-local beta-sheet protein with high stability and accuracy. *Nat Struct Mol Biol* **25**, 1028-1034 (2018).
303. P. S. Huang *et al.*, De novo design of a fourfold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* **12**, 29-34 (2016).
304. F. Sesterhenn *et al.*, Trivalent cocktail of de novo designed immunogens enables the robust induction and focusing of functional antibodiesin vivo. *bioRxiv*, 685867 (2019).
305. S. E. Boyken *et al.*, De novo design of tunable, pH-driven conformational changes. *Science* **364**, 658-664 (2019).
306. J. Dou *et al.*, De novo design of a fluorescence-activating beta-barrel. *Nature* **561**, 485-491 (2018).
307. R. A. Langan *et al.*, De novo design of bioactive protein switches. *Nature* **572**, 205-210 (2019).
308. B. Alberts, Molecular biology of the cell. (2015).
309. K. W. Plaxco, K. T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994 (1998).
310. H. Lechner, N. Ferruz, B. Hocker, Strategies for designing non-natural enzymes and binders. *Curr Opin Chem Biol* **47**, 67-76 (2018).
311. A. J. Burton, A. R. Thomson, W. M. Dawson, R. L. Brady, D. N. Woolfson, Installing hydrolytic activity into a completely de novo protein framework. *Nat Chem* **8**, 837-844 (2016).
312. N. F. Polizzi *et al.*, De novo design of a hyperstable non-natural protein-ligand complex with sub-A accuracy. *Nat Chem* **9**, 1157-1164 (2017).
313. N. H. Joh *et al.*, De novo design of a transmembrane Zn2+-transporting four-helix bundle. *Science* **346**, 1520-1524 (2014).
314. F. Sesterhenn *et al.*, Trivalent cocktail of de novo designed immunogens enables the robust induction and focusing of functional antibodies in vivo. *bioRxiv*, 685867 (2019).
315. S. O. Fedechkin, N. L. George, J. T. Wolff, L. M. Kauvar, R. M. DuBois, Structures of respiratory syncytial virus G antigen bound to broadly neutralizing antibodies. *Sci Immunol* **3**, (2018).
316. D. Y. Tian *et al.*, Structural basis of respiratory syncytial virus subtype-dependent neutralization by an antibody targeting the fusion glycoprotein. *Nat. Commun.* **8**, 7 (2017).
317. P. S. Kulkarni, J. L. Hurwitz, E. A. F. Simoes, P. A. Piedra, Establishing Correlates of Protection for Vaccine Development: Considerations for the Respiratory Syncytial Virus Vaccine Field. *Viral Immunol.* **31**, 195-203 (2018).
318. R. Arts *et al.*, Detection of Antibodies in Blood Plasma Using Bioluminescent Sensor Proteins and a Smartphone. *Anal Chem* **88**, 4525-4532 (2016).
319. M. Sadelain, R. Brentjens, I. Riviere, The Basic Principles of Chimeric Antigen Receptor Design. *Cancer Discov.* **3**, 388-398 (2013).
320. L. Morsut *et al.*, Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors. *Cell* **164**, 780-791 (2016).
321. M. Brenner, J. H. Cho, W. W. Wong, Synthetic biology Sensing with modular receptors. *Nat. Chem. Biol.* **13**, 131-132 (2017).
322. L. Scheller, T. Strittmatter, D. Fuchs, D. Bojar, M. Fussenegger, Generalized extracellular molecule sensor platform for programming cellular behavior. *Nat. Chem. Biol.* **14**, 723-+ (2018).
323. F. Sesterhenn *et al.*, Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen. *PLoS. Biol.* **17**, 27 (2019).
324. C. W. Wood *et al.*, CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics* **30**, 3029-3035 (2014).
325. A. C. Gavin *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147 (2002).

326. L. Giot *et al.*, A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-1736 (2003).
327. R. S. Syed *et al.*, Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature* **395**, 511-516 (1998).
328. N. F. Polizzi *et al.*, De novo design of a hyperstable non-natural protein-ligand complex with sub-angstrom accuracy. *Nat. Chem.* **9**, 1157-1164 (2017).
329. S. Cobey, S. E. Hensley, Immune history and influenza virus susceptibility. *Curr. Opin. Virol.* **22**, 105-111 (2017).
330. C. Henry, A. K. E. Palm, F. Krammer, P. C. Wilson, From Original Antigenic Sin to the Universal Influenza Virus Vaccine. *Trends Immunol.* **39**, 70-79 (2018).
331. M. Silva *et al.*, Targeted Elimination of Immunodominant B Cells Drives the Germinal Center Reaction toward Subdominant Epitopes. *Cell Reports* **21**, 3672-3680 (2017).
332. W. C. Ruder, T. Lu, J. J. Collins, Synthetic Biology Moving into the Clinic. *Science* **333**, 1248-1252 (2011).
333. M. Gossen, H. Bujard, Tight control of gene-expression in mammalian-cells by tetracycline-responsive promoters. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5547-5551 (1992).
334. L. Scheller, M. Fussenegger, From synthetic biology to human therapy: engineered mammalian cells. *Curr. Opin. Biotechnol.* **58**, 108-116 (2019).
335. A. Weiss, J. Schlessinger, Switching signals on or off by receptor dimerization. *Cell* **94**, 277-280 (1998).
336. K. Y. Chang *et al.*, Light-inducible receptor tyrosine kinases that regulate neurotrophin signalling. *Nat. Commun.* **5**, 10 (2014).
337. J. Schlessinger, Cell signaling by receptor tyrosine kinases. *Cell* **103**, 211-225 (2000).
338. I. Chung *et al.*, Spatial control of EGF receptor activation by reversible dimerization on living cells. *Nature* **464**, 783-U163 (2010).
339. K. Mohan *et al.*, Topological control of cytokine receptor signaling induces differential effects in hematopoiesis. *Science* **364**, 750-+ (2019).
340. A. M. Brzozowski *et al.*, Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **389**, 753-758 (1997).
341. H. F. Ye *et al.*, Pharmaceutically controlled designer circuit for the treatment of the metabolic syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 141-146 (2013).
342. L. Schukur, B. Geering, G. Charpin-El Hamri, M. Fussenegger, Implantable synthetic cytokine converter cells with AND-gate logic treat experimental psoriasis. *Sci. Transl. Med.* **7**, 11 (2015).
343. P. Saxena, G. Charpin-El Hamri, M. Folcher, H. Zulewski, M. Fussenegger, Synthetic gene network restoring endogenous pituitary-thyroid feedback control in experimental Graves' disease. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 1244-1249 (2016).
344. M. Xie *et al.*, beta-cell-mimetic designer cells provide closed-loop glycemic control. *Science* **354**, 1296-1301 (2016).
345. D. Bojar, L. Scheller, G. Charpin-El Hamri, M. Q. Xie, M. Fussenegger, Caffeine-inducible gene switches controlling experimental diabetes. *Nat. Commun.* **9**, 10 (2018).

Che Yang

yangche7@hotmail.com, +41-789647554, Taiwan

EDUCATION

2011-13 Master in Biochemical Science, National Taiwan University, Taiwan,

- GPA: 4.19/4.30; 4.22/4.3 (major); Rank: 2/21; Advisor: Prof. Rita P.-Y. Chen (Academia Sinica)
- Thesis: Exploring the α-to-β Structural Conversion Mechanism for Mouse Prion Protein

2007-11 Bachelor of Science in Life Sciences, National Central University, Taiwan,

- GPA: 3.9/4.0; 4.0/4.0 (major); Rank: 1/40

CURRENT AND PREVIOUS SCIENTIFIC ACTIVITIES

10/2015- Doctoral student, Ecole Polytechnique Fédérale de Lausanne, Switzerland

- Thesis: Developing next-generation vaccine through computational protein design

2015 Associate research fellow, Development center for biotechnology, Taiwan

- Design and characterize knobs-in-holes bispecific antibody against solid tumor

2011-13 Masters student, Biophysics/protein science lab., Academia Sinica, Taiwan

- Provide solid evidence for the structural conversion of prion protein by ESR approach

2012 Visiting Student, IMRAM (Instit. of Multidisciplinary Research), Tohoku University, Japan

- Demonstrate a correlation of structure dynamic and FRET efficiency in prion protein

2011 Summer student, Chemical/Pharmacological lab., National Taiwan University

- Identify potential compounds that inhibit SARS protease

2009-11 Undergraduate research student, National Central University

- Independently clarify the signaling pathway of one type of GPCR (G2A receptor)

Honors&Awards

2013 **Oral presentation award**, The 18th Biophysics Conference of R.O.C, Taiwan
2013 **Best poster award**, 8th Asian Biophysics Association Symposium, Korea
2013 **Travelling Grant**, Biophysical Society of R.O.C
2013 **Outstanding poster award**, The 28th Joint Annual Conference of Biomedical Science, Taiwan
2013 **First prize**, Master thesis in Institute of Biochemical Sciences, National Taiwan University
2011 **Dean's List**, Department of Life Science, National Central University
2011 **Outstanding undergraduate project**, National Science Council, Taiwan
2008-11 **Presidential Award**, National Central University

PUBLICATIONS

-Peer-reviewed articles, *equal contribution

- Sesterhenn F*, **Yang C*** *et al*., "Trivalent cocktail of de novo designed immunogens enables the robust induction and focusing of functional antibodies in vivo". *bioRxiv* 2019, doi:10.1101/685867
- Sesterhenn F, Galloux M, Vollers SS, Csepregi L, **Yang C** *et al*., "Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen". *PLOS Biology* 2019, doi: 10.1371/journal.pbio.3000164
- Bonet J*, Wehrle S*, Schriever K*, **Yang C*** *et al*., "Rosetta FunFolDes – A general framework for the computational design of functional proteins." *PLoS Comput Biol.* 2018, doi: 10.1371/journal.pcbi.1006623.
- **Yang C** *et al*., "Revealing structural changes of prion protein during conversion from α-helical monomer to β-oligomers by means of ESR and nanochannel encapsulation," *ACS Chemical Biology* 2014, doi: 10.1021/cb500765e

PATENTS

"Designed epitope-focused immunogens" Bruno Correia, Fabian Sesterhenn, **Che Yang**, Jaume Bonet. European patent application - EP19183026.4.

POSTER AND ORAL COMMUNICATIONS

2019 **Invited presentation**, 14th Asian Congress on Biotechnology, Taiwan

- Topic: A Bottom-up Functional De Novo Protein Design Approach: Reading from Functional Motif to Construct a Stable Protein Topology

2015 **Poster presentation**, Bioengineering Day, EPFL, Swiss

- Topic: Towards the rational design of improved immunogens for novel vaccines: developments on computational protein design and structural vaccinology

2013 **Poster presentation**, The 2nd NRPB International Symposium for Membrane Protein, Taiwan

- Topic: Exploring the α-to-β Structural Conversion Mechanism for Mouse Prion Protein

2013 **Poster presentation**, The Conference of Taiwan Society for Biochemistry and Molecular Biology

2012 **Oral presentation**, Tripartite conference of Hokkaido University, Tohoku University, Institute of Biological chemistry in Taiwan Academia Sinica, Japan

- Title: Misfolding of prion protein