Computational Design of Small Molecule Binding Proteins

Austin Day

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington 2015

Reading Committee:
David Baker, Chair
Patrick Stayton
Wendy Thomas

Program Authorized to Offer Degree:
Bioengineering

University of Washington

Abstract

Computational Design of Small Molecule Binding Proteins

Austin Lewis Day

Chair of the Supervisory Committee:
Dr. David Baker
Biochemistry

Protein design is still in its infancy, yet there have been many impressive examples of success in designing proteins to fold into a predictable structure [6, 4], to catalyse enzymatic reactions [10, 13], or to bind a specific protein [16], DNA sequence [17, 11], or small molecule target [8]. Each of these successes in the field is a major milestone, but protein design still lacks a generalized solution for reliably repeating these successes on future targets. The design of proteins capable of binding small molecules is particularly challenging due to the necessity to accurately understand and computationally model atomic scale physiochemical principles. We work towards this goal because being able to reliably design small molecule binders would allow for faster, and more efficient creation of detection elements for biosensors, sequestration proteins to aid in dialysis, and orthogonal binding tags for use in biotechnology applications. Even a modest advantage gained through computational design would allow for faster results when using more traditional directed evolution search methods. Since control of molecular specificity at the atomic level is essential

for diagnostic applications in which multiple similar molecules are present and require discrimination from each other, computational modelling can be especially useful because the desired molecular specificity can be explicitly incorporated into the design. Such cases exist with the detection of tetrahydrocannabinol (THC) from the non-psychoactive cannabidiol and downstream metabolites present in users of marijuana, and in the detection of 25-hydroxycholecaliferol from 25-hydroxyergocalciferol, a clinically important distinction of vitamin D3 metabolites where the two compounds differ by a single methyl group. With this particular goal in mind, we have developed a computational protocol, using the Rosetta software package, capable of designing protein models with good shape complementarity, favorable chemical environments, and designed molecular specificity for a target protein-ligand interaction. This protocol was optimized over many iterations and incremental successes into a final revision that is capable of creating protein binders for the ligands 25-hydroxycholecaliferol, the hormonally active form of vitamin D3, and tetrahydrocannabinol, the primary psychoactive ingredient in cannabis. In addition to learning how to make successful protein binding designs, we also attempted to recover non-functional designs through stabilization. Using an algorithm for inserting proline substitutions into failed designs, we believe we have identified a lack of stability as one potential cause for failed binding protein designs. The protocol improvements learned from both our successful and recovered function binders should move us towards a more generalizable and reliable method for designing future protein-ligand interactions.

0.1 Terminology, definitions, and abbreviations

Cholecalciferol: Vitamin D3.

Ergocalciferol: Vitamin D2.

25-hydroxycholecalciferol: The 25-hydroxylated form of vitamin D3 which is hormonally active.

25-hydroxyergocalciferol: The 25-hydroxylated form of vitamin D2.

THC: Tetrahydrocannabinol, the primary psychoactive ingredient in cannabis.

CBD: Cannabidiol, a chemical found in cannabis known for its potential health effects, related to THC but not psychoactive.

PDB: Protein Data Bank

Rotamer: One confirmation of an amino acid side chain.

"Functional" or "Binds": A protein that is referred to as functional or that binds its target is defined as showing a PE signal at least 2x greater than a negative control of identical cells incubated for equal time with no labelled ligand when assayed using flow cytometry and yeast surface display as described in the methods section of chapter 2. The concentration this assay is performed at varies depending on the availability of the labelled ligand target, but typically is the equivalent to between 1-20uM of non-avid ligand.

"Non-functional": A protein that is referred to as non-functional is defined as showing a PE signal less than 2x greater than a negative control of identical cells incubated for equal time with no labelled ligand when assayed using flow cytometry and yeast surface display. When a biotinylated ligand is not available, a BSA-Biotin conjugate is used in its place. The concentration this assay is performed at varies depending on the availability of the labelled ligand target and is the same as a positive or unknown binder it is being compared to.

"Recovered" or "Restored function" Binders: A protein that is designed to bind a target ligand but only gained function after incorporation of one or more mutations from either directed evolution or another computational method.

"Serendipitous" Binders: Proteins that were designed to bind a target ligand but experimentally bound an off-target ligand.

"Successful" Designs: Proteins that were designed to bind a target ligand and successfully did so without additional modification to the design.

Contents

Chapter 1: Introduction and Background

Small molecules are ubiquitous and critically important in biological processes due to their ability to cross membranes [52] and their natural use in energy storage and cell signalling. [85] They comprise both beneficial and harmful substances such as drugs, antibiotics, nutrients, toxins, pollutants, preservatives, and carcinogens. [87, 62] The importance of small molecules is undisputed and so the ability to measure concentrations or sequester small molecule targets in environmental or biological settings is crucial to further our understanding of basic biology, to aid in the creation of new therapeutics, and to protect ourselves from dangerous substances. Of critical importance to the quality of any application involving molecular detection is the affinity and specificity.

In all applications, the concentration of the target molecule in the unprocessed sample will determine what affinity a sensing protein must have in order to be useful. In the example of vitamin d3, concentrations in human blood are typically in the low nanomolar range [82], meaning that a binding protein must have a Kd in the same range in order to be useful. For oral fluid testing of THC, the concentration range can vary greatly. At 12 hours after smoking, the concentration of THC can be approximately 3nM, whereas 15 minutes after smoking oral fluids can contain as much as 18uM. [27] In other applications such as with creating tags for biotechnology applications, the ability to bind tightly and conditionally, such as in a certain pH range or temperature, may additionally be useful.

Existing technologies used to create sensing elements for small molecules, such as generating antibodies, selecting for RNA/DNA aptamers, or creating molecularly imprinted polymers, have advantages and disadvantages depending on the setting in which they are used. Our approach for creating high quality sensing elements is through the use of computationally designed proteins us-

ing the software package Rosetta. Proteins can be explicitly engineered with an environmental setting and application in mind and are readily evolvable to function in a wide variety of situations. Engineered proteins have the potential to provide solutions where other sensing elements cannot, and understanding that benefit in its context of competing technologies is important in knowing how and when computational design techniques hold the advantage.

1.1 Existing Molecular Recognition Elements

1.1.1 SELEX: DNA/RNA Aptamers

Systematic evolution of ligands by exponential enrichment, or SELEX, is a technique that can be used to isolate RNA or DNA aptamers capable of binding a desired target, such as a small molecule, protein, or cell surfaces. [78] In this technique, one starts with a large pool of randomized DNA sequences, puts them through repeated rounds of selection for binding against an immobilized target substrate, and PCR amplifies the enriched sequences. In this way, aptamers have been identified that are able to distinguish between individual functional groups on ligands, and in one example, with up to 10000 fold selectivity for a single methyl group. [67] Aptamers can also be very tight, capable of binding their targets with low nanomolar to picomolar affinities. [39, 45, 22] Aptamers can be chemically stable in more environments than proteins are due to their simple structural components and a strong evolutionary pressure for DNA to be chemically stable, but their structure is highly dependent on solution conditions. Additionally, binding of an aptamer to a target often results in large conformational changes, which may be a useful characteristic for using them in sensing applications, but also presents a potential unknown variable in an application setting since such movements are difficult to predict. [78]

Since RNA and DNA are easily degraded in blood, aptamers can suffer from

unwanted persistence issues if used in an *in vivo* setting. Aptamers often have a hard time with binding small molecules, although this is not necessarily a unique issue to aptamers. [55] Less than a quarter of existing apatmers, as of 2012, target small molecules, and the majority bind to larger targets such as proteins or cells. [76] Because aptamers are made up of nucleotide subunits, they have relatively similar shape and chemical properties when compared to amino acids. This limits the possible number of solutions for functional binding sequences and may result in it being more difficult to find binders of an adequate affinity or specificity. Additionally, RNA/DNA is negatively charged because they are linked by phosphate groups, which may not be well suited for binding certain classes of negatively charged targets. For example, fewer solutions may exist in RNA/DNA structure space for binding a non-polar molecule that will optimally require a very hydrophobic environment, due to the charged and polar nature of nucleotide subunits.

1.1.2 Molecular Imprinted Polymers

Molecular imprinted polymers (MIPS) are created by co-polymerizing the target analyte in the presence of the cross-linking monomers. This essentially creates an "imprint" of the target molecule in the polymer matrix. After the analyte is removed, a pocket remains that is capable of rebinding the target with very high specificity. [61] MIPS are unaffected by heat or pH to a much higher degree than biomolecule based binders. [47] Relative to most other methods for creating biomolecule sensing elements, creating MIPS binders are almost guaranteed because the process is so simple and straight forward. [48]

MIPS, however, often have high cross-reactivity and aren't readily amenable to chemical modifications. [68] The use of a homogeneous polymer matrix theoretically limits the diversity in the types of interactions that can be made to a particular ligand and in turn may limit the types of molecules that are able

to be bound at a needed affinity. This homogeneous matrix may additionally present specificity problems for ligands of certain shape or large size, such as large, flat ligands with few functional groups.

1.1.3 Antibodies

Antibodies are proteins naturally produced by plasma cells and are used by the immune system to bind and neutralize foreign substances. They are the most widely used binding protein in commercial applications and have been used for sensing applications since 1959. [21, 70] There is so much infrastructure set up to raise an antibody against a specific target that companies often need a very good reason to choose an alternative method when in need of a sensing or neutralizing compound. Some of the best antibodies can achieve dissociation constants in the femtomolar range [54] and are able to discriminate between single functional groups and chiral compounds. [83]. Of great benefit for *in vivo* applications, it is often much easier to go from inception to clinical trials with antibody products because of humanization technologies.

Antibodies arise through a selection mechanism within a host organism, and as a result, the desired specificity against a particular functional group of a target molecule isn't guaranteed. In fact, antibodies often cross react with many proteins other than the target it was raised against. [73] Additionally, because raising antibodies relies on the host animal's immune system, it is sometimes not even possible to raise immunogenic derivatives for all potential analytes. [81, 66] Antibodies often times have undesirable biophysical properties, such as poor stability or a high propensity for aggregation, which can limit where they can be used. [25, 46, 5, 38] Stability issues are usually more prevalent when using human antibody fragments, which can be a problem for therapeutic applications, but can sometimes be mitigated through considerable amounts of engineering and directed evolution. Antibodies are difficult to produce in

prokaryotic systems due in part to their complex folding requirements and post translational modifications. This can lead to difficulty in using them as the sensing element in *in vivo* detection systems.

1.1.4 Peptide Binders

Peptides are defined differently from proteins because of their relatively small amino acid length and inherently disordered structure. They're very interesting in that they can form various tertiary structures that can interact with many types of targets, almost acting like disembodied binding loop regions of antibodies. Because the binding of a peptide to a target is less affected by the need for a well defined structure, peptides can be exceptionally stable under a variety of conditions where antibodies, RNA, or other larger proteins would not be, such as in detection of soil samples treated with organic solvents. [3] Using selection techniques, such as phage display, one can fairly easily select for peptides that bind a target of choice from randomized, synthesized libraries in a process much less involved and less costly than antibody production. Additionally, peptides have relatively high biocompatibility and low immunogenicity. [84]

Peptides are by definition small, and so their potential is limited based on the maximum available surface area for potential interaction with a target. Being amino acid based, peptides may also be susceptible to proteases in the environment, as opposed to MIPS or DNA/RNA. Peptides generally aren't good at penetrating cells either, except for very specific sequences. [84] Additionally, any application where a defined three dimensional structure would be beneficial or required, such as in the detection of small molecules in high temperature settings where peptide flexibility may be amplified beyond acceptable limits.

1.1.5 Protein Binders

Protein binders are, for the purpose of this comparison, defined separately from antibodies. Even though antibodies consist of amino acids and are amenable to similar design-ability as other proteins, they are relatively restricted in their composition and biophysical characteristics. This restriction results in significantly different optimal use cases for antibodies when compared with other designed proteins, as described by the limitations of antibodies mentioned previously.

Proteins are used in nature to sense environmental conditions, utilize nutrients, are involved in metabolism, and are very amenable to evolution due to their high diversity and potential for chemical modification. Proteins can function in extreme environments and in blood [20], be very specific for their targets [8], and bind with femtomolar affinity [69]. This gives designed proteins an advantage for *in vivo* applications over aptamers, which often cannot survive well in blood, and antibodies, which are often generally unstable. [25] Proteins are the natural choice for small molecule sensing applications in biological systems.

Many techniques have been developed to screen large protein or peptide libraries for a desired activity such as biopanning [35], FACS [77], or *in vivo* survival assays. [17] All of these techniques share the commonality that they can be improved by smarter sequence sampling and library generation because of the large combinatorial search space. The aspect of using computational design methods to guide library creation adds value to using proteins as sensing elements. Random protein screening methods can therefore be viewed not as competition to computational protein design, but as a complement. As computational methods improve, the resources needed for these selection techniques will become less, saving time, money, and resources.

Computational protein design brings with it the advantage of exquisite con-

trol over specificity. Discrimination between similar small molecules can be explicitly designed from the onset to guarantee the specificity needed for a particular application. Designed binders can also be tailored for a specific binding application, as opposed to using or re-purposing a naturally occurring protein, which may have had alternative or less relevant evolutionary pressures dictating its function and specificity. Similar to SELEX and MIPS, designed proteins can target molecules that are non-immunogenic, toxic, or target very specific sub-regions that may be difficult to target using antibodies, due to the inherent epitopes an animal immune system may already contain. [55] However, one of the primary disadvantages of using proteins as a sensing element is that it requires an aqueous environment.

Having a design repertoire of 20 amino acids allows for high chemical diversity and the potential to create sequences with high complexity. An analysis on aptamer based small molecule binders found that more bits of information, as measured by Shannon Entropy [74], may be needed to bind smaller ligands, or ligands with higher entropy by measure of degrees of freedom, with the same affinity as their equivalent larger ligands. It suggests a sort of trade off between affinity, ligand entropy, and information content in aptamer binders, where higher information content is needed in order to obtain tighter binders for smaller ligands. [55] Because proteins have 20 amino acids in their repertoire, it can follow that proteins may encode information to a higher density than aptamers per subunit. If this trade-off holds for protein sequences as well, it suggests that proteins may be better suited to successfully bind smaller ligands at higher affinities than RNA/DNA based aptamers.

1.2 Rosetta

Rosetta is a software program that is being developed by many labs worldwide, but originated in the lab of David Baker at the University of Washington. [9] Rosetta is the primary software package used for our design and modelling efforts. Rosetta provides a framework for representing and manipulating protein structure and identity. It also contains many potential functions for computing interaction energies within and between represented macromolecules. Because the problem of predicting protein structure is NP-Complete [79], Rosetta also includes non-linear optimization methods for finding low energy configurations in many situations. Rosetta has already been successfully applied to many design problems including structure prediction, enzyme design, endonuclease design, RNA-folding, limited ligand-protein interactions and protein-protein interface design. [10, 13, 15, 7, 12, 6, 8]

The Rosetta score function is a metric used to approximate the free energy of macromolecule interactions. The score function consists of a Lennard-Jones potential that favors tightly packed residues, the Lazaridis-Karplus implicit solvation model [58] that favors hydrophobic amino acids in the interior of the protein and polar amino acids on the surface of the protein, an orientation dependent hydrogen bonding term [32], and torsion potentials derived from structures in the PDB. [26] There are also knowledge based terms such as the probability of observing a sequence given the structure and a weak electrostatics term that takes into account the probability of seeing two amino acid types near each other in native structures. [75]

There are assumptions made in order to simplify the calculation of the score function. These include scoring of specific protein states, rather than scoring through simulation via molecular dynamics, treating the solvent as a continuum, and using discrete, backbone dependent rotamer libraries. [26] Some of

the limitations to the score function includes an incomplete representation of entropy, simplified electrostatic terms due to induced polarization effects and pairwise calculations [14], and limited modelling of backbone movement.

1.3 Difficulty in Designing Protein Ligand Interactions

One of the primary driving forces in molecular recognition is the hydrophobic effect, a rationalization of the insolubility of hydrophobic molecules in aqueous solution. Although the mechanism of the hydrophobic effect isn't completely understood, [50, 40, 31, 63] it is thought to involve a gain in entropy by displacing ordered water molecules from hydrophobic surfaces at the binding interface and a favorable enthalpic change due to stronger hydrogen bonds made between water molecules in the bound state compared with the apo-state, as well as direct polar interactions between the ligand-protein complex. [80] The design approach we are adopting, follows the "lock and key" notion in which the hydrophobic effect is satisfied. In Rosetta, the hydrophobic effect is modelled through the use of an implicit solvation model that favors solvent exposed polar atoms and tightly packed hydrophobic atoms. [58] Entropic terms are partially incorporated via this solvation model and also through knowledge based score terms. Enthalpic effects are captured through hydrogen bonding scores [32] and approximated electrostatic calculations. [75]

Even knowing partially what drives the protein ligand interaction, protein design in general still suffers from a vast combinatorial search space. The design of protein-ligand interactions suffers from an incomplete understanding of the physical principles underlying molecular recognition as each small molecule target has unique chemical properties. The creation of proteins that bind small molecule targets is especially challenging compared with targeting proteins or cells, as evident by the limited success in both directed evolution experiments

[76] and computational design methods [8], when compared with binders for larger substances. [16, 18] [55, 43].

One theory as to why small molecules are difficult to bind relative to larger molecules, cells, or proteins, is in part because there is less surface area available for favorable interactions. In contrast, relatively large protein-protein interfaces can effectively sum together their interactions to create a very energetically favorable interface without the explicit need for any one individual interaction to be extremely strong. Binding a smaller target with the same specificity and affinity as a larger target would require that each interaction be energetically more favorable on average than the interactions in a larger interface. A random interaction between two compounds will more likely than not be a weak, non-specific interaction since if the opposite were true, we would instead have more difficulty in creating weak, non-specific interactions, which we do not see in practice. It follows that there must be fewer solutions in protein sequence space for high affinity, specific interactions than for weak, non-specific interactions. With fewer possible high affinity configurations available for protein-ligand interactions, the probability of finding such an interaction in a random library is therefore lower than for a larger molecule. This relates to difficulty in computational design efforts in that there is less room for error. The interactions made to a small target must be near perfect and any error will likely not be tolerated. Compounding this problem is the relative lack of information on ligand-protein interactions. Plenty of data is available in multiple databases related to protein-protein interactions, as those interactions essentially make up the core of all proteins. Protein-ligand interactions, however, have far fewer examples, but also the chemical diversity for small molecules is effectively infinite, making general interaction rules difficult to come by.

Chapter 2: Computational Methods for Designing Small Molecule Binding Proteins

The following protocol is given as a generalized procedure that can be applied to any ligand types and is the basis protocol that protocols discussed in later chapters will be compared with.

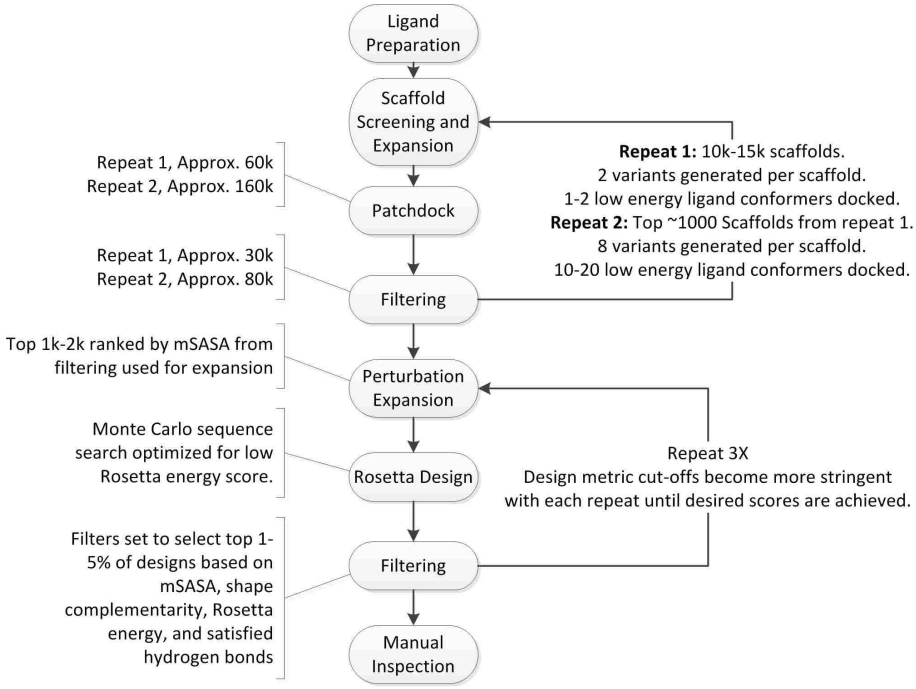

Figure 1: A schematic overview of the final iteration of the small molecule binder design protocol.

2.1 Ligand Model Preparation

PDB models for the target ligands are generated using Avogadro [51] and conformers are generated using Omega Openeye [49]. Our experimental assays require a PEG-Biotin linker attached to the target molecule, however, because of the flexibility of such a linker and the large number of additional degrees of freedom it would introduce into the model, a truncated form of the PEG-Biotin linker is generally modelled instead that consists of the first five atoms of the linker.

2.2 Scaffold Selection and Ligand Placement

In order to identify scaffolds with native binding pockets of appropriate size and chemical environment for our target ligands, we iterate twice through our scaffold selection protocol. The scaffolds used for iteration one are taken from the PDB based on the criteria below, with the purpose of identifying relatively well behaved proteins with appropriately sized native binding pockets, as well as characteristics that will aid in experimental validation of function.

1) The presence of native ligands with similar structural motifs as our target ligands, or the presence of native ligands of size within %40 of the target ligand by atomic weight.

2) An X-Ray structure resolution of below 2.5 Angstroms.

3) Consists of a single protein chain in its biological assembly.

4) Has been previously expressed in E. Coli.

5) Has a chain length of no more than 350 amino acids.

Scaffolds which have been previously shown in our lab to be capable of binding other ligands and that behave well when expressed are also included, as well as scaffolds in the same Pfam [2] family. These scaffolds include 1Z1S, 1OHO, 3FKA, 3HX8, and 3AKR. Additionally, all scaffolds in the MOAD [23] are also included. The approximate number for this initial set is generally in the many thousands, depending on which update of the PDB scaffolds you use.

This large set of scaffolds is expanded to include variants with all residues changed to alanine without mutating high structural residues such as aromatics, cysteine, proline, or glycine residues. This is done in order to ensure that scaffolds are searched based on a general pocket shape, instead of whatever shape the native functional residues bias them towards. These variants are then put through the Patchdock [71] procedure using the single lowest energy ligand conformer, or with a hand selected conformer that is chosen based on low energy as well as an extended configuration that maximizes the potential interaction area of the ligand with the protein. Pathdock will dock the target ligand into a scaffold using a fast geometry-based algorithm. The goal of this algorithm is to quickly identify scaffolds that offer good shape-complimentary towards our target ligands. Patchdock uses a Connolly dot surface representation [28] to divide up the protein and ligand into concave, convex, or flat patches. The surfaces

are then matched together in order to generate favorable ligand positions. The docked models are then filtered using the following three criteria.

1) The secondary structure content of the 8Åshell around the ligand must be greater than %80.

2) The direction of the vector between the last two atoms in the linker point away from the center of mass of the protein.

3) A temporarily modelled linker from the last two atoms of the linker model is able to extrapolate out 10Åand not come within 3.5Åof an atom of a secondary structure element in the scaffold protein.

4) A modified measure of solvent accessible surface area (mSASA) of the ligand in the docked model is greater then 0.6.

This mSASA extrapolates all outward facing, 3Åvectors to all corners and center-faces of a cube centered around all user defined atoms in the ligand and calculates the percentage of these points that are within 2Åof any atom in the protein. Docked models that pass all of these filters are ranked by mSASA and the top 2000 are used for a second Patchdock iteration that significantly increases the sampling. One of the primary differences is that the second iteration involves expanding each scaffold into eight variants instead of two. The types of modifications for each variant are described below.

Unless explicitly stated, native TRP, PHE, TYR, PRO, and GLY residues are kept the same.

Variant 1) Native scaffold.

Variant 2) Native with charged residues trimmed to closest non-charged

residue by approximate shape: Residues are native with the following changes.
GLU/ASP/ LYS /ARG
Changed to:
GLN/ASN/MET/GLN

Variant 3) Minimally sized residues with hydrophobic bias: Residues are native with the following changes.
GLN/SER/GLU/ASP/ARG/TYR/THR/HIS /ASN/MET/ILE /LYS/LEU
Changed to:
ALA/ALA/ALA/ALA/ALA/PHE/VAL/VAL/ALA/ALA/VAL/ALA/ALA

Variant 4) Minimally sized residues with hydrophilic bias: Residues are native with the following changes.
GLN/ALA/GLU/ASP/ARG/VAL/HIS/ASN/MET/ILE /LYS/LEU
Changed to:
SER/SER/SER/SER/SER/THR/THR/SER/SER/THR/SER/SER

Variant 5) Residues are changed to their closest hydrophobic equivalent: Residues are native with the following changes.
GLN / SER/ GLU/ ASP /ARG/TYR/THR/ HIS / ASN /LYS
Changed to:
MET/ALA/MET/LEU or MET/MET/PHE/VAL/PHE or ILE/MET or LEU/MET

Variant 6) Residues are changed to their closest hydrophilic equivalent: Residues are native with the following changes.
ALA/ARG/VAL/ HIS / MET /ILE/ LYS/ LEU
Changed to:

SER/GLN/THR/TYR or THR/GLN or GLU or LYS/THR/GLN/ASN or ASP

Variant 7) All residues are changed to alanine, including aromatics.

Variant 8) All residues are changed to alanine except for aromatics which are kept native.

The second important difference between the first and second Patchdock iterations is that the top 10-20 low energy ligand conformers, instead of just one or two, are each docked using the Patchdock protocol. After Patchdock, the models are again filtered using the same criteria for iteration 1 as mentioned previously. These models are then used as inputs for Ligand Perturbation Expansion.

2.3 Ligand Perturbation Expansion

The Ligand Perturbation Expansion protocol takes in a ligand-protein model and generates variants with slight translational and rotational perturbations to the ligand position. The amplitude of perturbation and density of sampling is varied depending on the stage of design. A more coarse grain sampling procedure is used initially and consists of three rotational perturbations starting from 0 degrees with 15 degree increments and two translational perturbations starting from 0Åwith 0.75Åincrements. In subsequent rounds, this sampling is increased to twenty rotational perturbations starting from 0 degrees with 2 degree increments and two translational perturbations starting from 0Åwith 0.25Åincrements. These perturbation variants are then each designed using an implementation of Rosetta Design achieved through the Rosetta Scripts inter-

face. [41] This design step uses alternating rounds of minimization, sequence redesign, and filtering based on a shape complementarity cut-off of 0.6, a solvent accessible surface area cut-off of 0.7, the presence of an appropriate number of hydrogen bonds being made to the ligand, and a Rosetta protein-ligand interface score of less than approximately -6 Rosetta energy units (REU), although this last metric varies depending on the size of the ligand and the -6 example was used for ligands THC and D3-OH. All of these filter metrics are meant to be rather lenient, only selecting against very bad designs. The shape complementarity is the only metric that is relatively difficult to pass and is effectively the more stringent selection criteria. The goal of this stage of design is to identify sequences that minimize the predicted energy between the ligand and the protein.

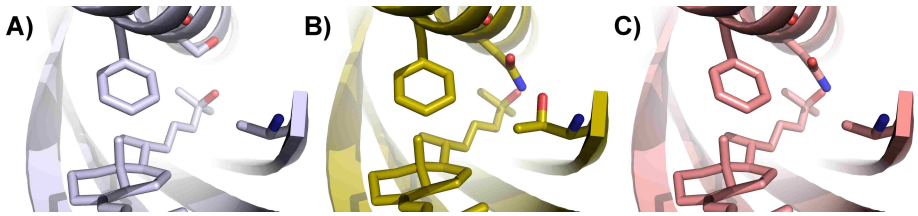

Figure 2: Simplified representation of the ligand perturbation protocol. A-C) Three selected perturbations where slight ligand movements results in discrete sequence changes after Rosetta design.

2.4 Filtering

After Rosetta Design, models are filtered based on criteria that can vary depending on the type of ligand being targeted. For THC and 25-D3, primarily hydrophobic ligands with one possible hydrogen bond, the primary criteria used is mSASA, the Rosetta interface score, shape complementarity, and the presence of an appropriate hydrogen bonding residue. For more hydrophobic ligands, this light filtering is generally sufficient to create designs that will at least show a binding signal. The challenge for hydrophobic ligands has been guaranteeing

that there is appropriate specificity, since hydrophobic interactions can be very general and therefore many ligand binding modes may exist since we do not currently include multi-state negative design against alternative ligand conformations.

For more hydrophilic ligands, such as biotin, with more than two potential hydrogen bonds, the filter for an appropriate number of hydrogen bonds is typically turned off and hydrogen bonds are instead designed in a semi-manual manner. Rosetta will often inaccurately model polar atoms and will conservatively place a hydrophobic residue near a polar group during design. With hydrophilic ligands, many examples where, say, an alanine or a valine can be changed manually to make a hydrogen bond with a serine or threonine substitution are seen and many designs can be saved by making these changes manually. In addition to more manual insertion of hydrogen bonding residues, more involved scoring metrics are used for design and filtering of hydrogen bonds and are described below.

2.4.1 Boltzmann Electrostatics Design

The Poisson-Boltzmann equation, described by Lu *et al*, [59] is used to calculate the electrostatic forces between molecules in ionic solutions. We implemented this model to calculate the electrostatic interactions between the surfaces of our ligand and the protein binding site for more hydrophilic ligands, like biotin, that can make 4 or more hydrogen bonding interactions. The electrostatics model calculates an all-body electrostatic field, as opposed to the standard pairwise calculation previously used through Rosetta. Our implementation does a scan of all residues within the binding pocket to find substitutions that stabilize the bound state based on the electrostatics score, the $\Delta\Delta$G, a measure of the difference in Gibbs free energy between bound and unbound ligand states, and

the total Rosetta score, but also does not destabilize the protein in the absence of the ligand.

2.4.2 Pareto Optimization

In an additional round of Rosetta design, we score each design based on several score metrics and select designs that are Pareto efficient, that is, on a multidimensional surface where each score metric is a separate dimension, all designs along a the leading edge of that multidimensional surface are considered efficient. Each design is evaluated by total Rosetta score, the solvent accessible surface area, hydrogen bonding scores, shape complementarity of the ligand, the Rosetta Holes packing score, and the interface energy. Designs that are determined to be Pareto efficient are kept. These designs are further filtered based on a very strict rotamer score cut-off, such that only near native configurations of individual side chains are allowed.

2.4.3 RosettaDock

RosettaDock is a protocol that allows us to computationally validate our designs by performing protein-ligand docking that explicitly models full side-chain, backbone, and ligand flexibility. [36, 37] Monte Carlo sampling is used to explore all associated degrees of freedom. Five thousand runs of this protocol for each of our designs allows us to generate a "docking funnel". This is a plot of how much our ligand has moved from it's starting point versus the Rosetta energy of the complex. In the cases where Rosetta finds a global minimum of energy and the ligand shows little to no movement from our initial model, we should see a "funnel" of data points in a plot of Rosetta energy vs RMSD from our initial dock positioning that will show that the lowest energy configuration is the one near the designed positioning. Native small molecule binders have this funnel character, as well as the well known example of the bioti-streptavidin

interaction. If our designs show a similar funnel without alternative low energy minima, then we can conclude that the design has successfully passed the docking validation criteria.

2.4.4 Molecular Dynamics

Our Rosetta design procedures use a fixed backbone approach and are therefore unable to realize changes in protein structure or solvent accessibility that occur from backbone motions. In order to partially account for this type of information in our final designs, and to get an idea if our protein-ligand complexes are stable after we make our functional mutations, we used the molecular dynamics package AMBER for simulations [34]. For each model, the protein is immersed in a box of up to 16,600 explicit water molecules and simulated for approximately 20ns at constant pressure with a periodic-boundary. Data collected from these runs include pair wise distance distributions, hydrogen bond directionalities, solvent accessibility, and root-mean-square displacements (RMSDs) relative to the original Rosetta design position. The molecular dynamics data is used to inform design mutations with the goal of stabilizing fluctuations of the entire protein or of the ligand in the binding pocket.

2.4.5 Manual inspection

Filtered designs are then inspected manually using Pymol [72] and FoldIt [29]. FoldIt can be thought of as a graphical interface for the Rosetta score function, design, and minimization. During this step, we filter out designs with obvious problems, such as when ligands are placed in non-native binding pockets, when ligand positions do not allow adequate space for a biotin conjugated linker to reach the solvent, and when voids around the ligand position are of insufficient volume to accommodate a discrete water molecule. During this step, we attempt to use our chemical intuition and experience looking at native binders to ensure

that there is nothing unexpected in what our automated design procedures have generated.

The most common type of manual changes are made for polar groups on ligands, as mentioned previously. Partly due to an incomplete electrostatics and solvent model, polar atoms are very difficult for Rosetta to design properly. Often designs will either not contain any hydrogen bonds to a polar ligand, even when upon manual inspection it can be accommodated with minor minimization. In other instances, a polar residue will be placed and will be making an interaction with the target ligand, but the residue will not be backed up by any other interactions in the protein. These unsatisfied polar groups are a problem because they are usually not present in native structures, but because we use an implicit solvent model, Rosetta may think there is enough room for a water molecule to make an interaction to satisfy a polar group even when there isn't, essentially creating a vacuum in the predicted model, a very unfavorable feature in a design.

2.5 In Silico Directed Evolution

Another strategy that has been implemented approximates a genetic algorithm type optimization. The designs are put through successive rounds of Ligand Perturbation and Rosetta design using progressively finer perturbations and score cut-offs during each progressive iteration. Each iteration essentially removed all but the top %20 of designs based on shape complementarity [60], Rosetta ligand-protein interface score, and solvent accessible surface area, although any set of weighted score metrics can be used. The remaining designs are then used for the next ligand perturbation expansion and Rosetta design iteration. After several iterations, designs are then manually inspected again and the top designs are selected. This type of iterative procedure generally results in designs

that have better selected score metrics. This procedure can often result in very subtle amino acid substitutions that wouldn't have been found with a coarse search protocol.

2.6 Revert to Native

Once we are satisfied with our designs and scores, we apply a reversion protocol that will revert any residue back to native so long as it doesn't negatively affect the score attributed to the protein-ligand interaction. Often our design procedure will mutate surface residues or positions that are not critical to the designed interaction. Due to the potential error or instability Rosetta suggestions may introduce, we opt to remove these extraneous mutations by reverting them back to their native residues. A final manual inspection is done after this step to ensure no key designed residues were reverted unintentionally.

2.7 Order Preparation

The final designs are converted into DNA sequences that are optimized for E. Coli codon expression by using DNAworks [57] and outsourced for synthesis.

2.8 Experimental Validation

2.8.1 Yeast Surface Display and FACS

Yeast surface display is a widely used method for maturation of antibody affinity, specificity, and stability. [24] This surface display system consists of the yeast agglutinin protein Aga2p, which is attached to the yeast cell wall through disulfide bonds to Aga1p. Aga2p is fused to the target design, which is also fused to a C-terminal c-Myc tag. The average number of surface complexes per cell is approximately 50,000. Cell-surface expression is accomplished by including the fusion protein encoded within a modified version of the pCTCON2

plasmid via a galactose-inducible promoter. Design plasmids are transformed into yeast strain EBY100 according to the Benatuil protocol. [19] A schematic of the labelled surface complex is shown below in figure 8.

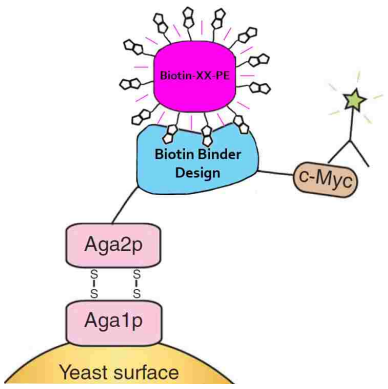

Figure 3: A schematic of the surface display complex for testing a biotin binding design. Biotin-XX-PE is the fluorescent molecule phycoerythrin that is conjugated to approximately 12 biotin molecules. This biotinylated label can be extended to other ligands through the use of a biotinylated target ligand conjugated with streptavidin-PE, which is commercially available. The terminal c-Myc tag is labelled with an anti-myc FITC conjugate and reads out as a surface expression level.

Cell-surface expression of the design is monitored by labelling the cells with a fluorophore that is attached to an anti-c-Myc antibody. Binding is detected by labeling with the target ligand attached to a second fluorophore. In the case of biotin, we use fluorescein isothiocynate (FITC) conjugated to an anti c-Myc antibody and a biotinylated phycoerythrin (PE) fluorophore, but the biotin-PE fluorophore can be exchanged for another biotinylated ligand target conjugated to streptavidin-PE. The FITC signal corresponds to the expression of the entire surface complex and the phycoerythrin signal corresponds to the binding of the design to the biotinylated label. For an example plot, see figure 15.

When testing our designs, we insert them into a pCTCON2 yeast surface display vector, transform into yeast strain EBY100 as mentioned previously. We culture and express the proteins in yeast according to the Wittrup protocol [24]

using an induction time of 24 hours at 30°C. This induction temperature can vary depending on the type of sort we are aiming to do. For early sorts where we are trying to detect a binding signal from new designs or from a library where there is no initial binder, we incubate at 4°C in order to potentially save less stable, but functional binders. In later sorts where we aim to optimize the binding signal, we increase the temperature up to 50°C in order to select against less stable variants. The buffer used in all the cell incubations is phosphate buffered saline at pH 7.5 with 1% bovine serum albumin and 100mM PEG 200 (PBSFP). PEG200 is included in our buffers in order to help complete with interactions against the PEG linker attached to our biotinylated target labels. For avid labelling conditions, either a ligand-biotin-BSA conjugate or a dextran conjugate is used are their compositions are described below.

For the ligand THC, a BSA conjugate is commercially available (Fitzgeral, US Biological). The commercially available BSA conjugate is biotinylated using EZ-Link Sulfo-NHS-Biotin (Thermo Fisher). The reaction is carried out in PBS at pH 8 using a ligand conjugated BSA concentration of 2mg/ml and adding 14.3ul of 10mM EZ-Link Sulfo-NHS-Biotin. The reaction is allowed to proceed for 2 hours on ice. The reaction is then dialysed and used for labelling with streptavidin PE as described previously.

A dextran conjugate can be prepared for all ligands for which a ligand-biotin conjugate is available by mixing 2.18ul of 20uM biotinylated ligand with 2.5ul 6.92uM 70K Biotin Dextran (Life Technologies). 3.82ul of 1mg/ml streptavidin PE (Invitrogen) is added to this mixture, mixed via vortexing, then allow to sit on ice for 1 minute. 1ul of anti-myc FITC antibody (ICLLAB) is added and PBSFP is added to a final volume of 26ul. This dextran conjugated label is then incubated with the designs for 2 hours at 4°C on a tube inverter. The cells are then spun down at 3000rpm for 3 minutes, washed twice with 1ml PBSFP,

re-suspended to a maximum of 5 million cells per ml in PBSFP and read on an a BD Accuri C6 flow cytometer. This dextran conjugate provides an avid label, and variants where 2 to 4 times the amount of biotinylated label are also used to increase avidity. The equivalent concentration of a non-avid label for the base dextran conjugate setup is about 30-50uM.

For affinity maturation using a non-avid label, the designs are incubated with biotin conjugated ligands for 2 hours at 4°C on a tube inverter. The cells are spun down at 3000rpm for 3 minutes and washed twice with PBSFP. The cells are then labeled for 1 minute using a mixture of PBSFP with 1:10 of 1mg/ml streptavidin PE and 1:20 of Anti-Myc FITC antibody in a volume of 20ul for 1 minute on ice. The cells are washed again with 1ml of PBSFP, then re-suspended in PBSFP to a density of about 5 million cells per ml. These cells are either sorted using a Sony SH800 Cell Sorter or BD Influx, or read on a BD Accuri C6 flow cytometer.

Designs are first typically tested with a high concentration of avid label before being evolved using lower concentrations and higher temperature conditions with a non-avid label. Library construction strategies such as error prone mutagenesis, combinatorial libraries based on Rosetta design suggested mutations, site saturation mutagenesis libraries, and guided combinatorial libraries based on Miseq data, using a Illumina Miseq desktop sequencer, for site saturation libraries are performed using yeast surface display and FACS sorting to increase the affinity and specificity of these initial binders. For specificity selection, an excess of the small molecules we aim to evolve specificity against is included in the incubation buffers. Kd approximation is done initially via yeast surface display as described by Wittrup *et al.* [24] Because of the slightly avid labelling conditions due to the streptavidin tetramer, and because the yeast surface can often be non-specifically sticky against the linker or PE fluorophore, more ac-

curate Kd measurement is needed. To obtain this, we purify the best variants to do either isothermal titration calorimetry or fluorescence polarization.

2.8.2 Protein Purification

The top designs from the library sorts are isolated and cloned into pET21(+) expression vectors. They are then transformed into E. Coli strain BL21, grown in TB at 37° (Fisher) until an OD600 of 1.2 is reached, and induced with 1ml 1uM IPTG. The protein is further purified using his-tag column purification and HPLC using a Superdex 200 column (GE Healthcare Life Sciences).

2.8.3 Equilibrium Fluorescence Polarization

Fluorescence polarization (FP) is a technique that relies on the fact that the degree of polarization of a fluorophore is inversely related to its molecular rotation. A measurement made in a solution of the difference in emission light intensity parallel and perpendicular to the excitation light plane, normalized by the total fluorescence emission intensity, can be related to the ability of the fluorophore to rotate. By using a ligand target conjugated fluorophore, we can get a measure of the difference in rotation due to the binding of the fluorophore conjugated target ligand to a target protein. [56] This difference can be used to generate a binding curve and approximate a Kd.

Fluorescence polarization is the preferred method for very hydrophobic ligands such as 25-hydroxycholecalciferol because of the unmodified ligand's low solubility in aqueous solutions. 25-Cholecalcifediol-TMR, a fluorophore conjugate, was provided to us by a collaborator from Kai Johnsson's lab at the École Polytechnique Fédérale de Lausanne (EPFL). To perform the assay, dilutions of purified protein in 40ul volumes are set up in flat bottom black polystyrene plates (Corning), and fluorophore-conjugate is added to a final concentration of 1uM. The mixture is allowed to mix on a plate shaker for 5 minutes before

measurements are made using a SpectraMax M5e(Molecular Devices) with an excitation of 540nm, emission of 580nm, and cut-off of 570nm. The average anisotropy measurements are analysed according to the method by Rossi *et al.* [30].

2.8.4 Isothermal Titration Calorimetry

Isothermal Titration Calorimetry (ITC) is a technique that can be used to quantitatively determine binding affinity and enthalpy changes. The technique directly measures the energy associated with a reaction when two chemical species are mixed together. In an experiment, one titrates one component into the other. This reaction is typically exothermic for a small molecule binding interactions with a protein. This energy can be related to the Gibbs free energy change and ultimately can be de-constructed into the entropic and enthanpic contributions via the standard thermodynamic expression delta G = -RT ln(Ka). [42] In our experiments, ligands which are soluble enough at the estimated Kd as measured by yeast surface display are used for Kd determination via ITC, as it is one of the most accurate methods.

2.8.5 Crystal Structure Determination

Protein is sent for crystal structure determination through a collaboration with Barry Stoddard's lab at the Fred Hutchinson research center.

Chapter 3: Computational Design of a 25-hydroxycholecalciferol Binding Protein with Low Nanomolar Affinity

3.1 Introduction

Cholecalciferol, more commonly known as Vitamin D3, is one of the most often prescribed diagnostics in medicine today. The standard method of measurement is done by taking a blood sample, sending it to a laboratory, and doing HPLC/mass spectroscopy to determine concentrations of the hormonally active form of vitamin D3, a hydroxylated version of vitamin D3 known as 25-hydroxycholecalciferol. This particular form is difficult to detect specifically because there often also exists cholecalciferol, which differs by one hydroxyl group, and ergocalciferol (vitamin D2), which differs by one methyl group from D3 and are also present in the blood sample. Antibody based measurement assays exist commercially, but often suffer from specificity issues due to these alternative forms of cholecalciferol and the difficulty in raising an antibody that recognizes all of the appropriate subtle differences. There currently exists a demand for a highly specific binder for 25-hydroxycholecalciferol that is able to distinguish between ergocalciferol, 25-hydroxyergocalciferol, and cholecalciferol, for incorporation into a commercial detection assay.

During our design efforts we created many designs that simply failed to show any binding signal on yeast surface display and flow cytometry. We created error prone libraries based on many of the more promising designs in an attempt to recover function and to hopefully learn why our designs were failing. This led us to a recovered design 2063.

Also in our early design attempts, we made a model named 4424, which was intended to bind the ligand tetrahydrocannabinol (THC). Testing of designs

against off target ligands was done in order to verify specificity for the designed ligand. In the case of 4424, the design did not bind the intended target of THC-BSA as shown using yeast surface display and flow cytometry, however, it did show a surprisingly strong signal for the target 25-hydroxycholecalciferol. Cross reactivity is generally expected, especially for chemically similar ligands such as these, but what was surprising was that after only one round of error prone evolution and one point mutation made from the 4424 initial design, we were able to isolate a variant with low micromolar affinity and approximately two orders of magnitude of specificity for 25-hydroxycholecalciferol over cholecalciferol, a difference of one hydroxyl group. Because of the excellent initial specificity of this design, we decided to follow up with further rounds of directed evolution and Rosetta aided design in order to create a binder with the appropriate Kd and specificity for application as a biosensor.

In addition to having this fairly good serendipitously discovered binder fall into our laps, we were also able to learn from this event, as well as failed designs and rescued binders, in order to improve our computational protocols. The design protocol improvements we worked out allowed us to create many additional, initially working binders with a success rate of approximately %25 for 25-hydroxycholecalciferol, as shown by yeast surface display and flow cytometry.

3.2 Methods

The failed, rescued, and serendipitous binders were all created using the previous protocol iteration that is slightly modified from the final protocol described in chapter 2. These differences are:

1) In the scaffold selection stage, the previous iteration screened PDBs based on the criteria mentioned in chapter 2 and used for Patchdock. In the final

iteration, scaffolds are screened multiple times and scaffold classes are identified in the first round so that sampling can be biased towards native scaffolds with more favorable chemical environments for the target ligand.

2) In the Patchdock step, the previous iteration used only the most energetically favorable ligand conformation is used during Patchdock into a native scaffold. In the final iteration, many more ligand conformations are used for the Patchdock step, only limited by computational resources. Native scaffolds are also expanded into many variants and each ligand conformation is docked into each scaffold variant to increase sampling.

3) The previous protocol iteration did not perform grid design whereas the final iteration is where the grid design protocol was first used.

4) In the previous iteration, design was performed with a minimal amino acid set that excludes charged residues, prolines, tryptophans, and glycines and allows for mutation of all amino acids excluding prolines or glycines. A procedure is used to design the binding pocket residues that greedily optimizes a metric based on a weighted combination of shape complementarity (10x), the Rosetta score packStat (5x), number and Rosetta energy of hydrogen bonds (5x), and Rosetta ligand interface energy by (2.5x). At most two iterations of Rosetta design and manual inspection were performed. In the final iteration, the design protocol no longer restricts the use of charged amino acids or aromatics and does not allow a change from a native aromatic residue to a non-aromatic. The greedy optimization protocol is removed in favor of lenient score cut-off that increases in stringency gradually over the iterations of grid design, manual inspection, and Rosetta design that gradually increase stringency of the score

cut-offs.

These improvements changed our success rate from near zero to approximately %25.

3.3 Results

3.3.1 Recovered Activity 25-Hydroxycholecalciferol Binders

Multiple designs targeting the ligand 25-hydroxycholecalciferol, in a scaffold with PDB ID 1WUB, were tested and found to be non-functional. Two of these designs are known as 2063 and 2064. Their sequences are located in appendix B. These two designs underwent error prone mutagenesis and selection in order to try and recover functional variants from these designs. This effort was was successful in finding two functional sequences with mutations: 1) R139P relative to design 2063 and 2) S48C, M126T, A131T, D141V, G153C relative to design 2064. A total of three rounds of directed evolution were done at ligand concentrations ranging from 21nM in the first round to 1nM in the final round. Only designs from base design 2063+R139P were enriched after the first round of directed evolution from a pool containing both error prone libraries.

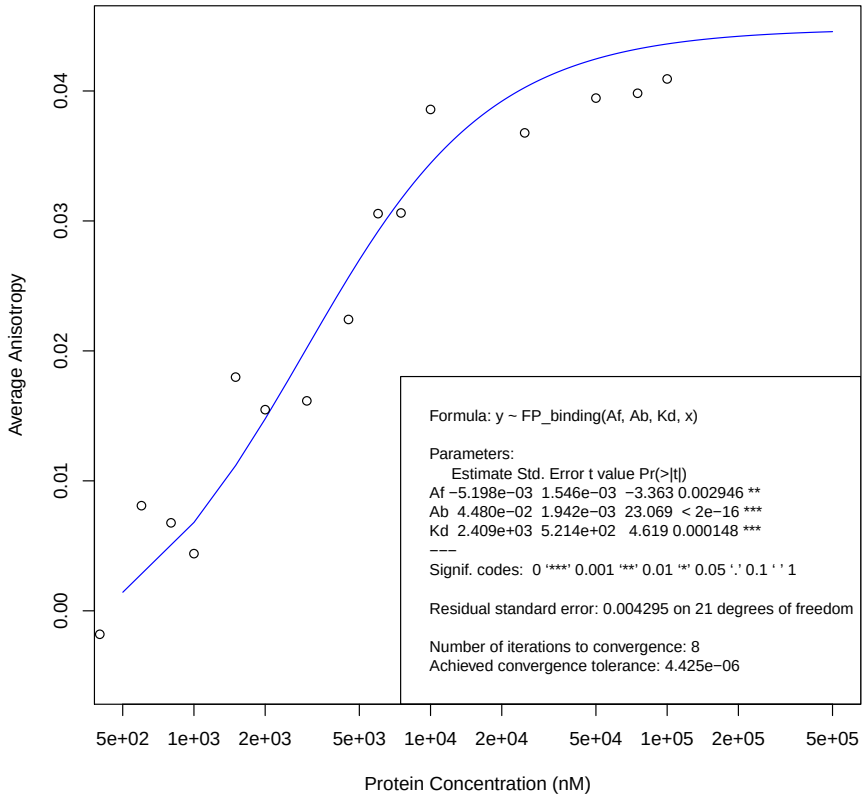

Figure 4: Equilibrium fluorescence anisotropy of 25-Cholecalcifediol-TMR mixed with purified 2063+R139P. The approximate Kd based on the non-linear fit is 2.4uM.

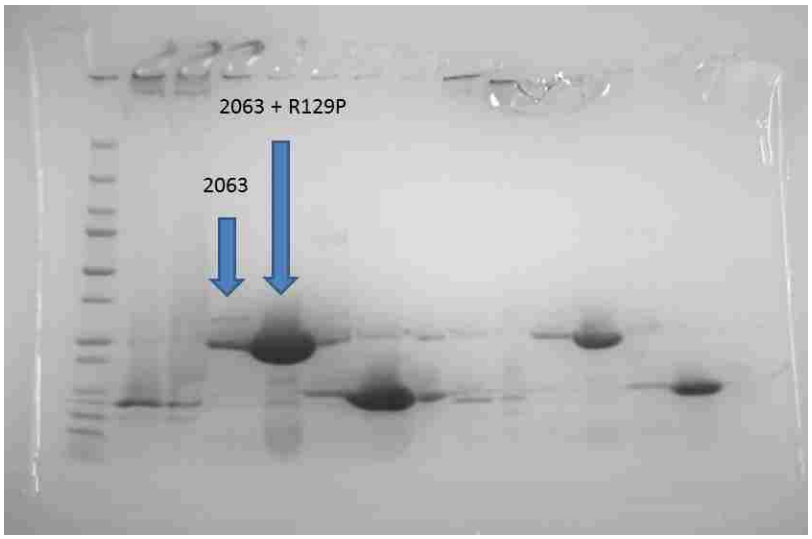

Figure 5: SDS PAGE gel comparing the levels of protein in the soluble fraction of the initial 2063 design and the 2063+R139P variant.

The second round of directed evolution was done at 21nM of vitamin D3. The mutations common among the final sort of error prone round 2 include: V24G, V170A, A101V. The third round of directed evolution was done based on the previous converged sequence from error prone round 2. The converged variant among the final sort of error prone round 3 include mutations: K15E, Y65H, S100N. During the third error prone screening, a selection was done to isolate designs that caused the evolved 1WUB variant to lose function. These mutations primarily occurred in the range of residues 50-57 and 141-145, both regions contain non-structured loop regions. A site saturation mutagenesis library based on variant 2063+R139P +V24G +V170A +A101V +K15E +Y65H +S100N was created and sorted at a concentration of 1nM for 2 rounds. The final variant from all this selection resulted in the following mutations relative to 2063: A4L, L11F, M18V, V23G, L27M, R34V, S45A, A47V, G63D, L79S, Q85L, I96V, F100V, A109V, G111V, M125T, M128V, V130A, M132A, R138P, L143M, V160M, F162M, A164L, A166I, L168A, V169A. The amino acid sequence for

this variant is located in appendix B and is named CM1-13 AD28.

3.3.2 Serendipitous 25-hydroxycholecalciferol Binder

A designed named 4424 in scaffold 3HX8, originally designed to bind the ligand THC, was also tested for binding against 25-Hydroxycholecalciferol and showed binding on yeast surface display. It's sequence is located in appendix B. This design did not bind the biotinylated form of the target ligand, THC, tested via yeast surface display and flow cytometry. Design 4424 was used as the basis for error prone directed evolution to bind 25-Hydroxycholecalciferol. After the first round, a glutamate was introduced in the binding pocket with mutation V106E. This mutation alone increased the specificity of 25-Hydroxycholecalciferol over cholecalciferol by approximately two orders of magnitude on yeast surface display and brought down the apparent Kd on yeast to the low micromolar range.

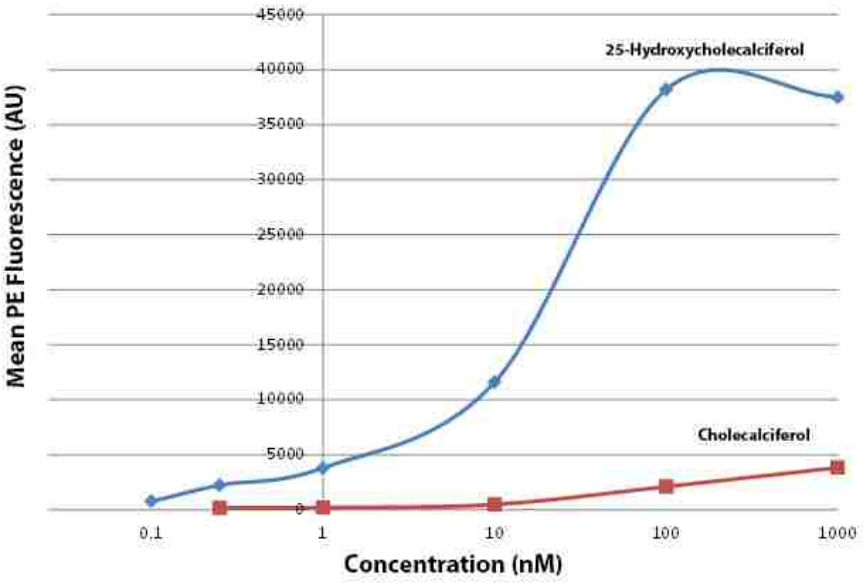

Figure 6: Titration data using yeast surface display and flow cytometry for design 4424 +V106E demonstrating the specificity between 25-Hydroxycholecalciferol and cholecalciferol.

Rosetta was used to create designs based on 4424+V106E that would further optimize the pocket for 25-Hydroxycholecalciferol. These Rosetta designs include the following variants:

4424 +V106E +T121A +S123A +V100M

4424 +V106E +T121A +S123 +V100I

4424 +V106E +T121V +S123 +V100I

4424 +V106E +V100I +S123A.

Two error prone libraries were created, one based upon the initial evolved variant 4424+V106E, and another based on the Rosetta improved variants. The best variant that converged from the two libraries was based on design 4424 +V106E +V100I +S123A, with final mutations 4424 +V106E +V100I +S123A +A36P +L66P +A80P.

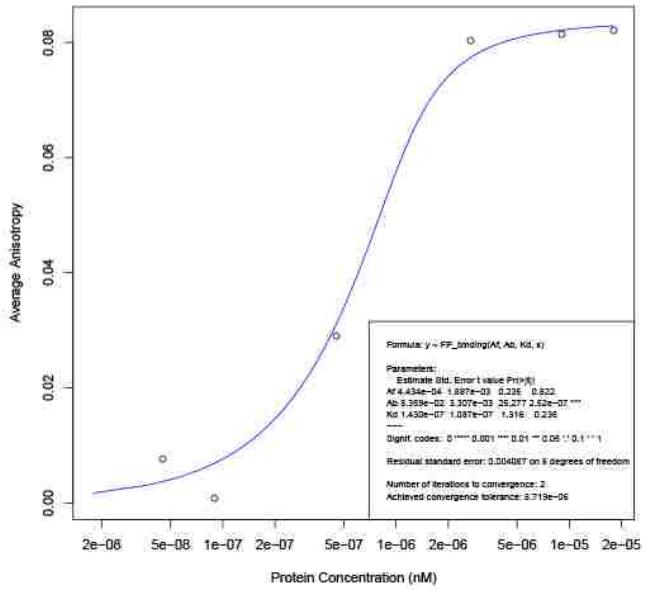

Figure 7: Equilibrium fluorescence anisotropy of 25-Hydroxycholecalciferol-TMR mixed with purified 4424 +V106E +V100I +S123A +A36P +L66P +A80P. The approximate Kd based on the non-linear fit is 140nM.

The next round of error prone yielded additional sequences with similar Kd

to 4424 +V106E +V100I +S123A +A36P +L66P +A80P. These mutants are named HH24, HH35.1, W19.1, and their sequences are located in appendix B.

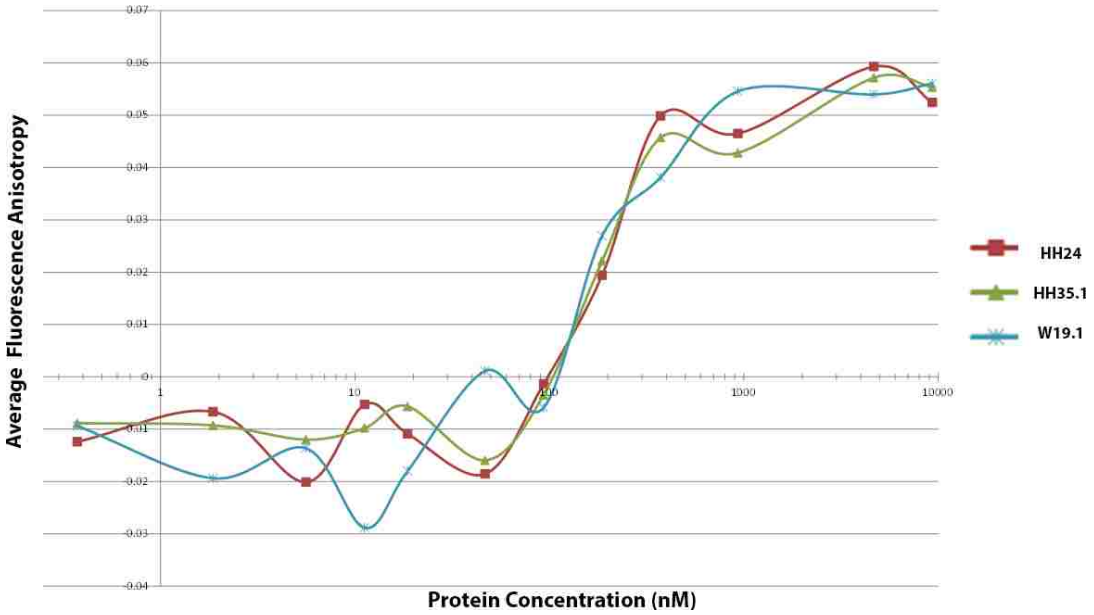

Figure 8: Equilibrium fluorescence anisotropy of 25-Hydroxycholecalciferol-TMR mixed with purified 4424 evolved variants HH24, HH35.1, W19.1 (Appendix B). The approximate Kd based on the non-linear fit is 110nM.

From these evolution variants, two crystal structures were solved of HH35.1 and W19.1, however they are very similar structurally and differ only by one or two surface residues. The structure for HH35.1 is shown below.

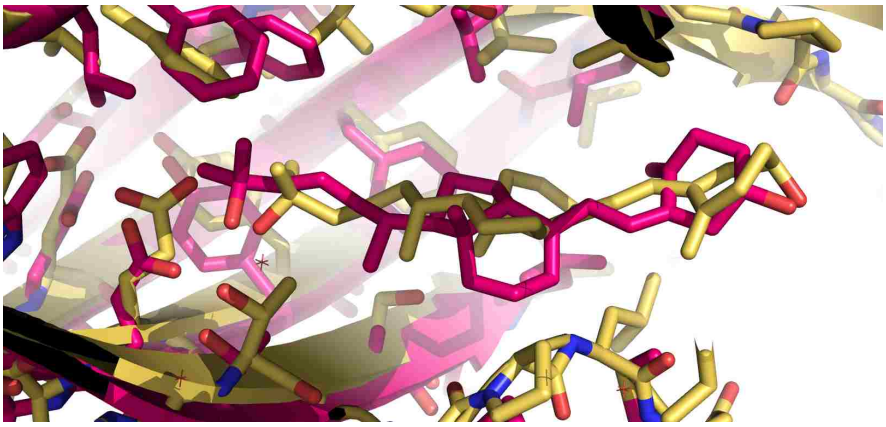

Figure 9: An overlay of the crystal structure for design HH35.1 with the best docked model. The RMS is calculated at 1.434Å. The pink model is the crystal structure and the tan model is the predicted dock design.

A site saturation mutagenesis library was created based on variant HH35.1. This library was sorted for 2 rounds at 1nM and sequenced using next generation sequencing technology. (Illumina Miseq) A heat map was generated to show enrichment of individual point mutations.

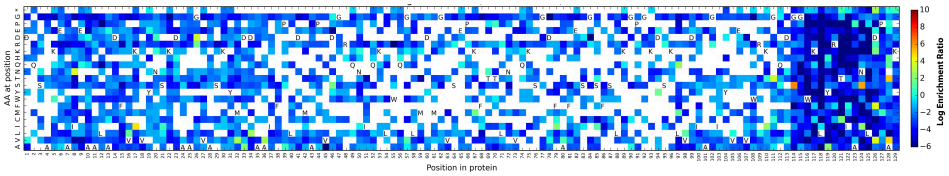

Figure 10: Heat map for the second sort at 1nM of a site saturation mutagenesis library where every position in design HH35.1 was mutated to every other position using NNK primers. The white colored positions represent substitutions where not enough data was gathered from the miseq run, possibly because the number of reads was insufficient or because the position was not present in the sample to a high enough degree.

The top give most enriched positions were combined to create a combinatorial library, which was sorted to convergence. The final sequence is named J1c-16, and contains the following mutations relative to the original 4424 design. G1D, A18V, R44P, L66P, L68F, D72N, E75K, D99G, V100I, K103I, V106E,

S123A, D126G. The final sequence is located in appendix B.

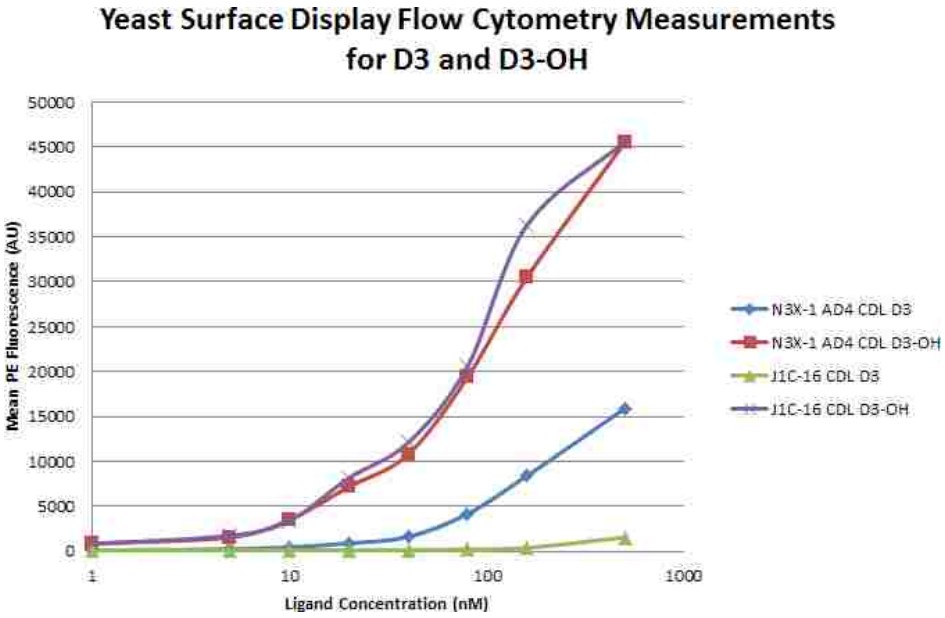

Figure 11: Yeast surface titrations were done on the J1c-16 and N3X-1 AD4. N3X-1 AD4 is another binder that is mentioned below. Both designs show specificity for 25-hydroxycholecalciferol over cholecalciferol, however J1C-16 is significantly more specific, likely due to its glutamate making the hydroxyl interaction, as opposed to a serine in N3X-1 AD4.

3.4 Successful 25-hydroxycholecalciferol Binders

Several designs were successfully created using the protocol outlined in chapter 2 to bind the ligand 25-Hydroxycholecalciferol, without modification, as shown by yeast surface expression and flow cytometry.

Table 1: Designations and sequence changes for designed and functional 25-Hydroxycholecalciferol binders. Full sequences are in Appendix B.

Design	Native PDB	Mutations from Native
6269	3ROB	W26F, L27V, L49Y, C65A, A76M, A78V, L96V, A115I, D132A A133Y, N134A
6326	1OHO	V19M, Y31F, N39L, Y56F, G59M, A67M, V87A, M89Y, D102A M115S, W119F
6218	1DMM	M12A, Y15M, V19L, V37L, D39A, M83A, M89T, D102S, M104S M115T, W119Y
6348	3NHX	Y13F, L17M, D37C, Y54F, S57A, L62V, V83A, N98A, M111S A113C
6264	3HX8	A37M, I63A, L66V, L68I, L88A, A90V, A100I
6344	3HX8	N12T, I63A, L66A, L68A, V106A, D121L
6220	1E3R	Y15L, Y31F, N39A, Y56F, V87I, D102S, M104I, I112V, W119F L124M
6234	1Z1S	S21A, L25V, W33F, L43M, W52Y, V75I, Y112F, Y126S, D128I
6258	3HX8	Y31F, W55F, A58S, I63A, L66V, L68M, L88A, A100W, L118M, L121S, W123F
6261	3HX8	Y31F, W55F, I63A, L66A, L88A
6333	2WC5	M6E, 9VM, I53L, I54L, M56A, S57A, M63L, D66N, V67N, R68A, H70N, Y77F,I78T, M91Q, H96A, E99A, D103A, E105V, D108A, R111A, V112A

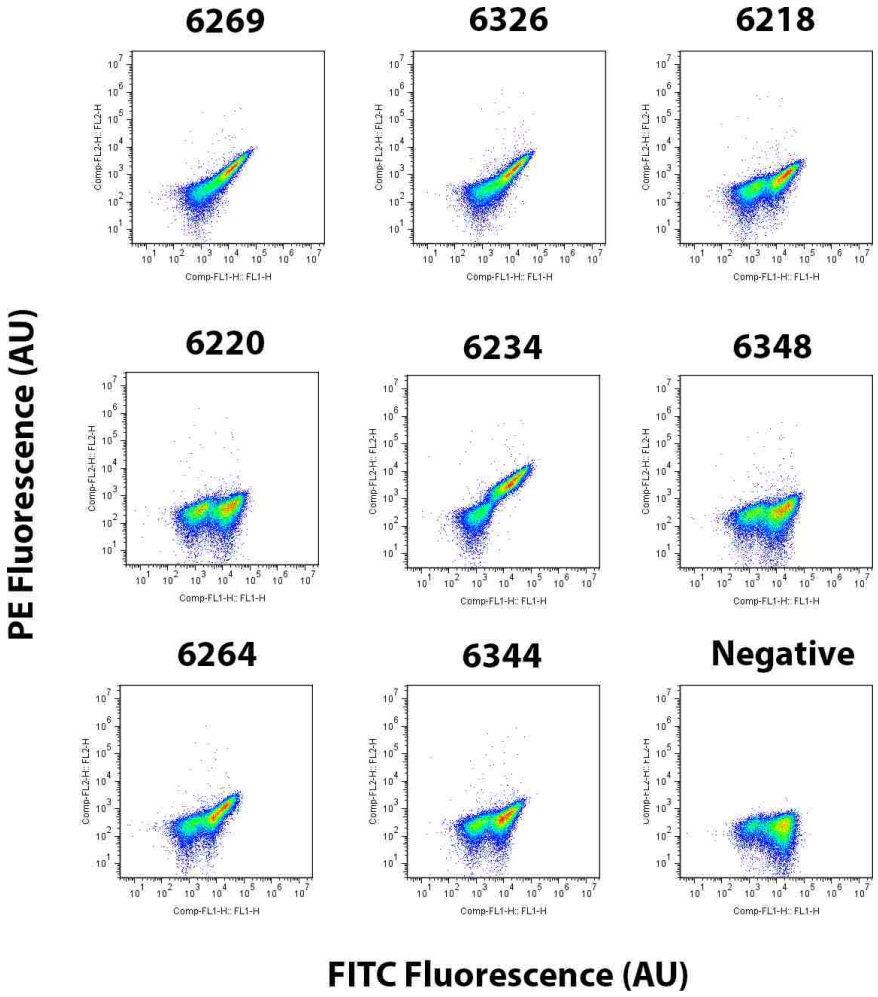

Figure 12: Yeast surface display and flow cytometry data for the designed, functional 25-Hydroxycholecalciferol binders. PE fluorescence represents binding to the target ligand and FITC fluorescence represents surface expression of the design on the yeast surface. The approximate concentration is 30uM of biotinylated 25-Hydroxycholecalciferol.

All of these initial binders were pooled and an error prone library screening was done to identify the tightest binder for further evolution and characterization. That selection converged to the design 6234 with additional mutations: S2N, S5R, P46S, H72P, G140V. This sequence is known as N3X-1 in appendix B. The yeast surface titration for this design is shown above in figure 11.

3.5 Discussion

3.5.1 Recovered Function Binder 2063

The design 2063 was recovered with an error prone library, yielding two functional sequences. 1) R139P relative to design 2063 and 2) S48C, M126T, A131T, D141V, G153C. Analysis of computationally generated scores can be insightful, but ultimately aren't reliable without crystal structure data. For this example, we found it interesting that the proline mutation found to recover activity is near the far side of the protein relative to the designed binding pocket. Proline residues are often associated with decreased flexibility and increased stability due to their restricted phi/psi angles. [64] When you consider the effect of this single proline mutation on the solubility or expression of E. Coli produced protein, the significantly increased levels of protein in the soluble fraction helps to support the idea that the protein is being stabilized and is regaining function as a result. (See figure 5) Additionally, the negative sort information collected during the second error prone library sort showed that the overwhelming majority of mutations that reduced binding existed in loops also on the "back side" of the scaffold, near this proline mutation. This also suggests that these loop regions are important to the stability of the protein, and that the protein itself may be marginally stable in its native form and requiring stabilization in order to gain function after our Rosetta design efforts add primarily functional mutations.

The final variant for this 2063 based design did not show as good of affinity or specificity for 25-hydroxycholecalciferol over cholecalciferol, and so it was not followed up on because the 3HX8 based serendipitous binder showed much better characteristics for eventual use in a biosensor application. Additionally, all the mutations that accumulated in design 2063 suggest that our initial design configuration is likely not what is actually taking place. Docking runs show that the ligand can enter from either end of the protein and mutations don't make a

clear distinction as to the orientation of the 25-hydroxycholecalciferol molecule. This may also suggest that the nature of this 1WUB scaffold makes it very amenable to binding long hydrophobic ligands, like it's native target, and that specificity in such situations is very difficult because there may be so many alternative, favorable conformations.

3.5.2 Serendipitous Binder 4424

The serendipitous binder 4424 was originally designed to bind the ligand tetrahydrocannabinol, but was tested against 25-hydroxycholecalciferol to test its cross reactivity. The two ligands are chemically similar because they both contain a long alkyl chain and are mostly hydrophobic. The fact that a designed protein can bind an off target ligand so readily without any positive design exemplifies to us that affinity is easier to obtain than specificity. The fact that we also gained very good specificity for the hydroxylated form of vitamin D3 with the V106E mutation is very fortunate, but that 3HX8 scaffold was included when designing our 25-hydroxycholecalciferol binders, so why weren't we able to come up with it? This is a very important piece of information we need to analyze and learn from.

The first lesson we learned is that the scaffold type may be playing a very important role, much more than we had previously realized. The native chemical environment in a designed scaffold may be determining the ability to bind certain ligand types and we should be taking advantage of this idea in order to gain an advantage. In previous design attempts with steroids [8], successful binders were easily obtained when using native scaffolds whose family type is known to bind other steroid molecules. In our example, the crystal structure of 3HX8 shows a tetraethylene glycol molecule binding in the pocket. This small molecule has a very similar chemical group compared to both vitamin d3 and THC and so should provide an advantage if used for designing binders to those molecules,

and also makes the cross reactivity of our design less surprising. This idea was known when creating our initial protocols, however, the extent may not have been fully appreciated due to our relatively even distribution of sampling spread across many different scaffolds. Diversity in protein folds was viewed by us as a good thing that was supposed to give us a higher chance of creating successful designs in case one scaffold type is generally unstable. That may not be a bad idea still, but scaffold bias for well expressing scaffolds with chemically similar ligand types is something that we have learned to take advantage of in our new protocols through the use of multiple Patchdock iterations as described in chapter 2.

So we learned that scaffold selection is much more important, but in this particular case, we did have the 3HX8 used in our design pipeline, so then why did we miss the functional design? After re-running our protocols, we found that we weren't sampling enough ligand conformers during the Patchdock step. The long alkyl chain of 25-hydroxycholecalciferol introduces many degrees of freedom into the ligand and we only choose to dock the most energetically favorable configuration because of limits on computational resources. The 25-hydroxycholecalciferol docked 4424 model found that a different conformer than the one we used during our Patchdock steps was the best fit. The 4422+V106E mutation should have been even easier to arrive at had we docked the ligand with the appropriate conformer. However, our previous protocol iteration excluded charged residues from design consideration due to the inability of Rosetta to properly weight the necessity for satisfying these charged residues, especially deep in a protein pocket. Additionally, when trying to allow Rosetta to design in the V106E mutation into design 4424, the proper rotamer was not found, underlining the fact that we still lack accurate modelling of electrostatic interactions, which is the initial reason we chose to omit charged residues. However, since

manual inspection is typically performed at some point during our protocols, it is not unreasonable to think that had we found the properly docked initial 4424 design, the V106E variant would have likely been created and ordered.

There is one final, important question that arises when analysing recovered, and serendipitous 25-hydroxycholecalciferol designs. What information can be learned and how can it be applied or generalized for creating initially functional binding designs? In light of all this information, we modified our protocol into the final iteration by including an additional Patchdock step that first casts a wide net on many diverse scaffold types at a lower resolution, before identifying those scaffolds with chemically favorable chemical environments and biasing the subsequent sampling steps towards them. The amount of sampling could also be increased since we would then be working with a smaller, more targeted set of scaffolds. We expanded each scaffold to include variants with different types of substitutions in order to provide additional starting points for the Patchdock search, and we also expanded the number of ligand conformers which are used in each Patchdock run up to the capacity of our computational resources.

Suggested by the 2063 recovered binder information, stability may be playing a role in the ability of our designs to function properly. To try an learn from this, we introduced additional restrictions during the scaffold variation generation steps and Rosetta design steps that would conserve aromatic, CYS, and PRO residues. These residues are often responsible for core packing and secondary structure, so by removing the potential instability resulting from changing such residues, we would be decreasing our changes of causing unwanted backbone movement in our final designs that may be resulting in non-functional designs. Additionally, this points to the need to explore orthogonal methods of stabilization, such as the design of disulfide bonds or stabilization through proline inclusion. A separate experiment was done in order to test these techniques and

is described in more detail in chapter 6.

By including these learned protocol modifications, we came up with our final protocol iteration which is the basis for chapter 2. This protocol allowed us to create create many
25-hydroxycholecalciferol designs that bound the intended ligand in two different scaffold families without any additional modification, providing evidence that our protocol improvements may have been partially solving systematic design problems.

In addition to specificity between the two hydroxyl variants of vitamin D3 and vitamin D2, the real test of specificity will be in the context of where the binder will be used in any potential application. This brings up a concern with non-specific binding, since in human blood or serum, there may be many large hydrophobic molecules that would be competing with vitamin D3 for the binding pocket. Until we are able to test all of the known, common cross reactive components, simply doing selection in human serum may be the best way to both guide any future evolution away from non-specific components, and to test the ultimate specificity of the vitamin D3 binders.

Chapter 4: Computational Design of Tetrahydrocannabinol and Cannabidiol Binding Proteins

4.1 Introduction

Tetrahydrocannabinol (THC) is the primary psychoactive compound found in marijuana and cannabidiol (CBD) is another non-psychoactive compound that is thought to play a role in mediating the effects of THC. THC is used for both recreation and medicine, whereas CBD seems to have primarily interesting medicinal properties. CBD has been referenced as a neuroprotective antioxidant [86], is known to affect the effects produced by THC [53], and has been studied for anti psychotic effects [88].

As more local governments consider legalization of marijuana use, there is a growing interest in being able to quantify the amount of THC and CBD in a sold product, or to conduct DUI type testing for cannabinoid compounds and metabolites that may affect one's ability to drive. The first step for any type of sensor is to have a recognition element, such as a protein, that is able to detect THC specifically from CBD and other metabolites. As mentioned previously, a useful sensor for THC would need to operate in the 3nM to 18uM range [27] for oral fluid detection. Using the protocols laid out in chapter 2, we have been able to successfully create binders in this range.

4.2 Methods

The design protocol used is exactly as it is presented in the chapter 2 outline.

4.3 Results

4.3.1 Recovered Activity THC Binders

Multiple designs targeting the ligand tetrahydrocannabinol were created in multiple scaffolds. Two designs, one referred to as 4423 in PDB ID 2Z77, and 4572 in PDB ID 3HX8 were non functional for THC binding, but after an error prone library selection, showed binding.

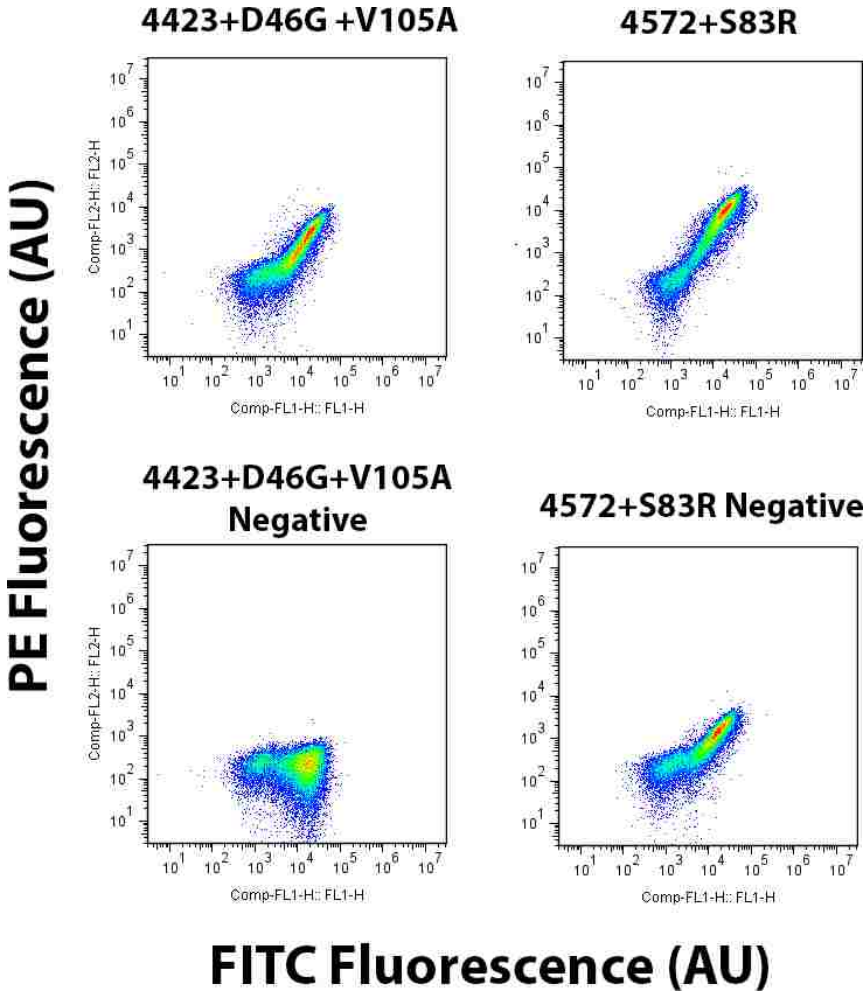

Figure 13: Flow cytometry results for recovered binders based on THC designs 4572 and 4423. The restorative mutation was S83R for design 4572, and D46G and V105A for design 4423. The negative control in both cases is a biotin-BSA label, as opposed to the THC-BSA-Biotin label in the positive tests. The 4423 based design shows off target binding against biotin in this assay, although the positive signal is approximately six times higher for the intended THC target.

4.3.2 Serendipitous THC Binders

Since THC and 25-hydroxycholecalciferol share similar structural and chemical characteristics, we also tested many 25-hydroxycholecalciferol binders against THC in order to discover additional starting points. This is a lesson we learned

from our serendipitous 25-hydroxycholecalciferol binder. The 25-hydroxycholecalciferol binder 6234 in scaffold 1Z1S appeared to bind THC fairly well, and so a site saturation mutagenesis library was created based on 6234 in order to see if we could evolve the initial design into specific 25-hydroxycholecalciferol, THC, or CBD binders. The best variant for THC binding that came from this selection is named NTH1 and has the following mutations from the original 6234 design: S2N, S5R, P46S, P50G, H72P, S126I, G140V.

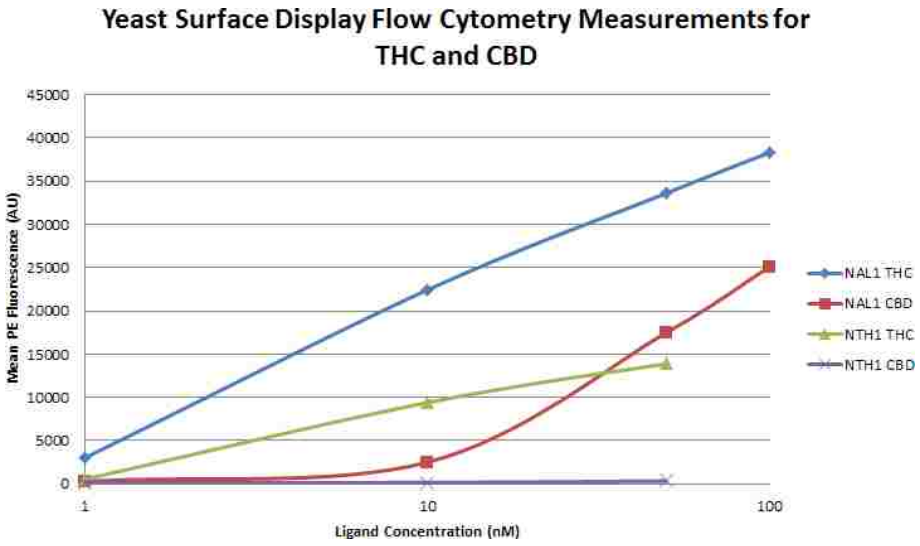

Figure 14: Yeast surface display and flow cytometry data for the designs NAL1 and NTH1. Both designs show a specificity preference for THC over CBD, although the design NTH1 doesn't bind CBD at all in the concentration ranges tested and thus has much higher specificity than the arguably tighter binder NAL1.

4.3.3 Successful THC Binders

Several designs were created using the protocol outlined in chapter 2 to bind the ligand tetrahydrocannabinol, and did so successfully, as shown by yeast surface expression and flow cytometry using a BSA conjugated THC label.

Table 2: Designations and sequence changes for designed and functional Tetrahydrocannabinol binders

Name	Native PDB	Mutations from Native
6285	2BVV	5YF, Q7L, N35A, V37A, F69Y, Y80F, R112M, S117A I118A, D119N, D120S, A165S, R172I
6290	2OVD	L33I, V36I, T53F, V66A, T68F, R70V, Y83F, L94N, R100Q H104I, V105A, L118V, L120T, L129Q, Y131F
6304	3AKR	S16L, Y17F, N44L, V46I, E86A, Y88F, R122A, I128M, Q136M Y171F, E177G
6314	4F6B	Q44V, K47Q, T48Q, T49F, D51A, Y52F, A62F, L65A, R68A H69S, R72L, V74I, I102V, I105V

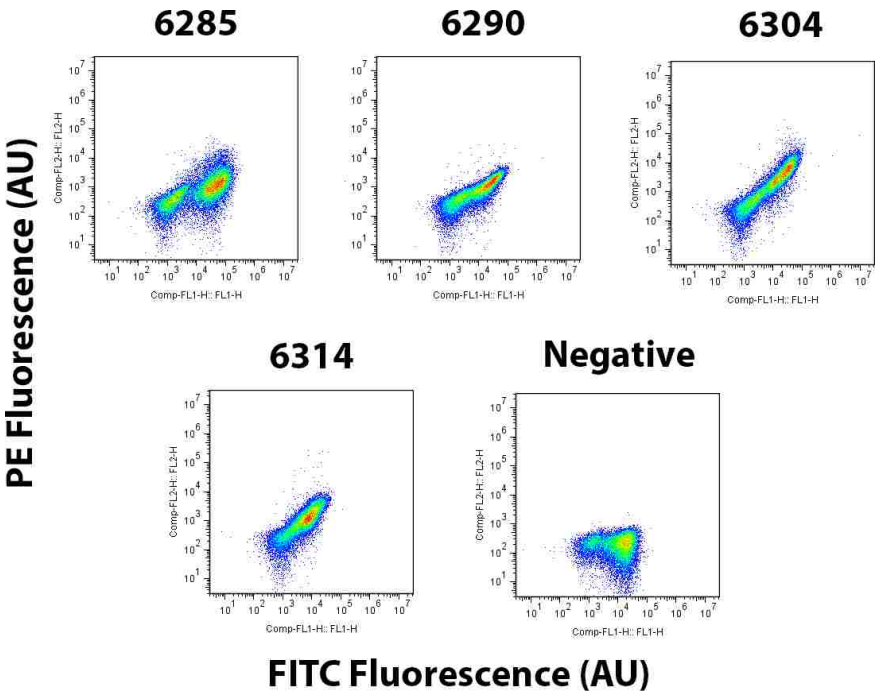

Figure 15: Yeast surface display and flow cytometry data for the designed, functional tetrahydrocannabinol binders labeled with a Biotin-THC-BSA conjugate. PE fluorescence represents binding to the target ligand and FITC fluorescence represents surface expression of the design on the yeast surface. The negative control is the design 6304 with no THC-BSA added during labeling.

All of these initial binding hits were pooled and an error prone library screen-

ing was done to identify the tightest binder for further evolution and characterization. This evolution was done using a non-avid biotin conjugated THC molecule. The selection converged to the design 6304 with additional mutations, N124A. A combinatorial library using design variant 6304 +N124A as the base was generated and resulted in a variant named NAL1 with the following mutations from the original design 6034: D20H, N69Y, I85F, M128I, M136I. This amino acid sequence is located in appendix B.

4.4 Discussion

4.4.1 Serendipitous THC Binders

The design NTH1 came from a design that initially bound 25-hydroxycholecalciferol. Without any prior knowledge, one would likely think that a protein designed for a particular target molecule would provide a better starting point than one found serendipitously. In the case of the 3HX8 D3-OH binder, this was not the case, as all successful designed binders for 25-hydroxycholecalciferol showed lower affinity and specificity so far. In this case, however, the best successfully designed THC binder in scaffold 6034 shows the higher affinity for THC, with slight preference for THC over CBD. The serendipitous THC binder NTH1 doesn't show as high of affinity for THC, but it's specificity over CBD is much better in the concentration ranges we measured. It seems that serendipitous binders have just as good of a chance to have excellent affinity as designed binders do. There may be a slight advantage in specificity seeing as all of the THC designs tested have a slight preference for THC over CBD. Knowing this, it may seem that we still do not know how to predict the evolve-ability for optimal affinity of an initial binder. We may be able to provide a slight specificity preference, but seeing now NTH1 shows better specificity without having been designed for it flies in the face of our computational design methods. I would

argue, however, that we simply have incomplete information on how to predict the evolve-ability of an initial binding protein, and so the more important factor in creating a useful binding protein is the number of starting points. With an approximately %17 success rate, we at least seem to do better than randomly screening proteins. (Unpublished data)

Chapter 5: Computational Design of a Biotin Binding Protein

5.1 Introduction

The first small molecule target we explored making binders for was biotin. The biotin molecule consists of a polar tetrahydroimidizalone ring fused to a relatively non-polar tetrahydrothiophene ring with an attached valeric acid substituent. Binding of biotin to the streptavidin tetramer is mediated by an extensive hydrogen bonding network, as well as Van Der Waals interactions. The interaction at the ureido oxygen is polarized enough that it gives the carbonyl group SP3 character, allowing it to form three hydrogen bonds. Additionally, when biotin forms a complex with streptavidin, two disordered loop regions become ordered. Mutational studies have demonstrated, and we have verified in our assays, that there is still detectable binding even when two out of the five head group hydrogen bonds are removed and an unfavorable polar interaction is introduced at the ureido oxygen. [65]

Biotin was chosen as our first target because of it's well known interaction with streptavidin. [69] The biotin streptavidin interaction is very well studied and high resolution crystal structures of native and mutant variants are readily available. Since we know exactly what interactions were needed to create a femtomolar interaction, we thought it might be relatively easy to essentially graft that binding site into another scaffold and get a binder. We thought wrong. It turns out this is extremely difficult because of the very unique, specific, and complex interactions biotin has with streptavidin.

Nonetheless, we learned about many limitations in our design protocol and were able to improve it significantly after attempting many, many years of design work on biotin. Ultimately, we were able to create a biotin binder, as shown on

yeast surface display and flow cytometry, but the binder was extremely weak and only showed a signal when using a highly avid label.

5.2 Methods

Designs to bind biotin were created very early on in our protocol development and what we learned from these attempts helped to inform improvements. Relative to the protocol described in chapter 2, the biotin design protocol varied in the following ways:

1) In the scaffold selection step, the biotin design protocol only used scaffolds found in the binding MOAD (mother of all databases) [23], which is a database containing high-quality examples of ligand-protein binding models. We then selected PDB files that are monomers, have a similar ligand size to the target ligand, have been expressed in *E. coli*, have binding pockets that consist primarily of α-helices and β-sheets, have ligands that are at least partially solvent accessible to accommodate the linker used in experimental validation, and do not contain structural metal ions. After filtering for these criteria, we then searched the protein database for structural homologs of those structures that passed the filters and added them to our set. This filtering process leaves us with several thousand starting scaffolds in the case of the biotin ligand. In the final protocol, a much less filtered set of scaffolds are screened multiple times and scaffold classes are identified in the first round so that sampling can be biased towards native scaffolds with more favorable chemical environments for the target ligand in subsequent rounds.

2) In the ligand placement step, the biotin design protocol used both Rosetta Match and Patchdock were the primary protocol used for generating initial

placements of the biotin ligand into the scaffold pocket. For more information on Rosetta Match, see appendix A. In the final protocol, Patchdock is used exclusively and initial scaffolds are expanded into many variants and each ligand conformation is docked into to increase sampling.

3) The biotin design protocol did not use grid design whereas the final protocol did.

4) In the Rosetta Design step, the biotin design protocol used a wide variety of filters and variations on their strictness and how they were applied. These additional filters are described more in depth in appendix A. The primary method of applying these filters was in a weighted greedy optimization step where all residues within approximately 6Åof the ligand is mutated to every other residue and scored using many weighted score terms. These score terms were derived from native binder examples in the Community Structure-Activity Resource (CSAR) and the Cambridge Crystallographic Data Centre (CCDC). [**?** 1] From these collections, we extracted crystal structures for a diverse set of protein-ligand complexes, many of which have measured Kd values. These examples of native binding proteins are hand curated by us to select structures with only one ligand in the binding pocket, structures in which the ligand binding site is not at the interface between two subunits or two proteins, and structures with no metal ions or water molecules that participate in the binding interaction. These criteria are used to avoid scoring complications and to make the results more directly meaningful to our design cases. All of these native scaffolds are subjected to a Rosetta minimization protocol to alleviate any clashes and are then scored by all available metrics. The size of this benchmark set is 49 and the scores are shown in table 3.

Metric	Average	Std Dev	Median
Interface Energy (IE) (REU)	-10.962	7.633	-9.537
Interface Area (IA) (Å^2)	528.027	251.06378	548.112
IE / IA (REU/Å^2)	-0.0208	0.843	
Fractional SASA	0.854	0.130	0.896
Rosetta Holes	0.690	0.046	0.679
Shape Complementarity	0.743	0.083	0.743
Ave. H-bond Energy (REU)	-0.651	0.606	0.607

Table 3: Score Metrics of Native Binding Proteins: Interface energy is a score defined by the sum of Rosetta energy terms over all pair-wise interactions between the ligand and protein. Because the interface energy scales with the size of the ligand, it must be normalized by the interface area for comparison between ligands. The fractional solvent accessible surface area (SASA) is a zero-to-one measure of how exposed the ligand is where zero is completely exposed and one is completely buried. Rosetta Holes is a zero-to-one measure of empty space in the non-solvent exposed portions of the protein, where zero is completely devoid of atoms and one is completely packed. Shape complementarity is a zero-to-one measure of how well the contours of the ligand match up geometrically with the scaffold along the protein-ligand interface, where zero is un-complementary and one is perfectly complementary. The hydrogen bond energy is determined by the distance and orientation between the acceptor and donor atoms. It typically ranges from zero to negative two.

Metric	Average	Std Dev	Median
Interface Energy (IE) (REU)	-9.173	1.854	-9.766
Interface Area (IA) (Å^2)	389.75	94.077	404.812
IE / IA (REU/Å^2)	-0.0235	0.314	
Fractional SASA	0.899	0.100	0.899
Rosetta Holes	0.643	0.058	0.066
Shape Complementarity	0.698	0.077	0.722
Ave. H-bond Energy (REU)	-1.120	0.599	-1.215

Table 4: Example of Score Metrics of Designed Proteins

Another difference between the biotin design protocol and the final protocol is that the biotin design protocols allowed for mutation from and to all residues. In the final protocol, residues are restricted to leave residues that may contribute to scaffold stability as native, such as glycines, prolines, aromatics, and cysteines, untouched during design. The final protocol also got rid of the majority of complicated score terms and relied more on manual inspection of designs to inform the *in silico* evolution steps.

5.3 Results

5.3.1 UM_37 Recovered Binder

The designs created using the first protocol iteration described previously did not yield any functional biotin binders. It did, however, result in one error prone recovered design named UM_37 in scaffold 1OHO and its evolved variant WBD1-V3. The sequence changes between the two are: V62W, V64I, C77R, Y82W, A94C. Sequences for these two variants are located in appendix B.

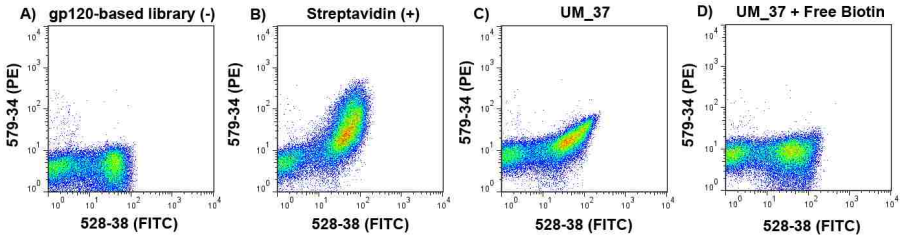

Figure 16: Flow cytometry data for controls and the successful biotin binding design UM_37 found via library selection. The Y-axis represents binding to our biotinylated fluorophore and the X-axis represents cell surface expression. All samples are incubated with 16.6μM Biotin-XX-PE in 40μL at 4 C for 3 hrs and washed once with 200μL PBSF immediately before reading. A) This negative control is an orthogonal gp120-based library available in the Baker lab (S2). B) Streptavidin positive control. C) Biotin binding design named UM_37 in scaffold 1OHO. D) Same design in (C) but with >100-fold molar excess of free biotin.

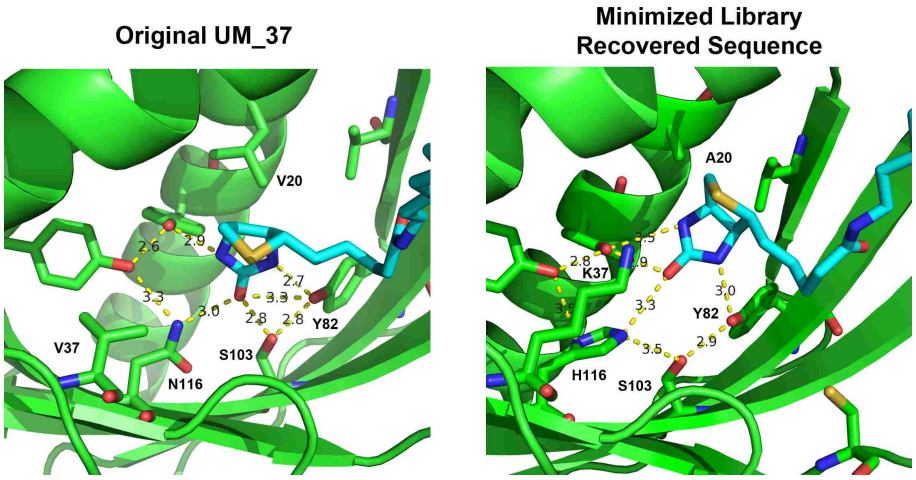

Figure 17: Structures of the original design model of UM_37 and the minimized model with mutations found from library screening. The sequence found from library screening shows binding activity, whereas the original model does not.

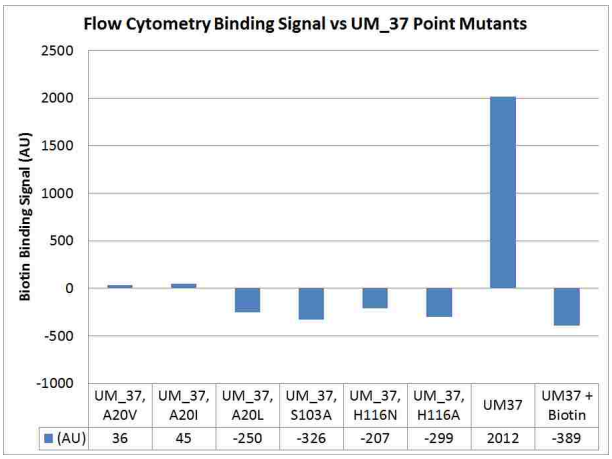

Figure 18: The Y-axis represents the binding of the a UM_37 mutant to our biotinylated fluorophore, phycoerythrin. It is calculated by measuring the mean PE signal in the expressing yeast population and subtracting the PE signal from the non-expressing population. All mutations that increase the size of position 20 or remove potential hydrogen bonding residues knock out binding activity.

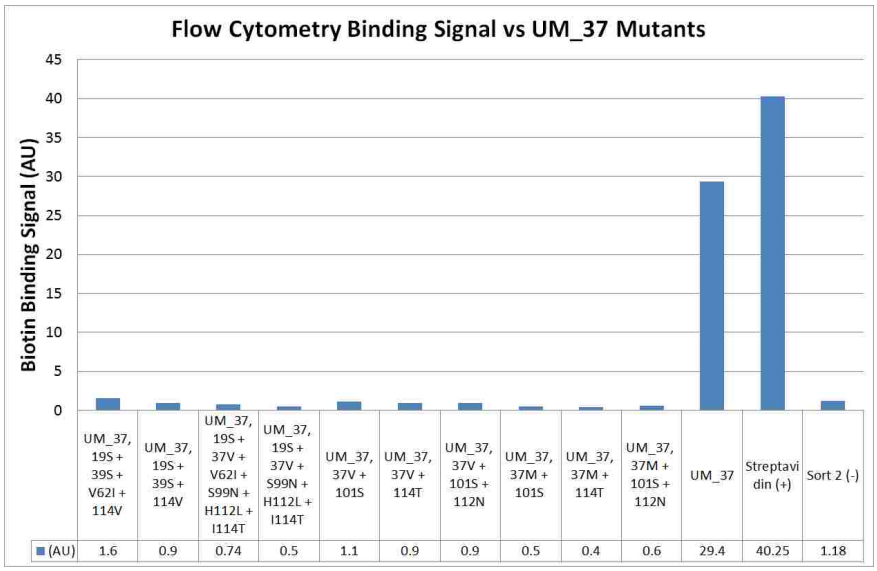

Figure 19: The Y-axis represents the binding of the a UM_37 mutant to our biotinylated fluorophore, phycoerythrin. It is calculated by measuring the mean PE signal in the expressing yeast population and subtracting the PE signal from the non-expressing population. Every mutation that involves changing lysine 37 to either a valine or methionine knocks out binding activity.

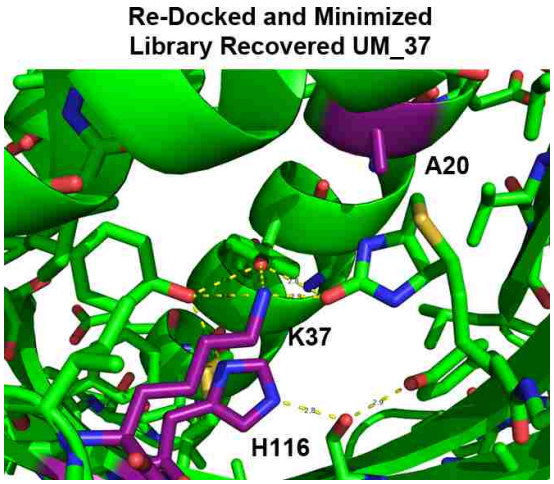

Figure 20: A model of re-docked UM_ that best matches the sequence of the binder found after library screening. The residues in purple show positions where the active site changes from the initial non-binding model to the successful binding design obtained through library screening.

To try and find additional mutations that improve binding affinity and to better explore the new docked model's sequence space, we created and screened an NNK library that will exhaustively sample all positions near the active site for each amino acid non-combinatorially. We identify two point mutations, A97C and Y82W, that show enhanced binding. A titration curve for this comparison is shown in figure 21.

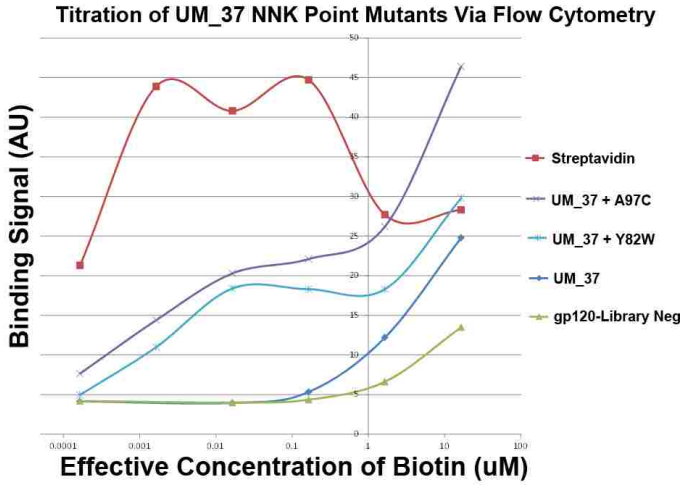

Figure 21: The Y-axis represents the binding of the a UM_37 mutant to our biotinylated fluorophore, phycoerythrin. It is calculated by measuring the mean PE signal in the expressing yeast population and subtracting the PE signal from the non-expressing population. Concentrations of our label were not sufficient to achieve saturation, however, this figure suggests that the A97C change results in the mutant with the highest relative affinity. Both mutants show increased binding over the original UM_37 design found in the initial library screen.

A combination mutant that contains both of these mutations is created and shows a binding signal via flow cytometry. Although it has not had its relative affinities measured compared to the single A97C and Y82W point mutants. (See figure 22.)

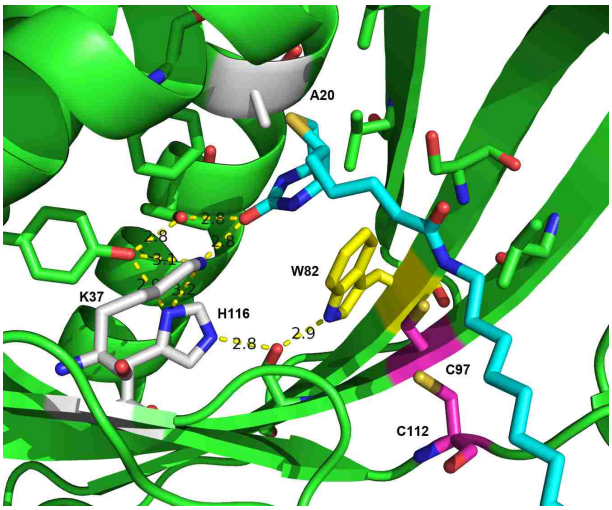

Figure 22: The image shows the structure of the UM_37 + A97C + Y82W mutant. The mutations from our original unsuccessful model to this optimized successful binder are V37K, N116H, Y82W, A97C, W112C, V20A.

5.3.2 2082 Successful Biotin Binder

A successful binder was made after protocol modifications and is named 2082. The evolved variant of 2082 is named P2 and contains the mutations S5G, M59V, and E67G. The amino acid sequence for both 2082 and P2 are located in appendix B.

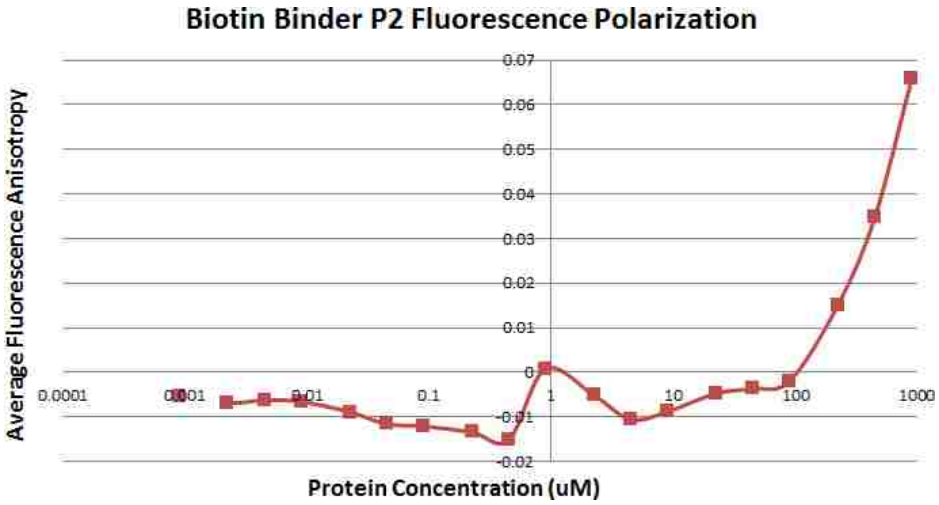

Figure 23: Equilibrium fluorescence anisotropy of Biotin-FITC mixed with purified P2. The Kd cannot be reasonably determined from this data because the binder does not saturate at 1mM concentration of Biotin-FITC, suggesting that the label is binding non-specifically to the protein, or the binder has a Kd of greater than 500uM.

In addition to these two binders, we were able to pull out many others from a large set of designs that underwent error prone mutagenesis. These following binders were not followed up on because the docked models looked and scored like non native binders, suggesting that all of these binders were likely not binding in any way similar to our predicted designs.

5.4 Discussion

5.4.1 UM_37 Recovered Binder

The knock-out mutants for recovered variant UM_37 suggest that the orientation of biotin in our original model is incorrect. Although the loss of activity by mutating serine 103 to alanine seems reasonable, the mutation at position 116 from an asparagine to a histidine requires an unfavourable histidine rotamer to maintain a direct hydrogen bond to biotin in our model.

The sequence of the successful UM_37 binder recovered from the library has

several mutations in the binding site compared to our original model, so we proceeded to verify if the ligand was still binding in the originally predicted orientation. We created single point mutations with the goal of testing the proximity of the ligand to position 20 and the importance of positions 116 and 103 to binding. The mutation of position 37 from a valine to a lysine seemed unfavourable according to our original model so we attempted to revert residue 37 back to a hydrophobic valine or methionine residue. These results are shown in figures 18 and 19.

Minimizing biotin in the library recovered sequence design results in a model with a void near position 20 because of the introduced V20A mutation. Point mutants shown in figure 18 were made to try and fill this void by substituting in a larger isoleucine, valine, or leucine residue. All of these mutants showed no binding activity, suggesting that these mutants are causing steric clashes with biotin, which is much closer to position 20 than our original model predicts.

Residue 37 changes from a non-interacting valine in our original model to a charged lysine residue in the library recovered binding sequence. This lysine residue seems to be unfavourable and not able to make a hydrogen bond to biotin, so we create many mutants in which residue 37 is changed to either the original valine or a methionine residue. All of these mutants show complete loss of binding and suggest that this lysine is actually making a critical interaction with biotin. These results are shown in figure 19.

Because of the inconsistency of our model with these mutational data, we computationally re-dock biotin in the new sequence with a much higher sampling rate and more extensive backbone minimization. This results in a model that is more consistent with the mutational data because the biotin ligand re-orients the ureido oxygen towards lysine 37, making a hydrogen bond, and the ligand moves closer to position 20. This re-docked model is shown in figure 20.

The point mutation A97C introduces a potential disulphide bond with C112. The addition of a disulphide bond is possibly stabilizing a flexible loop that interacts with the biotin linker and may be telling us that there are subtle backbone movements playing a role in binding. The A97C mutation alone gives the greatest increase in affinity among the NNK library mutants. The second best mutation is Y82W, which increases the packing against the ligand, and suggests that Rosetta may be under packing our initial designs.

It is interesting to note that a histidine residue is not predicted to be participating in the biotin binding interaction. We tested binding for pH dependence, as such a product may be commercially interesting, however, no pH dependent binding was seen, suggesting that even our current model may not be accurate. Due to the nebulous nature of this binder, we decided to move on to making new designs based on an improved protocol.

5.4.2 2082 Successful Biotin Binder

Many of the early biotin binding designs were created overwhelmingly through human intuition and non-repeatable design procedures. Emphasis was put on hydrogen bonding interactions and initial ligand placement was done only using Rosetta Match with native or idealized hydrogen bonding constraints. This led us to designs that were very polar in nature and often sacrificed good packing and shape complementarity interactions in order to create numerous, but unrealistic and relatively high energy hydrogen bonding networks. Learning from these mistakes, we created the biotin design protocol described above, which put a much stronger emphasis on hydrophobic interactions and shape complementarity by utilizing the Patchdock protocol for initial ligand placement and by requiring fewer hydrogen bonds be made to our ligand during early rounds design. In addition to these major changes, many minor changes were made, for example, the final biotin design protocol introduced automated scripts for

filtering out ligand placements that are too buried and result in mutation of structurally important core hydrophobic residues, and we have created filters that look for unrealistic threading of our ligand linker through secondary structure elements to solvent instead of through native cavities.

It wasn't until these changes were made to the protocol that we were able to achieve any binding signal for a biotin design. The major changes changes were:

1) We switched to using Patchdock as the primary ligand placement method.

2) Design was stripped down to only use non charged residues and aromatic residues were not allowed to mutate.

3) Manual inspection was used to make minor changes and to introduce hydrogen bonds into the designs. (ALA to SER, PHE to TYR, VAL to THR).

After applying these changes, the successful 2082 binder was made in scaffold 3FKA.

Chapter 6: Functional Recovery of Computationally Designed Small Molecule Binding Proteins via Proline Stabilization

6.1 Introduction

The majority of the small molecule binder designs are non-functional. This is a troublesome issue because often times it isn't obvious why a certain design will not work as intended. Sometimes the reason may be obvious, such as the decision to order both variants of a possibly questionable mutation during manual inspection. Other times, it may be a Rosetta favored mutation that passes through the filters because of an assumption made by the Rosetta software, such as assuming a fixed backbone, the use of discrete rotamers, implicit solvent, score terms based on empirical crystal structure data, or the simplified, pairwise electrostatics model. These assumptions may cause mutations and designs to appear favorable by Rosetta score when in reality they should not be. Detecting these kinds of problems would require very meticulous manual inspection and are not always found before ordering designs.

Several other reasons for designs failing to function may include inherently unstable native scaffolds, destabilized scaffolds or unpredicted backbone movement due to inserting designed mutations, or our expression system may not be optimized or have the correct chaperones to handle a particular protein fold. Because of limited resources, the difficulty in solving crystal structures of non-functional designs of potentially low significance and because it is very difficult to extract any predictive structural information based on yeast binding assay data alone, it must be pointed out that much of the analysis to follow is highly speculative, but nonetheless interesting.

That being said, one theory as to why some of our designs are non-functional

is based on the fact that most proteins in nature are marginally stable. [44] In other words, it seems that functional mutations often come at the cost of overall protein stability. If a protein is not globally stable enough, it may become disordered, unfolded, or badly behaved. Our design methods to date have only dealt with functional mutations, often only modifying amino acids in the proposed binding pocket. We hypothesized that by doing this, we may be destabilizing our native scaffold starting points so much that they become distorted and no longer capable of folding into the predicted functional design. In order to test this hypothesis, we took all designs from our latest design round that did not show a significant signal using yeast surface display and flow cytometry and attempted to stabilize them to regain or improve function. If stability is indeed an issue in a subset of our designs, then by including prolines at these positions, we would either restore function or significantly improve the ability of those designs to bind the target ligand.

6.2 Methods

The design protocol used is exactly as it is presented in the chapter 2 outline but then applied a round of proline stabilization design. To do this, we considered combinations of proline substitutions at all positions that fall into the "proline rule".[64] These are the i + 1 position of type I and II beta-turns and position i in type II beta-turns. It was shown that the addition of prolines at these positions in many examples improved stability of the protein, so we computationally modelled all of these proline substitutions combinatorially. Combinations that showed a decrease in global Rosetta energy after proline substitution, minimization, and backbone relaxation, were ordered and tested for binding improvement or recovered function.

6.3 Results

The scores associated with all computationally modelled proline substitutions are plotted in the following histogram.

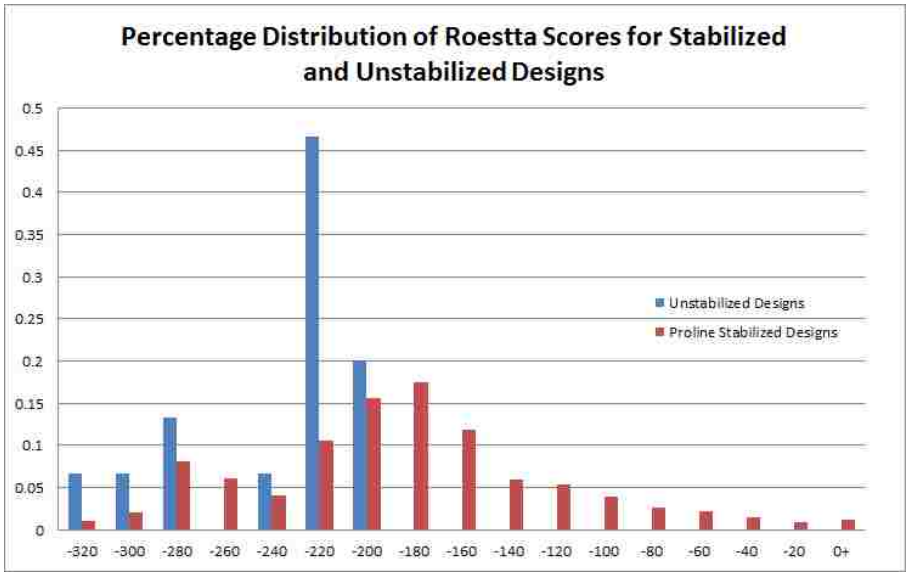

Figure 24: A histogram showing the distribution of percentages of Rosetta scores for all un-stabilized and stabilized 25-Hydroxycholecalciferol binders that failed to show a significant signal initially on flow cytometry. There are 15 original un-stabilized designs and 792 stabilized variants.

Of these proline variants, three initial designs were found that had a significantly improved binding signal due to one or more proline substitutions.

These designs that showed improved function with proline substitutions are named CS1, CS3, and CS4. CS1 and CS3 are in scaffold 3HX8 and CS4 is in scaffold 2WC5. CS1 is based on initial design 6258. CS3 is based on initial design 6261, and CS4 is based on initial design 6333. The sequences for all of these variants are located in appendix B. The proline substitution positions for CS1 are K93P and Q112P. The substitution positions for CS3 are also K93P and Q112P, although this design has a different initial design for the same scaffold. The substitution position for CS4 is H70P only.

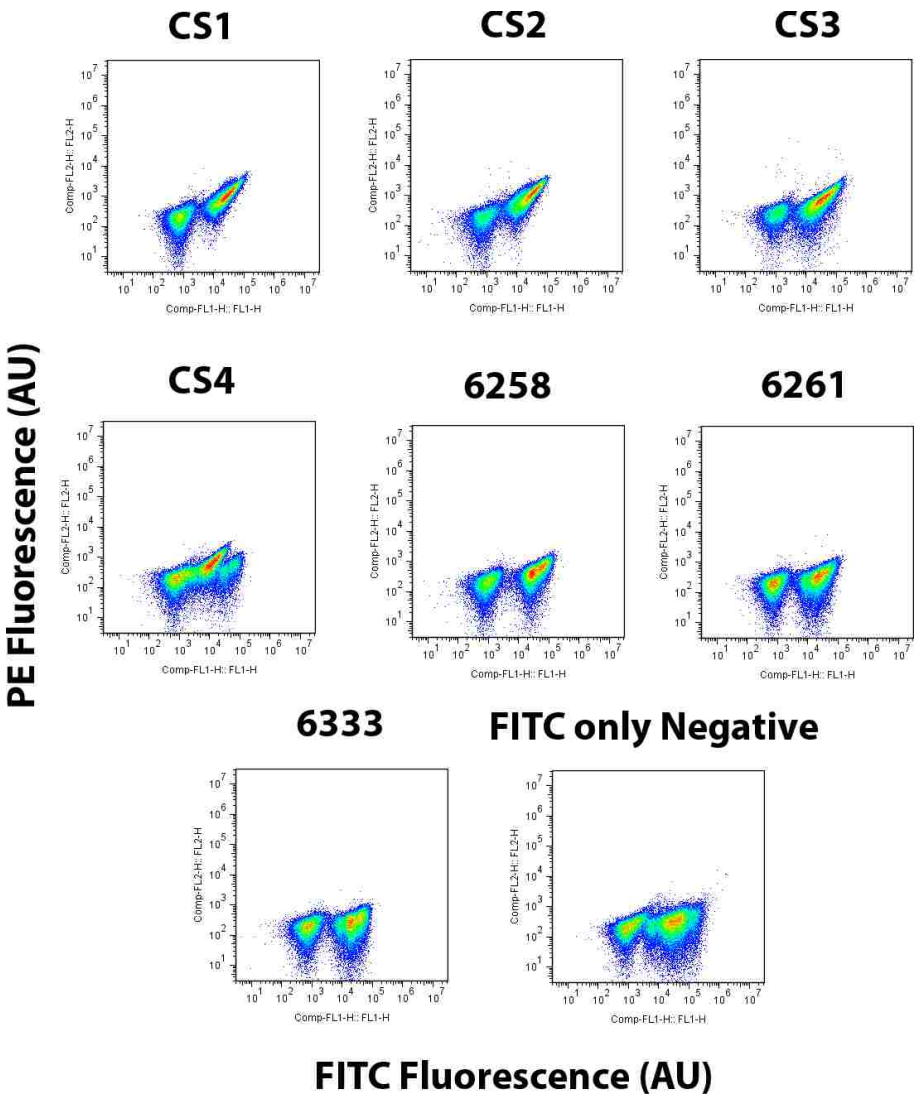

Figure 25: Yeast surface display and flow cytometry data for proline stabilized variants and their initial designs. CS1 is a proline variant of 6258. CS3 is a proline stabilized variant of design 6261, and CS4 is a proline stabilized variant of design 6333.

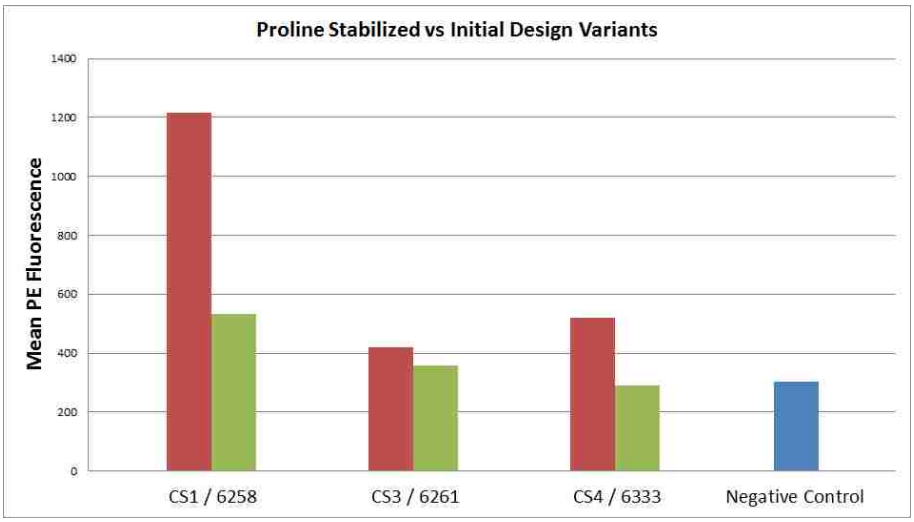

Figure 26: Mean PE fluorescence as shown by yeast surface display and flow cytometry for proline stabilized variants and their initial designs. CS1 is a proline variant of 6258. CS3 is a proline stabilized variant of design 6261, and CS4 is a proline stabilized variant of design 6333. All three of these variants increased the mean PE fluorescence signal over their non-proline stabilized variants. In the case of CS4 and 6333, the design went from undetectable binding to detectable.

6.4 Discussion

We have seen several cases during our directed evolution that mutations in the binding site are often either accompanied, or followed up in the next evolution round, by mutations located well outside of the binding site. An example we encountered was with design 2063+R139P. The native design 2063 alone did not show any binding activity for 25-D3OH, but after the introduction of one proline mutation on the opposite side of the designed binding site, binding activity was restored. Another design from the same error prone round and scaffold that had restored function was 2064 +S48C +M126T +A131T +D141V +G153C. Several of these mutations occurred in the designed binding site for 25-D3OH, suggesting that the ligand is located near the designed binding site and not near the R139P position. The R139P mutation is likely contributing to the restoration of binding function either through a long range backbone movement

that modifies the designed active site, or through some contribution to global stabilization that restored the protein from an unstable, non-functional state.

We saw another example with the serendipitous binder for 25-D3OH model 4424. After the first round of directed evolution, Rosetta was used in an attempt to rationally design mutations into the active site that would contribute to better binding function, without any regard to global stability. All of these design attempts resulted in designs that showed a lower signal on yeast than the initial starting point, 4424+V106E. This initial starting point had, in the previous evolution round, introduced a charged acid at position 106, which is in the protein core. As charged residues are not usually favorable in such a deep position, and since we know from crystal structure data that this acid does in fact interact with 25-D3OH and is critical for contributing both an increased affinity and specificity towards the ligand, it is most likely a primarily functional mutation. In the context of the marginally stable protein theory, it seems that the Rosetta improved designs based on 4424+V106E were pushing the protein towards instability, which may have been reflected in reduced activity when tested on yeast. In the round of error prone selection based on these potentially destabilized designs, as well as the initial 4424+V106E variant, the protein with the highest activity found was based on one of the Rosetta modified designs, and contained a three proline mutations: A36P + L66P + A80P. These positions didn't agree with the positions for increasing protein stability predicted by Fu *et al.* [64], however prolines in general tend to restrict the number of degrees of freedom due to its having relatively restricted dihedral angles, which may be contributing to increasing the global stability of the protein. It is interesting to note that the best sequences were all based on Rosetta designs that showed a lower initial signal than other starting points in the library. The library screening is not a complete sampling, but the results support the idea that Rosetta

may have potentially introduced functional mutations accurately, but required stabilizing mutations to accompany them in order to make their modelled contribution to binding. These potentially stabilizing mutations may be a critical complement to the functional Rosetta predicted mutations in many of our designs. These results, although with relatively low statistical significance, led us to the proline stabilization experiment mentioned in the introduction to this chapter.

When we scored proline variants of existing non-functional designs, we noticed that the majority of them seemed to make the Rosetta energy worse. Figure 24 shows this, suggesting that the types of proline mutations that we are able to score well are only the subset of substitutions that only require minimal backbone movement in order to improve the Rosetta score. This might seem restricting, but if larger changes to the backbone are expected as a result of substitutions, the scaffolds may be more stable, but may also be unable to maintain the designed overall structure in the original model, so this restriction may actually be considered a feature in our computational screening method.

Chapter 7: Acknowledgments

Here comes 6 years of gratitude compacted into one paragraph. I'd like to say thanks to Per Greisen for general "encouragement" and for fixing all my bugs over the years. I basically wasn't getting anywhere with design until Per came along and helped me turn ideas into code, and code into working designs. Jiayi Dou has provided me with great encouragement and has also been a great sounding board for all my woes, science related or otherwise. Christy Tinberg provided me with lots of experience and tons of help with all my experimental debugging needs. Matt Bick has been a great help in keeping me grounded and maintaining perspective on science and what comes afterwards. He also doesn't judge me when I go on an expletives rampage late at night in the lab, in fact, he encourages me. Thanks to David La for telling me late night stories about ebola and interface design. He's also been a fun person to bounce far fetched ideas off of. I have to thank Yifan Song because, well, he's Yifan and he bought me sushi. He's also awesome. I'd like to thank Lance Stewart for all his motivating words, help with brainstorming applications for this work, and for help with the THC binding project. He's been great with trying to find funding and getting expensive biotin conjugates for THC. Thanks to Alberto Sencha and Kai Johnsson for providing us with small molecule targets and their conjugates for our assays. A big thank you to Lindsey Doyle and Barry Stoddard for helping me get some awesome crystal structures of my binders. Thanks to Nephi Stella for the materials to get the THC binding project off the ground. I want to thank *almost* everyone in the lab for providing such a nice work environment. (You know who you are!) Thanks to all my friends and family who kept me sane and helped me realize that there may be slightly more to life than working in a lab. I'd like to thank the faculty who helped me over the years with my departmental exams: Wendy Thomas, Pat Stayton, Wenqing

Xu, Georg Seelig, and Herbert Sauro. And finally, I'd like to thank David, the bringer of funds, the PI of pain, the protein prodigy, the man who makes Bill Nye the Science Guy look like a toddler with a learning disability, Baker, for spoiling my graduate school experience with such an awesome lab space, my own desk and bench, and practically unlimited funding.

References

[1] The cambridge crystallographic data centre, 2012.

[2] Finn RD Mistry J Tate J Coggill P Heger A Pollington JE Gavin OL Gunasekaran P Ceric G Forslund K Holm L Sonnhammer EL Eddy SR Bateman A. The pfam protein families database. *Nucleic Acids Res*, 38:D211–22, 2010.

[3] Ellen R. Goldman Mehran P. Pazirandeh J. Matthew Mauro Keeley D. King Julie C. Frey George P. Anderson. Phage-displayed peptides as biosensor reagents. *Journal of Molecular Recognition*, 13:382–387, 2000.

[4] Sabine Kaltofen Chenge Li Po-Ssu Huang Louise C. Serpell Andreas Barth Ingemar André. Computational de novo design of a self-assembling peptide with predefined structure. *Journal of Molecular Biology*, 427(2):550–562, 2015.

[5] Ng AHC Uddayasankar U Wheeler AR. *Anal Bioanal Chem*, 397:991–1007, 2010.

[6] Brian Kuhlman Gautam Dantas Gregory C. Ireton Gabriele Varani Barry L. Stoddard David Baker. Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302(5649):1364–1368, 2003.

[7] Carol A. Rohl Charlie E.M. Strauss Kira M.S. Misura David Baker. Protein structure prediction using rosetta. *Methods in Enzymology*, 383:66–93, 2004.

[8] Christine E. Tinberg Sagar D. Khare Jiayi Dou Lindsey Doyle Jorgen W. Nelson Alberto Schena Wojciech Jankowski Charalampos G. Kalodimos Kai Johnsson Barry L. Stoddard David Baker. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*, 501:212–216, 2013.

[9] D Baker. Prediction and design of macromolecular structures and interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1467):459, 2006.

[10] Daniela Röthlisberger Olga Khersonsky Andrew M. Wollacott Lin Jiang Jason DeChancie Jamie Betker Jasmine L. Gallaher Eric A. Althoff Alexandre Zanghellini Orly Dym Shira Albeck Kendall N. Houk Dan S. Tawfik David Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453:190–195, 2008.

[11] Justin Ashworth James J. Havranek Carlos M. Duarte Django Sussman Raymond J. Monnat Jr Barry L. Stoddard David Baker. Computational redesign of endonuclease dna binding and cleavage specificity. *Nature*, 441:656–659, 2006.

[12] Kim T. Simons Rich Bonneau Ingo Ruczinski David Baker. Ab initio protein structure prediction of casp iii targets using rosetta. *Proteins: Structure, Function, and Bioinformatics*, 37(3):171–176, 1999.

[13] Lin Jiang Eric A. Althoff Fernando R. Clemente Lindsey Doyle Daniela Röthlisberger Alexandre Zanghellini Jasmine L. Gallaher Jamie L. Betker Fujie Tanaka Carlos F. Barbas III Donald Hilvert Kendall N. Houk Barry L. Stoddard David Baker. De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387–1391, 2008.

[14] Ora Schueler-Furman Chu Wang Phil Bradley Kira Misura David Baker. Progress in modeling of protein structures and interactions. *Science*, 310(5748):638, 2005.

[15] Paul M. Murphy Jill M. Bolduc Jasmine L. Gallahere Barry L. Stoddard David Baker. Alteration of enzyme specificity by computational loop remodeling and design. *PNAS*, 106(23):9215–9220, 2009.

[16] Sarel J. Fleishman Timothy A. Whitehead Damian C. Ekiert Cyrille Dreyfus Jacob E. Corn Eva-Maria Strauch Ian A. Wilson David Baker. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, 332(6031):816–821, 2011.

[17] Summer B. Thyme Jordan Jarjour Ryo Takeuchi James J. Havranek Justin Ashworth Andrew M. Scharenberg Barry L. Stoddard David Baker. Exploitation of binding energy for catalysis and design. *Nature*, 461:1300–1304, 2009.

[18] Timothy A Whitehead Aaron Chevalier Yifan Song Cyrille Dreyfus Sarel J Fleishman Cecilia De Mattos Chris A Myers Hetunandan Kamisetty Patrick Blair Ian A Wilson David Baker. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature Biotechnology*, 30:543–548, 2012.

[19] J. M. Belk J. Hsieh C. M. Benatuil, L. Perez. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng Des Sel*, 23(4):155–9, 2010.

[20] Rainer Jaenickea Gerald Böhm. The stability of proteins in extreme environments. *Current Opinion in Structural Biology*, 8(6):738–748, 2002.

[21] Carl A.K Borrebaeck. Antibodies in diagnostics – from immunoassays to protein chips. *Immunology Today*, 21(8):379–382, 2000.

[22] R. Welz R. R. Breaker. Ligand binding and gene control characteristics of tandem riboswitches in bacillus anthracis. *RNA*, 13(4):573–582, 2007.

[23] L Hu ML Benson RD Smith MG Lerner HA Carlson. Binding moad (mother of all databases). *Proteins*, 60:333–40, 2005.

[24] W. L. Hackel B. J. Sazinsky S. L. Lippow S. M. Wittrup K. D. Chao, G. Lau. Isolating and engineering human antibodies using yeast surface display. *Nat Protoc*, 1(2):755–68, 2007.

[25] D. Lowe K. Dudgeon R. Rouet P. Schofield† L. Jermutus D. Christ. Aggregation, stability, and formulation of human antibody therapeutics. *Advances in Protein Chemistry and Structural Biology*, 84:41–61, 2011.

[26] R. L. Dunbrack Jr F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–1681, 1997.

[27] Marilyn A. Huestis1 Edward J. Cone. Relationship of delta 9-tetrahydrocannabinol concentrations in oral fluid and plasma after controlled administration of smoked cannabis. *Journal of Analytical Toxicology*, 28(6):394–399, 2004.

[28] ML Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 19:709–713, 1983.

[29] F. Treuille A. Barbero J. Lee J. Beenen M. Leaver-Fay A. Baker D. Popovic Z. Players F. Cooper, S. Khatib. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–60, 2010.

[30] Rossi AM1 Taylor CW. Analysis of protein-ligand interactions by fluorescence polarization. *Nature Protocols*, 3:365–387, 2011.

[31] Chandler D. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–7, 2005.

[32] Kortemme T Morozov AV Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J Mol Biol.*, 326(4):1239–59, 2003.

[33] Zanghellini A Jiang L Wollacott AM Cheng G Meiler J Althoff EA Rothlisberger D Baker D. New algorithms and an in silico benchmark for computational enzyme design. *Protein Science*, 15(12):2785–2794, 2006.

[34] III T.A. Darden R.E. Duke T.J. Giese H. Gohlke A.W. Goetz N. Homeyer S. Izadi P. Janowski J. Kaus A. Kovalenko T.S. Lee S. LeGrand P. Li T. Luchko R. Luo B. Madej K.M. Merz G. Monard P. Needham H. Nguyen H.T. Nguyen I. Omelyan A. Onufriev D.R. Roe A. Roitberg R. Salomon-Ferrer C.L. Simmerling W. Smith J. Swails R.C. Walker J. Wang R.M. Wolf X. Wu D.M. York P.A. Kollman D.A. Case J.T. Berryman R.M. Betz D.S. Cerutti T.E. Cheatham. Amber. *University of California, San Francisco*, 2015.

[35] Maryam Hamzeh-Mivehroud1 Ali Akbar Alizadeh Michael B. Morris W. Bret Church Siavoush Dastmalchi. Modeling structurally variable regions in homologous proteins with rosetta. *Drug Discovery Today*, 18(23/24):1144–1157, 2004.

[36] I. W. Baker D. Davis. Rosettaligand docking with full ligand and receptor flexibility. *J Mol Biol*, 385(2):381–92, 2008.

[37] I. W. Raha K. Head M. S. Baker D. Davis. Blind docking of pharmaceutically relevant compounds using rosettaligand. *Protein Sci*, 18(9):1998–2002, 2009.

[38] Taylor SL Nordlee JA Niemann LM Lambrecht DM. *Anal Bioanal Chem*, 395:83–92, 2009.

[39] M. Jo J. Y. Ahn J. Lee et al. Development of singlestranded dna aptamers for specific bisphenol a detection. *Oligonucleotides*, 21(2):85–91, 2011.

[40] Meyer EA Castellano RK Diederich F. Interactions with aromatic rings in chemical and biological recognition. *Angew Chem Int Ed Engl.*, 42(11):1210–50, 2003.

[41] A. Corn J. E. Strauch E. M. Khare S. D. Koga N. Ashworth J. Murphy P. Richter F. Lemmon G. Meiler J. Baker D. Fleishman, S. J. Leaver-Fay. Rosettascripts: a scripting language interface to the rosetta macromolecular modeling suite. *PLoS One*, 6(6):e20161, 2011.

[42] Stephanie Leavitt Ernesto Freire. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current Opinion in Structural Biology*, 11(5):560–566, 2001.

[43] P. Pfeffer H. Gohlke. Drugscorerna—knowledgebased scoring function to predict rna—ligand interactions. *Journal of Chemical Information and Modeling*, 47(5):1868–1876, 2007.

[44] Darin M. Taverna Richard A. Goldstein. Why are proteins marginally stable? *Proteins: Structure, Function, and Genetics*, 46:105–109, 2002.

[45] J. H. Niazi S. J. Lee Y. S. Kim M. B. Gu. *Bioorg. Med. Chem*, 16:1254–1261, 2011.

[46] Kuramitz H. *Anal Bioanal Chem*, 394:61–69, 2009.

[47] B. T. S. Bui K. Haupt. Molecularly imprinted polymers: synthetic receptors in bioanalysis. *Analytical and Bioanalytical Chemistry*, 398(6):2481–2492, 2010.

[48] Yannick Fuchsa Olivier Sopperab Karsten Haupta. Photopolymerization and photostructuring of molecularly imprinted polymers for sensor applications—a review. *Analytica Chimica Acta*, 717:7–20, 2012.

[49] A.G. Warren G.L. Ellingson B.A. Stahl M.T. Hawkins, P.C.D. Skillman. Conformer generation with omega: Algorithm and validation using high quality structures from the protein databank and the cambridge structural database. *J. Chem. Inf. Model.*, 50:572–584, 2010.

[50] Noel T. Southall Ken A. Dill A. D. J. Haymet. A view of the hydrophobic effect. *J. Phys. Chem. B*, 106:521–533, 2002.

[51] Marcus D Hanwell Donald E Curtis David C Lonie Tim Vandermeersch Eva Zurek Geoffrey R Hutchison. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, 4(17), 2012.

[52] M. J. Cho R. Juliano. Macromolecular versus smallmolecule therapeutics: drug discovery, development and clinical considerations. *Trends in Biotechnology*, 14(5):153–158, 1996.

[53] A. W. Zuardi I. Shirakawa E. Finkelfarb I. G. Karniol. Action of cannabidiol on the anxiety and other effects produced by 9-thc in normal subjects. *Psychopharmacology*, 76(3):245–250, 1982.

[54] Boder ET Midelfort KS Wittrup KD. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proc Natl Acad Sci*, 97(20):10701–10705, 2000.

[55] James M. Carothers Jonathan A. Goler Yuvraaj Kapoor Lesley Lara Jay D. Keasling. Selecting rna aptamers for synthetic biology: investigating magnesium dependence and predicting binding affinity. *Nucleic Acids Research*, 38(8):2736–2747, 2010.

[56] Wendy A. Lea and Anton Simeonov. Fluorescence polarization assays in small molecule screening. *Expert Opin Drug Discov*, 6(1):17–32, 2012.

[57] David M. Hoover Jacek Lubkowski. Dnaworks: an automated method for designing oligonucleotides for pcr-based gene synthesis. *Nucleic Acids Research*, 30(10):e43, 2002.

[58] Lazaridis T Karplus M. Effective energy function for proteins in solution. *Proteins*, 35(2):133–52, 1999.

[59] Benzhuo Lu Deqiang Zhang J. Andrew McCammon. Computation of electrostatic forces between solvated molecules determined by the poisson–boltzmann equation using a boundary element method. *Journal of Chemical Physics*, 122:214102, 2005.

[60] Peter M. Colman Michael C. Lawrence. Shape complementarity at protein/protein interfaces. *JMB*, 234:946–950, 1993.

[61] Karsten Haup Klaus Mosbach. Molecularly imprinted polymers and their use in biomimetic sensors. *Chem. Rev*, 100:2495–2504, 2000.

[62] T. Roemer J. Davies G. Giaever C. Nislow. Bugs, drugs and chemical genomics. *Nature Chemical Biology*, 8(1):46–56, 2012.

[63] Ball P. Water as an active constituent in cell biology. *Chem Rev.*, 108(1):74–108, 2005.

[64] Hailong Fu Gerald R. Grimsley Abbas Razvi J. Martin Scholtz C. Nick Pace. Increasing protein stability by improving beta-turns. *Proteins: Structure, Function, and Bioinformatics*, 77(3):491–498, 2009.

[65] title = Patrick S. Stayton Stefanie Freitag Lisa A. Klumb Ashutosh Chilkoti Vano Chu Julie E. Penzotti Richard To David Hyre Isolde Le Trong Terry P. Lybrand Ronald E. Stenkamp.

[66] Carter PJ. *Nat Rev Immunol*, 6:343–357, 2006.

[67] R. D. Jenison S. C. Gill A. Pardi B. Polisky. High resolution molecular discrimination by rna. *Science*, 263(5152):1425–1429, 1994.

[68] Z. X. Xu H. J. Gao L. M. Zhang X. Q. Chen X. G. Qiao. The biomimetic immunoassay based on molecularly imprinted polymer: a comprehensive review of recent progress and future prospects. *Journal of Food Science*, 76(2):R69–R75, 2011.

[69] Uri Piran William J. Riordan. Dissociation rate constant of the biotin-streptavidin complex. *Journal of Immunological Methods*, 133(1):141–143, 1990.

[70] S.A. Berson R.S. Yalow. Assay of plasma insulin in human subjects by immunological methods. *Nature*, 184:1648–1649, 1959.

[71] D. Inbar Y. Nussinov R. Wolfson H. J. Schneidman-Duhovny. Patchdock and symmdock: servers for rigid and symmetric docking. *Nucleic Acids Res*, 33(Web Server issue):W363–7, 2005.

[72] Schrödinger, LLC. The AxPyMOL molecular graphics plugin for Microsoft PowerPoint, version 1.0. May 2010.

[73] Gregory A Michaud Michael Salcius Fang Zhou Rhonda Bangham Jaclyn Bonin Hong Guo Michael Snyder Paul F Predki1 Barry I Schweitzer. Analyzing antibody specificity with whole proteome microarrays. *Nature Biotechnology*, 21:1509–1512, 2003.

[74] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[75] Ruczinski I Kooperberg C Fox BA Bystroff C Baker D. Simons KT. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, 34(1):82–95, 1999.

[76] Regina Stoltenburg Christine Reinemann Beate Strehlitz. Selex—a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering*, 24(4):1–20, 2007.

[77] W. A. Bonner H. R. Hulett R. G. Sweet and L. A. Herzenberg. Fluorescence activated cell sorting. *Rev. Sci. Instrum*, 43(3):404–409, 1972.

[78] Andrew D. Ellington Jack W. Szostak. In vitro selection of rna molecules that bind specific ligands. *Nature*, 346(30), 1990.

[79] Berger B Leighton T. Protein folding in the hydrophobic-hydrophilic (hp) model is np-complete. *J Comput Biol.*, 5(1):27–40, 1998.

[80] Berger B Leighton T. Phillip w. snydera jasmin mecinovića demetri t. moustakasa samuel w. thomas iiia michael hardera eric t. macka matthew r. locketta annie hérouxb woody shermanc george m. whitesides. *PNAS*, 108(44):17889–17894, 2011.

[81] Boehm T. *Nat Rev Immunol*, 11:307–317, 2011.

[82] Michael F. Holick Rachael M. Biancuzzo Tai C. Chen Ellen K. Klein Azzie Young Douglass Bibuld Richard Reitz Wael Salameh Allen Ameri Andrew D. Tannenbaum. Vitamin d2 is as effective as vitamin d3 in maintaining circulating concentrations of 25-hydroxyvitamin d. *JCEM*, 93(3):677–681, 2013.

[83] FF Richards WH Konigsberg RW Rosenstein JM Varga. On the specificity of antibodies. *Science*, 187(4172):130–137, 1975.

[84] Elena Pazos Olalla Vázquez José L. Mascareñas M. Eugenio Vázquez. Peptide-based fluorescent biosensors. *Royal Society of Chemistry*, 38:3348–3359, 2009.

[85] T. A. Walsh. The emerging field of chemical genetics: potential applications for pesticide discovery. *Nature Chemical Biology*, 63(12):1165–1171, 2007.

[86] A. J. Hampson M. Grimaldi J. Axelrod D. Wink. Cannabidiol and − delta-tetrahydrocannabinol are neuroprotective antioxidants. *Nat Protoc*, 95(14):8268–8273, 1998.

[87] M. L. Ashour M. Wink. Genus bupleurum: a review of its phytochemistry, pharmacology and modes of action. *Journal of Pharmacy and Pharmacology*, 63(3):305–321, 2011.

[88] A. W. Morais S. L. Guimarães F. S. Mechoulam R. Zuardi. Action of cannabidiol on the anxiety and other effects produced by 9-thc in normal subjects. *Journal of Clinical Psychiatry*, 56(10):485–486, 1995.

Chapter 8: Appendices

8.1 Appendix A: Alternative Methods and Filters

8.1.1 Scaffold Selection: Rosetta Match

The Rosetta Match algorithm [33] is another method for placing the ligand into our target scaffolds. Rosetta Match searches a set of scaffolds for a pre-defined configuration of binding residues. We have used three different approaches for matching: Matching for side chains that recapitulate the geometries observed in native ligand binding sites, matching for side chains that make idealized bonding interactions according to hybridization theory and/or maximized computed van der Waals interactions, and matching for side chain positions obtained by computationally placing disembodied amino acids near the ligand and minimizing the system. Iterative side chain placement and six-dimensional geometric hashing techniques are used to search the scaffold for these side chain configurations in linear time. In a variation of this approach, we also perform a Patchdock run prior to matching in order to pre-define a smaller search area within the shape-complimentary scaffolds. This latter approach enables higher density sampling.

8.1.2 Design: Rosetta Scripts

Rosetta Scripts is a implementation of Rosetta that uses an XML-like syntax for specifying modelling protocols. [41] Previous to Rosetta Scripts, one would only have access to individual protocols, such as docking, sequence redesign, interface design, etc... one at a time. Rosetta Scripts allows the user to mix and match these protocols, easily create variations, apply filters, and do fast testing and optimization of procedures without having to recompile. Using Rosetta Scripts, we have created a ligand binding design algorithm that involves multiple

rounds of minimization and sequence redesign with alternating hard and soft repulsive van der Waals penalties. The goal of this stage is to identify sequences within the identified binding sites that maximize affinity between the ligand and the protein. We use multi-stage filtering to stop bad trajectories early and save computation time. We also use a multi-criterion optimization procedure based on shape complementarity, ligand interface energy, Rosetta Holes packing measures, solvent accessible surface area, and hydrogen bond score terms.

8.1.3 Design: Boltzmann Electrostatics

The Poisson-Boltzmann equation, described by Lu *et al*, [59] is used to calculate the electrostatic forces between molecules in ionic solutions. We are implementing this model to calculate the electrostatic interactions between the surfaces of our ligand and the protein binding site. The model calculates an all-body electrostatic field, as opposed to the standard pairwise calculation previously used through Rosetta. Our implementation does a scan of all residues within the binding pocket to find substitutions that stabilize the bound state based on the electrostatics score, the $\Delta\Delta$G, a measure of the difference in Gibbs free energy between bound and unbound ligand states, and the total Rosetta score, but also do not destabilize the protein in the absence of the ligand.

8.1.4 Validation: RosettaDock

RosettaDock is a protocol that allows us to computationally validate our designs by performing protein-ligand docking that explicitly models full side-chain, backbone, and ligand flexibility. [36, 37] Monte Carlo sampling is used to explore all associated degrees of freedom. Five thousand runs of this protocol for each of our designs allows us to generate a "docking funnel". This is a plot of how much our ligand has moved from it's starting point versus the Rosetta energy of the complex. In the cases where Rosetta finds a global minimum of energy

and the ligand shows little to no movement from our initial model, we should see a "funnel" of data points leading to that low energy configuration. Native small molecule binders have this funnel character, as well as streptavidin. If our designs show a similar funnel without alternative low energy minima, then we can conclude that the design has successfully passed the docking validation criteria.

8.2 Appendix B: Sequence Data for Referenced Binders

2064: MKWNLDPSHTSFDFKVRHMGIASVRGSMKILSGSVETDEAGRP
IQAEAVFDAASIATGEPQRDGHLRSADFLHAEQYPESRFVSTQIEPLGG
NRYRVQGNATIRDITKPVTVEAEVSAPIKDPWGMQRVAASASGQINRKD
WNMTWNQVLELGALLVGEEMKMNLEIEAVAPAPVAAQ

2065: MKWNLDPSHTSFDFKVRHMGIASVRGSLKILSGSVETDEAGRP
IQAESVIDAASIATGEPQRDGHLRSADFLHAEQYPEIRFVSTQIEPLGG
NRYRVQGNLTIRDITKPVTIEAESSAPIKDPWGMQRAAASASGQINRKD
WNLTWNQVLELGALLVGEEVKFNAEIEAVAPAPVAAQ

4574: MTQTTQSPALIASQSLWRCAQAHDREGFLALMADDVVIELPIG
KSVSNPDGSGIKGKEAVGAFFDTAIAANRLTVTCEETFPSSSPDEIAHI
LVLHVEFDGGFTIEVRGVFTYRVNKAGLITNMRGYWNLDMMTFGNQE

4424: GQSAKEAIEAALADFVKAYNSKDAAGVASKYMDDAAIFPLDMA
RVDGRQNIQKLWQGLMDMGVSELKLTTLDVQESGDFAFESGSFSLKGPG
KDSKLVDVAGKYVVVWRKGQDGGWKLYRTISNLDPAK

HH24: GQSAKEAIEAALADFVKAYNSKDAAGVASKYMDDAAIFPLDMA

PVDGRQNIQKLWQGLMDMGVSEPKFTTLNVQESGDFAFESGSFSLKGPG
KDSKLVDIAGIYVEVWRKGQDGGWKLYRTIANLDPAK

HH35.1: DQSAKEAIEAALADFVKVYNSKDAAGVASKYMDDAAIFPLD
MAPVDGRQNIQKLWQGLMDMGVSEPKFTTLNVQKSGDFAFESGSFSLKG
PGKDSKLVDIAGIYVEVWRKGQDGGWKLYRTIANLDPAK

W19.1: LPTAHEAIEAALADFVKVYNSKDAAGVASKYMDDAVIFPLDM
ARVDGRQNIQKLWQGLMDMGVSEPKFTTLNVQESGDFAFESGSFSLKGP
GKDSKLVDIAGIYVEVWRKGQDGGWKLYRTIANLDPAR

6218: NLPTAQEVQGLAARMIELLDVGDIEAIVQMYADDATLEAPFGQ
PPIHGREQIAAFFRQGLGGGKVRACLTGPVRASHNGCGAAPFRVETVWN
GQPCALDVISVSRFDEHGRIQTTQAYYSEVNLSVREPQ

6220: NLPTAQEVQGLMARLIELVDVGDIEAIVQMFADDATVEAPFGQ
PPIHGREQIAAFFRQGLGGGKVRACLTGPVRASHNGCGAMPFRIEMVWN
GQPCALDVISVIRFDEHGRVQTMQAYFSEVNMSVREPQ

6234: GSSSSGREQGHMNAKEILVHALRLVENGDARGFCDLFHPEGVM
EFPYAPPGYKTRFEGRETIWAHMRLFPEHLTIRFTDVQFYETADPDLAI
GEFHGDGVATVSGGKLAQDFISVLRTRDGQILLSRIFWNPLRHLEALGG
VEAAAKIVQGA

6264: GQSAKEAIEAANADFVKAYNSKDAAGVASKYMDDAAMFPPDMA
RVDGRQNIQKLFQGSMDMGASEVKITTLDVQESGDFAFESGSFSAKVPG
KDSKLVDIAGKYVVVVWRKGQDGGWKLYRDIFNSDPAK

6269: SNAMSGNVGAGRHADELAIRTVQYRFVEATRKFDRQVLSSLMT
DDVVFYTPGRLPFGKEEFLAAAEQNDQRVIIEMSVTFEEIVIVEPMAYT
RTHVHIKVTPRSGGAVRELAGHIMSIFRRSMFGEWQLARAYALVVPI

6326: NLPTAQEVQGLMARFIELMDVGDIEAIVQMFADDATVELPFGQ
PPIHGREQIAAFFRQMLGGGKVRMCLTGPVRASHNGCGAMPFRAEYVWN
GQPCALDVIAVMRFDEHGRIQTSQAYFSEVNLSVREPQ

6344: GQSAKEAIEAATADFVKAYNSKDAAGVASKFMDDAAAFPPDMA
RVDGRQNIQKLWQGAMDMGASEAKATTLDVQESGDFAFESGSFSLKAPG
KDSKLVDAAGKYVAVWRKGQDGGWKIYRLIFNSDPAK

6348: NTPEHMTAVVQRFVAAMNAGDLDGIVALFADDATVECPVGSEP
RSGTAAIREFFANALKLPVAVELTQEVRAVANEAAFAFTASFEYQGRKT
VVAPIAHFRFNGAGKVVSSRCLFGEKNIHAGA

P2: MTTGEHIAALTALVETYVMALTRGDRPALERIFFGKASSVGHYEG
ELLWNSRDAFIAVCEDAADAGTDPFWAISSVSVQGDIAMLHVELDWAGM
RFDVFLTVLLHEGSWRIVSSVYRIR

2082: MTTSEHIAALTALVETYVMALTRGDRPALERIFFGKASSVGHY
EGELLWNSRDAFIAMCEDAADAETDPFWAISSVSVQGDIAMLHVELDWA
GMRFDVFLTVLLHEGSWRIVSSVYRIR

UM_37: NLPTAQEVQGLMARYIELADVGDIEAIVQMYADDATKELPF
GQPPIHGREQIAAYFRAGGKVRVCLTGPVRASHNGCGAMPYRSETVWNG

QPAAVDAISVMRFDEHGRIQTHQIYCTAVKVSV

WBD1-V3: NLPTAQEVQGLMARYIELADVGDIEAIVQMYADDATKELP
FGQPPIHGREQIAAYFRAGGKWRICLTGPVRASHNGRGAMPWRSETVWN
GQPAAVDCISVMRFDEHGRIQTHQIYCTAVKVSV

N3X-1: GNSSRGREQGHMNAKEILVHALRLVENGDARGFCDLFHPEGV
MEFSYAPPGYKTRFEGRETIWAHMRLFPEPLTIRFTDVQFYETADPDLA
IGEFHGDGVATVSGGKLAQDFISVLRTRDGQILLSRIFWNPLRHLEALV
GVEAAAKIVQGA

CM1-13 AD28: KWNLDPSHTSFDFKVRHVGIASGRGSMKILSGSVET
DEAGRPIQAEVVFDAASIATGEPQRDDHLRSADFLHAEQYPESRFVSTL
IEPLGGNRYRVQGNVTIRDITKPVTVEAEVSAPIKDPWGTQRVAASASG
QINPKDWNMTWNQVLELGALLVGEEMKMNLEIEAAAPAPVAAQ

J1c-16: DQSAKEAIEAALADFVKVYNSKDAAGVASKYMDDAAIFPL
DMAPVDGRQNIQKLWQGLMDMGVSEPKFTTLNVQKSGDFAFESGSFSL
KGPGKDSKLVGIAGIYVEVWRKGQDGGWKLYRTIANLGPAK

NAL1: ETIQPGTGYNNGYFYLFWNHGHGGVTYTNGPGGQFSVNWSNS
GLFIGGKGWQPGTKNKVINFSGSYNPYGNSYLSVYGWSRNPLFAYFIV
ENFGTYNPSTGATKLGEVTSDGSVYDIYRTQAVNQPSIIGTATFYIYW
SVRRNHRSSGSVNTANHFNAWAQQGLTLGTMDFQIVAVGGYFSSGSAS
ITVS

CS1:GQSAKEAIEAANADFVKAYNSKDAAGVASKFMDDAAAFPPDMAR

VDGRQNIQKLFQGSMDMGASEVKMTTLDVQESGDFAFESGSFSAKAPG
PDSKLVDWAGKYVVVWRKGPDGGWKMYRSIFNSDPAK

CS3:GQSAKEAIEAANADFVKAYNSKDAAGVASKFMDDAAAFPPDMAR
VDGRQNIQKLFQGAMDMGASEAKLTTLDVQESGDFAFESGSFSAKAPG
PDSKLVDAAGKYVVVWRKGPDGGWKLYRDIWNSDPAK

CS4:TAEVESHMTAHFGKTLEECREESGLSVDILDEFKHFWSDDFDVV
HRELGCALLCAANKFSLLDPNNAMNPVNMDEFTKSFPNGQVLAEKQVK
LIANCAKQFATVTDACTAAVKVAACFKEDSRKEGIAPEVAMVEAVIEK
Y

6258:GQSAKEAIEAANADFVKAYNSKDAAGVASKFMDDAAAFPPDMA
RVDGRQNIQKLFQGSMDMGASEVKMTTLDVQESGDFAFESGSFSAKAP
GKDSKLVDWAGKYVVVWRKGQDGGWKMYRSIFNSDPAK

6261:GQSAKEAIEAANADFVKAYNSKDAAGVASKFMDDAAAFPPDMA
RVDGRQNIQKLFQGAMDMGASEAKLTTLDVQESGDFAFESGSFSAKAP
GKDSKLVDAAGKYVVVWRKGQDGGWKLYRDIWNSDPAK

6333:TAEVESHMTAHFGKTLEECREESGLSVDILDEFKHFWSDDFDV
VHRELGCALLCAANKFSLLDDNNAMNHVNMDEFTKSFPNGQVLAEKQV
KLIANCAKQFATVTDACTAAVKVAACFKEDSRKEGIAPEVAMVEAVIE
KY