Exploring the Molecular Design of Ligand Binding Sites by Computational Protein Design

Jiayi Dou

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

David Baker, Chair

Barry L. Stoddard

James D. Bryers

Program Authorized to Offer Degree:

Department of Bioengineering

University of Washington

Abstract

Exploring the Molecular Design of Ligand Binding Sites by Computational Protein Design

Jiayi Dou

Chair of the Supervisory Committee:
Professor David Baker
Department of Biochemistry

Ligand binding sites in natural proteins, with diverse structural details, provide the foundation for enzymatic activity, antibody-antigen recognition, ligand-induced pathway activation and drug discovery in general. The work presented in this dissertation seeks to understand the general design principles of the molecular details revealed in the ligand-protein complex structures. An engineering approach based on computational protein design was taken to expand the boundary of our current knowledge. By combining computational structural modeling and protein biochemical characterization, computational design of ligand binding proteins iterates between structure-based design hypotheses and experimental validation. This research scheme was applied to two related topics: 1) re-purposing natural ligand binding sites and 2) designing *de novo* ligand binding proteins. Representative small molecules, steroids (digoxigenin, 17-αhydroxylprogesterone, cortisol) and an environmentally sensitive fluorophore (DFHBI), were

chosen as design targets. High-resolution X-ray crystal structures of the engineered proteins were obtained and analyzed for modeling feedback. Binding affinity and specificity, protein stability and function, as well as modeling challenges were discussed in each case. The design methods developed and tested in this work represent a systematic way of engineering small molecule binding sites and can be expanded to broad applications. As a rigorous test of our current knowledge, computational design of ligand-binding proteins presented in this work emphasizes the high precision required for accurate ligand positioning and protein conformation modeling.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I want to thank my colleagues, friends and family for their support over the years, without whom, this thesis would not have been possible.

First and foremost, I thank my thesis advisor, David Baker, for supporting my scientific growth during my time at UW. David fostered a professional and friendly lab environment where I learned not only how to think like a scientist but also how to find my own voice and stand up for it. His continuous encouragement for scientific conversations between diverse groups of research often leads to interesting collaborations; his constant excitement for new ideas and discoveries inspires us to thrive with creativity; his generous support for lab entertainment makes a welcoming and delightful work environment. I vividly remember seeing him set up protein gel electrophoresis and wait for centrifuge with a pair of pink gloves in lab. Together with several graduate students in lab, we once came to his desk to admire the protein models he made and ask him which following experiments he had in mind. With an 80-people lab to manage, he always tries to have his own projects to work on. David always reminds me of all the joys and excitements in the pursuit of science.

I also thank my professors at UW, who opened the door of structural biology and bioengineering for me. In particular, my committee member Barry Stoddard provided important feedback on my projects. His technician Lindsey Doyle contributed to all of my projects by setting crystal trays and solving the structures of proteins I designed. Wendy Thomas took me as a rotation student and supported me in my second year in graduate school. James Bryers and Wenqing Xu gave me suggestions on every stage of my graduate school. It is such an honor to have four of you in my committee!

I also thank the numerous students and postdocs that I had the opportunity to interact with in the Baker lab. In particular, I thank past and current members of our subgroup of small-molecule binding design, including Christy Tinberg, Austin Day, Sagar Khare, Per Greisen, Matt Bick, Jason Klima and Ralph Cacho. I enjoyed discussing science and becoming friends with them. Christy

mentored me when I joined the lab as an intimidated first-year graduate student from a foreign country. Her warm personality helped me become a part of big Baker lab. She taught me not only the experimental skills but also the integrity for rigorous science. Thank you Christy!

During my last two years I had the opportunity to work with a talented postdoc in the lab, Anastassia Vorobieva. I could not have asked for a better collaborator: Anastassia's professionalism and love of science helped drive the project presented in Chapter 2. I greatly enjoyed learning with Anastassia. I thank her for her friendship, scientific discussions, and help with my experiments. I will not forget the summer BBQ we had in your beautiful garden, the oyster party in your basement apartment, and sailing with you on Lake Washington! I will also remember the excitement we shared for making the first *de novo* small-molecule binding protein!

Last but not least, I could not have made it through without the support of my beloved family and friends. I want to thank my parents, Xiaofeng Dou and Shujuan Wang, and my brother, Jialong Dou, for their constant and unyielding support during this time. My friends, Jingda Wu, Yu Jin, Jinchao Huang and Fan Zheng, are great advocates of me. Thank you all for providing a constant source of love and happiness in my life!

INTRODUCTION

1. INTRODUCTION TO MOLECULAR INTERACTIONS BETWEEN PROTEINS AND SMALL MOLECULES

Molecular recognition is at the root of all the essential processes happening in biology. Specific interactions between proteins and small molecules are associated with enzymatic activity, antigen-antibody interaction, ligand-induced pathway activation and drug discovery in general. The ability of a protein to bind selectively and with high affinity to a small molecule depends on the formation of a set of weak non-covalent interactions such as van der Waals attractions, electrostatic attractions and hydrogen bonds. Since each individual interaction is weak, effective binding occurs only when many of these interactions form cooperatively. Such a combination of interactions is possible only if the surface contour of the ligand molecule fits very closely to the protein. Additionally, in aqueous media, where biology happens, non-covalent interactions must compete with the substantial solvation energy of ions and polar groups. The de-solvation penalty from burying polar protein side chains usually requires a compensation by protein folding energy; for a water-soluble small molecule, the de-solvation penalty upon binding needs to be paid by the newly formed interactions with its protein receptor. The free energy of protein-small molecule association comes from a "visible" enthalpy, where all the new interactions contribute and an "invisible" entropy, which benefits from freeing constrained water molecules upon desolvation[1]. On the other hand, kinetics of ligand-binding interactions mostly depends on the activation barrier for association and dissociation (Figure 0.1).

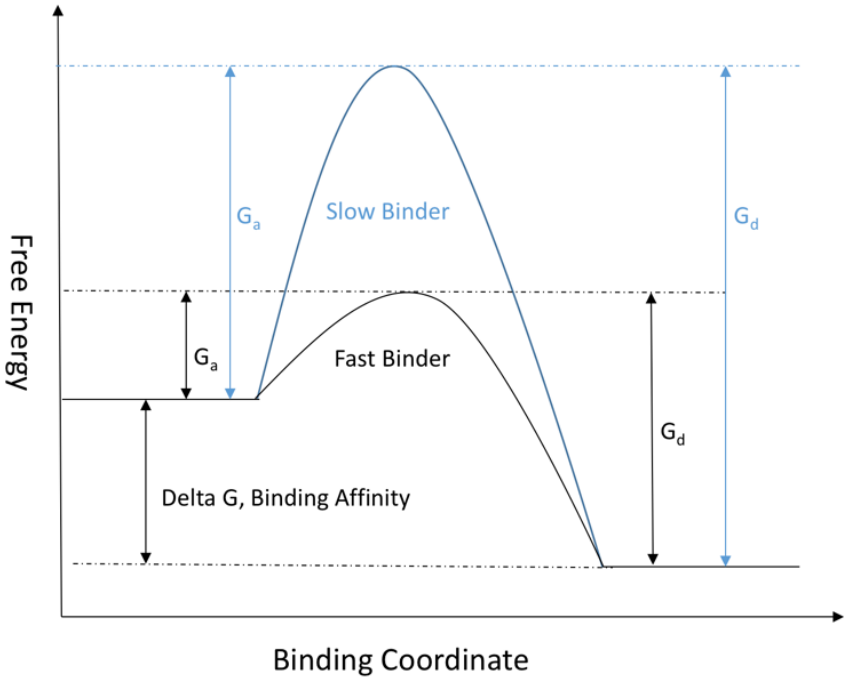


Figure 0.1. Free-energy diagram of ligand-protein interactions.

Structural biology has largely expanded our knowledge of protein-ligand interactions and has allowed us to study these binding interactions with atomic details. Many high-resolution crystal structures of ligand-bound protein complexes have revealed the specific physicochemical details of interacting groups. For example, an early crystal structure of streptavidin-biotin complex observed "an extensive pattern of hydrogen bonds", which led the authors to the conclusion that "the unusual high affinity of streptavidin for biotin reflects participation of a number of factors, the analogs of which have been previously encountered individually in other protein-ligand interactions"[2]. Based on the structures of both unbound and bound conformations, the extremely slow off-rate ($2.4 \times 10^{-6} s^{-1}$)[3] of streptavidin-biotin interaction was attributed to a ligand-induced conformation change where an eight-residue loop becomes structured upon biotin binding and acts like a lid to cover the ligand[4]. Another excellent example is the discovery and establishment of

cation-pi interaction: a wide range of results from structural biology, together with the studies of organic model systems, strongly supported the importance of cation-pi interactions in a variety of proteins that bind cationic ligands[5]. In many other cases, a high-resolution structure coupled with mutagenesis and biochemical studies was able to relate a structure feature to its contribution for binding properties (Figure 0.2).

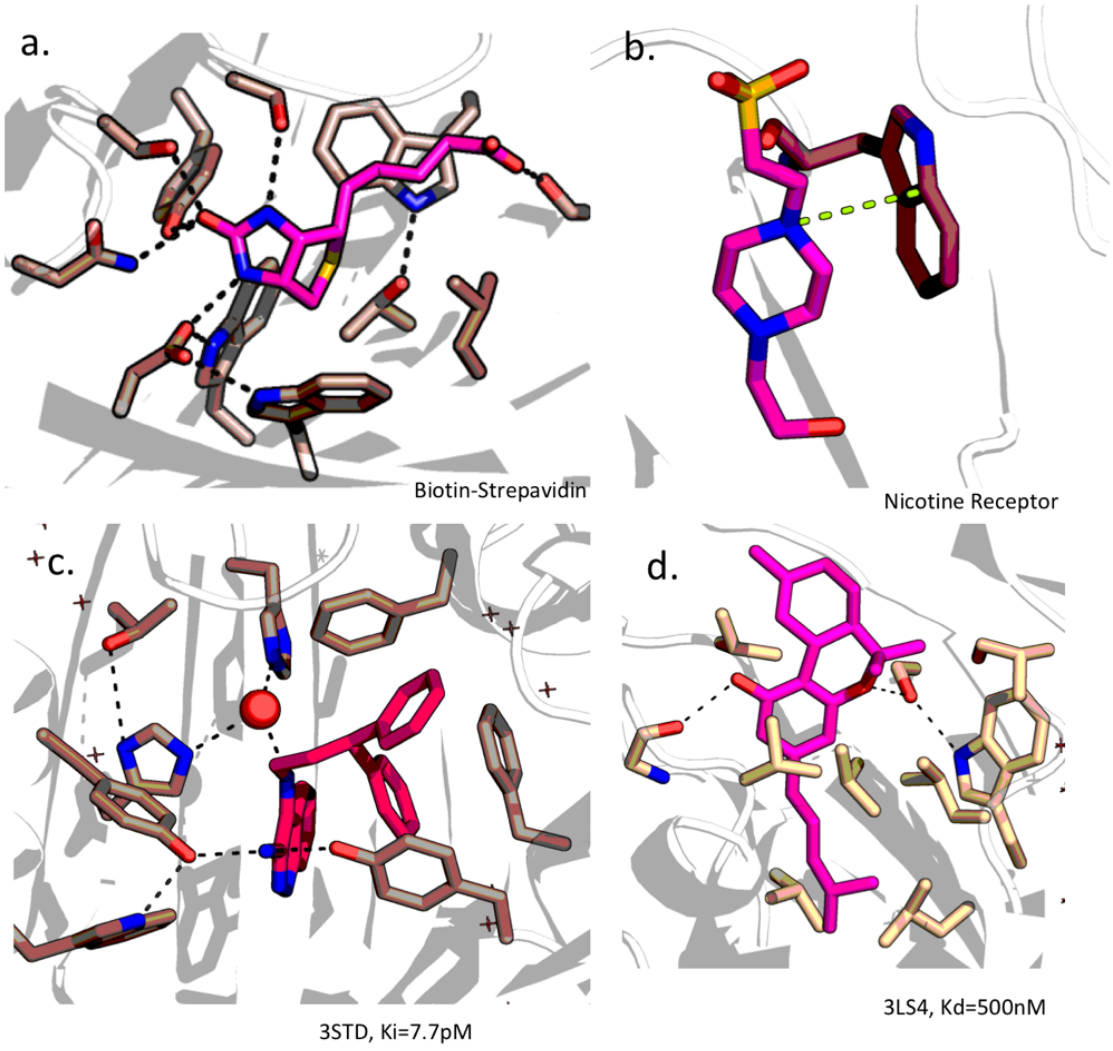


Figure 0.2. Examples of the molecular details of natural ligand binding proteins. **a.** Binding interface of streptavidin and biotin (PDB ID: 1STP). Hydrogen bonding interactions are highlighted as black dashed lines. All the polar groups in a biotin molecule are satisfied by a complementary hydrogen bonding side chain in streptavidin. **b.** Cation-pi interaction observed in

nicotine receptor (PDB ID: 1I9B). **c.** Bound structure of scytalone dehydratase with a designed inhibitor (PDB ID 3STD). Extensive pi-pi interactions and water-bridged hydrogen bonding interactions were observed on the interface. **d.** Anti-tetrahydrocannabinol (THC) Fab fragment in complex with THC (PDB ID: 3LS4). Hydrophobic interactions were mediated through small aliphatic sidechains rather than aromatic packing.

Structural evidences derived from antibodies in their free and bound forms shed light on the general mechanistic studies of protein-ligand interactions, which preferentially promoted the idea of induced-fit ligand-binding mechanism[6]. Following structural studies on antibody maturation from weak binding (micro-molar affinity) to high affinity binding (nano-molar affinity) have shown that a number of factors are involved in improving affinity across this range. Several engineering studies pushed the binding affinity of natural antibodies to femto-molar range or beyond. In those cases, pre-organization of protein binding pockets and "lock-and-key" fitting mechanism were highly emphasized for affinity improvement based on the structural snapshots during the course of affinity improvement[7–10].

As more and more high-quality structural data become available along with the constantly increasing power of computers, the modeling and prediction of protein-small molecule interactions becomes possible and continues growing as a thriving field[1]. It has become widely accepted that the general principles derived from rich structural data can be used to demarcate paths and develop strategies for a range of computational modeling and design problems.

2. INTRODUCTION TO COMPUTATIONAL PROTEIN DESIGN

Ever since the determination of the first protein structure[11], the computational modeling of macromolecules has grown into an important aspect of structural biology[12], serving as a complement to conventional biochemical experiments and structural determination. The two main

families of simulation technique are molecular dynamics (MD)[12] and Monte Carlo (MC)[13]. Molecular dynamics(MD) simulations solve the equations of Newtonian motion for all the atoms in the system with a potential energy force field. The force fields for MD, aimed at accurately reflecting the physics, often include classical terms as well as many cross-terms, where the parameters are typically determined by quantum chemical calculations combined with thermophysical and phase coexistence data. As its simulation consists of the numerical, step-by-step, solution of the classical equations of motion, MD's predictive power is limited by the system size and the time scale of the simulation. In contrast, Monte Carlo(MC) simulations are less restricted by the simulation scale. A typical MC simulation run is a series of random steps in conformation space, each perturbing some degrees of freedom of the system. A step is accepted with a probability that depends on the change in value of an energy function. It has been shown that a long MC simulation with the Metropolis criterion and an appropriate step generator produces a distribution of accepted conformations that converges to the Boltzmann distribution[14]. For MC simulations, proteins are represented as long kinematic chains with random perturbations from defined degrees of freedom in either torsion space or Cartesian coordinate space. While MD simulations have been widely used for elucidate the protein dynamics in very short time scale, MC simulations have proved quite efficient for sampling a large conformational space. For this reason, MC simulation methods have been adopted for protein structure prediction and designing new proteins.

The general goal of computational protein design is to generate amino acid sequences to adopt a specific three-dimensional structure and to realize a specified function. Although the search space associated with such design problem is vast and combinatorial in nature, there has been many successful cases in the design of *de novo* protein structures, enzymatic catalysts, self-assembling

protein cages, and binding proteins for protein targets and small molecules of interest[15–20] . These successes have occurred largely because of many simplifying assumptions that can drastically reduce the effective search space, making rational design computationally feasible. For most protein design problem, once the protein backbone is defined, either from existing protein structures[21] or from *de novo* modeling[15,22], the backbone conformation is modeled as fixed scaffold to reduce the size of searching space. Even with this fixed-backbone assumption, the continuous conformations of each amino acid type at each position can similarly increase the search space beyond what we are capable of modeling. Another assumption to remedy this problem is the use of discrete side chain conformations called rotamers[23]. These assumptions reduce the design problem to the problem of finding the right amino acid rotamer combinations for a pre-defined backbone conformation. The rotameric protein models are then scored using an energy function developed to reproduce the features of natural functional proteins. Most energy functions use a combination of physical-based and knowledge-based energy terms, each of which describes one particular interaction. Typically, van der Waals interactions are approximated by Lennard-Jones potential, and hydrogen bonds are mostly geometric-based. Statistical potentials are used to describe backbone conformations derived from the Ramachandra diagram. Although interactions between a protein and its aqueous environment are crucial, protein design cannot afford modeling water molecules explicitly. Therefore, solvation effects are often described implicitly[24]. Electrostatic attraction is described by either statistical potential or simple physical approximation. Most energy functions have an additional reference energy term to balance the overlap between different terms. Energy functions are usually further simplified to compute the interaction energy between each pair of rotamers, as well as the energy between each rotamer and the protein backbone. The sum of pairwise rotamer-rotamer energies and rotamer-backbone energies is used

as a relative standard for optimization. Based on certain benchmarks, energy function can be modified or extended for a specific design problem[25]. These simplifications and pair-wised score terms effectively reduce the protein design problem to a mathematic optimization problem in a combinatorial searching space.

Our lab along with a collaborating scientific community develops a protein-modeling software suite known as Rosetta[26]. This software offers various modular functions that aid the design and optimization of protein structure and function. Rosetta uses a Monte Carlo searching algorithm that can sequentially optimize protein sequences by making random perturbations and accepting or rejecting them based on an energy evaluation. It is essentially stochastic and is able to give a result regardless of the difficulty, and there is no guarantee that the global minimum has been reached. Other searching algorithms like greedy algorithm are used in Rosetta for local sequence searching too. Deterministic optimization algorithms, like dead-end elimination[27], are implemented in other protein design programs.

Computational design of ligand binding proteins, develops and experiments with computational methods for: 1. docking a small molecule into binding sites of protein scaffolds and 2. optimizing the amino acids composition on the binding interface to make optimal interactions. Computational models are subject to a serial of biochemical tests regarding protein stability, binding thermodynamics and kinetics. The designed protein-ligand interactions are further perturbed by mutagenesis and binding selection to better understand the fitness landscape of the designed binding interaction. Structural information is considered the most valuable data to confirm and further study the designed interactions. Once a functional binding protein is experimentally confirmed for a chosen small-molecule target, further optimization is usually required to satisfy a specific application. (Figure 0.3).

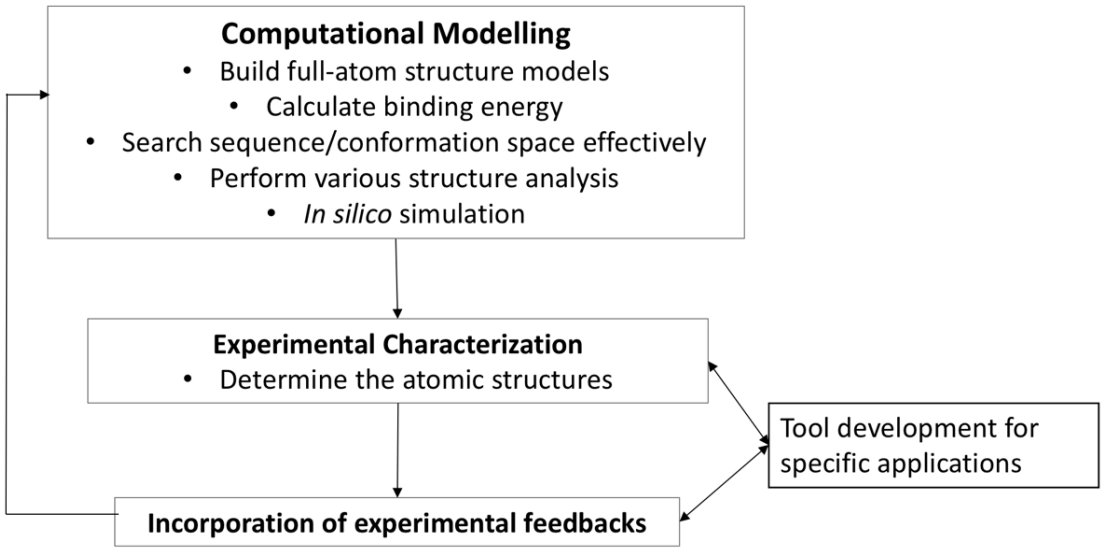


Figure 0.3. General working scheme of computational protein design

3. APPLICATION OF ENGINEERING PROTEIN-SMALL MOLECULE INTERACTIONS

In his seminal paper published in 1981[28], K. Eric Drexler proposed the idea of manipulating molecules with atomic precision and coined the word "protein design" for molecular engineering. As a visionary engineer, he concluded the paper by saying:

"Development of the ability to design protein molecules will, by analogy between features of natural macromolecules and components of existing machines, make possible the construction of molecular machines. … This capability has implications for technology in general and in particular for computation and characterization, manipulation, and repair of biological materials."

Although this sounded quite outlandish at the time, many technologies he proposed have come to fruition in the past two decades. Engineering natural proteins for various applications has become a routine for many research labs. Protein-based molecular devices for detecting and reporting have been used not only as research tools, but also are being developed for wearable devices for disease

monitor[29]. Protein engineering, especially antibody engineering, also holds great promise for revolutionizing medicine as the molecular recognition power of proteins can be manipulated to perform therapeutic tasks. In fact, there has been many FDA-approved antibody therapeutics in the past decades[30]. For small-molecule ligands, as this dissertation is mostly concerned, a protein binding domain can be used for 1. disease diagnostics and therapy[31], 2. detection and quantification of drugs and narcotics, and 3. applications for food storage and environmental monitoring[32]. There are also many exciting research applications in need of ligand-binding proteins to guide and manipulate cell signaling and cell fate[33].

While it has been proved quite effective to repurpose natural proteins, conventional experiment-based engineering approaches lack the atomic control. It is fair to say that we are still far from "positioning reactive groups to atomic precision" as K. Eric Drexler stated in 1981. As I will elaborate in the following chapters, computational protein design aims to integrate computational modeling methods and experimental approaches to provide generality and high accuracy for the general field of protein engineering.

4. SPECIFIC TOPICS IN THIS DISSERTATION

The work of this dissertation comprises two major topics: 1. Re-purposing natural ligand binding sites by computational design and 2. Designing a *de novo* protein with ligand binding functionality. The detailed method development and experimental results are presented in Chapter 1 and Chapter 2, respectively. In Chapter 3, I explored the latest technology development in gene synthesis and next-generation sequencing and presented a side project for high-throughput experimental assessment of designed proteins.

Chapter 1. RE-PURPOSING NATURAL LIGAND BINDING SITES BY COMPUTATIONAL PROTEIN DESIGN

1.1 INTRODUCTION

1.1.1 *Natural proteins as a great source of inspiration for protein engineering*

Proteins as the workhorses in biological systems can perform incredibly diverse tasks faced by different organisms in the process of natural selection. In comparison with manmade macroscopic devices, proteins do their jobs with atomic precision and high efficiency on the molecular level. For protein engineers, natural proteins represent a great source of admiration and inspiration for protein engineers. The repertoire of natural ligand-binding proteins contains the immunoglobulin antibody and a list of proteins from other fold families. The implementation of DNA mutagenesis in a large scale combined with high-throughput functional selection has enabled protein engineers to repurpose many natural proteins and produce ligand-binding proteins beyond what nature has evolved. The representative ligand-binding proteins generated by combinatorial library selection include single-chain-variable fragments of antibodies, “Anticalin” lipocalin proteins, and “Affibodies” helix bundles. The engineering work of repurposing those proteins proved the possibility of protein re-engineering and laid the foundation for computational protein design.

Antibodies are host proteins that are naturally produced by the mammalian immune systems in response to foreign molecules that enter the host body. As the most versatile recognition proteins seen in nature, antibodies have a wide range of applications in biochemistry research and medicine. For this reason, the animal immune systems have been harnessed to manufacture custom antibodies routinely since 1980s. Although the antibody production techniques have been matured in biotech industry for several decades, the time and effort commitment to make a monoclonal

antibody is still tremendously large[34]. The size of natural antibodies is around 150kDa with two identical heavy chains and two identical light chains connected by disulfide bonds (Figure1.1.a). The binding regions include six variable loops referred to as the complementarity determining regions (CDRs). The loop flexibility enables antibody to adapt to a variety of antigens while hindering the modeling and rational engineering. Therefore, the generation of new antibodies still relies largely on the mammalian immune system. Directly inherited from antibody structure, single-chain variable fragment (scFv) proteins are recombinant polypeptides, composed of an antibody variable light chain tethered to a variable heavy chain by a designed linker[35] (Figure1.1.b). They have the same specificities and affinities for their antigens as the monoclonal antibodies whose heavy and light chains were used to construct the recombinant genes. The smaller size has made scFv better recognition proteins for some applications. In addition, laboratory directed evolution has pushed the binding ability of scFv beyond the natural monoclonal antibodies[7]. The discovery of single-domain antibodies (Nanobodies) from camelids that contain only heavy chain variable fragments has enabled the development and engineering of even smaller immunoglobulin (IgG)-based recognition proteins[36].

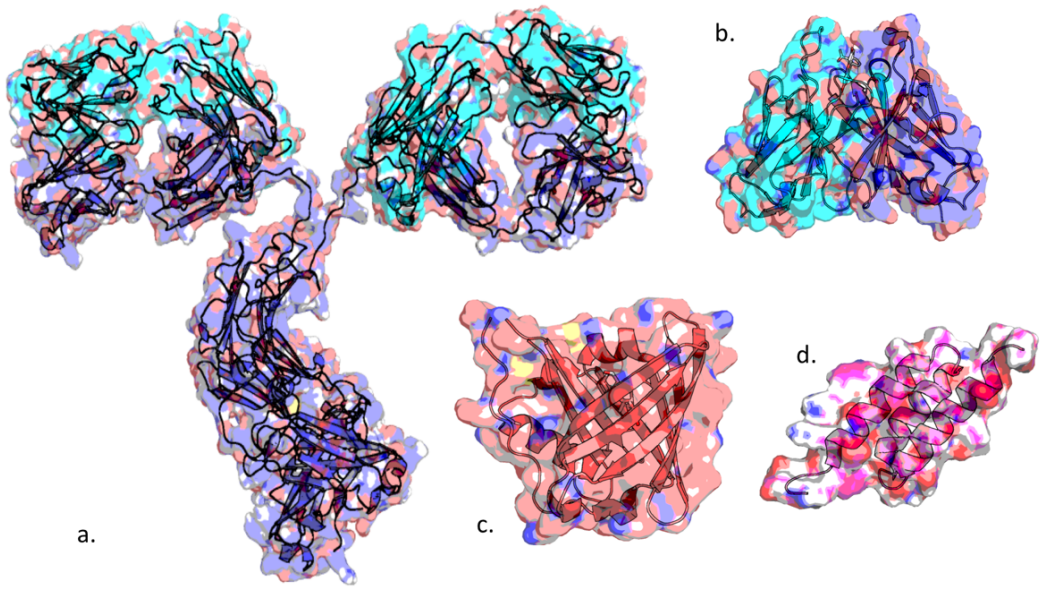


Figure 1.1.Representative ligand binding proteins from combinatorial selection. **a.** An antibody structure that contains two identical heavy chains (purple) and two identical light chains (cyan). Heavy and light chains are connected by disulfide bonds. The molecular weight of an antibody molecule is around 150kDa. **b.** The single-chain variable fragments derived from antibody structure. Variable regions of heavy and light chains are fused together by a flexible linker. **c.** An example "Anticalin" lipocalin protein with anti-parallel beta-barrel structure. **d.** An example "Affibody" molecule that contains three helices.

The exploration of ligand-binding ability beyond immunoglobulin (IgG) fold gained a lot of attention after the success of scFv[37]. One particular fold is the helix bundles proteins named as Affibodies[38]. Affibodies are small, engineered protein domains that are capable of specific binding to target ligand. They were selected from randomization of solvent-accessible surface residues of a stable alpha-helical bacterial receptor domain Z of protein A. Most available Affibodies target proteins or peptides instead of small molecules due to the simplicity of its structural fold (Figure 1.1.c). Another group of engineered ligand-binding proteins are the Anticalins from lipocalin proteins[39] (Figure 1.1.d). This protein family shares a conserved barrel of eight antiparallel beta-

strands and four of the loop regions at one end of the barrel form the binding sites. They normally serve for the storage or transport of physiologically important compounds. Member of this protein family have been engineered to bind several different small molecules by combinatorial library construction and selection[39]. However, the presence of multiple disulfide bonds and solubility issues have limited its broad applications.

1.1.2 *Methods for repurposing the ligand-binding sites in natural proteins*

a. Directed evolution

Directed Evolution is a general term used to describe various molecular biology techniques that mimic the natural process of mutation and selection, techniques that have been proved extremely useful for protein engineering[40]. These techniques involve introducing mutations (randomly or rationally) at the genetic level followed by selection for the desired characteristics at the protein level. A variety of methods are available to perform site-directed or random mutagenesis. Most of them are well established and easily adapted for parallelization. Since DNA synthesis cost has dropped dramatically, very large libraries can be constructed either from DNA assembly with doping oligos, or direct gene synthesis. Most laboratory directed evolution work is limited by the selection throughput. Display systems have been developed to connect the DNA variants to the diverse proteins expressed and displayed on the surface of phage viruses, bacteria, yeasts, or ribosomes. For example, the yeast display system has been optimized for constructing large libraries containing up to ~10^8 variants[41]. When labeled by compatible fluorescent reagents, the binding properties of displayed protein can be tested by flow cytometer or selected by fluorescence activated cell sorting (FACS). The iteration of *in vitro* DNA mutagenesis and high throughput FACS selection can sample a particular protein sequence space quite effectively and find out variants with improved binding property. Recently, the next- generation sequencing (NGS)

technique has provided a thorough approach to map the mutation effect. Benchtop NGS sequencers can sequence the entire library and generate a complete sequence-function mapping[42].

b. Computational modeling as general engineering tool for constructing new ligand binding sites

As the selection assay often requires a detectable signal and a serial of gradual improvements is necessary, laboratory directed evolution work can be quite laborious for realizing a dramatic change[43]. Given the astronomic number of protein sequences available for sampling, the screening effort can become tremendous without additional structural guidance. Computational modeling, in contrast, requires structural information to start with and aims to provide a small set of possible solutions for experimental verification. By effectively avoiding non-productive variants based on energy calculation, computational modeling can help reduce the lab work and potentially find novel solutions beyond what directed evolution can reach. Since it is not limited by the pre-existing activity and it is guided by structural modeling, computational design provides a general approach to build new ligand binding sites based on proteins of known structure.

The process of computational design of ligand binding consist two steps: 1. placing the new ligand into the binding site of existing protein structure and 2. changing the surrounding amino acid identities to better accommodate the new ligand. To circumvent the protein folding problem and reduce the sampling space, the overall protein backbone is kept fixed in most design algorithms (as a general assumption for protein modeling and designing). Together with the assumption of discrete rotameric states for the protein side chains, the design problem become a search for a constellation of backbone positions that can accommodate the proper sidechain rotamers that can make favorable interactions with the target ligand. Based on these two assumptions, computational design of protein-metal interactions with clear geometric requirements are among the early successes[44,45].

An early success of computational design of small molecule binding came from the computational re-design of bacterial Periplasmic Binding Proteins (PBPs). Looger and colleagues used the three-dimensional structure of PBP proteins and superimposed the target ligands (serotonin, ribose, TNT) onto the native substrate and redesigned the binding sites to form stereochemically complementary "lock-and-key" interfaces[46]. The combinatorial search algorithm used in this case simultaneously optimizes sequence choice and ligand configuration. The binding activity of re-designed PBPs was interrogated indirectly by an *in vivo* fluorescent assay and was later challenged by biochemical and structural studies[47]. The computational design of ligand binding sites has thus been entitled "an unsolved problem".

1.1.3 *Computational methods for enzyme design*

Enzymes with their remarkable catalytic activity are attractive targets for engineering work in biotech industry as well as in academic research labs. Designing enzyme molecules to perform chemical catalysis has become the Holy Grail for computational protein design. While many natural enzymes have very sophisticated multi-step catalysis mechanisms, the non-covalent interaction between an enzyme and its small-molecule substrate is governed by basic physical forces and thus share a general design principle. In fact, the generation of catalytic antibodies and computational enzyme design mainly focused on the binding interactions between the protein and the transition-state analog of the enzymatic reaction[17]. Thus, many enzyme design methods can be adapted for designing and testing small molecule binding.

For most computationally generated enzymes, the catalytic efficiency is quite low in comparison with the natural enzymes. Directed evolution is almost always applied to boost the activity. While many natural enzymes display the diffusion-limited catalysis potency, designed protein catalysts usually require high substrate concentrations to reach maximum reaction velocity (high K_m). The

major caveat of enzyme design appears to be due to the low binding affinity of substrate recognition. It seems reasonable that the effort to improve the design precision for ligand recognition can boost the catalysis efficacy of designed enzymes.

1.2 METHODS

1.2.1 *Scaffold library curation*

Computational design of ligand binding proteins by repurposing natural proteins requires a curated library of suitable protein structures for providing the "scaffolds". A general scaffold library includes all the high-quality protein-small molecule complex structures available in RCSB Protein Data Bank (PDB). The curation criteria usually consider the quality of X-ray diffraction data and the protein oligomerization state. For specific computational benchmark calculations, a scaffold library should exclude structural homologs to avoid statistically meaningless duplicates. Structural alignment scores based on backbone similarity are used to further filter out similar scaffolds. Protein structures solved by nuclear magnetic resonance (NMR) are not included in most scaffold libraries since the assemble models are ambiguous in the sense of atom coordinates. Also, most NMR structures in PDB are from small single-domain proteins that do not have a pre-existing binding site; therefore, they are not suitable for redesigning ligand binding sites.

A chosen protein-small molecule complex undergoes several simple pre-treatments before becoming a scaffold. The first thing is to eliminate non-amino acid atoms in the structures. Co-crystalized salt ions, bound water molecules, and the native ligands are removed to simplify the computational modeling. The structures are further relaxed with heavy-atom coordinate constraints to fit well into the Rosetta energy landscape. A systematic assessment of computational protocols has been done to ensure the pretreatment modeling can eliminate local clashes while maintaining

the fidelity of the experimental data[48]. A companying file that defines the pocket position is often required for downstream design calculation. The simplest and most reliable way to define a pocket is to assume the native binding site. Other methods that can detect a secondary or "hidden" binding pocket can be adapted from drug discovery for this purpose as well.

Binding MOAD (Mother Of All Databases)[49] is a general binding database that maintains the largest collection of well resolved protein crystal structures with clearly identified biologically relevant ligands annotated with experimental determined binding data. Entries from MOAD can be further prepared and relaxed according to the Rosetta energy function. A general large-scale computational design usually starts with MOAD. Since multiple mutations introduced by computational design usually destabilize the protein, a suitable protein scaffold should also be stable enough to tolerate arbitrary mutations. With many rounds of experimental feedbacks, the Baker group has accumulated experimental data on *Escherichia coli (E.coli)* protein expression and protein stability. For different engineering purposes, specialized scaffold libraries have been curated based on thermostability, protein fold family, or known conformational changes upon binding.

1.2.2 *Computational docking methods*

RosettaMatch[50] (first developed for enzyme design) has been adapted for designing protein-small molecule interactions. For a chosen small molecule, RosettaMatch requires a set of pre-defined side chains centered on the target ligand. Those interactions can be described by geometric elements that are used for looking for accommodating backbone positions and later used as constraints for energy calculation. For every scaffold in the scaffold library, RosettaMatch places each pre-defined interacting side chain to one of the pocket position and positions the ligand according to the defined geometry. The compatibility of the next pre-defined protein side chains

at certain backbone position can be assessed quickly by checking if the target ligand still stays the same position (Figure 1.2.a). The hashing algorithm implemented in RosettaMatch allows very fast searching through hundreds of scaffolds. While it allows precise control over the geometry of interactions and enables designing selectivity for a specific chemical moiety, RosettaMatch heavily relies on the geometric descriptions and can be potentially biased towards some non-productive ligand configuration. Other small molecule docking methods developed for drug discovery can also be used to determine an initial ligand position. For example, PatchDock[51], a fast low-resolution docking method that emphasizes the shape complementarity between protein receptor and the ligand has the advantage of putting ligand into a complementary pocket and allows further design steps to make specific interactions. (Figure 1.2.b)

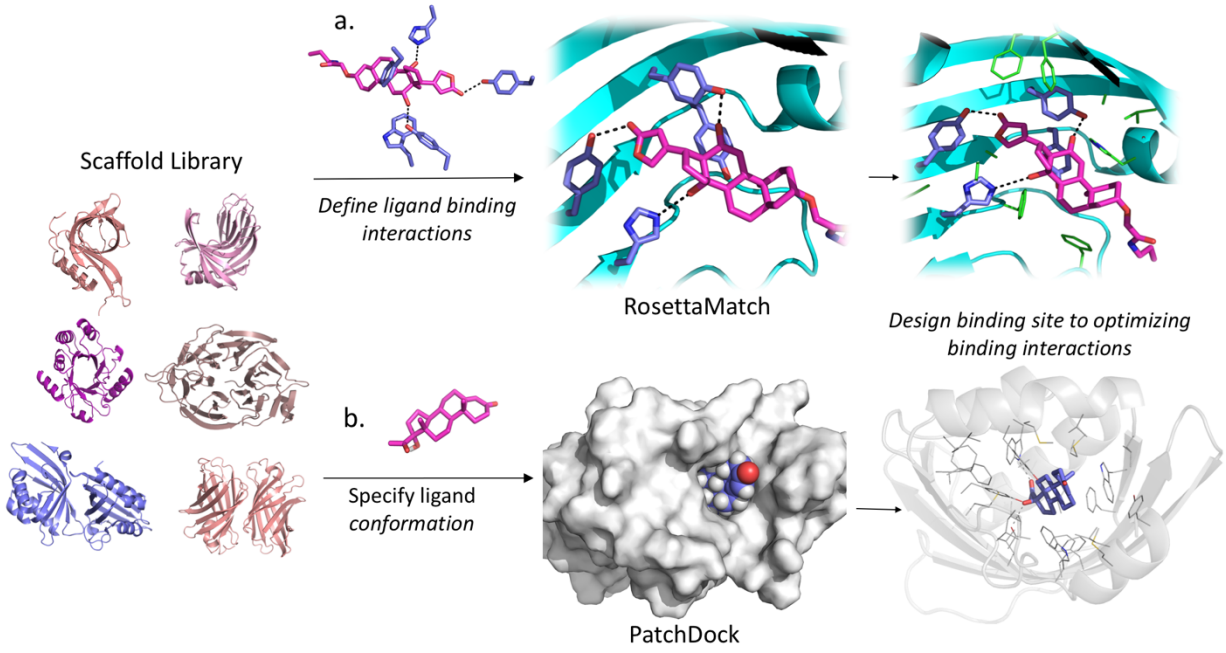


Figure 1.2. Computational methods for re-purposing natural ligand-binding proteins. **a.** RosettaMatch-based design protocol; **b.** PatchDock-based design protocol

1.2.3 *Computational sequence designing methods*

Monte Carlo optimization algorithm is at the core of the fixed-backbone sequence design process in Rosetta[26]. A specialized program (*mover*) has been designed to sample the rotamer composition around the docked ligand to optimize the ligand interacting energy. Greedy algorithm is implemented as a local sampling method in Rosetta[52]. It tries every point mutation and combines the beneficial ones for a secondary combinatorial search. Modifications made to the energy function can sometimes prevent the optimization process from being trapped in local minimums. A damped repulsion term has been used for evaluating the binding-site rotamer composition during MC searching, followed by a Newtonian energy minimization with a full-weight repulsion term.

1.2.4 *Ligand docking and Molecular dynamics simulations for in silico verification*

To test model consistency, design models generated by Rosetta are verified by ligand docking simulation using a similar energy function. Inside the Rosetta ligand docking simulation, a coordinate grid is used to sample the ligand position and the movement is introduced as MC perturbation. Other standard small molecule docking software are used for model verification, as an independent sampling and scoring validation. Glide[53] and Vina[54] use distinct sampling and scoring methods from Rosetta, respectively. The agreement between difference simulations is often seen as an indication of effective energy optimization. Molecular dynamics (MD) simulations are also used for *in silico* verification. Since the design calculation and docking simulations are done with an implicit solvent model, the explicit water molecules in MD can help validate the model stability within a more realistic setting.

1.2.5 *Experimental validation and optimization*

Yeast display (Figure 1.3)

Yeast display is used as a cell-base assay to test the ligand binding activity, where the yeast secretion system is hijacked to display proteins of interest. By knocking out the genes encoding Aga1 and Aga2 and supplying the cells with their copies in a nutrition-selective plasmid, proteins of interest can be fused to the C-terminus of Aga2 gene and become displayed by the secretion system and forming complex with Aga1 on the cell surface. The special yeast strain EBY100 and supplementary plasmid pCTCON were obtained from Wittrup lab[41]. Yeast cells displaying the designed proteins can be labeled by fluorophore-conjugated antibodies to test their expression and binding by flow cytometry. Together with Aga2, the design protein contains a C-terminus Myc peptide that can be labeled by fluorescein isothiocyanate (FITC)-conjugated anti-cMyc antibodies. To interrogate the binding activity, the target small molecule in conjugation with biotin is non-covalently bound to streptavidin fused with fluorescent protein phycoerythrin (PE) (Figure 1.3a). In this way, the protein display is represented by FITC fluorescent signal and the binding event is correlated to PE fluorescence signal (Figure 1.3b). Coupled with yeast display and fluorescence labeling, fluorescent-activated cell sorting (FACS) can be used to select out the protein designs with different binding affinities based on their fluorescence intensity (Figure 1.3c). Complex library of mutants can be quantitatively interrogated by FACS and analyzed by NGS sequencers.

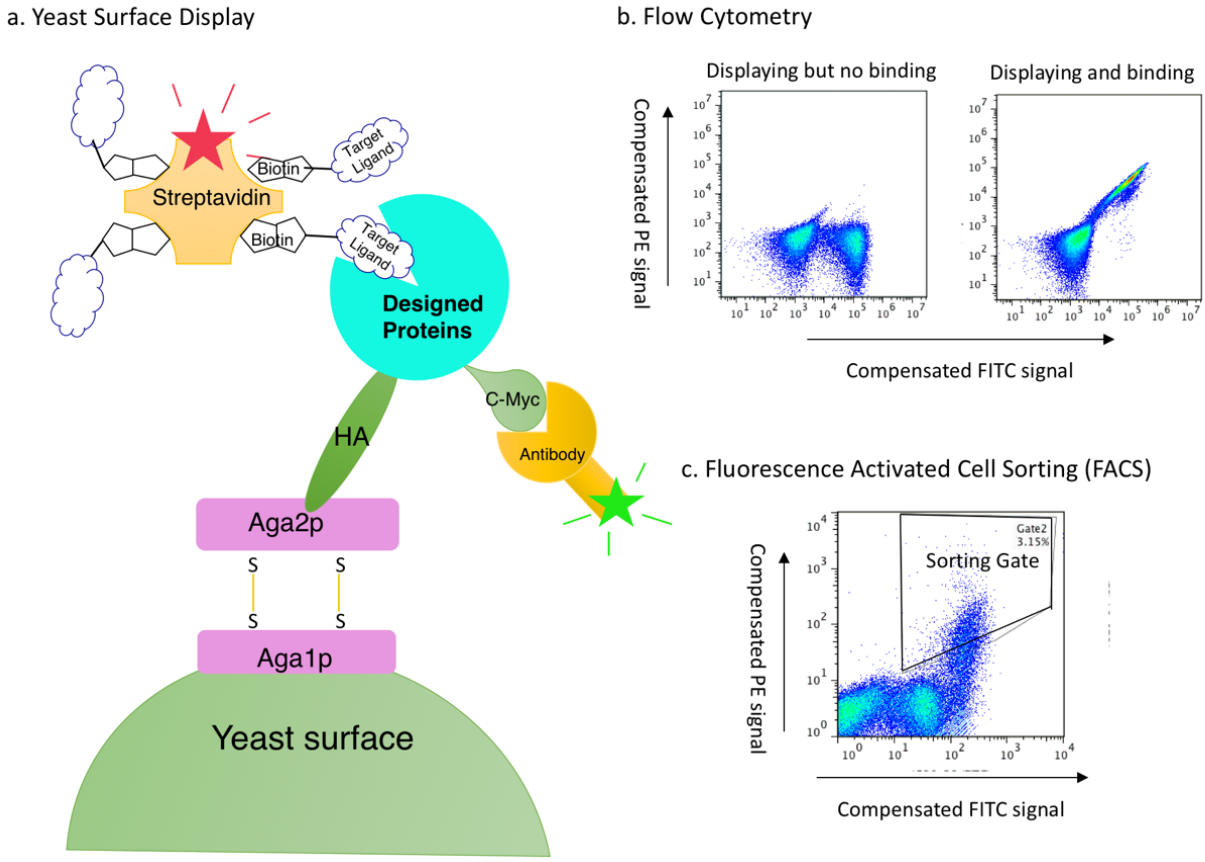


Figure 1.3. Yeast surface display combined with flow cytometry assays for detecting ligand-binding proteins. **a.** Yeast display system with labelling antibodies for detecting ligand-binding activities. FITC-conjugated anti-cMyc antibody (with the green star) is used for detecting display level; PE-fused streptavidin is used for detecting the binding event mediated by biotinylated ligand. **b.** Flow cytometry for detecting FITC and PE signals. **c.** FACS can be used for selecting the binders from a library of protein variants.

Protein production

Standard *E.coli* protein expression and purification is used for producing designed proteins for biochemical characterization. T7 expression system including *E.coli* BL21 strain and pET-serial plasmids with built-in His tag are used for combinatorial expression. Cells are cultured batchwise in LB medium to the exponential growth stage with O.D. of 0.6-0.8 at 37C before IPTG is added to induce the protein expression at 18C overnight. Cells are collected by centrifuging at 8000rpm

followed by sonication in the presence of protease inhibitor phenylmethylsulfonyl fluoride (PMSF) and DNAse. Protein purification is carried out by flowing supernatant through Ni-NTA gravity column(Qiagen) and eluting with buffer containing 200mM imidazole. For crystallization application, eluted proteins are further purified by sizing-exclusion column in a Fast Protein Liquid Chromatography (FPLC) machine.

Biophysical characterization

For experimental characterization of designed ligand-binding proteins, three techniques are generally used to characterize the dynamics as well as the kinetics of the protein-ligand interaction. They are: 1. *Isothermal titration calorimetry* (ITC) is the most widely accepted technique to determine the thermodynamic constants of a binding reaction. It measures the heat change associated with reactions in solution at a constant temperature. Thermodynamic parameters can be determined by a sequential addition of ligand to the protein solution. Binding constants can be derived from the resultant titration curve. 2. *Surface plasmon resonance* (SPR)-based biosensors, such as BIAcore, are commonly employed to determine kinetic constants for biomolecular interactions. Measuring binding reactions using a SPR biosensor requires one of the binding partners to be immobilized onto a surface. The second reactant is then flowed across the surface, thus the interaction of the soluble reactant with the immobilized partner is observed continuously and directly. The mechanism of detection is based on the fact that the adsorbing molecules changes in the local index of refraction, changing the resonance conditions of the surface plasmon waves. Quantitative kinetic data can be obtained from the primary response from which binding constants can be derived. 3. *Fluorescence polarization* (FP) provides another way to quantify ligand-protein association[55]. FP assay is applicable to any purified ligand-binding site for which an appropriate

fluorescent ligand is available. FP measures the light emitted from a fluorescent ligand in two planes (horizontal and vertical) after excitation with plane-polarized light in one of these planes. The speed at which the flourophore tumbles determines the ratio of emitted light monitored. The binding of fluorophore to the protein confines its tumbling and shows as signal in the detection. It is a real-time measuring process and easier to expand to high throughput.

1.3 RESULTS

1.3.1 *Steroids as a model system*

Steroid compounds are widely seen in nature as well as in synthetic drugs. The structural rigidity and diversity make them the ideal targets for study protein-ligand interactions. The signature four-ring core provides a relatively large surface for making hydrophobic interactions; the polar branches from the core structure can be used for exploring hydrogen bonds and binding specificity (Figure 1.4). In fact, steroids were used as conjugated haptens for early-day antibody generation and the following structural studies have offered insight on general ligand-protein interactions[56]. Besides being a model example for testing the computational design methods, designed protein binders for steroid hormones can be useful as sensor proteins for prognosis and diagnosis. Together with the newly-developed fluorescence detection systems[29], smaller hormone binders with high affinity and specificity are desired protein tools to monitor the concentration of steroid hormones in blood samples.

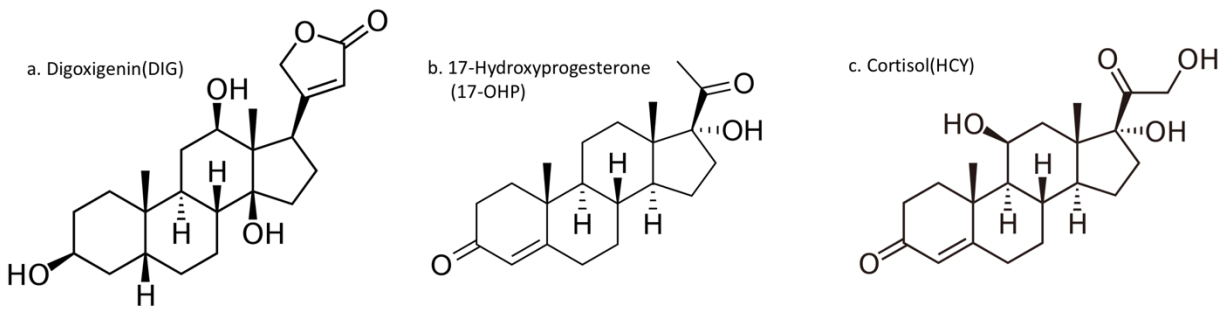


Figure 1.4. Steroids as model system for computational design of ligand binding **a.** Digoxigenin (DIG) is an anolog of a cardiac drug with a narrow therapeutic window; it is also used as a biochemical marker for DNA labeling. b.17-hydroxylprogesterone (17-OHP) is a natural hormone that has been identified as a biomarker in the diagnosis of congenital adrenal hyperplasia. **c.** Cortisol (HCY) is widely used as a biomarker for measuring exposure to chronic stress.

Design and Characterization of Digoxigenin Binding Proteins

The steroid digoxigenin (DIG) (Figure 1.4.a) is the aglycone of digoxin, a cardiac glycoside used to treat heart disease, and a commonly used non-radioactive biomolecular labeling reagent. Anti-DIG antibodies are routinely administered to treat overdoses of digoxin[31], which has a narrow therapeutic window, and are used widely to detect biomolecules in applications such as fluorescence *in situ* hybridization[57].

Computational Design

The scaffold library used for designing DIG-binding sites was inherited from the previous enzyme design work[17]. RosettaMatch was used to place DIG into the natural binding sites together with pre-defined hydrogen bonding and hydrophobic interactions, for which we examined the crystal structures of anti-DIG antibody (K_d~0.1 nM, PDB ID:1IGJ) and an engineered DIG-binding lipocallin (K_d~ 30.2 nM, PDB ID:1LKE) (Figure1.5). DIG-binding lipocalin uses histidine and

tyrosine for hydrogen bonding and phenylalanine and tryptophan for hydrophobic packing (Figure 1.5a); in contrast, anti-DIG antibody does not make polar contact with DIG and the binding interface is made of mainly hydrophobic contacts (Figure 1.5b). Three different sets of ligand interactions were defined for RosettaMatch, each of which utilize tyrosine, histidine, and asparagine for hydrogen bonding interaction and phenylalanine and tryptophan for hydrophobic packing. After the first round of searching by RosettaMatch, several protein folds were found producing more matching solutions than other folds. A focused homologous search was done by using Dali server[58] to further enrich those preferred scaffolds. (Dali server provides a network service for comparing protein structures in 3D). All 114,680 "matchers" were used for energy-based sequence optimization. During Rosetta sequence design, the "matched" residues were fixed in the first round of optimization; surrounding residues were optimized to better accommodate both the ligand conformation and the rotameric state of matched residues. In the second round of sequence design, all the pocket residues were allowed to change and the new design models were examined by hydrogen bonding satisfaction. We did the second round of sequence optimization to release the bias introduced by the limited number of pre-defined interactions. In the end, the top 599 designs ranked by interface energy and shape complementarity were inspected manually. Since previous designs made for serotonin (unpublished results in Baker lab) suffered expression problems and insolubility, some "revert-to-native" mutations were introduced manually to maintain the folding stability of scaffold proteins. In total, seventeen designs were ordered as synthetic genes from Genscript (New Jersey) and tested for DIG binding by yeast display.

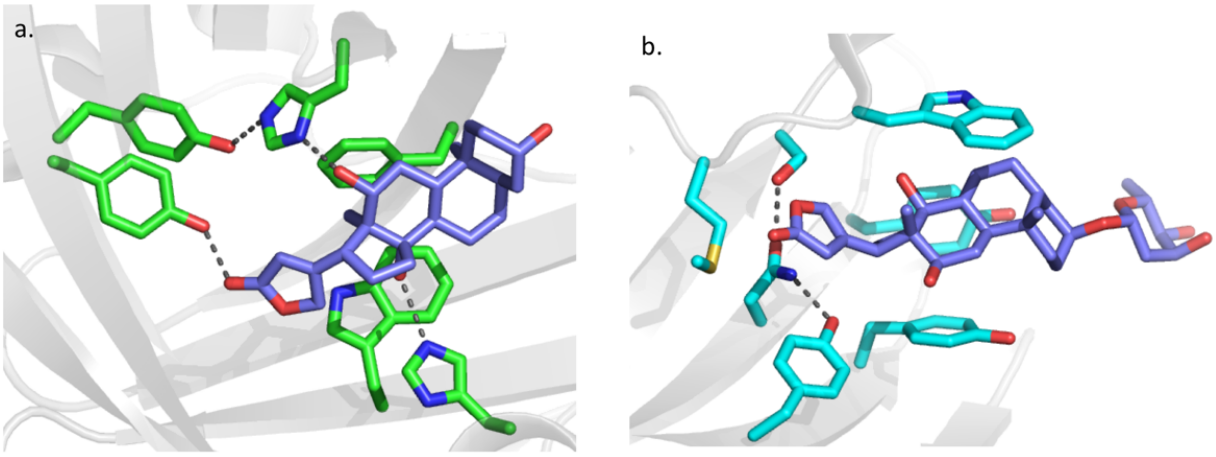


Figure 1.5. Binding sites of DIG-binding proteins in existing structures. **a.** The binding site in DIG-binding lipocalin (PDB ID: 1LKE). **b.** The binding site in anti-DIG antibody (PDB ID: 1IGJ).

Experimental Characterization

Among the seventeen protein designs ordered for binding DIG, three of them showed a binding signal on the yeast surface: DIG5, DIG8, DIG10. Notably, both DIG5 and DIG10 were designed from the same natural protein with unknown biological function (PDB ID: 1Z1S). DIG5 appeared red when purified from *E. coli* Terrific Broth (TB) culture. UV-Vis absorption spectrum indicates that DIG5 could bind enzyme cofactor heme with an incorporated iron (Figure1.6).

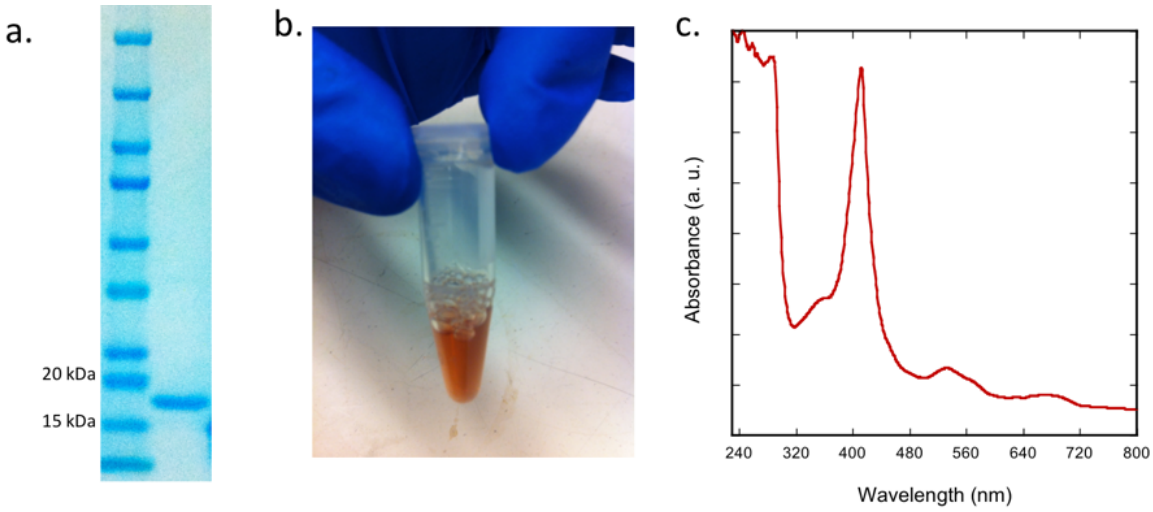


Figure 1.6. DIG5 binding heme co-factor from *E. coli*. **a.** Protein purification gel of DIG5 protein after Ni-NTA column. **b.** Purified DIG5 protein in an Eppendorf tube appearing red. **c.** UV-Vis absorption spectrum of purified DIG5. The absorption peaks at 400nm and 550nm and a tail around 650nm match the signature of heme-iron absorption.

Point mutants that knockout each individual designed interaction were constructed and tested to perturb the binding site (Figure 1.7). Site-Saturated Mutagenesis (SSM) libraries that contains all the possible single mutants in the binding site were constructed and tested for all three confirmed DIG binders. After two to three rounds of fluorescent labeling and FACS selection, cells displaying the highest binding fluorescent signal were plated on the selective agar plates. Yeast colonies were sent to Genewiz (South Plainfield, NJ) for Sanger sequencing. Beneficial mutations that increase the binding signal were further combined to construct a combinatorial library.

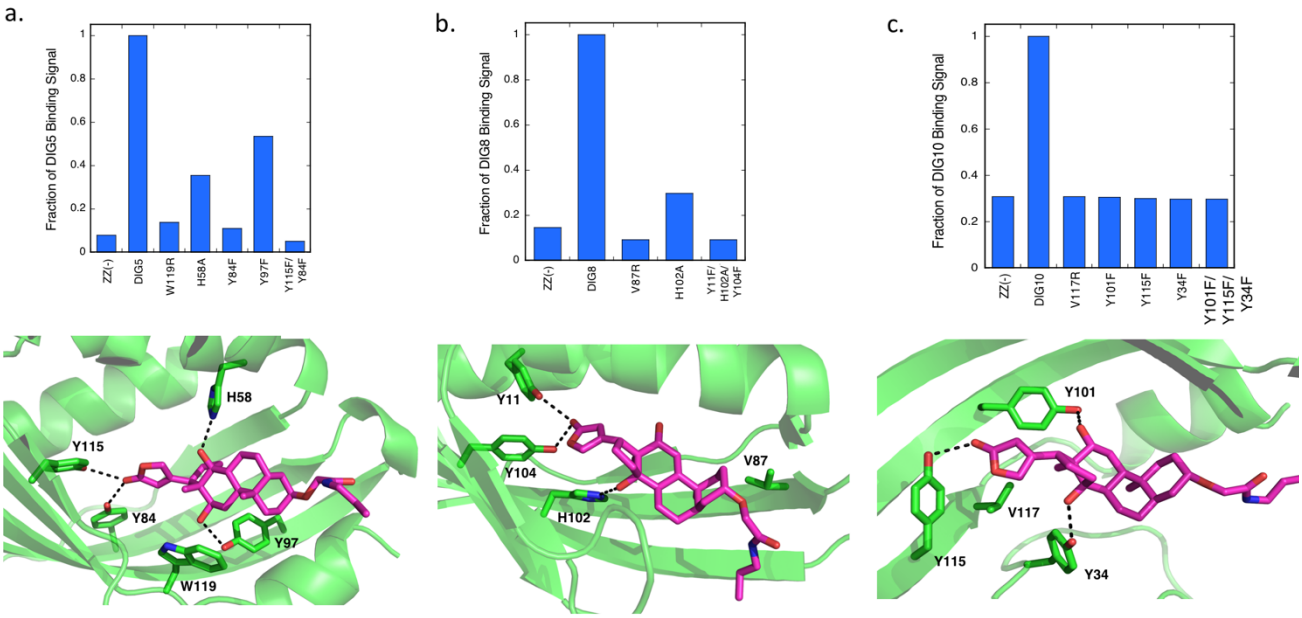


Figure 1.7. Mutational binding data of DIG5(**a**.), DIG8(**b**.) and DIG10(**c**.) The decreased binding signals indicate that the designed binding residues contribute to the binding activity.

For DIG5, the SSM library was able to identify one mutant that Y115L that correct the original hydrogen bonding pattern that appears to cause repulsion between two Tyrosine residues (Figure 1.8a). By combining all the top beneficial mutations, the DIG5 combinatorial library was built with Y115L. DIG5 combinatorial library was sorted to convergence and the best variant DIG5.1 with four additional mutations was identified (F34Y, V64W, F101Y, A104I). The beneficial mutants from DIG8 SSM library did not seem compatible with the design model. Hence DIG8 was not pursued further. For DIG10, three continuous rounds of mutagenesis and selection were performed. The directed evolution of DIG10 yielded higher-affinity binders: DIG10.1, DIG10.2 and DIG10.3. The characterization of the DIG10 series was described in our published paper[19].

Structural Validation

All three confirmed binders and their evolved variants can be purified as soluble proteins for crystallization. The co-crystal structures of DIG5.1, DIG10.2 and DIG10.3 were solved to the resolution of 2.06 Å, 2.05 Å and 3.2 Å respectively. DIG10.2 and DIG10.3 structures have been

published[19]. The structure of DIG5.1-DIG complex matches the its design model, with average all-atom and ligand RMSD values of 0.53 Å and 1.37 Å, respectively, over all four copies in the asymmetric unit. With all Y34, Y104 and Y84, adopting their predicted rotameric conformations (Figure1.8c). In all four copies of the asymmetric unit, clear density is observed for a water molecule that makes interactions with DIG lactone oxygen, hydroxyl group of S10, and pyrrole nitrogen of W22 (Figure1.8b). This water molecule adopts the role intended for the Y115 hydroxyl sidechain in DIG5 before it evolved to a Leu. In DIG5 design model, Y115 and Y84 are both designed to make H-bonds with the lactone carbonyl of DIG but this solution is likely suboptimal due to electrostatic repulsion between the lone pairs of these tyrosines in the bound state with atomic distance of 3.5 Å (Figure 1.8a). As the crystal structure has revealed, during directed evolution, this electrostatic sub-optimality is overcome by introducing a water molecule that can be placed at a longer distance from Y84 (5.1 Å) while still donating a hydrogen bond to the ligand. Water-mediated hydrogen bonds are a common feature of natural protein-small molecule interfaces, and DIG5 affinity maturation results demonstrate how they can be incorporated facilely by evolution to improve binding affinities.

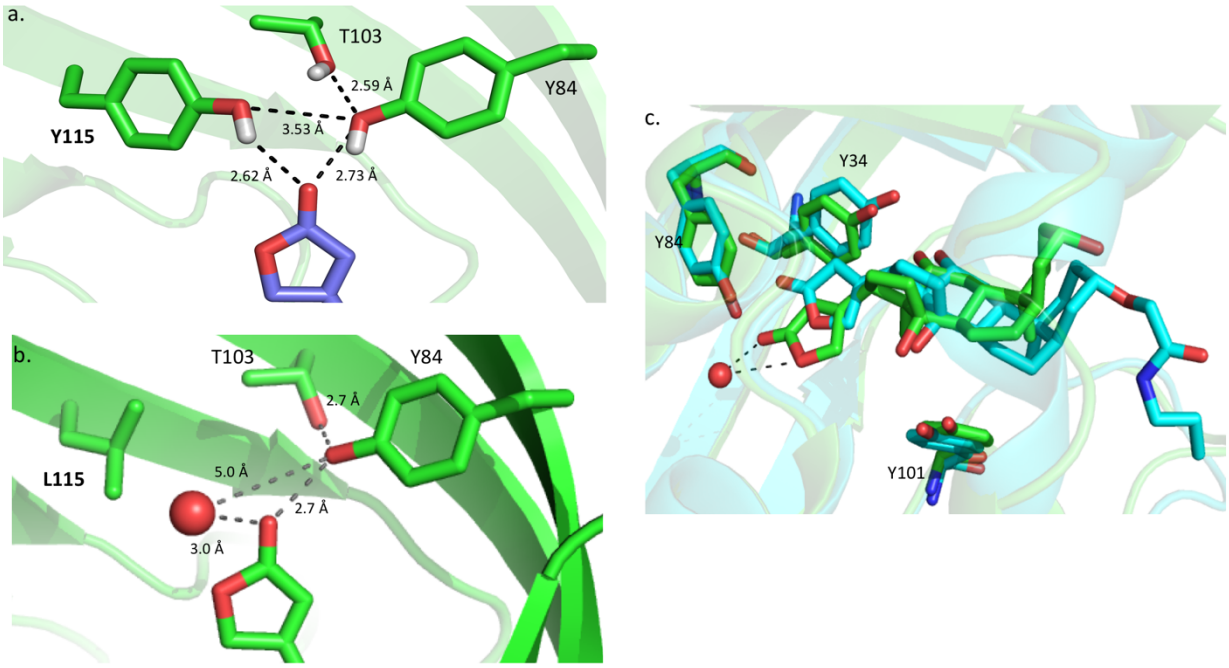


Figure 1.8. Crystal structure of DIG5.1 and a water molecule in the binding site. **a.** The designed hydrogen bonding network in DIG5 for hydrogen bonding lactone carbonyl of DIG. The distances between polar atoms are labeled with dashed lines. **b**. The crystalized binding interactions in DIG5.1. A water molecule is seen in the position of design Y115. **c.** Superimposed structures of DIG5.1 design model (cyan) and crystal structure (green).

A striking similarity between the binding sites of DIG5.1 and DIG10.2 shows up in the design models after DIG5 adopted the same tyrosine hydrogen bonding interactions (Y101 and Y34) during directed evolution. Comparison of the crystal structures of the evolved variants DIG10.2 and DIG5.1 shows a similar overall ligand binding mode with an average ligand RMSD of 0.88 Å, in spite of the 18 amino acid differences between them (Figure 1.9). In contrast, in the unevolved low-affinity computationally designed complexes, the ligand was designed to bind in very different modes (ligand RMSD = 2.46 Å) enforced by distinct sets of hydrogen bonding groups (Figure 1.9). This structural convergence suggests that while there may be multiple lower

affinity binding mode solutions, there exists a unique high-affinity binding mode solution for this scaffold-ligand pair.

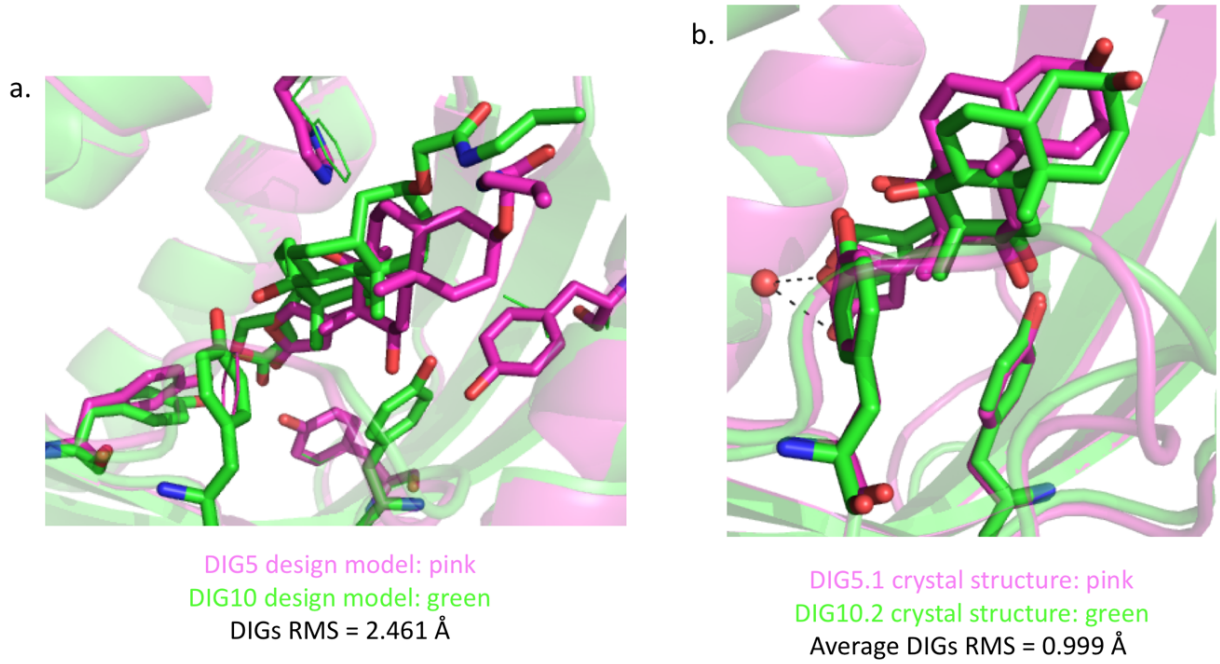


Figure 1.9.Comparison between DIG5 and DIG10 binding modes. **a.** Comparison between DIG5 design model (pink) and DIG10 design model (green). **b.** Comparison between crystal structures of DIG5.1 (pink) and DIG10.2 (green).

Design and Characterization of 17α-hydroxylprogesterone Binding Proteins

The hormone 17α-hydroxylprogesterone(17-OHP) (Figure 1.4.b) is a biomarker of a group of autosomal recessive disorders called congenital adrenal hyperplasia(CAH)[59]. Quick and accurate monitoring of blood concentration of 17-OHP is critical for disease diagnosis and treatment. Newborn screening based on elevated levels of 17-OHP is performed in many countries for early diagnosis of CAH. The most widely-used 17-OHP-based CAH diagnosis methods are radioimmunoassays, enzyme-linked fluoroimmunoassays, gas chromatography mass spectrometry(GC-MS), and liquid chromatography linked with tandem mass spectrometry(LC-

MS/MS)[60]. There is a need for highly specific 17-OHP screening methods with lower cost and improved sensitivity.

Computational Design

We used crystal structures of NTF2-like proteins from the RCSB Protein Data Bank (PDB) as starting scaffolds. In our previous work with DIG binding protein design, we used RosettaMatch to place the ligand relative to the scaffold backbone such that amino acids could be placed to make hydrogen bonds with each of the DIG polar groups. As 17-OHP has fewer polar groups, we experimented with approaches in which ligand placement is primarily determined by shape complementarity with the scaffold. We developed protocols that first, identify shape complementary placements of the ligand in the pocket and second, search for hydrogen bonds with sidechains from neighboring backbone segments (Figure 1.2b). For the first step, we used PatchDock, and for the second, either HBnet or Rosetta design allowing small perturbations of the ligand rigid body orientation. Sixteen designs with favorable binding energy and high shape complementarity were selected for experimental characterization. Synthetic genes encoding the 16 designs were obtained and the proteins were displayed on the yeast surface (Figure 1.3) Eight of the designs showed binding signal in the flow cytometry assay(Figure1.10). Mutations that knockout the designed hydrogen bonding and major hydrophobic packing interactions decrease the binding signals (Figure 1.11). Six of eight binders can be purified as soluble proteins from *E. coli*. Fluorescence polarization assay using the Alexa488-conjugated probe and purified proteins showed the binding affinity was in the low to high micro-molar range (Figure 1.12 and Figure 1.13).

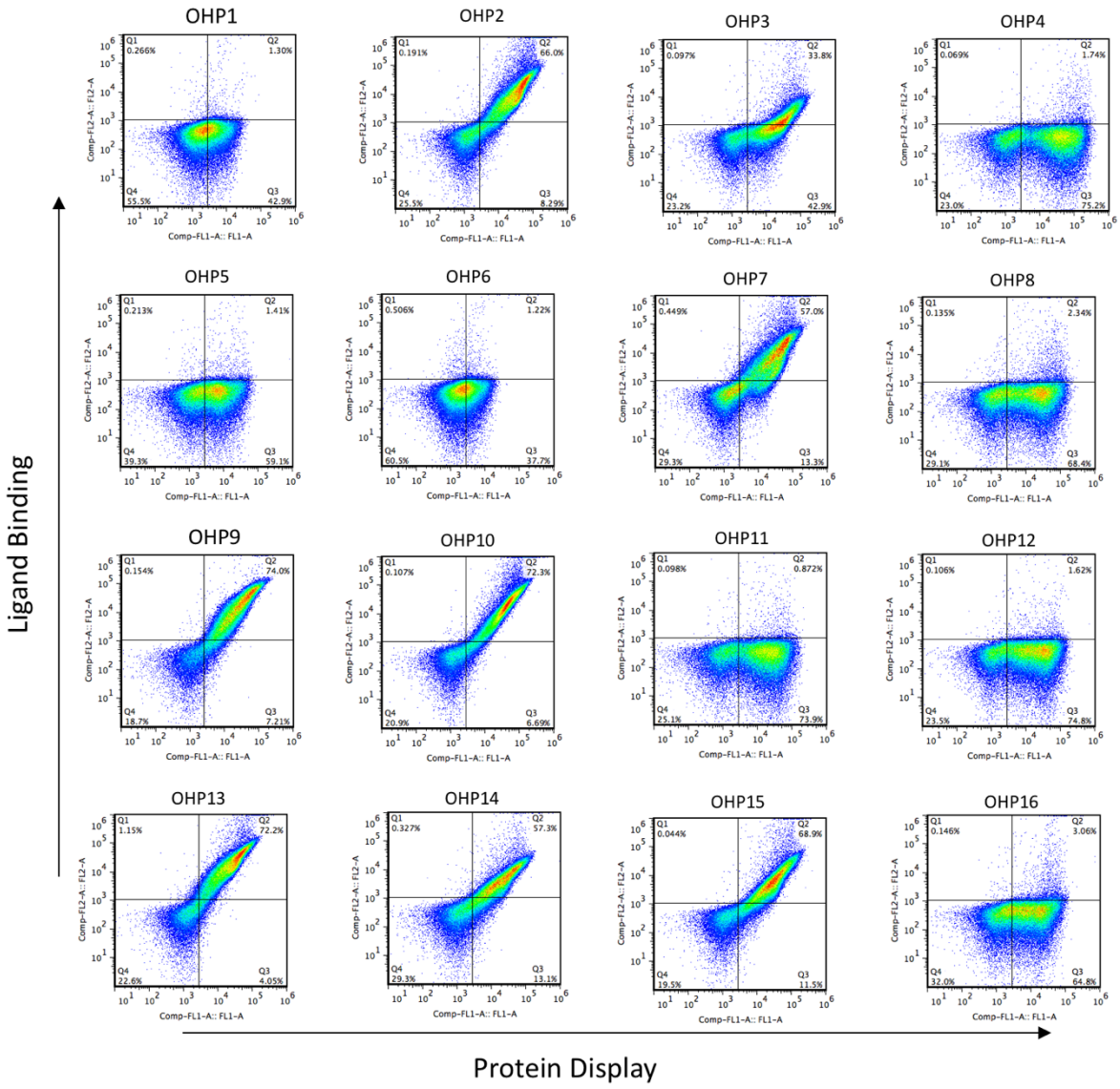


Figure 1.10. Flow cytometry binding data of sixteen 17-OHP designs

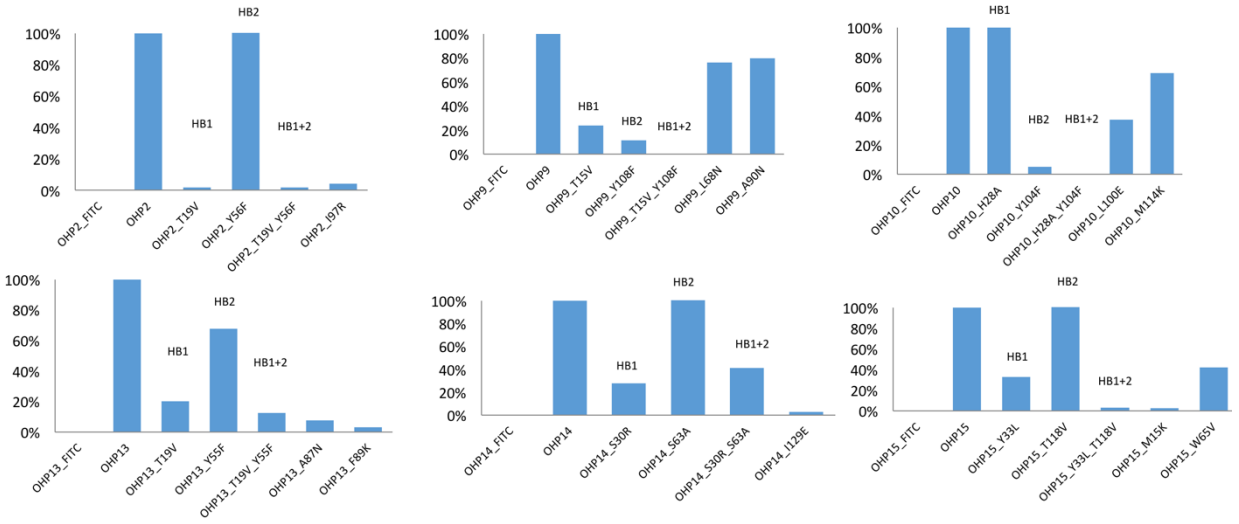


Figure 1.11. Single mutations for six 17-OHP binders. Each chart presents the relative binding signal (y axis) of individual mutations of OHP2, OHP9, OHP10, OHP13, OHP14 and OHP15, respectively.

OHP9 Crystal Structure and Binding Landscape

OHP9 is based on the *Mycobacterium tuberculosis* protein RV0760, which currently has no annotated biological function[61]. Six substitutions were introduced by Rosetta into that protein scaffold in the design calculations (Figure 1.12). RV0760 does not bind 17-OHP in the yeast display assay (Figure 11). OHP9 was expressed and purified from *E. coli*, and the binding affinity for OHP was estimated by fluorescence polarization to be around 10μM. The crystal structure of OHP9 in complex with 17-OHP was solved at 2.0Å resolution. Ligand density is clear and unambiguous in all four copies in the crystallographic asymmetric unit (Figure 1.13). While the protein backbone in the crystal structure is similar to that in the design model (Cα RMSD $\leq$ 0.52Å, Table 2), there are considerable differences in both the rotameric state of binding site sidechains and the ligand placement. The four individual copies of the ligand found in the crystallographic asymmetric unit display three distinct configurations that each interact with different sidechain rotameric states and make different water-mediated hydrogen bonds (Figure 1.14). The observed

presence of several similar, but distinct binding modes within the crystal implies that sufficient energetic, and kinetic barriers (and structural differences) separate each state to drive consistent positioning of the distinct modes in the crystallographic asymmetric unit.

In the first configuration (crystal chain A), 17-OHP is flipped 180° in comparison with the design model (Figure 1.14). Thr15, originally designed to form a hydrogen bond to the ligand hydroxyl group (atom name O1) instead turns to form an inter-residue hydrogen bond with Trp23, leaving its methyl group facing the pocket preferring hydrophobic interaction (Figure 1.13). Without Thr15 hydrogen bonding donor to distinguish its two polar ends, 17-OHP's ketone oxygen atom O3 on the other end recapitulates the designed hydrogen bond with Tyr108. Two methyl groups (C12 and C16) switch their positions in the flipped orientation making hydrophobic interactions with Leu119, Phe106, Thr95 and Trp123 (Figure 1.13). For all of these residues, except Phe106, the rotameric states are modeled incorrectly in the design model. In the second configuration (crystal chain C), 17-OHP is rotated another 180° along the longer axis while maintaining the flipped orientation seen in the first configuration (Figure 1.14). The protein pocket stays almost exactly the same in the first and second configurations except slight backbone movements on Leu68 and Val102. In the third configuration (crystal chain B and chain D), the indole nitrogen proton of Trp12 faces solvent and forms water-mediated hydrogen bonds with the unpaired beta-strand backbone, leaving an open space in the pocket. Water-mediated hydrogen bonds are formed between 17-OHP and Tyr108. 17-OHP is tilted along the longer axis in comparison with the first configuration (Figure 1.13). Since the probes we used in the binding assays have a PEG linker, the open space in crystal chain B and chain D could allow the linker to exit (this is speculative since the compound used in the crystal structure determination does not have the linker). In addition to

the Trp12 conformation change, Leu68 and Ser92 adopt different rotamers in the third configuration compared with the first and second configuration.

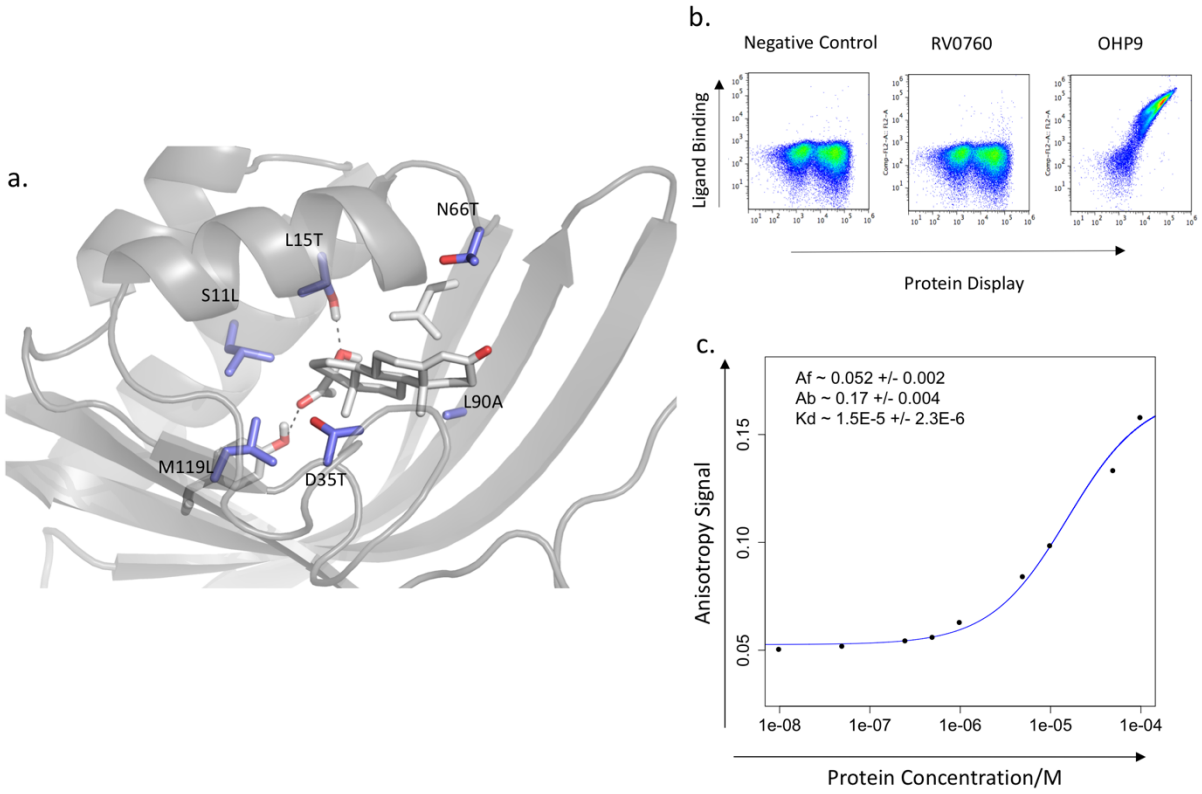


Figure 1.12 OHP9 design model and binding data. **a.** The design model of OHP9. Mutations made to the scaffold protein RV0760 are shown as purple sticks. Hydrogen bonds are indicated as dashed lines. **b.** Flow cytometry data showing OHP9 binds biotinylated 17-OHP while RV0760 does not. **c.** Titration curve of fluorescence polarization data for determining the dissociation constant (K_d) of OHP9 binding 17-OHP. The Kd value is estimated to be $15 \pm 2.3\mu M$.

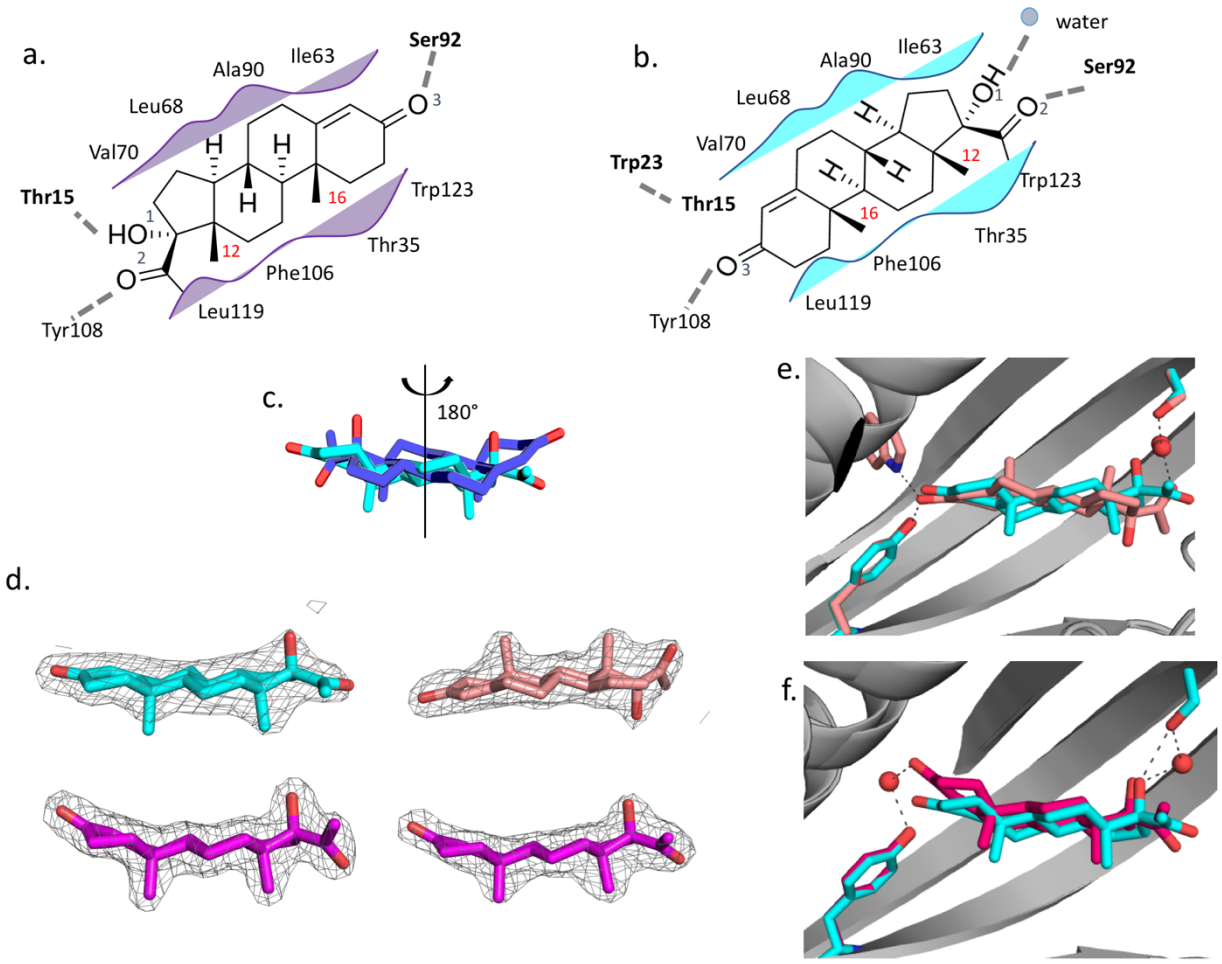
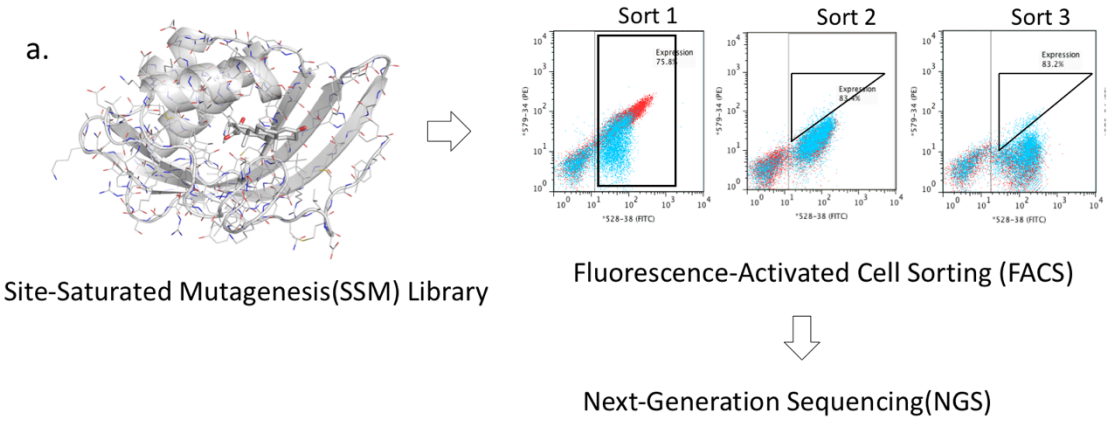


Figure 1.13 Crystal structure of OHP9 in comparison with OHP9 design model **a.** 2D representation of designed interactions around 17-OHP. Hydrogen bonding interactions are highlighted as dashed grey lines. Hydrophobic packing interactions are represented as purple shakes. **b**. 2D representation of interactions in OHP9 crystal chain A in the same fashion as in **a**. A water molecule is represented as a grey dot. Observed intra-protein hydrogen bond between Thr15 and Trp23 is highlighted. Atom O1, O2, O3, C12 and C16 of 17-OHP are labeled explicitly for direct comparison. **c**. Superimposed ligands upon aligned protein backbones. 17-OHP in design model is shown in purple. The same ligand in crystal chain A is in cyan and its orientation deviates by a 180-degree rotation. d. 2Fo-Fc density maps of four ligand copies in OHP9 crystal: chain A in cyan; chain B and chain D in magenta; chain C in salmon. **e**. and **f**. Superimposed ligands upon protein alignment. Same color representations as in **d**.

To probe the sequence-structure-function relationships underlying the three different ligand conformations, we performed a scanning mutagenesis analysis spanning the entire sequence of OHP9. Every residue in OHP9 was mutated in parallel to each of the twenty amino acids. Yeast cells displaying the mutant pool were incubated with the biotinylated probe, labeled with streptavidin-PE and anti-cMyc FITC conjugated antibody for fluorescence-activated cell sorting (FACS) (Figure 1.3). Naïve and selected pools were deep sequenced and mutation counts were summarized as enrichment values (Figure 1.14).



b.

$$\Delta \mathbf{E}^{x} = \boldsymbol{log}_{\mathbf{2}}\left[\frac{f^{x,sel}}{f^{x,unsel}}\right] - \boldsymbol{log}_{\mathbf{2}}\left[\frac{f^{WT,sel}}{f^{WT,unsel}}\right]$$

Figure 1.14. Mutagenesis Scanning experiment for mapping OHP9 binding landscape. **a.** OHP9 SSM library was constructed as described previously and genes were transformed into EBY100 via electroporation with more than 10^6 transformants. Yeast library was first sorted for protein display(Sort 1), followed by two rounds of sorting for binding and expression(Sort 2 and Sort 3). Cells were labeled with decreasing amount of 17OHP-PEG3-biotin: 5μM for Sort 1, 50nM for Sort 2, 50nM and 10nM for Sort 3, respectively. 17OHP binding incubation was followed by a quick cold PBSF washing and 10min SAPE and anti-Myc-FITC labeling. Red dots represent the flow chart of the OHP9 cells and cyan dots are for the library cells. Sorting gates

were indicated as black square or triangular windows. Unselected naïve library and selected libraries were sequenced using an Illumina MiSeq Sequencer and sequencing data were processed. **b.** The effect of the single mutant was expressed as the enrichment value ΔE^x : the log frequency of the mutant x in the selected ($f^{x,sel}$) versus the unselected library($f^{x,unsel}$), relative to the wild type(WT) OHP9 design.

The hydrogen bonding residues Tyr108 and Thr15 are highly conserved after three rounds of functional selection; major packing residues introduced by computational design also were conserved (Figure 1.15). Overall, the mutational data are compatible with both our design model and the crystal structure. Although the ligand configurations are quite different, the designed pocket residues play similar structural and chemical roles conferring binding function. Several peripheral mutations are better understood based on the crystal structure. Small amino acids (Gly and Ala) are highly preferred at position 12, where the linker of the biotinylated probe must exit from the binding pocket. The preference of position 13 for Tyr, Trp, and Phe and favorable mutations seen at position 16 might affect stability of the exposed conformation of Trp12 in chain B and chain D of the crystal structure (Figure 1.15). Using this comprehensive mutational map as a guide, we constructed two combinatorial libraries to explore the synergy between the beneficial mutations found at either the periphery or pocket positions (Figure 1.15c and Figure 1.15d).

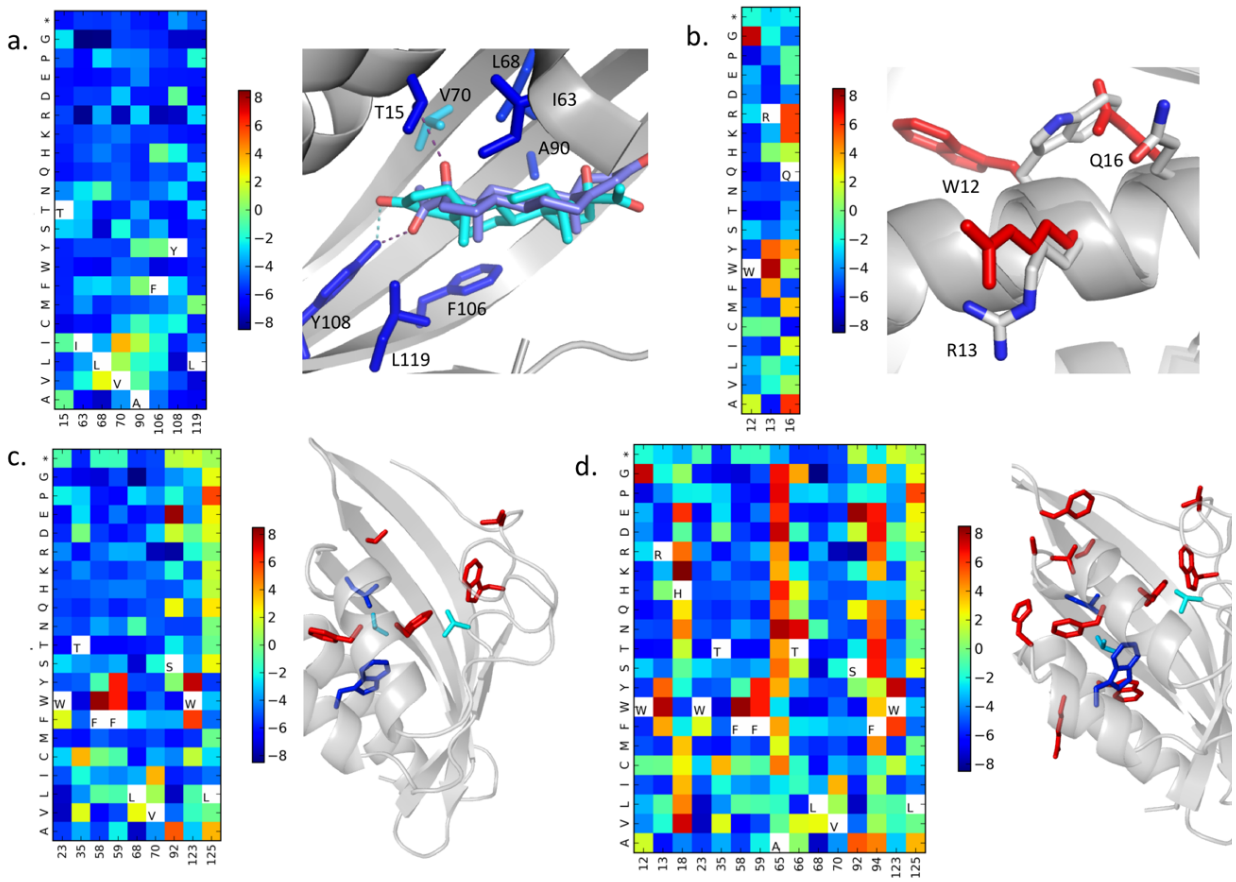
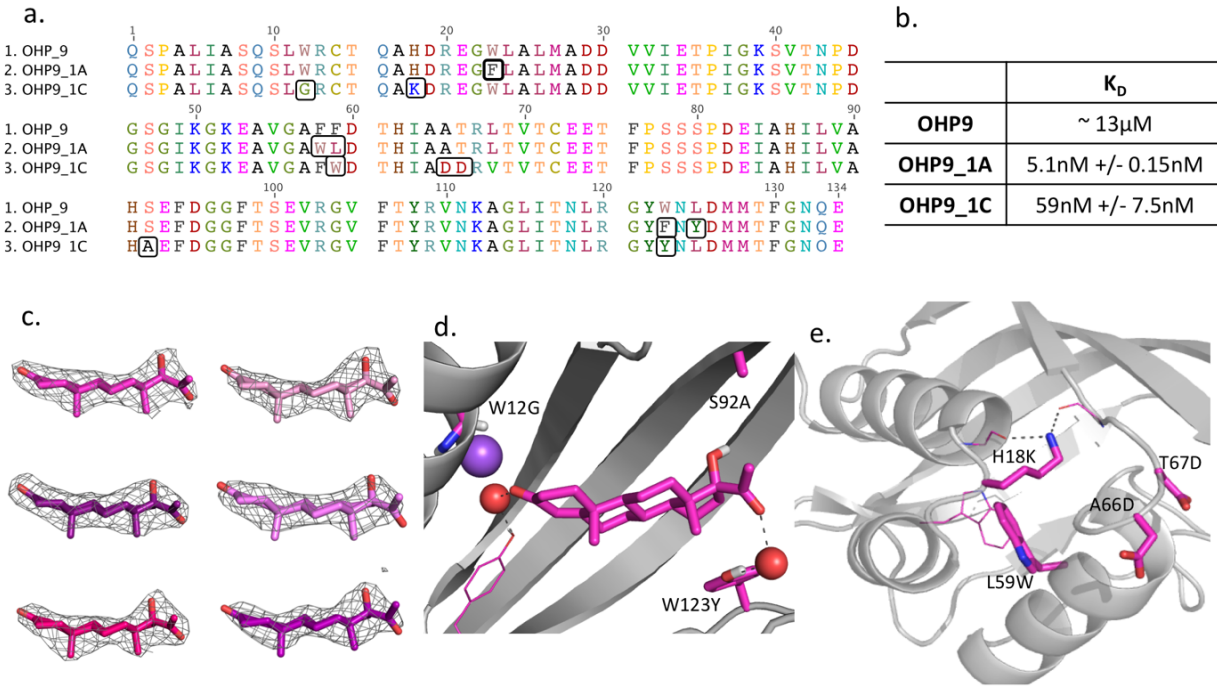


Figure 1.15. Binding landscape of OHP9. The effect of each amino acid substitution(Y axis) at selected protein positions(X axis) was assessed by calculated enrichment values(ΔE^{x}, Figure 1.14). Colored grids represent single mutant substitutes, where red and blue indicate high enrichment and depletion respectively after three rounds of selection for better binding. The initial OHP9 amino acid at each position is indicated in by its one-letter amino acid code in a white grid. **a.** Designed interacting residues in OHP9 are highly conserved during the affinity selection. Few or no substitutions are enriched shown in the colored data matrix. Residue positions are mapped on to the OHP9 structure, where design ligand(purple) and crystal chain A ligand(cyan) are superimposed, and their hydrogen bonds are indicated by dash lines in the same color. **b.** Periphery beneficial mutations that seemingly conflict with the design model (red sidechains) can be partially explained by the crystal chain B and D conformation (grey sidechains). **c.** Beneficial mutations in close vicinity to the ligand were included for constructing a pocket-only combinatorial library of OHP9. The nine positions are mapped onto the OHP9 structure. **d.** Top enriched substitutions, mostly in the periphery region of the protein, were

included for making a general combinatorial library. In total, fifteen positions were mutated all around the protein.

The libraries were sorted to convergence to identify the highest affinity binders. OHP9_1A, isolated from the pocket-mutant-only library, carries five mutations. Of particular note, Trp23 (which appears to prevent Thr15 from forming a hydrogen bond to the ligand) is mutated to Phe in OHP9_1A (Figure 1.16). Its dissociation constant (K_D) for binding 17-OHP was determined (via fluorescent polarization) to be 5.1nM, an estimated 2000-fold affinity increase from OHP9. Attempts to obtain high-resolution diffracting crystals of OHP9_1A were unsuccessful. OHP9_1C, which displayed a K_D value of 50nM, was identified from the library designed to combine all of the top beneficial point mutants. It carries seven mutations from OHP9, none of which is seen in OHP9_1A (Figure 1.16). The co-crystal structure of OHP9_1C in complex with 17-OHP was solved at 2.5Å, and the ligand was found in a configuration similar to that in the chain B and D of OHP9 crystal structure, with a water-bridged hydrogen bond between Tyr108 and 17-OHP (Figure 1.16d). All six copies of ligand in the crystal structure adopt the same configuration (Figure 1.16c). The average backbone Cα RMS change from OHP9 to OHP9_1C is 0.305Å. Mutation of Trp12 to Gly likely opens an exit for the probe linker (a sodium ion is present in the empty space in the crystal structure). The Ser92-to-Ala mutation releases a structured water molecule in OHP9. The Trp123-to-Tyr mutation enables another water-bridged hydrogen bond between Tyr123 and 17-OHP and helps lock the ligand in the same configuration in all six copies of crystallographic asymmetric unit (Figure 1.16d). The other four mutations are located at least 10Å away from the binding pocket. Lys18 caps the C-terminus of the first helix and connects the helix to the unpaired beta-strand backbone. Asp66 forms water-bridged hydrogen bond to cap the third helix (Figure 1.16d). The convergence on a single binding mode and the increase in affinity thus likely result

from a combination of backbone changes favoring the selected binding configuration and small adjustments to the ligand binding site.



	K_D
OHP9	~ 13μM
OHP9_1A	5.1nM +/- 0.15nM
OHP9_1C	59nM +/- 7.5nM

Figure 1.16. High-affinity variants based on OHP9. **a.** Sequence alignment of OHP9, OHP9_1A and OHP9_1C. Black windows mark the positions mutated in OHP9_1A and OHP9_1C. **b.** Equilibrium dissociation constants of OHP9, OHP9_1A and OHP9_1C determined by fluorescence polarization assays. **c.** 2Fo-Fc density maps of six ligand copies in OHP9_1C crystal. **d.** The converged ligand binding configuration in OHP9_1C. Mutations inside the binding site are labeled and highlighted by magenta sticks. Water molecules are shown as red spheres and the sodium ion as a purple sphere. **e.** Periphery mutations in OHP9_1C in magenta sticks mapped onto the crystal structure. Hydrogen bonds are represented by grey dash lines.

OHP9 Structure Analysis and Modeling Feedbacks

As noted above, a central flaw in the original computational design calculations was the failure to recognize that the side chain of Thr15, initially designed to form a hydrogen bond with the hydroxyl group in 17-OHP, instead can assume an alternate rotameric conformation forming an inter-residue hydrogen bond with Trp23. This failure could be due to either a backbone or

sidechain sampling problem or to energy function inaccuracy. To address this question, we systematically evaluated the energies of each rotameric state of Trp23, allowing all neighboring sidechains to reconfigure into their lowest energy states for each rotamer choice. When the backbone of the crystal structure is used in the calculations, the Trp23-Thr15 is the lowest energy state, as in the crystal structure. However, when the design model backbone was used, the Trp23-Thr15 hydrogen bond is not observed in the low energy ensemble (Figure 1.17). The shortcoming thus arises from the fixed-backbone approximation used in our design protocol: local refinement by short molecular dynamics (MD) simulation was able to sample a subtle change in the N-Cα-Cβ bond angle at position 23 which allows Trp23 to hydrogen bond with Thr15 (Figure 1.17).

To investigate whether the unanticipated intra-protein hydrogen bond between Trp23 and Thr15 can explain the lack of observation of the designed lig and orientation in the crystal structure, we performed ligand docking simulations starting with both design model and the crystal structure using Rosetta, Glide and Vina (we used Glide and Vina in addition to Rosetta to reduce bias arising from the use of the same energy function for design and docking). The docking solutions generated by Glide and Vina with the design model closely matched the designed ligand orientation, suggesting that the designed binding configuration is indeed a deep local minimum given the design model backbone (Figure 1.17c, upper panel). Starting from the crystal structure, Glide correctly generates the flipped orientation with ligand position moved 2.2Å away from the crystalized configuration; the best model generated by Vina from the crystal structure is still in the (incorrect) designed configuration. (Figure 1.17c, lower panel). Rosetta ligand docking with flexible sidechains (Glide and Vina were run with fixed sidechains) samples all three of the ligand configurations observed in the crystal structure, but the lowest energy configuration is that of the design model for both the model backbone and crystal structure backbone (Figure 1.17). In the

crystal structure backbone, the lower energy of the incorrect designed ligand configuration comes from both electrostatic and hydrogen bonding interactions.

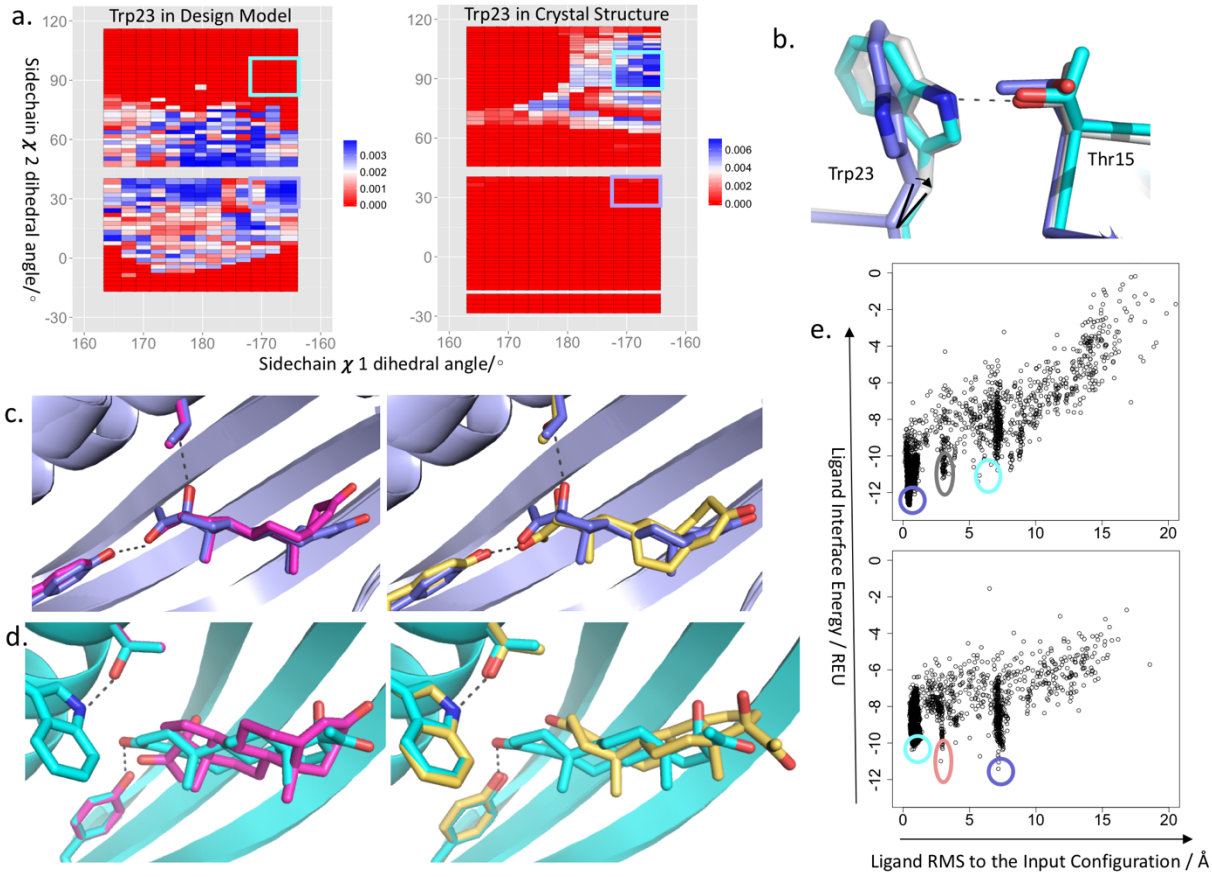


Figure 1.17. Structural Analyses based on OHP9 crystal structure. **a.** Computed sidechain rotamer distribution for Trp23 in OHP9 design model backbone and OHP9 crystal backbone. Each colored grid represents one conformational state of Trp23 with sidechain $\chi 1$ and $\chi 2$ angles indicated by X and Y axes respectively. Colorimetric scale is based on Boltzmann probability calculated from Rosetta energy term, where blue representing high-probability(low-energy) states and red representing low-probability(high-energy) states. The designed Trp23 rotamer is indicated by a purple window($\chi 1 \sim -170°$, $\chi 2 \sim 30°$), and crystalized Trp23 rotamer by a cyan window($\chi 1 \sim -170°$, $\chi 2 \sim 90°$). **b.** Change of Cα-Cβ vector in molecular dynamics simulation that captures the Trp23-Thr15 hydrogen bond. Design model shown in purple serves as the starting point for simulation; representative MD model after a short simulation shown in grey closely matches the conformations in crystal structure(cyan). **c.** and **d.** Docked ligand conformations using Vina and Glide. OHP9 design model in purple on the upper panel where

Vina ligand (pink) and Glide ligand (yellow) are superimposed with the design ligand(purple) for comparison; (**d.**) OHP9 crystal chain A in cyan on the lower panel where Vina ligand (pink) and Glide ligand(yellow) overlay with the chain A ligand(cyan). Ligand hydrogen bonds are highlighted by grey dash lines. **e.** Ligand energy landscapes generated by Rosetta ligand docking. OHP9 design model was used as the input conformation for the docking simulation summarized on the upper panel, where the purple color circles the design ligand configuration; cyan circle is close to the crystal chain A ligand configuration; grey color circles the ligand configuration that is 180-degree rotated from chain C ligand (with two polar groups inside the protein pocket); For the lower-panel docking landscape, crystal chain A was used as the input docking conformation where salmon circle represents the crystal chain C ligand configuration. The same colors are used for indicating design(purple) and chain A(cyan) ligand configurations.

OHP13 Binding Landscape

Among eight identified 17-OHP binders, OHP13 showed the highest binding affinity towards 17-OHP. The scaffold protein (PDB ID: 3T8N) is a ketosteroid isomerase with two catalytic residues Y16 and D103 mutated to Alanine[62]. OHP13 contains thirteen mutations from 3T8N and pocket mutations were shown contributing to the binding activity, except that Y55 designed for hydrogen bonding with the hydroxyl group of 17-OHP might not be the optimal solution (Figure 1.18b). The scaffold protein 3T8N when displayed on yeast surface binds 17-OHP with a lower affinity (Figure 1.18c). A competitive binding assay based fluorescence polarization was designed to test binding specificity towards progesterone missing one hydroxyl group from 17-OHP (Figure 1.18e). The results indicate that OHP13 binds progesterone with higher affinity, which seems to be compatible with the Y55F mutation data.

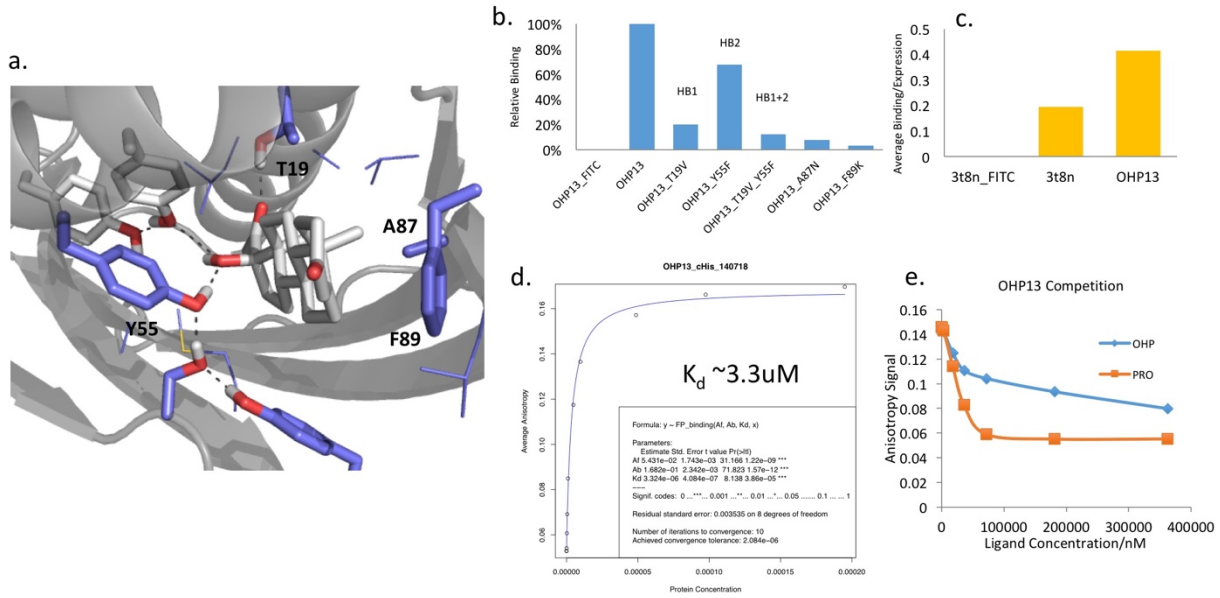


Figure 1.18. OHP13 design model and binding data **a.** The design model of OHP13. Mutations from the scaffold protein (PDB ID: 3T8N) are shown as purple sticks. Designed hydrogen bonds are represented as dashed grey lines. **b.** Binding data for functional knockout mutations of OHP13. HB1 and HB2 indicate the first and second designed hydrogen bonds shown in **a. c.** Binding data showing that the scaffold protein 3T8N binds 17-OHP with a lower binding signal than OHP13. **d.** The titration curved fitted to fluorescence polarization data for determining the dissociation constant of OHP13 binding 17-OHP. The K_d value is estimated to be $3.3\mu M$. **e.** Binding assays using free 17-OHP and progesterone (PRO) compounds to compete off the fluorophore-linked 17-OHP. Progesterone(orange) appears to be a better competitor than 17-OHP(blue).

To further perturb the local sequence space, we mapped its binding landscape by the single-mutation scanning (Figure 1.14a). The naïve and the selected libraries were sequenced on a benchtop sequencer Miseq (Illumina, San Diego). Sequencing data were processed by calculating the enrichment value for each mutation (Figure 1.14c). The designed interacting residues tend to be the optimal choice for the position except M117 can be a bigger aromatic residue Y or F. The top enriched positions were highlighted in Figure 1.19a and mapped to the design model in Figure

1.19b. G63P showed up as the most enriched single mutation and its binding affinity was confirmed as individual clone. However, OHP13_G63P also tends to bind progesterone better (Figure 1.19c).

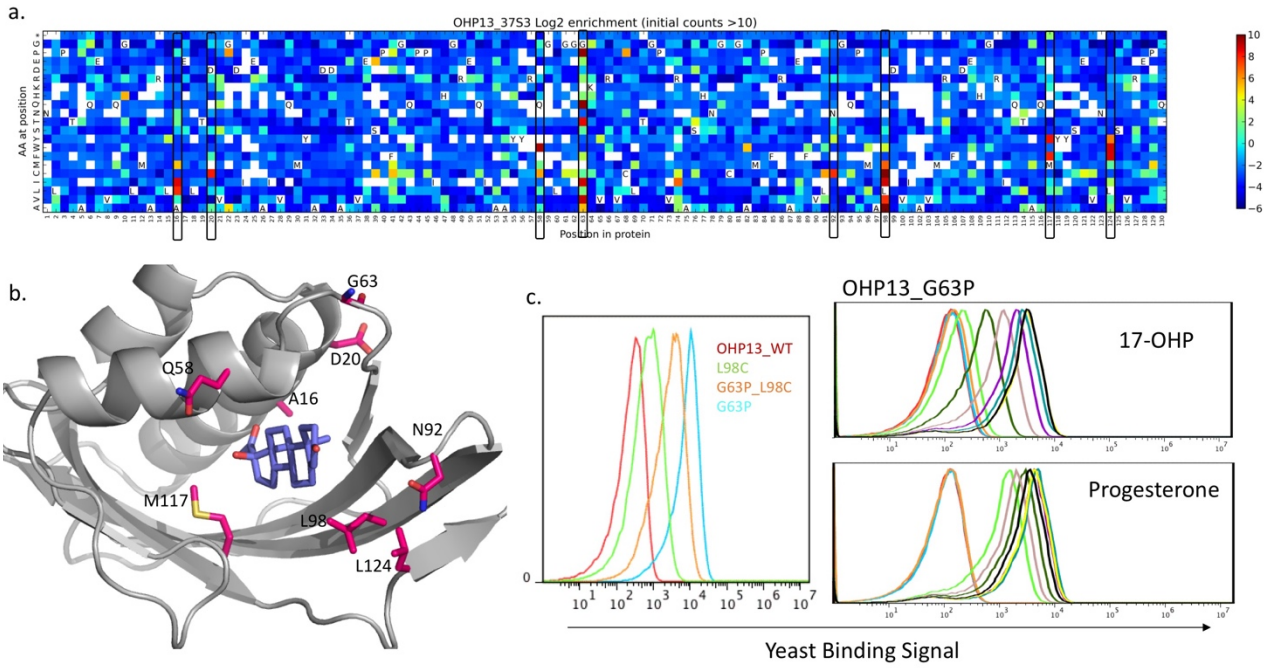


Figure 1.19. OHP13 binding landscape. **a.** The effect of each amino acid substitution(Y axis) at selected protein positions(X axis) was assessed by calculated enrichment values(ΔE^{x}, Figure 1.14). Colored grids represent single mutant substitutes, where red and blue indicate high enrichment and depletion respectively after three rounds of selection for better binding. The initial OHP13 amino acid at each position is indicated in by its one-letter amino acid code in a white grid. Mutations that do not have enough sequencing reads (10 counts from the unselected library) are represented as empty grids. **b.** The most enriched mutations are mapped to OHP13 design model as magenta sticks, which are highlighted in black windows in **a.** 17-OHP is shown as purple sticks. **c.** Binding data of the top enriched mutant L98C, G63P and their combination G63P_L98C. Flow cytometry data are shown as histograms of different colors (OHP13: red; L98C: green; G63P: cyan; G63P_L98C: orange). Titration data (right panel) show that OHP13_G63P binds progesterone with higher affinity.

Design and Characterization of Cortisol Binding Proteins

Cortisol (HCY) (Figure 1.4.c) is an important steroid hormone controlling the blood sugar level. Cortisol binders can be used for constructing protein-based biosensors to monitor stress level. Its synthetic analog, dexamethasone (DEX), is widely used in racehorses, therefore, frequently detected in anti-doping tests. Specific cortisol binding proteins can be later repurposed for detecting DEX.

Computational Design and Experimental Test

Two rounds of computational design were carried out to design a cortisol binder. In the first round, we used the design methods (Figure 1.2b) as described for OHP binding design. Six HCY conformers were generated and used during design calculation. Eighteen designs with top interface energy and shape complementarity were chosen for experimental evaluation. Synthetic genes were ordered from GEN9 (Boston, MA) and transformed into yeast cells for surface display. None of the eighteen designs showed a binding signal on yeast surface (Figure 1.20a). In a second round of design, MOAD library was used for scaffold searching. Both RosettaMatch and Patchdock were used to place the ligand into the scaffold binding sites (Figure 1.2). In total, fifty designs were ordered as synthetic genes from GEN9 (Boston, MA) and transformed into yeast cells as a pool. After fluorescent labeling for binding and expression, FACS was used to select out any potential binder. However, the design pool did not show a binding signal (Figure 1.20b).

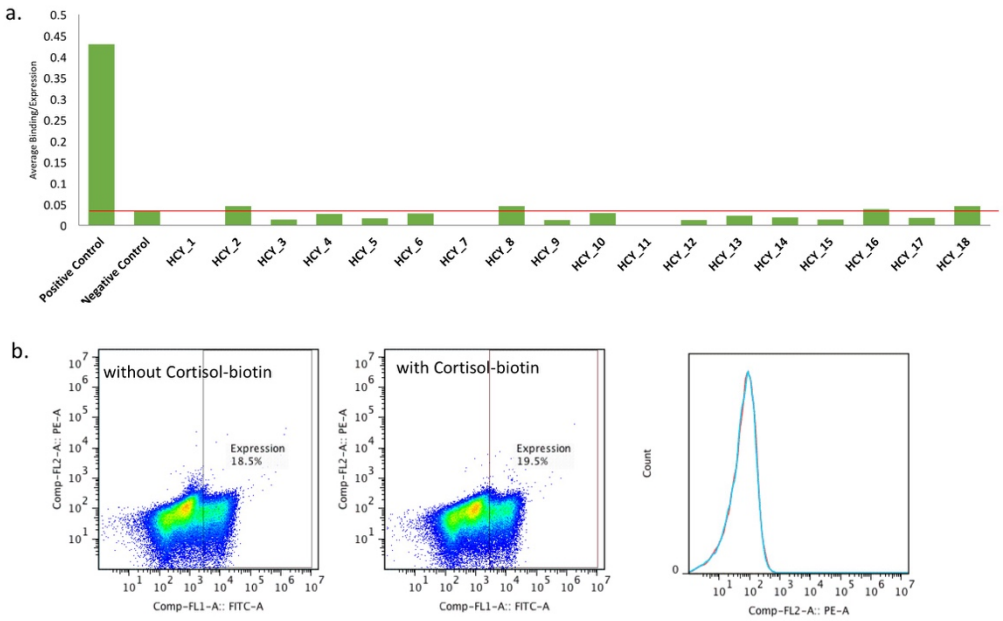


Figure 1.20. Two rounds of computational design for cortisol binding. **a.** Flow cytometry data of eighteen cortisol designs tested for cortisol binding. Red line represent the background binding level. **b.** Flow cytometry charts showing the yeast pool of fifty cortisol designs does not show binding signals above the background level.

1.3.2 *DFHBI fluorescence*

3,5-difluoro-4- hydroxybenzylidene imidazolinone (DFHBI) (Figure 1.21) is a derivative of GFP fluorophore that is commercially available and has been used successfully to screen for a fluorescent RNA aptamer[63]. It is cell permeable and exhibits negligible cytotoxicity. Notably, DFHBI is exclusively in the deprotonated phenolate form that resembles the fluorophore state of enhanced GFP. The intrinsic fluorophore of GFP is fixed in the center of a beta-can structure both covalently and via a hydrogen bonding network. The intramolecular anchoring largely reduces the radiationless relaxation by inhibiting *cis-trans* isomerization and suppressing exocyclic torsional deformations. Also, inside the protein, the fluorophore is protected from quenching by jostling water dipoles and paramagnetic oxygen molecules.

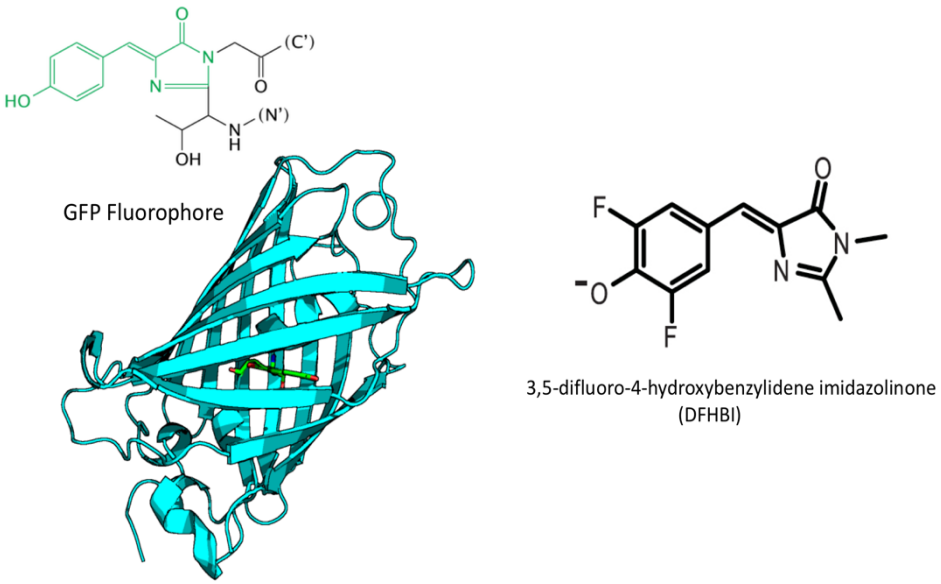


Figure 1.21.GFP fluorophore and its derivative DFHBI.

Computational Design

The same NTF2 scaffold library was used to design a non-covalent DFHBI binder to activate its fluorescence. Both Matcher and Patchdock were used to place the ligand into the native binding site of the scaffold protein. In order to activate the DFHBI through non-covalent binding, we defined four sets of anchoring interactions and used RosettaMatch to place those interacting side chains inside a recessed binding site of scaffold proteins (Figure 1.22a). As a complementary method, we also used the Patchdock to find a shape-complementary binding site to start with (Figure 1.22b). In total, 50 designs were ordered from GEN9 as synthetic genes.

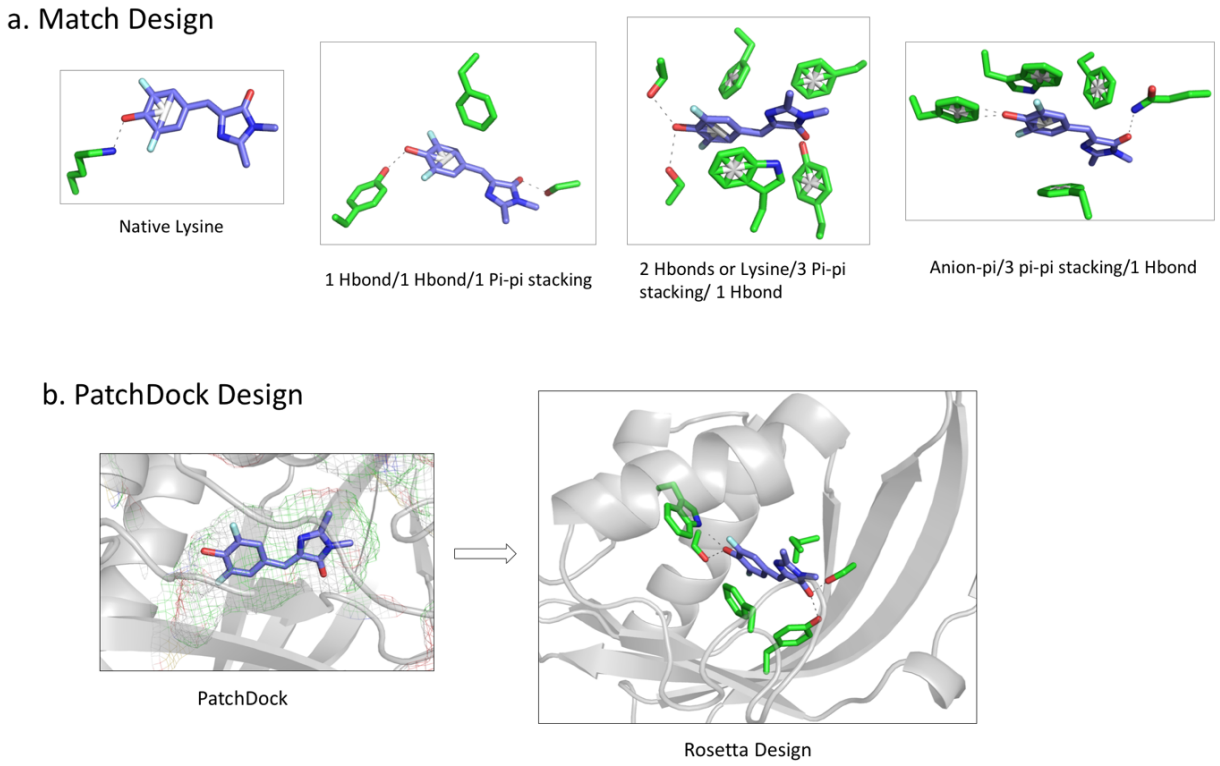


Figure 1.22.Computational design for DFHBI binding. **a.** Four sets of defined interactions used for binding DFHBI. **b.** Illustration of PatchDock design. Ligand is place into a scaffold binding site based on calculated shaped complementarity, followed by energy-based sequence design in Rosetta.

Experimental Test

Gene fragments were cloned into pET21b vector via Gibson assembly[64]. Validated plasmids were transformed into *E. coli* BL21 strain for IPTG-induced protein expression. 5mL of induced cells of each design were washed twice by PBS and re-suspended in 500μL PBS. 199μL washed cells and 1μL 100mM DFHBI in DMSO were added to each well of CORNING fluorescent plate and incubated for 2 hours with gentle shaking in 4 C° cold room. Fluorescent signals were recorded by Molecular Device M5e Plater reader with excitation wavelength at 480nm and emission at 520nm. Cells without DFHBI and free DFHBI in PBS were used as negative controls. Six designs showed positive fluorescent signals after background subtraction (Figure 1.23a). Six designs were

expressed and purified at a large scale from 1L *E. coli* culture. HBI_3 and HBI_36 were purified as soluble proteins (Figure 1.23b), but only HBI_3 could activate DFHBI fluorescence in solution.

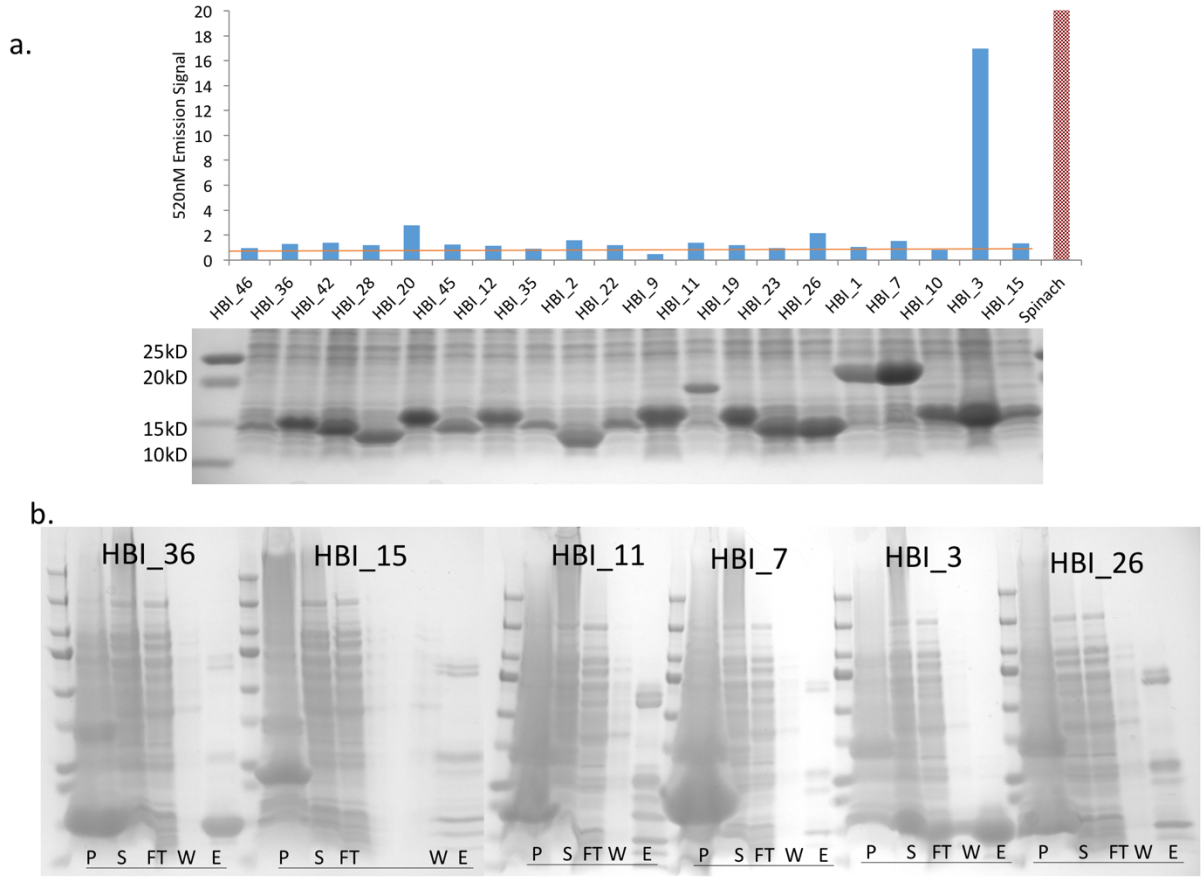


Figure 1.23.Binding results of DFHBI designs. **a.** The bar-graph representation of fluorescence signals of designed DFHBI binding proteins tested. Corresponding protein expression gel is shown on the lower panel. **b.** A SDS-PAGE gel image showing each fraction during purification (P: insoluble pellets; S: soluble supernatant; FT: Ni-NTA flow-through; W: Ni-NTA wash-through; E: Ni-NTA elution). HBI_36 and HBI_3 yielded soluble proteins.

HBI_3 is based on the *Mycobacterium tuberculosis* protein RV0760 (PDB ID: 2A15, 2Z76, 2Z77)[61]. It is the same scaffold protein OHP9 was based on and it currently has no annotated biological function. HBI_3 contains 8 mutations from RV0760 (Figure 1.22a). The ligand DFHBI was placed into the binding site by Patchdock followed by energy-based sequence design (Figure

1.22b). Binding data of the pocket mutants appear to be compatible with the design model (Figure 1.24b).

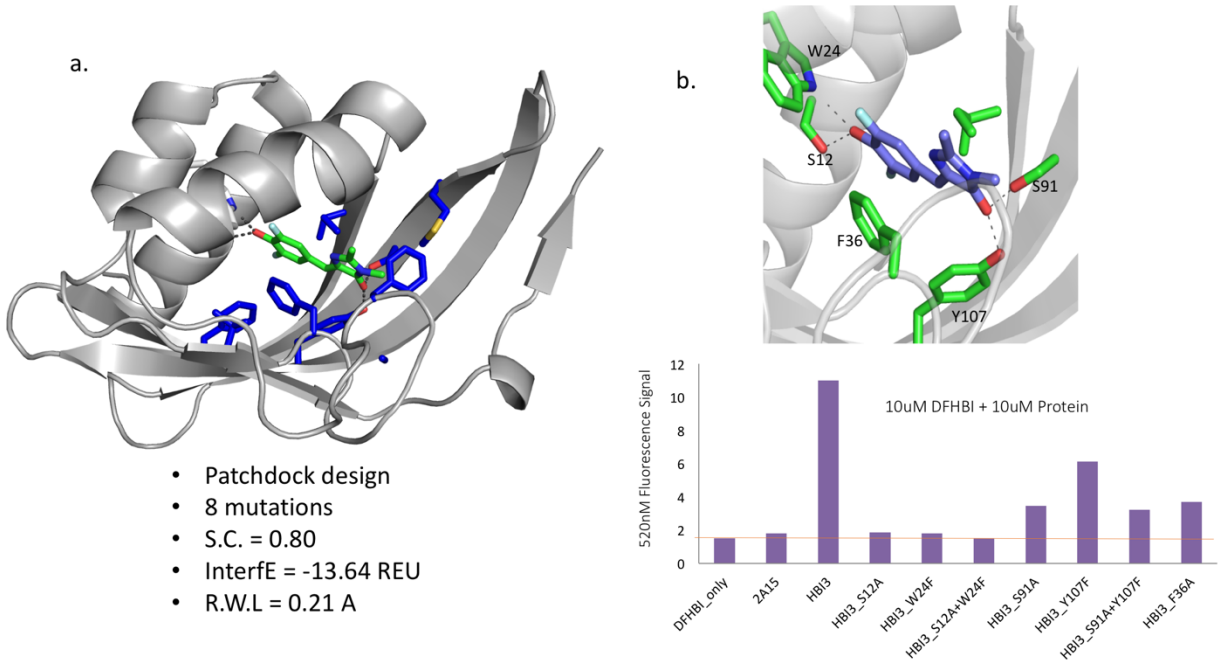


Figure 1.24. HBI_3 design model and binding data. **a.** Design model of HBI_3. Eight mutations on scaffold protein are shown as blue sticks, while ligand molecule DFHBI is shown as blue sticks. **b.** Binding site of HBI_3 with interacting residues labeled and shown as green stick. Functional knockout mutations (down panel) show lowered binding signals compared with the original design HBI_3.

Structural Validation

The gene encoding HBI_3 was amplified by PCR and inserted into pCDB24 vector with a N-terminus cleavage His-SUMO tag. A 2.9Å X-ray diffraction dataset was collected for HBI_3. The overall protein structure matches the design model. Side chain rotameric states in the binding site are clear and match the design model. However, the ligand density is not strong enough to determine the exact orientation and configuration. This might be due to the torsional flexibility of DFHBI, of which, the low-affinity binder HBI_3 cannot fully constrain the vibration.

1.4 DISCUSSION

Three steroid compounds with different polarity and shape were chosen as a model system for methodology development of computational design of ligand-binding proteins. All three targets are relatively rigid and hydrophobic. DIG has three dispersed polar moieties, for which RosettaMatch was used to design hydrogen bonding interactions. 17-OHP has two polar moieties clustered at one end of the molecule and another similar one on the other end. Cortisol has four polar moieties with three of them clustered at one end. DIG has the biggest surface area with an additional ketone ring attached to the steroid ring. 17-OHP and cortisol are about the same size (Figure 1.4). We developed two complementary computational protocols to generate customized binding proteins for each of three chosen steroids. The rich mutagenesis data and crystal structures revealed both the advances and limitations of current methods for computational ligand binding design.

1.4.1 *Success of DIG binding*

DIG10 and its evolved variants highlight the importance of binding-site pre-organization.

Comparison of the properties of successful and unsuccessful designs can verify the hypotheses underlying the design methodology. Although all 17 DIG-binding designs had high computed shape complementarity to DIG by construction, the DIG10 design, which had the highest affinity for DIG, had the most favorable computed protein-ligand interaction energy and its hydrogen bonding residues were predicted to be the most pre-organized by discrete rotameric energy calculation[19], suggesting that these attributes should continue to be the focus of future design methodology development. The emphasis of binding-site pre-organization was also seen during three rounds of DIG10 directed evolution. A Proline residue was introduced in the DIG10.2 and

DIG10.3 to rigidify a loop conformation; side-chain conformational flexibility of Y115 is selected against during the final affinity maturation[19]. Directed evolution achieved binding-site pre-organization by introducing second-shell interactions (W105 and A103) to buttress side chains making key ligand contacts. This result indicates that to construct a high-affinity ligand binding protein, not only the binding sites need to be redesigned, the structural elements and second-shell residues all need to change to accommodate the new ligand. However, natural proteins with marginal stability may not be able to afford arbitrary drastic changes.

DIG5 and DIG10 converge on hydrogen bonding patterns

DIG5 and DIG10 were designed from the same scaffold protein with distinct ligand positioning. The fact that DIG5.1 turns to adopt the hydrogen-bonding pattern designed for DIG10 indicates that DIG10 hydrogen bonding pattern represents a more optimal ligand configuration. Interestingly, the pre-defined hydrogen bonding residues H32, Y14, H124 were all discarded in the second round of design. Residues Y34 and Y101 were introduced to make new hydrogen bonds (Figure 1.25). When a new set of pre-defined interactions utilizing only tyrosines for hydrogen bonding was used for matching, the Y115/Y34/Y101 hydrogen bonding pattern can be found for the same scaffold (unpublished data). This result suggests that Rosetta energy-based design is able to find novel interactions despite the fact that the pre-defined interactions were not optimal. A general docking algorithm that does not confine the specific interactions might help Rosetta find more scaffold-compatible interactions. This idea lead to the development of PatchDock based design protocol.

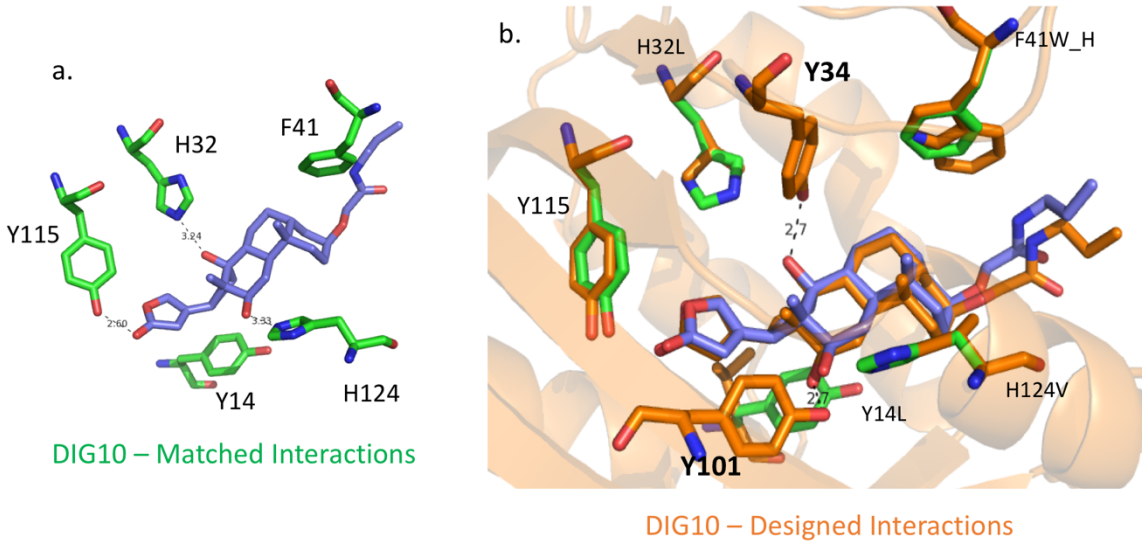


Figure 1.25.DIG10 binding pocket from Matcher to Rosetta design. **a.** The pre-defined interactions found for DIG10. **b.** New residues were introduced in the second round of design calculation. Pre-defined interacting residues are shown as green sticks; newly-introduced residues are shown as orange sticks.

Scaffold Selection

Success of DIG binders lead to the discovery of NTF2-like scaffolds. Four out of seventeen DIG designs can be expressed as soluble proteins from *E. coli*, all of which are from NTF2-like scaffold proteins (Figure 1.26). The NTF2-like fold family represents a classic example of divergent evolution wherein the proteins have the common structural details but diverge greatly in their function. The shared fold architecture contains a conical α/β roll with an N-terminal helical region packed against a dramatically curved antiparallel β sheet. Based on their structural characteristics and experimental evidence from DIG designs, we hypothesize that NTF2-like proteins can serve as stable scaffold proteins for large-scale computational protein design. This also leads to the computational design of *de novo* NTF2-like scaffolds[65].

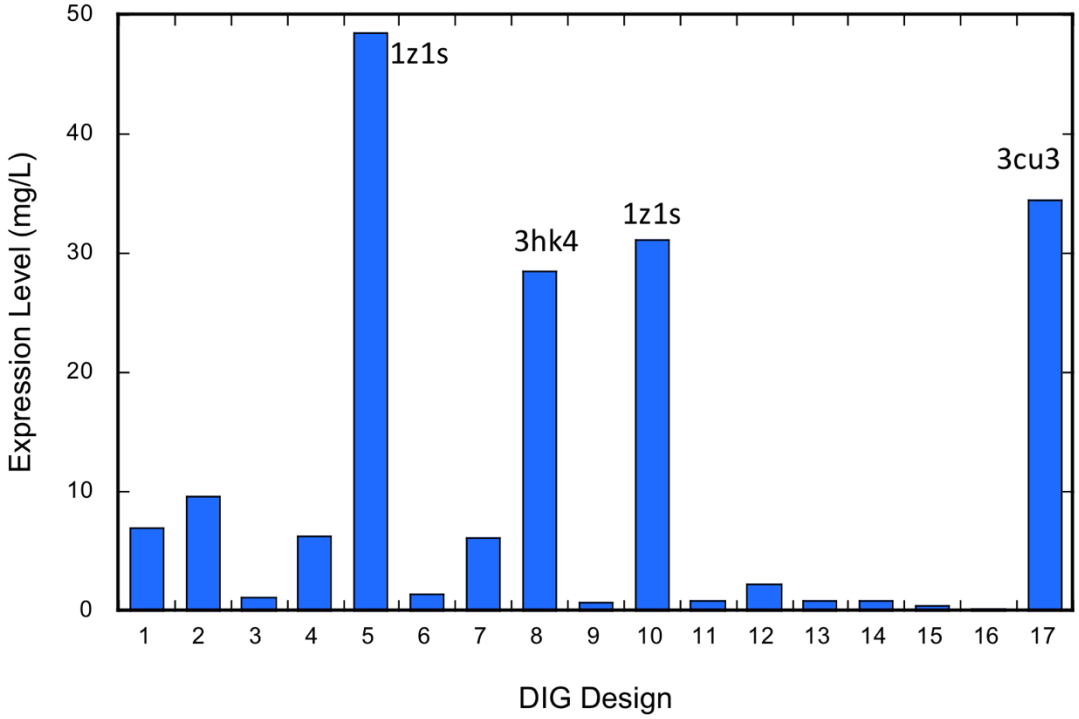


Figure 1.26.*E.coli* expression level of DIG designs. The top four designs with the highest expression level are labeled by the scaffold PDB IDs. All of them are NTF2-like proteins.

1.4.2 *Errors in OHP9 modeling*

The discrepancy between the position of the 17-OHP in the design model of OHP9 and the actual ligand positions observed in the crystal structure highlights several of the critical challenges to computational design of small molecule binding proteins. The first challenge is accurate energy evaluation. In the OHP9 case, in the design model the more polar end of 17-OHP is buried such that the polar groups make hydrogen bonds with designed side chain residues. In the crystal structure, the ligand configurations are flipped with the more polar portion sticking out into solvent, suggesting that the cost of desolvating these groups outweighs the energy gain from hydrogen bond formation. The energy function used in design clearly underestimates the desolvation penalty; this is also the case for Vina.

The second challenge is to properly model intra protein sidechain and backbone conformational changes. As noted above, a slight change of the N-Cα-Cβ bond angle of residue Trp23 allows it to hydrogen bond to Thr15 instead of interacting with 17-OHP. In addition, several hydrophobic sidechains adopt different rotamers to accommodate the configurations of the ligand observed in the crystal structure. These intraprotein conformational rearrangements disfavor the design target binding mode and favor the alternative observed modes. These challenges are particularly acute when designing binders for hydrophobic small molecules, where there are only one or two polar moieties to make directional interactions and provide "handles" for residues in the binding pocket to grab on to.

It is instructive to compare 17-OHP to DIG: the latter has two additional hydrogen bonding groups that the high affinity DIG binding exploits to achieve both specificity and precision. The general problem of specifically orienting non–polar ligand is exacerbated for 17-OHP by its near two-fold symmetry—a flip around the pseudo two-fold axis does not change the steric and non-polar interactions significantly and hence it is challenging for designs to strongly favor one of the two orientations. There are several clear routes forward. First, the error in energy evaluation of ligand-protein interactions results mainly from the improper balance between the unfavorable cost of desolvating ligand polar groups and the favorable hydrogen bonding interaction. Future work on improving the solvation model will help correct those errors. Second, improved sampling methods coupled with more computing resources and an improved energy function should be able to recapitulate the subtle backbone changes that enable multiple sidechain rearrangement.

1.4.3 *Failure to generate a Cortisol binder*

The current design methods we developed were not able to generate a cortisol binder, despite the fact that three proteins designed for binding Artemisinin bind cortisol accidentally (Chapter 3). In

comparison with DIG and 17-OHP, cortisol has the most polar groups with three polar groups condensed at one end of the molecule, which we tried to bury inside the protein binding site. As it is noted for OHP9 modeling errors, the de-solvation penalty is underestimated in Rosetta energy evaluation. Three polar side chains were introduced on the base of the binding site for hydrogen bonding with cortisol. The repulsion between the polar side chains is also beyond what Rosetta energy function can capture –-- a similar situation is seen for DIG5 Y115L mutation. Natural cortisol binders form bi-dentate hydrogen bonding interactions using Asp and Thr to satisfy all four polar groups in cortisol. This could be a new approach to make cortisol binders.

1.4.4 *Conclusion*

In this chapter, we developed two complementary computational methods for re-purposing ligand binding sites in natural proteins. Both methods start with a pre-curated library of existing structures with ligand binding sites. Those structures provide the framework for binding-site engineering; thus they are referred to as "scaffolds". The first method based on RosettaMatch is focused on satisfying pre-defined interactions and is well-suited for specifying specific coordinating interactions. However, with a limited number of scaffolds, the geometric solutions from RosettaMatch either do not exist (fail to find a matching solution) or they are not compatible with the overall protein stability of the scaffold protein, which leads to insoluble or unstable proteins. The second method circumvents this "scaffold destabilization" problem by first searching for ligand configurations that fit the native binding sites with stereo-complementarity, so that some native side chains can be re-used as anchoring points for binding the new ligand. This method usually introduces fewer mutations to the scaffold protein and yield more soluble proteins. However, to satisfy the polar groups with directional requirements become a limiting step. We were able to design binding proteins for several hydrophobic small molecules using the second

method. With limited structural feedbacks and cross-binding data (Chapter 3), it appears that hydrophobic ligands are relatively easy to gain binding affinity but the binding selectivity is hard to achieve (unpublished data from Baker lab). This is clearly exhibited in the case of OHP9 where multiple ligand configurations were observed.

The upcoming challenge for computational ligand binding design is to incorporate the structural principles of ideal nonfunctional proteins[65] into backbone sampling and explore a much larger protein conformation space that is not limited by the finite number of suitable scaffolds in nature. Together with the improvement of scoring accuracy, the iterative sampling strategy[15] can potentially drive the backbone conformation to better accommodate the destabilizing functional sites. Potentially, with more sampling control over the protein backbone and increasing scoring accuracy, a new generation of *de novo* proteins can be computationally designed with high affinity and specificity.

Chapter 2. *DE NOVO* DESIGN OF LIGAND BINDING PROTEINS

2.1 INTRODUCTION

2.1.1 *De novo Protein Design*

De novo proteins are defined as new proteins generated by computational design on the basis of physical principles with sequences unrelated to those in nature[44,66]. Computational design of *de novo* proteins serves as a rigorous test of our basic understanding of protein biochemistry and biophysics. More importantly, the ability of making arbitrary protein structures allows systematic engineering of protein-based research tools, drugs, machines and materials. In the past ten years, computational protein design has advanced to the point that many pre-defined protein topologies can be designed from scratch[66]. Unlike the fixed-backbone sequence design problem described in Chapter 1, *de novo* protein design does not start with a pre-existing backbone. Usually, a description of protein topology is required as the starting point to construct the backbone from scratch. Sequence-independent design principles regarding the lengths and positioning of structural elements (helices, strands, loops, beta bulges and glycine kinks) are often used as the hypothesized guidance for constructing the initial backbone. Currently most topologies are assembled from short peptide fragments with the exception that helical structures can be modeled parametrically. The process continues with several rounds of iteration between fixed-backbone sequence design and full-sequence structure refinement. The problem of *de novo* protein design has been defined as a search for low-energy sequence-backbone pairs in a topology-constrained conformational space[15].

As the structural principles of protein topology become manifest in the process of *de novo* design, design of functions becomes the next great challenge. With sufficient modeling accuracy and efficient sampling strategy, it should be possible now to generate customized protein binders from scratch to recognize small molecules. In theory, the innumerous topologies *de novo* proteins can sample will open new possibilities for scaffold construction for ligand binding. In reality however, success rate of *de novo* design is still quite low for most topologies and the generation of arbitrary protein structure is still beyond the design capability. Specifically, for designing a *de novo* ligand binding protein, there are multiple challenges: 1. Ligand binding sites in single-domain proteins often require recessed cavities, which will lead to instability in overall protein folding. So far, the unprecedented stability and rigidity of *de novo* proteins are mostly due to the ideality of their structures; every component in the protein is optimized to maintain the maximum folding stability, consequently, closed voids and buried unsatisfied polar groups are strictly excluded. 2. Small molecule binding design requires high precision of backbone conformation and side chain positioning (as shown in the previous work of OHP9 binder). While the computational design process reflects the sequence-structure co-dependent relationship, the flexible backbone conformation during the iterative optimization brings uncertainty to the construction of binding sites. Improved computational search algorithms with less dependency on backbone accuracy are needed to identify suitable initial backbones for ligand binding. 3. The robust generation of *de novo* protein scaffolds is still an unmet goal. To increase the structural diversity of backbone scaffolds is an absolute requirement for making ligand binding proteins for a variety of small molecules.

2.1.2 *Protein Stability versus Function*

The first challenge stated above appears to be crucial based on our initial attempt to design *de novo* NTF2-like ligand-binding proteins. While most natural NTF2-like proteins exist as dimers with relatively large open ligand-binding sites in each subunit, we were able to design fully-filled monomeric NTF2-like proteins by controlling the beta bulge positions[65]. A similar design method was used to construct binding sites in natural NTF2-like topology for various small molecules: biotin, DFHBI, nicotinamide and adenine. The scaffold library contains 2453 NTF2-like backbones constructed from peptide fragment assembly. In total, 559 designs were ordered from Agilent as pooled double-strand DNA and transformed into the EBY100 yeast strain for surface display. We used a high-throughput proteolysis assay to assess the folding stability of displayed proteins (Gabe Rocklin *et al*, accepted to Science). The protease-resistant genes were selected by fluorescence-activated cell sorting (FACS) and sequenced by benchtop sequencer Miseq (Illumina). The stability of the displayed proteins shows a negative correlation with the size of ligand binding sites and strong positive correlation with the size of hydrophobic packing core (Figure 2.1). The stable designs from the design pool tend to have very small binding sites and do not show binding signals towards the any of the small-molecule targets (unpublished results). While it is difficult to interpret the negative results for a complex problem like this, one lesson to be learned from this initial attempt is that a stabilizing mechanism is needed in the design process to ensure the protein folding before any specific binding function can be introduced.

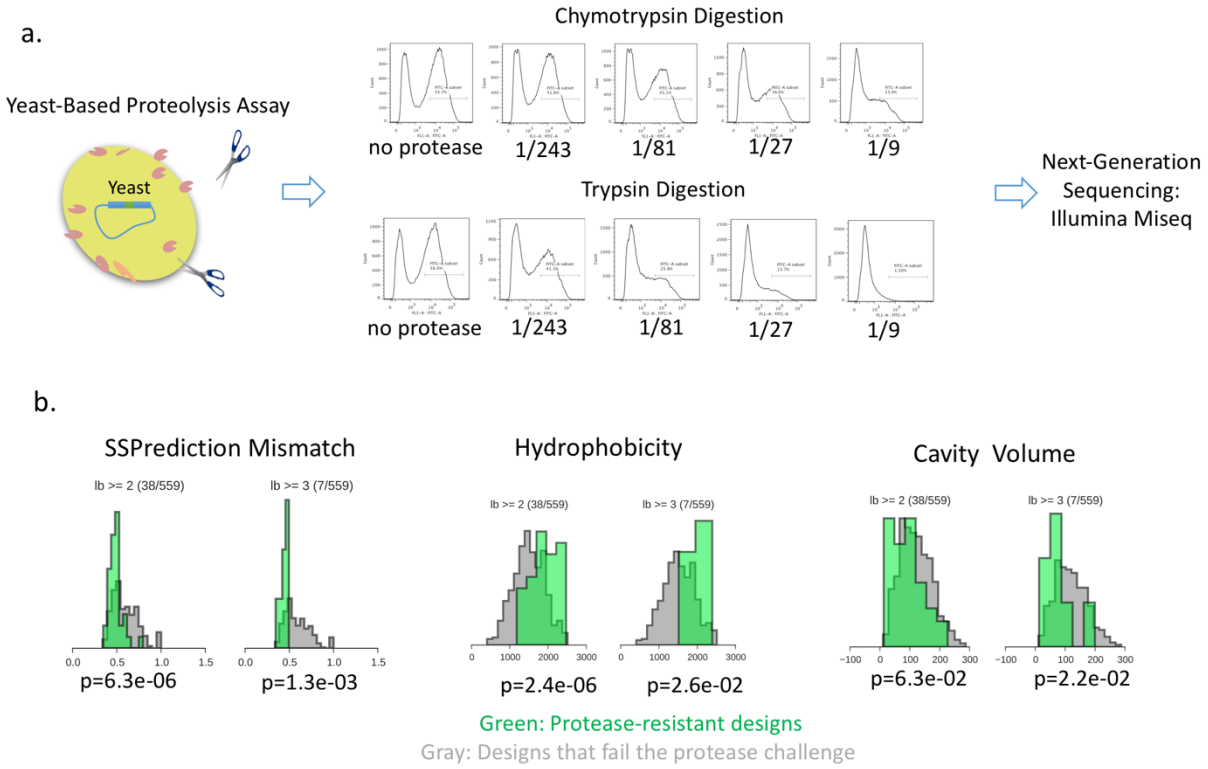


Figure 2.1.High-throughput proteolysis assay for assessing folding stability of ligand-binding designs based on *de novo* NTF2-like scaffolds. **a.** The yeast-based proteolysis assay used to assess the protein stability of designed proteins. Yeast cells displaying designed proteins are challenged with chymotrypsin and trypsin at various concentrations (dilution factors: $^1/_{243}$, $^1/_{81}$, $^1/_{27}$, and $^1/_9$, with increasing concentrations of protease). Yeast cells displaying un-digested proteins are selected using FACS and identified by sequencing. **b.** Histogram comparisons of calculated metrics for protease-resistant designs (green) and digested designs (grey). Three metrics are identified with statistically significant correlation: Secondary-structure prediction (SSPredition) mismatch, Hydrophobicity and Cavity volume.

Natural proteins reach the delicate balance of structure stability and high-efficiency function by natural evolution and the solutions are manifested in diverse formats, as Nature always does. Many studies have been done to explore the underlying biophysics of the trade-off between protein stability and function[67]. It is widely accepted that the specific organization of functionality

characteristics of natural proteins has come at a considerable cost of protein stability. In order to form a guiding hypothesis for *de novo* ligand binding design, it is necessary to review the typical stabilizing strategies Nature has explored. Natural ligand binding proteins (reviewed in Introduction, Chapter 1) display diverse structural features with distinct or combined stabilizing mechanisms, which can be summarized into three categories: 1. *Utilization of irregular loop structures without carving out the hydrophobic packing core*. Antibodies' complementarity-determined regions (CDRs) can form diverse binding sites but they do not always oppose the overall protein folding stability (with extra disulfide bonds stapling together the structures). For a stable protein structure, the longer loop regions usually fold passively and they can be used for constructing binding sites with a smaller penalty for stability; 2. *Using disulfide bonds to help protein folding*. Extracellular proteins often use disulfide bonds to help protein folding. This is especially predominant for ligand binding proteins. Both lipocallin proteins and antibodies use one or more disulfide bonds; 3. *De-localization of packing core from functional sites*. Oligomer interfaces can be a major source of folding stability[68]. Both streptavidin and avidin with extraordinary binding strength towards biotin form tetramers and it has been shown that an engineered monomeric subunit is not as stable nor high-affinity. Dimer interfaces of NTF2-like proteins are found to be crucial for its stability[69] too. The opposite is also seen in Nature; some multi-oligomer proteins form binding sites on the oligomeric interfaces while each subunit has a fully packed hydrophobic core[68]. For single-domain monomeric ligand binding proteins, an alternative packing site is often found aside from the binding site[70]. *De novo* design of ligand binding proteins can be based on an individual or a combination of those stabilizing mechanisms.

2.1.3 *Significance of designing de novo ligand binding proteins*

Computational protein design seeks to reach a full understanding and manipulation power over protein structure and function by combining computational modeling and structural validation. Specific ligand binding sites in natural proteins represent the most delicate atomic arrangement of natural evolution, which involve both local side chain preorganization and global protein stability. A general computational method that designs the entire protein sequence to fold and accommodate a ligand binding site could be very powerful for a couple reasons. First, such a method will end the longstanding dependency on natural proteins and enable protein engineers to craft new molecules for a chosen small-molecule target. Second, the customized ligand-binding proteins would lay the foundation for constructing more complicated protein-based systems like enzymes and synthetic pathways. Lastly, the ability to design weak interactions can extend our understanding of molecule association force and will have broad impact in related fields, like computational drug discovery and bio-macromolecule recognition.

2.2 METHODS

2.2.1 *Computational Methods*

Protein backbones assembled from short peptide fragments with geometric constraints were prepared and organized as a scaffold library using the published protocol[65]. A target ligand was placed into the defined binding sites using a newly developed method called Rotamer-Interaction-Field (RIF) docking. Docked ligand-backbone complexes were further designed and refined using a two-step design protocol. Detailed method development is described in the Results part of Chapter 2.

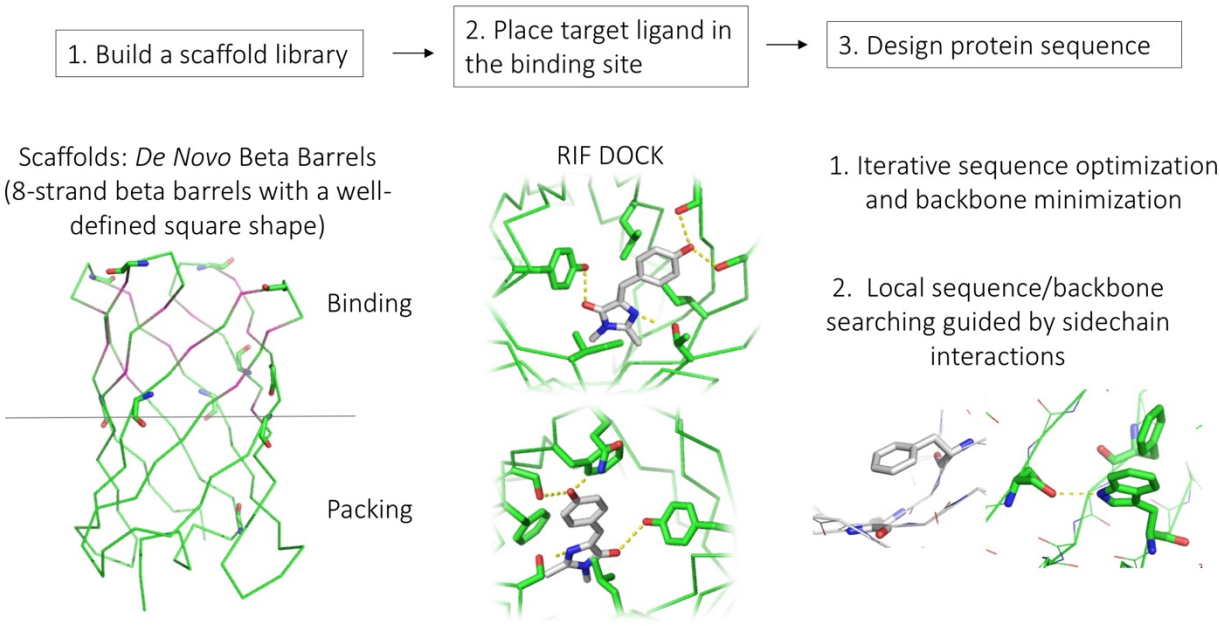


Figure 2.2. Overall scheme of design strategy.

2.2.2 *Experimental Characterization Methods*

Protein purification

Genes encoding the designed proteins were synthesized by Genscript Inc. (New Jersey) and cloned into pET29b vector with a C-terminal 6His tag and transformed into *E. coli* BL21 strain. Disulfide-bond variants were later transformed into a SHUFFLE strain (NEB) for promoting disulfide bond formation in the intracellular environment. Proteins were expressed internally in 50mL LB medium by standard IPTG induction. Induced cells were harvested by spinning down at 4000rpm in a benchtop centrifuge for 30min and homogenized by FastPrep homogenizer (MP Biomedicals). Soluble supernatants were incubated with Ni-NTA resin and purified by gravity column elution. Each fraction recovered during the purification procedure was analyzed by running SDS-PAGE gel.

Size-exclusion chromatography

Proteins eluted from Ni-NTA columns were concentrated into less than 1mL volume and loaded onto a ATKA Pure FPLC machine (GE Healthcare Life Sciences, Pittsburgh, PA). A Superdex 75 Increase™ column (GE, Pittsburgh) was used for all size-exclusion chromatography.

Far-UV circular dichroism(CD) measurement

Circular dichroism (CD) wavelength and temperature scans were recorded on an AVIV model 420 or Jasco J-1500 CD spectrometer. For thermal denaturation, protein samples were prepared at 1.0–1.5 mg ml^{-1} final concentration in PBS buffer (pH 7.0). Wavelength scans from 195 nm to 260 nm were recorded at 25 °C, 75 °C, 95 °C and again after cooling back to 25 °C. For thermal denaturation experiments, CD signal at 226nm was recorded during temperature increases. The denaturing data were fitted into a two-stage equilibrium model to estimate the T_m values.

Fluorescent Binding Assay

DFHBI fluorescent signals were recorded using Synery™ Neo2 plate reader (BioTek, Winooski, VT). Protein samples and DFHBI solution were added to a 96-well CORNING fluorescent plate to reach the final volume of 200 μL per well, and mixed and incubated at room temperature at benchtop for about 1.5 hours. Excitation wavelength was set at 450nm and emission was monitored at 500nm and 510nm. The protein-activated fluorescent signal was determined by comparing with DFHBI-only fluorescence in the same plate at the same concentration.

Isothermal titration calorimetry (ITC) measurement

ITC experiments were carried out using Microcal Auto ITC200 (Malvern Instruments, Westborough, MA). Protein samples were dialyzed overnight into PBS buffer with 2% DMSO. The dialysis buffer was used to dissolve DFHBI powder into 1mM final concentration. The initial data were obtained by titration 1mM DFHBI into 100uM protein sample in the ITC cell. The

titration contains 40 injections of 1 μL ligand. Data were processed using Origin software (Northampton, MA).

2.3 RESULTS

We aim to develop a general computational design method that utilizes *de novo* protein backbones to construct customized ligand-binding proteins. Among the diverse scaffolds that can be built from scratch (Figure 2.3), we chose anti-parallel beta-barrel topology for its potential capacity to maintain a recessed cavity. A derivative of GFP chromophore, 3,5-difluoro-4-hydroxybenzylidene imidazolinone (DFHBI) (Figure 1.22), is chosen for the proof of concept and fluorescent imagining applications. Due to its exocyclic torsional deformation, free DFHBI in solution does not exhibit effective fluorescence; a binding site reducing its torsional vibration can enable its fluorescence. A new computational docking method that constructs and places a ligand-centric "rotamer interaction field" (RIF) onto chosen scaffolds was developed and applied to the initial placement of DFHBI. We then performed Rosetta computational design calculation to optimize both the protein sequence and backbone conformation in the presence of target ligand. For the top-ranking complexes, protein models of unbound state (*Apo* models) were built and ligand docking simulations were used to check model consistency. Among fifty-six designs experimentally tested, twenty-two designs formed monomeric beta structure with one of them confirmed by X-ray crystallography to be a beta barrel; two monomeric DFHBI fluorescent binders were identified with binding affinity in the range of low-to-high micromolar.

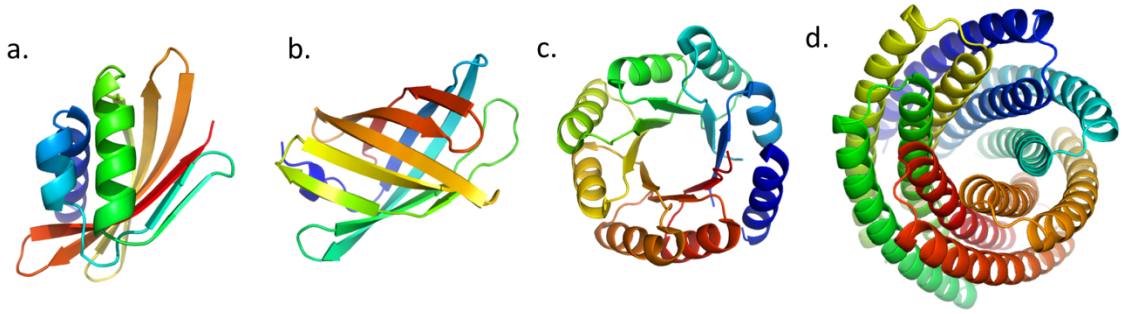


Figure 2.3.*De novo* scaffolds that can be used for small molecule binding. Examples structures for **a.** NTF2-like proteins; **b.** anti-parallel beta-barrel proteins; **c.** symmetric TIM-barrel proteins; **d.** porous helical bundles.

Small Molecule Targets

3,5-difluoro-4-hydroxybenzylidene imidazolinone (DFHBI) is a derivative of the GFP chromophore with excellent cell permeability and negligible toxicity[63]. It has two fluorine substitutions on the benzyl ring and resides exclusively in the deprotonated phenolate form that resembles the fluorophore state of enhanced GFP (Figure 1.22). Similar to the GFP chromophore, free DFHBI does not exhibit effective fluorescence due to molecular torsional deformation between its benzyl and imidazole rings. We aim to build a *de novo* protein that forms non-covalent binding interactions with DFHBI that reduce the inner-molecule torsion vibration and enable its fluorescence.

A DFHBI fluorescent binder can have broad applications in fluorescent imaging. There is a clear need of improved photoactivatable proteins for both single-molecule biophysics and high-resolution imaging fields. Small-molecule fluorophores are widely used as labeling reagents in a variety of biotechnological applications, but the lack of labeling specificity often hinders their usage in complex systems that contain multiple cell types. Natural fluorescent proteins and their derivatives need additional post-translational chemistry to mature. These processes require time

and conditions that might not be satisfied in certain experiments. It has been found that many GFP superfamily members exhibit intensity fluctuations and poor photostability owing to the dynamics of covalent fluorophore fixation in protein backbone. Also, the large size of the GFPs has been a concern for a long time. A genetically encodable DFHBI-binding protein can overcome these limitations and provide more flexibility for labeling applications.

Scaffold Selection

A number of ideal protein topologies have been designed from scratch with defined structural "rules"[16,65]. *De novo* protein design has generated porous helical bundles, concavely-curved repeat proteins, oligomers with interface grooves, and curved beta sheets, all of which have the potential to bind a small molecule (Figure 2.3). In general, curved beta structures have the geometric advantages of wrapping around a small molecule with dense side-chain interactions. This structure-function feature is seen in the major families of natural ligand binding proteins like lipocalin and NTF2-like proteins. We chose beta-barrel topology to bind DFHBI for two reasons: 1. the ideal *de novo* beta barrel topology has a relatively small size (106 amino acids) and 2. the barrel shape with big interior volume can be segmented to have a bottom packing core and an upper binding site (Figure 2.4).

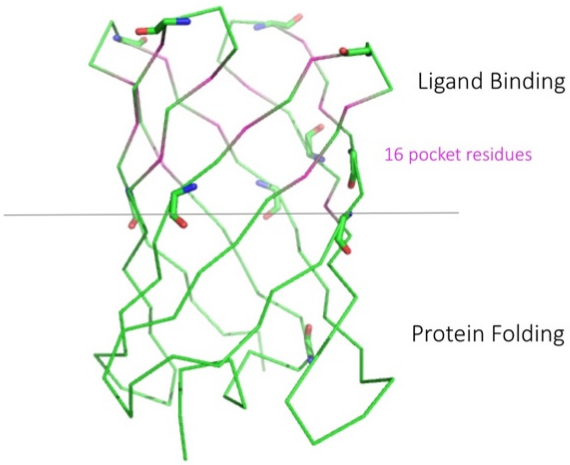


Figure 2.4. *De novo* anti-parallel beta-barrel scaffold compartmented for ligand binding.

108 anti-parallel beta-barrel backbones were assembled from short peptide fragments with geometric constraints to define the closed topology. To reduce the modeling artifacts from sequence-anonymous geometric constraints, the initial backbone models were subject to one round of sequence design and backbone refinement before being used as scaffolds for ligand binding. We noticed that this one round sequence design diversifies the phi psi distributions in the allowed Ramachandra region for beta conformation. A side-by-side comparison of sequence-anonymous "raw" scaffolds and pre-designed scaffolds shows that this backbone dihedral diversity helps to generate more "good models" in the following design step (Figure 2.5).

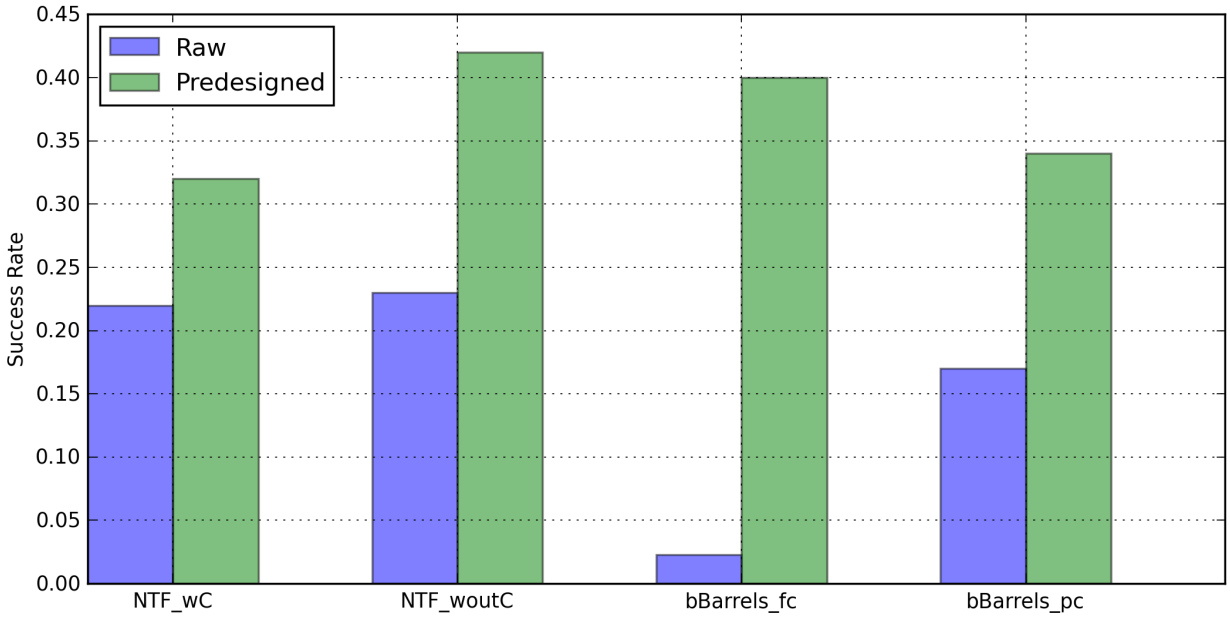


Figure 2.5. An *in silico* benchmark test for assessing scaffold quality. Two sets of scaffolds, "raw" and "pre-designed", are prepared for four categories of scaffold topologies. Using the same design methods, "pre-designed" scaffolds have a higher success rate for generating "good" design models (arbitrary definition based on computed metrics of protein stability and ligand binding energy).

RIF Docking

Previous ligand docking algorithms for fixed-backbone protein design used high-quality crystal structures of natural proteins as scaffolds. For example, the Matcher algorithm grows pre-defined interacting side chains from the scaffold backbone and shows high sensitivity for slightly different backbone conformations[48]. The initial *de novo* backbones assembled from short peptide fragments do not have the same level of fidelity as the crystal structures – the backbone conformations get optimized during sequence design. This is especially true for structures with curved beta sheets that are assembled with additional topology constraints[65]. A new ligand docking algorithm, named as *rotamer-interaction-field* (RIF) docking, focuses on searching for optimal side chain arrangements and uses their discrete backbone dihedral combination to choose the suitable

scaffolds. By reverse the searching order from backbone-sidechain-ligand to ligand-sidechain-backbone, RIF docking can sample more combinatorial discrete sidechain interactions that Matcher does and thus it is less restricted by backbone precision.

To maintain a hydrophobic folding core of the beta-barrel structure, only the top half of the barrel (contains 16 positions) are utilized for placing functional residues during RIF docking (Figure 2.4). Deprotonated DFHBI has five hydrogen bonding acceptors. A minimum number of three hydrogen bonds are required during RIF docking. The amino acid sequences from the pre-treatment sequence-structure optimization were mutated back to Alanine before RIF placed the interacting side chains. 1079 ligand-scaffold complexes with both polar and non-polar interacting side chains around DFHBI were used for further sequence design and structure optimization.

Sequence-Structure Optimization

Unlike the fixed-backbone ligand binding design using natural protein scaffolds, a computational method for designing customized *de novo* ligand binders needs do three things at once: 1. optimize the binding pocket to accommodate the target ligand; 2. optimize the rest of the protein sequence to adopt a lower-energy backbone conformation; and 3. ensure the final backbone conformation is compatible for both ligand binding and protein folding. A two-step design strategy was developed to satisfy these three criteria. First, the side chains in close vicinity of the target ligand (within 10Å and pointing into the ligand) are optimized for lower ligand binding energy while the interacting residues introduced by RIF docking are kept fixed. This is done in a fixed-backbone setting. Second, the remaining protein positions are optimized for lower complex energy while the pocket residues designed in the first step and RIF residues are kept fixed. Once the entire sequence has been assigned to the initial backbone, the complex conformation is refined with backbone flexibility towards a lower total energy. The updated backbone conformation is again used to

optimize the ligand binding energy and the complex energy sequentially. Each design run contains three iterative rounds of sequence-backbone updates.

The design models were ranked by total energy of the complex, ligand binding energy, ligand-protein shape complementarity, and backbone hydrogen bonding. It has been noticed that the omega torsion is distorted by the beta-pairing constraints during the initial backbone assembly and the following iteration of sequence-structure design can correct this with a compatible lower-energy sequence. Hence omega torsion score was used as a calibration of modeling correctness. After each design run, the average line of each of these metrics were used as filters to select models for further refinement. The final top-ranking designs were naturally clustered by the initial ligand docking configuration. Another two rounds of profile-based design were performed to sample the local combinatorial sequence space[71].

In silico Verification and Analysis

To further estimate the folding stability and binding pre-organization, *apo* models for the design proteins are needed. The *ab initio* folding simulations used for validating nonfunctional ideal *de novo* proteins may not apply to a functional protein with a carved-out binding site. Indeed, only a small fraction of natural functional proteins can pass the folding simulation due to their non-ideal structural elements. We pre-assume that the sequences with total energy that is low enough in comparison with the successful non-functional beta barrel design can sample the designed conformation. Hence, the *apo* state should stay fairly close to the bound state. Based on these two assumptions, we used Rosetta structure refinement protocol[72] to model the *apo* state of the designed proteins starting from the bound protein conformation. The modeling convergence is used for stability assessment. Interface pre-organization is estimated by the structural fluctuation around the binding site. The top five low-energy *apo* models as well as top five *apo* models with

closest RMSD to the complex design model were used as the starting structures for ligand docking simulation. Rosetta ligand docking protocol[73] with backbone and side chain flexibility in the binding site was run in parallel for each starting structure to generate five thousand docking trajectories. (Figure 2.6)

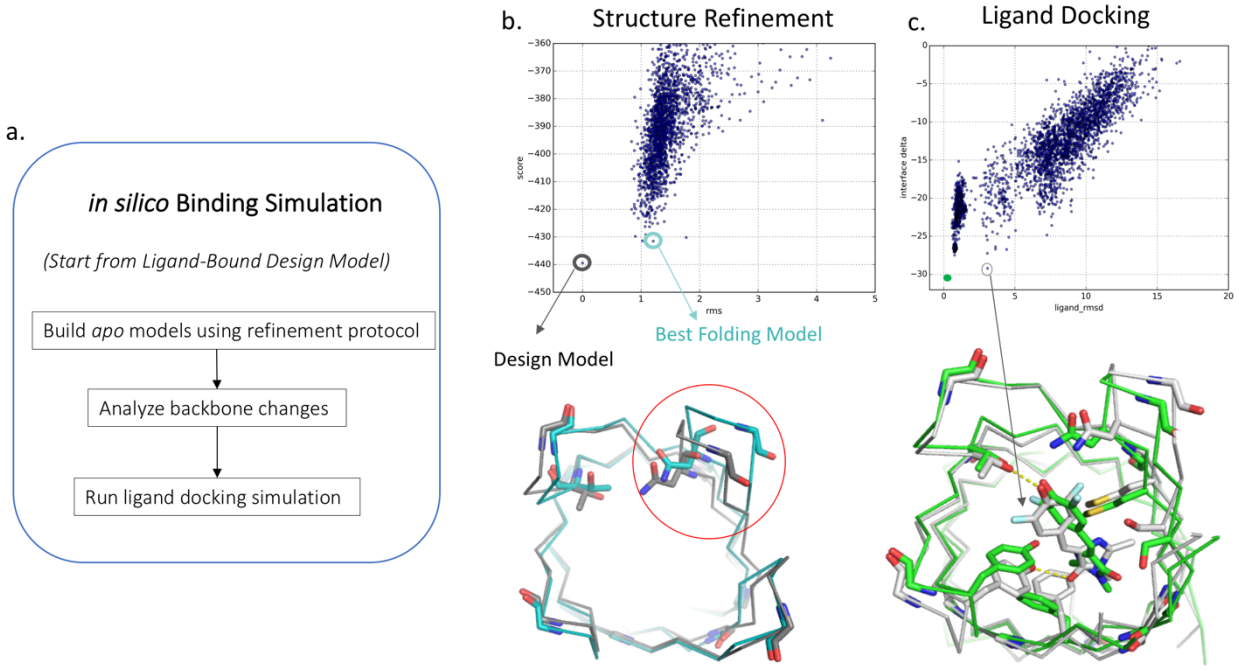


Figure 2.6. *In silico* verification by structure refinement and ligand docking. **a.** The general scheme for model validation. **b.** Example results for building *apo* structures using structure refinement protocols. The converged *apo* models are compared to the design model (bound state). **c.** Example results for docking the target ligand into five top-ranked *apo* models from **b.**

Experimental Characterization of Designed Proteins

The top 42 designs from twenty different original backbones were chosen for experimental assessment. To enhance the folding stability, disulfide bonds were added to eight of those designs. In total, 56 designs were ordered as synthetic DNA for experimental verification. Among all 56 designs tested, 48 can be expressed in *E. coli*; 38 of them can be purified as soluble proteins. By size-exclusion chromatography and circular dichroism (CD) measurement, twenty-two designs

can form monomeric beta proteins, among which, six are perfect monodisperse monomeric proteins. Disulfide bonds in general help monomeric formation. We tested binding activities of the monomeric portion of twenty-two designs by measuring fluorescent signals upon adding DFHBI. Two designs, HBI_b_32 and HBI_b_11, were identified as fluorescent-activating DFHBI binders. HBI_b_32 (Figure 2.7) can be purified as a monodisperse monomeric protein. Far-UV CD spectrum shows that the secondary structure in HBI_b_32 is mainly anti-parallel beta sheet. A positive signal at 226nm is used to monitor the thermal stability of the overall structure. The melting temperature of HBI_b_32 is determined to be around 80℃. The binding effect of point mutants inside the pocket appear to be consistent with the design model (Figure 2.9). The designed hydrogen bonds and major hydrophobic packing interactions are shown to be crucial for activate the fluorescence. The dissociation constant measured by isothermal titration calorimetry(ITC) of HBI_b_32 binding DFHBI is estimated around ~800μM.

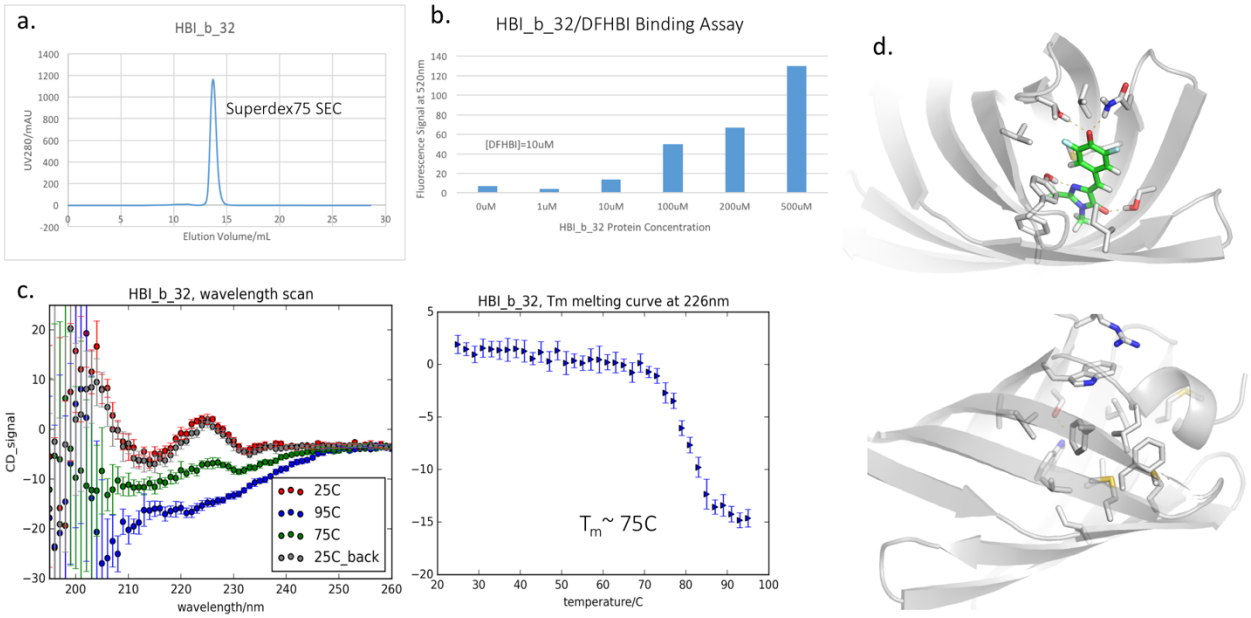


Figure 2.7 Design model and experimental characterization of HBI_b_32 **a.** Size-exclusion chromatography profile of HBI_b_32 after Ni-NTA purification. **b.** HBI_b_32 can activate DFHBI fluorescence in a concentration-dependent manner. **c.** CD spectrum of HBI_b_32. d.

Design model of HBI_b_32. Ligand is shown as green sticks. Binding site is shown on the upper panel while the hydrophobic packing is on the lower panel.

HBI_b_11(Figure 2.8) contains one disulfide bond that connects its C-terminal capping helix with one of strand residues. It is purified as a mixture of oligomers and monomer. Monomeric species can be separated by sizing-exclusion chromatography and it stays as a monomer after separation. After denaturing in 6M GnCl in a diluted oxidizing buffer environment, the oligomeric mixture can be refolded into major monomers. Far-UV CD measurement confirms that HBI_b_11 contains major anti-parallel beta-sheet structures. HBI_b_11 can activate DFHBI fluorescence in a concentration-dependent manner. By removing the designed hydrogen bonds, the fluorescence activation is either abolished or diminished. ITC measurement determines the dissociation constant of HBI_b_11 binding DFHBI is around 18 μM.

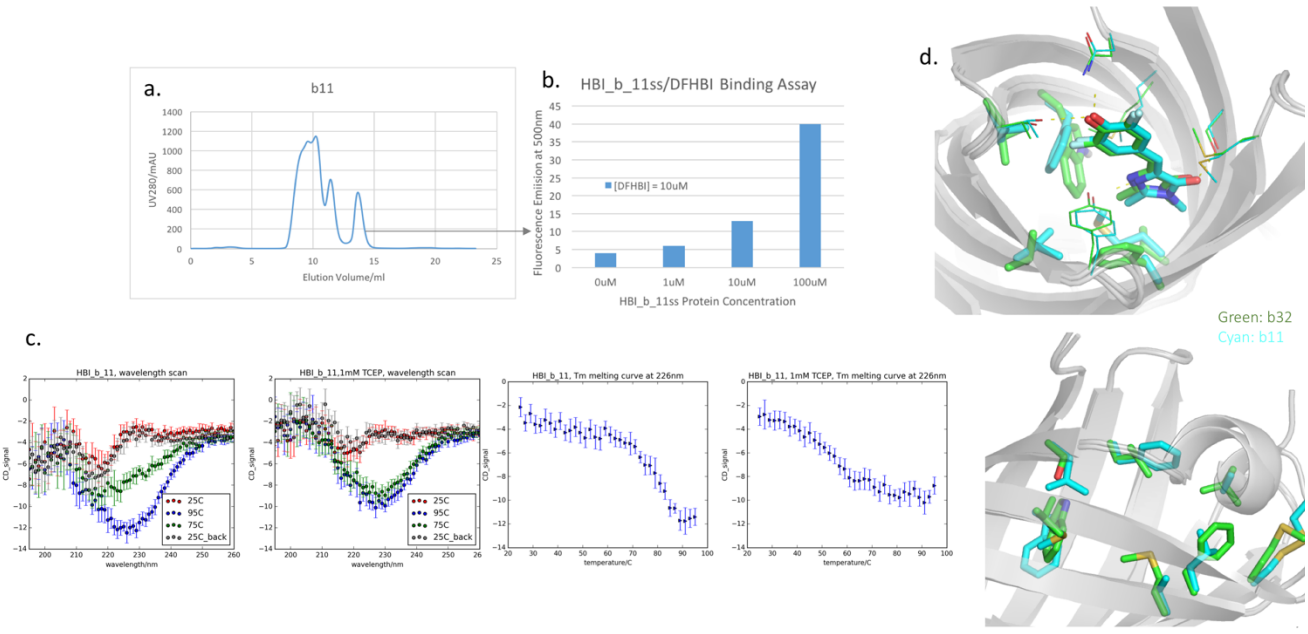


Figure 2.8.Design model and experimental characterization of HBI_b_11. **a.** Size-exclusion chromatography profile of HBI_b_11 using Superdex 75 column. The monomeric peak is around 14mL. b. The monomeric peak is separated for binding assay. Fluorescence signals upon binding

DFHBI are protein concentration dependent. **c.** CD spectra of HBI_b_11 with and without 1mM TCEP. d. HBI_b_11 design model (cyan) is superimposed with the design model of HBI_b_32 (green).

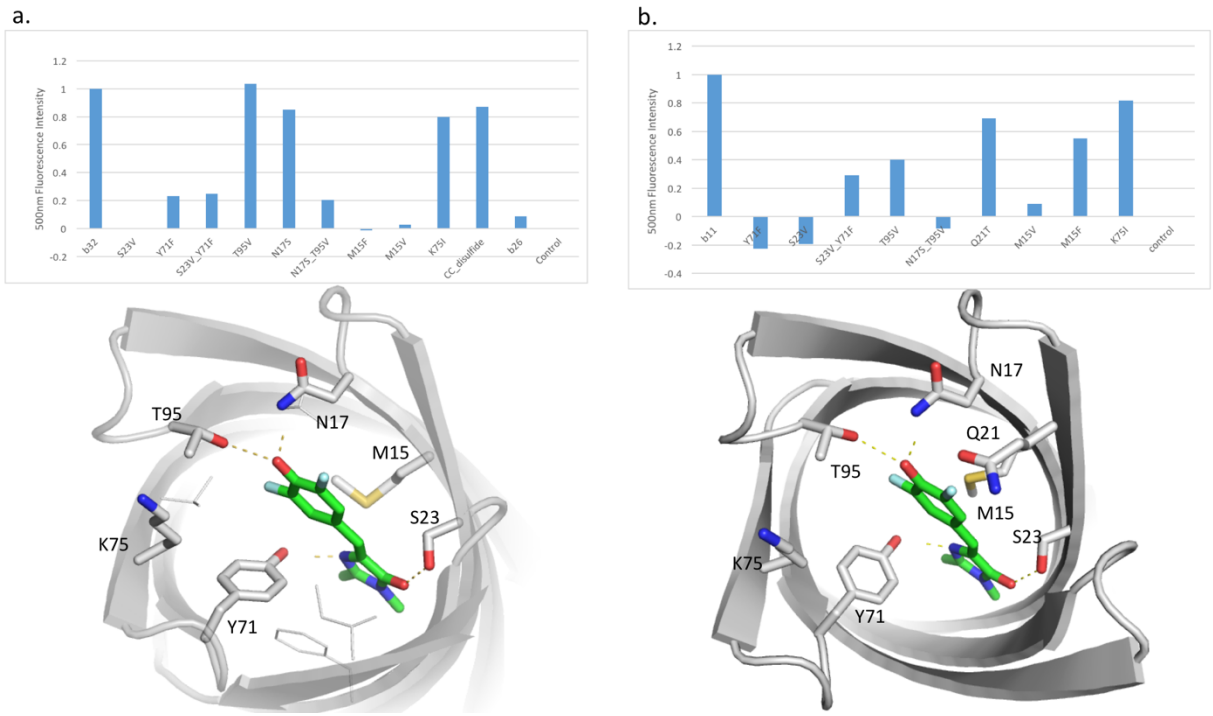


Figure 2.9. Single-mutant functional mapping for HBI_b_32 and HBI_b_11. **a.** Normalized fluorescent signals of HBI_b_32 mutants. **b.** Normalized fluorescent signals of HBI_b_11 mutants. Mutates residues are labeled in the design models on the lower panel.

To gather structural feedbacks for this design calculation and further improve the method development, we set up crystallization screenings for all the monomeric designs. A 1.6Å dataset was collected for HBI_b_10 (Figure 2.10). HBI_b_10 is a monomeric anti-parallel beta-sheet protein as confirmed by sizing-exclusion chromatography and CD measurement. However, it does not activate DFHBI fluorescence even at a very high concentration. In the crystal structure, HBI_b_10 forms an 8-strand 10-residue anti-parallel beta barrel and the overall structure is very close to the design model with Cα backbone RMSD of 0.575Å. All the key structure elements are confirmed in HBI_b_10 crystal structure. The design binding cavity is well preserved and three

water molecules are seen in the pocket. In comparison with the designed complex model, Phe101 in the binding pocket adopts a different rotameric state in crystal structure, which might hinder the entrance of DFHBI. A subtle movement of loop 3 backbone is also observed in crystal structure. Thus we speculate that the failure to activate the DFHBI fluorescence might be due to the subtle backbone flexibility and side chain rotameric fluctuation. This structure highlights the atomic precision required for ligand binding design (Figure 2.11).

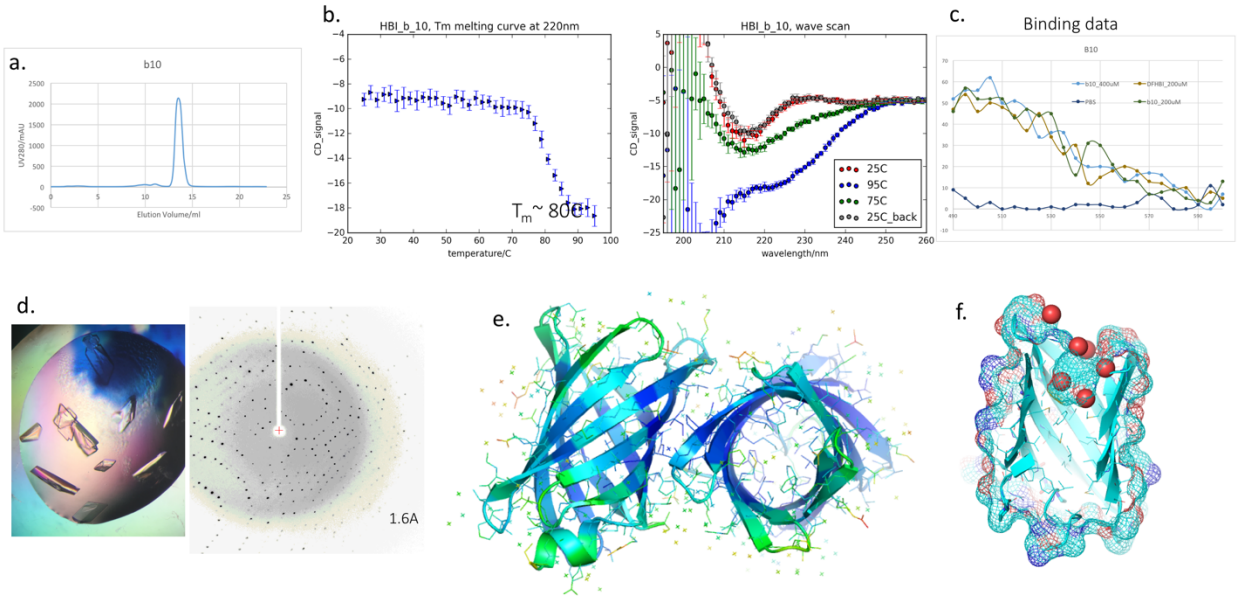


Figure 2.10 Experimental characterization and crystal structure of HBI_b_10. **a.** Size-exclusion chromatography profile of HBI_b_10 using Superdex 75 column. **b.** CD spectra of HBI_b_10. **c.** Fluorescence emission spectrum of HBI_b_10 upon adding DFHBI. HBI_b_10 does not activate DFHBI fluorescence when compared with free DFHBI. **d.** Crystals of HBI_b_10 protein and the X-ray diffraction pattern. **e.** The asymmetric unite of HBI_b_10 crystal after final refinement. f. HBI_b_10 crystal structure has an open binding site with several water molecules.

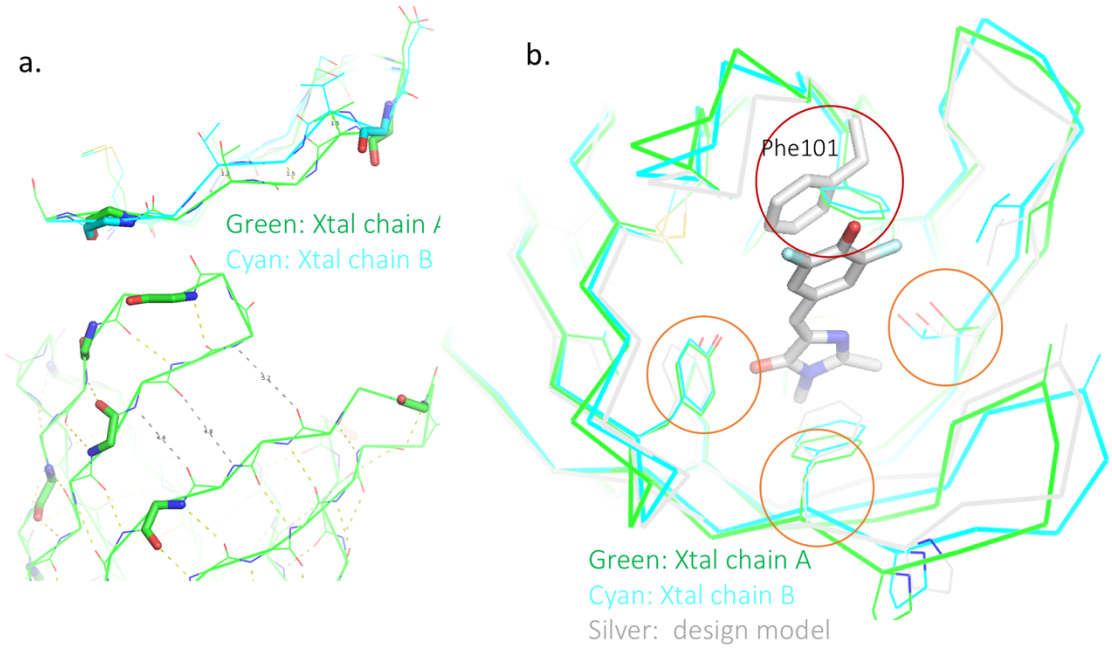


Figure 2.11 Crystal structure of HBI_b_10. **a.** Comparison between crystal chain A (green) and chain B (cyan). Part of chain-A barrel has non-ideal hydrogen-bond pairing. **b.** Comparison of ligand binding sites. Design model with DFHBI ligand is shown in silver. Phe101 adopts a different rotameric conformation from predicted model.

Towards Improving the Binding Affinity

The initial binding affinity of our best fluorescent-activating DFHBI binder, HBI_b_11, is estimated to be 18 μM by ITC. The fluorescent signal starts to becomes undetectable when DFHBI concentration is lower than 1 μM. For imaging applications, this is far below the useful minimum for cellular imaging. From the perspective of rational protein design, the determinants of high-affinity ligand binding (K_d of nano-molar to pico-molar) are still beyond our current computational calculations. The typical low-micromolar binding affinities of designed ligand binding proteins raises the issue of missing components producing high-affinity binding proteins. Extensive structural studies have been carried out to address this question. Early structural and mutagenesis studies of biotin-streptavidin interaction suggest that the extended hydrogen bonding network,

comprehensive hydrophobic packing and a "sequestering" loop contribute to the unusual binding strength[4]. The structural changes from low-affinity germline antibodies to monoclonal antibodies of nanomolar affinity indicate a mechanistic transition from "induced-fit" to "lock-and-key"[8]. *In vitro* directed evolution experiments that improve antibody-antigen interaction to pico or femto molar binding also support the underlying "lock-and-key" mechanism[74]. Based on those results, we hypothesize that in order to improve the binding affinity of computational designed binders, we need the following: 1. more comprehensive packing interactions and 2. pre-organization of interacting residues in the *apo* state. One engineering approach to test this hypothesis is to build well-ordered loop elements into the original beta-barrel scaffolds.

Loop structures in natural proteins are very diverse and flexible; they often adopt specific conformations that are critical for protein function. Without being locally constrained by regular backbone hydrogen bonding patterns, loops reflect highly intertwined sequence-structure relationships: the backbone conformation seen in a loop is largely dictated by its side chains and highly sensitive to sequence changes. To circumvent this complexity, we manually curated a collection of well-ordered beta-turn-beta loop structures from PDB (Figure 2.12a). Side chains that are involved in satisfying the backbone hydrogen bonds are kept fixed as parts of the fragment information. We used Rosetta Remodel to search and insert loops that can fit into the barrel structure. Based on the design models, loop 3, loop 5 and loop 7 of HBI_b_32 and HBI_b_11 are used individually to explore "loop insertion" for improving the binding affinity towards DFHBI. Three iterative rounds of sequence design and backbone refinement were performed after loop insertion. With fixed structural side chains, the rest of the loop sequence was optimized for ligand binding interaction; relaxed and minimized for total complex energy. To validate the model consistency, the lower-energy loop conformations were subject to structure prediction by

kinematic loop closure (KIC)[75]. The design models with longer loops showed improved ligand binding energy and interface shape complementarity (Figure 2.12c). In total, 46 loop variants of HBI_b_32 and 36 loop variants of HBI_b_11 were ordered as synthetic genes and tested in experiments.

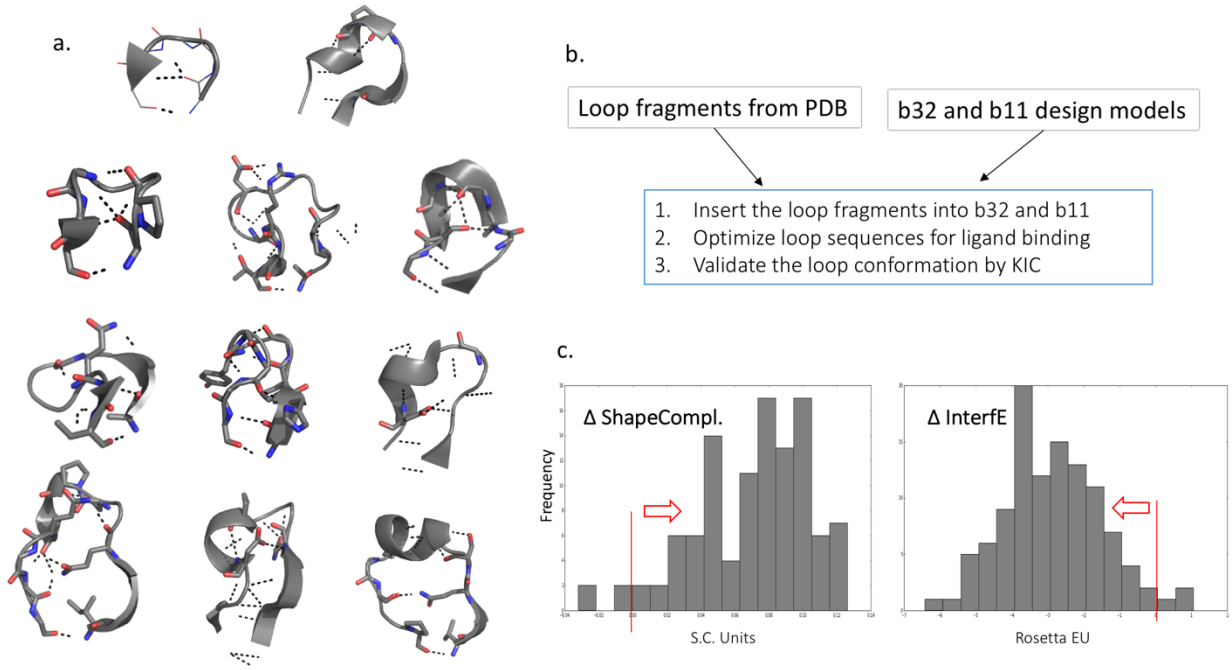


Figure 2.12 Computational design of structural loops to boost binding affinity. **a.** Various types of loop structures collected for grafting onto HBI_b_32 and HBI_b_11. **b.** The computational protocol used for inserting and designing the loops. **c.** Histograms of changes of calculated shape complementarity (left) and interface energy (right) after adding the loops.

Since HBI_b_32 can be easily produced from *E. coli* as a mono-dispersed monomeric protein, we tested 46 loop variants of HBI_b_32 by protein expression and purification, followed by sizing-exclusion chromatography and CD measurement. Purified proteins were used for testing DFHBI binding activity by detecting fluorescent signal upon adding DFHBI to the protein solution (Figure 2.13a). 36 loop variants based on HBI_b_11 were directly transformed into EBY100 yeast strain

for surface display and the binding activity was tested by flow cytometry. Loop variant b11_L5F shows improved binding affinity. (Figure 2.13b)

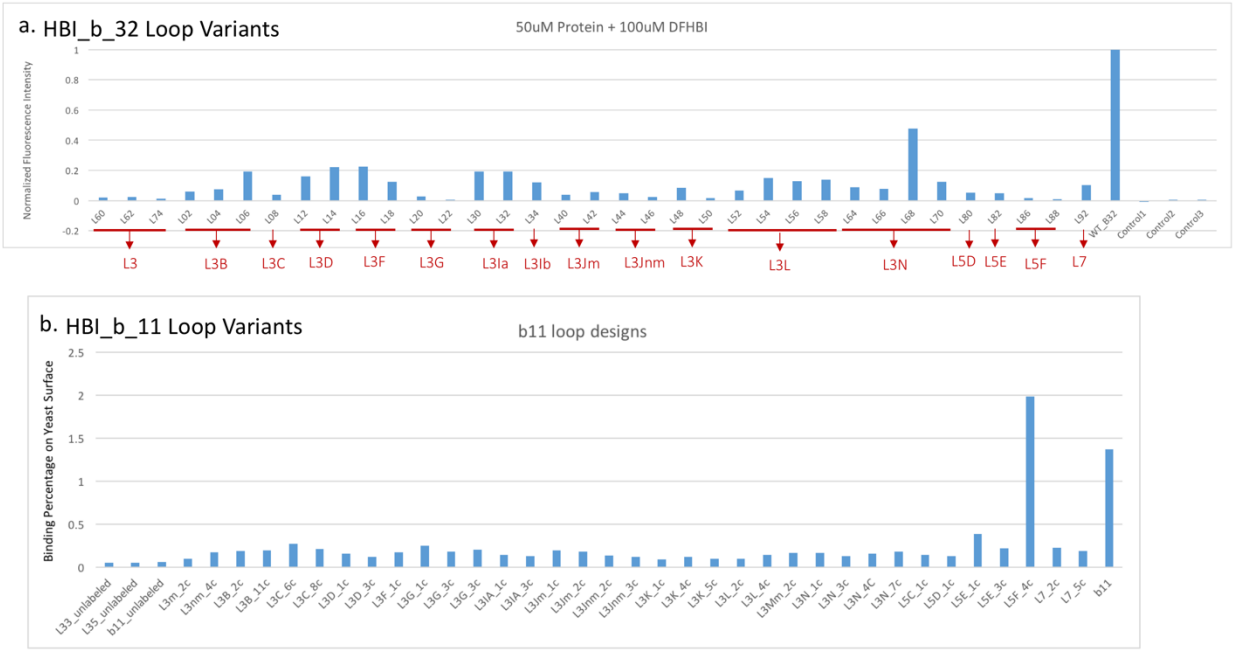


Figure 2.13 Fluorescence binding data for loop variants of HBI_b_32 and HBI_b11. **a.** Normalized fluorescence signals of designed loop variants for HBI_b_32. **b.** Fluorescent signals detected by flow cytometry of yeast cells displaying designed loop variants for HBI_b_11. L5F_4c appears to be the only that improves the binding affinity.

2.4 Discussion

Most current protein engineering work relies on re-purposing natural proteins to realize a new function. While there is tremendous diversity to explore and learn from, natural proteins may not be the best starting point to engineer novel functions. For small molecule binding and enzyme design, the precise placement of interacting side chains requires a compatible backbone and coherent second-shell supporting residues. To achieve this, extensive changes need to be introduced, which may not be compatible with the native function and built-in stabilizing

mechanism of existing natural proteins. The work presented in this chapter explored the possibility of designing an entire protein from scratch to fold and form a specific binding site with a coherent hydrophobic core.

The results show that we were able design a monomeric anti-parallel beta-barrel protein that fold and activate DFHBI fluorescence by forming non-covalent interactions. By compartmenting the hydrophobic packing core from the dedicated binding site, we were able to generate stable well-folded monomeric beta barrels, confirmed by high-resolution crystal structure of HBI_b_10 (Figure 2.10). Two fluorescence-activating binders, HBI_b_11 and HBI_b_32, were identified with low-to-high micromolar binding affinity to DFHBI (Figure 2.7 and Figure 2.8).

Two binders are based on the same starting scaffold and the same docking interactions with sequence identity of 72%. This convergence reflects the dual requirements for both backbone conformation and sidechain positioning. The majority of designs tested form partial or complete beta aggregates (20/42), some of which can be rescued by a single disulfide bond. While the sample size is still too small to draw out significant discriminating factors, we saw that monomeric formation positively correlates with the volume of hydrophobic packing. To further improve the success rate of monomer formation, a new topology with more hydrophobic packing interactions or a de-localized secondary packing core could help. For the current beta-barrel topology, another option is to move the binding site to the peripheral loop regions thus more residues in the barrel can be utilized for hydrophobic packing.

While several of loop variants show an increase for binding affinity, all the loop variants for HBI_b_32 decreased the fluorescence activation. We suspect that the subtle change of the bound conformation of DFHBI with an additional loop might be the reason. The plasticity of the side chain arrangement in the pocket can distort the electron conjugation of DFHBI and lead to non-

productive binding. Similar results were seen for HBI_b_11 loop variants. 35 out 36 HBI_b_11 loop variants showed decreased fluorescence by yeast display. The only one variant L5F_4c that increases the fluorescence also improved the binding affinity by about two fold. We are in the process of generating crystal structures of any loop variants to understand the mechanism.

Chapter 3. HIGH-THROUGHPUT ASSAY FOR DETECTING PROTEIN-LIGAND INTERACTIONS

3.1 INTRODUCTION

3.1.1 *Gene Synthesis and Next-generation Sequencing*

Because the proteins that we are designing do not exist in nature, genes that encode their amino-acid sequences cannot be amplified from existing genomes. To produce designed proteins as recombinant proteins in *E. coli*, synthetic genes that encode the designed amino acid sequences must be manufactured in the first place. DNA synthesis techniques have improved dramatically in the past ten years, greatly reducing the cost of synthesizing *de novo* DNA and increasing the number of computational designs that can be tested experimentally. The traditional column-based oligo synthesis can routinely produce up to ~100 nucleotides (nt) with error rates of 1 in 200 nt or better. Oligo pools with degenerate codes have enabled the construction of large site-directed mutagenesis libraries. However, the length of 100 nucleotides is too small to cover most designed proteins. The commercialization of array-based oligo synthesis with enzymatic error correction has enabled the large-scale gene synthesis for designed proteins[76]. By the time this work here was done, Gen9 was founded to produce low-cost high-quality synthetic genes (Gen9 has recently been acquired by GinkoBioworks Inc. Boston, MA). Agilent Technologies (Santa Calara, CA), Twist Bioscienc (San Francisco, CA) and other biotech companies equipped with the latest generation of gene synthesis technique have emerged to provide even cheaper synthetic genes. Nowadays, thousands of designed proteins can be easily synthesized as synthetic genes for expression and testing. This allows biochemists and protein engineers to develop new experiments to tackle protein design problems.

On another field of battle, the release of the first high-throughput sequencing platform in the early 2000s announced a 50,000-fold drop in the cost of human genome sequencing and this led to the moniker: next-generation sequencing(NGS)[77]. To date, the capacity of NGS technologies has increased by a factor of 100-1,000. The cost of sequencing a human genome is now down to around US$1,000 and DNA sequencing has become a clinic tool in major research hospitals[78]. For basic research, NGS platforms provide unprecedented technologies to tackle the major problems in

genetics and molecular biology. While it has propelled exciting genetic research that was considered impossible only a few years ago, the application of NGS technologies in protein engineering has been mostly limited to detecting mutations in an oligo-constructed gene library[79].

3.1.2 *Interrogating Small Molecule-Protein interactions*

While many high-throughput methods exist for the characterization of protein-protein and protein-DNA/RNA interactions, far fewer approaches exist to elucidate interactions between proteins and small molecules. Current experimental approaches to interrogate the small molecule-protein interactions (SMPIs) are mostly based on affinity-chromatography separation or relying on detecting difference in the stability between unbound and bound proteins. Both approaches gain high-throughput by downstream mass spectrometry experiments and have proved particularly powerful for proteomics studies. However, for engineered proteins, multiplex detection system that can interrogate hundreds to thousands of designed proteins against multiple small molecule targets for both affinity and specificity measurements would be preferred.

Protein display and functional selection have been used to repeatedly select for a desired functionality from a diverse gene library until convergence on a small number of protein variants is achieved. With the advent of NGS, the frequency of many variants can be evaluated in parallel in pools selected under different conditions. The strength of selective pressure for the desired functionality determines the diversity of the gene pool after selection; depending on the selective pressure, weak or dysfunctional variants will be depleted from the pools. The diversity that can be assessed in detail depends on the numbers of genes or gene fragments that can be sequenced; currently even a simple benchtop sequencer, such as the Miseq (Illumina), can obtain up to 35 million sequences in a single run and the numbers are increasing due to constant improvements of the technology. Functional selection combined with NGS has been used to obtain detailed protein fitness landscapes for a variety of proteins, and to probe individual residue contributions to specificity of interactions[79,80].

Here, we describe an approach to quickly identify not only new protein-protein interactions (PPI) and small molecule-protein interactions (SMPI), but also to comprehensively monitor promiscuous binding behavior and off-target interactions. We combine proteins to be queried into single pools and carry out selections against a series of target molecules. The selected pools as well as the starting library are sequenced and analyzed to obtain a binding profile that contains

both affinity and specificity information. This new approach will enable a high-throughput assessment of thousands of ligand binding proteins against multiple small molecule targets from computational design.

3.2 METHODS

Gene Synthesis and Cloning

Gene fragments were synthesized (Gen9 Inc.) with 5' and 3' additions homologous to the pETCON plasmid allowing recombination into the expression vector within the yeast cell. pETCON plasmid was digested with NdeI and XhoI restriction enzymes. A total of 5 μg of the gene fragments and 1 μg digested pETCON vector were co-transformed into yeast EBY100 cells[41]. The transformation yielded 1 x 10^7 transformed cells. Yeast cells containing the synthetic gene libraries were grown overnight at 30°C in 50 mL minimal medium containing 2% glucose, but lacking tryptophan and uracil. For induction, cells were adjusted to an optical density of 1, sub-cultured into SGCAA and grown at 22°C for another 18 h. Before incubating with respective target molecules, yeast cells were washed twice with PBSF and normalized to an O.D. of 1.

For the selection of small molecule binders, three different modifications of small molecules were utilized: category 1. Monovalent biotinylated ligands, 2. biotinylated BSA conjugated to the ligand, and 3. biotinylated 70K-dextran conjugated the ligand. Fluorescent detection was enabled by incubation with SAPE and anti-myc Fitc conjugated antibody. For each ligand in category 1, 4 μM was pre-incubated with 1 μM SAPE to create additional avidity. For category 2, 2.5 μM biotinylated BSA ligand conjugates were mixed with 627 nM SAPE. For category 3, we combined 640 nM biotinylated 70K-dextran conjugates with 487 nM SAPE. To every 50 μL reaction, 1 μL anti-Myc-FITC was added. Cells (5 x 106) were labeled at room temperature for 2.5 hours while rotating and washed once with ice-cold PBSF before sorting. For each target, at least 1 million cells were sorted using fluorescence-activated cell sorting (FACS) on a BD Influx sorter. Gates were drawn based on the signals observed (Figure 3.3 and Figure 3.4); cells that were only labelled with anti-cMyc-Fitc conjugated antibody were used as a reference for background fluorescence at 580 nm and gates were drawn just above these populations.

DNA Preparation and Next-Generation Sequencing

Plasmids from cells of the starting and selected pools were extracted as previously described[81]. Following a QIAgen PCR clean-up step producing a 30 μL DNA solution, 15 μL were subjected to PCR for the addition of selection-specific barcodes and flow cell adapters. For that, two PCR steps were performed. Plasmid-specific primers at the 5' (upstream of the *NheI* site) and 3' site (including the *XhoI* site) were used for the first PCR to amplify the gene. To add the Illumina flow-cell adapters and selection-specific barcodes, a second PCR step with a higher annealing temperature (64°C) was performed using primers overlapping the end of the first PCR. Miseq chip adapter and selection-specific barcoding sequence were added to the flanking regions of the primers. 2 μL of the first reaction served as template for the second reaction. All primers were PAGE purified. The first PCR step was performed for 14 cycles, whereas the second PCR was performed for 15 cycles. However, the cycles necessary for the first reaction depend on the efficiency of the DNA preparation from the yeast cells and may need more cycles, which can be monitored using qPCR. Resulting DNA fragments were gel purified and amounts were quantified by qPCR as instructed (Illumina qPCR manual). For sequencing whole genes (SMPI libraries, from pool 2 selection), 300 forward and 300 reverse reactions were performed and 5 times the amount of DNA was used for the unselected pool to get a complete coverage.

3.3 RESULTS

The experiments were designed for simultaneously assessing binding of thousands of proteins to dozens of small molecule targets. The proteins of interest are displayed on the surface of yeast and evaluated for binding to biotinylated target molecules. DNA fragments encoding the protein of interest with flanking regions containing a short sequence for homologous recombination into the surface expression vector were co-transformed into yeast cells with linearized plasmid DNA and evaluated in pools for display and binding to a range of target molecules. To compensate for possible overrepresentation of individual genes, the fraction of each clone in the starting pool was determined. Plasmid DNA for the unsorted gene library and each selected pool was isolated and tagged *via* PCR with sequencing-chip tethering adapters and individual barcodes for each sorting experiment. After next-generation sequencing, sequences were counted, and gene frequencies were normalized to their corresponding reference pool. The enrichment values provide information on the affinity and specificity of each queried protein for each target and thereby portray specificity

profiles. We then used the method to assess binding affinity and specificity of designed binding proteins for a set of small molecule targets. (Figure 3.1)

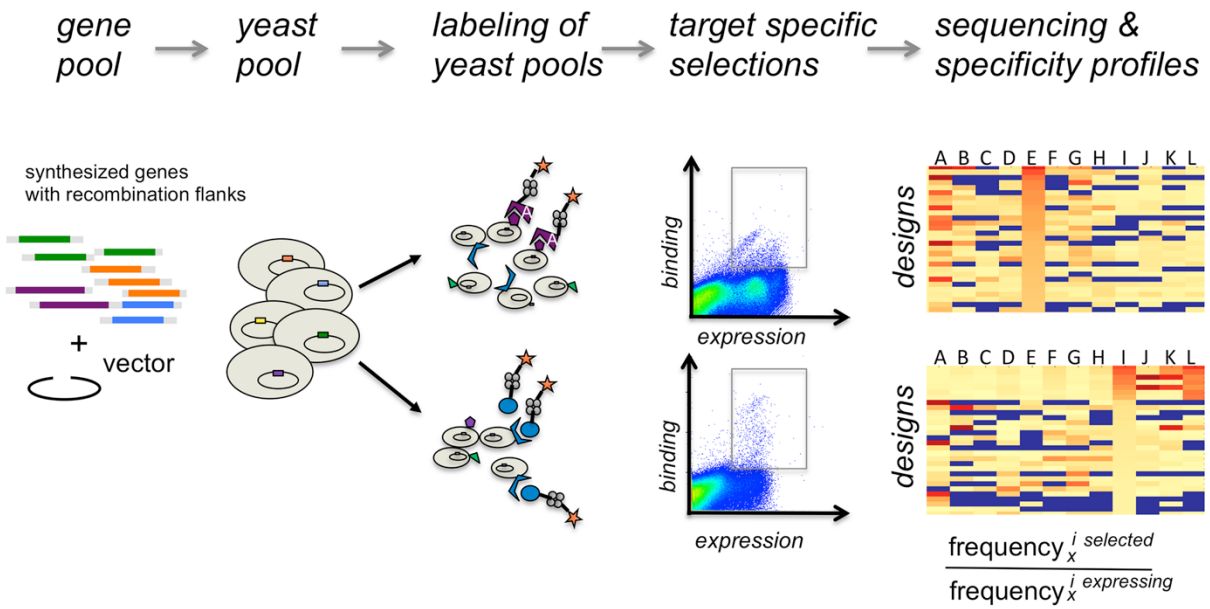


Figure 3.1.Experimental scheme for high-throughput binding assessment of designed proteins. Genes synthesized as a pool are transformed into yeast cells for surface displaying. The resulted yeast pools are labeled with different targets at various conditions to enable conditional FACS selection. DNA from the selected and unselected yeast pools are extracted and sequenced. Sequencing results are analyzed to get an enrichment factor for each gene at different conditions.

3.3.1 *Selections of Small Molecule Binders*

We interrogated the ligand binding properties of a pool of 228 designed proteins to 8 conjugated small molecule ligands: cortisol (HCY), 17-hydroxyprogesterone (OHP), vitamin D (VitD), Mycophenolic acid (MPA), Apixaban (APN), artemisinin (ART), Fentanyl (FEN) and biotin (BTN). With the exception of HCY and OHP (Figure 1.4), these small molecule targets are quite different in their molecular properties (Figure 3.2). Ligands were either conjugated directly to biotin, to biotinylated BSA or to dextran as a "carrier" molecule to increase avidity effects. Fluorescence activated cell sorting (FACS) was carried out with each ligand conjugate. To increase the signal to noise ratio, we performed two rounds of sorting (Figure 3.3). Using the procedure described above, we identified several new binding proteins and assessed their binding selectivity.

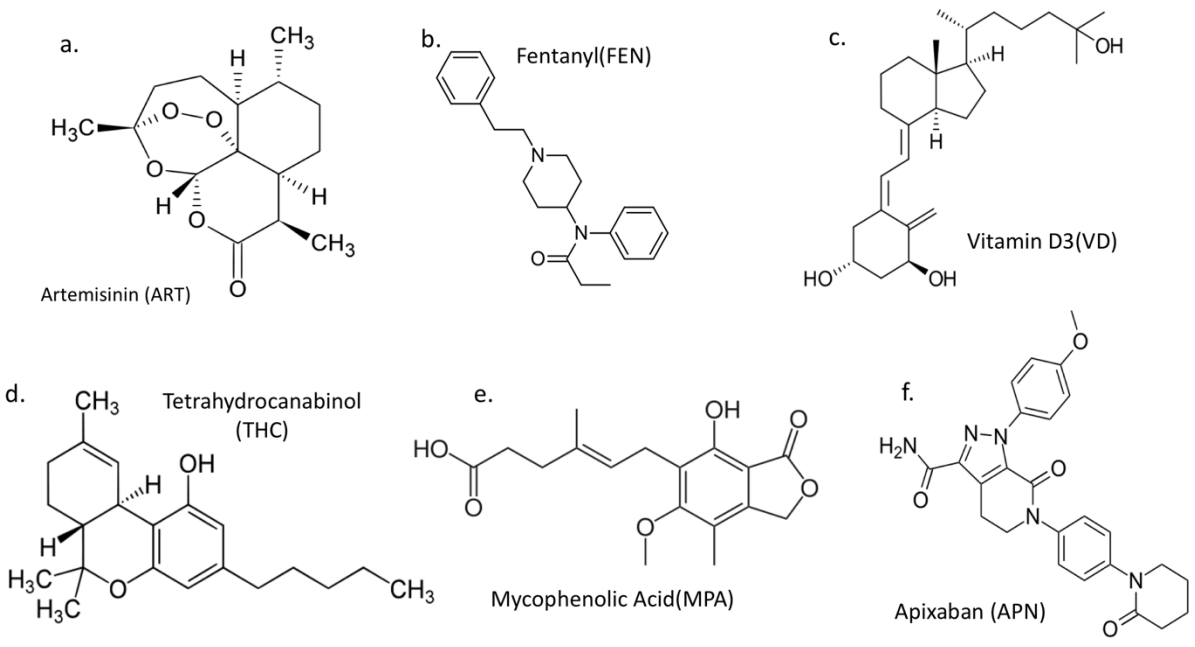


Figure 3.2. Small-molecule targets tested in the parallel screening. **a.** Artemisinin **b.** Fentanyl **c.** Vitamin D3 **d.** Tetrahydrocanabinol **e.** Mycophenolic acid **f.** Apixaban

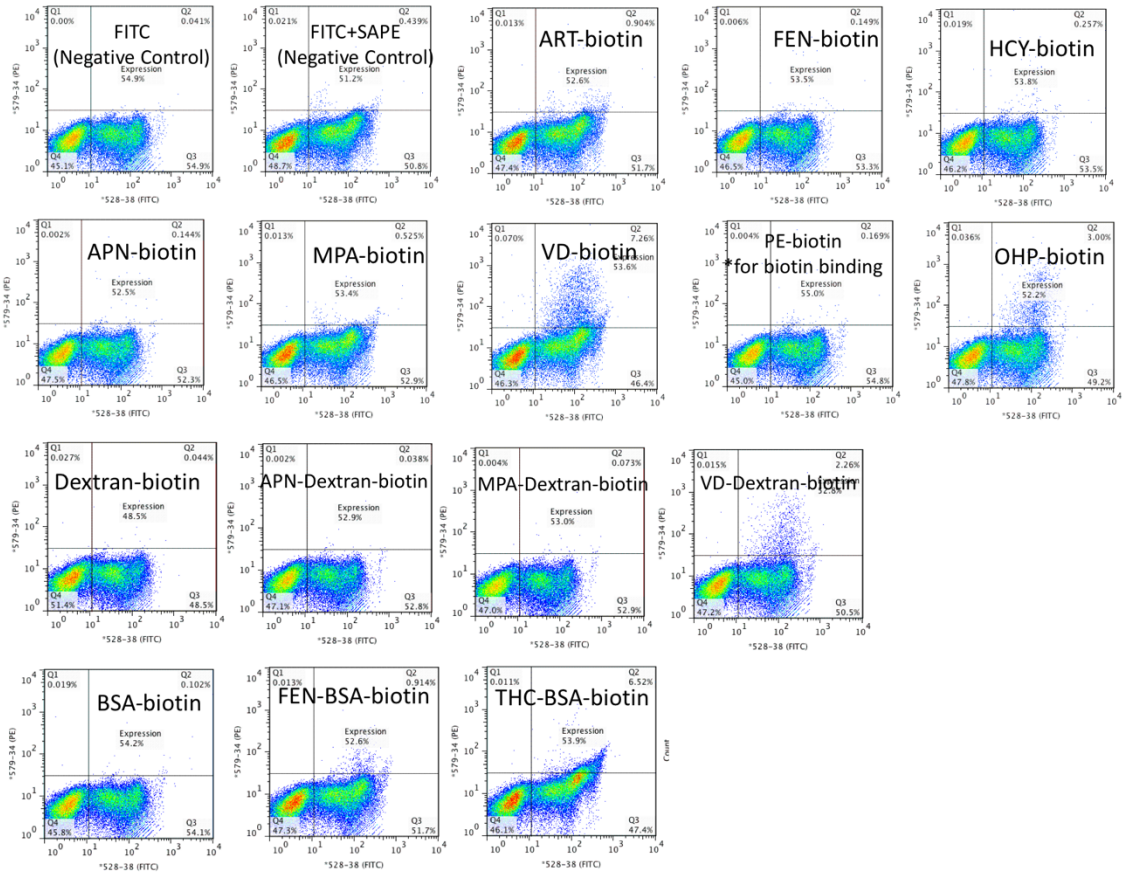


Figure 3.3.First round of fluorescent labelling and FACS selection towards different targets.

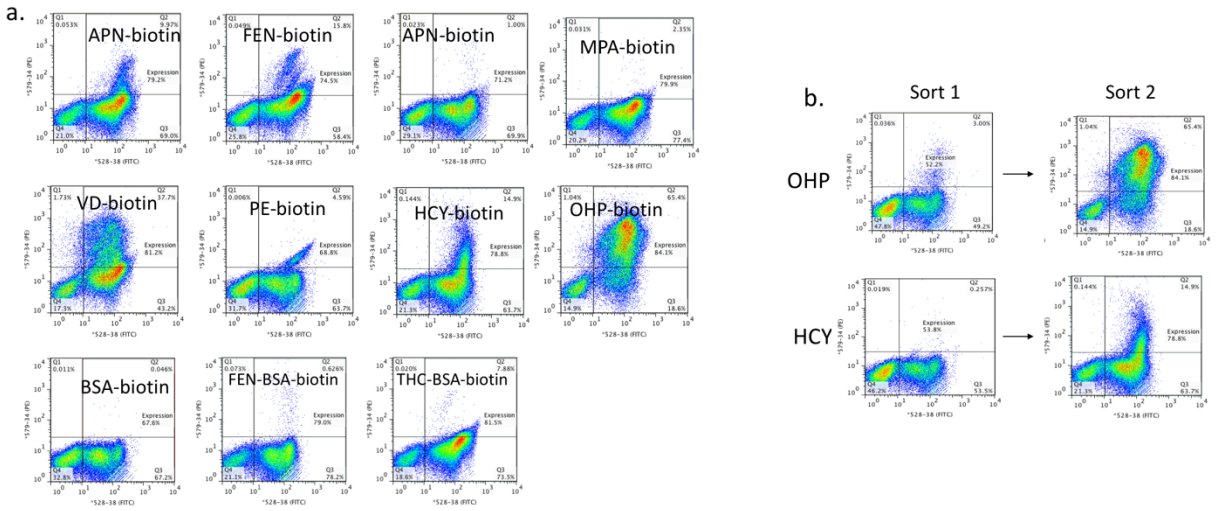


Figure 3.4.Second round of fluorescent labeling and FACS selection to enhance the positive binding signals. **a.** FACS signals of selected pools labeled with their targets for the second round

selection. **b.** Two rounds of FACS signals for target 17-hydroxylprogesterone (OHP) and cortisol (HCY) show increased binding signal after the first selection (Sort 1).

All sixteen designs for OHP binding (Chapter 1) were included in the gene pool. Our high-throughput assay was able to identify seven out of eight binders towards OHP. In addition, binding selectivity among the list of small molecule targets (Figure 3.2) was interrogated in parallel (Figure 3.4). In general, designed OHP binding proteins showed a specific signal for OHP while OHP10 also showed a positive signal against vitamin D. To verify that the enriched proteins indeed bind to the corresponding small-molecule target, we tested the designs as individual clones (Figure 3.5). All three designs showed a clear binding signal to 1 μM biotinylated OHP but not PE or BSA controls. OHP10 indeed binds 1 μM biotinylated vitamin D, for which, OHP13 also shows a weak binding signal that corresponds to a lower enrichment value (Figure 3.6).

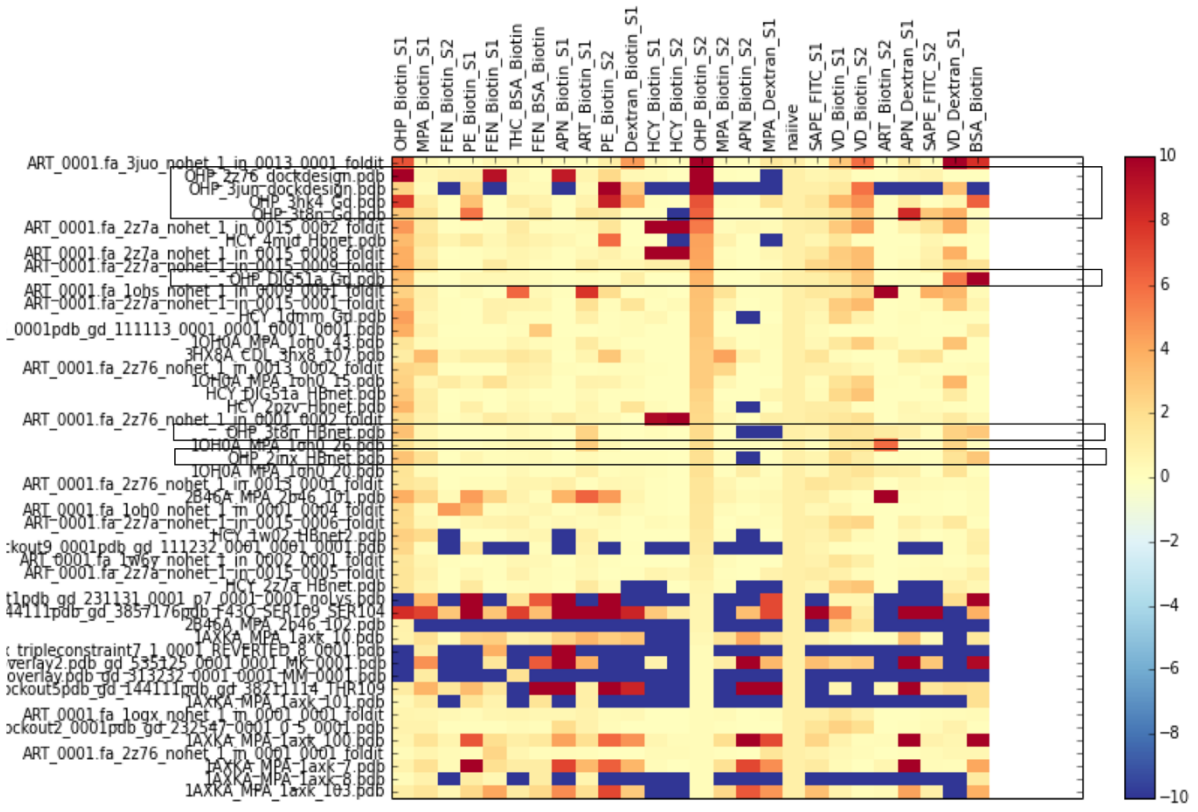


Figure 3.5.Enriched designs after two rounds of FACS sorting towards OHP binding. The horizontal axis lists all the small-molecule targets interrogated. The vertical axis lists each design in the pool. The color matrix represents the calculated enrichment values of individual design

against each target. For this plot, the designs are ranked by their enrichment values in the second round of binding selection for 17-hydroxylprogesterone (OHP_Biotin_S2).

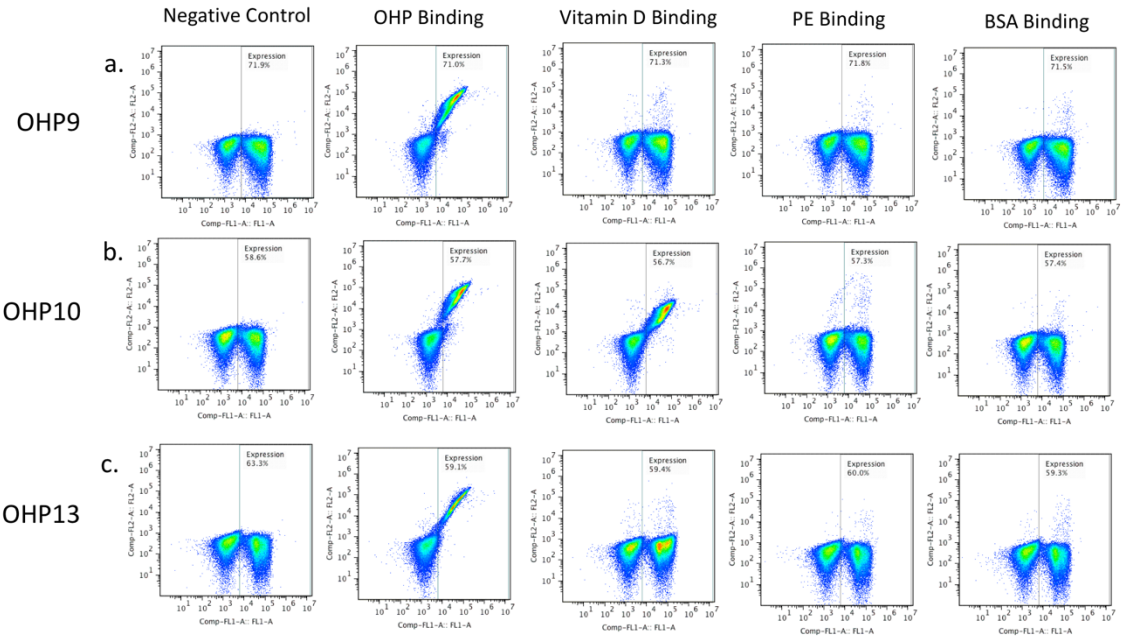


Figure 3.6.Three representative OHP binders tested as individual clones. **a.** OHP9-displaying yeast cells are tested for binding 17-hydroxylprogesterone (OHP), Vitamin D, Phycoerythrin (PE), and BSA. OHP9 only binds OHP. **b.** OHP10 binds both OHP and Vitamin D. **c.** OHP13 binds OHP and shows a weak binding signal for Vitamin D.

Three designed proteins for binding Artemisinin showed positive enrichment values for cortisol binding (Figure 3.7), which appeared quite selective towards cortisol. Interestingly, those three designs were based on the natural protein RV0760[61], which is the same scaffold for OHP9 (Chapter 1) and HBI3 (Chapter 2). ART_1_2, ART_15_8 and ART_5_2 were further tested as individual yeast clones . The scaffold protein does not bind cortisol at the same testing condition; Several point mutants were made to identify the binding site (Figure 3.8).

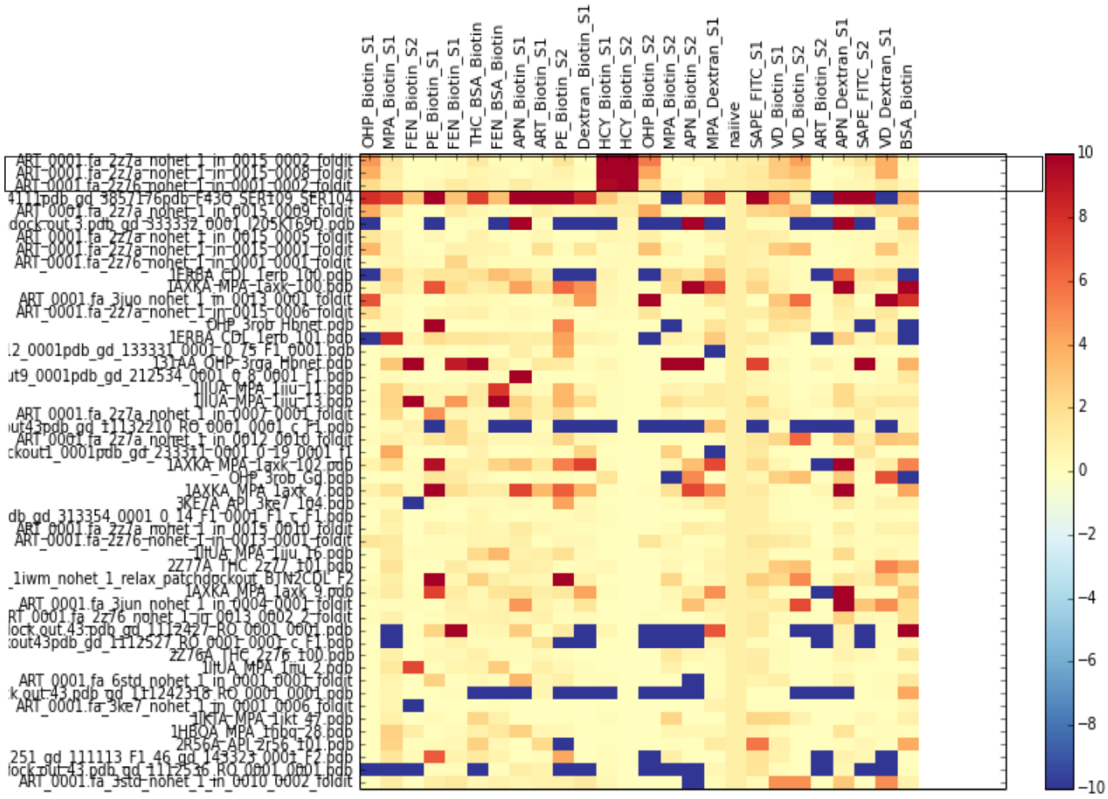


Figure 3.7. Enriched designs after two rounds of FACS sorting towards cortisol binding. The horizontal axis lists all the small-molecule targets interrogated. The vertical axis lists each design in the pool. The color matrix represents the calculated enrichment values of individual design against each target. For this plot, the designs are ranked by their enrichment values in the second round of binding selection for cortisol(HCY_Biotin_S2). Three designs for binding artemisinin (ART) show up on top of the list.

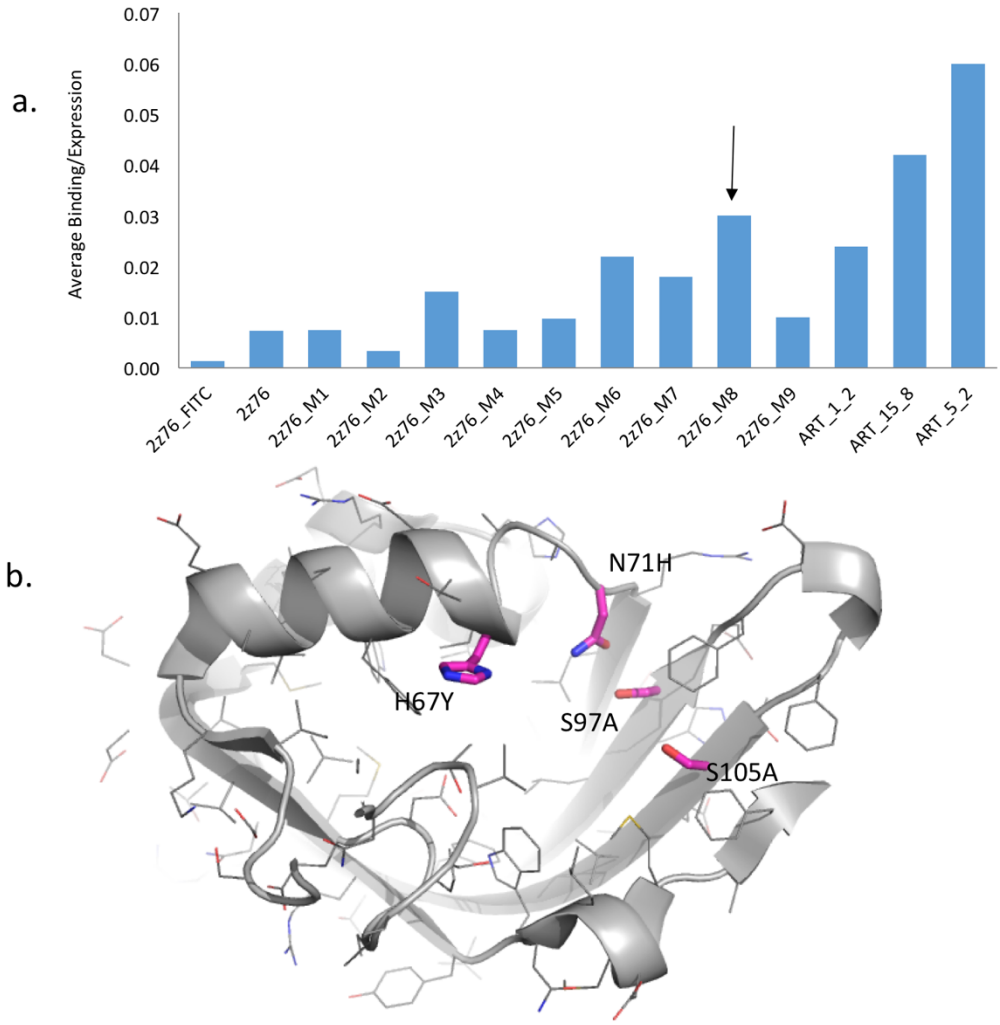


Figure 3.8.Mutational mapping for the accidental cortisol binders. **a.** Three artemisinin designs, ART_1_2, ART_15_8 and ART_5_2, are found to bind cortisol in the large screen are tested as individual clones. The scaffold protein (PDB ID: 2Z76) does not bind cortisol. Nine mutants, M1 to M9, are constructed to map the binding sites. M8 shows highest binding signal among all the mutants, as indicated by a black arrow. **b.** Four mutations that M9 carries to bind cortisol. They are highlighted as pink sticks.

3.3.2 *Bioinformatics Tool Development*

Sequencing reads were split into the different populations based on their 12 bp selection-specific barcodes. Pools were treated identically for analysis and quality filtration. All sequences with an

average quality score below 20 or if they contained any position with a score lower than 12 were rejected. Sequences between the restrictions sites were extracted and counted (For genes longer than the sequencing cycle, we utilized only the forward sequences for counting). The frequency of each gene in each selected pool was normalized by its frequency in the reference pool and described as enrichment values. To provide an estimate of the reliability of the data, we incorporated a bootstrapping re-sampling step; we oversampled the library size by pulling 20,000 sequences 50 times randomly from the raw sequencing data split into corresponding selection pools. This reduces artificially high appearing enrichment values caused by low sequencing coverage in the reference pool; spread of data can be monitored for any given input gene. For each draw, frequencies from the reference pool (either starting library or cells selected for expression on the yeast surface) were then used to normalize all frequencies of each gene in any given selection.

3.4 DISCUSSION

We demonstrate that pooling a variety of unrelated genes and selecting for binding of their surface expressed proteins to multiple fluorescently labeled targets by flow cytometry enables a rapid assessment of relative affinities and specificity profiles. Such selections distinguish proteins that bind specifically to a desired molecule from those that bind non-specifically to multiple targets. In case of designed proteins, off-target binding can indicate problems with the structural integrity of the protein and exposure of hydrophobic core residues. Monitoring of off-target interactions is also crucial for the development of novel protein-based therapeutics, diagnostics and synthetic sensors or regulators, for recombinant proteins or antibodies. Our assay can be used in early discovery steps to facilitate decisions on lead candidates.

As gene synthesis is becoming cheaper and genomic libraries are becoming readily available, high throughput analysis of protein interactions is becoming increasingly powerful. Highly parallel analyses as described here provide an effective way to extract maximum information content on binding affinity and specificity.

BIBLIOGRAPHY

1. Mobley, D. L. & Dill, K. a. Binding of small-molecule ligands to proteins: 'what you see' is not always 'what you get'. *Structure* **17,** 489–98 (2009).
2. Weber, P. C., Ohlendorf, D. H., Wendoloski, J. J. & Salemme, F. R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **243,** 85–8 (1989).
3. Piran, U. & Riordan, W. J. Dissociation rate constant of the biotin-streptavidin complex. *J. Immunol. Methods* **133,** 141–3 (1990).
4. Freitag, S., Le Trong, I., Klumb, L., Stayton, P. S. & Stenkamp, R. E. Structural studies of the streptavidin binding loop. *Protein Sci.* **6,** 1157–66 (1997).
5. Dougherty, D. a. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* **271,** 163–168 (1996).
6. Wilson, I. a & Stanfield, R. L. Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.* **4,** 857–867 (1994).
7. Boder, E. T., Midelfort, K. S. & Wittrup, K. D. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 10701–5 (2000).
8. Wedemayer, G. J., Patten, P. a, Wang, L. H., Schultz, P. G. & Stevens, R. C. Structural insights into the evolution of an antibody combining site. *Science (80-.).* **276,** 1665–1669 (1997).
9. Yang, P. L. & Schultz, P. G. Mutational analysis of the affinity maturation of antibody 48G7. *J. Mol. Biol.* **294,** 1191–201 (1999).
10. Midelfort, K. S. & Wittrup, K. D. Context-dependent mutations predominate in an engineered high-affinity single chain antibody fragment. *Protein Sci.* **15,** 324–34 (2006).
11. KENDREW, J. C. *et al*. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181,** 662–6 (1958).
12. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9,** 646–652 (2002).
13. Li, Z. & Scheraga, H. a. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 6611–5 (1987).
14. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21,** 1087 (1953).
15. Kuhlman, B. *et al*. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302,** 1364–8 (2003).
16. Koga, N. *et al*. Principles for designing ideal protein structures. *Nature* **491,** 222–227 (2012).
17. Jiang, L. *et al*. De novo computational design of retro-aldol enzymes. *Science* **319,** 1387–91 (2008).
18. King, N. P. *et al*. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* **336,** 1171–4 (2012).
19. Tinberg, C. E. *et al*. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501,** 212–6 (2013).
20. Fleishman, S. J. *et al*. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* **332,** 816–21 (2011).

21. Dahiyat, B. I. & Mayo, S. L. De Novo Protein Design: Fully Automated Sequence Selection. *Science (80-.).* **278,** 82–87 (1997).
22. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. High-Resolution Protein Design with Backbone Freedom. *Science (80-.).* **282,** 1462–1467 (1998).
23. Dunbrack, R. L. & Karplus, M. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.* (1993).
24. Lazaridis, T. & Karplus, M. Effective energy function for proteins in solution. *Proteins* **35,** 133–52 (1999).
25. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12,** 6201–6212 (2016).
26. Leaver-fay, A. *et al.* R OSETTA 3 : An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. **487,** 545–574 (2011).
27. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356,** 539–542 (1992).
28. Drexler, K. E. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U. S. A.* **78,** 5275–5278 (1981).
29. Griss, R. *et al.* Bioluminescent sensor proteins for point-of-care therapeutic drug monitoring. *Nat. Chem. Biol.* **10,** 598–603 (2014).
30. Tomlinson, I. M. Next-generation protein drugs. *Nat. Biotechnol.* **22,** 521–2 (2004).
31. Chan, B. S. H. & Buckley, N. A. Digoxin-specific antibody fragments in the treatment of digoxin toxicity. *Clin. Toxicol. (Phila).* **52,** 824–36
32. de Wolf, F. a & Brett, G. M. Ligand-binding proteins: their potential for application in systems for controlled delivery and uptake of ligands. *Pharmacol. Rev.* **52,** 207–36 (2000).
33. Wu, C.-Y., Roybal, K. T., Puchner, E. M., Onuffer, J. & Lim, W. A. Remote control of therapeutic T cells through a small molecule-gated chimeric receptor. *Science (80-.).* **350,** aab4077-aab4077 (2015).
34. Liu, J. K. H. The history of monoclonal antibody development - Progress, remaining challenges and future innovations. *Ann. Med. Surg.* **3,** 113–116 (2014).
35. Bird, R. E. *et al.* Single-chain antigen-binding proteins. *Science* **242,** 423–6 (1988).
36. Wang, Y. *et al.* Nanobody-derived nanobiotechnology tool kits for diverse biomedical and biotechnology applications. *Int. J. Nanomedicine* **Volume 11,** 3287–3303 (2016).
37. Binz, H. K., Amstutz, P. & Plückthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.* **23,** 1257–68 (2005).
38. Nygren, P.-A. Alternative binding proteins: affibody binding proteins developed from a small three-helix bundle scaffold. *FEBS J.* **275,** 2668–76 (2008).
39. Beste, G., Schmidt, F. S., Stibora, T. & Skerra, a. Small antibody-like proteins with prescribed ligand specificities derived from the lipocalin fold. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 1898–1903 (1999).
40. Arnold, F. H. How proteins adapt: lessons from directed evolution. *Cold Spring Harb. Symp. Quant. Biol.* **74,** 41–6 (2009).
41. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat. Protoc.* **1,** 755–68 (2006).
42. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships.

Nat. Methods **7,** 741–6 (2010).
43. Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science* (2012). doi:10.1126/science.1231434
44. DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F. & Lombardi, A. De novo design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* **68,** 779–819 (1999).
45. Dwyer, M. a, Looger, L. L. & Hellinga, H. W. Computational design of a Zn2+ receptor that controls bacterial gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11255–60 (2003).
46. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. Computational design of receptor and sensor proteins with novel functions. *Nature* **423,** 185–90 (2003).
47. Schreier, B., Stumpp, C., Wiesner, S. & Ho, B. Computational design of ligand binding is not a solved problem. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 18491–18496 (2009).
48. Nivo, L. G. A Pareto-Optimal Refinement Method for Protein Design Scaffolds. **8,** 1–5 (2013).
49. Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G. & Carlson, H. a. Binding MOAD (Mother Of All Databases). *Proteins* **60,** 333–40 (2005).
50. Zanghellini, A., Jiang, L. & Wollacott, A. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* 2785–2794 (2009). doi:10.1110/ps.062353106.to
51. Duhovny, D., Nussinov, R. & Wolfson, H. Efficient Unbound Docking of Rigid Molecules. *Algorithms Bioinforma.* 185–200 (2002). doi:10.1007/3-540-45784-4_14
52. Nivón, L. G., Bjelic, S., King, C. & Baker, D. Automating human intuition for protein design. *Proteins* 1–9 (2013). doi:10.1002/prot.24463
53. Friesner, R. A. *et al.* Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein - Ligand Complexes. 6177–6196 (2006).
54. Trott, O. & Olson, A. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comput. Chem.* **31,** 455–461 (2010).
55. Rossi, A. M. & Taylor, C. W. Analysis of protein-ligand interactions by fluorescence polarization. *Nat. Protoc.* **6,** 365–87 (2011).
56. Jeffrey, P. D. *et al.* 26-10 Fab-digoxin complex: affinity and specificity due to surface complementarity. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 10310–10314 (1993).
57. Höltke, H. J. *et al.* The digoxigenin (DIG) system for non-radioactive labelling and detection of nucleic acids--an overview. *Cell. Mol. Biol. (Noisy-le-grand).* **41,** 883–905 (1995).
58. Holm, L. & Rosenström, P. Dali server: Conservation mapping in 3D. *Nucleic Acids Res.* **38,** 545–549 (2010).
59. Speiser, P. W. & White, P. C. Congenital Adrenal Hyperplasia. *N. Engl. J. Med.* **349,** 776–788 (2003).
60. Choi, J. Recent advances in biochemical and molecular analysis of congenital adrenal hyperplasia due to 21-hydroxylase deficiency. 1–6 (2016).
61. Cherney, M. M., Garen, C. R. & James, M. N. G. Crystal structure of Mycobacterium tuberculosis Rv0760c at 1.50A resolution, a structural homolog of delta5-3-ketosteroid isomerase. *Biochim. Biophys. Acta - Proteins Proteomics* **1784,** 1625–1632 (2008).

62. Schwans, J. P., Sunden, F., Gonzalez, A., Tsai, Y. & Herschlag, D. Evaluating the Catalytic Contribution from the Oxyanion Hole in Ketosteroid Isomerase. *J. Am. Chem. Soc.* **133,** 20052–20055 (2011).
63. Paige, J. S., Wu, K. Y. & Jaffrey, S. R. RNA mimics of green fluorescent protein. *Science* **333,** 642–6 (2011).
64. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–5 (2009).
65. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science (80-.).* **355,** 201–206 (2017).
66. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537,** 320–327 (2016).
67. DePristo, M. a, Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6,** 678–87 (2005).
68. Miller, S., Lesk, A. M., Janin, J. & Chothia, C. The accessible surface area and stability of oligomeric proteins. *Nature* **328,** 834–6 (1987).
69. Feng, J. *et al.* A general strategy to construct small molecule biosensors in eukaryotes. 1–23 (2015). doi:10.7554/eLife.10606
70. Henry, J. T. & Crosson, S. Ligand-Binding PAS Domains in a Genomic, Cellular, and Structural Context. *Annu. Rev. Microbiol.* **65,** 261–286 (2011).
71. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* **12,** 29–34 (2015).
72. Song, Y. *et al.* High-Resolution Comparative Modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).
73. Combs, S. A. *et al.* Small-molecule ligand docking into comparative models with Rosetta. *Nat. Protoc.* **8,** 1277–1298 (2013).
74. Midelfort, K. S. *et al.* Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *J. Mol. Biol.* **343,** 685–701 (2004).
75. Mandell, D. J., Coutsias, E. A. & Kortemme, T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6,** 551–552 (2009).
76. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11,** 499–507 (2014).
77. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–45 (2008).
78. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17,** 333–351 (2016).
79. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–6 (2010).
80. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491,** 138–42 (2012).
81. Whitehead, T. a *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30,** 543–8 (2012).

VITA

Jiayi Dou was born in Tianshui, a beautiful city in west China. She studied chemistry and physics at the University of Science and Technology of China and earned a Bachelor of Science in Chemistry in 2010. She moved to the United States of America for continuing graduate-school study at the University of Washington, Seattle. In 2017, she earned a Doctor of Philosophy in Bioengineering in the Biological Physics, Structure and Design program, under the advisement of Dr. David Baker.