2017

Computational Approaches For Designing Protein/inhibitor Complexes And Membrane Protein Variants

Krishna Gajan Vijayendran
University of Pennsylvania, kvija@mail.med.upenn.edu

Computational Approaches For Designing Protein/inhibitor Complexes And Membrane Protein Variants

Abstract

Drug discovery of small-molecule protein inhibitors is a vast enterprise that involves several scientific disciplines (i.e. genomics, cell biology, x-ray crystallography, chemistry, computer science, statistics), with each discipline focusing on a particular aspect of the process. In this thesis, I use computational and experimental approaches to explore the most fundamental aspect of drug discovery: the molecular interactions of small-molecules inhibitors with proteins.

In Part I (Chapters I and II), I describe how computational docking approaches can be used to identify structurally diverse molecules that can inhibit multiple protein targets in the brain. I illustrate this approach using the examples of microtubule-stabilizing agents and inhibitors of cyclooxygenase(COX)-I and 5-lipoxygenase (5-LOX).

In Part II (Chapters III and IV), I focus on membrane proteins, which are notoriously difficult to work with due to their low natural abundances, low yields for heterologous over expression, and propensities toward aggregation. I describe a general approach for designing water-soluble variants of membrane proteins, for the purpose of developing cell-free, label-free, detergent-free, solution-phase studies of protein structure and small-molecule binding. I illustrate this approach through the design of a water-soluble variant of the membrane protein Smoothened, wsSMO. This wsSMO stands to serve as a first-step towards developing membrane protein analogs of this important signaling protein and drug target.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Jeffrey G. Saven

Second Advisor
Amos B. Smith

Keywords
drug design, experimental biophysics, medicinal chemistry, organic chemistry, physical chemistry, protein design

Subject Categories
Computer Sciences | Organic Chemistry | Physical Chemistry

COMPUTATIONAL APPROACHES FOR DESIGNING PROTEIN/INHIBITOR COMPLEXES AND MEMBRANE PROTEIN VARIANTS

Krishna Gajan Vijayendran

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

____

Dr. Jeffrey G. Saven, Ph.D.

Professor of Chemistry

Graduate Group Chairperson

____

Dr. Li-San Wang, Ph.D.

Associate Professor of Pathology

Dissertation Committee

Dr. Kim Sharp, Ph.D. (Associate Professor Of Biochemistry and Biophysics)

Dr. Gideon Dreyfuss, Ph.D. (Isaac Norris Professor of Biochemistry and Biophysics)

Dr. Ravi Radhakrishnan, Ph.D. (Professor of Bioengineering)

`

COMPUTATIONAL APPROACHES FOR DESIGNING PROTEIN/INHIBITOR COMPLEXES AND MEMBRANE PROTEIN VARIANTS

`

ABSTRACT

COMPUTATIONAL APPROACHES FOR DESIGNING PROTEIN/INHIBTOR COMPLEXES AND MEMBRANE PROTEIN VARIANTS

Krishna Gajan Vijayendran (Advisor: Dr. Jeffrey G. Saven)

The discovery of small-molecule inhibitors of protein targets is a vast enterprise that involves several scientific disciplines (i.e. genomics, cell biology, x-ray crystallography, chemistry, computer science, statistics), with each discipline focusing on a particular aspect of the process. In this thesis, I use computational and experimental approaches to explore the most fundamental aspect of drug discovery: the molecular interactions that take place between drugs and their protein targets.

In **Part I: Designing a Drug**, I describe how computational docking approaches can be used to identify structurally diverse molecules that can inhibit multiple protein targets in the brain. I illustrate this approach using the examples of microtubule-stabilizing agents and inhibitors of cyclooxygenase (COX)-I and 5-lipoxygenase (5-LOX).

In **Part II: (Re)Designing a Drug Target**, I focus on membrane proteins, which are notoriously difficult to work with due to their low endogenous levels, low yields from protein expression systems, and propensities to aggregate in solution. I describe a general approach for designing water-soluble variants of membrane proteins, for the purpose of developing a cell-free, label-free, detergent-free, solution-phase assay for assessing ligand-binding events. I illustrate this approach though the design of a water-soluble variant of the membrane protein Smoothened, wsSMO. wsSMO stands as a first-step towards developing membrane protein analogs of this important signaling protein and drug target.

`

TABLE OF CONTENTS

PART I: DESIGNING A DRUG

`

PART II: (RE)DESIGNING A DRUG TARGET

CHAPTER II: Application of Docking to Alzheimer's Disease Drug Targets

`

CHAPTER III: Computational design of membrane protein water-soluble variants

CHAPTER IV: Experimental characterization of wsSMO

`

APPENDIX

`

LIST OF TABLES

`

LIST OF FIGURES

`

`

`

`

FORWARD

My motivation for a research career is driven by the promise of science to develop effective therapeutics against diseases that currently plague us. Over the past several years, it dawned on me that the pathophysiology of many diseases, particularly in the fields of infectious diseases and oncology, can be reduced to a joint genetic-molecular basis. Examples of this include genetic mutations that result in acquired drug resistance via binding site mutations (i.e. HIV protease mutations that were identified in the 1990s, and Smoothened mutations that were identified in the 2010s in basal-cell carcinoma); and the acquisition of genetic elements that result in resistance to certain antibiotics (i.e. methicillin-resistant *Staphylococcus aureus*). Identifying the genetic states that exist within a disease context, and identifying drugs that are effective (and ineffective) towards each state, can allow for the development of effective therapeutic regimens. This is illustrated by the emergence of second-line HIV protease inhibitors, and by the use of antibiotic combinations to treat dangerous drug-resistant bacterial strains. From these examples, a general approach to drug design can be formulated: use biological experiments and structural analyses to determine the role of a protein target in a disease; from this information, use chemistry to design molecules that can inhibit this protein's activity; and use genomic analyses to identify patients whose disease is driven by this protein. Extrapolating this approach to several disease-relevant protein targets can allow for the identification of regimens consisting of drugs that attack multiple protein targets.

During my Ph.D. years, my goal was to develop a core background that would allow me to one day become involved in this type of drug design process. The work described in **Part I: Designing a Drug** allowed me to identify drug candidates via computational structure-based drug design, and to make a subset of these candidates using synthetic organic chemistry techniques. The work described in **Part II: (Re)Designing a Drug**

Target, spanned the field of physical chemistry, protein design and biochemistry. It allowed me to understand the principles underlying protein structure; the experimental approaches for protein purification; and the different biophysical approaches to test ligand-binding. Both projects existed at the interface between medicine, chemistry and computer science.

I thank you for taking the time out to read about my research, and to learn about the research approach that I will be dedicating my career to.

Krishna Vijayendran
December 19th, 2016
Philadelphia, PA

`

INTRODUCTION

Drug discovery of small-molecule protein inhibitors is a vast enterprise that involves several scientific disciplines. These disciplines include genomics (identification of potential drug targets from clinical samples of a particular disease), biology (experimental validation of these targets in cellular and animal models), x-ray crystallography (acquisition of the target's macromolecular structure), computational chemistry (prediction of the activity of drug candidates based on the target's structure), statistics (analysis of high-throughput screening data of millions of compounds), organic chemistry (synthesis and optimization of candidate compounds obtained from a screen), pharmacology (evaluation of bioavailability and toxicity of candidate compounds in animal models) and, lastly, clinical medicine (conducting clinical trials of candidate drugs in human beings).

Each of these disciplines approaches the drug discovery process from a unique perspective, and are all vitally important to the end goal of identifying compounds that can successfully treat a disease. In the following chapters, I use computational and experimental approaches to explore what I consider to be the most fundamental aspect of the drug discovery process: the molecular interactions of small-molecule inhibitors with proteins.

In **Part I: Designing a Drug (Chapters I and II)**, I focus on the computational design of drugs, and the modeling of drug-protein complexes. I describe how computational docking approaches can be used to not only develop brain-penetrant inhibitors, but to also identify compounds that can inhibit multiple targets in the brain. In **Part II: (Re)Designing a Drug Target (Chapters III and IV)**, I switch my focus from drugs to the protein targets themselves. Membrane proteins are a class of protein targets that are notoriously difficult to work with and develop drugs against. Using principles from statistical mechanics, I developed an algorithm that can take an input membrane protein structure and output a water-soluble variant of it. Such a variant would enable one to test drug binding in a solution-phase assay, without the use of detergents, fluorescent labeling or radio-labeling.

Part I: Challenges with designing brain penetrant inhibitors

For the first part of my thesis, I start off by focusing on (re)designing small-molecule inhibitors for a new disease context. The vast majority of drugs inhibit protein targets that are found in peripheral tissues. However, many of these peripheral targets have also been found to be involved in the pathophysiology of different neurological disorders. A first attempt at inhibiting one of these proteins in the brain would be to administer the same drug that inhibits the target peripherally. However, due to the blood brain barrier, many of these compounds (which often have very polar functional groups) have poor brain penetrance and are ineffective. From this, two questions arise. Can one take a small-molecule inhibitor that is known to bind to a particular target, and modify it's functional groups so that now it not only crosses the blood-brain barrier, but can still inhibit it's target? Going one step farther, can one take a known brain-penetrant inhibitor that binds a particular target, and repurpose it to inhibit new targets in the brain? Using computational approaches (i.e. chemical enumeration to design large compound libraries, and molecular docking algorithms) and synthetic organic chemistry, I test these two hypotheses through the design of two classes of candidate Alzheimer's Disease inhibitors: microtubule-stabilizing agents, and inhibitors of cyclooxygenase (COX)-I and 5-lipoxygenase (5-LOX).

`

Part II: Challenges with directly probing small-molecule/membrane protein interactions

For the second part of my thesis, I explore inhibitor-target molecular interactions from the other direction, and focus on the drug targets themselves. For many protein targets, drug discovery has been accelerated due to the availability of assays that allow for the direct determination of drug-protein interactions (i.e., in vitro kinase assays for cancer drug development). However, for membrane proteins (the largest class of drug targets), such a direct method does not exist. Current membrane protein binding methods are either cell-based, competition-based, or require the use of detergents. Cell-based assays are expensive, time-consuming, difficult to scale up, and provide little information about the molecular interactions involved in drug binding. Furthermore, developing such an assay often requires knowledge of a protein's downstream pathway in order to obtain a read-out, which may not be possible for newly discovered membrane protein targets. Competition-based assays test whether a candidate molecule is able to displace binding of a radioactive or fluorescently labeled known binder in cells over-expressing the membrane protein, or in crude isolated membrane protein fractions. However, these methods are unable to identify molecules that bind outside of the labeled-protein's binding site, making them inappropriate for identifying compounds that target protein-protein interactions or drug resistant mutants. Detergent-based assays are technically challenging, require multiple rounds of optimization of experimental conditions, and are prone to problems with protein stability and protein aggregation.

The reason why it is difficult to obtain a quick and direct membrane protein/drug-binding assay is due to the inherent structural features of membrane proteins. In contrast to water-soluble globular proteins, which on average tend to have hydrophobic residues packed in the interior and hydrophilic residues on the exterior exposed to solvent,

`

membrane proteins have large numbers of exterior, lipid-contacting hydrophobic residues on the transmembrane region. These hydrophobic residues make them difficult to isolate from a protein expression system (i.e. E. coli, yeast or Sf9 cells) due to toxicity and the formation of inclusion bodies. In the event that one is able to isolate them, their solubility is poor, which makes them prone to aggregation and difficult to reconstitute in native forms.

From these problems, a question arises: can one computationally design a membrane protein variant that is water-soluble, yet retains the wild-type's structural and ligand-binding properties? Such a variant would allow for the development of cell-free, label-free, detergent-free, solution-phase studies of protein structure and small-molecule binding. I test this hypothesis and illustrate this approach through the design of a water-soluble variant of the membrane protein Smoothened, wsSMO. wsSMO stands to serve as a first-step towards developing membrane protein analogs of this important signaling protein and drug target.

`

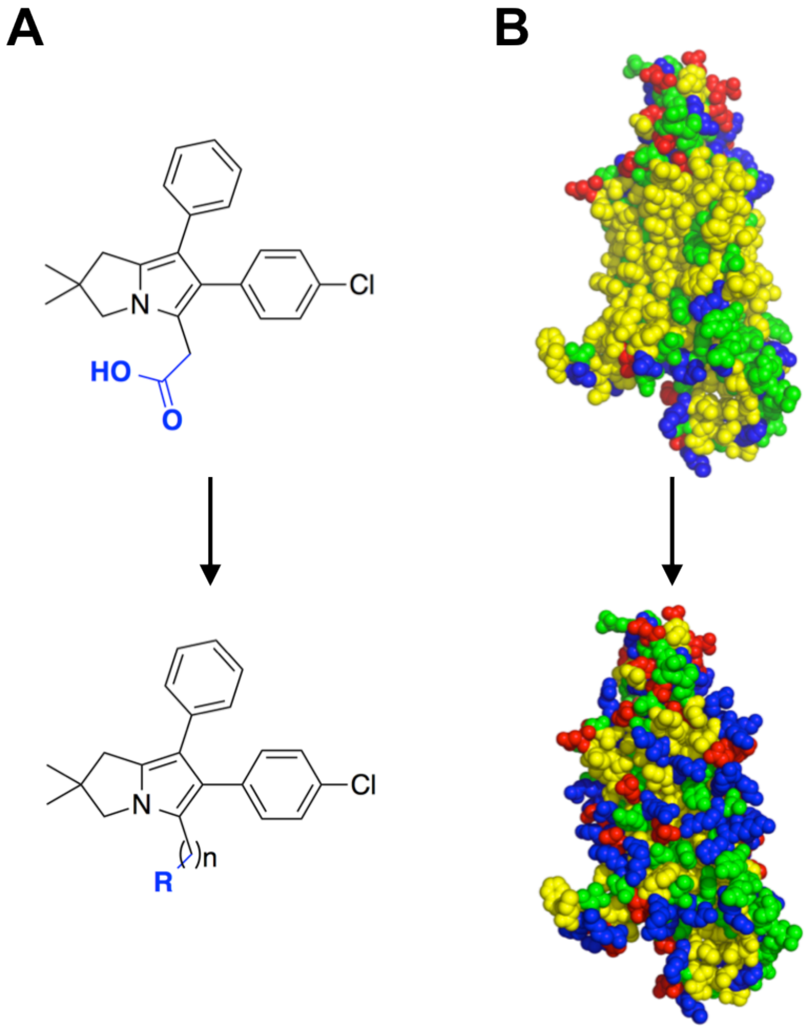


FIGURE 0: OVERVIEW OF THIS THESIS

In this thesis, I focus on drug design from two directions: the drug, and the drug target. I address the following questions in these sections:

A. Part I: Designing a Drug: Can one take an existing drug that does not penetrate the brain, and redesign it so that it can not only cross the blood-brain barrier, but still inhibit its target there? Can one repurpose an existing drug so that it can inhibit multiple targets in the brain?

B. Part II: (Re)designing a Drug Target: Can one take a membrane protein and design a water-soluble variant of it that retains the wild-type's structural and ligand-binding properties, yet can be used in solution-phase binding assays?

`

PART I: DESIGNING A DRUG

CHAPTER I

MOLECULAR DOCKING IN STRUCTURE-BASED DRUG DESIGN

Molecular docking is a computational modeling tool that is widely used in both academic and industrial drug development programs. This approach aims to calculate the binding energy between a candidate compound and a target protein, using principles from theoretical physical chemistry and algorithmic methods from computer science. This approach has led to the development of several FDA-approved drugs[1-4], and is the approach that we used to develop candidate brain-penetrant inflammation inhibitors and microtubule-stabilizing agents for Alzheimer's Disease (**Chapter II**). In the following sections, I will provide explain the theory behind Autodock, one of the most widely-used docking algorithms in the field that has been refined and improved upon since it's introduction over 25 years ago[5]. Though Autodock is one of many docking algorithms, understanding the basis of it can provide a good starting point for understanding these other approaches. First, I will describe the principals underlying structure-based drug design and structure-activity relationship (SAR) optimization of candidate drugs. Next, I will explain how docking algorithms such as Autodock attempt to automate this process, and describe how they calculate the free-energy of binding between a candidate drug and a protein's binding site. Afterwards, I will explain how Autodock represents the drug and the protein, and present two methods that it can use to "search" a protein's binding-site for locations where a drug (at a specific conformation) can bind favorably. Lastly, I will explain how different metrics can used to determine if docking hits are "drug-like," in order to help prioritize compounds for synthesis.

`

1.1. MOLECULAR INTERACTIONS IN DRUG DESIGN

Binding of a small-molecule inhibitor to a protein target is facilitated by specific attractive molecular interactions between atoms of the inhibitor and atoms of the protein binding site residues. The goal of structure-based drug design is to use structural information from a protein target (or a closely related homologue) to identify specific molecular interactions that candidate inhibitors can exploit, and to synthesize these candidates and test them for activity. This process is an iterative process: based on the activity data, the medicinal chemist will go back to the structure and attempt to understand why certain functional groups on the compounds worked and why certain ones did not.

Based on this analysis, they will devise additional compounds to synthesize. This entire process is known as structure-activity-relationships, or SAR, optimization. Traditionally, the structure that is used is an x-ray crystallographic structure of the protein bound to an endogenous ligand, or to a previous small-molecule inhibitor that the medicinal chemist is trying to improve upon.

This process can be illustrated by the following example. **FIG 1-1** shows the x-ray structure of the G Protein-Coupled Receptor Smoothened (SMO) bound to the small-molecule inhibitor taladegib[6]. Upon a close analysis of the structure, different molecular interactions between the inhibitor and the protein can be inferred:

- the bulky phthalazine ring is buried deep in the pocket, making hydrophobic interactions with residues such as L522, W281, and I389
- the more polar amide and 4-fluoro-2-trifluoromethylphenyl groups exist nearer to the solvent- exposed mouth of the cavity
- hydrogen bond interactions exist between the phtalazine nitrogens and R400, and between the amide carbonyl carbon and N219

`

- a $\pi - \pi$ interaction exists between F483 and the phenyl ring of the 4-fluoro-2-trifluoromethylphenyl group

Using these observations, a few key principles emerge that can be used to develop second-line SMO inhibitors. The scaffold of the drug needs to contain two aromatic groups flanking a central polar nitrogen. One of the aromatic groups will be buried (to presumably form favorable hydrophobic interactions), and another will be more solvent exposed. These general principals are validated when one looks at structures of ligands that have been found to bind to SMO with nanomolar affinity[7] **(FIG 1-2)**. With the exception of sonidegib, these compounds have been crystallized with SMO. The groups that are buried in the cavity (orange) or more solvent-exposed near the mouth of the cavity (blue), together with the central nitrogen (red), have been indicated.

The SAR process using structural information is powerful because it allows a medicinal chemist to make explore the chemical space around an inhibitor in a reasoned and justifiable manner. However, this process is very slow and arduous. Each functional group has its own unique chemical properties, requiring a plethora of reaction conditions that need to be optimized and, in many cases, devised from scratch for that particular ligand scaffold. This can involve very complex synthetic routes to make compounds whose potential activity is uncertain. Further difficulty comes from the process of inferring the molecular interactions themselves. As we saw in **FIG 1-1**, a plethora of interactions can be inferred between a structure and a ligand, but it is not clear that all of them are relevant for binding. For example, the fluourine group substituted on the phenyl ring is often added late in compound optimization to increase the lipophilicity of compounds that suffer from poor biodistribution in animal models, or to block liver metabolism and enhance absorption; it is not clear that these fluorines were placed for

`

enhance molecular interactions with Smoothened. Hence, trying to infer every single interaction and perform SAR optimization on each of them can quickly turn into a labor-intensive, time-consuming and expensive process.

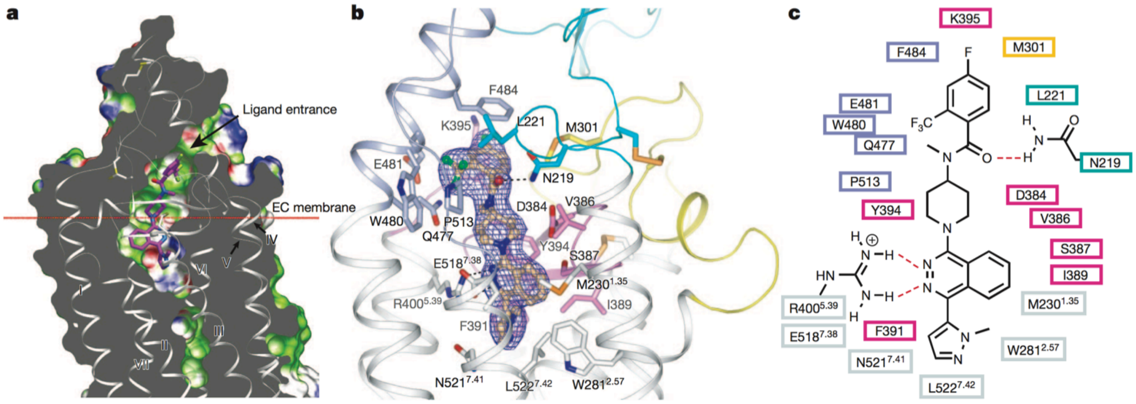


FIGURE 1-1: X-ray structure of taladegib bound to Smoothened

Molecular interactions made between the small-molecule inhibitor taladegib and the Smoothened binding site reveal several molecular interactions that can be used to develop second-line agents. From Figure 3 of Wang et al. 2013[6]

`

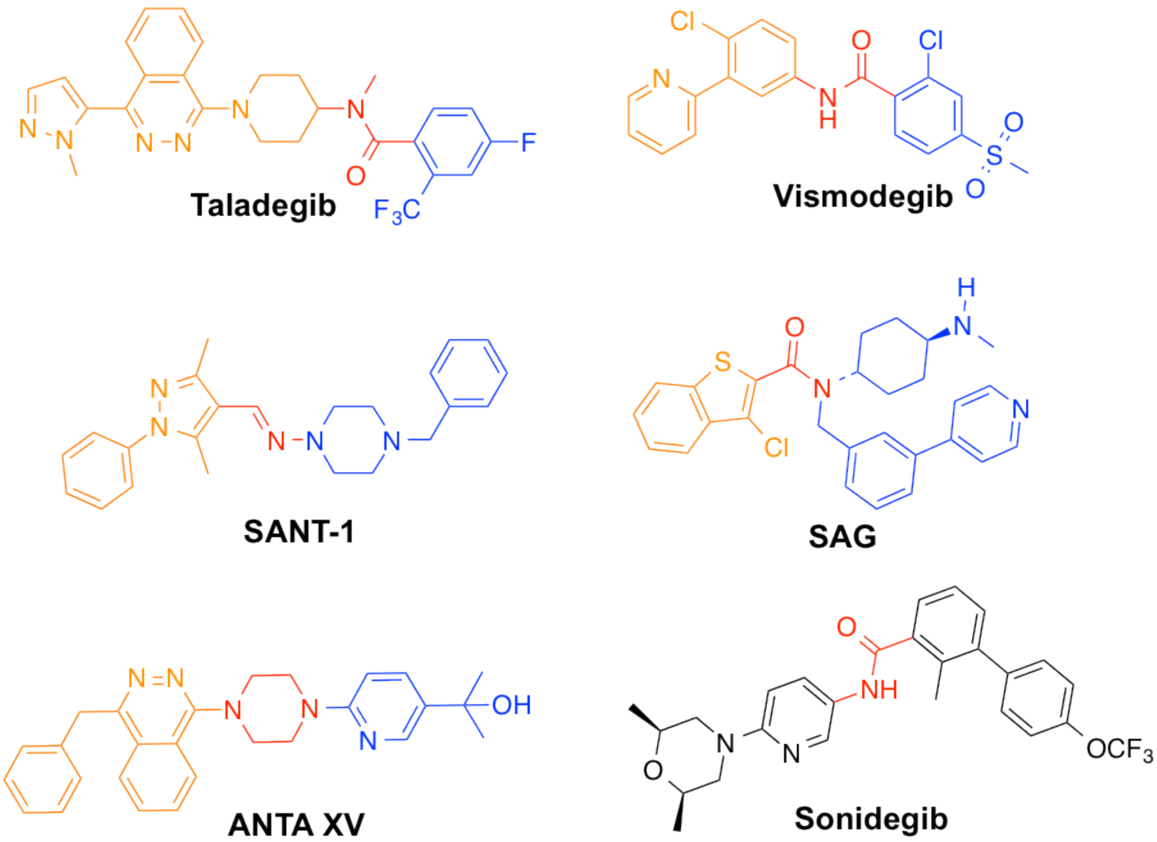


FIGURE 1-2: Pharmacophores of Smoothened ligands

Known SMO compounds have several structural features in common[6,7]: a polar nitrogen group (red) flanked by an aromatic ring racing the mouth of the binding-cavity (blue) and a ring buried deeper into the pocket (orange). These three regions represent pharmacopores that can be varied to create new compounds. Note: since the x-ray crystal structure of sonidegib with Smoothened has not been solved, locations of the aromatic rings are not currently known.

`

The application of molecular docking algorithms to structure-based drug discovery has provided medicinal chemists with a greater amount of versatility for dealing with these aforementioned difficulties. Rather than using qualitative guesses about functional group substitution to prioritize compounds to synthesize, docking algorithms can quickly test whether a large number of diverse candidate ligands will bind to a protein target of interest **(FIG 1-3)**. From these results, the medicinal chemist can have an idea of the types of ligand scaffolds or functional group replacements to use for SAR optimization. This can save time over the traditional method of synthesizing a diverse set of ligands with different substitutions in hopes of hitting upon one with a favorable binding energy. In essence, molecular docking allows one to automate a large part of the SAR optimization process that was just described.

There are various types of docking algorithms that take different philosophical approaches to solving the protein-drug interaction problem. Despite these differences, there are key features that all good docking algorithms must be able to do[8]:

- Determine compounds from a library that can theoretically inhibit a protein target
- Identify the **true pose** for a known protein-drug interaction, when docking that drug into the apo form of the protein
- Output the **top candidates** from a ligand library
- Calculate the binding score **rapidly**

To understand how these goals are achieved, we need to look inside how the algorithms work, and how they make use of principles from thermodynamics and apply them to the small-molecule inhibitors. The key questions that need to be addressed regarding these docking algorithms are:

`

- How is the **binding energy** between a protein and a drug calculated?
- How are the drug and the protein **represented**?
- How does the algorithm **dock** drug to proteins?
- How does one **evaluate** the hits, and prioritize the ones to actually synthesize?

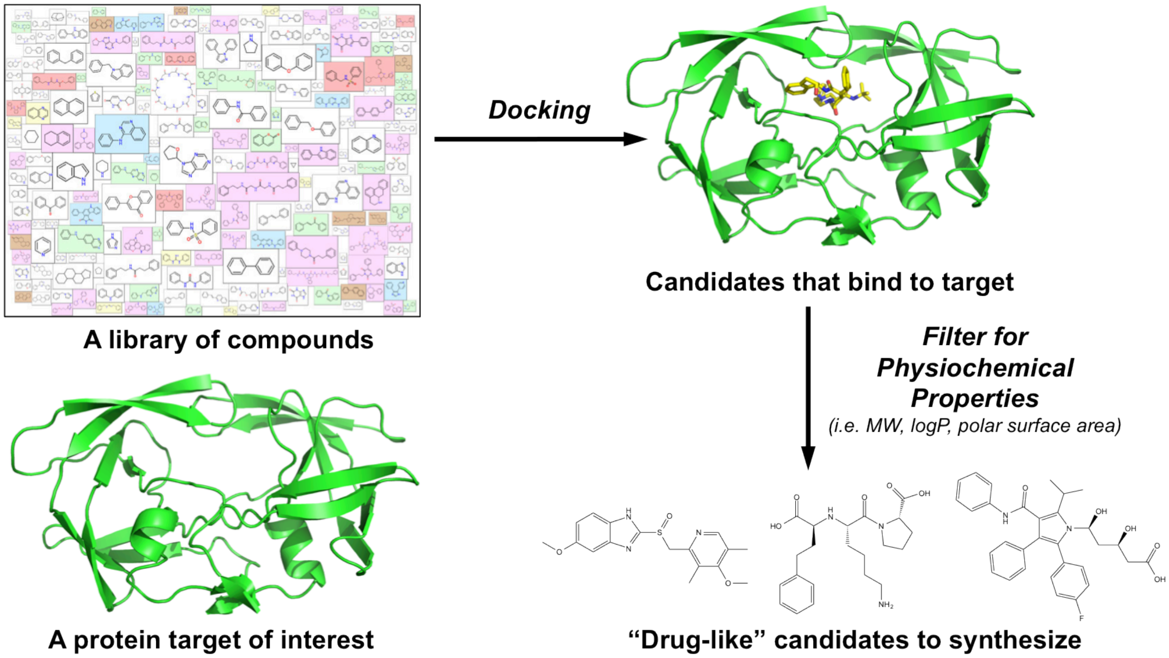

A library of compounds
Docking
Candidates that bind to target
Filter for Physiochemical Properties
(i.e. MW, logP, polar surface area)
A protein target of interest
"Drug-like" candidates to synthesize

FIGURE 1-3: Overview of the molecular docking process

Starting with a protein structure of interest (usually an x-ray crystal structure, but alternatives include an NMR structure, cryo-EM structure or homology model), and a library of computationally-designed compounds, one can use docking algorithms to identify candidate compounds that are predicted to bind to this target. Afterwards, the docking hits can be filtered according to physiochemical properties that are highly correlated with "drug-like" chemical space (i.e. Lipinski's Rule of 5, cut-offs involving lipophilicity and total polar surface area, ligand efficiency) in order to come up with a final list of candidate drugs to synthesize.

`

1.2. BINDING FREE ENERGY MODEL FOR MOLECULAR DOCKING

Though there are examples of drugs that make covalent bonds with their targets (i.e. irreversible HIV protease inhibitors[9]), the majority of identified drug-target interactions occur via non-covalent interactions such as van der Waals interactions, hydrogen bonds, electrostatic interactions, and hydrophobic interactions[10]. In this section, I will introduce the concept of the binding free-energy; explain the nature of these non-covalent interactions; and explain how Autodock mathematically represents these interactions.

a. Binding free energy, enthalpy and entropy

When a protein binds to a drug, a protein-drug complex is formed; this complex exists in equilibrium with the individual protein and drug components[8] (**FIG 1-4**).

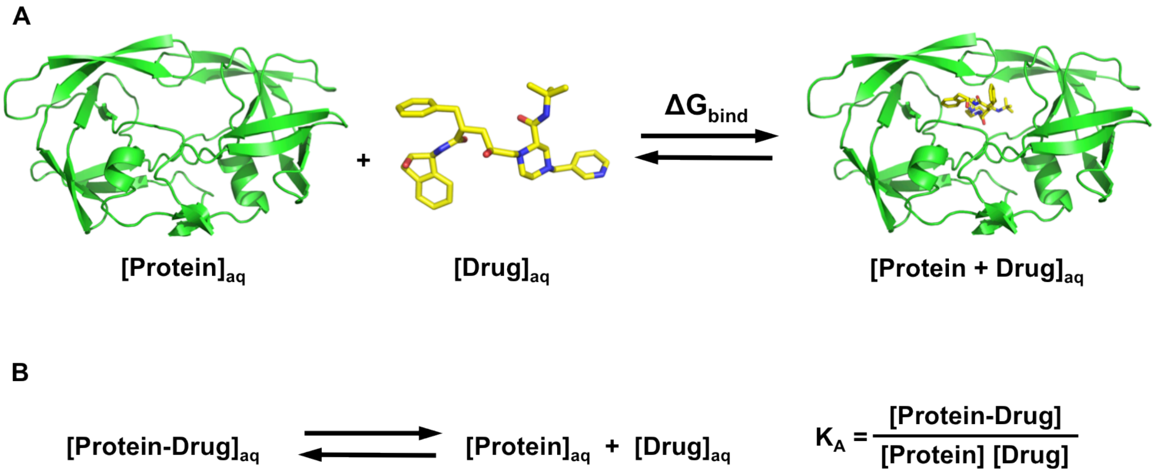


FIGURE 1-4 Thermodynamics of drug-protein target interactions

A: Interaction of a free protein, $[\text{Protein}]_{aq}$, and a free drug, $[\text{Drug}]_{aq}$, can result in the drug binding to the protein and a protein-drug complex, $[\text{Protein + Drug}]_{aq}$, being formed in the event that there is a favorable free energy of binding, ΔG_{bind}.

B. Definition of the association constant, K_A, of the protein-drug complex.

`

The free energy of binding (*ΔG*$_{bind}$) is related to the binding affinity K_A:

$$K_A = e^{\frac{-\Delta G_{bind}}{RT}} \quad \mathbf{(1.1)}$$

or, equivalently,

$$\Delta G_{bind} = -RTln(K_A) \quad \mathbf{(1.2)}$$

When an inhibitor is present in a protein's binding pocket, a binding event takes place only when it is associated with a high binding affinity and, correspondingly, a favorable (negative) binding free energy (*ΔG*$_{bind}$). According to Gibb's Free Energy, this binding free energy is the sum of an enthalpic component an enthalpic component (*ΔH*) and an entropic component (-*TΔS*)[11]:

$$\Delta G_{bind} = \Delta H - T\Delta S \quad \mathbf{(1.3)}$$

The enthalpic component refers to the change in heat that occurs when the drug binds to the protein[11]. For example, with hydrogen bond interactions (**section 1.D**), the optimal distance and angle between the hydrogen bond donors and acceptors will lead to an optimal hydrogen bonding energy, which in turn will contribute favorably to the binding free energy. The entropic[12] component refers to the change in disorder to the entire system (i.e. the drug, protein binding site residues, water molecules, co-factors and ions that are present in the binding pocket) that occurs as a result of the binding event.

There are two major terms that contribute to binding entropy: the entropy change related to the conformations of the ligand and the protein residues (conformational entropy), and the entropy change related to the desolvation of the drug's polar atoms and protein's polar residues. Conformational entropy is related loss of rotational and translational degrees of freedom of both the drug and the protein upon binding. Desolvation entropy is related to the hydrogen bonding interactions that water molecules make with atoms of

`

the drug and protein binding site residues. If the hydrogen bond interactions between the protein residues and ligand are weaker than those each makes to water molecules (i.e. due to suboptimal distances and angles between the protein hydrogen bond donors and drug hydrogen bond acceptors), then the enthalpic component of hydrogen bonding will be unfavorable. However, if the opposite is true, then water molecules will be released (as hydrogen bonds form between the protein and the drug) and the entropy of the system increases since the waters gain rotational and translational degrees of freedom. In this setting, the desolvation entropy increases.

The situation is different for hydrophobic groups on the drug[13]. These groups are incapable for forming hydrogen bonds with water, and when they are introduced to solvent a disruption occurs in the water-hydrogen bonding network. The waters will orient themselves along the surface of the hydrophobic groups in order to maximize their distance and orientation for hydrogen bonding with other waters and to minimize this disruption. The result will be a structured cage, or solvation shell, around the nonpolar surface. Unlike the prior situation where the water molecules had more conformational freedom to move around and hydrogen bond with different partner water molecules, the water molecules in this solvation shell are more ordered and have more restricted mobility, and have stronger bonding and ordering around the hydrophobic solute. Hence, there is a loss in entropy when waters are in the presence of hydrophobic groups, and the overall free energy is increased. However, in situations where the drug's hydrophobic groups go from solvent exposed to buried (as in the case of taladegib's aromatic groups from **(FIG 1.1)**, the water molecules are unable to form a solvation shell around these groups, and the entropy of the system is increased.

`

Both the entropic and enthalpic terms of Gibb's Free Energy need to contribute favorably in order for high affinity binding to occur. Autodock approximates these components using the following additive equation[14]:

$$\Delta G_{bind} = \Delta G_{VDW} + \Delta G_{Electrostatics} + \Delta G_{Hb} + \Delta G_{Desolvation} + \Delta G_{Torsions} \quad \mathbf{(1.4)}$$

According to this equation, a drug's biological activity is determined by the sum of its van der Waals interactions, electrostatic interactions, hydrogen bonding interactions, desolvation enthalpy, and torsional free energy. This model is known as the Bohm scoring function[15], and it commonly used in different docking algorithms. In the following sub-sections, I will explain each of the terms in this equation, and describe how they are calculated.

b. E_{VDW}

van der Waal (VDW) forces are the result of oscillations in an atom's electron charge distribution[16]. A non-polar atom's electron charge will, overtime, be distributed uniformly around its nucleus. However, there will be instances where a nonsymmetrical distribution of electron density will be present. These temporary dipoles of one atom can induce instantaneous, opposite dipoles on neighboring atoms, resulting in a temporary attraction between the two atoms. This process will take place amongst all of the atoms in a given system. Though individual VDW interactions are weak and significant only when atoms are close, when one considers all of the VDW interactions taking place between all of the atoms in a system, they can add up and contribute significantly.

The VDW contacts made between atoms of the ligand and atoms of binding site residues is represented by a 12-6 Lennard-Jones potential[17], which provides a description between the attractive forces (due to instaneous dipoles) and repulsive forces (due to steric clashing when atoms are very close to one another):

$$E_{VDW} = \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \quad \mathbf{(1.5)}$$

where r_{ij} is the distance between a pair of non-bonded atoms *i* and *j*, and *A* and *B* are atom-specific parameters obtained from the Amber force-field. According to this equation, the van der Waals energy between atoms *i* and *j* is determined by a balance of attractive forces (the r_{ij}^{6} term) and repulsive forces (the r_{ij}^{12} term)[17]. For example, when two atoms are very close together, the r_{ij}^{6} term will be smaller than the r_{ij}^{12} term, which will result in a large value for E_{ij} and hence a dominating repulsive interaction. The Lennard-Jones potential is illustrated in **FIG 1-5A**.

c. $E_{Electrostatic}$

At physiological pH of 7.4, basic amino acids (i.e. ARG, LYS, HIS) are protonated and positively charged, while acidic amino acids (i.e. ASP, GLU) are deprotonated and negatively charged. These charged residues can make electrostatic interactions with drug atoms that have the opposite charge. **FIG 1.7** shows an example of an electrostatic interaction between the negatively-charged carboxylate of the drug Indomethacin and an Arginine located at the mouth of the binding cavity of cyclooxygenase (COX)-1[18]. The electrostatic interaction is very strong, and is estimated to contribute ~5 kcal/mol to ΔG_{bind}.[10]. The representation of electrostatic interactions takes the form of pair-wise columbic interactions[19] (illustrated in **FIG 1-5B**)

$$E_{Electrostatics} = \sum_{i,j} \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \quad \mathbf{(1.6)}$$

where q_i and q_j are the partial atomic charged on atoms *i* and *j* that are a distance r_{ij} from each other, and ϵ_0 is the vacuum permittivity constant.

`

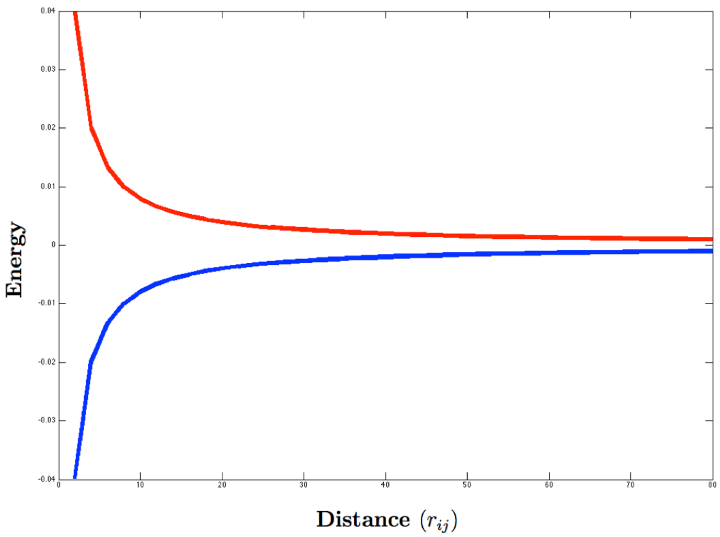


FIGURE 1-5A: Electrostatic interaction energy

Electrostatic interaction energy ($E_{elecrostatics}$) for atoms with opposite charges (blue, favorable interaction energy) and similar charges (red repulsive interactions), as a function of distance

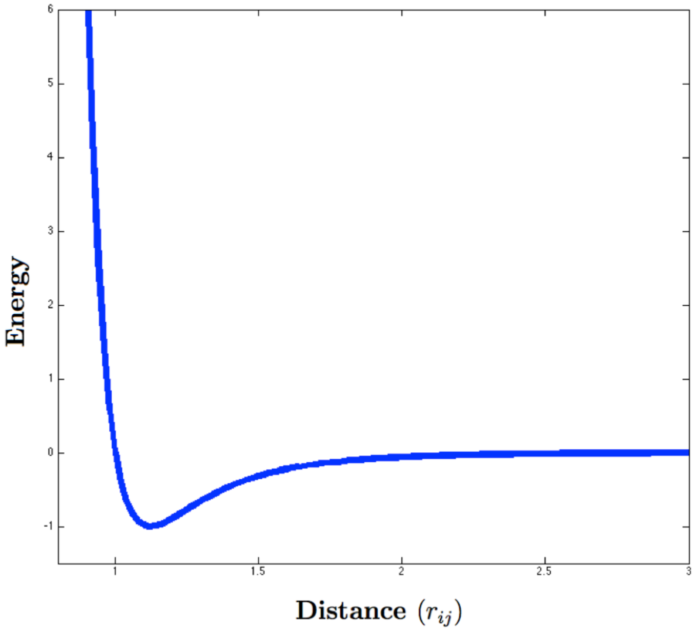


FIGURE 1-5B: van der Waals interaction energy

The Lennard-Jones potential, for the case where the well-depth ϵ_0 = - 1, and σ = 1 (the distance at which the potential is zero)

`

d. E_{Hbond}

Hydrogen bonding interactions occur when an H atom is bound to a small, electronegative atom with lone pairs (predominantly, N, O or F) that serves as a hydrogen bond donor (HBD). Since the H-F, H-O and H-N bonds are very polar, electron density is drawn away from H, conferring it with a partial positive charge. This positively charged hydrogen can interact with an atom with a partial negative charge (the hydrogen bond acceptor, or HBA), resulting in a stable interactions that depends on both the distance and the angle between the donor and the acceptors **(FIG 1-6)**. Mathematically, this relationship is represented as a 12-10 potential[17]:

$$E_{Hb} = \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) cos(\theta) \quad \mathbf{(1.7)}$$

where E_{Hb} is the hydrogen bonding energy between atom *i* and atom *j* that are a distance r_{ij} from each other; and C and D are atom-specific parameters obtained from the Amber force-field. E_{Hb} will be zero when θ = $\underline{+}90^0$ (when the hydrogen bond donor and hydrogen bond acceptor atoms are orthogonal to each other), and -1 when θ = $\underline{+}180^0$ (when the two groups are directly in front of each other).

`

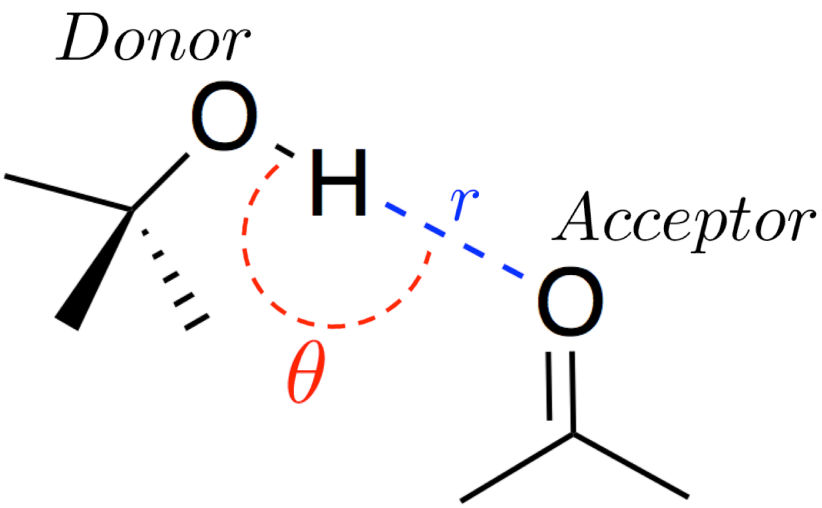


FIGURE 1-6: Hydrogen Bond interactions

Representation of the hydrogen bond distance *r* and angle *θ* between a hydrogen bond donor and a hydrogen bond acceptor.

`

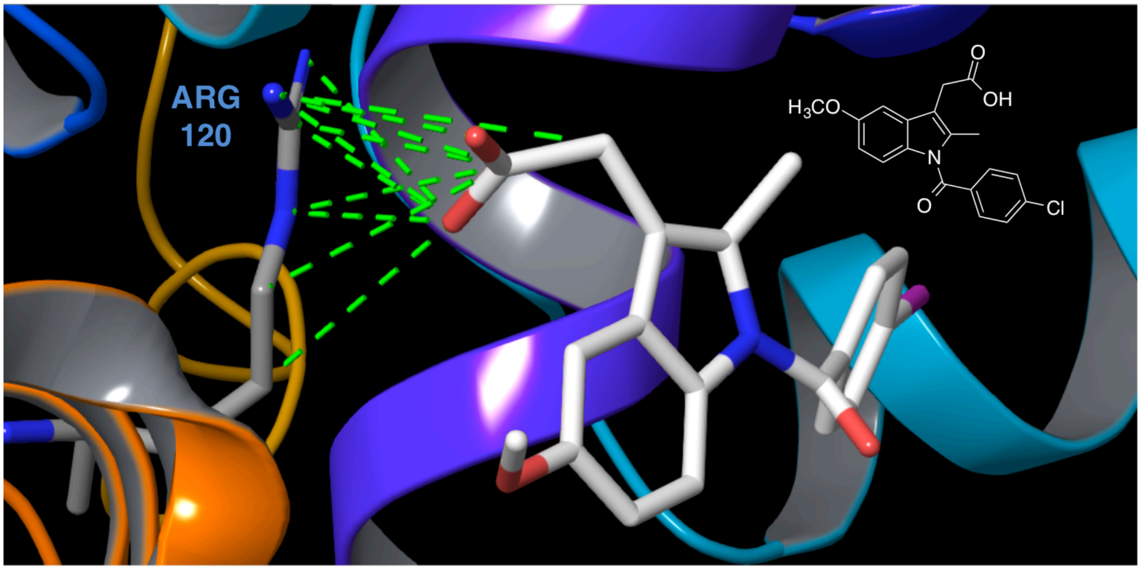


FIGURE 1-7: Indomethacin in the binding pocket of COX-1

Due to its carboxylic acid group, indomethacin exists as an anion at physiological pH 7.4. This allows it to bind and inhibit COX-1 via an electrostatic interaction with ARG 120 at the mouth of the binding site. Green lines indicate electrostatic interactions between the negatively-charged indomethacin anion (red atoms) and the positively-charged ARG 120 nitrogen atoms (blue atoms).

`

e. $E_{Desolation}$

As described in the binding free energy section, during a drug-protein binding event there is an enthalpy change associated with the desolvation of polar atoms belonging to both the protein and the ligand. Autodock's desolvation term[20] assumes that the desolvation energy of a ligand atom is related to the degree by which the atom is exposed to solvent, which in turn is defined according to the percentage of volume around the atom that is "empty" (and hence, available for water to be present in). Mathematically, desolvation is represented as:

$$E_{Desolvation} = \sum_{i,j}(S_iV_j + S_jV_i)e^{\frac{-r_{ij}^2}{2\sigma^2}} \qquad \mathbf{(1.8)}$$

where S_i is the solvation term for ligand atom *i*, S_j is the solvation term for are the protein atom *j*, V_i is the atomic fragmental volume for protein atom *i*, V_j is the atomic fragmental volume for protein atom *j*, r_{ij} is the distance between atom *i* and atom *j*, and σ is a gaussian distance constant. σ is set to 3.5 angstroms due to the fact that this distance roughly corresponds to the VDW potential for two heavy atoms[20].

The atomic solvation parameters, S_i and $\mathrm{S_j}$, are specific for each atom, and are calculated via:

$$S_i = a_i + k|q_i| \quad \mathbf{(1.9)}$$

where *qi* is the partial atomic charge of atom *i*, *k* is the charge-based atomic solvation parameter, and a_i is the atomic solvation parameter for atom *i*. According to this model, solvation is seen as a force that pushes the polar atoms into the solvent (i.e. in the direction of low occupancy), and pulls non-polar atoms in towards the interior. Hence, there will be favorable energetics for desolvating carbon atoms and unfavorable energetics for desolvating polar and charged atoms.

`

f. $E_{Torsions}$

The torsional free energy[21] term is related to conformational entropy (discussed in section **a**), which decreases when the drug binds to the protein due to the loss of torsional and rotational degrees of freedom. Mathematically, Autodock describes this torsional free energy as:

$$E_{Torsions} = N_{Torsions} \quad \mathbf{(1.10)}$$

where N_{tor} is the number of sp^3 rotatable bonds in the drug.

g. The complete free energy model

The complete binding free energy model that Autodock[14] uses is a weighted combination of the terms that were described in sections **b-f**:

$$\Delta G_{bind} = w_1 \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + w_2 \sum_{i,j} \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) + w_3 \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) cos(\theta) + \\ w_4 \sum_{i,j} (S_i V_j + S_j V_i) e^{\frac{-r_{ij}^2}{2\sigma^2}} + w_5 N_{Torsions} \quad \mathbf{(1.11)}$$

where the summations are performed over all pairs of ligand atoms *i,* and protein atoms, *j.* The free-energy of binding is a weighted-combination of the different calculated interaction energies described in **b-f**. The *w* coefficients were calculated via a linear regression analysis of a set of protein-inhibitor x-ray structures with known binding constants.

An illustration of how the E_{VDW} is calculated for a given drug within a binding site is provided in **FIG 1.8.**

`

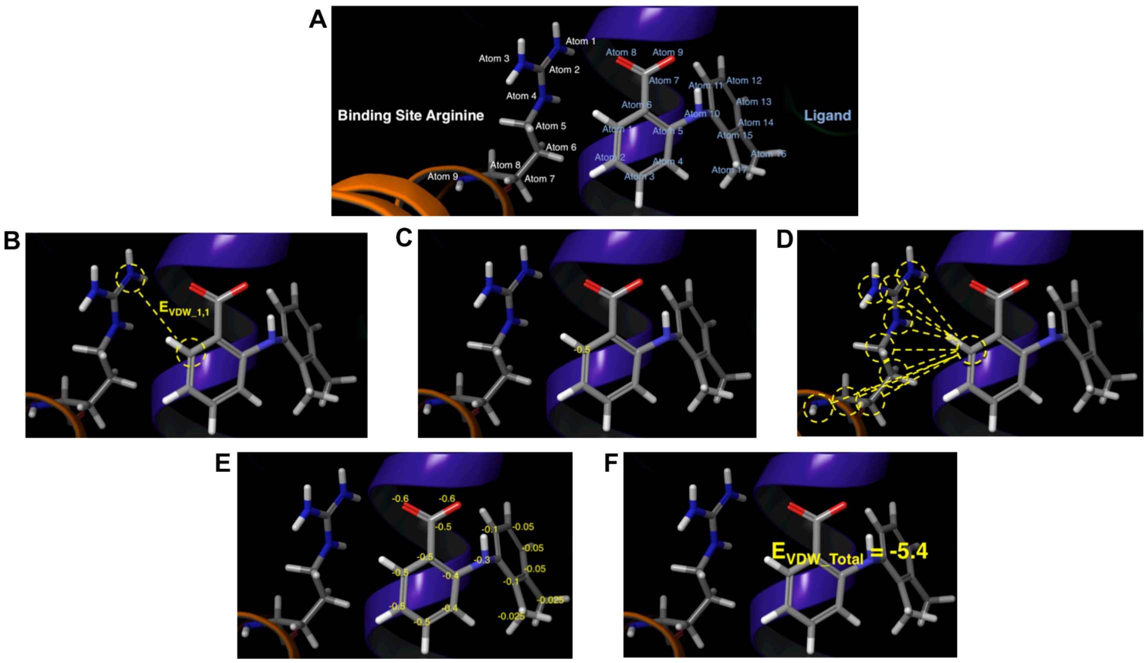


FIGURE 1-8: Example of how E_{VDW} is calculated

According to Equation 5, the E_{VDW} is calculated by summing the energy of interaction between all atoms in the compound and all of the atoms in the binding site residues. This figure shows a simplified example where there is only one binding site residue. In **A**, all of the ligand atoms and residue atoms are denoted. The E_{VDW} is calculated for the first ligand atom and the first residue atom (**B, C)**, and then between the first ligand atom and the rest of the residue atoms (**D**). Repeating this process for all of the ligand atoms results in VDW energies being assigned to each of the ligand atoms (**E**), which are then summed up to result in the total E_{VDW} for the ligand and the residue.

`

1.3. Protein and ligand representation

In the last section, we went over the different molecular interactions that Autodock uses to calculate the binding free energy. Now, we are faced with a different problem: how does Autodock represent the ligand and the protein? Proteins residues and ligands are not static, stationary molecules; rather, they are constantly in flux, existing in many different conformations. Somehow, a docking algorithm needs develop a way to deal with the conformational dynamics of both the ligand and the protein. In this section, I will explain how Autodock represents them.

a. Ligand degrees of freedom

The conformational degrees of freedom of a ligand are represented by three parameters[22] **(FIG 1-9):**

- the **position** in the Cartesian plane (i.e. the ligand can move in the *x,y* and *z* directions in the binding site)
- the **orientation** in the binding site (i.e. rotation of the ligand about its axis), defined as a quaternion (a vector having an axis of rotation, with an associated angle of rotation about this axis)
-the ligand's **torsion** angles (as defined by the number of rotatable sp^3 bonds)

Together, the total degrees of freedom of a ligand in the protein binding site is 3 + 3 + n.

A ligand structure, like a protein structure, is represented as a .pdb file, which shows the Cartesian coordinates of each atom, that in turn define the ligand's degrees of freedom. For a given drug, varying the Cartesian coordinates will result in different conformations of the drug (i.e. versions of the drug with unique set of position, orientation and torsion values), which are referred to as "poses" in Autodock. As will be described in **Section V**,

`

Autodock represents the conformational dynamics of a drug by generating several (i.e. thousands to hundreds of thousands) of poses for a drug and calculating the binding energy of each pose.

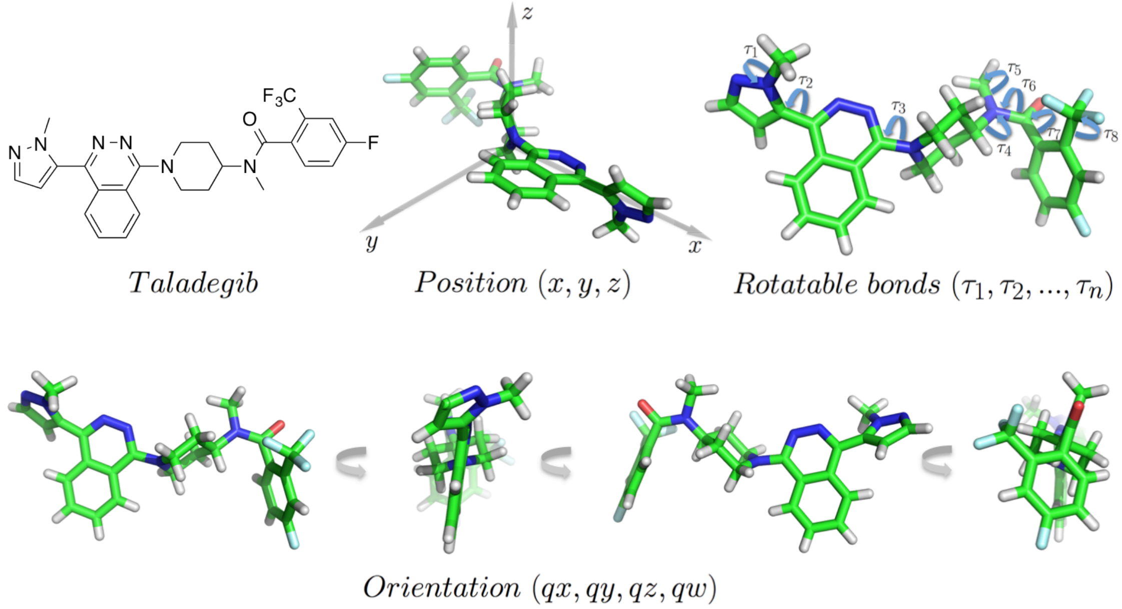

F3C
O
N
N
N-N
N
N
F
Taladegib
z
y
x
Position (x, y, z)
τ1
τ2
τ3
τ4
τ5
τ6
τ7
τ8
Rotatable bonds (τ1, τ2, ..., τn)
Orientation (qx, qy, qz, qw)

FIGURE 1-9: Drug degrees of freedom

The degrees of freedom of a drug, illustrated by the example of the SMO drug taladegib.

`

b. Grid representation of the protein

During a docking run, Autodock explores the binding site by varying the degrees of freedom of a given drug to produce different poses, and calculating the binding energy of each pose. However, a problem arises if the algorithm attempted to calculate the atomic interaction energies for every single pose. Suppose a ligand exists in one pose, and the VDW interactions are calculated between its atoms and the atoms of a binding site residue (**FIG 1-10).** Next, Autodock slightly varies the coordinates to create a second pose, and again to create a third pose. The problem here is obvious: by performing calculations for each pose of a drug, the algorithm ends up performing redundant calculations for atoms that had the same Cartesian coordinates in all three poses. This redundancy will add to the run time of the algorithm. To get around this problem, there has to be a way to get around these redundant calculations and to calculate them only once.

Suppose that, for a given atom in a drug, rather than calculating E_{VDW} for every single position that it exists in among all of the poses, one simply pre-calculated this energy at pre-determined points along the binding site, stored these values in a table, and simply looked up the E_{VDW} from this table for the different poses. The procedure, known as the grid representation method[23], has allowed Autodock and other docking algorithms to quickly calculate the interactions energies in a drug (**FIG 1-11A).** In this method, a 3D grid with equally-spaced intervals is placed over the binding site. A probe atom is then placed at each grid point, corresponding to one of the atoms in the ligand. The energy of interaction between the probe atom and the protein will be calculated, and assigned to that particular grid point. This will be done for the entire grid, and the resulting the matrix of values will then be stored in memory. This matrix, then, can be used as a lookup table to evaluate the interaction energy rapidly.

`

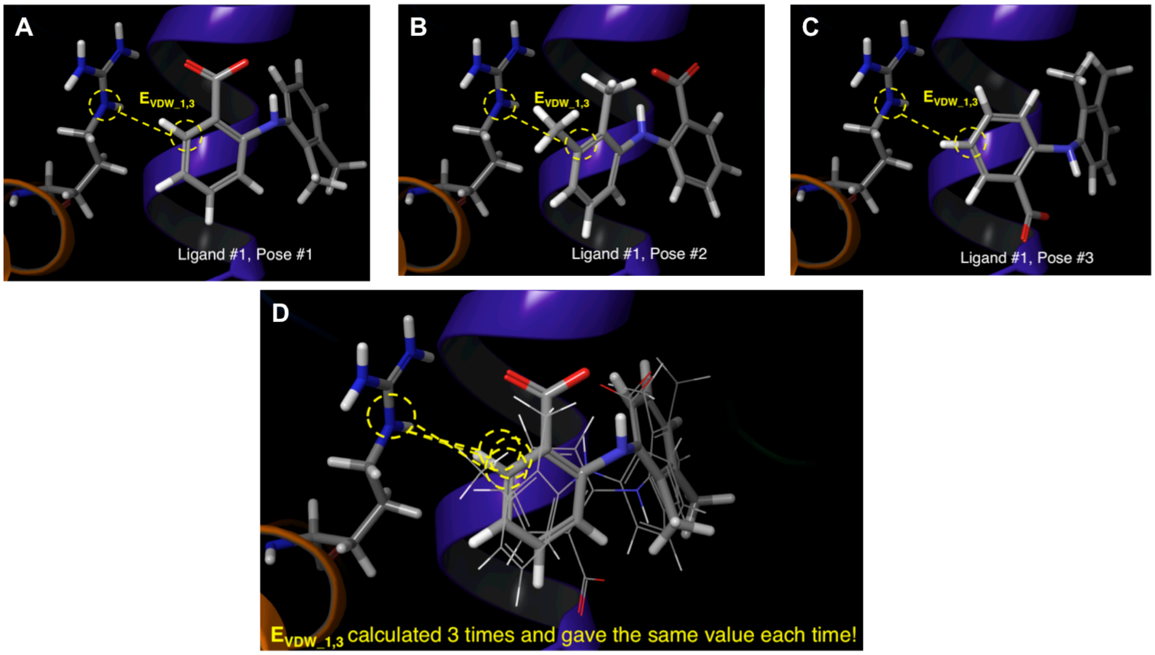


FIGURE 1-10: Possibility of redundant calculations during docking

Redundant calculations can be made if the binding energy is calculated for every atom of every drug pose that a docking algorithm tests.

`

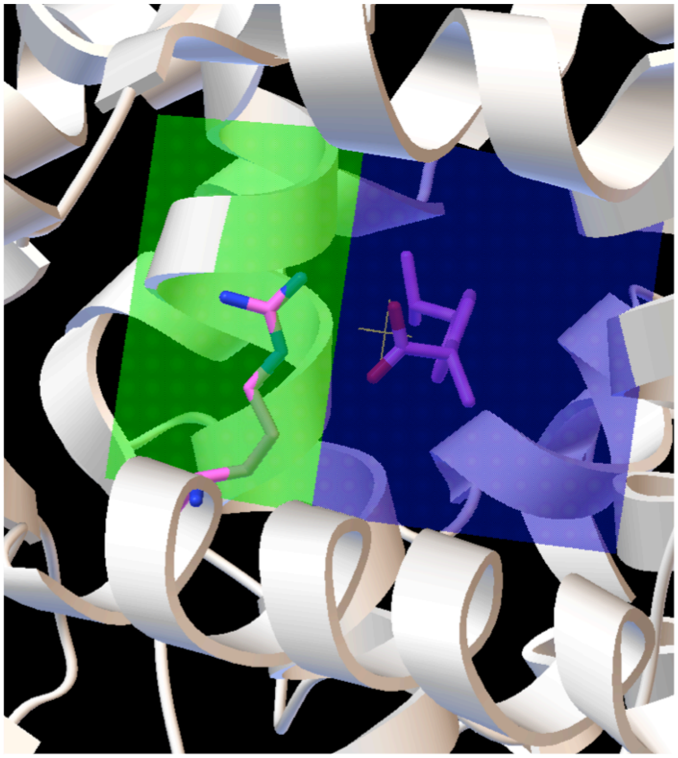
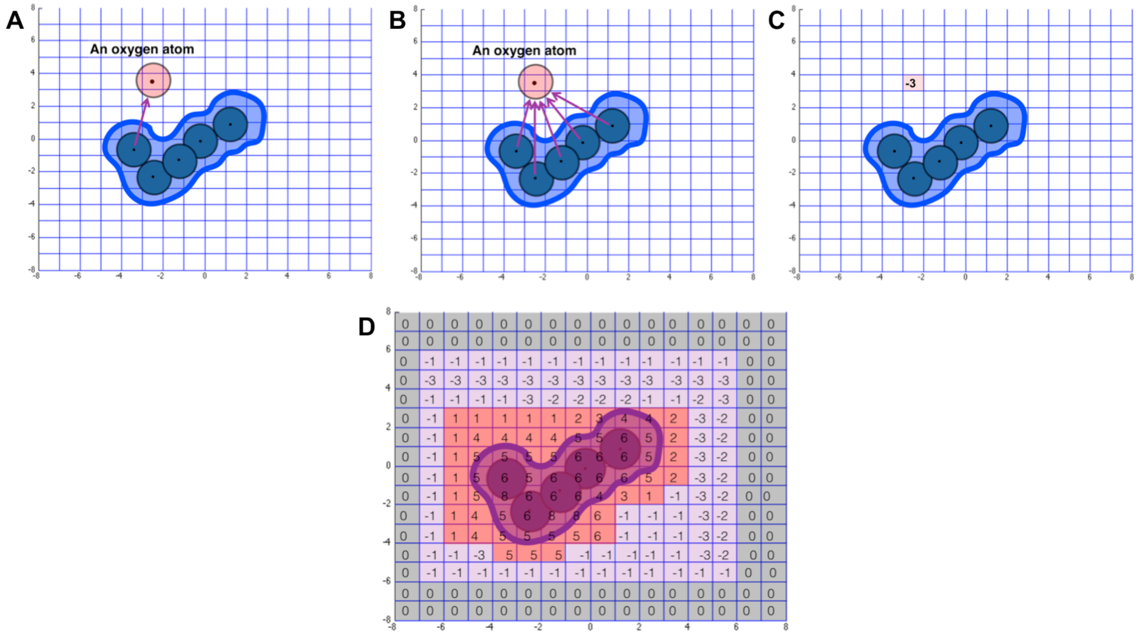


FIGURE 1-11A: Grid representation of the binding site

For Autodock and other docking algorithms, the user defines the coordinates of a grid box that will serve as the search-space for the compound poses.

`



0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0
0	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	0	0
0	-1	-1	-1	-1	-3	-2	-2	-2	-2	-1	-1	-2	-3	0	0
0	-1	1	1	1	1	1	2	3	4	4	2	-3	-2	0	0
0	-1	1	4	4	4	4	5	5	6	5	2	-3	-2	0	0
0	-1	1	5	5	5	5	6	6	6	5	2	-3	-2	0	0
0	-1	1	5	6	5	6	6	6	6	5	2	-3	-2	0	0
0	-1	1	5	8	6	6	6	4	3	1	-1	-3	-2	0	0
0	-1	1	4	5	6	8	8	6	-1	-1	-1	-3	-2	0	0
0	-1	1	4	5	5	5	5	6	-1	-1	-1	-3	-2	0	0
0	-1	-1	-3	5	5	5	-1	-1	-1	-1	-1	-3	-2	0	0
0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 1-11B Grid representation of a protein binding site: energy contours

A large 3D grid-box is placed over the binding site, and divided into small grid-boxes along evenly-spaced increments. At a given grid box, the total binding energy is measured between a probe atom (red sphere) and each of the binding site residue atoms (blue spheres) (**A-B**), resulting in a total binding energy assigned to that grid point (**C**). This process is repeated for every grid box, resulting in energy-contours (**D**) that determine the positions in the binding site where the energy is favorable (violet), unfavorable (red) or negligible (grey).

`

The use of the grid in docking can be illustrated as follows (**FIG 1-11: B-D**). Suppose there exists a hypothetical binding site that consists of only one residue, and a hypothetical ligand that consists of a sole oxygen atom. One wants to calculate the E_{VDW} between the ligand and the residue. A grid of equally-spaced points will be placed over this amino acid. At every small grid box, an oxygen probe atom is placed. The E_{VDW} is calculated between the oxygen atom and the atoms of the amino acid; these energies are then summed, and the total energy between the probe atom and the atoms of the amino acid is then assigned to that grid box. This action is repeated for every grid box, until the E_{VDW} for the entire grid is calculated.

After all of these energies are calculated, energy-contours will be present (**FIG 11C**). The violet contours represent regions where the energy is negative and favorable (i.e the negative regions in the E_{VDW} graph from **FIG 1-5a**); The areas of unfavorable energy (i.e. regions that are too close to the amino acid and which have steric clashing) are represented in red; and the areas in grey are distant regions for which the interactions energy is calculated as zero. This grid energy contour provides us with view of the different locations where the oxygen atom can placed and have favorable or unfavorable VDW interaction energies with the binding site.

Using the above look-up table, we can now calculate the interaction energy of a hypothetical molecule, such as methanol, in the binding site (**FIG 1-12**). For the methanol oxygen atom at a certain position in the grid, the algorithm will use the stored oxygen VDW table to look-up the value of oxygen's interaction energy at its current position. Different algorithms use different methods to calculate this energy. Autodock takes the closest eight points and performs tri-linear interpolation of these eight different interaction energies to produce an average score for the oxygen atom at that location.

`

This average interaction energy is then assigned to the methanol oxygen atom. This procedure is performed for the rest of the atoms of methanol, using precalculated look-up tables for carbon and hydrogen. After this is performed, the interaction energies for the oxygen atom and the carbon atom are then summed, and a total VDW interaction energy is assigned to methanol for that specific orientation in the binding site.

c. Caveats with the grid representation

For Autodock, the coordinates of the grid are input by the user based on the 3D-structure of the protein. First, the size of the pre-defined grid is important. Since the algorithm will only explore within the predefined grid, boxes that are to small can lead to false-negatives, while boxes that are too large can dramatically increase the run time. Second, under the grid representation, the binding site is conformationally rigid, and only the ligand varies. Biologically, this is not accurate, as both the drug and the protein residues are varying. As a result, if the x-ray crystal structure that one is using for docking is not appropriate for the drug that one is docking, the false-negatives will also be produced.

`

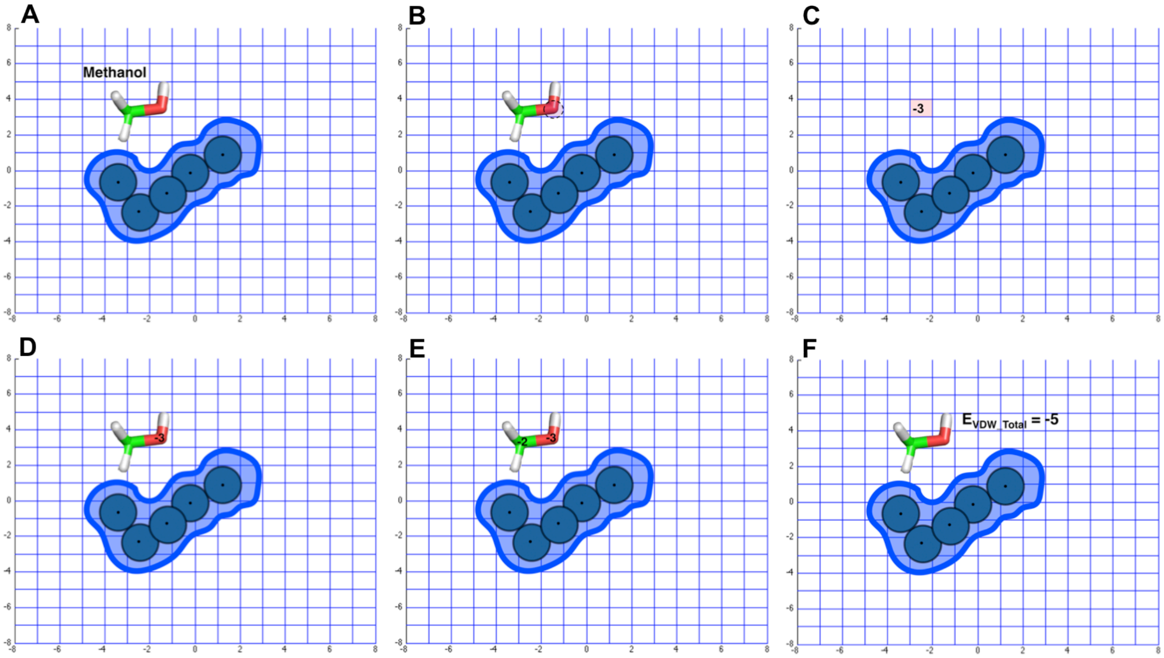


FIGURE 1-12: Example of E_{VDW} calculation for a ligand, calculated via grid method

Using the simple example of methanol (3 atom types), the binding energy for each atom is calculated via trilinear interpolation of the nearest 8 grid boxes. These energies are then summed to give the total VDW energy for the ligand at the given location in the binding site.

`

1.4. How Autodock docks drugs to proteins

So far, we have seen how Autodock quantitates the interaction energy between a ligand and a protein, and how it uses a grid-based method to represent the protein's binding site in order to save valuable computational time and memory. Now, the last step is to determine how exactly the algorithm takes a drug and determines the location within the binding site, and the conformation of the drug at this location, that corresponds to the most favorable binding free-energy for that particular ligand-protein pair. This is a difficult optimization problem, due to the numerous combinations of displacement-torsion angle-orientation values the drug can take on. The docking algorithm must find a way to sample enough of this vast space, without naively considering each and every possibility (which would make this problem computationally intractable). To achieve this goal, Autodock has stochastic search algorithms, which were first pioneered in the field of artificial intelligence. In the following sections, I will describe two such search algorithms: the simulated annealing algorithm, and the lamarkian genetic algorithm.

a. Simulated annealing Stochastic Search

Overview

A protein binding site often does not have one single region where a drug can bind to with favorable binding energy; rather, there can be several local minima "valleys" separated from each other by steep energetic "hills," and one deep global minimum. **FIG 1-13** shows a visual representation of such a landscape. In simulated annealing, a ligand will "explore" the free-energy landscape by performing a random-walk through the protein binding site grid. While exploring, it will search for locations and poses that confer favorable binding energies. After performing this exercise, the algorithm will output the best poses from its search.

`

The Algorithm In-depth

A simulated annealing[24] run is broken up into different rounds, and each round consists of different steps. During the first step of the first round, the algorithm will assign random values to each of the ligand's degrees of freedom (i.e. random displacement, torsion and rotation), resulting in a drug pose. The binding energy of this pose is calculated. For the second step, small changes will be made to each degree of freedom, resulting in a new pose. The binding energy of this second pose will be calculated. If the current pose's binding energy is more favorable, then it will be accepted, and its degrees of freedom will be randomly changed to create a new pose for the next step. If, however, the current pose's binding energy is higher, it will not automatically be rejected. Rather, it will be accepted or rejected via a probabilistic expression of acceptance, known as the Metropolis criteria:

$$P(\Delta E) = e^{\frac{-\Delta E}{k_b T}} \quad \mathbf{(1.12)}$$

where *ΔE* is the energy difference between the current pose and the previous pose; K_b is Boltzmann's constant; and *T* is a user-defined temperature. During each round, a randomly generated number between 0 and 1 is generated that serves as the "probability of acceptance," or *P(Acceptance)*. If *P(ΔE)* ≥ *P(Acceptance),* then the previous pose will be discarded and the current pose will be selected for the next step. If *P(ΔE)* ≤ *P(Acceptance),* the current pose is rejected, and the prior pose will be send into the next step. Due to the randomized nature of the acceptance probability, the algorithm will mostly accept poses with a lower energy that the previous pose, but occasionally accept poses with a higher energy. This allows the algorithm to climb over high-energy barriers that may exist between minima, and can prevent a situation where the ligand is stuck in a local minima.

`

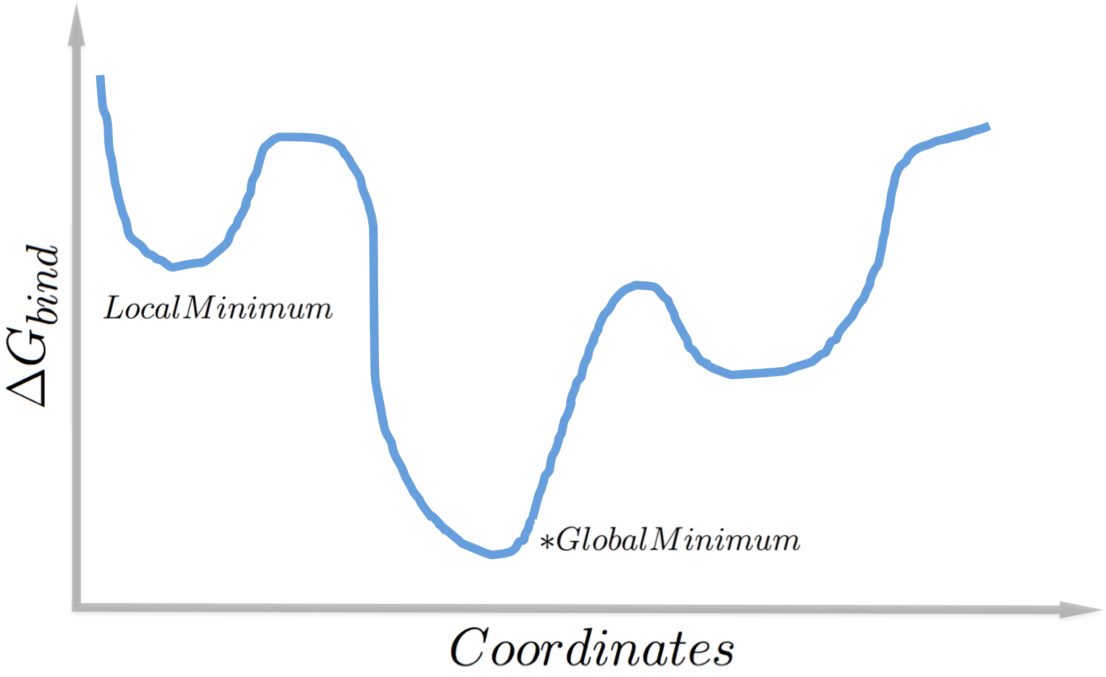


FIGURE 1-13: A hypothetical free-energy landscape of a protein binding site

Within a protein's binding site, there are several regions where the drug can bind with a favorable binding energy (local minima), but one region that where the binding energy is the most favorable (global minimum. The goal for a docking algorithm is to search the binding site for the global minimum, and to avoid becoming trapped in local minima.

`

The Temperature Parameter *T*

T controls the stringency of acceptance. When *T* is high, nearly all poses are accepted; as *T* decreases, the probability of acceptance decreases. At the beginning of each round, *T* is decreased and remains constant for the duration of that round according to

$$T_i = gT_{i-1} \ \mathbf{(1.13)}$$

where T_i is the temperature of step *i* and *g* is a constant between 0 and 1. High values of *T* during the early rounds allow the algorithm to search the binding-site grid widely along the free-energy landscape (**FIG 1-13**); in essence, it is performing a global search. As a result, poses can occupy high-energy regions in the binding site, and can climb over these "peaks" on the way to finding energetic minima "valleys". As *T* decreases during subsequent rounds, the algorithm performs more of a local search: the lower-energy states become more probable, and the algorithm will refine the ligand's degrees of freedom in the current valley **(FIG 1-14).** When the temperature is zero, the ligand will, theoretically, be in a global minimum energy pose.

The Reduction factor

In addition to *T* varying during each cycle, the amount by which the pose's degrees of freedom are randomly changed are also varied. A reduction factor between 0 and 1 is multiplied to the translation, and conformation values from the previous step; this factor can decrease at the beginning of every round. As a result, the changes in the ligand degrees of freedom will be subtler as the algorithm continues. For example, with regard to the ligand's translation coordinates, the ligand will jump to different random locations in the binding-energy landscape. During later rounds, as both *T* and the reduction factor decrease, the ligand will undergo local optimization.

`

Algorithm Parameters

Typically, each simulated annealing docking run is performed with 50 rounds, with a maximum of 3,000 steps (poses) accepted or rejected per round. A temperature of 616 cal/mol is used during the first round and is reduced linearly after each round so that it reaches 0 by the last round. For the reduction factors, the translation maximum steps are reduced linearly from 3.0 angstroms in the first round to 0.2 angstroms in the last round, and the torsion angle maximum step sizes are decreased linearly from 24 degrees to 5 degrees.

b. Genetic Algorithm and Lamarckian Genetic Algorithm

Overview of the genetic algorithm

A genetic algorithm[14] represents the ligand's degrees of freedom (translation, orientation, and conformation) as genes on a chromosome, with each gene given a numerical value:

- three genes for ligand translation: <x,y,z>
- four genes for orientation (quaternion): <qx, qy, qz, qw>
- n genes for conformation, corresponding to each of the n torsion angles present in the
  ligand: <τ_1, τ_2, τ_3,… τ_n>

Together, the genes make up the ligand's "chromosome." The ligand's pose corresponds to its "phenotype," and the pose's calculated binding energy is referred to its "fitness." Each time step of the algorithm is called a "generation." During a generation, a population of ligand poses ("individuals") will exist, and the binding energy will be calculated for the entire population. Poses that have the highest fitness (i.e. the most favorable binding energy) will "mate" with each other to produce offspring that will have a combination of genes inherited from the parents. To add additional variation, a

`

percentage of the offspring will undergo random changes in their genes ("mutations"). The algorithm will run for several generations, and as time progresses, the poses with the best binding energy will be selected at higher rates, while those that don't be become "extinct."

Algorithm In-depth

Initially, a population of ligand individuals will be generated to form the first generation; the user determines the size of the population. For each individual, the genes will be given randomly generated values:

- the three translation genes will be given a uniformly distributed random value between the minimum and maximum of the binding site grid
- the four orientation genes will be a uniformly distributed random rotation angle (θ, such that $-180^0 < \theta < +180^0$), and a random unit vector
- the torsion angle genes, if present, will be given random values between -180^0 and 180^0

During each generation, give events will occur to the individuals of the population:

- fitness evaluation
- selection
- crossing-over and mutation
- elitist selection
- local search

`

Fitness evaluation and selection

First, the binding energy of every individual in the population will be calculated. During the **selection** phase, the number of offspring that an individual will be allowed to produce, n_i, is given by

$$n_i = \frac{f_w - f_i}{f_w - <f>}, f_w \neq < f > \ \mathbf{(1.14)}$$

where f_i is the fitness of individual i; f_w is the fitness of the weakest individual in the population (i.e. the pose with the worst binding energy) over last N generations (where N is usually set to 10); and $<f>$ is average fitness of the population. According to this equation, individuals whose fitness f_i is better than the average fitness of the population $<f>$ will receive proportionally more offspring. (For individuals whose fitness is less than the population average, the algorithm will automatically allocate one offspring). The algorithm converges when $<f> = f_w$.

Crossing-over, mutation and elitism

During **crossing-over**, two individuals chosen at random will "mate" by a two-point cross-over of their genes. For example, suppose that individual 1's translation genes are given by $\mathbf{<x_1, y_1, z_1>}$ and individual 2's translation genes are given by $\mathbf{<x_2,y_2,z_2>}$. If crossing-over occurs within the translation genes, then the resulting off-spring can be represented as:

$$\mathbf{<x_1,y_1,z_1> + <x_2,y_2,z_2> \longrightarrow <x_1,y_2,z_2> + <x_2,y_1,z_1>}$$

where the red genes indicate those that were part of the crossing-over event. Additionally, randomly selected off-spring produced from the crossing-over events will undergo genetic mutation. This is performed by adding a randomly generated real number drawn from a Cauchy Distribution to the numerical value of the gene:

`

$$C(\alpha, \beta, x) = \frac{\beta}{\pi(\beta^2+(x-\alpha)^2)}, \text{ where } \alpha \geq 0, \beta > 0, -\infty < x < \infty \quad \mathbf{(1.15)}$$

where α and β are parameters that affect the mean and variance of the distribution. The purpose of this mutation is to add introduce a small amount of variability into the population: small changes to genes can be made that would not be present in the population by crossing-over, which inherits already existing gene values.

Elitism is a user-defined parameter that defines the number of individuals, n_{elite}, who will survive into the next generation. At the end of a generation, all of the individuals in the population (the newly produced off-spring, along with the individuals that did not mate) will be sorted according to their fitness binding energy, and the top n_{elite} individuals will go on to the next generation.

Lamarckian Genetic Algorithm

A Lamarckian Genetic Algorithm (LGA) differs from a regular GA that, before the mating and mutation steps, a user-defined fraction of the population will undergo a local search (via changes in its degrees of freedom) around their current location in the binding site in order to find a local minimum. After finding this minimum, the orientation, torsion and rotation values that correspond to this pose will be mapped to its genotype. From here, the algorithm will proceed in the same manner as the GA, with crossing-over, mutation, and elitism.

Output of the GA/LGA

The algorithm will run until either the maximum number of pre-defined generations is met, or until fitness convergence is reached via <*f*> = *fw*. The algorithm will continue according to the number of generations that the user has pre-defined. In the end, the genetic algorithm will output the genotype (coordinates), phenotype (3D structure) and calculated free-energy of binding for the top performing pose.

Algorithm parameters

Typically, LGA is set-up to perform a local search on a small percentage of the population (i.e. 7%). The average worse value, f_w, is taken over ten generations. One run consists of X generations, and for one ligand, 20 runs are performed (resulting in 20 different ligand poses).

c. Differences between the Simulated Annealing and Genetic Algorithm

Ligand complexity

Though simulated annealing was used in the earliest versions of Autodock, it was found that it only performs well in cases where the ligand has 8 or less rotatable bonds.

Comparing multiple poses during each round

In a simulated annealing round, a ligand poses undergoes random changes, and the new pose is compared to the old pose. In LGA, an entire population of poses are compared to each other through the via the selection (**EQUATION 1-14**) and elitism steps. In doing so, it is comparing poses that exist in different regions across the free-energy landscape. This allows it to traverse greater space in the free-energy landscape during each round, and hence allow it to discover a global minimum. With the exception of early rounds where both T and the reduction factor are high, simulated annealing compares poses that exist in the same free energy location.

Finding global minima

`

As described by **FIG 1-13**, there are multiple locations in the binding site where ligand can bind, and multiple conformations that the drug can exist in within these binding sites. Since these minima are separated by very steep hills in the free-energy landscape, a global search method that coarsely changes the ligand's values and quickly traverse these high-energy hills is better-suited for finding the global minimum that a local search algorithm. However, once global search finds the global minimum, the question now turns to the ligand conformation that fits as well as possible within this location. In this case, local search subtly tunes and refines the ligand's pose within this region, in order to come up with the best geometric fit and optimize the molecular interactions that are being made. Hence, a local search performed in on a ligand that exists in an energy minima can output its best conformation for that location.

LGA has an advantage over the Simulated Annealing because it performs both a global search (due to different individuals present in different regions of the free-energy landscape) and a local search (performed on a subset of individuals during the beginning of each generation) during each round. Simulated Annealing, on the other hand, performs global search during the beginning rounds and local search in later rounds; during the later rounds, it is unable to make large jumps across the free-energy landscape. It is for this reason that the LGA has, over the years, been the default search algorithm: it can sample distant regions of the binding pocket, and once a minimum is found, use local search to further refine the ligand.

d. Caveats with stochastic search methods

There are two main caveats with the use of stochastic search methods in docking. The first is related to protein dynamics. As previously mentioned, a ligand stabilized a subset of conformations that the receptor is sampling. An x-ray crystal structure that is used for docking is a single snap-shot of one conformation that the ligand stabilized. The big assumption with performing docking on a single x-ray is that this conformation is one of the best conformations for a drug-target interaction, and that if a ligand is theoretically able to bind favorably into the same conformation as the ligand in the crystal structure, then it represents a structure that, with high probability, will be successful. However, this is not always a correct assumption: the x-ray structure is stabilizing an very specific set of interactions with specific atoms. A new ligand, with the same scaffold as the x-ray structure but with different functional groups, may have different interactions and hence may not be stabilizing the particular conformation that is present in that x-ray snapshot.

One way around this protein dynamics problem is to vary the side-chains of the binding site in addition to varying the conformation of the ligand. This is known as flexible docking[26], and has recently been incorporated into current iterations of Autodock and other commonly used docking algorithms such as Glide[27]. The trade-off is time: the algorithm needs to explore different conformational states of both the ligand and the receptor, which can exponentially increase the running time. However, such a method will be necessary for cases where the ligand library contains ligands that are dramatically different than the ligand that the protein structure is crystallized with; or in cases where the binding site has been structurally solved without a ligand; or, in the most complicated cases, when a structure has been solved but the binding site has not been determined, and the researcher wants to search the entire structure for a new binding site or allosteric site.

The second caveat with the use of stochastic search involves water molecules[28]. Like receptor side-chains that exhibit conformational flexibility, many binding sites have buried water molecules that make hydrogen bonding interactions with the ligand that

`

serve as "bridges" with the binding site residues. These water molecules, like the ligand, are present in different locations in the binding site at different times. For interactions that depend on these water molecules, it would be necessary to take into account these interactions with water molecules.

1.5. Determining whether docking hits are "drug-like"

Thus far, I have described how Autodock can take a library of ligands and predict whether a subset of them can bind favorably to a drug target. However, once theoretical candidates emerge, another question arises: would these candidates be actually good drugs if they were given to a person? Would they be absorbed into the bloodstream? Would they be stable in the bloodstream, and make it into the target tissues they are intended for (i.e. would they be bioavailable?) Would they be toxic?

The properties of bioavailable small-molecule drugs tend to be confined in a small, narrow range of physiochemical space called the "drug-like" space[29]. A little over twenty years ago, 39% of clinical trial drugs were halted due to poor bioavailability and pharmacokinetics; hence, much attention is currently given to developing drugs that exist within this drug-like space as early in the drug discovery pipeline as possible[30]. This is relevant to not only large-scale screening libraries performed in pharmaceutical companies, but to docking screens as well: after obtaining a series of Autodock hits, the medicinal chemist needs to first determine if the hits can conceivably function as drugs before taking the time to synthesize and test them.

Though there are a large numbers of physiochemical properties that one can calculate for a candidate drug, those that are related to size, polarity, lipophilicity and the number of hydrogen bonds have been found to correlate the best with bioavailability. In this

`

section, I will review criteria that take these properties into account: **Lipinski's Rule of 5**, and **Ligand Efficiency**. First, I will discuss an important property that plays a role in both of these metrics: lipophilicity.

a. Lipophilicity, logP and logD

Lipophilicity of a drug refers to its property of being dissolved in hydrophobic solvents such hexane and toluene[31]. In a sense, every drug candidate that targets an intracellular process needs to be lipophilic to some degree in order to pass through the cell membrane. For example, though indomethacin's carboxylic acid group exists as a carboxylate at physiological pH[18] **(FIG 1-6);** however, it also possesses a hydrophobic aromatic ring that confers lipophilicity and allows it to cross the lipid membrane on its way to the COX binding site. Without this hydrophobic group, the anionic group would block it from entering cells. The opposite problem arises for drugs that are too lipophilic. Such compounds have been associated with increased toxicity and promiscuity[31]; increased liver metabolism and plasma-protein binding (limiting the chances that the drug can make it out of the systemic circulation to its target tissues); and decreased water-solubility, a limiting factor for GI absorption. Hence, lipophilicity is a property that needs to be strongly considered when optimizing a lead compound.

The common method of measuring lipophilicity is through a calculation of logP[32], the solubility of a compound in the organic solvent 1-octanol (which simulates the cell membrane) relative to water:

$$log\ p_{(octanol/water)} = log\left(\frac{[solute]_{octanol}}{[solute]_{water}}\right) \quad \mathbf{(1.16)}$$

`

Experimentally, *P* is calculated by placing a compound in a separatory funnel with 1-Octanol and water, mixing the two components, and determining the concentration of the compound in each layer via HPLC. If the compound is more soluble in water than in 1-octanol, *P<1* and *logP* will be negative. Hence, the larger the value of *log P*, the more lipophilic the compound.

Since many drugs have ionizable groups that cause them to be charged at physiological pH, it would be inappropriate to measure the logP at pH 7.4 due to the fact that the charged form would not enter the 1-Octanol layer during the mixing experiment. For these cases, the distribution coefficient, logD[33], describes the logP of the compound where the aqueous phase is adjusted to a specific pH:

$$log\ D_{(octanol/water)} = log\left(\frac{[solute]_{octanol}}{[solute]_{water(ionized)}+[solute]_{water(non-ionized)}}\right) \quad \mathbf{(1.17)}$$

b. Lipinski's Rule of Five

Lipinski's Rule of 5[34] focuses on the drug-like space related to bioavailability: will a compound likely have absorption problems because of poor solubility and/or permeability? For his analysis, Lipinski and colleagues analyzed the physiochemical properties of ~2,000 drugs and clinical trial candidates that were orally active. After this analysis, they concluded that 90% of these compounds four properties in common that could be used to predict if they would be membrane permeable and easily absorbed in the body:

- molecular weight < 500 daltons
- log P $\leq$ 5 (octanol-water partition coefficient)

`

- H-bond donors ≤ 5 (i.e. the total number of nitrogen-hydrogen and oxygen-hydrogen bonds)
- H-bond acceptors < 10 (all nitrogen and oxygen atoms)

These properties emphasize the synthesis of compounds that are not too large, floppy and polar, but that are also not too lipophilic. Of note, these rules only apply to drugs that are meant to be bioavailable that undergo passive diffusion through cell membranes; compounds that are actively transported, for example, are exempt from these criteria.

In 2002, Veber et al.[35] made two additions to the Lipinski's Rules in 2002. In their work, they analyzed rat bioavailability data of over 1100 clinical candidates. From there analysis, they found that the best predictors of bioavailability, independent of molecular weight, are:

- less than 10 rotatable bonds (due to reduced molecular flexibility of the ligands)
- polar surface area < 140 Å^{2} (defined as the area on the surface of the drug that is contributed by polar atoms, calculate via an atom-based method)

One of the caveats with using the Lipinski/Veber rules is that they have been used as strict deterministic rules instead of guidelines in many drug discovery programs[36]. This has resulted in many promising drug discovery candidates to be shelved. Strictly interpreting these rules has questionable value. Undesirable drugs can barely pass all of these cut-offs and be considered success with respect to physiochemical properties yet fail in actual clinical testing, while a desirable drug can barely miss just one of the cut-offs and not even make it into the clinical testing phase. It is estimated that 16% of orally available, approved drugs violate at least one of these criteria, and 6% violate two or more[16]. Notably, one of the all-time biggest drug sellers, atorvastatin, fails the Lipinski's

`

Rules, and would never have been advanced into the clinical if these guidelines were in place during its development.

A second caveat with the Lipinski's Rules is that they do not apply to natural products or natural product derivatives[38]. In the original Lipinski paper, it was notes that these drugs are often bioavailable despite violating the Rule of Five. Evolutionary, this makes sense: these compounds were carefully developed over millions of years of evolutionary history for specific biological purposes. Many of these compounds bind to extracellular membrane receptors and cell transporters. Despite violating the rule of 5, natural products make up over 30% of all approved small-molecule drugs; within oncology drugs, this percentage raises to 47%.

A last caveat that I will raise with the Rule of 5 is that it was obtained for drugs that inhibit targets that are different often different from the targets that we are going after now. It has been argued that most of the "low-hanging fruit" drug discovery has been taken up; what we are left with are very difficult drug targets (i.e. large transcription factor complexes, protein-protein interactions) that do not have traditional binding pocket grooves that a small-molecule inhibitor can neatly fit into. Hence, blindly obeying the Rule of 5 may cause drug discovery programs to abandon promising, unconventional approaches. Many drugs under development have a molecular weight over 500 daltons, have been designed to form large macrocycle complexes.

c. Ligand Efficiency

Over the past twenty years, the mean molecular weight and lipophilicity of experimental compounds have risen over 100 Daltons, but these values for newly-approved drugs have not. This trend is concerning, given that (as discussed in section a) these two properties have been found as having a higher probability of failure in clinical trials. One reason that can account for this trend is that large, highly lipophilic compounds are more likely to be detected during HTS screening due to the having the ability to form more interactions with the target. Large-scale screens are identifying them as hits, leading to their subsequent SAR optimization and experimental testing. This same principal holds for molecular docking screening as well: as we have seen in **EQUATION 1-5**, VDW interactions are calculated by summing across all of the atoms in a ligand. A large, bulky, lipophilic molecule will then have a higher chance of having a favorable VDW energy and hence a favorable free energy of binding, just in lieu of its size.

To overcome screening biases towards large lipophilic compounds, two metrics have been proposed. The first metric argues that during the clinical development of a molecule, one needs to look at not only the binding energy of the ligand, but instead the binding energy relative to the number of atoms (ligand efficiency, or LE) or the lipophilicity (LLE)[39]. Ligand efficiency metrics quantify how effectively the ligand is binding to the target. It is argued that this combined metric, rather than just potency, is what should be optimized rather than just potency alone.

The simpest LE metric is to take the molecule's calculated free binding energy and dividing it by the number of heavy atoms:

$$\mathrm{LE} = -\frac{\Delta G_{bind}}{HA} \quad \mathbf{(1.18)}$$

One caveat of this approach is that is it treats different atom types (i.e. carbon, nitrogen, oxygen, sulfur, halogens) in the same manner, despite the fact that there are marked differences in size and polarity between them. Furthermore, this equation assumes that all of the atoms in the ligand are involved in binding the target. This is often not the case

`

for halogens such as Cl and F, which are often placed on benzene rings to aid increase logD and permeability.

The Lipophilic Ligand Efficiency Index (LE_{lipo}) is given by:

$$LE_{lipo} = log\left(\frac{-\Delta G}{P}\right) \ \mathbf{(1.19)}$$

Data has shown that using LE_{lipo} during the process of drug optimization has led to compounds with increases binding affinity without increased lipophilicity.

d. Conclusion

Thus far, we have seen how to design a library of ligands, calculate their theoretical binding energy, and, once candidates have been revealed, use guidelines to determine if the candidates have physiochemical properties that would confer a high failure rate. Now, I am going to discuss how these principles were used in the computational design and synthesize of brain-penetrant inflammation inhibitors for Alzheimer's Disease.

`

CHAPTER II: APPLICATIONS OF DOCKING TO ALZHEIMER'S DISEASE DRUG TARGETS

2.1. BACKGROUND

a. Neuroinflammation in Alzheimer's Disease

In normal settings, microglial cells have a protective function in the brain: they surround neurons and involved in maintenance or tissue homeostasis; synaptic remodeling; secretion of neurotropic factors, and the scavenging for infectious pathogens[40]. However, during certain adverse situations (i.e. systemic inflammation, brain trauma, and presence of amyloid plaques), microglia become activated and release pro-inflammatory molecules (eicosanoids) in an attempt to clear damaged cells and amyloid plaques. While this may be beneficial in the short-term for clearing away damaging molecules in the brain, in the long-term this response can become uncontrolled. The sustained neuroinflammation from microglia themselves result in neuronal damage and, subsequently, the formation of amyloid-beta (Aβ) peptide-containing senile plaques that are associated with Alzheimer's Disease (AD).

The eicosanoids released by microglia are produced by the membrane protein cyclooxygenase-1 (COX-1) and the cytosolic protein 5-lipoxygenase (5-LOX) (**FIG 2-1**). In AD patients, COX-1 (and COX-derived PGE2 and TXA2 metabolites) is increased in the brain relative to age-matched, non-AD brains. PGE2 is also elevated in the cerebrospinal fluid of AD patients[41]. Experimental work from the Lee and Trojanowski labs has provided additional evidence in support of the role of inflammation in AD[42,43]. First, knockout of the PGE2-binding receptors (EP1, EP2, EP3 or EP4) in transgenic mice expressing the human amyloid precursor protein (from which Aβ is derived) results in a reduction of senile plaque formation. Second, genetic or pharmacologic modulation of 5-LOX activity results in a reduction of senile plaque burden in amyloid precursor protein transgenic mouse models. Third, activation of the thromboxane receptor by TXA2, the EP1 and EP3 receptors by PGE2, and the CysLT1 receptor by LTD4, all result in increased amyloid precursor protein expression and amyloid beta release. Collectively, these clinical findings and experimental results suggest that therapeutic strategies to reduce the levels of inflammatory eicosanoids in the brain via direct inhibition of COX and 5-LOX can lead to a decrease in Aβ release and senile plaque formation.

Due to the possible role of neuroinflammation in AD onset and progression, our goal was to develop brain penetrant compounds that could inhibit both COX-1 and 5-LOX. Currently, dual 5-LOX/COX inhibitors are being investigated for peripheral inflammation. Notably, the compound licofelone is in Phase III clinical trials for osteoarthritis[44] (**FIG 1-2**). Licofelone was found to have similar efficacy as other COX inhibitors, but improved side-effect profile. The drawback with attempting to use licofelone as an AD treatment, however, is that it has limited blood-brain barrier permeability due to its carboxylic acid group. At physiological pH, this group exists as a carboxylate and cannot cross the blood-brain barrier. The carboxylate is responsible for the mechanism of action of not only licofelone, but all of the FDA-approved COX-1 inhibitors[45]: it makes a strong electrostatic with ARG 120 in the mouth of the COX-1 binding cavity, locking it into an inactive state and preventing it from metabolizing Arachidonic Acid (**FIG 1-6**). Hence, our goal was to replace this carboxylic acid with a functional group that could still maintain COX/LOX inhibition (presumably through hydrogen bonding interactions rather than electrostatic interactions), but would be lipophilic enough to confer brain penetrance. To

`

move towards this goal, we made use of the concept in medicinal chemistry known as bioisosterism.

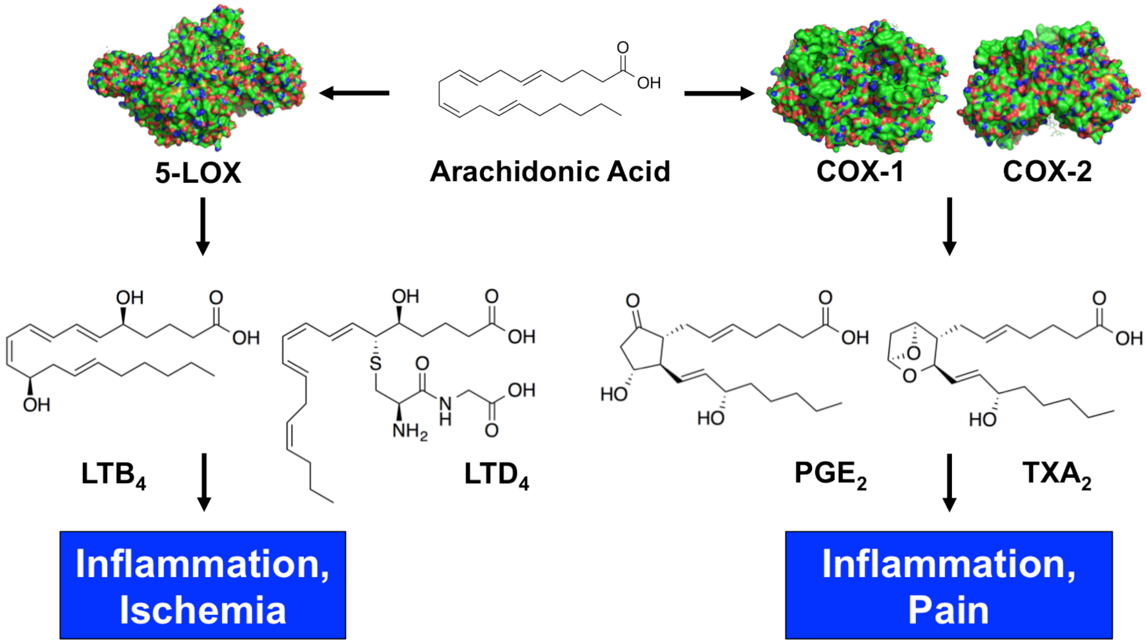


FIGURE 2-1: Overview of the COX and LOX pathways

The 5-lipoxygenase (5-LOX) leads to the production of leukotrienes (i.e. LTB$_4$ and LTD$_4$) that are involved in the recruitment of innnate immune system cells that release pro-inflammatory cytokines, resulting in inflammation and ischemia (via vasoconstriction). The cyclooxygenase (COX) pathways result in the production of prostaglandins (i.e. PGE$_2$ and TXA$_2$), which also result in inflammation in addition to the sensation of pain. Both pathways have been found to be hyperactive in Alzheimer's Disease patients, leading to the hypothesis that inhibiting them cause decrease neuroinflammation and decrease Alzheimer's Disease progression.

`

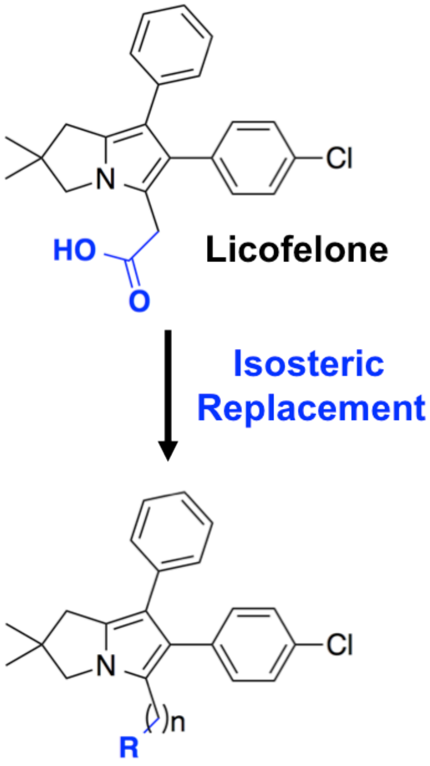


FIGURE 2-2: A dual COX/LOX inhibitor, Licofelone

Though, systemically, licofelone has been found to be an effective dual COX/LOX inhibitor, its carboxylic acid group is negatively charged at physiological pH. This limits its ability to cross the blood-brain barrier and inhibit COX and LOX in the brain. Our goal is to perform isosteric replacement of this carboxylic acid group, in order to synthesize a brain-penetrant licofelone analogue.

`

b. Bioisosterism in medicinal chemistry

Bioisosteres are functional groups that have similar physical and chemical properties and which produce similar biological effects[46]. In medicinal chemistry, bioisosteres are commonly used to modify the biological activity, pharmokinetic profile (i.e. absorption of the compound, transport through the bloodstream, excretion), and toxicity of a compound during SAR optimization. Bioisosteric replacement can have profound effects on a drug's structural parameters (i.e. size, shape and electronic distribution of charge) and pharmacokinetic parameters (i.e. water solubility, pKa, polarizability). Historically, there are two groups of bioisosteres: classical and non-classical. Classical isosteres (**FIG 2-3**) are groups that have the same number of valence electrons (they may, however, have different numbers of atoms). Nonclassical bioisosteres often produce the same biological effects, but do not have the same number of valance electrons (**FIG 2-4**). An example where bioisosterism can dramatically change the activity or potency of a compound is shown for the example of an angiotensin II receptor antagonist (ARB)[47] **(FIG 2-5)**. When the lead compound has a carboxylic acid group, the Ic_{50} is 275 nM. However, isosteric replacement with tetra-fluoryl sulfonamide, squaric acid, and tetrazole groups resulted in gradually increasing potency.

2.2. DESIGNING A CARBOXYLIC ACID ISOSTERE LIBRARY

a. Carboxylic Acid bioisosteres

The carboxylic acid group is an important drug pharmacopore that is present is over 450 marketed drugs[48.] Its ubiquity is due to its ability to make electrostatic interactions and hydrogen bonding interactions with protein binding site residues, and due to fact its solubility-conferring effects on the compound. However, drawbacks with this group include metabolic instability, toxicity via glucuronidation (which can lead to the production of metabolites that covalently bond to other proteins), and, as mentioned,

`

limited permeability across biological membranes. Our goal of identifying carboxylic acid isosteres that could overcome these barriers and confer brain penetrance led our group to synthesize 36 carboxylic acid isosteres spanning different functional group classes[48] (**FIG 2-6A-B**). The 3-phenylpripionic acid group (**FIG 2-6C**) was chosen as the scaffold due to its minimal molecular weight (for detection and avoidance of volatility); UV activity (allowing for detection via thin-layer chromatography and high-pressure liquid chromatography (HPLC) purification), and the presence of a spacer between the benzyl group and the functional group (to prevent the aromatic ring from interfering with physiochemical measurement of the isostere). The organic syntheses of isosteres that I synthesized (acylsulfonamide, hydroxamic acid, 3-Isoxazolol, and Sulfonylurea) are shown in **FIG 2-7**, and described fully in **APPENDIX A-D**. I was able to crystallize the sulfonylurea compound, and submit it for x-ray crystal structure determination (**APPENDIX A**).

After synthesis of the library was completed, key physiochemical properties relevant to medicinal chemistry were experimentally determined (**TABLE 1**): lipophilicity (via $\log D_{7.4}$ calculation), acidity (through pK_a determination), permeability (using a Parallel Artificial Membrane Permeability Assay (PAMPA) to determine the permeability coefficient, P_{app}) and plasma protein binding. For brain penetrance, the most important physiochemical properties are $\log D_{7.4}$, $\log P_{app}$, and pK_a. These values are plotted together in **FIG 2-7** for the library. In general, more lipophilic and less acidic compounds exhibit higher rates of membrane permeability. Isosteres in the right-upper quadrant (**FIG 2-8**) represent those that are more permeable (lower $\log P_{app)}$ and more lipophilic (higher $\log D_{7.4}$) than the reference carboxylic acid (of 3-phenylpropionic acid). These data suggested that isosteres in this quadrant, compared to the others that could be made, can possibly

`

confer brain penetrance when substituted onto an NSAID scaffold. These isosteres are being weighted heavily for our ongoing synthesis efforts.

`

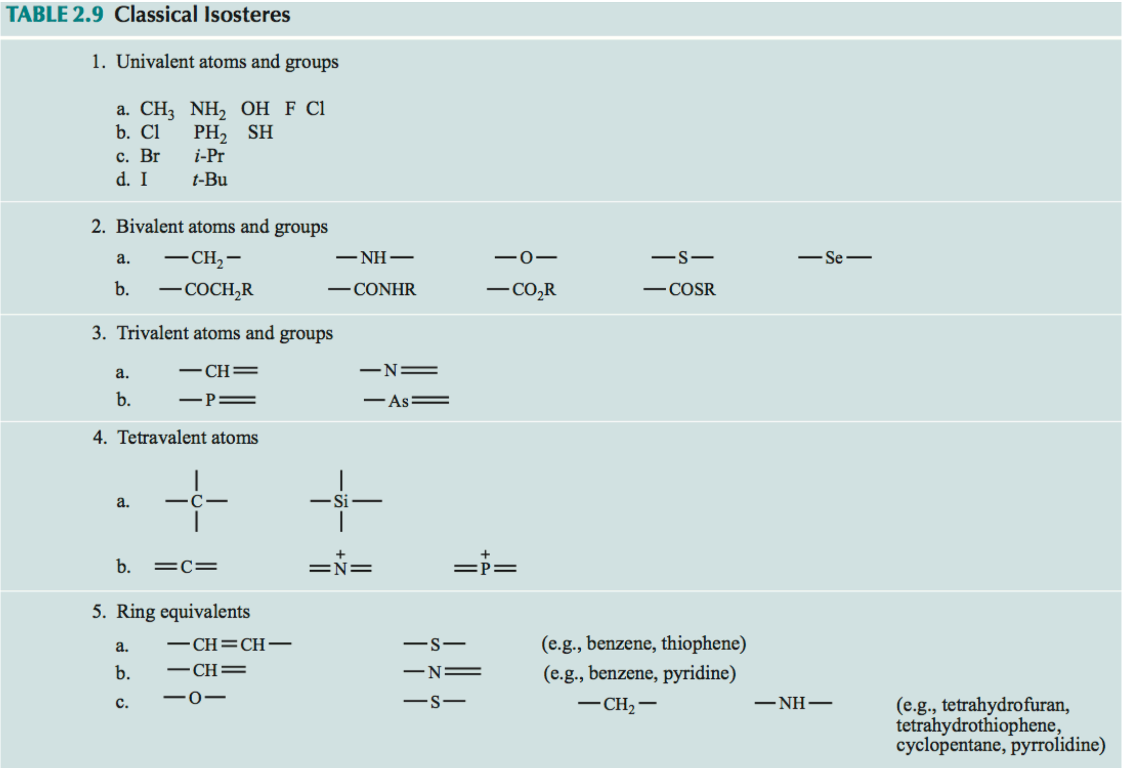


TABLE 2.9 Classical Isosteres

1. Univalent atoms and groups					
a.	CH_3	NH_2	OH	F	Cl
b.	Cl	PH_2	SH		
c.	Br	*i*-Pr			
d.	I	*t*-Bu			
2. Bivalent atoms and groups					
a.	$-CH_2-$	$-NH-$	$-O-$	$-S-$	$-Se-$
b.	$-COCH_2R$	$-CONHR$	$-CO_2R$	$-COSR$	
3. Trivalent atoms and groups					
a.	$-CH=$	$-N=$			
b.	$-P=$	$-As=$			
4. Tetravalent atoms					
a.	$-\overset{\vert}{\underset{\vert}{C}}-$	$-\overset{\vert}{\underset{\vert}{Si}}-$			
b.	$=C=$	$=\overset{+}{N}=$	$=\overset{+}{P}=$		
5. Ring equivalents					
a.	$-CH=CH-$	$-S-$	(e.g., benzene, thiophene)		
b.	$-CH=$	$-N=$	(e.g., benzene, pyridine)		
c.	$-O-$	$-S-$	$-CH_2-$	$-NH-$	(e.g., tetrahydrofuran, tetrahydrothiophene, cyclopentane, pyrrolidine)

FIGURE 2-3: Examples of classical isosteres

Table from:

Silverman and Holladay. "The Organic Chemistry of Drug Design and Drug Action (Third Edition)." Chapter 2: Lead Discovery and Lead Modification. Elsevier (2014).

`

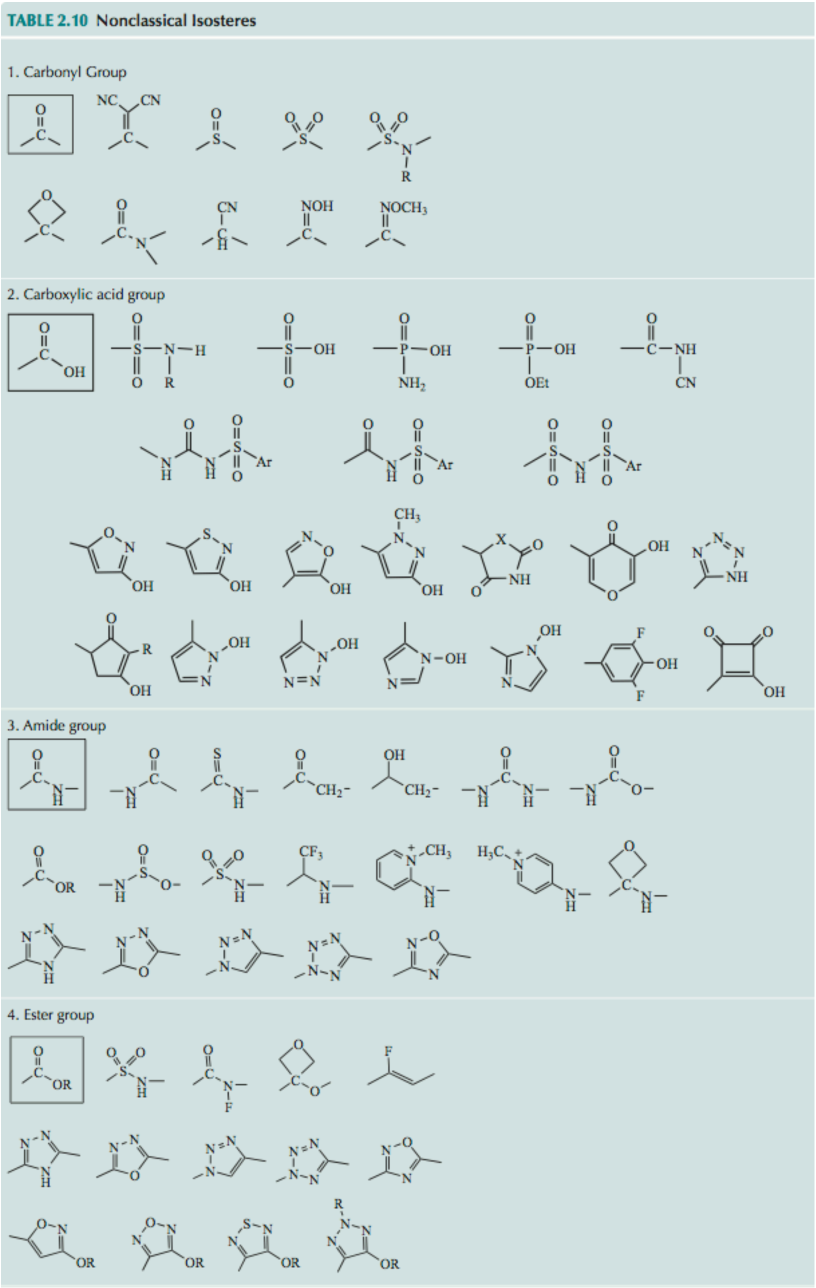


TABLE 2.10 Nonclassical Isosteres

1. Carbonyl Group

2. Carboxylic acid group

3. Amide group

4. Ester group

FIGURE 2-4: Examples of non-classical bioisosteres

Table from:

Silverman and Holladay. "The Organic Chemistry of Drug Design and Drug Action (Third Edition)." Chapter 2: Lead Discovery and Lead Modification. Elsevier (2014).

`

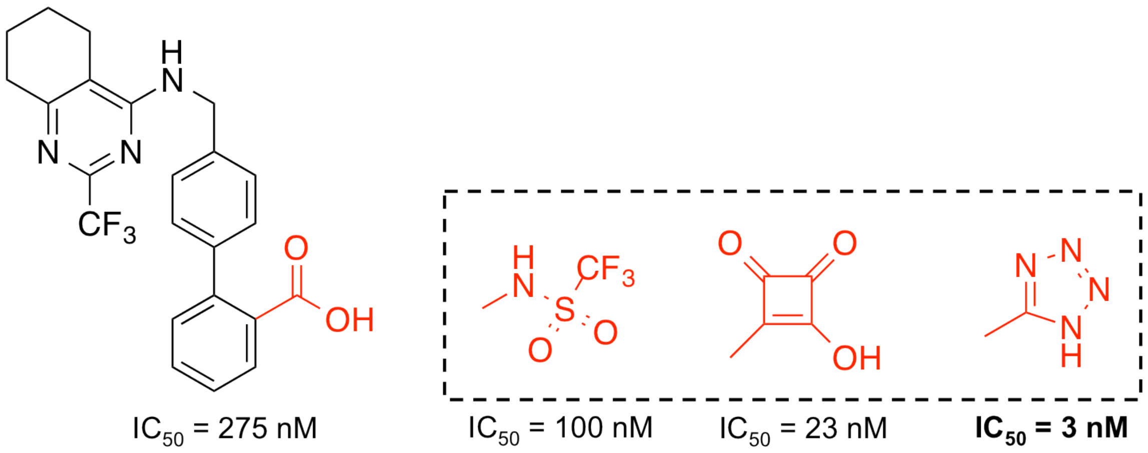


FIGURE 2-5: SAR optimization of the hypertensive drug losartan

Beginning with the carboxylic acid scaffold, substitutions to different carboxylic acid isosteres led increasingly better potency. The tetrazole group ended up being the isostere that conferred the biggest increase in potency, and was the compound that subsequently cleared clinical trials and acquired FDA-approval.

`

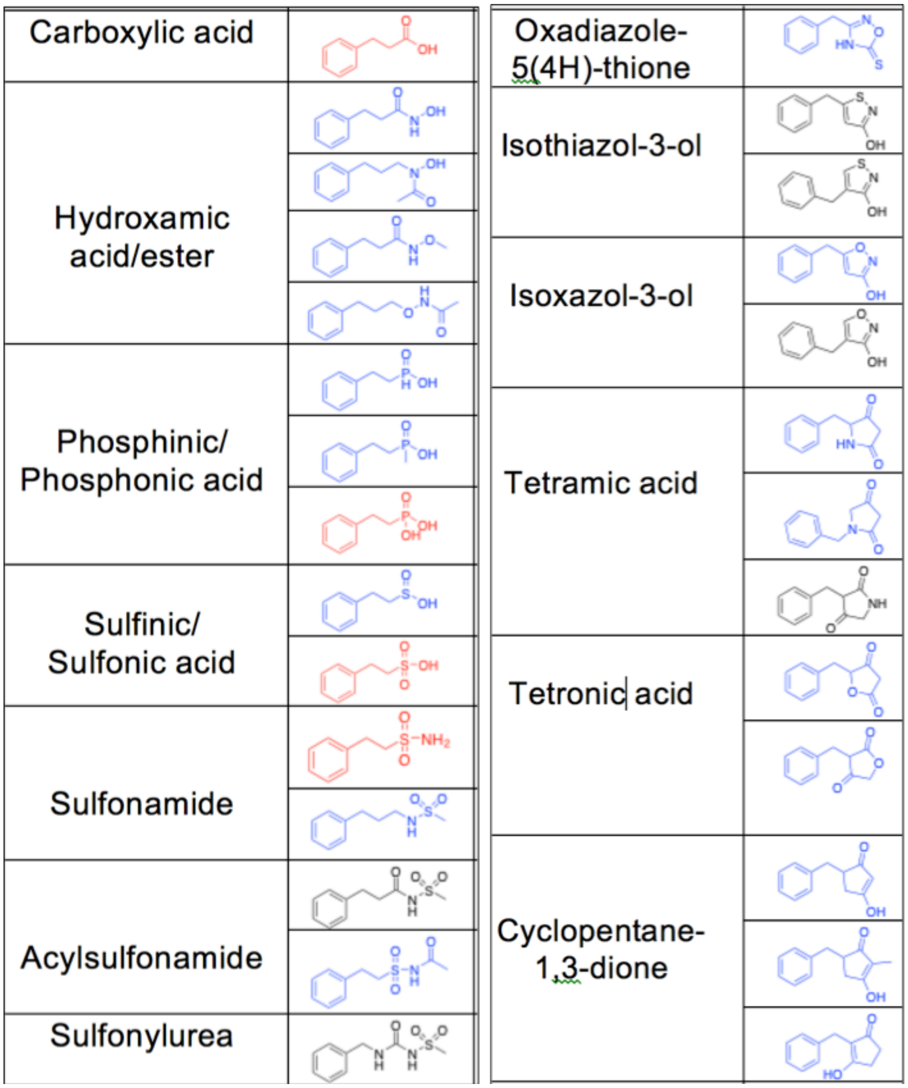
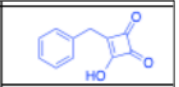
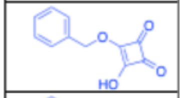
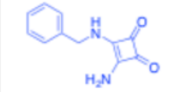
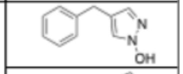
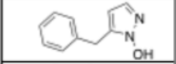
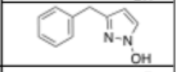
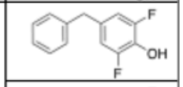
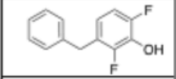
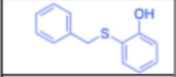
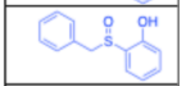
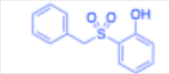


FIGURE 2-6A: Subset of the 36 carboxylic acid isosteres that were synthesized

`

Squaric acid and derivatives	
	
	
Pyrazol-1-ol	
	
	
Substituted phenol	
	
	
	
	

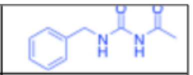
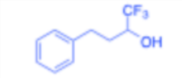
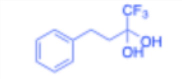
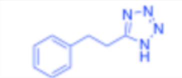
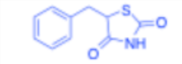
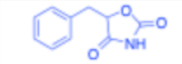
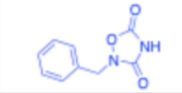
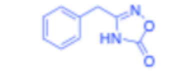
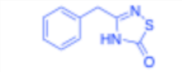
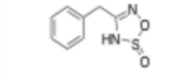
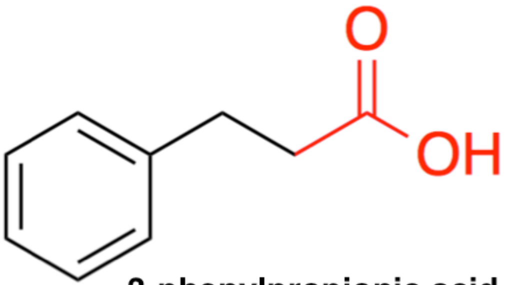
Acylurea	
Trifluoromethyl alcohol	
Trifluoromethyl ketone	
Tetrazole	
Thiazolidine-dione	
Oxazolidine-dione	
Oxadiazolidine-3,5-dione	
Oxadiazol-5(4H)-one	
Thiadiazol-5(4H)-one	
Oxathiadiazole 2-oxide	

FIGURE 2-6b: Second subset of the 36 carboxylic acid isosteres that were synthesized

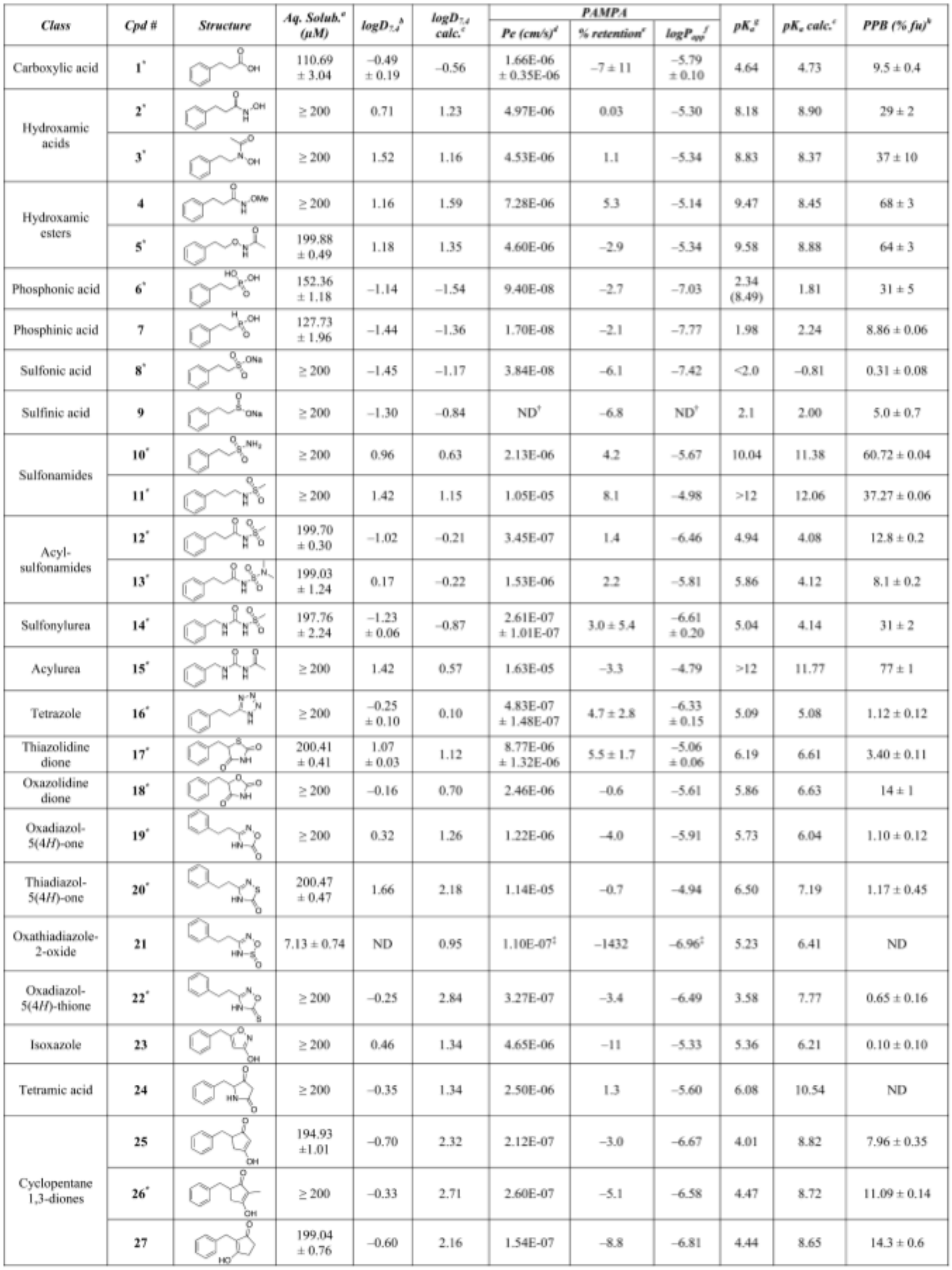


3-phenylpropionic acid

FIGURE 2-6C: The 3-phenylpropionic acid scaffold

`

The 3-phenylpropionic acid scaffold was chosen because of its low molecular weight, UV activity, and spacer between the aromatic group and the functional group.



Class	*Cpd #*	*Structure*	*Aq. Solub.*^{a} *(μM)*	*logD*$_{7.4}^{b}$	*logD*$_{7.4}$ *calc.*^{c}	*PAMPA*			*pK*$_a^{g}$	*pK*$_a$ *calc.*^{c}	*PPB (% fu)*^{h}
						Pe (cm/s)^{d}	*% retention*^{e}	*logP*$_{app}^{f}$			
Carboxylic acid	$\mathbf{1}^*$		110.69 ± 3.04	–0.49 ± 0.19	–0.56	1.66E-06 ± 0.35E-06	–7 ± 11	–5.79 ± 0.10	4.64	4.73	9.5 ± 0.4
Hydroxamic acids	$\mathbf{2}^*$		≥ 200	0.71	1.23	4.97E-06	0.03	–5.30	8.18	8.90	29 ± 2
	$\mathbf{3}^*$		≥ 200	1.52	1.16	4.53E-06	1.1	–5.34	8.83	8.37	37 ± 10
Hydroxamic esters	**4**		≥ 200	1.16	1.59	7.28E-06	5.3	–5.14	9.47	8.45	68 ± 3
	$\mathbf{5}^*$		199.88 ± 0.49	1.18	1.35	4.60E-06	–2.9	–5.34	9.58	8.88	64 ± 3
Phosphonic acid	$\mathbf{6}^*$		152.36 ± 1.18	–1.14	–1.54	9.40E-08	–2.7	–7.03	2.34 (8.49)	1.81	31 ± 5
Phosphinic acid	**7**		127.73 ± 1.96	–1.44	–1.36	1.70E-08	–2.1	–7.77	1.98	2.24	8.86 ± 0.06
Sulfonic acid	$\mathbf{8}^*$		≥ 200	–1.45	–1.17	3.84E-08	–6.1	–7.42	<2.0	–0.81	0.31 ± 0.08
Sulfinic acid	**9**		≥ 200	–1.30	–0.84	ND†	–6.8	ND†	2.1	2.00	5.0 ± 0.7
Sulfonamides	$\mathbf{10}^*$		≥ 200	0.96	0.63	2.13E-06	4.2	–5.67	10.04	11.38	60.72 ± 0.04
	$\mathbf{11}^*$		≥ 200	1.42	1.15	1.05E-05	8.1	–4.98	>12	12.06	37.27 ± 0.06
Acyl-sulfonamides	$\mathbf{12}^*$		199.70 ± 0.30	–1.02	–0.21	3.45E-07	1.4	–6.46	4.94	4.08	12.8 ± 0.2
	$\mathbf{13}^*$		199.03 ± 1.24	0.17	–0.22	1.53E-06	2.2	–5.81	5.86	4.12	8.1 ± 0.2
Sulfonylurea	$\mathbf{14}^*$		197.76 ± 2.24	–1.23 ± 0.06	–0.87	2.61E-07 ± 1.01E-07	3.0 ± 5.4	–6.61 ± 0.20	5.04	4.14	31 ± 2
Acylurea	$\mathbf{15}^*$		≥ 200	1.42	0.57	1.63E-05	–3.3	–4.79	>12	11.77	77 ± 1
Tetrazole	$\mathbf{16}^*$		≥ 200	–0.25 ± 0.10	0.10	4.83E-07 ± 1.48E-07	4.7 ± 2.8	–6.33 ± 0.15	5.09	5.08	1.12 ± 0.12
Thiazolidine dione	$\mathbf{17}^*$		200.41 ± 0.41	1.07 ± 0.03	1.12	8.77E-06 ± 1.32E-06	5.5 ± 1.7	–5.06 ± 0.06	6.19	6.61	3.40 ± 0.11
Oxazolidine dione	$\mathbf{18}^*$		≥ 200	–0.16	0.70	2.46E-06	–0.6	–5.61	5.86	6.63	14 ± 1
Oxadiazol-5(4*H*)-one	$\mathbf{19}^*$		≥ 200	0.32	1.26	1.22E-06	–4.0	–5.91	5.73	6.04	1.10 ± 0.12
Thiadiazol-5(4*H*)-one	$\mathbf{20}^*$		200.47 ± 0.47	1.66	2.18	1.14E-05	–0.7	–4.94	6.50	7.19	1.17 ± 0.45
Oxathiadiazole-2-oxide	**21**		7.13 ± 0.74	ND	0.95	1.10E-07‡	–1432	–6.96‡	5.23	6.41	ND
Oxadiazol-5(4*H*)-thione	$\mathbf{22}^*$		≥ 200	–0.25	2.84	3.27E-07	–3.4	–6.49	3.58	7.77	0.65 ± 0.16
Isoxazole	**23**		≥ 200	0.46	1.34	4.65E-06	–11	–5.33	5.36	6.21	0.10 ± 0.10
Tetramic acid	**24**		≥ 200	–0.35	1.34	2.50E-06	1.3	–5.60	6.08	10.54	ND
Cyclopentane 1,3-diones	**25**		194.93 ±1.01	–0.70	2.32	2.12E-07	–3.0	–6.67	4.01	8.82	7.96 ± 0.35
	$\mathbf{26}^*$		≥ 200	–0.33	2.71	2.60E-07	–5.1	–6.58	4.47	8.72	11.09 ± 0.14
	27		199.04 ± 0.76	–0.60	2.16	1.54E-07	–8.8	–6.81	4.44	8.65	14.3 ± 0.6

Table 2-1. Experimental properties of a sub-set of synthesized isostere library[48]

`

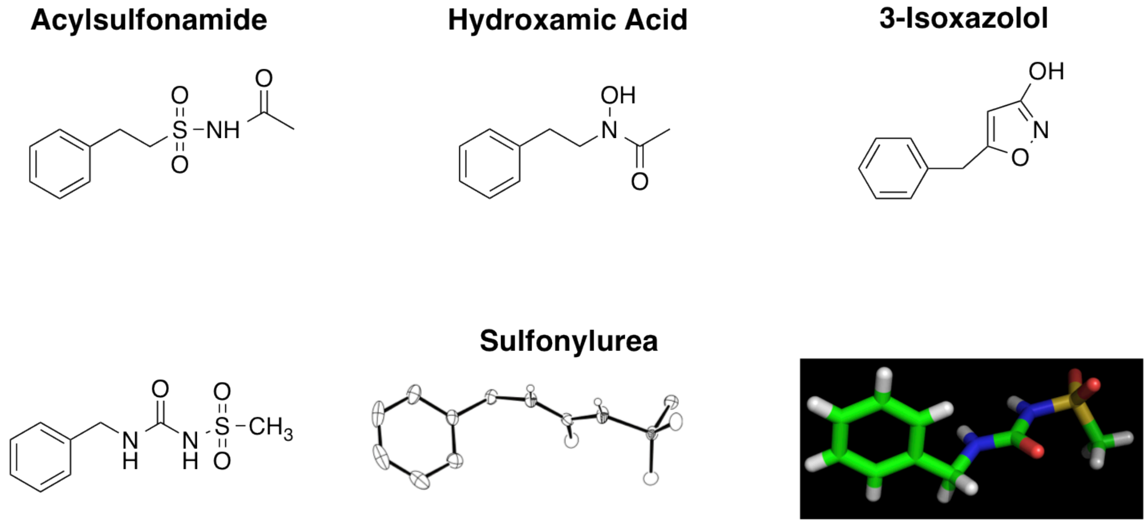


FIGURE 2-7: Carboxylic acid isosteres that I synthesized

The schemes for these syntheses are found in **APPENDIX A-D**.

`

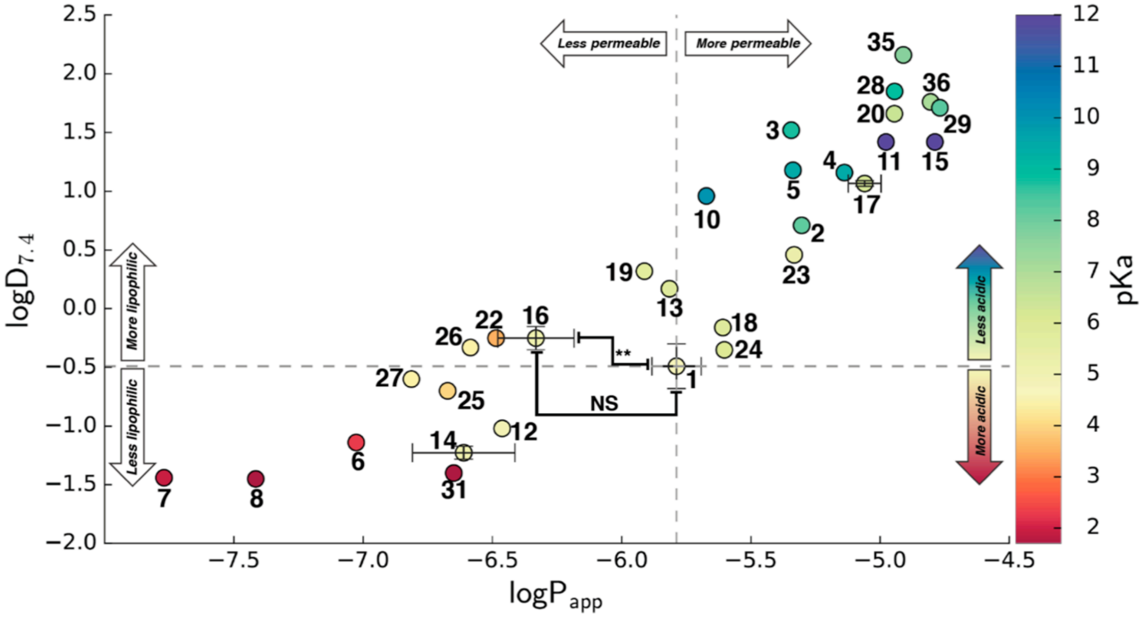


FIGURE 2-8: Plot of log P_{app}, log$D_{7.4}$, and pKA for isostere library

This plot demonstrates the values of log P_{app}, log$D_{7.4}$, and pKA relative to the carboxylic acid (centered as compound 1).

`

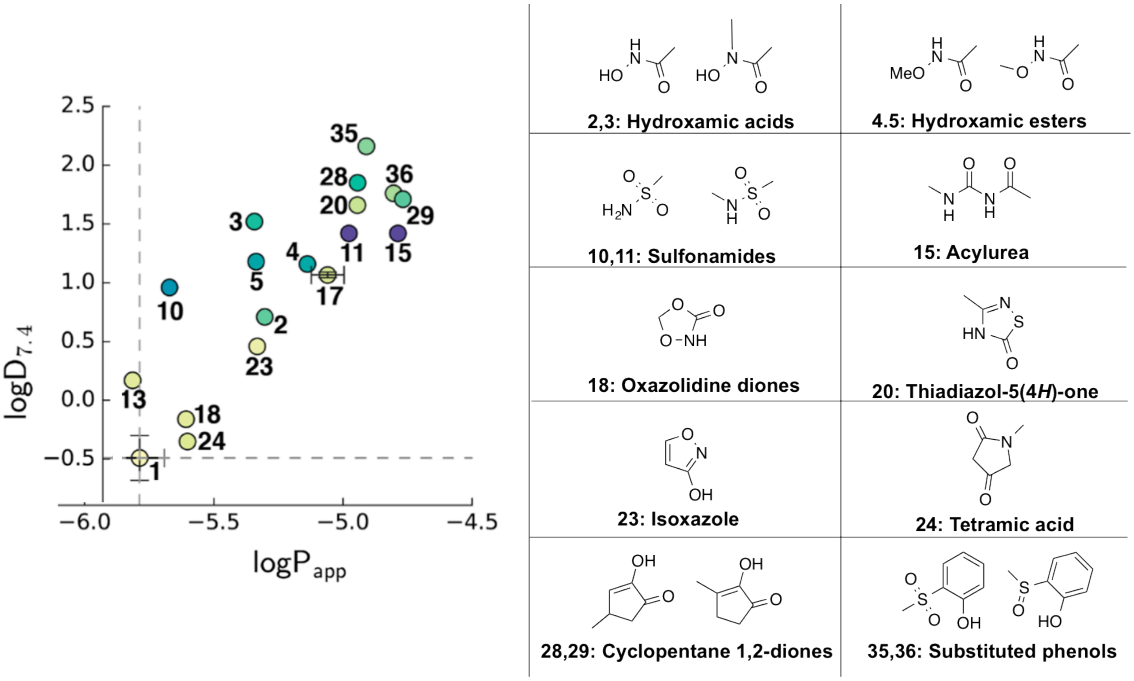


FIGURE 2- 9: Lipophilic and Permeable Isostereres

Carboxlic acid isosteres that are more permeable and lipophilic than the reference carboxylic acid (compound 1).

`

b. Chemical enumeration to produce an NSAID isostere library

Our goal was to develop carboxylic acid isosteres of not only licofelone, but of 11 other nonsteroidal anti-inflammatory drug (NSAID) scaffolds **(FIG 2-12)**. This variation in scaffolds would give us a greater chance of finding a series of compounds that with both brain penetrance and COX/LOX activity. However, our goal of synthesizing an entire library of isosteres from these 12 scaffolds was challenging due to the large isostere chemical space that we needed to explore. In addition to using experimentally-determined physiochemical properties as criteria for brain penetrance, we wanted additional criteria to predict whether a given isostere was going to be active against COX/LOX. Our final choice of compounds to synthesize would take into account promising results from both criteria. To predict activity, we used molecular docking. To design our compound library, we used the chemical enumeration algorithm MARVIN[49] (Chemaxon) to vary three positions of each NSAID scaffold (**FIG 2-10)**: the carboxylic acid group (substituted for a wide range of carboxylic isosteres of different classes); the halogen substituent on the benzyl ring; and the length of the alkyl chain spacer between the pyrazole ring and the carboxylic acid group (to vary the conformational flexibility of the isostere interacting with binding site residues). Varying these positions provided us with an enumerated representation of all of the NSAIDS; the example of licofelone (1292 compounds generated from enumeration) is shown in **FIG 2-11**. The same procedure was repeated for 11 other NSAID scaffolds (**FIG 2-12).** In the end, there were a total of 23,851 NSAID isosteres output by the enumeration algorithm.

c. Filtering compounds based on physiochemical properties

As discussed in **Chapter 1.VI**, the goal is to have molecules that are as close to the "drug-like" space as possible. The following criteria were used to filter our compound libraries:

`

- MW < 500 Daltons
- Total polar surface area < 20 Å^2
- Hydrogen bond donors < 2

After performing this filter, we ended up with 21,048 compounds for our docking library (**TABLE 2-2,** and **FIG 2-12**).

`

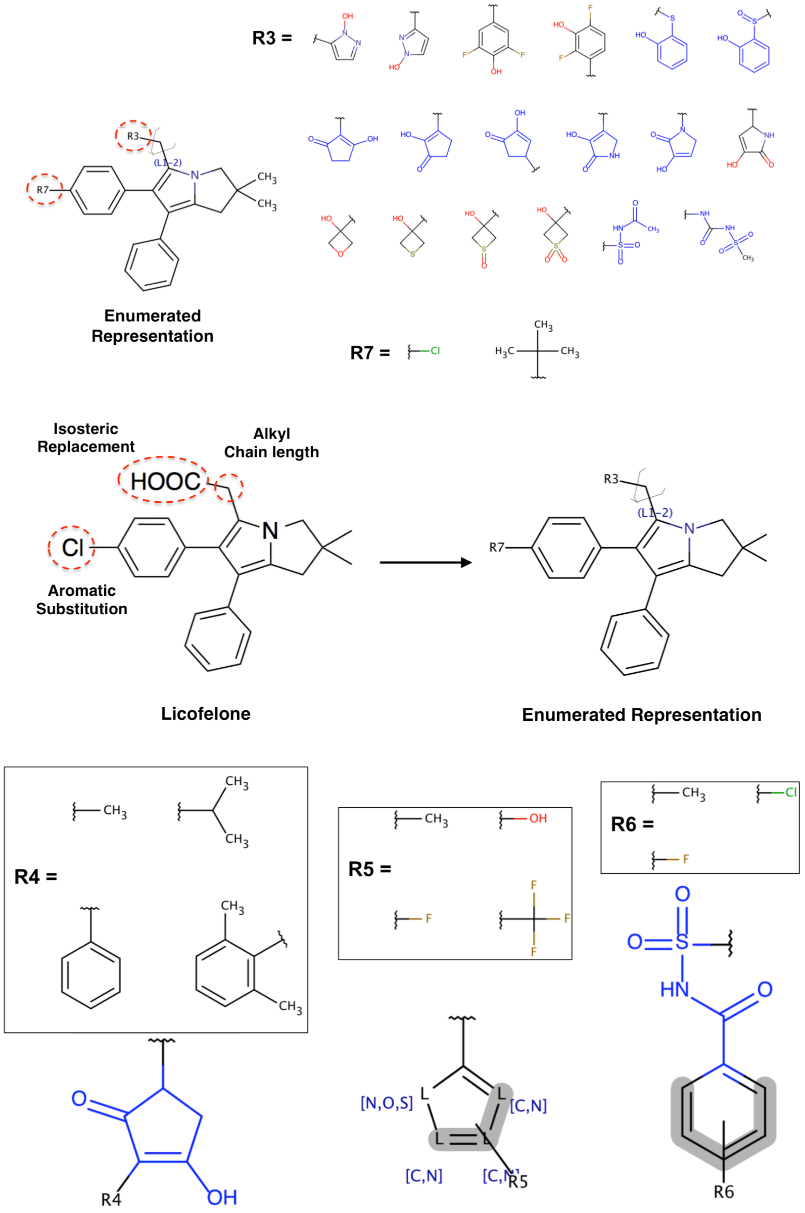
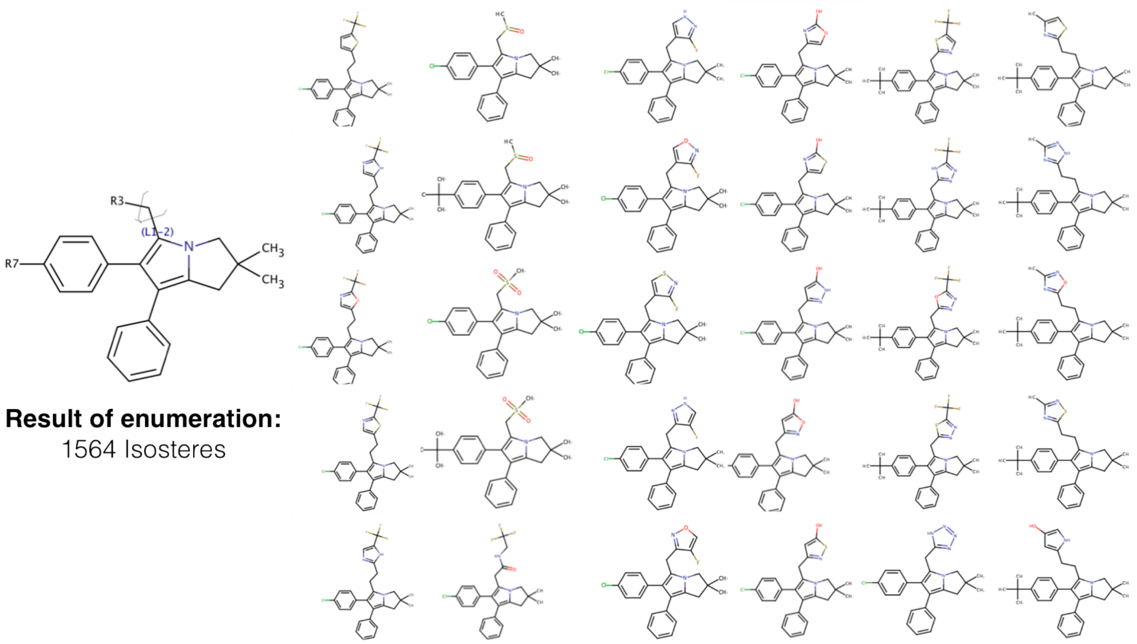


FIGURE 2-10: Chemical enumeration of three different licofelone sites

Substituting different isosteres and changing the length of the alkyl chain spacer at position R3; varying the phenyl substituent R7; and adding additional variation on the R3 substitutions led to the creation of a diverse library of licofelone isosteres (1292 compounds) from one scaffold. This process was performed for all of the NSAID scaffolds in **FIG 2-12** to give us a library of 23,851 NSAID isosteres.

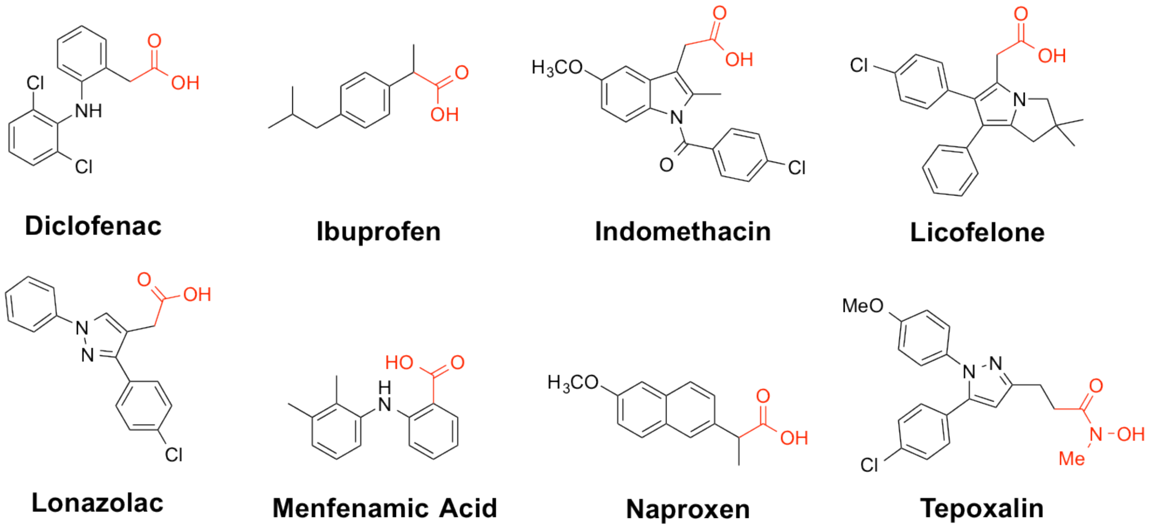
`



Result of enumeration:
1564 Isosteres

FIGURE 2-11: Example output after licofelone chemical enumeration

`



Diclofenac

Ibuprofen

Indomethacin

Licofelone

Lonazolac

Menfenamic Acid

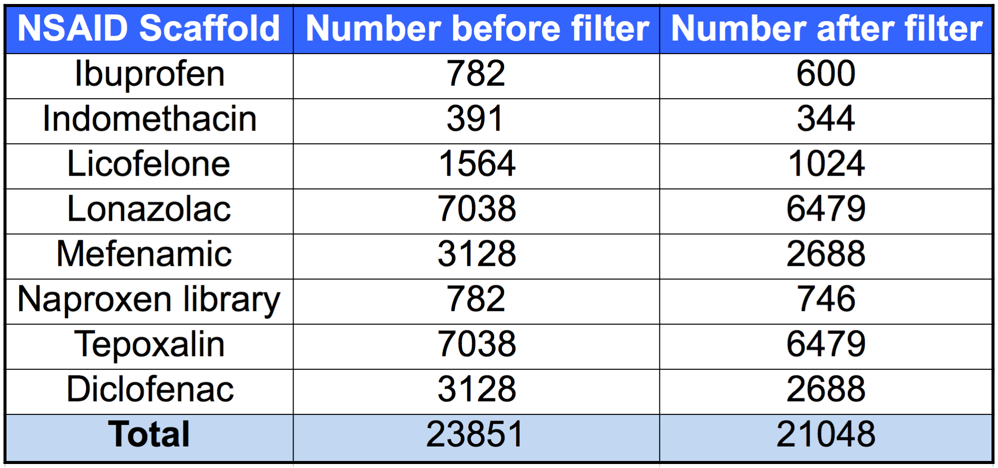
Naproxen

Tepoxalin

FIGURE 2-12:

The different NSAID scaffolds that chemical enumeration was performed on, resulting in an NSAID isostere library.

`



NSAID Scaffold	Number before filter	Number after filter
Ibuprofen	782	600
Indomethacin	391	344
Licofelone	1564	1024
Lonazolac	7038	6479
Mefenamic	3128	2688
Naproxen library	782	746
Tepoxalin	7038	6479
Diclofenac	3128	2688
Total	23851	21048

TABLE 2-2: NSAID isostere docking library after physiochemical property filter

Using criteria to filter-out compounds from our isostere library that had detrimental properties for bioavailability, brain penetration and COX/LOX inhibition (MW < 500 Daltons; Total polar surface area < 20 Å^2; Hydrogen bond donors < 2, we were able to eliminate compounds from our library and decrease the number of compounds to dock and subsequently analyze.

`

III. MOLECULAR DOCKING OF NSAID ISOSTERES

In the following sub-sections, I will describe the workflow used for docking the NSAID isostere library onto COX-1.

a. Selecting a COX-1 crystal structure: general ideas

With a large-library of NSAID isosteres on hand, our goal was to next use docking algorithms to determine if the compounds could theoretically inhibit COX-1. Though we were targeting both COX-1 and 5-LOX, studies have shown that 5-LOX inhibitors work by binding to the protein 5-lipoxygenase-activating protein (FLAP), and possibly to a complex of FLAP/5-LOX[50]. Because of this, and because there did not exist an x-ray structure of an NSAID bound to 5-LOX, we decided to focus our SAR efforts on COX-1 inhibitors. This approach was justified by findings that isosteric replacement of NSAID scaffolds can convert a COX-1 inhibitor to a 5-LOX inhibitor[51].

First, we wanted to choose an appropriate x-ray crystal structure of COX-1 for our docking experiments. This task was complicated by the fact that there are 20 different x-ray structures of COX-1, bound to various inhibitors and endogenous ligands and existing in a variety of conformations (**FIG 2-13**). When one encounters several structures of the same target, selection of the proper one for docking is important. As discussed in **Chapter 1**, proteins exist in a variety of conformational states, and each x-ray structure provides a snapshot of thousands of possibilities. The goal of structure selection is to select the one that is in a conformation that is best suited for the query compound. This principle can be illustrated by comparing the x-ray structures of COX-1 bound to indomethacin[18] and ibuprofen[52] . Though indomethacin and ibuprofen differ in size, COX-1 binds to both of them with high affinity[18,52], and x-ray crystal structures have been solved for both of them (PDB code **1EQG** for ibuprofen, and code **1PGG** for

`

indomethacin). However, if one were to remove from 1EQG and attempt to dock indomethacin into it, the binding energy would be -5 kcal/mol. This is in contrast to the -8 kcal/mol that would be output if indomethacin were docked to 1PGG. The reason for this is that 1EQC and 1PGG represent different conformations of the binding site. 1EQC has a tighter binding site (presumably due to an induced fit made with Ibuprofen), and when indomethacin is placed there, steric clashing between the bulky aromatic rings and the residue side-chains contribute unfavorably to the calculated binding energy. This example illustrates that if an unsuitable structure is used to dock one of our scaffold libraries, potentially tens of thousands of potential hits would be lost due to false-negative elimination based on poor calculated binding energies.

b. Validating the docking algorithm

Before doing any docking experiment, the first step is to determine whether the docking algorithm is able to verify a known result i.e. recapitulate the binding mode of the ligand that it was crystallized with. To do this, a 2D SDF representation of each scaffold compound was generated. Next, each scaffold was docked back into its respective x-ray structure (that had been stripped of the compound). If the docking algorithm is performing its stochastic search properly, it would provide the same (or very similar) pose as the x-ray structure's pose. If this is not the case, then there is a technical flaw that needs to be addressed (i.e. the size of the user-defined grid box is not appropriate, or there are binding site water molecules that assist in binding that need to be incorporated). Using this procedure, Autodock was able to correctly output the correct scaffold pose as the x-ray structure's pose.

c. Selecting an x-ray crystal structure for licofelone and other NSAIDS

Licofelone, unlike other NSAIDs that we built enumerated libraries for, does not have an x-ray crystal structure with COX-1. Hence, it is important to pick the appropriate structure due to variation in binding site conformations that was described in **a**. **FIG 2-12** shows the structure of licofelone in comparison to six other NSAIDS that were crystallized with COX-1. Structurally, licofelone scaffold has a central pyrazole ring with meta and para-substituted phenyl rings, and a ortho-substituted carboxylic acid group. Compared to these other structures, the compound that structurally-resembles it the most is indomethacin. When licofelone is docked into indomethacin's structure (**1PGG**), it exhibits the same binding pose as indomethacin (**FIG 2-14**): its carboxylic acid group faces ARG 120, and its bulky aromatic rings face down towards the deep end of the binding pocket.

`

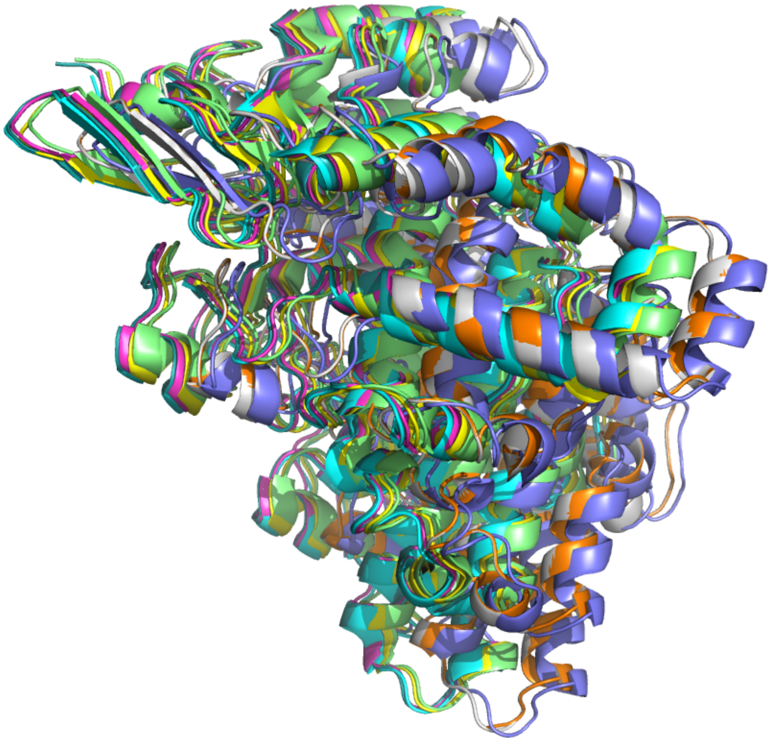


FIGURE 2-13: Multiple x-ray crystal structures for COX-1

Since various structures of COX-1 had been solved (bound to various ligands small-molecules), it was important to find the structure that was in the conformation best suited for the drug scaffold we were docking into it.

`

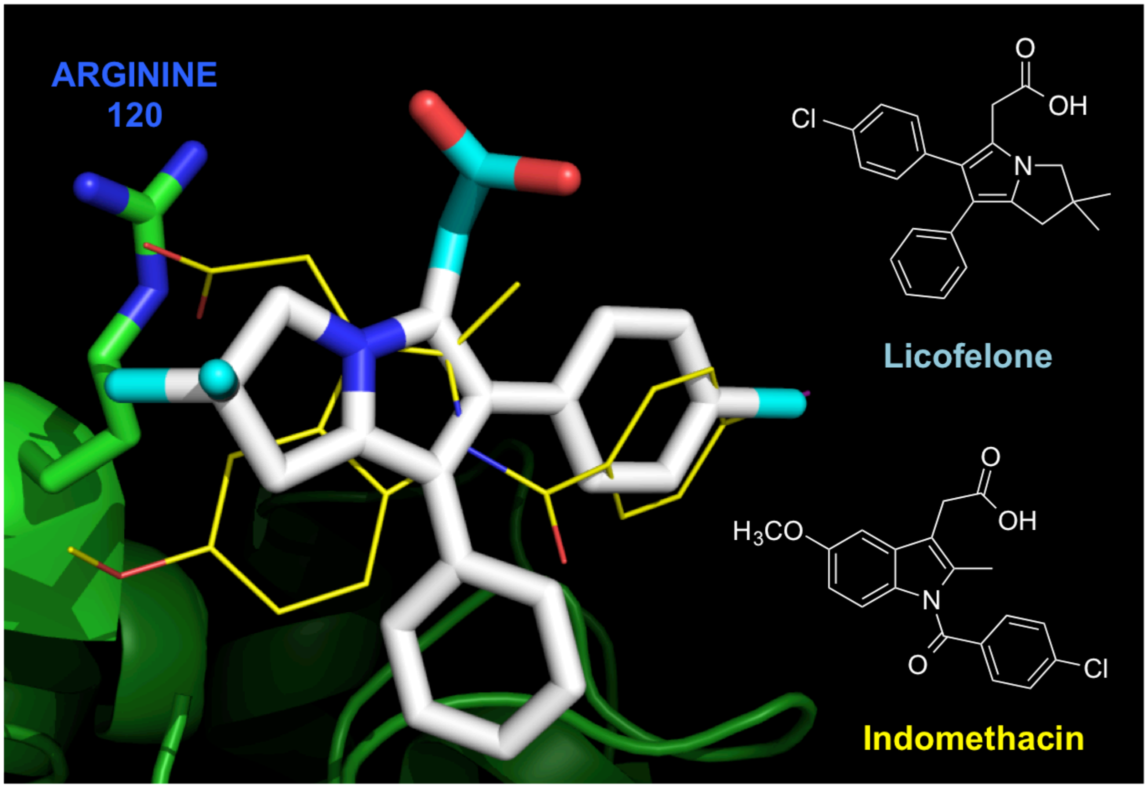


FIGURE 2-14: Licofelone docked onto the crystal structure of indomethacin

We choose the COX-1 structure 1PGG (bound to indomethacin) for screening of the licofelone isosteres. This was due to the fact that (a) licofelone had a similar scaffold to indomethacin and (b) licofelone existed in the same general pose as indomethacin when it was docked in, with high calculated binding energy.

`

interaction with ARG 120, while the two aromatic rings face deep into the cavity. The calculated binding energy is -9.0 kcal/mol, which is not far from the calculated binding energy of indomethacin when it is docked into its own structure (-10.3 kcal/mol). When attempting to dock licofelone into the x-ray structure of ibuprofen (**1EQG**), the same result occurred as when indomethacin was docked into it, and the binding energy was worse than when it was docked to 1PGG (-5.0 kcal/mol). This is presumably due to the same issues as with indomethacin, related to the size of the ibuprofen binding size and the occurrence of steric clashing with licofelone's aromatic groups.

d. Docking screen of NSAID isosteres

After exploring different possibilities for a common COX-1 structure to dock the entire library into, I decided to dock each NSAID isostere library into it's own x-ray crystal structure (and licofelone into the indomethacin structure **1PGG**). Two criteria were used to determine if a compound was a docking hit. First, compound's binding-energy had to be within 2 kcal/mol of the binding energy of the native ligand (when re-docked into its own x-ray structure). Second, the orientation of the compound had be consistent with the known COX-1/NSAID binding mode, such that the isosteric replacement group was adjacent to ARG 120, and the aromatic groups were buried into the binding cavity. **TABLE 2-3** provides a summary of the binding energy cut-offs, and **FIG 2-15** provides representative hits for each NSAID scaffold.

`

Compound	Binding Energy (kcal/mol) in X-ray Structure	Lower Bound (kcal/mol)
Diclofenac	-7	-5
Ibuprofen	-7.4	-5.4
Indomethacin	-10.4	-8.4
Ionazolac	-9.6	-7.6
Menfenamic Acid	-8.1	-6.1
Naproxen	-8.4	-6.4
Tepoxalin	-8.8	-6.8

TABLE 2-3: Binding energy cut-offs for each of the docking scaffolds

The binding energy when each NSAID was re-docked into its own x-ray structure was used as a baseline to compare the performance of a scaffold's isostere analogues. Due to the absence of a crystal structure with COX-1, licofelone used Indomethacin's binding energy cut-offs.

`

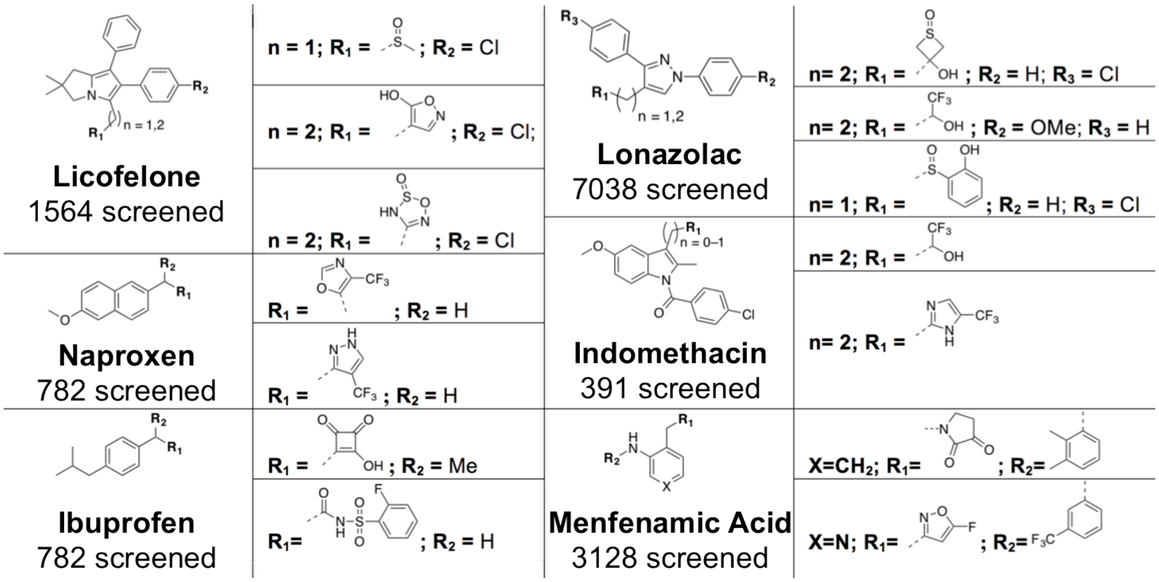


FIGURE 2-15: Representative hits for each scaffold from COX-1 docking screen

`

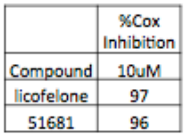
e. Synthesis and experimental validation of a docking hit

To help determine if the assumptions that we made during our docking screens (i.e. design of enumerated library and selection of isosteres; filtering of library based on physiochemical values; and choice of COX-1 structure to dock), I synthesized a docking hit for licofelone: a hydroxamic acid isostere (CNDR-51681) that was also, from **FIG 2-3**, predicted to confer brain penetrance via its $\log D_{7.4}$ and $\log P_{app}$ values (**FIG 2-16**). The synthesis for this compound is shown in **APPENDIX F.**

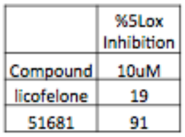
In general, the assay that provides a read-out of whether our compounds are inhibiting the COX or LOX pathways is a rat basophilic leukemia (RBL-1) cell-based assay for specific assay conditions)[53]. In this assay, cells are plated in 24-well plates with growth medium, and incubated at 37^0 C with either DMSO control or 10 uM of a candidate COX or LOX inhibitor. After 2 hours, 12 uM calcium ionopore is added, which results in the induction of both the 5-LOX (leading to an increased secretion of the leukotriene LTB_4) and COX-1 pathways (leading to the secretion of prostaglandin PGD_2). The supernatants of these cells are then collected, spun down, and dried under vacuum. The dried samples are then dissolved in 50% acetonitrile and analyzed by LC-MS, to determine the levels of PGD_2 and LTB_4. Changes in these metabolite levels, when compared to the DMSO control, can provide a read-out of the percent inhibition of both pathways.

This assay demonstrated that CNDR-52681 was able to inhibit COX-1 in an analogous manner to licofelone (96% inhibition of COX-1 metabolites, vs. 97% by licofelone). However, CNDR-52681 was also able to inhibit 5-LOX pathway compared to licofelone (91% inhibition, vs. 19% for licofelone). This data suggests our use of molecular docking to prioritize compounds for synthesis was sound.

`



	%Cox Inhibition
Compound	10uM
licofelone	97
51681	96



	%5Lox Inhibition
Compound	10uM
licofelone	19
51681	91

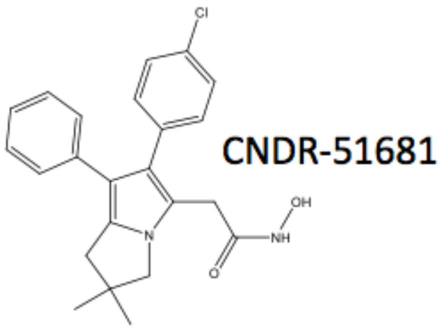


FIGURE 2-16: RBL cell COX/LOX assay to test activity of CNDR-51681

10 uM of the licofelone hydroxamic acid isostere were tested in an RBL cell-based COX1/LOX assay. In this assay, addition of calcium ionopore induces the stimulation of these pathways and the corresponding production and secretion of leukotrienes and prostaglandins, which can be monitored by LC-MS analysis of the cell supernatant. Incubation of the cells with a COX or LOX inhibitor prior to calcium ionopore induction will result in a decrease in LC-MS signal for prostaglandins or leukotrienes, measured relative to a DMSO control. These data demonstrate that CNDR-51681 inhibit both COX and LOX pathways.

Experiments and data analysis were performed by Vishruti Makani, Yuemang Yao, and Michael James (labs of Dr. Virginia Lee and Dr. John Q. Trojanowski, Center for Neurodegenerative Disease Research).

`

IV. IDENTIFICATION OF TRI-ACTIVE COMPOUNDS THAT STABILIZE MICROTUBULES AND INHIBIT COX/LOX PATHWAYS

As the synthesis of docking hits continued, another question arose: can our computational docking approach (and the assumptions that we made) be used to predict if a compound can bind to a target from another class of compounds, and potentially inhibit multiple targets? This question is relevant in light of the pathophysiology of Alzheimer's disease, where multiple independent pathways have been linked to disease onset and progression[40,54]. In the following section, I will describe our attempts to use computational docking to develop compounds that target COX-1, 5-LOX and which stabilize structures in the brain called microtubules.

a. Background: Microtubules in Alzheimer's Disease

Microtubules (MTS) are hollow, 24 nm diameter tubes made up of α and β-tubulin heterodimers[54]. In the cytoskeleton of eukaryotic cells, MTs play essential structural and regulatory roles, such as intracellular transport, cell division, and the maintenance of cell shape. In the axons of neurons, the protein tau stabilizes microtubules and plays a key role in axonal transport. In the class of disorders called tauopathies[55], tau function is lost (i.e due to protein misfolding, sequestration into insoluble aggregations i.e. neurofibrillary tangles, and hyperphosphorylation) and become detached from MTS, resulting in axonal transport deficiencies occur that can have been implicated in the development Alzheimer's Disease and fronto-temporal dementia (FTLD). Currently, brain-penetrant MT-stabilizing agents are developed (i.e. Epothilone D, currently in phase 1B clinical trials for AD) with the aim of restoring tau function in order to restore effective axonal transport[56].

`

As discussed in **section 2a**, studies have shown that Alzheimer's Disease initiation and progression occurs through multiple co-occurring pathophysiological pathways, such as loss of Tau function along with neuroinflammation. Hence, one possible therapeutic avenue would be to target multiple pathways simultaneously, whether through a combination of drugs or one compound that has multiple targets. Since we identified a possible pipeline (through isosteric replacement, chemical enumeration, and computational) for prioritizing compounds to synthesize for dual COX/LOX inhibition, a new question arose: could we identify compounds that could not only theoretically inhibit COX/5-LOX, but which can also serve as microtubule-stabilizing agents?

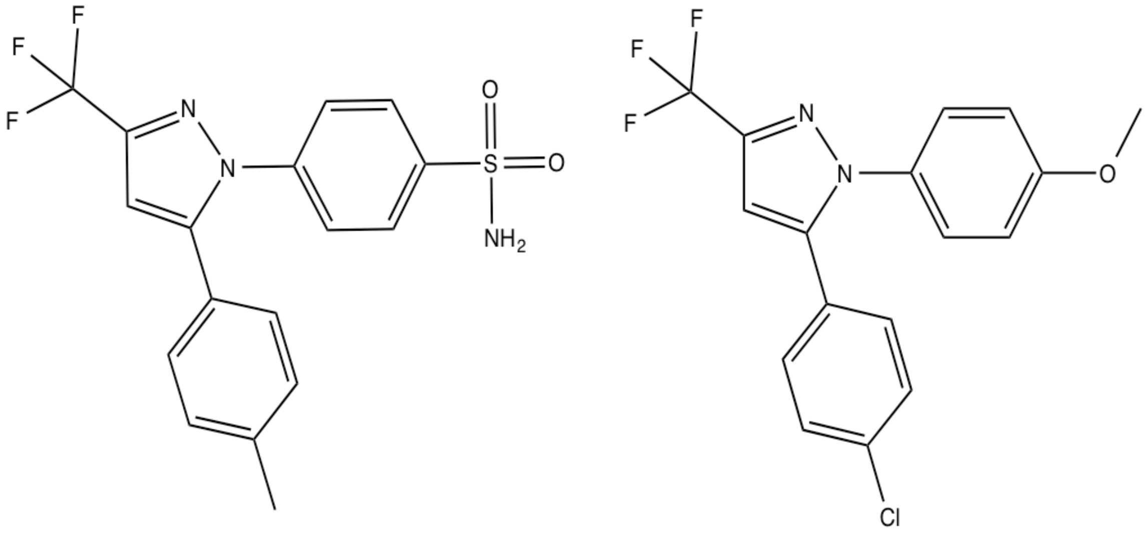
b. Using docking to identify tri-active compounds

Our goal of developing lead compounds with microtubule stabilizing activity and with activity against COX and 5-LOX was motivated by an observation that many of the compounds that were synthesized in literature with reported microtubule-stabilizing activity contained a central scaffold that was similar to the COX inhibitors celecoxib[57] and SC-560[58] (**FIG 2-17**). These two compounds contain a central pyrazole ring (with a trifluoryl methyl group in the 3’ position) that is flanked with hydrophobic phenyl groups in the 1' and 5' positions. According to the x-ray crystal structure of celecoxib bound to COX-1[57], the nitrogen of the sulfonamide makes key hydrogen bond interactions with GLN 192 and LEU 252 (**FIG 2-18**), while the 2' phenyl ring makes hydrophobic contacts with PHE 518 and other hydrophobic residues adjacent to it (i.e. LEU 352, ILE 517) deeper in the cavity. In the absence of an x-ray structure of SC-560 bound to COX-1, we presumed that the methoxy phenyl group of SC-560 makes similar interactions as celecoxib's sulfonamide group.

`

From the literature and chemical databases, we compiled 114 cases of MT-stabilizing agents that contained an di-phenyl-substituted imidazole or pyrazole ring scaffolds[59](**APPENDIX G**). Next, we needed to choose a structure to dock our compounds against. Since SC-560 most closely resembled celecoxib, we first docked it onto the x-ray structure of celecoxib bound to COX-1 (PDB code: 3KK6). SC-560 bound with a calculated binding energy of -10.3 kcal/mol, close to the binding energy of celecoxib redocked to 3KKG (-10.0 kcal/mol). Furthermore, SC-560 was in the same pose as the celecoxib crystal structure pose (**FIG 2-18**), suggesting that it shared the same binding mode. Due to these results, 3KK6 was selected as the structure to dock our compounds onto.

`



Celecoxib

SC-560

FIGURE 2-17: Celecoxib and SC-560 structures

`

For each docked compound, the poses were analyzed to determine if they were in the same general orientation as celecoxib, and making the same interactions with GLN 192, LEU 252 and PHE 518 **(FIG 2-18)**. Those compounds were then chosen as candidates, and sorted according to their theoretical binding energy. After analyzing our docking results, we initially selected 5 compounds that we had already synthesized to test for COX and LOX inhibition activity. 4 of these compounds were among the docking top scorers with respect to calculated binding energy (**FIG 2-19**), while the 5th compound (CNDR-51665) was an attempt to perform SAR optimization on the pyrazole ring by removing the 1'chloro group from the pyrazole scaffold of CNDR-51735.

These 5 compounds were tested in the RBL COX/LOX LC-MS assay[53] (described in the previous section) (**FIG 2-19**). With the exception of CNDR-51672, all five compounds exhibited decreases in LTB4 synthesis (indicating 5-LOX inhibition). With respect of COX inhibition, four out the 5 compounds conferred decreases in PGD2 synthesis, indicating COX inhibition (CNDR 51672 had no change with respect to vehicle). Hence, four out of the five compounds that we tested directly from the docking output were experimentally validated in both assays. One short-coming of our approach was based on the performance of CNDR-51665 in the COX inhibition assay: despite that its docking energy was -7.9 kcal/mol, it featured the best inhibition of PGD2 synthesis out of the entire series.

`

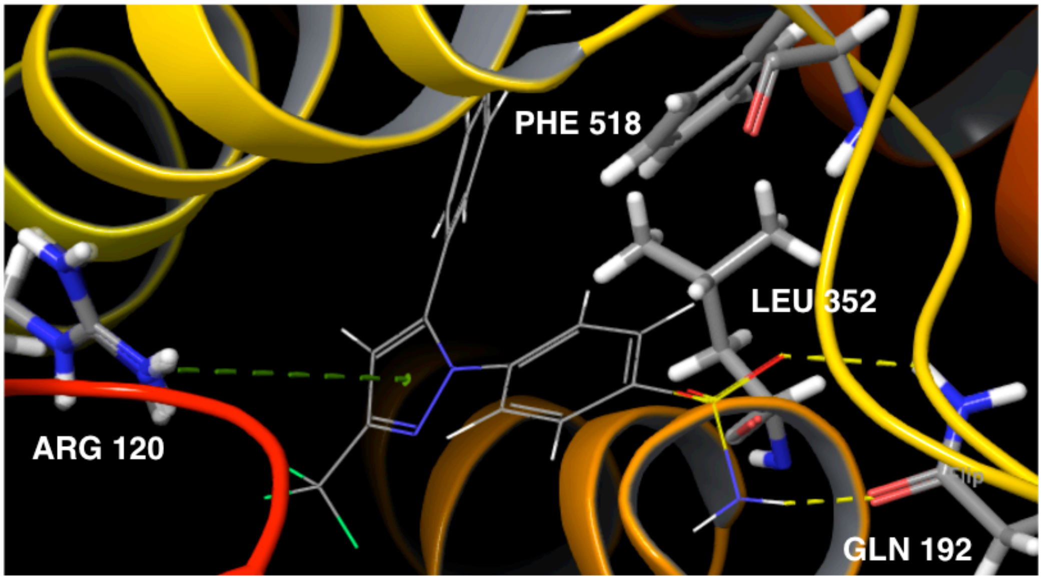


FIGURE 2-18: Celecoxib interactions in the COX-1 binding site

Celecoxib's sulfonamide nitrogen makes key hydrogen bond interactions with GLN 192 and LEU 252, while the 2' phenyl ring makes hydrophobic contacts with PHE 518 and other hydrophobic residues adjacent to it (i.e. LEU 352, ILE 517) deeper in the cavity.

`

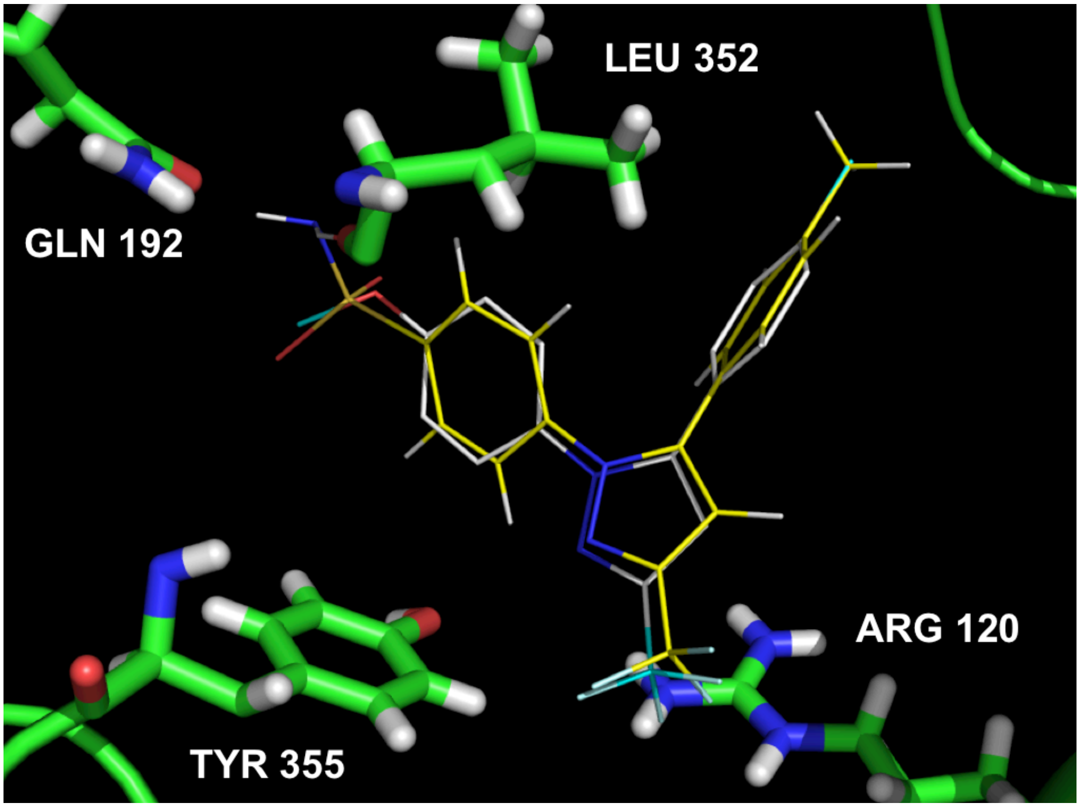


FIGURE 2-19: SC560 docked in Celecoxib-COX-1 x-ray structure

Docked pose of SC560 (in grey, overlayed with celecoxib, in yellow) in 3KKG COX-1 x-ray structure

`

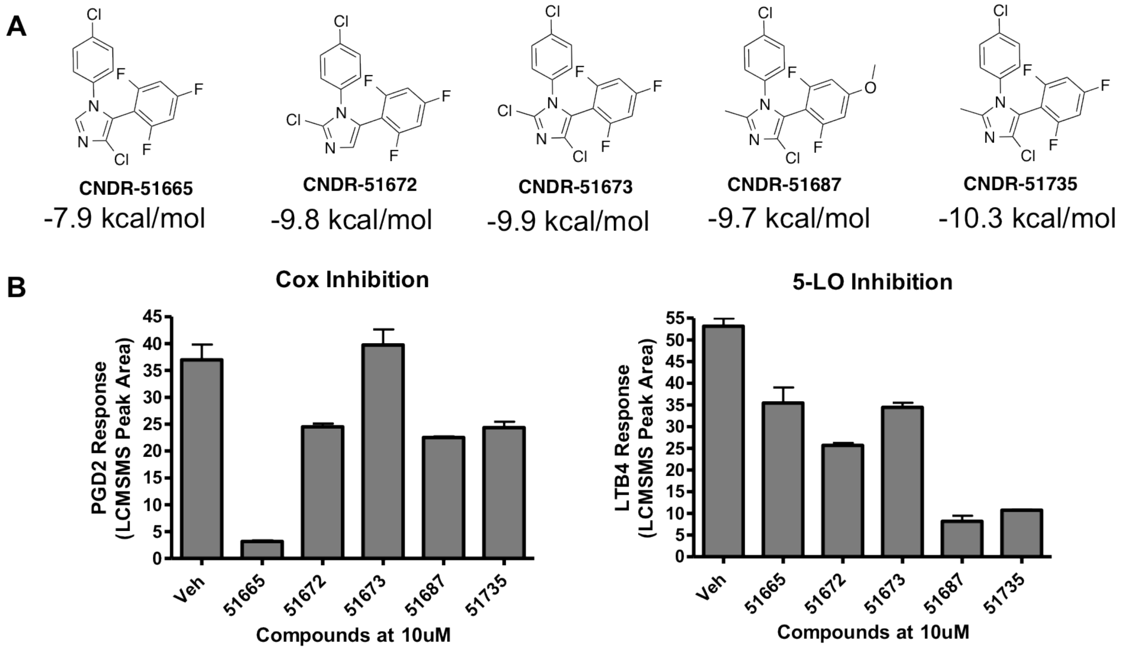


FIGURE 2-20: COX and LOX activity data for selected docking hits

A: Compounds initially selected for testing, with Autodock binding energies

B: Experimental activity results for both COX and LOX inhibition.

Experiments and data analysis were performed by Vishruti Makani, Yuemang Yao, and Michael James (labs of Dr. Virginia Lee and Dr. John Q. Trojanowski, Center for Neurodegenerative Disease Research).

`

Based on our initial results, we synthesized 38 additional compounds **(FIG 2-21)**. 27 of these were directly from the docking screen (red compounds were docking hits because they met our binding energy criteria, while the pink compounds, despite not meeting our binding energy requirements, were synthesized because they were easily derived from the precursors of the red compounds). The other 11 compounds (blue) were slightly modified analogues of the docking hits. **FIG 2-22** and **FIG 2-22** demonstrate the activity results of these compounds in the LOX and COX inhibition assays.

When combining these results from the initial results in FIG 2-20, we found that there were 11 experimentally-validated COX/LOX hits. Seven of these hits came directly from the docking screen (with no optimization), while three of these hits were slightly modified from the docking hits. These results demonstrate the power of the molecular docking approach for identifying candidate compounds from a large set of possibilities, and provide a starting point for tri-active Alzhimer's Disease drugs.

`

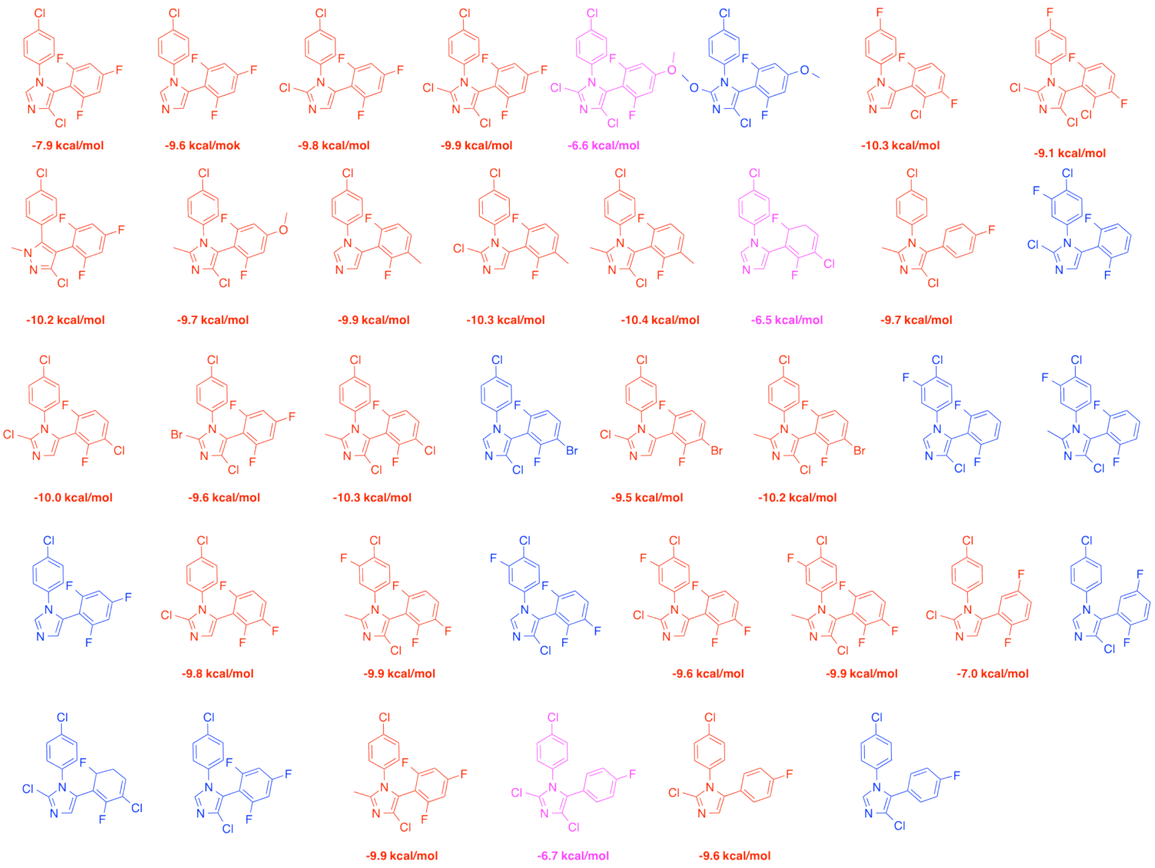


FIGURE 2-21: Additional microtubule-stabilizing agents synthesized from docking hits

Red compounds were hits directly from the docking screen. **Pink compounds** were compounds that were not hits (i.e. did not make our binding energy cut-off), but which were easily obtained from the synthetic schemes of other compounds. **Blue compounds** were slight modifications of the docking hits.

`

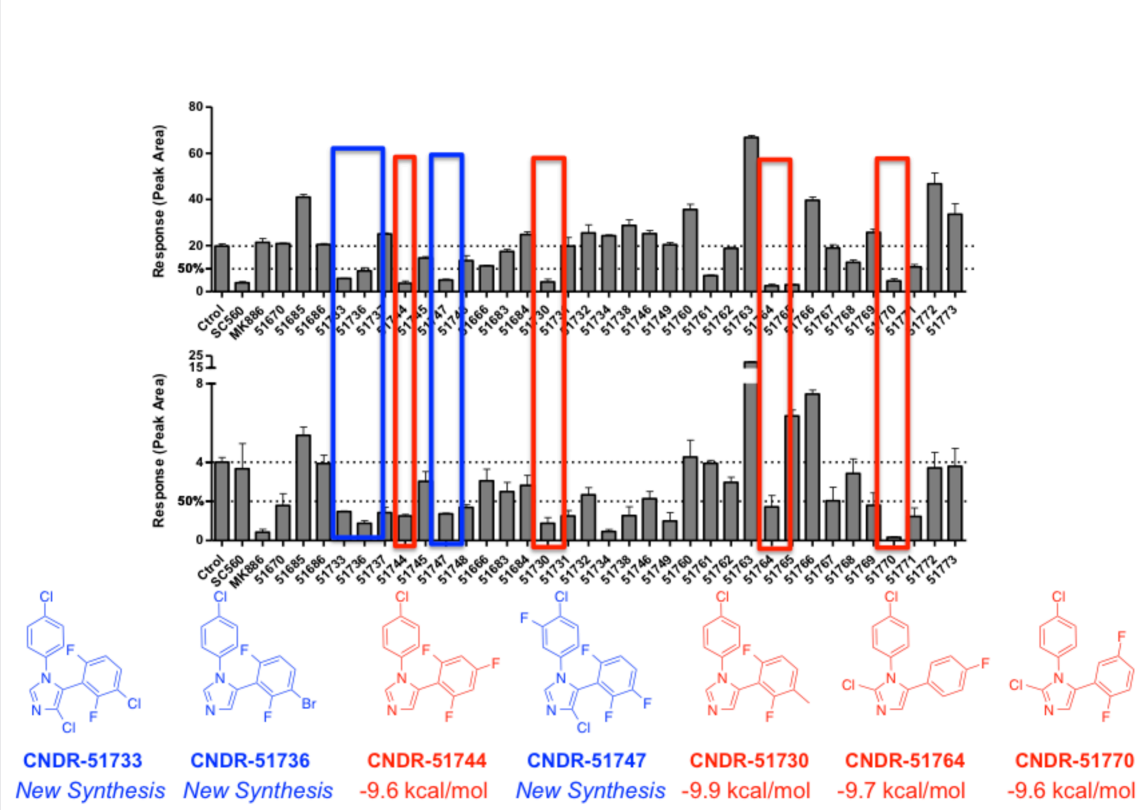
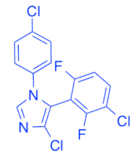


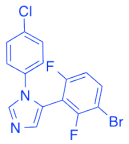
FIGURE 2-22: Results of COX/LOX activity experiments of additional microtubule-stabilizing agents from FIGURE 2-21

`

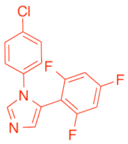
CNDR#	Cox %Inhibition	Lox %Inhibition
51730	79	78
51733	71	63
51736	54	78
51744	82	69
51747	74	66
51764	87	57
51770	76	96



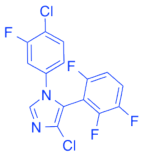
CNDR-51733
New Synthesis



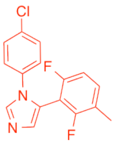
CNDR-51736
New Synthesis



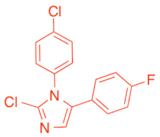
CNDR-51744
-9.6 kcal/mol



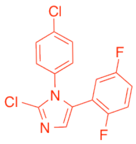
CNDR-51747
New Synthesis



CNDR-51730
-9.9 kcal/mol



CNDR-51764
-9.7 kcal/mol



CNDR-51770
-9.6 kcal/mol

FIGURE 2-23: Results of COX/LOX activity experiments of additional microtubule-stabilizing agents from FIGURE 2-21

`

PART II: (RE)DESIGNING A DRUG TARGET

CHAPTER III. COMPUTATIONAL DESIGN OF WATER-SOLUBLE VARIANTS OF MEMBRANE PROTEINS

3.1. MOTIVATION

As discussed in the main introduction, many of the barriers to membrane protein drug discovery are due to the lack of a simple, direct method for testing the interactions between a small-molecule inhibitor and a membrane protein. This is largely due to hydrophobic residues found on the exterior of membrane proteins, which make them difficult to express, isolate, and perform experimental assays with. Ideally, a given membrane protein could be isolated and its interactions with small-molecules probed directly. Our goal for this project is develop a cell-free, label-free, detergent-free solution phase binding assay for membrane proteins **(FIG 3-1).** We want this assay to serve as a quick and simple way for the drug discovery community (in both academia and in the pharmaceutical industry) to test compounds for activity. To achieve this goal, we devised a general algorithm for designing water-soluble variants of membrane proteins. Our hypothesis is that a designed water-soluble variant would retain the wild-type protein's and ligand-binding properties, and would allow for the direct testing of biological interactions and drug-binding events.

In the past, our group has used a combination of computational and rational protein design approaches to develop water-soluble variants of the bacterial potassium channel KscA[60], the nicotinic acetylcholine receptor[61], and the mu-opioid receptor[62]. However, these past studies required multiple rounds computational design and experimental testing, and ultimately required the proteins to be isolated and refolded from bacterial inclusion bodies. In the end, this was a very time-consuming and labor-intensive process. Our goal was to take everything that we learned about membrane protein solubilization in our past attempts, and to develop method that could produce a water-soluble variant in one attempt that could be isolated directly from the soluble fraction of *E. coli*.

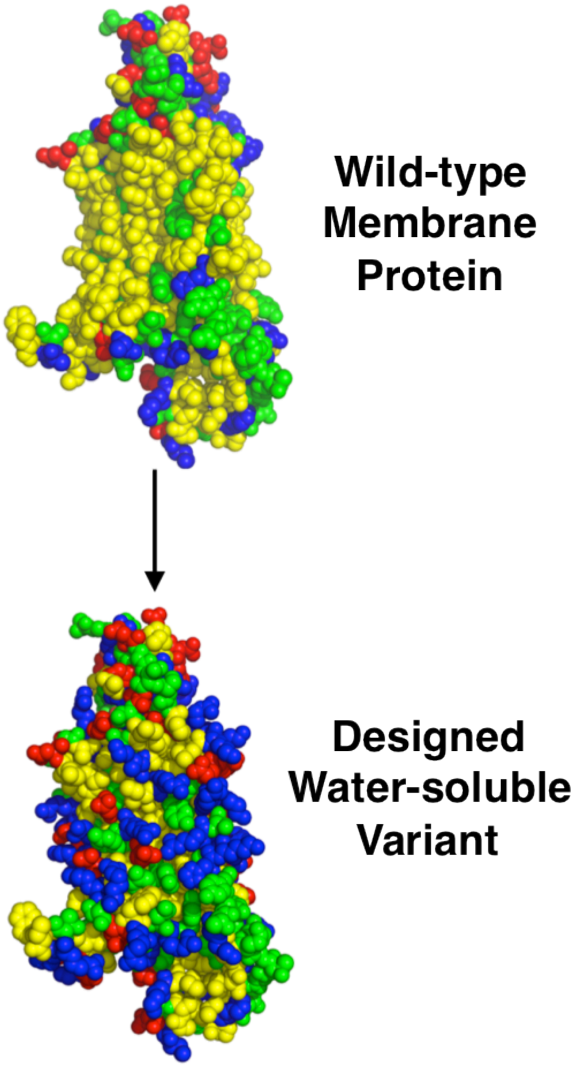


FIGURE 3.1 Computational design of water-soluble variants of membrane proteins

Our goal is to develop an algorithm that can modify an input membrane protein so that it can be used for cell-free, label-free, drug binding studies.

`

In the following section, I will describe the development of a computational approach that (i) identifies solvent-exposed, transmembrane hydrophobic residues and (ii) mutates these residues so as to confer water-solubility while retaining the protein's overall 3D-folded structure. This approach implements molecular mechanics force fields and statistical energy functions inferred from a library of water-soluble globular proteins. As proof-of-concept of this approach, we designed a water-soluble variant of the G Protein-Coupled Receptor Smoothened. Smoothened was selected based on the depth and breadth of knowledge that was available for it. There were five x-ray crystal structures in a variety of conformations[6,7], bound to a different agonists and antagonists. Several tool compounds existed for it (antagonists and agonists)[63]. Multiple parts of its intracellular pathway were deconvoluted over the past forty years in multiple model systems, from *Drosophila* to mammalian cells[64]. Lastly, it is deeply relevant to human health, as several clinical trials are underway for various SMO-driven cancer types[65].

3.2. BACKGROUND: SMOOTHENED AND THE HEDGEHOG SIGNALING PATHWAY

The Hedgehog signaling pathway is involved in cell proliferation, including the regulation of stem cell proliferation, during development and tissue repair during adulthood[64]. This pathway consists of four main components: the Hedgehog extracellular ligands (Sonic Hedgehog, Indian Hedgehog and Desert Hedgehog); the twelve-membrane spanning Patched Receptors (PTCH1 and PTCH2), which negatively regulate the pathway; Smoothened (SMO), the seven-transmembrane G Protein-Coupled Receptor that serves as the key transducer of this pathway; and the GLi transcription factors (which regulate target genes via a specific consensus sequence). Though the mechanism of SMO inactivation via PTCH1 has been the subject of much controversy, it is generally accepted that PTCH1 inhibits SMO by transporting small-molecule SMO agonists out of the cell[7,8]. Upon binding of Hedgehog to PTCH1, PTCH1 becomes internalized, resulting

`

in the intracellular build-up of SMO agonists subsequent SMO activation. Subsequently, via a still undetermined mechanism, SMO activated GLi1, which in turn expresses target genes that are involved in cell proliferation and survival.

Aberrant activity of the Hedgehog pathway has been implicated in a variety of cancer types[64,65]. One of the first links between this pathway and cancer came from the inherited disease Gorlin’s syndrome (also known as basal cell nevus syndrome). Patients with this disease are highly predisposed to the development of basal cell carcinomas (skin), medulloblastomas (brain), meningiomas (brain) and rhabdymyosarcomas (muscle). It was later discovered that Gorlin's syndrome patients had either loss-of-function mutations of PTCH1 or activating mutations in SMO. Activating mutations in SMO were later found to be implicated in other cancer types, notably medulloblastoma (MB) and basal cell carcinoma (BCC), leading to the development of several SMO inhibitors. In 2009, one such inhibitor, vismodegib, was found to be effective in phase II clinical trials for locally advanced and metastatic basal cell carinoma[66-68]. In this trial, 30% of patients with locally advanced basal cell carcinoma responded to treatment (21% of whom had a complete response) and 43% of patients with metastatic basal carcinoma responded. Since then, there have been 24 clinical trials for SMO inhibitors in a variety of cancer types. In 2012, vismodegib was approved by the FDA for treatment of BCC based on its favorable outcomes compared to chemotherapy. Despites vismodegib’s promise, a subset of patients who initially respond to treatment subsequently relapse[69]. One mechanism of vismodegib resistance is via mutations in the drug’s binding site. This phenomenon has been observed with other clinical trial compounds and represents a huge challenge to the development and widespread clinical use of SMO inhibitors.

`

One approach for developing second-line agents to counteract SMO drug resistant mutations is to develop compounds that target sites that are distinct from the main binding site that all of the clinical trial drugs bind to. These interactions, once identified, could serve as potential targets for drugs with novel mechanisms of action (i.e. protein-protein interactions, novel allosteric sites regulatory sites). However, this is challenging in SMO because, as discussed in the previous section with membrane proteins, many of the key direct intracellular and extracellular interactions that allow SMO to transduce signals are not clearly understood. For example, the hypothesized endogenous SMO agonist has not been discovered, and it is not known whether SMO activates oncogenic transcription factors through a direct interaction or indirectly through other proteins. The lack of a simple experimental system for testing SMO binding has prevented progress in our understanding of these key molecular interactions.

To facilitate the study of SMO and the identification of small molecules that impact its activity, we hypothesize that a computationally-designed water-soluble variant of SMO (wsSMO) can be developed and used in simple solution-phase biophysical ligand-binding assays to test SMO drug-binding events. This variant would retain key structural and functional features of wild-type SMO, yet will be readily isolatable from a heterologous protein expression system (i.e. *E. coli*). wsSMO and the associated biophysical assays will represent a first-step toward developing membrane protein analogs that stand to accelerate drug development. In the following sections, we will describe our efforts to develop wsSMO. First, I will provide an overview of the Saven protein design methology, which uses principles from theoretical physical chemistry to calculate the site-specific probabilities of sites in a protein structure, subject to different constraints imposes on the sequence. Afterwards, I will describe the development of an

`

algorithm that, for an input membrane protein structure, can select the exterior hydrophobic residues that, if mutated, can confer solubility to the protein as a whole.

III. COMPUTATIONAL DESIGN OF MEMBRANE PROTEINS

a. Overview of protein design and membrane protein design

The field of protein design aims to identify the physical properties that dictate protein folding and to synthesize novel proteins from theoretically designed target structures. These tasks are challenging due to the sheer complexity of protein folding. A protein can contain tens to thousands of amino acid residues, and for a single protein sequence there are a plethora of available conformations due to variation in backbone and side-chain degrees of freedom[72,73.] The possibilities are even greater when the non-covalent interactions that stabilize a protein's folded state (i.e. van der Waals, hydrophobic interactions, electrostatic interactions, and hydrogen-bonding interactions) are taken into account. Furthermore, there is complexity with respect to the number of candidate sequences a target protein structure can take: for example, a 100 amino acid protein made up of the 20 naturally occurring amino acids would have more than 10^{130} possible sequences[72,73]. The protein design field has made great strides to overcome these challenges via computational methods (guided by statistical mechanical theory) for predicting how a linear amino acid sequence will fold and, conversely, the possible amino acid sequences that will form a target structure of interest. These methods have resulted in many recent breakthroughs, such as the creation of enzymes to catalyze chemical reactions not performed by naturally occurring proteins (i.e. the Retro-aldol reaction[74], the Kemp elimination[75], and the Diels-Alder reaction[76]), and the reengineering of macromolecular interaction interfaces (i.e. the DNA-binding interface of the endonuclease I-MsoI[77]).

`

The goal of work in the following sections is to "redesign" a membrane protein so that it is now water-soluble. There are two main steps needed to solubilize a membrane protein. First, one has to identify the solvent exposed, hydrophobic residues that exist on the transmembrane domain of the protein. Second, these residues need to be mutated in such a way that the protein as a whole is now hydrophilic, *and* in such a way that the total conformation of the wild-type protein is conserved (one wouldn't want to introduce mutations that would alter fundamental characteristics of the protein's structure). Though this seems like a simple process, there is a tremendous amount of complexity involved in these two steps. How does one identify the hydrophobic residues to mutate? Do all of them get mutated, or do we let some of them remain? Once we have identified sites to mutate, what residues do we mutate them to? Does everything get changed to lysine, or to glutamic acid?

To redesign membrane proteins, we devised an algorithm that can take an input protein structure and identify solvent-exposed hydrophobic residues. Once a set of residues has been identified for redesign, the next step is to determine what mutations to make at each variable site. The Saven group has developed a probabilistic approach to computational protein design and redesign that I will explain in the next section.

b. Saven's computational design methodology

The computational design methodology is an entropy-based, probabilistic approach for designing proteins. It takes two inputs: a target three-dimensional structure (i.e. the protein to be redesigned), and energy functions that will be used to quantify sequence-structure compatibility. The site-specific probabilities of amino acids at variable positions in a protein are calculated as those that maximize an effective entropy function, subject to constraints on the sequences. To determine these probabilities, the methodology

`

makes use of the theoretical chemistry concept of Effective Sequence Entropy[78,79]. Entropy refers to the amount of "microstates" in a system (i.e. a particular microscopic arrangement of atoms or molecules of the system that corresponds to the given state of the system). For example, if one has a glass jar full of gas molecules and heats it up, the gas molecules will move around more, will occupy more positions in space, and hence will occupy more microstates in the jar. The greater number of microstates that the molecules occupy results in a greater the entropy of the system. This equation is given by the following equation:

$$S = kln(W) \ \ \mathbf{(3.1)}$$

where *W* is the number of microstates and *k* is Boltzmann's constant. This equation shows that entropy is proportional to the number of microstates in a system. For our purposes, we are factorizing the probabilities of amino acid sequences at every site in a protein into this form of the entropy equation. This provides us with an Effective Sequence Entropy function:

$$S = -\sum_{i,\alpha,k} w_i(\alpha, r_k(\alpha)) ln(w_i(\alpha, r_k(\alpha))) \quad \mathbf{(3.2)}$$

where *i* is the site in the protein; α is the amino acid and *k* is a conformation (rotamer) that this amino acid can take on. $wi(\alpha,r_k(\alpha))$ is the probability of a specific amino-acid/rotamer pair at some site *i* in the protein.

Our goal is to maximize the Effective Sequence Entropy function, which will in turn maximize the probability of certain amino acids (at certain rotameric states) at each site. However, we want our Effective Entropy to be maximized such that it provides tolerable amino acid changes in the protein; for example, we would not want to mutate the exterior residues in such a way that the water-soluble protein folds in a different way than the

`

wild-type, or introduce mutations. In order to do this, we use the mathematical method of Lagrange multipliers to constrain two key quantities: the Conformational Energy (E_c)[78,79] and the Environmental Energy (E_{env}). Using these two Lagrange multiplier constraints, the calculated amino acid probabilities at each variable site in the protein will be used to determine mutations at that site that retain the wild type structure while increasing the water-solubility of the protein.

(1) The Conformational Energy constraint (E_c): the overall energy calculated using a molecular potential energy, e.g., the Amber force-field parameters, to approximate the energy of all of the amino acid interactions (i.e. van der Waals interactions, electrostatic interactions and hydrogen bonds) and recover the chemical and shape complementarity usually observed in folded proteins. Mathematically, it is expressed as:

$$E_c = \sum_{i,\alpha,k} \epsilon_i(\alpha, r_k(\alpha)) w_i(\alpha, r_k(\alpha)) + \sum_{i,j>i,\alpha,\alpha',k,k'} \epsilon_{i,j}(\alpha, r_k(\alpha); a', r_{k'}(\alpha')) w_i(\alpha, r_k(\alpha)) w_j(\alpha', r_{k'}(\alpha')) \quad \mathbf{(3.3)}$$

where ε is the energy of interaction. The first term of the equation is the “one-body” term: it quantifies the energy ε_i of an amino acid side-chain α at site i interacting with its own backbone. The second term of the equation is the “two-body” term: it quantifies the interaction energy ε_{ij} that an amino acid i has with amino acid j. The summations in the two-body term are pair-wise across all protein sites (i and j), amino acids at these sites (α and α'), and rotameric states of these amino acids (k and k'). These one-body energies and two-body energies are calculated using the AMBER force-field parameters.

(2) Total Environmental Energy score constraint (E_{Env}): This constraint will be described fully in **section III.** Briefly, due to the presence of solvent-exposed hydrophobic residues, the total E_{Env} score of a membrane protein will differ significantly from that of a water-soluble protein of the same size. While maximizing the total

`

Effective Entropy, the total E_{Env} score will be constrained to have a value consistent with a soluble protein at the same size of the input membrane protein.

c. Total Environmental Energy score (E_{Env}): The Parameters

The Environmental Energy is used to quantify the hydrophobicity of a given site in a protein. The key parameters for calculating it are the local beta carbon density (C_β) and the solvent-accessible surface area of an amino-acid side-chain[80,81] (V_{access})

I. C_β Density

The local beta carbon density, ρ, approximates how buried a particular residue is given its location within the protein:

$$\rho(\alpha) = \frac{n_\beta}{V_{sphere} - V_{access}} = \frac{n_\beta}{\frac{4}{3}\pi R_c^3 - V_{access}} \quad \mathbf{(3.4)}$$

In this equation we are taking a given amino acid side-chain and placing it within a theoretical 8 Å sphere that is centered at the side-chain's geometric center of mass (**FIG 3-2**). This sphere represents that amino acid's "local environment." Beta Carbons belonging to nearby residues in the proteins will be contained within this sphere. Hence, the density of beta carbons within this sphere provides a quantitative readout for how buried or exposed a given residue is.

`

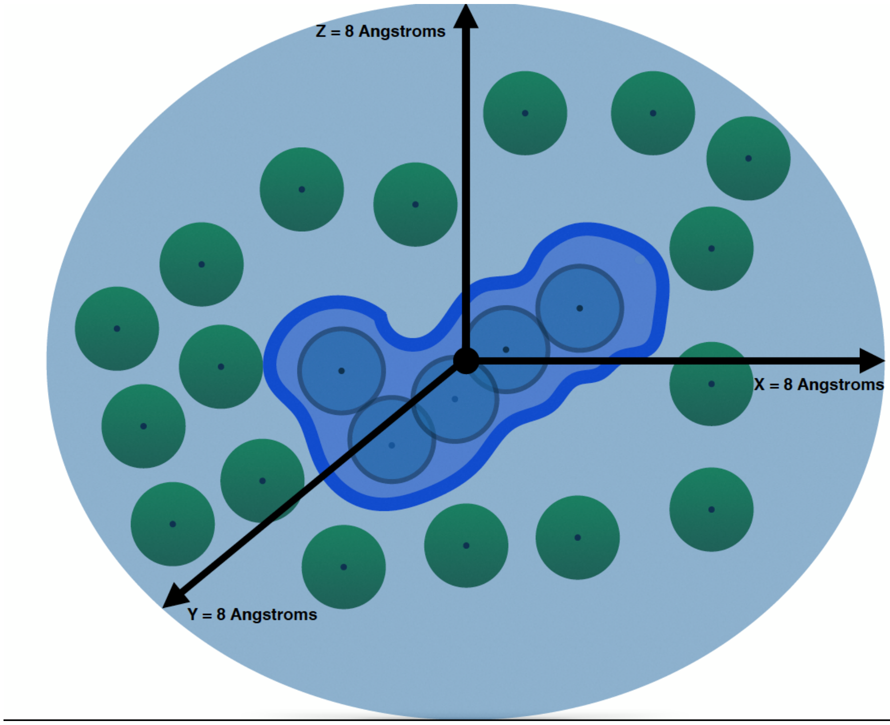


FIGURE 3-2: The CB Density of an amino-acid side-chain

The CB Density, ρ, of an amino-acid side-chain (represented in blue) is calculated by dividing the number of beta carbons (nβ, represented in green) in it's local environment (defined as V_{sphere}, an arbitrary 8 Å sphere centered at α's geometric center) by the volume of the sphere that is taken up by α itself (accessible volume, or V_{access}).

`

d. Calculation of solvent-accessible surface area (V_{access})

The volume of an amino-acid side-chain will vary according to its different rotameric states. To approximate the volume, we made use of a grid-approximation method that was pioneered by Lee and Richards. In this method, a small methyl-probe "rolls around" the surface of a side-chain. The volume that the probe is not able to access is used to approximate the total volume of the side-chain. To calculate V_{access}, we take our amino-acid side chain and place it in the center of a large grid with sides of length *A* (**FIG 3-3**). This grid consists of *M* voxels placed in uniformly-spaced grid points. Since the total volume of the grid is A^3, the volume of one voxel, denoted by $\delta_{V,}$ is given by $\frac{A3}{M}$.

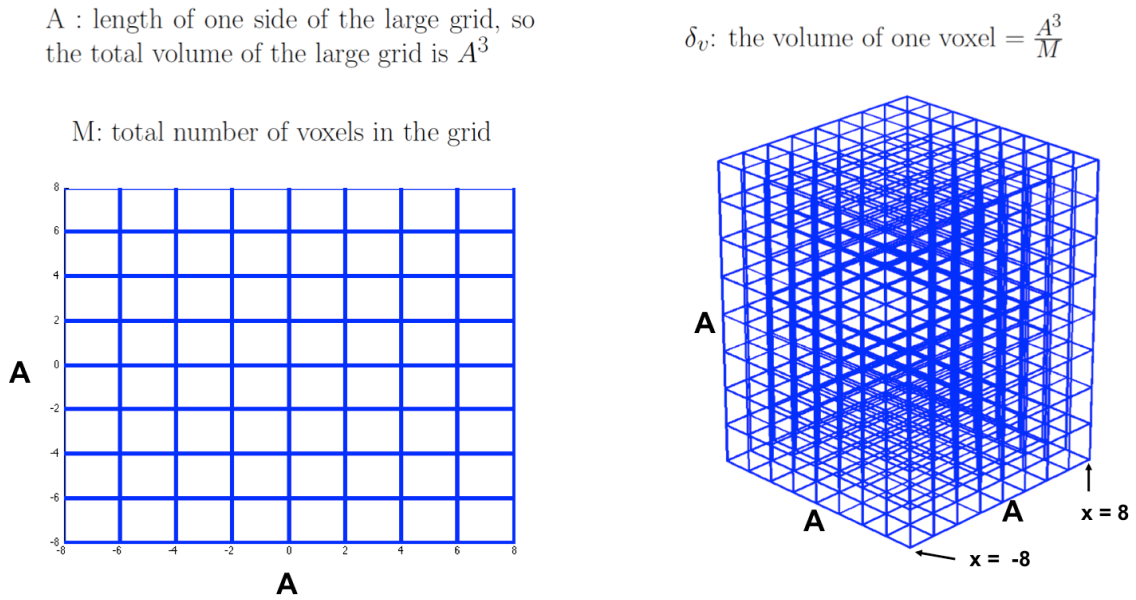

A : length of one side of the large grid, so the total volume of the large grid is A^3
M: total number of voxels in the grid
δ_v: the volume of one voxel $= \frac{A^3}{M}$
8
6
4
2
0
-2
-4
-6
-8
-8
-6
-4
-2
0
2
4
6
8
A
A
A
A
A
x = 8
x = -8

FIGURE 3-3 Structure of the grid used for V_{access} calculation

The protein-side chain is placed in a large grid, that that is used to for the calculation of the side-chain's solvent-accessible surface area, V_{access}

`

The procedure for calculating V_{access} is illustrated in **FIG 3-4**. In this figure, $\vec{r}_\alpha$ are the coordinates for voxel *α*, $\vec{r}_i$ are the coordinates for side-chain atom *i,* and $\vec{r}_b$ are the coordinates of the methyl probe *b* that will "roll-around" the side-chain to determine it's solvent accessible surface area. The distance between *α* and *i* is then given by

$$|\overrightarrow{r}_\alpha - \overrightarrow{r}_i| \quad \mathbf{(3.5)}$$

If σ_i is the radius of atom *i*, and σ_β is the radius of the methyl probe, then the distance between the center of the methyl probe and the center of the side-chain atom *i* is given by

$$\sigma_\beta + \sigma_i \quad \mathbf{(3.6)}$$

Total number of voxels interior to the residue, N_{int}, will then be given by the step function:

$$N_{int} = \sum_\alpha \sum_i \Theta(\sigma_\beta + \sigma_i) - |\overrightarrow{r}_\alpha - \overrightarrow{r}_i|) \quad \mathbf{(3.7)}$$

If the distance between the C_β probe and the side-chain atom *i* is greater than the distance between the voxel *α* to atom *i,* then the voxel will count as being interior to the residue. By multiplying N_{int} with the volume of one voxel, δ_V, an estimate of the interior volume of the residue is given.

`

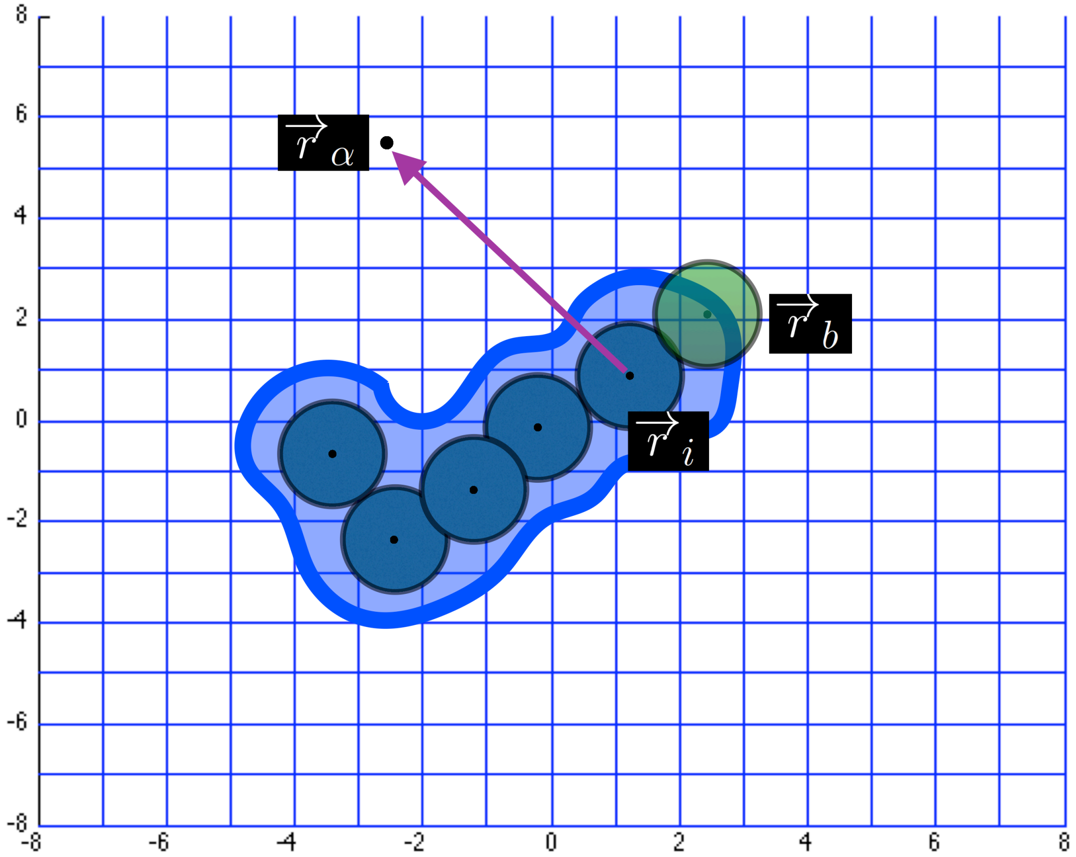


FIGURE 3-4: Grid approximation method for calculating V_{access}

Grid-approximation method for calculating a side-chain solvent accessible surface area, V_{access}

The distance between a probe atom and a given amino-acid side-chain atom is calculated. If the distance between one of the small boxes (voxels) on the grid and the side-chain atom is less than the distance between the probe and the side-chain atom, then the voxel can be considered to occupy space within the side-chain. By repeating this procedure for all of the side-chain atoms, and adding up the volumes of the voxels that occupied space within the side-chain, the side-chain's total volume can be approximated.

`

3.4. DEVELOPMENT OF THE ENVIRONMENTAL ENERGY (E_{ENV}) MODEL

Using the C_β density as the key parameter for estimating whether a residue is buried or exposed, we wanted to determine if there existed *quantifiable* differences between each of the twenty amino acids with respect to there location in a water-soluble globular protein's structures. If we were somehow able to analyze the structures of every single water-soluble globular protein in existence, would there be amino acids that we would almost always observe to be buried, or almost always observe to be exposed to solvent? Our sample for water-soluble globular proteins is a training set of 423 water-soluble globular proteins. This training set consists of x-ray crystallographic structures from the protein classes that are known to function in the cytosol (hydrolases, transferases, isomerases, ligases, oxidoreductases); which have only one chain; which have ≤ 2 Angstrom resolution; and which have a chain length > 40 **(FIG 3-5)**. Furthermore, to prevent one family of proteins from biasing our model, we stipulated that the maximum sequence identity between any two sequences as ≤ 30%. I calculated ρ for every amino acid in this training set, and used these values to generate

corresponding Environmental Energy scores via

$$\epsilon(\alpha, \rho) = -ln(\frac{f(\alpha,\rho)}{f(\alpha)f(\rho)}) \quad \mathbf{(3.8)}$$

Here, *ε(α,ρ)* is the environmental energy score ε for an amino acid α that has a beta carbon density *ρ*. *ε* is calculated by dividing the frequency of times *α* and *ρ* are observed together, *f(α,ρ),* by the frequency of times they are observed independently (given by the product of *f(α)* and *f(ρ)*). *f(α,ρ)*, *f(α)* and *f(ρ)* were calculated from the training-set. A visual explanation of *ε(α,ρ)* is given in **FIG 3-6.** A negative score indicates that the amino acid has a strong preference for a particular value of ρ.

`

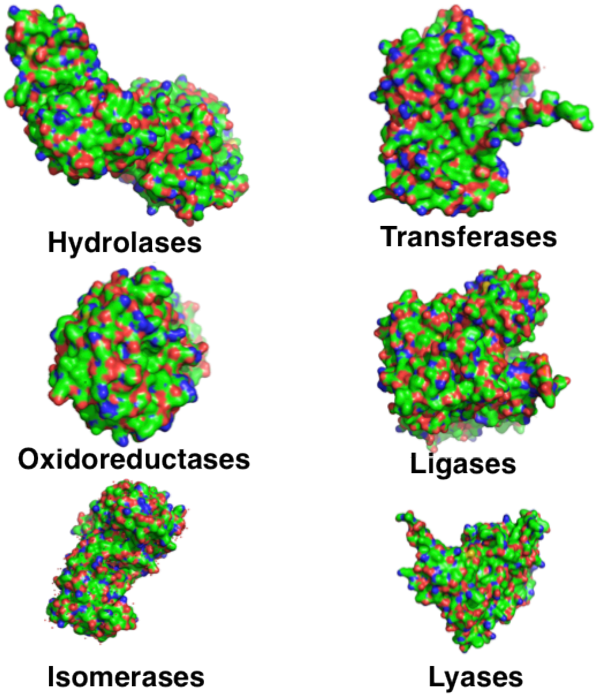


FIGURE 3-5: Learning from a set of globular proteins

To ensure that our EEnv model was not biased by one family of proteins (which would have similar structural features), we carefully selected a diverse array of proteins meeting the following critiera: containing chain; $\leq$ 2 Angstrom resolution; chain length $\geq$ 40; and percent similarity between any two structures as $\leq$ 30%. Using this training set of 423 protein structures, we sought to determine (a) the statistical propensity for each amino acid type to be found at a specific location in water-soluble protein; and (b) the amino acid types that are, statistically, more likely to be found exposed on the exterior and buried on the interior.

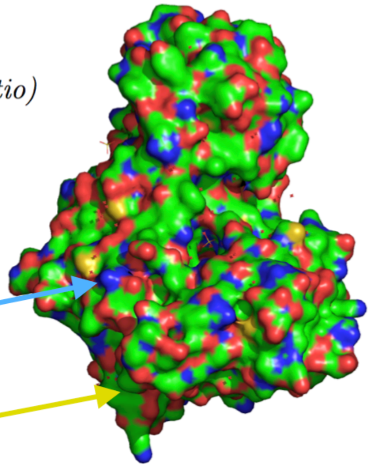
`

$$\epsilon(\alpha, \rho) = -ln(\frac{f(\alpha,\rho)}{f(\alpha)f(\rho)}) \quad \textit{(log-odds ratio)}$$

where:
α = an amino acid type (i.e. **LYS** or **LEU**)
ρ = solvation status (i.e. **BURIED** or **EXPOSED**)

ε(**LYS**, **SOLVENT EXPOSED**) = NEGATIVE VALUE

ε(**LEU**, **SOLVENT EXPOSED**) = POSITIVE VALUE



Hexokinase

FIGURE 3-6: Example of E$_{Env}$

The EEnv is the logarithm of the ratio between the frequency by which one would observe an amino acid alpha at some solvation state, and the frequency by which one would observe that amino acid and that solvation state independently. For example, suppose that there are only two solvation states, buried and exposed, and two amino acids, LYS and LEU. If one looked across the training set of globular proteins, they would find that frequency of LYS in solvent-exposed areas is greater than the product of the frequency of observing LYS and observering a solvent-exposed site; due to the negative logarithm, this value for the (LYS,solvent exposed) pair will be negative. The opposite will be said for LEU in solvent-exposed regions: we would expect the (LEU,solvent-exposed) pairing to be a rare occurance. Hence, f(LEU,solvent-exposed) < f(LEU)f(solvent-exposed), and the corresponding EEnv(LEU,Solvent-exposed) will be positive.

`

For each amino acid in the training set, we generated scoring values for every value of ρ that was observed from the training set and fit these data to polynomial regression curves (**FIG 3-7**). This yielded site-specific scoring functions for amino acids and their propensities to be in particular local environments, where the propensities are derived from the protein training set. These regression curves demonstrate that for the "hydrophobic" residues (green), maxima are obtained at low C_β Densities (i.e. when the residues are exposed to solvent), and minima are observed at high C_β densities (when the residues are buried in the protein core). The reverse holds true for the "hydrophilic residues." Our E_{env} models for each amino acid (taken at a value of ρ that indicates a buried state) correlated well with two widely-used amino-acid hydrophobic scales (**FIG 3-8**). Mean fractional area loss (MFAL)[82] measures the volume of an amino-acid side-chain R that is buried when it goes from a standard-state (calculated from a theoretical GLY-R-GLY tripeptide) to a folded-protein state, while the Fauchere[83] hydrophobic scale measures the partitioning of R (within an N-acetyl-amino acid amide, $CH_2CO\text{-}NH\text{-}CH(R)\text{-}CONH_2$) between octanol and water. The graphs in **FIG 3-7**, and the correlation with known hydrophobic scales, validate our choice of C_β Density as a parameter, and demonstrate that the Environmental Energy model provides a sensitive approach for quantifying the propensity of an amino acid to exist in a particular location in a protein.

`

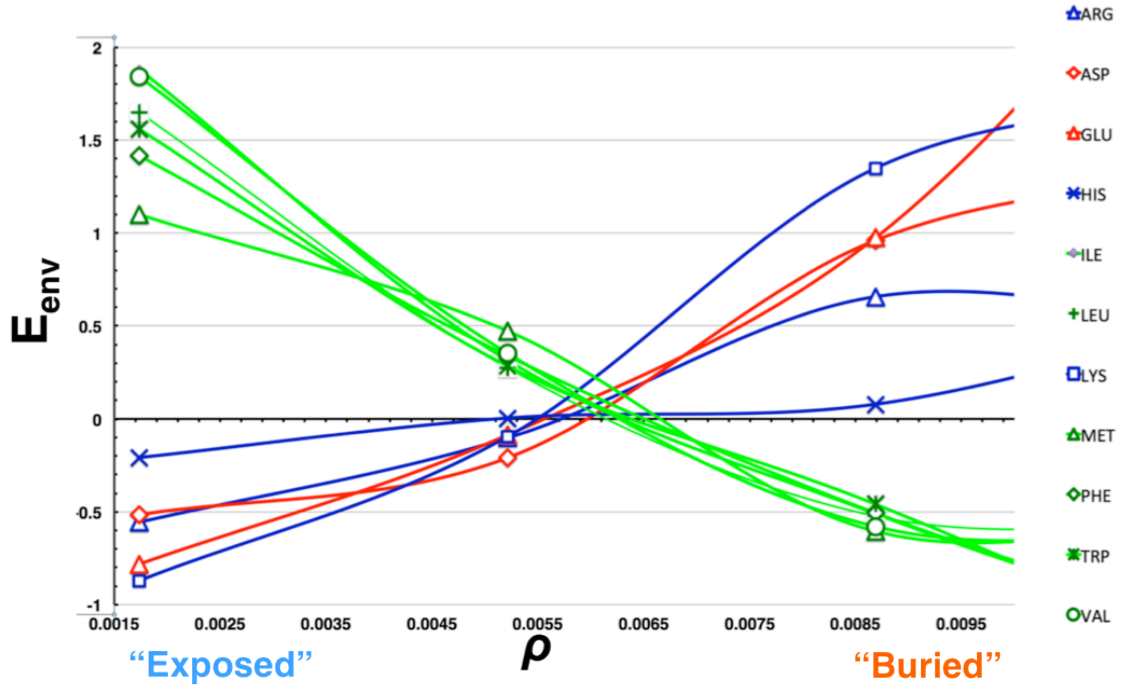


FIGURE 3-7 The E_{env} model for all 20 amino acids

By calculating the environmental energy score for every amino acid across a range of beta carbon densities, and fitting these data polynomial regression curves, we observe that the residues that we conventionally consider to be hydrophobic (TRP, PHE, MET, ILE, LEU, VAL) have their environmental energy minima when they are present in local environments that have a high beta carbon density (i.e. in "buried" states); the inverse is true for the hydrophilic residues (ARG, LYS, ASN, GLN, GLU).

`

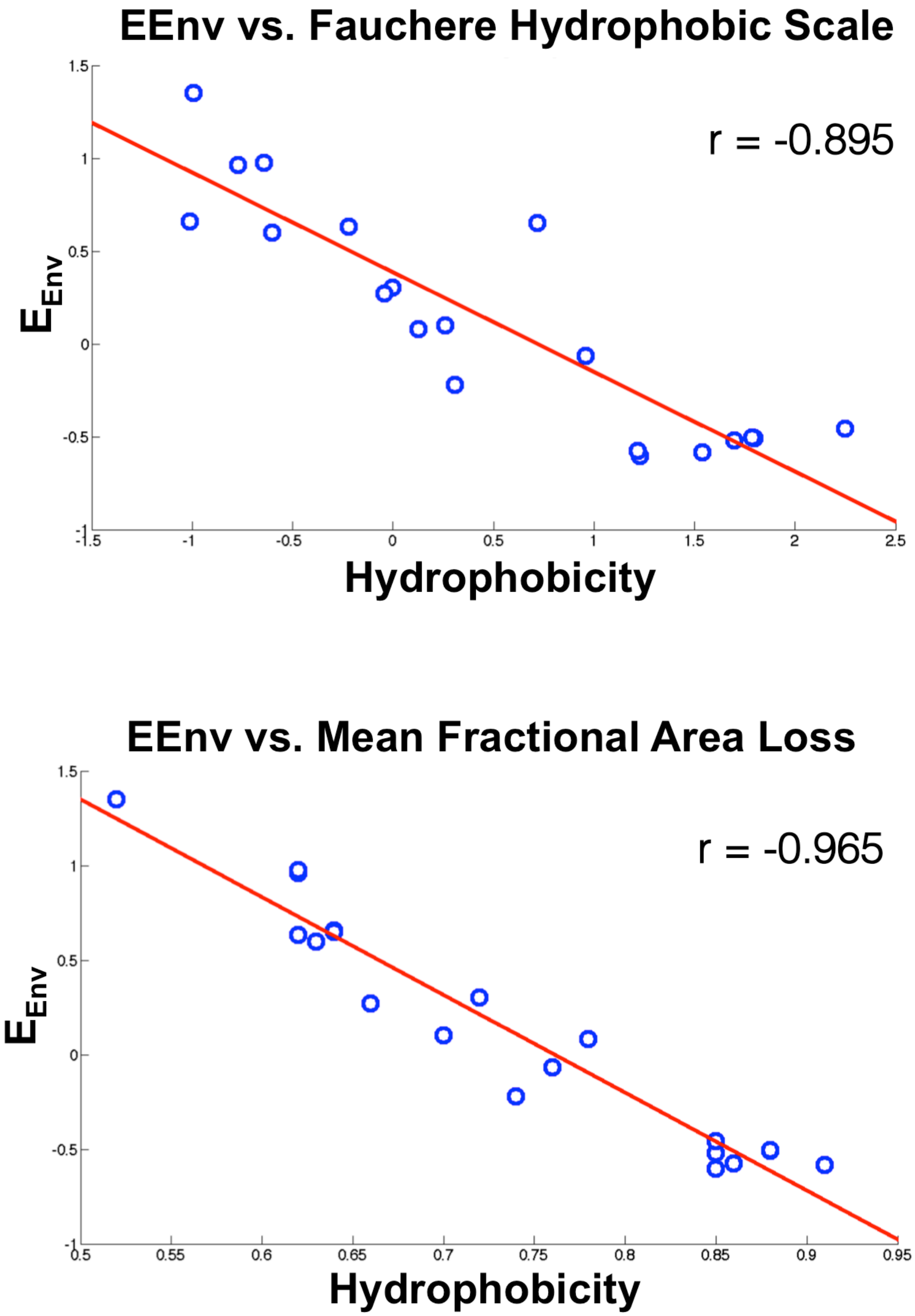


FIGURE 3-8: E_{Env} vs Mean Fractional Area Loss and the Fauchere hydrophobic scale

`

From the training set, we can quantify the *total* E_{Env} score for a protein by summing up the E_{Env} scores at every site. These summed scores can then be plotted against the protein’s chain-length (**FIG 3-9**). A linear relationship was found between the protein's chain length and its Environmental Energy score. As we will demonstrate in **Section 3.V**, this graph and the preceding Environmental Energy vs. CB Density regression curves graph will be pivotal for our membrane protein redesign efforts.

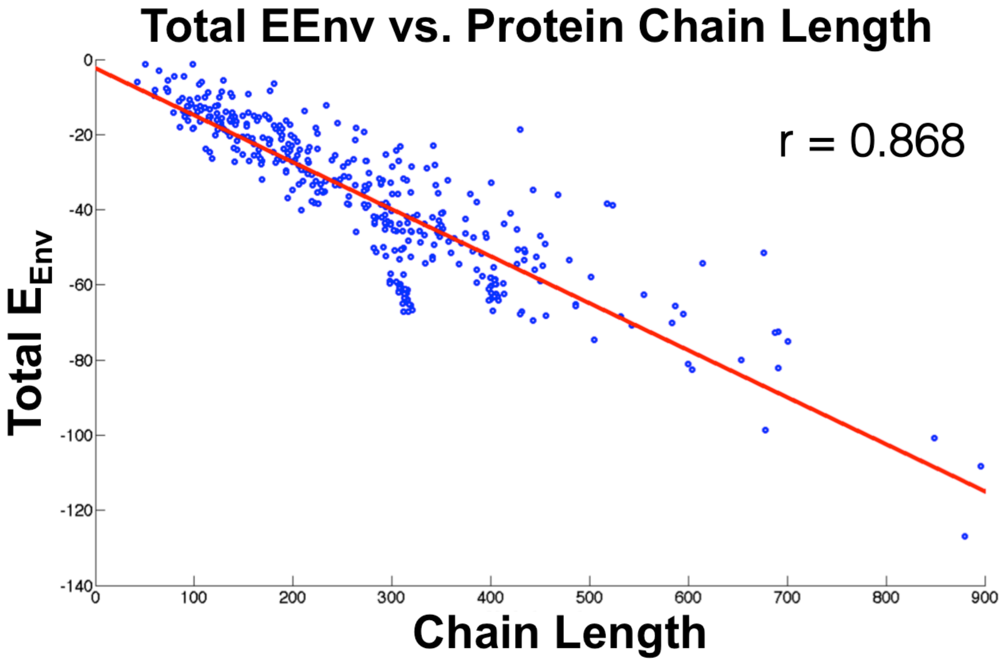


FIGURE 3-9: E_{Env} vs Chain Length for the training set proteins

Using the E_{Env} models for every amino acid, one can calculate the total E_{Env} score for a protein by calculating the environmental score for every residue in the protein and summing the scores up. For the training set protein structures, these values were plotted against the total chain length. In globular proteins, most of the hydrophilic residues tend to be solvent exposed and most of the hydrophobic residues tend to buried at the core, resulting in a total E_{Env} that will be negative for a given protein (a trend that scales linearly with the size of the protein in the training set).

`

3.5. Identifying candidate sites to mutate to confer water-solubility

The following described an algorithm for identifying solvent-exposed, hydrophobic residues in a membrane protein that we want mutate in order to confer solubility to the entire protein. This algorithm takes as input a membrane protein PDB structure, and locations of the transmembrane domain residues, and outputs candidate solvent-exposed hydrophobic sites to mutate to confer water-solubility. These sites are then input into the Saven lab protein design methology to output an amino-acid sequence that is consistent with a water-soluble structure. In this section, this process will be illustrated with the GPCR Smoothened.

a. C_β Density criteria for solvent-exposure criterion

The C_β density was used to determine solvent-exposed residues. To determine the CB density cut-off to use, GETAREA[84] was used to determine the solvent accessible surface area (SASA) for each residue in the training set of 423 water-soluble globular proteins. This algorithm determines the residues that are buried and the residues that are solvent exposed based on a method that geometrically approximates the solvent accessibility of an atom based on the surface area of adjacent atoms that overlap it (where the atoms are represented as spheres). This analysis provided a set of residues that the GETAREA algorithm determined as “solvent exposed” (f >50%). f obtained by calculating the SASA score for a given amino acid α within the reference structure, and dividing it by the calculated SASA score of α that is contained within a Glycine-α-Glycine tripeptide reference structure (an isolated, capped amino acid). f can be formulated mathematically as:

$$f = \frac{SASA(\alpha, protein)}{SASA(\alpha, reference\ structure)} \quad \mathbf{(3.9)}$$

`

Using this ratio, a set of residues were determined as “buried” (f < 30%). For each of these sets, histograms were developed of the GETAREA SASA score as a function of the C_β density. These histograms show that there is a range of C_β densities, $0.004 < \rho_{C\beta} < 0.008$, that is shared and intermediate between the distributions of buried residues and the distribution of solvent exposed residues **(FIG 3-10).** We decided to use a $\rho_{C\beta} < 0.0061$ value as our criteria for solvent exposure; this choice subtends most of the residues classified as solvent exposed. Hence, residues with C_β density < 0.0061 were considered as solvent-exposed.

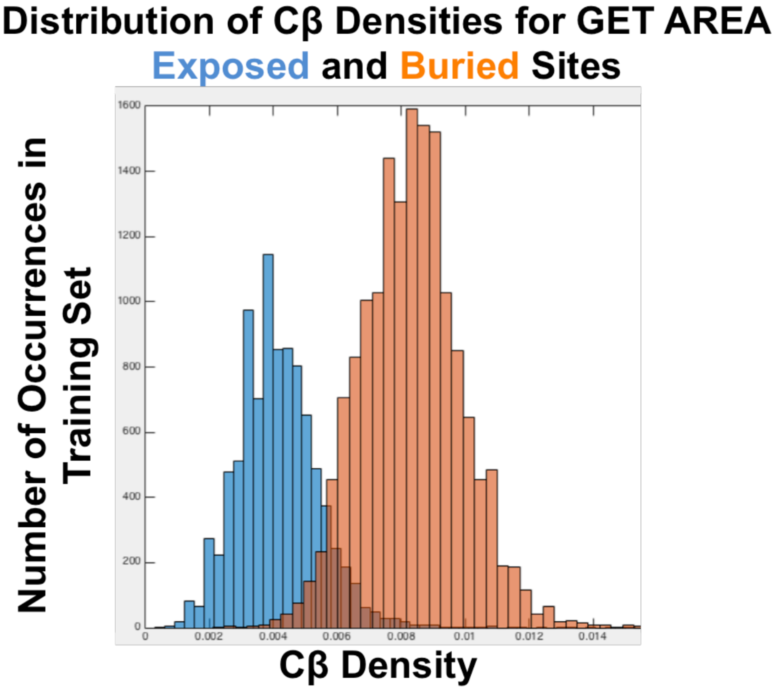


FIGURE 3-10: Relationship between calculated C_β density and GETAREA score.

The set of amino-acid residues that were defined as solvent-exposed by GETAREA are colored in blue, while the set of residues defined by GETAREA as buried are colored in orange.

`

b. E_{Env} Criteria

Based on **Equation 4,** a site in a protein with a positive $E_{Env}(\alpha) > 0$ value contains an amino acid residue (α) that is statistically *unlikely* to be present at that location if that protein were a water-soluble, globular protein. Of the residues that were identified as solvent-exposed via the C_{β} value cut-off in **Criterion 1**, we selected the residues in this set that had positive $E_{Env}(\alpha) > 0$ scores.

c. Amino Acid Identify Criteria

Though the criteria **1** and **2** provided us with solvent-exposed hydrophobic residues, there are certain residues that we did not want to mutate for structural reasons, given our goal of retaining as much of the wild-type protein's inherent structure as possible. For example, Proline and Glycine form structurally important kinks in helices, and mutating those residues may cause structural alterations in the core transmembrane-region. Alanine is found at a high frequency in α-helices, and hence which may confer a degree of helicity to the transmembrane regions. For this reason, we excluded sites with these residues as candidate sites to mutate for our solubilization efforts.

d. Consistent Residues in different conformations criteria

To date, there exist four SMO x-ray crystal structures in complex with the small-molecules LY2940680 (4JKV)[6]; SANT (4N4W)[7]; SAG (4QIN)7; and ANT (4QIM)[7]. These small-molecules bound to different sites in SMO and were found to stabilize different conformations. We felt that the information from these four structures would be informative for our efforts to identify solvent-exposed, hydrophobic residues. It could be the case that a residue is solvent-exposed in one conformation, yet buried in another conformation. We wanted to avoid a situation where we mutated a residue that appeared to be innocuous and solvent-exposed in one case, but which actually was involved in the

`

protein’s conformational dynamics. Hence, we applied the criteria from Steps 1 to 3 to all four structures individually. Residues that met these criteria in *all four structures* were chosen as candidates sites to mutate.

e. Transmembrane domain criterion

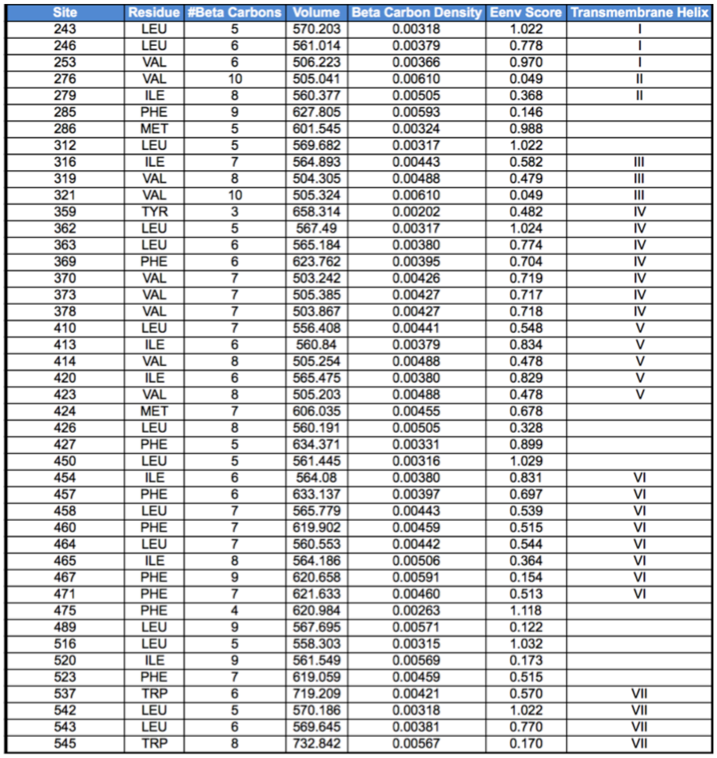
We initially limited our candidate sites to those that are located in the literature-annotated transmembrane domain of SMO (as referenced in the Uni-ProtK website). Applying this criteria to the sites that we identified in steps 1-4 narrowed down our candidate sites to 32. However, upon visually identifying the structure, we noted that there existed solvent-exposed hydrophobic residues that were close to the transmembrane domain boundaries. We identified an additional 12 sites that, while not in the transmembrane domain, met all of the above criteria from steps 1-4. This yielded a total of 44 candidate sites. All of the candidate sites that we identified using Criteria 1-5 were independently identified by the GETAREA algorithm as solvent-exposed hydrophobic sites. This validated our algorithm’s selection of these candidate sites.

f. Computational redesign of protein exterior to confer solubility

We selected the above sites for mutation and a protein with all of the 44 sites selected for mutation **(FIG 3-11).** For these designs, the *β value* was constrained to a value of 0.5; the E_{Env} score was constrained to a value of -35.0, consistent with the size of a water-soluble protein with the same size as SMO **(FIG 3-12)**; and the total net charge of the protein was constrained to a value of 0 and the total net charge of the protein was constrained to a value of 0. The calculated amino acid probabilities at each variable site in the protein were used to determine mutations at that site that retain the wild type structure while increasing the water-solubility of the protein. The specific identified sequence is shown in **FIG 3-13**, with their alignment to the wild-type structure. A

`

rendering of SMO in the wild-type form (left) and the computationally designed water-soluble (wsSMO), is given in **FIG 3-14**.



Site	Residue	#Beta Carbons	Volume	Beta Carbon Density	Eenv Score	Transmembrane Helix
243	LEU	5	570.203	0.00318	1.022	I
246	LEU	6	561.014	0.00379	0.778	I
253	VAL	6	506.223	0.00366	0.970	I
276	VAL	10	505.041	0.00610	0.049	II
279	ILE	8	560.377	0.00505	0.368	II
285	PHE	9	627.805	0.00593	0.146	
286	MET	5	601.545	0.00324	0.988	
312	LEU	5	569.682	0.00317	1.022	
316	ILE	7	564.893	0.00443	0.582	III
319	VAL	8	504.305	0.00488	0.479	III
321	VAL	10	505.324	0.00610	0.049	III
359	TYR	3	658.314	0.00202	0.482	IV
362	LEU	5	567.49	0.00317	1.024	IV
363	LEU	6	565.184	0.00380	0.774	IV
369	PHE	6	623.762	0.00395	0.704	IV
370	VAL	7	503.242	0.00426	0.719	IV
373	VAL	7	505.385	0.00427	0.717	IV
378	VAL	7	503.867	0.00427	0.718	IV
410	LEU	7	556.408	0.00441	0.548	V
413	ILE	6	560.84	0.00379	0.834	V
414	VAL	8	505.254	0.00488	0.478	V
420	ILE	6	565.475	0.00380	0.829	V
423	VAL	8	505.203	0.00488	0.478	V
424	MET	7	606.035	0.00455	0.678	
426	LEU	8	560.191	0.00505	0.328	
427	PHE	5	634.371	0.00331	0.899	
450	LEU	5	561.445	0.00316	1.029	
454	ILE	6	564.08	0.00380	0.831	VI
457	PHE	6	633.137	0.00397	0.697	VI
458	LEU	7	565.779	0.00443	0.539	VI
460	PHE	7	619.902	0.00459	0.515	VI
464	LEU	7	560.553	0.00442	0.544	VI
465	ILE	8	564.186	0.00506	0.364	VI
467	PHE	9	620.658	0.00591	0.154	VI
471	PHE	7	621.633	0.00460	0.513	VI
475	PHE	4	620.984	0.00263	1.118	
489	LEU	9	567.695	0.00571	0.122	
516	LEU	5	558.303	0.00315	1.032	
520	ILE	9	561.549	0.00569	0.173	
523	PHE	7	619.059	0.00459	0.515	
537	TRP	6	719.209	0.00421	0.570	VII
542	LEU	5	570.186	0.00318	1.022	VII
543	LEU	6	569.645	0.00381	0.770	VII
545	TRP	8	732.842	0.00567	0.170	VII

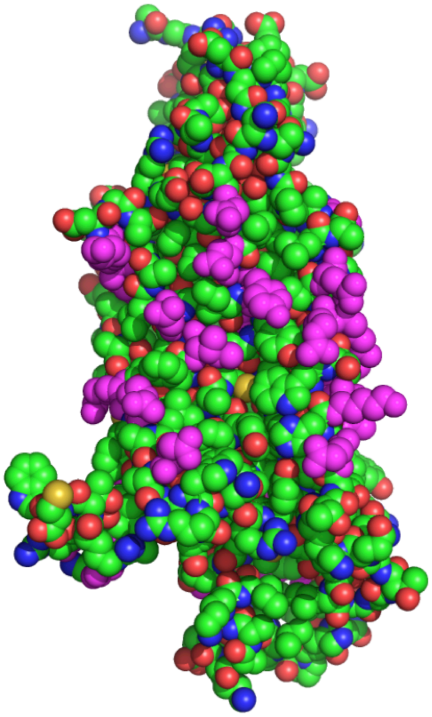


FIGURE 3-11 : **Candidate solvent-exposed hydrophobic sites to mutate that were output by our criteria**

`

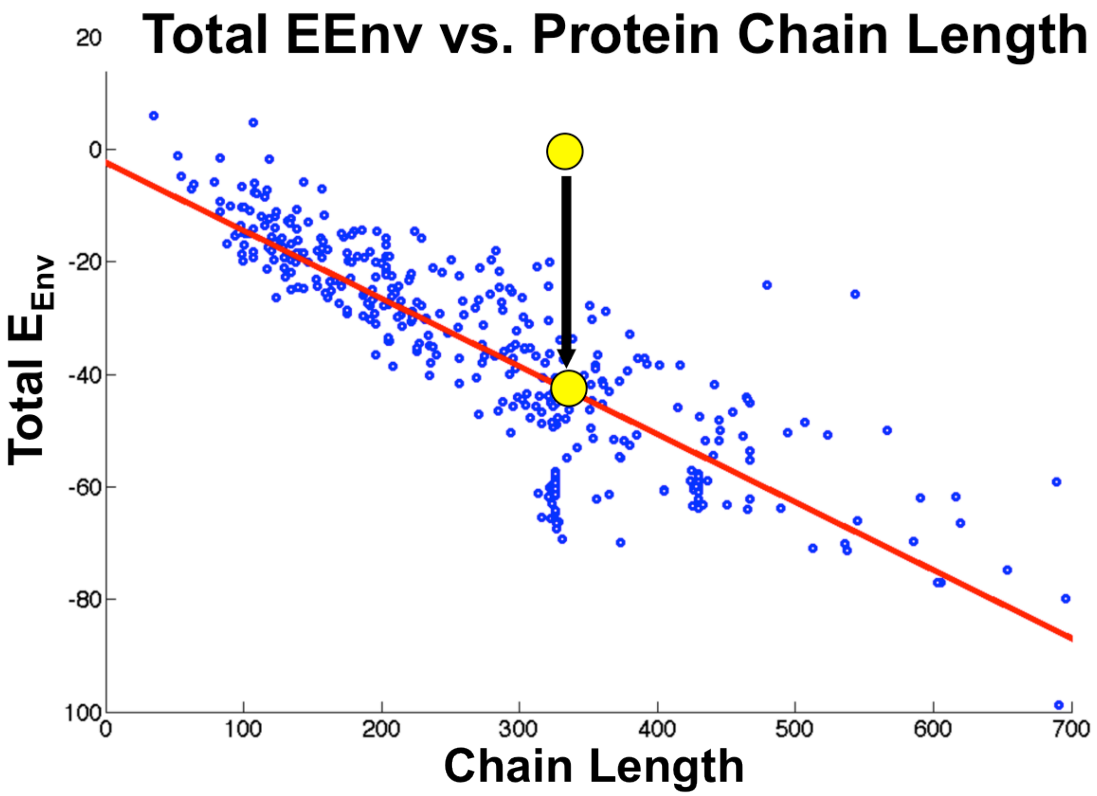
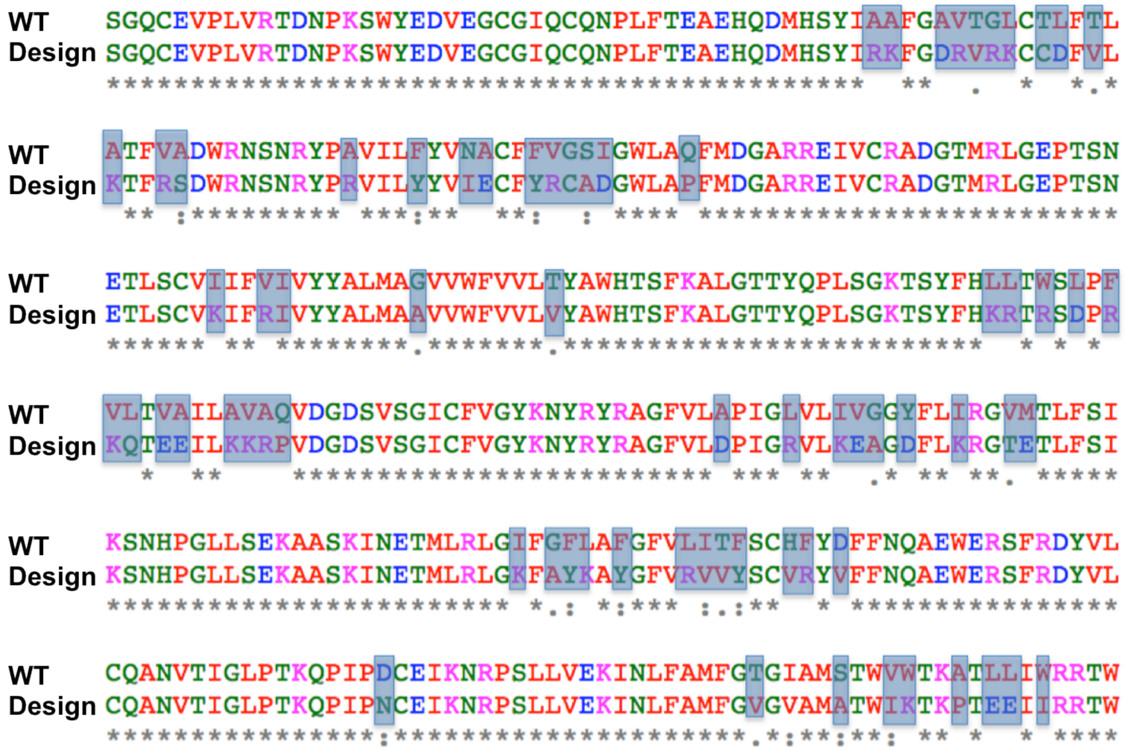


FIGURE 3-12: Total E_{Env} score of Wild-type SMO

Wild-type SMO protein is 346 residues, and the Total E_{Env} score of a training set water-soluble protein with this size is -44.276. Due to SMO's exterior hydrophobic residues, its Total E_{Env} score is 0.921.

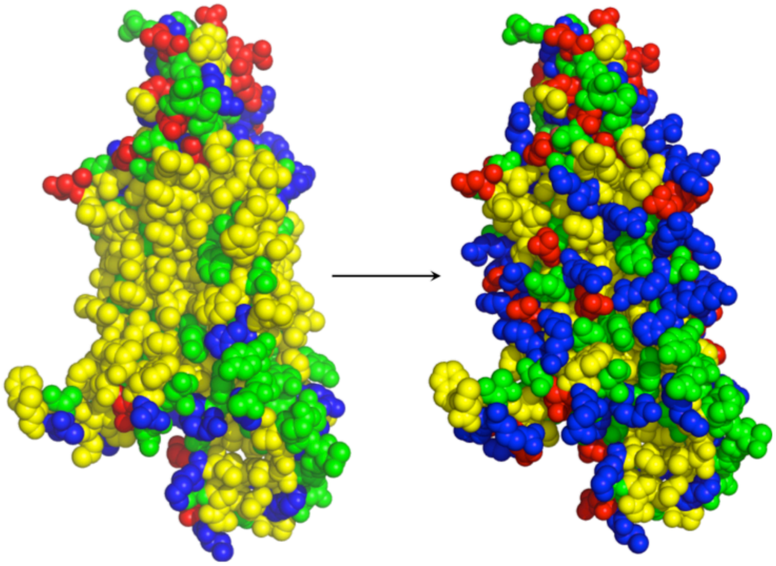
`



```
WT     SGQCEVPLVRTDNPKSWYEDVEGCGIQCQNPLFTEAEHQDMHSYIAAFGAVTGLCTLFTL
Design SGQCEVPLVRTDNPKSWYEDVEGCGIQCQNPLFTEAEHQDMHSYIRKFGDRVRKCCDFVL
       *********************************************  **  .  *  *.*

WT     ATFVADWRNSNRYPAVILFYVNACFFVGSIGWLAQFMDGARREIVCRADGTMRLGEPTSN
Design KTFRSDWRNSNRYPRVILYYVIECFYRCADGWLAPFMDGARREIVCRADGTMRLGEPTSN
        ** :********* ***:**  **:  : **** *************************

WT     ETLSCVIIFVIVYYALMAGVVWFVVLTYAWHTSFKALGTTYQPLSGKTSYFHLLTWSLPF
Design ETLSCVKIFRIVYYALMAAVVWFVVLVYAWHTSFKALGTTYQPLSGKTSYFHKRTRSDPR
       ****** ** ********.*******.*************************  * * * 

WT     VLTVAILAVAQVDGDSVSGICFVGYKNYRYRAGFVLAPIGLVLIVGGYFLIRGVMTLFSI
Design KQTEEILKKRPVDGDSVSGICFVGYKNYRYRAGFVLDPIGRVLKEAGDFLKRGTETLFSI
         *  **    ************************* *** **  .* ** **. *****

WT     KSNHPGLLSEKAASKINETMLRLGIFGFLAFGFVLITFSCHFYDFFNQAEWERSFRDYVL
Design KSNHPGLLSEKAASKINETMLRLGKFAYKAYGFVRVVYSCVRYVFFNQAEWERSFRDYVL
       ************************ *.: *:*** :.:**  * ****************

WT     CQANVTIGLPTKQPIPDCEIKNRPSLLVEKINLFAMFGTGIAMSTWVWTKATLLIWRRTW
Design CQANVTIGLPTKQPIPNCEIKNRPSLLVEKINLFAMFGVGVAMATWIKTKPTEEIIRRTW
       ****************:*********************.*:**:**: ** *  * ****
```

FIGURE 3-13: wsSMO and wild-type SMO protein sequence alignment

`



Site	Wildtype Residue	New Residue
243	LEU	ARG
246	LEU	ARG
253	VAL	GLU
276	VAL	ILE
279	ILE	ASP
285	PHE	PHE
286	MET	LYS
312	LEU	ASP
316	ILE	ASP
319	VAL	TYR
321	VAL	VAL
359	TYR	ASP
362	LEU	LYS
363	LEU	ASP
369	PHE	ARG
370	VAL	LYS
373	VAL	GLU
378	VAL	THR
410	LEU	ARG
413	ILE	THR
414	VAL	ILE
420	ILE	ARG

Site	Wildtype Residue	New Residue
423	VAL	THR
424	MET	GLU
426	LEU	GLU
427	PHE	ARG
450	LEU	LYS
454	ILE	ASP
457	PHE	TYR
458	LEU	GLU
460	PHE	TYR
464	LEU	GLU
465	ILE	VAL
467	PHE	TYR
471	PHE	ASP
475	PHE	ARG
489	LEU	ILE
516	LEU	GLU
520	ILE	ARG
523	PHE	ARG
537	TRP	LYS
542	LEU	GLU
543	LEU	GLU
545	TRP	ILE

FIGURE 3-14 Comparison of wild-type SMO and wsSMO

A summary of what all the sites were mutated to by the Saven protein design methodology is shown in the left table, along with a rendering of both the wild-type and the water-soluble variant.

`

CHAPTER IV: EXPERIMENTAL-CHARACTERIZATION OF WSSMO

In this section, I will describe our efforts to experimentally characterize wsSMO. First, I will the scheme that we used to clone express and purify wsSMO protein. Next, I explain the techniques that we used to verify our protein's identity. Lastly, I will describe the biophysical techniques that we are using to test for wsSMO's ability to bind to ligands that are known to bind to wild-type SMO.

4.1. CLONING, EXPRESSION AND PURIFICATION OF WSSMO

a. Preliminaries: FPLC, and the different columns that were used

Protein purification is performed using a fast-pressure liquid chromatography (FPLC) system. There are four main protein columns that I used:

- Affinity columns: Maltose-Binding Protein, Histidine-Tag, and Anion-exchange columns

A maltose-binding protein (MBP) column is used to isolate proteins that have a maltose-binding protein affinity tag. This column consists of packed dextrin sepharose beads that are coated with amylose resin. When cell lysate is introduced to the column, non-MBP tagged protein will flow-through, and the MBP-tagged species will stick to the amylose. Elution with 10 mM maltose, which will outcompete with amylose resin for binding to the MBP-binding site, will result in the release of MBP-tagged protein from the column. The Histidine-tag column works in the same way, except that its sepharose beads are coated with Nickle ions that will bind to a 6X-Histidine protein tag; elution of the bound protein occurs with 500 mM imidazole. An Anion-exchange column has positively-charged tertiary amine molecules bound to its beads that will bind to negatively-charged proteins; elution takes place along an NaCl gradient (usually from 0 M NaCl to 75 mM NaCl over a

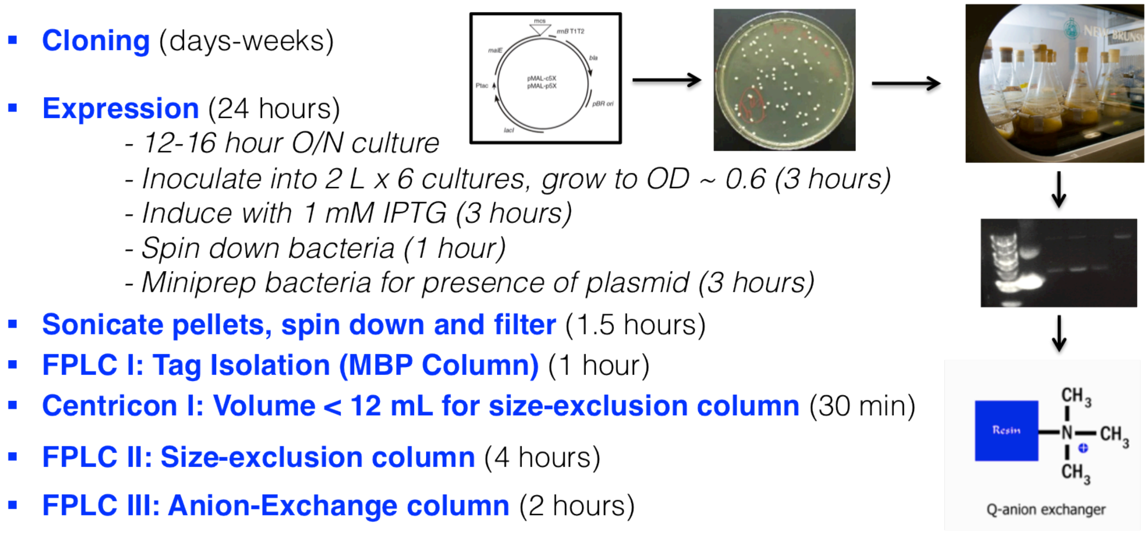
`

50 mL range). Proteins will elute from the column according to their net negative charge at the pH of the elution buffer (pH 8.0): positively charged species will pass through the column and end up in the flow-through.

- Size-exclusion Chromatography

Size-exclusion chromatography separates proteins via their ability to move through a series of porous beads. Small species will be able to fit into these porous locations, and will take longer to come out that a larger protein. Hence, proteins will elute from the column in decreasing order. Due to the sensitivity of this column, a volume less than 12 mL needs to be injected.

The pipeline for cloning, protein expression and FPLC purification is shown in **FIG 4-1**. These steps will be fully-explained in the forthcoming sections.



- **Cloning** (days-weeks)
- **Expression** (24 hours)
  - *- 12-16 hour O/N culture*
  - *- Inoculate into 2 L x 6 cultures, grow to OD ~ 0.6 (3 hours)*
  - *- Induce with 1 mM IPTG (3 hours)*
  - *- Spin down bacteria (1 hour)*
  - *- Miniprep bacteria for presence of plasmid (3 hours)*
- **Sonicate pellets, spin down and filter** (1.5 hours)
- **FPLC I: Tag Isolation (MBP Column)** (1 hour)
- **Centricon I: Volume < 12 mL for size-exclusion column** (30 min)
- **FPLC II: Size-exclusion column** (4 hours)
- **FPLC III: Anion-Exchange column** (2 hours)

FIGURE 4-1: Pipeline for pMAL-MBP-wsSMO protein purification

`

b. Selection of expression vector, cell-line, and growth conditions

Selection of affinity tag

In previous water-solubilization efforts, expression and purification was difficult, and the final proteins needed to be refolded from the insoluble pellets of the *E.coli* protein expression systems we were using. For wsSMO, the maltose-binding affinity tag (MBP) was chosen due to its ability to increase the solubility, rate-of-folding, and yield of recombinant proteins in protein expression systems. The presence of the affinity tag would also us to isolate the protein from the lysate via an MBP column.

Selection of cell-line

Though there are various *E. coli* protein expression systems, we had to be careful about the one we chose due to the presence of disulfide bonds. For many proteins, disulfide bonds have been found to be critical for stability, activity, and proper folding. The structure of Smoothened is notable because it contains 4 extracellular disulfide bonds that have been functionally implicated in SMO activation. In *E. coli*, disulfide bond formation is challenging due to the presence of thiol oxidative proteins in the cytoplasm (i.e. thioredoxin reductase and glutathione reductase); for endogenous *E. coli* proteins, disulfide bond formation takes place in the periplasm. This feature makes it problematic for using *E. coli* to produce proteins with multiple disulfide bonds, as recombinant proteins are expressed in the cytoplasm. As a result, a portion of proteins that contain disulfide-bonds end up being misfolded or poorly expressed in *E. coli*. Misfolded proteins can result in insoluble aggregates that are toxic to *E. coli* cells, leading to their death and effectively decreasing protein yield. In the case where *E. coli* are able to survive after the formation of these aggregates, these aggregates will become part of inclusion bodies upon cell lysis.

`

To help overcome the hurdles associated with over-expressing proteins with disulfide bonds, an *E. coli* cell line was engineered to overexpress disulfide bond isomerase C (DsbC), a protein found in *E. coli* periplasm that catalyzes disulfide bond formation, and which contained mutated and inactive forms of the cytoplasmic thioredoxin reductase (trxB) and glutathione reductase (gor) proteins. This cell line was subsequently commercialized (Shuffle T7 cells, New England Biolabs). We selected this cell line for our large-scale protein purification efforts in order to give us the best chance of obtaining a high-yield of MBP-wsSMO protein in the soluble fraction.

c. Recombinant DNA cloning and pilot protein expression experiment

The cDNA for the designed wsSMO sequence from **Chapter III** was obtained (DNA 2.0, Menlo Park, CA) and cloned into a pMAL-c5X expression vector via the *NdeI* and *Hind III* restriction sites (**FIG 4-2A**). pMAL-c5x contains an isopropyl β-d-1-thiogalactopyranoside (IPTG) inducible promoter flanking the multiple cloning site; the *malE* gene encoding an N-terminal maltose-binding protein (MBP) tag; and the *amp*^{r} drug selection marker. Theoretically, pMAL-MBP-wsSMO expression in *E. coli* would result in the production of MBP-wsSMO fusion protein (**FIG 4-2B**), which can be isolated via an MBP column. The MBP tag can be cleaved by Factor Xa protease (FXa), due to a FXa recognition site in the linker region between MBP and wsSMO in the fusion protein.

Before attempting to purify protein MBP-wsSMO on a large scale, the optimal isopropyl β-d-1-thiogalactopyranoside (IPTG) concentration needed to induce expression of pMAL-MBP-wsSMO in *E. coli* needed to be determined. NEB Turbo Express competent *E. coli* cells were transformed with the pmAL-MBP-wsSMO construct, and the transformants were grown on LB Ampicillin plates at 37^0 C for 12 hours. Resulting colonies were then inoculated into 5mL LB-Amp starter cultures and grown for 12 hours.

200 uL of each culture were then inoculated into a 20 mL culture of LB-Ampicillin. These cultures were grown with 250 RPM shaking at 37^0 C to an optical density of 0.6, and then induced with a range of IPTG concentrations for 3 hours. After the induction period, the cultures were centrifuged, filtered and sonicated.

The soluble fractions and the insoluble fractions (obtained via incubation of the bacterial pellet with 8M urea for 2 hours) corresponding to each IPTG concentration were run on an SDS-PAGE denaturing gel (**FIG 4-3)**. The coomassie-stained gel demonstrates that a concentration of 1 mM IPTG results in the highest expression of protein species near the theoretical MBP-wsSMO fusion protein size of 85 kDA. However, there was also an increase in the expression of protein at ~42 kDA, the size of the MBP tag, as well as other proteins at various sizes. These other proteins could represent native *E. coli* proteins or MBP-wsSMO degradation products.

`

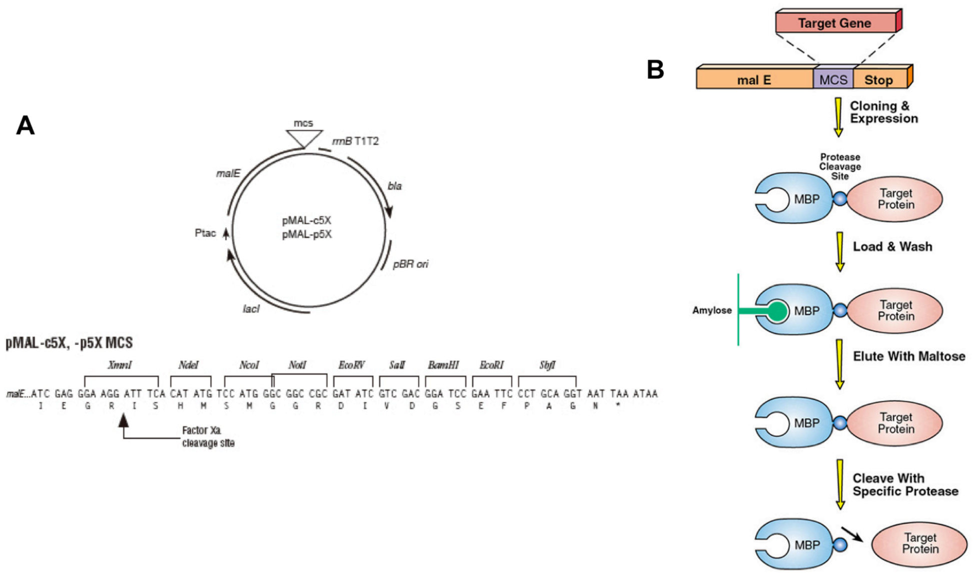


FIGURE 4-2: pMAL expression system

a. Using *NdeI* and *HindIII* restriction sites, wsSMO was cloned into the pMAL-c5x vector. Upstream of the cloning site is a *malE* gene that encodes for MBP protein, resulting in an N-terminal MBP tag on wsSMO connected by a short linker sequence.

b. Expressed of pMAL-MBP-wsSMO in an *E. coli* expression system can theoretically result in the production of MBP-wsSMO protein. In addition to being used an an affinity tag, MBP has been also reported to aid with protein-folding.

`

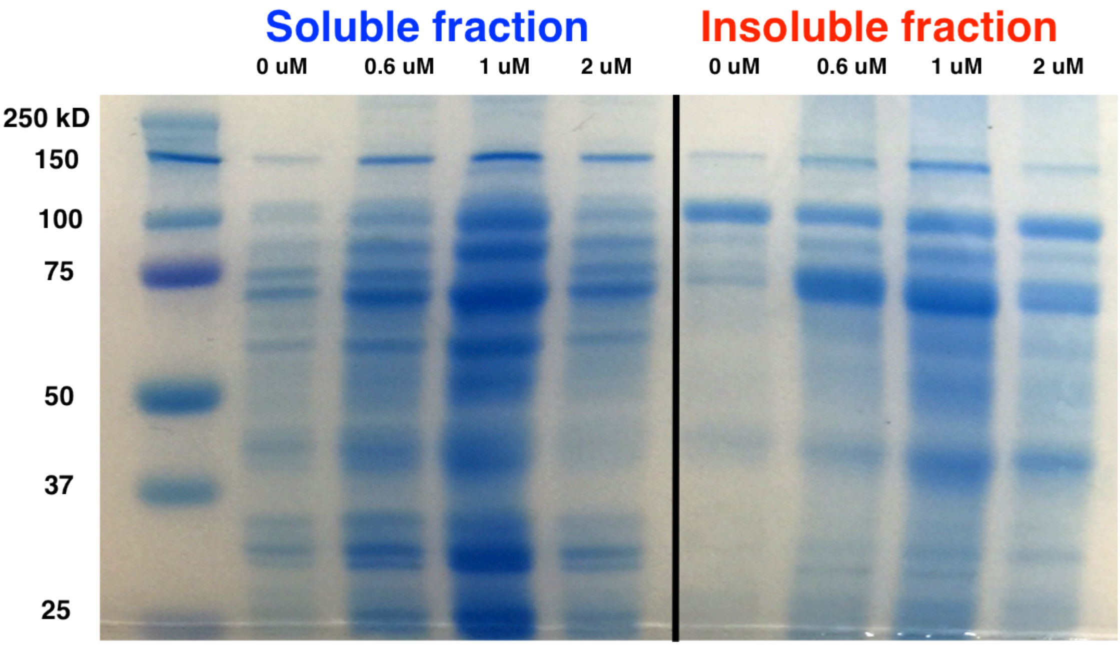


FIGURE 4-3. IPTG induction experiment for optimal MBP-wsSMO expression

A pilot experiment was performed to determine the proper concentration of IPTG for MBP-wsSMO protein expression. We found that 1 mM IPTG gave high signal of protein expression in the region between 100 kDA and 75 kDA, where we expected MBP-wsSMO to be present (theoretical size: 85 kDA). We found protein species that corresponded to this theoretical size in both the soluble and insoluble fractions.

`

d. Protein Purification and Expression of wsSMO

From the pilot experiment, it appeared that 1mM IPTG was the optimal concentration for protein expression. From here, a scale-up experiment to 2 L cultures was performed. pMAL-MBP-wsSMO DNA was transformed into Shuffle T7 competent *E. coli* cells and incubated on LB-Ampicillin plates at 37^0 C for 24 hours. Individual colonies were then isolated and inoculated into 10 mL Terrific Broth-Ampicillin liquid starter cultures at 30^0C with 250 rpm shaking for 16 hours. These starter cultures were then inoculated into 2 L Terrific Broth-Ampicillin flasks and grown with shaking at 30^0 C to an optical density of 0.6. MBP-wsSMO protein expression was induced with 1 mM IPTG for 4 hours. The bacteria were then pelleted and lysed via sonication. To prevent the potential formation of insoluble aggregates from forming due to disulfide-bonds via surface-exposed cysteines, 1 mM DTT was added to the sonication buffer to keep these cysteines in a reduced state.

After sonication, the cell lysate was filtered and run through an FPLC MBP-column (**FIG 4-4A**). Bound protein was eluted with 10 mM maltose, resulting in a prominent elution peak. Running the eluted protein on an SDS-PAGE denaturing gel revealed multiple protein species (**FIG 4-4B)**. To separate these species, the MBP column eluted protein was run through a Superdex 75 size-exclusion column, resulting in a two peaks (**FIG 4-5**). Fractions corresponding to the first peak (C11-D14) contained protein that was near the expected size of MBP-SMO (85 kDA), while the fractions corresponding to the second peak (fractions E1-F2) contained protein near the size of the MBP tag (43 kDA). To separate the fusion protein band from Peak 1 from the other protein bands below it, its fractions were collected and injected into an anion-exchange column (**FIG 4-6**). Individual peaks were then run on SDS-PAGE. Fractions C4-C12 were found to pure protein at the size of MBP-SMO. The range of salt concentrations in this collection was

`

534 mM NaCL to 750 nM NaCl. UV-VIS determination of the collected fractions revealed a yield of 283.61 ug of protein from 2 L of bacteria. Notably, combining fractions B9-C3 (which contained a mixture of our suspected MBP-wsSMO fusion and a contaminant around 60 kDA) revealed a yield of 2.3 mg.

`

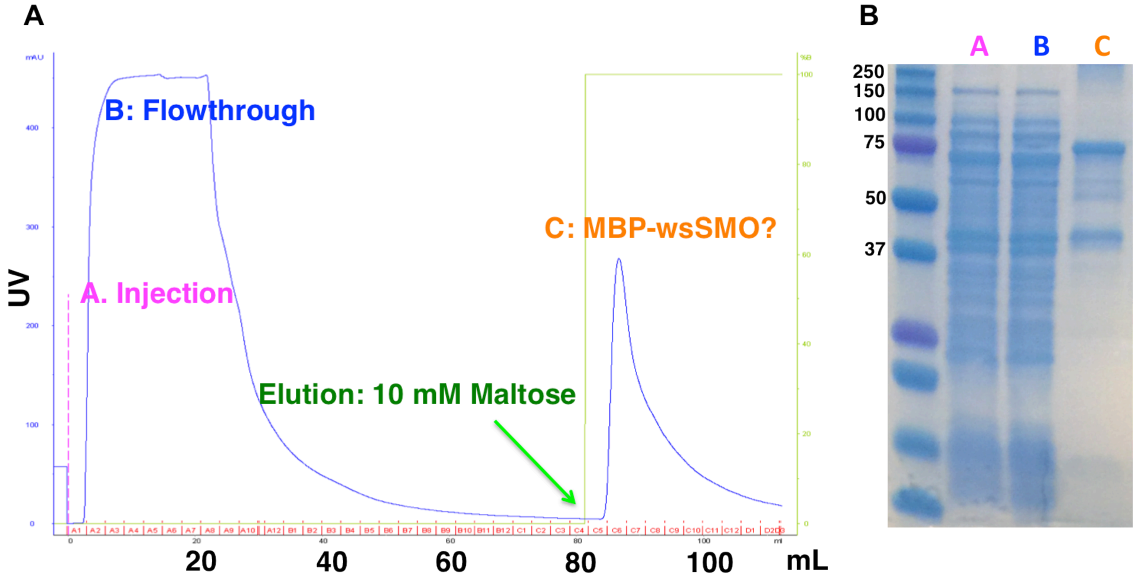


FIGURE 4-4: MBP column purification

A) Results of the MBP column chromotography demonstrate a strong protein signal upon elution with 10 mM maltose, indicating the presence of MBP-tagged protein species.

B) SDS-Page gel of the injection, flow-through and elution peaks. Lane C indicates multiple isolated protein species from MBP column elution, with the top species near to the theoretical size of MBP-SMO (85 kDA)

`

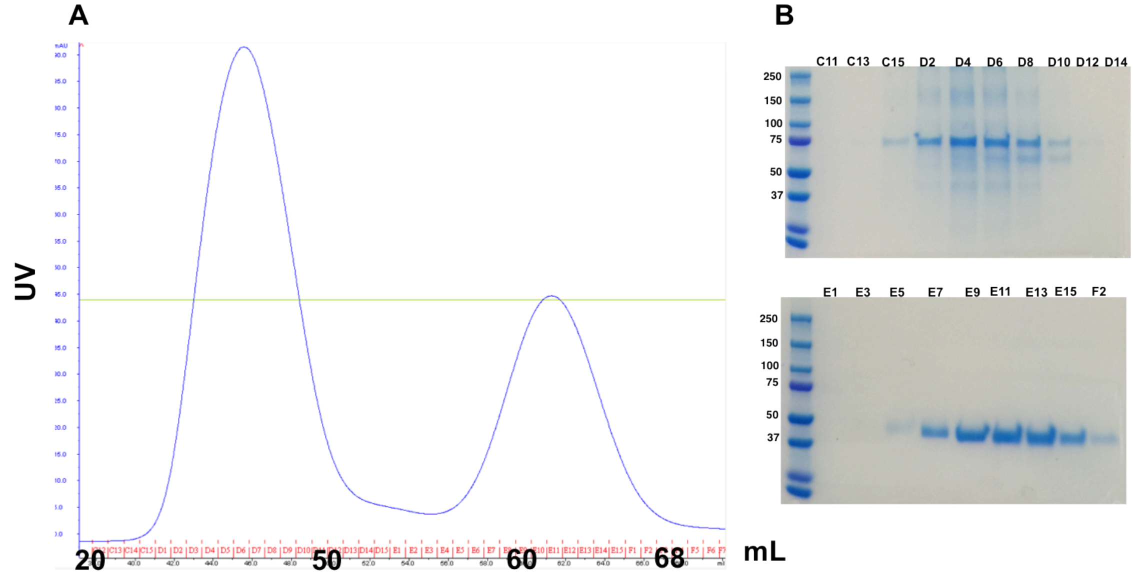


FIGURE 4-5: Size-exclusion column purification of MBP column elution

A: Size-exclusion chromatography of the MBP elution fractions from **FIGURE 4-6** reveals two peaks.

B: SDS-Gel of all the frations reveals that the first peak contains the two larger species, while the second peak contains what is likely the MBP tag (43 kDA)

`

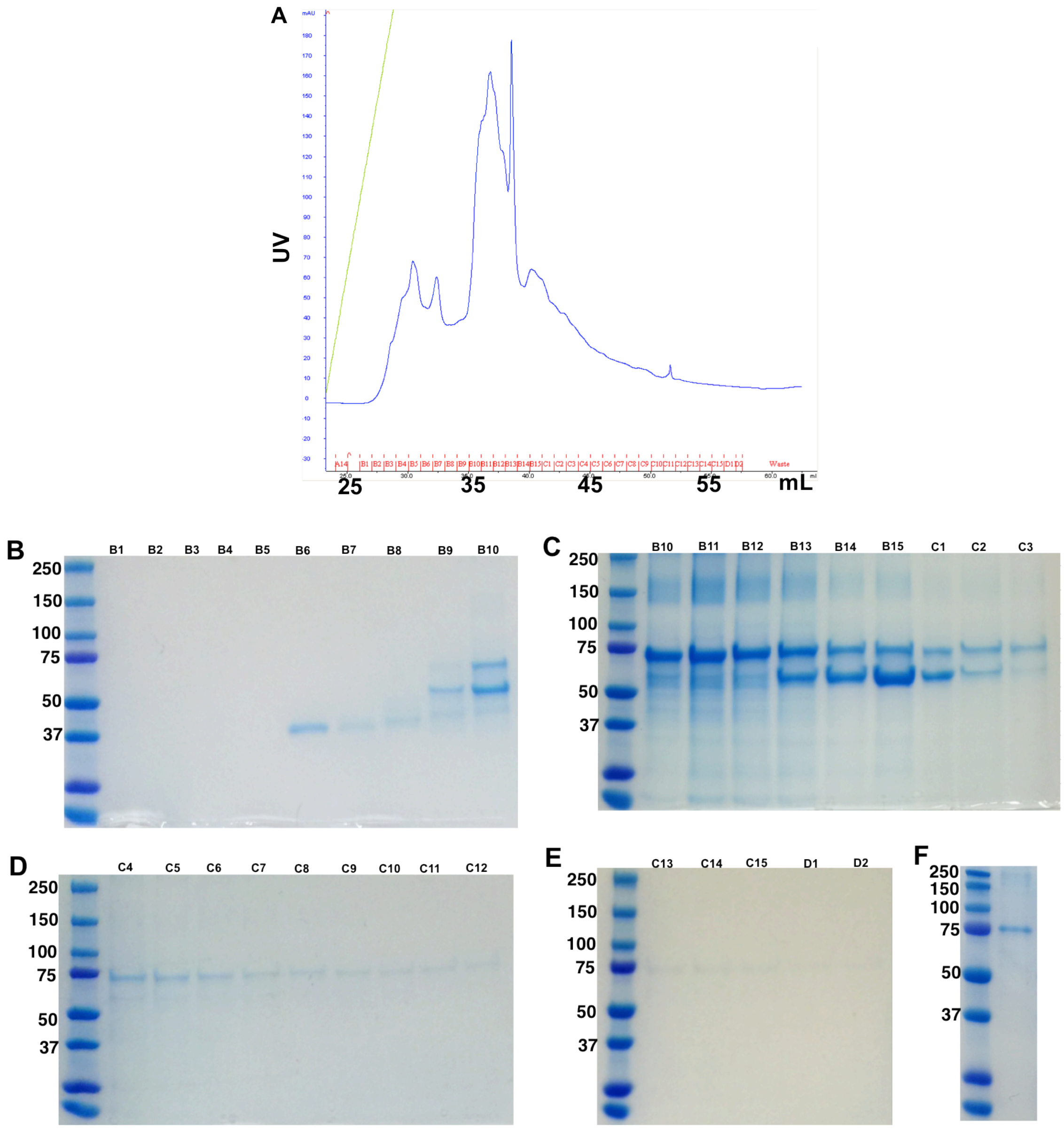


FIGURE 4-6: Anion exchange column of size-exclusion void peak

A: Anion-exchange chromatography to the size-exclusion peak 1 from **FIGURE 4-5** reveals two peaks that are not that well-resolved.

B-E: SDS-PAGE of individual anion-exchange fractions; pure MBP-wsSMO fusion protein was detected in fractions C4-D12

E: SDS-Page of the suspected MBP-wSMO fusion protein, after fractions C4-C12 were collected and centrifuge-concentrated

`

e. Primary and Secondary Structure verification

Secondary structure verification by circular dichroism

Chiral molecules absorb left-handed circularly polarized and right-handed circularly polarized light to different extents, an effect that can be measured as a function of wavelength known circular dichroism (CD). CD can be used to determine the secondary structural elements of proteins, such as alpha helices and beta sheets[86]. The percentages of secondary structural elements in a protein can be detected by observing the far-UV range (190-260 nm). For example, alpha-helical structures exhibit negative signal at 222 and 208 nm, while proteins with anti-parallel beta-sheets have negative bands at 218 nm and positive bands at 105 nm. CD spectra of the isolated fusion protein exhibited strong 209 and 222 minima, consistent with the presence of α-helical content (**FIG 4-7**). This spectra indicates that the protein is folded properly. If it were not folded properly, it would have CD spectra that is consistent with random coil.

`

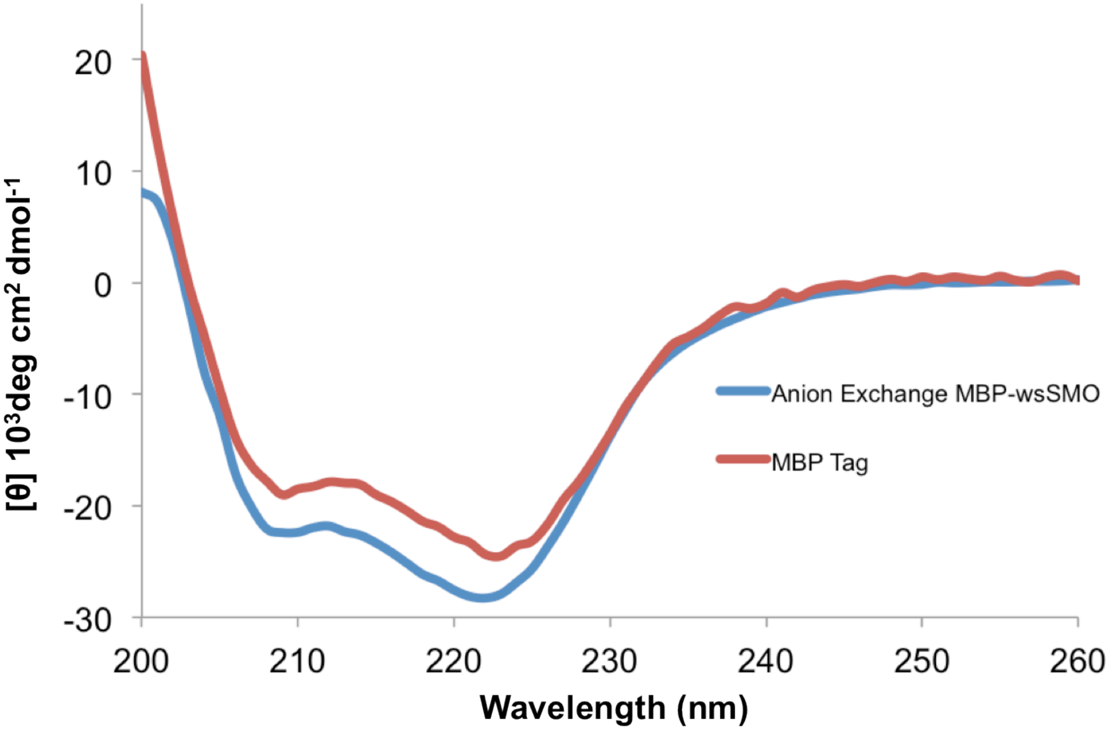


FIGURE 4-7: CD spectra of the isolated protein from FIGURE 4-6

The CD spectra of the isolated MBP-wsSMO protein indicates minima at 209 nm and 222 nm, indicating alpha-helical structure and suggesting that the protein had folded properly. MBP tag's CD spectra is shown for reference.

`

Primary structure verification by M/S proteomic sequencing

We attempted to obtain M/S sequencing on purified, untagged wsSMO. First, Factor XA cleavage was performed directly on the MBP column elution fractions. Factor XA cleavage was performed directly on the MBP column elution fractions (12 hours, 30 degrees celsius), and the resulting cleavage products were injected into the MBP column **(FIG 4-8).** This resulted in prominent flow-through (possible untagged wsSMO) and elution (possible MBP tag) peaks. The flow-through products were then loaded them onto a size-exclusion column, resulting in four distinct peaks **(FIG 4-9A).**

An SDS-gel of these four peaks (**FIG 4-9B**), along with the MBP flow-through and elution products after Factor XA cleavage, revealed the size-exclusion Peak 2 and Peak 4 size-exclusion bands were close to the theoretical size of wsSMO (41.5 kDA). Peak 1 was the void peak, and Peak 3 was likely to be a degradation product. A CD analysis of the size-exclusion Peak 2 and Peak 4 protein samples demonstrated that both species had strong α-helical signal, indicating that both are properly folded.

`

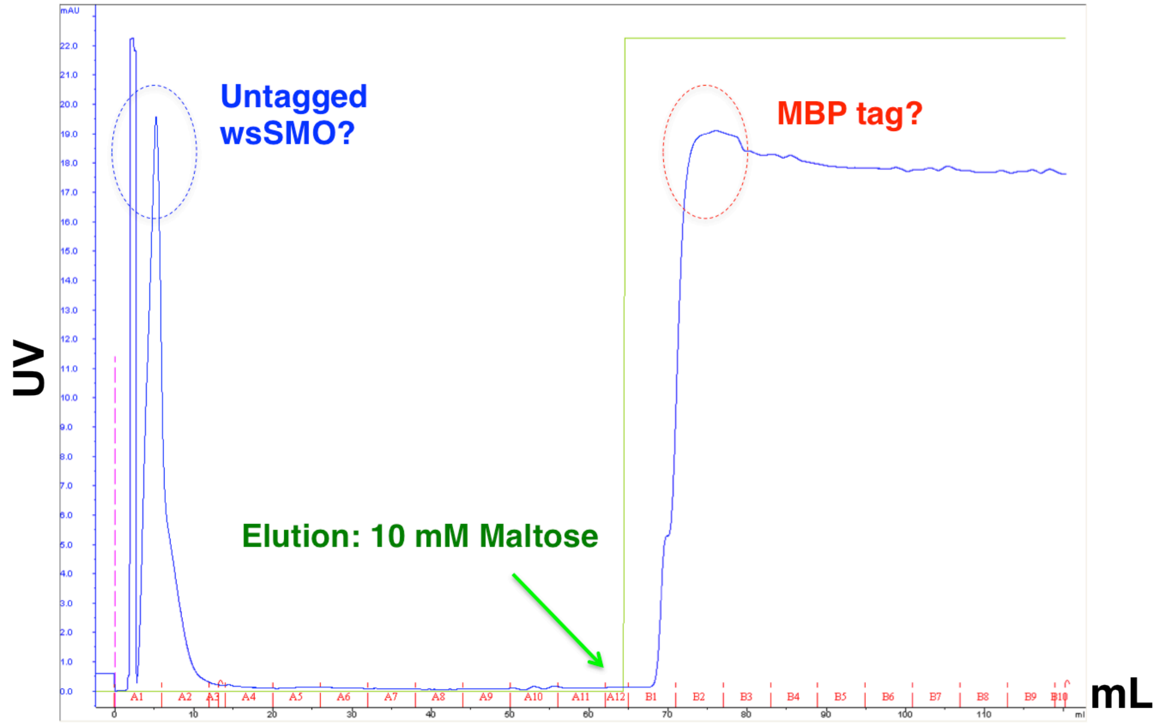


FIGURE 4-8: FXA cleavage products run on the MBP column

FXA cleavage of the MBP-wsSMO fusion protein will theoretically result in two proteins: an isolated MBP tag and an isolated wsSMO. We performed FXA cleavage on the MBP-wsSMO protein, and ran the cleavage products on an MBP column. Theoretically, the wsSMO protein should be found in the flow-through, while the MBP tag will be eluted from the column.

`

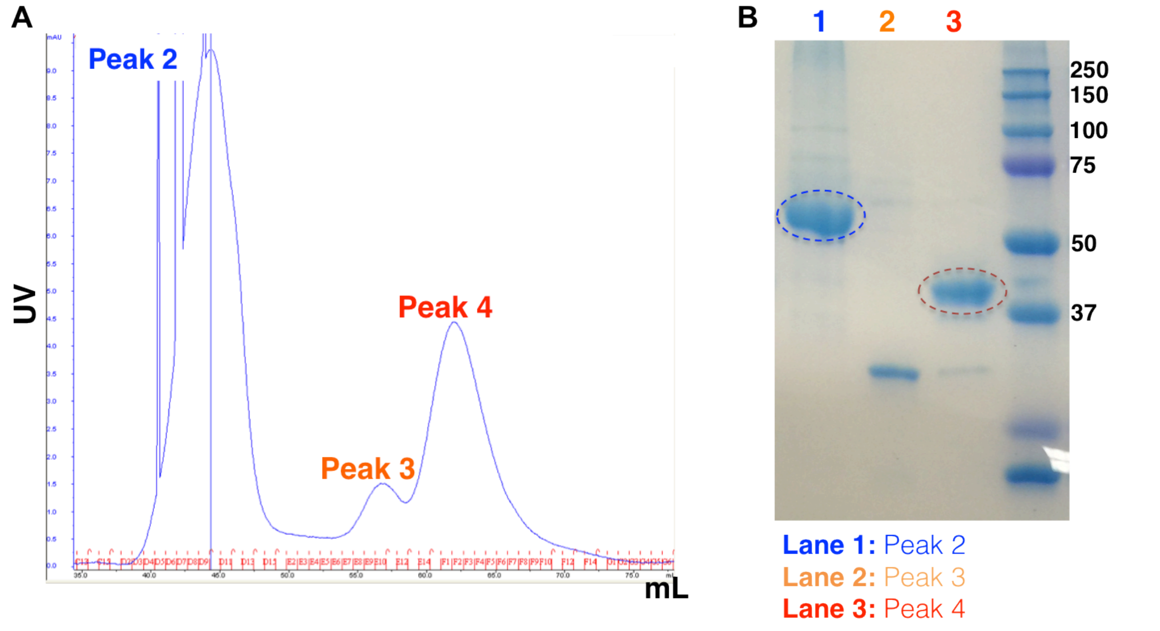


FIGURE 4-9: Size-exclusion run of FXA cleavage products

A: Running the MBP column flow-through from **FIGURE 4-8** through a size-exclusion column results in four peaks; Peak 1 was an injection bubble peak and was cut-off from the chromatograph for clarity purposes.

B: SDS-Page of the different peaks reveals that Peak 4 is close to the theoretical size of untagged wsSMO (41.5 kDA), while Peak 2 is a larger product.

`

To confirm if Peak 2 or Peak 4 was wsSMO, both samples were submitted to Dr. Ben Garcia for M/S protein identification. Using this technique, M/S fragments can be mapped to different regions of the MBP-wsSMO protein sequence.The MBP-wsSMO fusion protein is 763 amino acid longs, with residue 378 representing the division between MBP and wsSMO. For Peak 2, only peptide fragments corresponding to MBP (before residue 378) were found, amounting to a 10.22% total coverage of the MBP-wsSMO protein sequence. Subsequently, the Peak 4 sample was analyzed. Initial digestion with chymotrypsin did not reveal any wsSMO signal; there was only 19.92% coverage of the entire MBP-wsSMO protein, with all of signal corresponding the MBP region (**FIG 4-10A**). However, this M/S spectra was not consistent with the size of the protein on the SDS-gel: it was giving a clean band at ~42 kDA. It was reasoned that, due to inherent structural characteristics of wsSMO, it was not being digested properly by chymotrypsin (Dr. Garcia had observed this in past attempts to sequence membrane proteins). Next, the Peak 4 protein sample was digested with trypsin instead of chymotrypsin, and subjected M/S (**FIG 4-10B**). This digestion resulted in ion fragments that mapped to the wsSMO region. This verified that the size-exclusion Peak 4 sample was indeed wsSMO, and that the MBP-wsSMO fusion was being expressed.

`

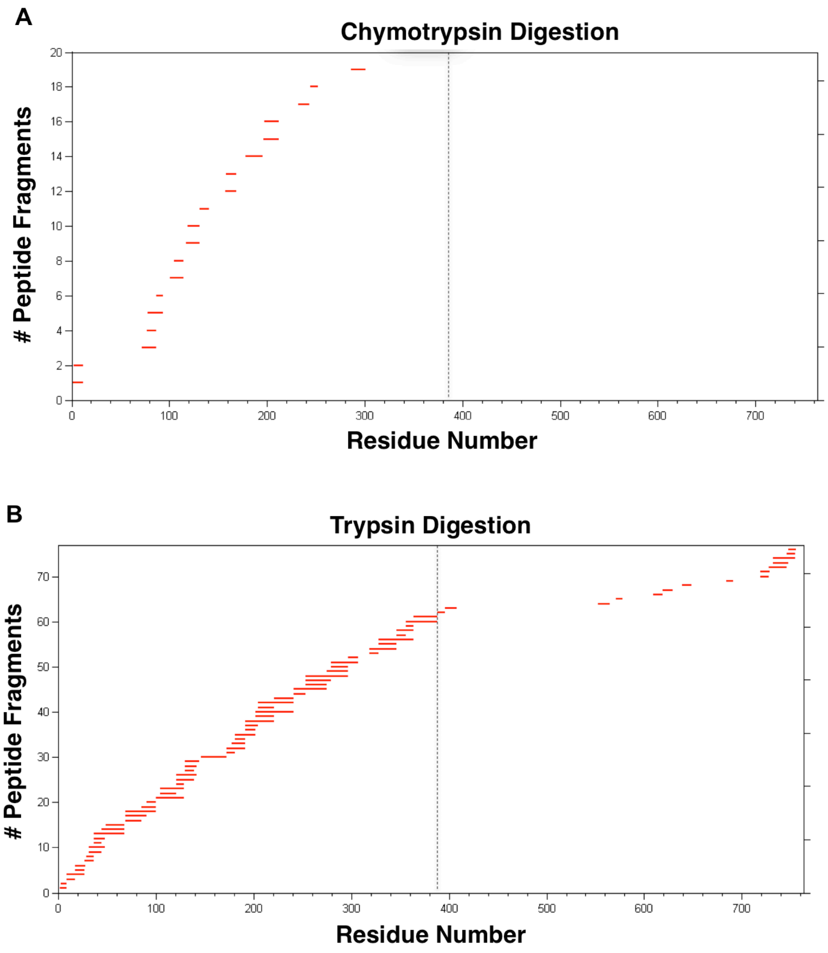


FIGURE 4-10 Peak 4: M/S Protein Identification

The MBP-wsSMO fusion protein is 763 amino acids. Residue 378 (dotted line) represents the division between MBP and wsSMO. Peak 4 sample from Figure 4-9 was subjected to M/S protein identification through digestion of either chymotrypsin or trypsin.

9A: Chymotrypsin digestion of Peak 4 resulted in 19.92% coverage of the entire fusion, with all of the fragments (red) mapping to the MBP region of the sequence (left of the dotted line)

9B: 63.83% coverage of the entire MBP-wsSMO sequence was obtained when Peak 4 was digested with trypsin, with fragments mapping to the wsSMO portion of the sequence (right of the dotted line)

`

The difficulty with trypsin in digesting wsSMO can be accounted for by its high hydrophobic content. In the data, the sequence corresponding to the first 150 residues was not represented by ion fragments. A structural analysis of the designed wsSMO indicates that this region belongs to helix I of the transmembrane domain, and has a high content of hydrophobic residues. Hence, Dr. Garcia believes that chymotrypsin was unable to adequately cleave this region and product ionizable fragments.

A second inconsistency in the M/S data is the presence of MBP ion fragments. The Peak 4 protein sample was subjected to the MBP column, and all of the cleaved off MBP tag should theoretically have stuck to the column. Furthermore, it is not the case that the full MBP-wsSMO was the species that was sequenced, because a strong 85 kDA band was not present in the SDS-gel. This led me to hypothesize that there was a subset of MBP protein that, after cleavage, was not binding to the MBP column and hence was entering the flow-through. This can occur if the MBP tag itself was aggregating or was not properly folded after Factor Xa cleavage, due to the fact that MBP needs to be properly folded in order to bind to the column. From this data, it was concluded that a secondary tag was needed to purify wsSMO from the MBP tag after Factor Xa cleavage.

Western Blots with Anti-MBP Antibody

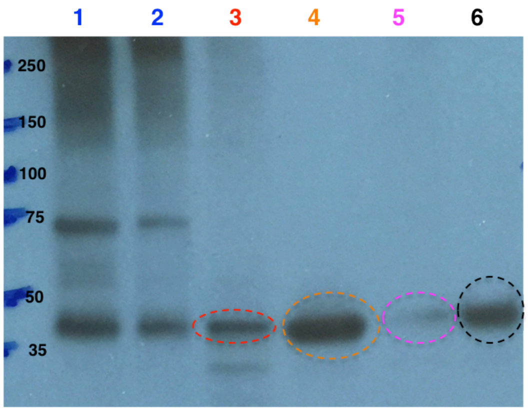
To test the hypothesis that residual cleaved MBP-tag was mixed with untagged-wsSMO after Factor Xa cleavage and MBP column purification, I performed a western blot analysis using rabbit anti-MBP antibody (New England Biolabs; 1:10,000 dilution in BSA overnight; 1:2000 dilution for 2^0 rabbit antibody, one hour incubation at room temperature) (**FIG 4-11**). This western blot demonstrates that, as expected, there is a band in the MBP elution column (**lane 4**) after Factor Xa cleavage that is running at the

`

same size as the MBP control (**lane 6**); both are running at approximately the expected size of MBP, 42 kDA. This is consistent with the MBP tag being separated from the untagged protein species. However, there is also residual MBP found in the MBP-column flow-through (lane 3) and the HIS column flow-through (lane 5). This western blot verifies the M/S sequencing results, which reported a mixture of MBP and wsSMO.

`

Lane 1: MBP column eluted run #1 (pre-FXA)
Lane 2: MBP column eluted run #2 (pre-FXA)
Lane 3: MBP column flow-through (post-FXA)
Lane 4: MBP column eluted (post-FXA)
Lane 5: MBP-SMO-HIS (post-FXA, HIS column FT)
Lane 6: Commercial MBP (positive control for antibody)



- Rabbit Anti-MBP (NEB)
- 1:10,000 dilution for 1^0 in BSA O/N, 4^0C
- 1:2000 dilution for 2^0 Rabbit, 1 hour, RT

FIGURE 4-11: Western Blot analysis of the MBP-wsSMO before and after FXA cleavage

Lane 1 and 2 correspond to MBP-wsSMO fusion protein that is eluted from an MBP column. MBP-wsSMO protein after FXA cleavage (**Figure 4-9**) and injection into the MBP column is shown in Lane 3 (MBP column flow-through) and Lane 4 (MBP column eluted). Theoretically, Lane 3 should be pure untagged protein and Lane 4 should be pure MBP. Using an anti-body against MBP protein (NEB, rabbit primary antibody, 1:10,000 dilution), MBP signal was found in the MBP column's eluted fractions (consistent with MBP-wsSMO protein). However, when MBP signal was also detected in the MBP flow-through fractions of Lane 3. This indicates that the MBP column was not separating the MBP tag from the wsSMO untagged protein after FXA cleavage. Pure MBP protein is shown for reference in Lane 6.

`

4.2. TERTIARY STRUCTURE VERIFICATION VIA BIOPHYSICAL LIGAND-BINDING STUDIES

To test whether wsSMO is in its proper tertiary structure within the MBP-wsSMO fusion construct, we tested its ability to bind three small-molecule inhibitors that are known to bind to wild-type Smoothened (vismodegib, taladegib, sonidegib, and SANT-1)(**FIG 1-2**) using three biophysical approaches: circular dichroism thermal melt, CPM dye thermal melt, and nuclear magnetic resonance (NMR) ^{1}H and ^{19}F spectroscopy. As of writing, the CPM dye and NMR studies are still in progress.

a. Circular Dichroism thermal shift assay

Thermal shift assays are used to monitor protein unfolding as a function of temperature[89] (usually, across a range of 2^{0}C- 98^{0}C). There are several variations of this assay. One variant monitors the protein's circular dichroism signal at 222 nm during the temperature range; loss of signal corresponds to a loss of helicity as the protein unfolds[90]. Another variant of this assay uses a fluorescent dye[91] (i.e. SYPRO orange, 1-anilinonaphthalene-8-sulfonic acid (ANS), 7-diethylamino-3-(40-maleimidylphenyl)- 4-methylcoumarin (CPM)) that binds to hydrophobic residues in the protein. As a protein unfolds and interior hydrophobics become more exposed, the dye molecules will bind and an increase in fluorescent signal can be detected. In addition to being used to determine the melting temperature (T_m) of a protein, these approaches are used to as a read-out for ligand binding. When a ligand binds to a protein and stabilizes it, a shift in melting temperature relative to the apo protein is expected.

For our initial experiments, we used a ratio of 2.35 uM MBP-wsSMO (purified from Anion-Exchange, as described in **section b**) to 200 uM SANT-1. This ratio is used in the pharmaceutical industry as a starting point in high-throughput drug screening using

`

thermal shift assays [92]. First, a 20 mM ethanol stock solution of SANT-1 was made, and this stock was added to MBP-wsSMO in 20 mM Tris-200 mM NaCl (pH 7.4) buffer to a final volume of 250 uL (1% ethanol). This mixture was incubated at 23^0C for two hours, with gentle mixing every 30 minutes, and then subjected to CD thermal shift analysis. These results, compared to Apo 2.35 uM MBP-wSMO in 1% ethanol (obtained from the same batch of protein earlier in the day), are shown in **FIG 4-12**. These data demonstrate that Apo MBP-wsSMO appears to be more thermostable than the MBP tag alone: while MBP undergoes a dramatic unfolding event starting at ~50^0C, MBP-wsSMO unfolding occurs more gradually over a longer temperature range. Furthermore, the comparison of Apo and 200uM MBP-wsSMO samples demonstrates that drug incubation appears to stabilize MBP-wsSMO at temperatures above 70^0C. To test if SANT was binding non-specifically to the MBP tag, 10 uM of MBP was incubated with 200 uM SANT (under the aforementioned incubation conditions) and subjected to a thermal shift assay (**FIG 4-13**). These results indicate that MBP incubated with SANT is not more thermostable than Apo MBP during the temperature range it unfolds (between 50^0C and 60^0C).

`

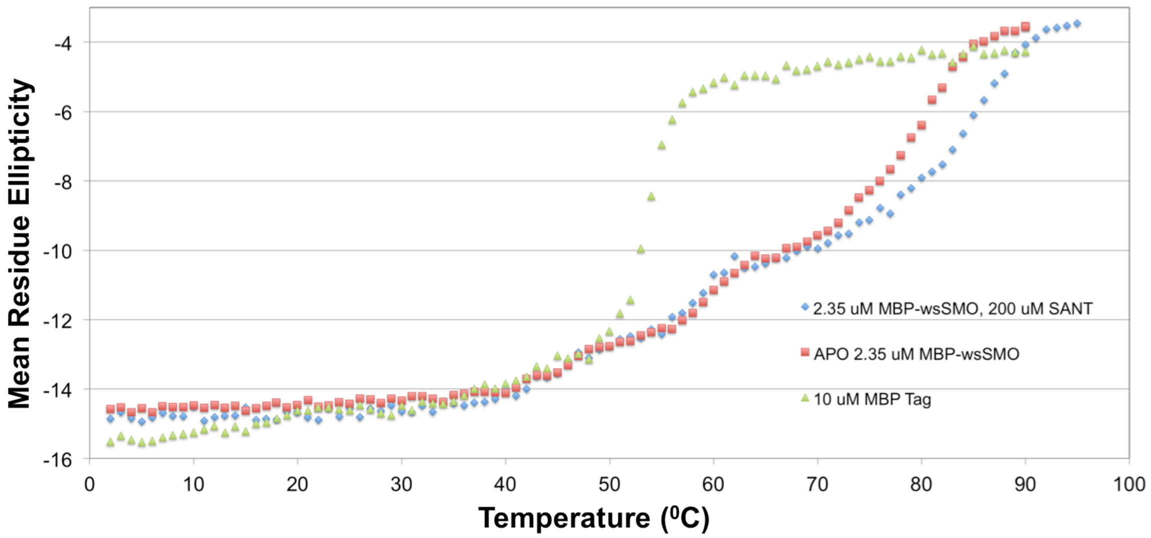


FIGURE 4-12: Tmelt spectra of the isolated protein from FIGURE 5 (vs. MBP)

The CD thermal profile of 2.35 uM MBP-wsSMO vs. 20 uM MBP demonstrate that MBP-wsSMO is more thermally-stable than MBP, and has a melting/unfolding period over a longer temperature range. Comparison of Apo MBP-wsSMO with MBP-wsSMO incubated with 200 uM SANT-1 drug indicate a shift in melting temperature and stabilization of helicity in the presence of drug.

`

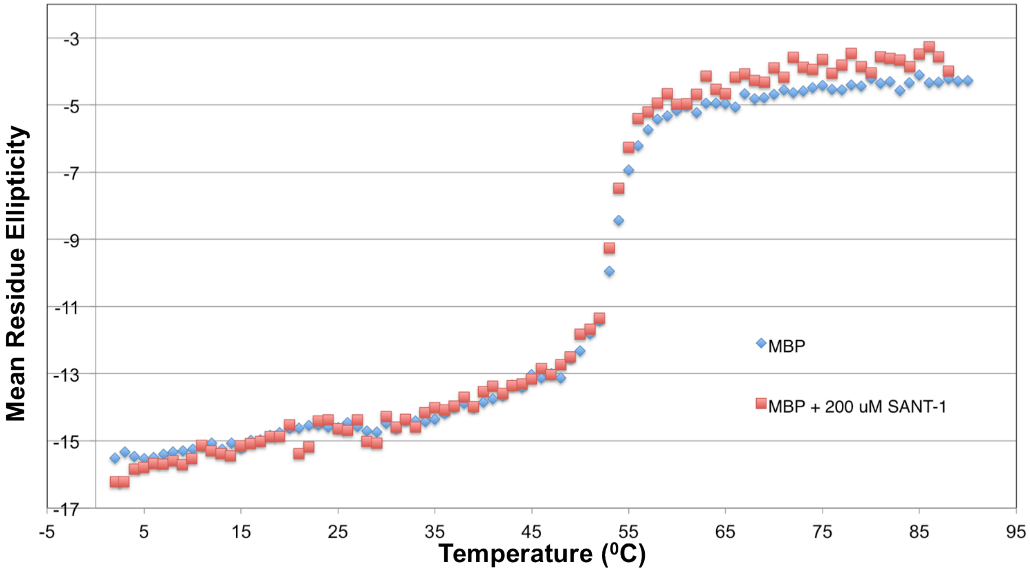


FIGURE 4-13: MBP and MBP-SANT

To verify that SANT was not binding to MBP, a thermal melt was performed with 10 uM MBP in the absence (blue points) and presence of 200 uM SANT (red points). These data indicate that 200 uM SANT does not induce a melting temperature shift in MBP, nor cause an increase in helical character over the range where MBP melts.

b. NMR-based binding experiments

In this section, I describe how ^{19}F and ^{1}H NMR were used to determine if known SMO inhibitors can bind to MBP-wsSMO: SANT-T, Vismodegib, and Taladegib.

NMR-based ligand-binding methods are currently being used in the pharmaceutical industry in drug screening assays and, in particular, fragment-based screening assays[93]. An example of this is shown in **FIG 4-14.** Here, a ^{19}F-NMR spectra was obtained for small-fragments that contain fluorine groups (top panel, in black). After 10 uM of protein target of interest, BACE-1, is added to the fragment mixture, the resulting NMR spectra

`

demonstrates a broadening and intensity decrease of one of the peaks (bottom panel, in red). The same type of experiment has been shown with ^{1}H NMR as well[94] **(FIG 4-15)**

The explanation for the NMR-spectra differences in peak width and peak intensity between the free drug and bound drug states can be explained as follows. The peak broadening is caused by differences in the small-molecule's local environment. When the unbound molecule is in solution, its molecular tumbling rate (i.e. its rotations and translations about its axis) is high[95].Since a molecule's NMR spectra represents a population-weighted average of the different electronic environments for each functional group, a molecule with a high tumbling rate will have sharp, narrow peaks. In contrast to small-molecules, proteins are much larger and have lower molecular tumbling rates. When a small-molecule associates with a protein's binding site, its tumbling rate will be equivalent to the protein's tumbling rate. The resulting NMR spectra will reflect a population-weighted average of the free and bound forms. As a result, during a binding event, one will observe peak broadening (due to the averaging of the bound molecules' low tumbling rate and the unbound molecule's high tumbling rate) in the NMR spectra that for the functional groups that are bound to the protein. The decreases in peak intensity can be attributed to the fact that, when the compound is bound to the ligand, some of its functional groups are buried and not as exposed to the magnetic field versus if it were free in solution.

`

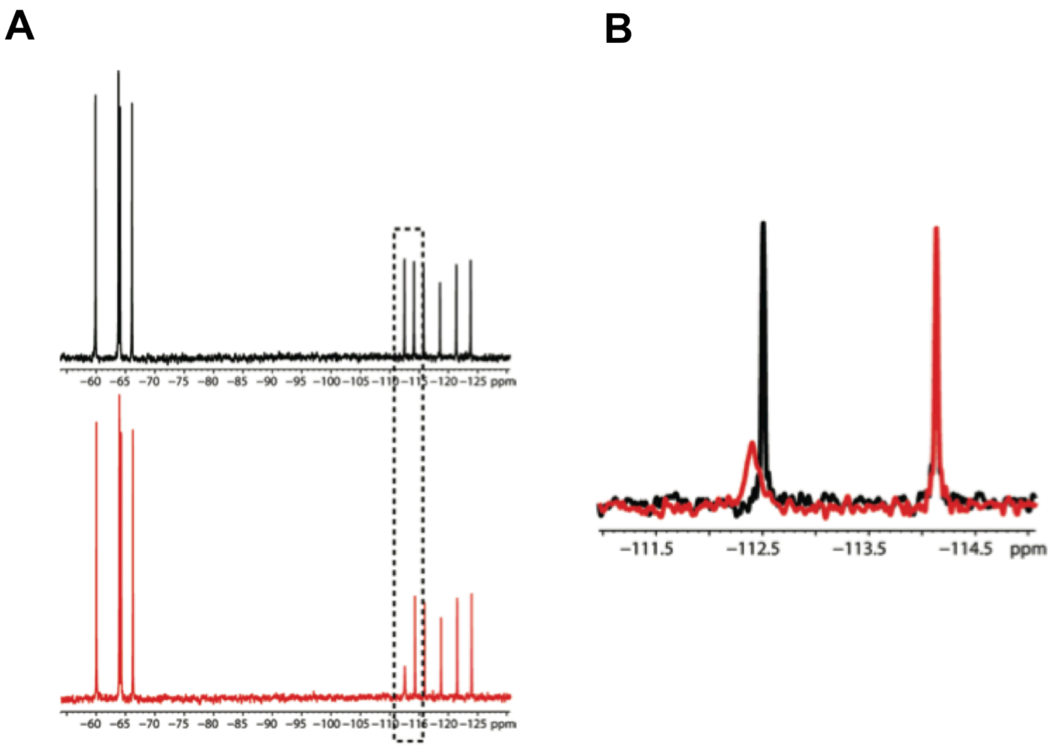


FIGURE 4-14: ^{19}F NMR Fragment Screen Spectra

A) Black panel (above) demonstrates NMR spectra of multiple ^{19}F fragments. The red panel (below) demonstrates the NMR spectra of the fragments incubated with 10 uM of a target protein of interest, BACE-1.

B) Zoom-in and overlap of the spectra before and after protein incubation can allow for the determination of fragments that are binding to the protein, via decrease in peak intensity and a peak shift.

From the paper:

Jordan et al. Fragment-based Drug Discovery: Practical Implication Based on ^{19}F Spectroscopy. Journal of Medicinal Chemistry 2012, 55, 678-687

`

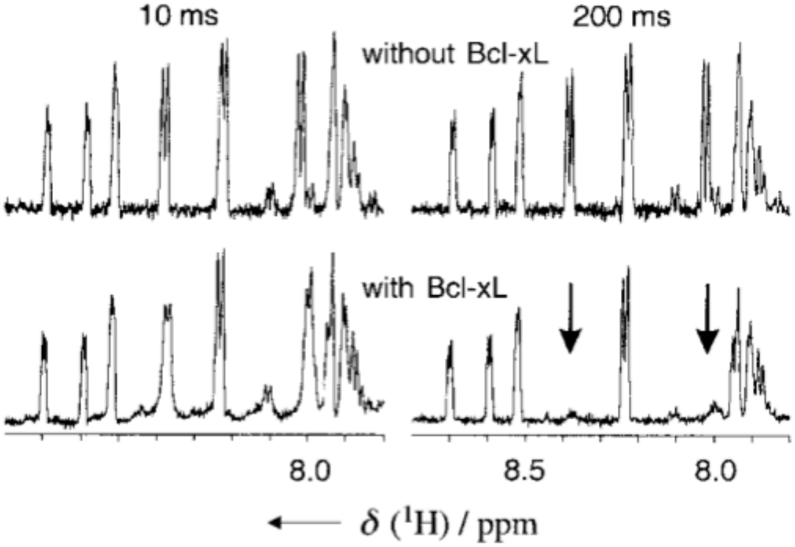


FIGURE 4-15: ^{1}H NMR fragment screen spectra for BCL-XL inhibitors

Top panel demonstrates ^{1}H NMR spectra of compound library without protein target (top), and incubated with Bcl-xL protein (bottom). A decrease in intensity of portions of the NMR spectra in the presence of protein (arrow) allows one to determine the compound that is binding to Bcl-xL.

From the paper:

Meyer and Peters. NMR Spectroscopy Techniques for Screening and Identifying Ligand Binding to Protein Receptors. Angew. Chem. 2003, 42, No.8

`

^{1}H binding studies with SANT-1 and Vismodegib

To determine if SANT-1 binds to MBP-wsSMO via ^{1}H NMR, we first performed a buffer exchange of pure MBP-wsSMO from its storage buffer (20 mM Tris-HCl, 200 mM NaCl, pH 7.4) to the equivalent deuterated buffer (20 mM deuterated- Tris, 200 mM NaCl, pH 7.4 in deuterated H_20) using Amicon centrifugal filters. For this experiment, protein sample in the storage buffer were placed in the filter and centrifuged for 15 minute cycles at 6000 rpm (5450 rcf). After each cycle, a volume of deuterated buffer was added to the protein that was equivalent to the volume of normal buffer that passed through the filter. These volume changes were used to track the volume percent of H_2O in solution. After seven cycles, the H_20 percentage was reduced to < 0.5% by volume, which was adequate performing ^{1}H NMR experiments without water contaminating the spectra.

Next, we wanted to determine if there were differences in the ^{1}H NMR spectra of SANT-1 alone and SANT-1 incubated with MBP-wsSMO. First, we obtained the ^{1}H spectra of 200 uM SANT-1 (**FIG 4-16**, aromatic group between 7-8 ppm circled in red). A series of titrations were performed, from a 200:1 ratio of SANT-1: MBP-wsSMO down to 20:1 (**FIG 4-17**). For these experiments, the appropriate volume of MBP-wsSMO protein and 200 uM drug were added together to a total volume of 600 uL, and incubated at 23^{0}C for 1 hour. Next, the incubation mixture was transferred to an NMR tube, and ^{1}H spectra acquisition was obtained. The data in **FIG 4-17** demonstrate that, as MBP-wsSMO protein is titrated into 200 uM SANT-1 at increasing concentrations, the peaks that correspond to the SANT-1 aromatic groups exhibit widening and a decrease in intensity; the extent of these changes increases with increasing protein concentration. Taken together, these data indicate that SANT-1 is binding to MBP-wsSMO. We next performed a 20:1 (drug : protein) experiment with MBP-wsSMO and the SMO inhibitor

`

vismodegib **(FIG 4-18)**, an FDA-approved compound that is currently indicated for treatment of basal-cell carcinoma. Incubation of 200 uM vismodegib with 10 uM MBP-wsSMO resulted in a dramatic intensity decrease of the aromatic peaks (**FIG 4-19**), indicating that vismodegib is binding to MBP-wsSMO.

`

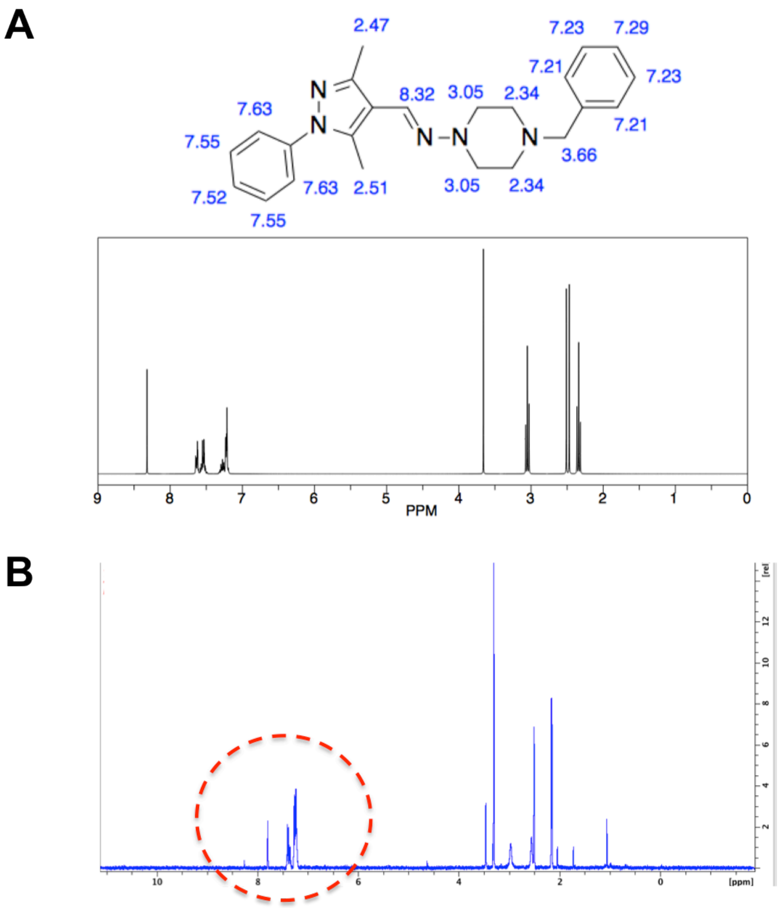


FIGURE 4-16: ^{1}H NMR Spectra of 200 uM SANT-1

A) Predicted ^{1}H NMR spectra for SANT-1 (ChemDraw)

B) Actual ^{1}H NMR spectra for 200 uM SANT-1. Aromatic region of interest is circled in red.

`

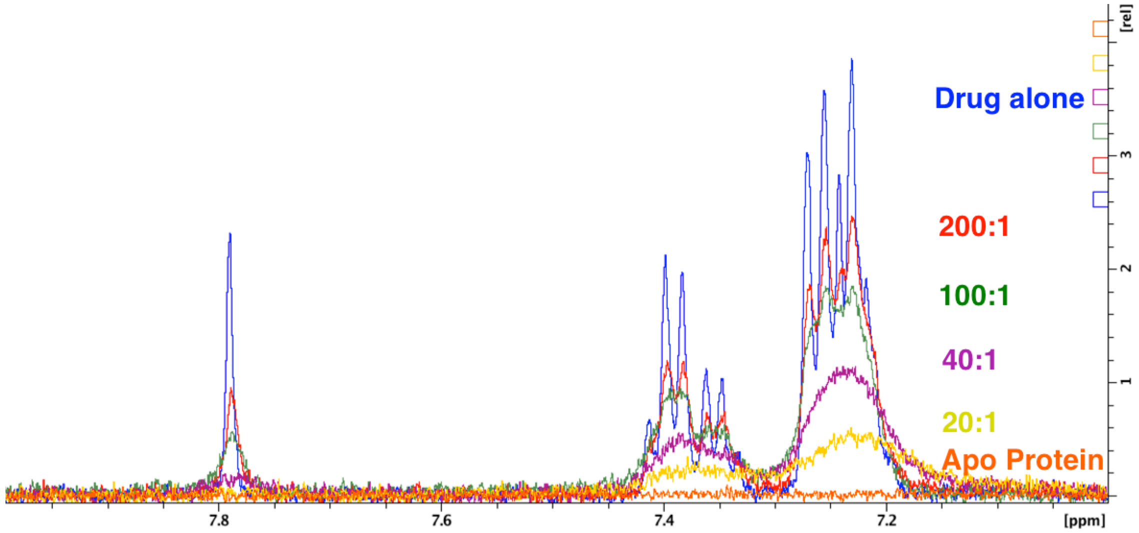


FIGURE 4-17: ^{1}H NMR spectra of SANT-1 incubated with MBP-wsSMO Protein

A) ^{1}H NMR spectra for 200 uM SANT-1 incubated with various concentrations of MBP-wsSMO protein: **SANT-1 alone**, **200:1 (1 uM MBP-wsSMO)**, **100:1 (2 uM MBP-wsSMO)**, **40:1 (5 uM MBP-wsSMO)**, **20:1 (10 uM MBP-wsSMO)**, **Apo Protein (10 uM MBP-wsSMO)**. A dose-dependent decrease in peak intensity, and peak broadening, is observed.

`

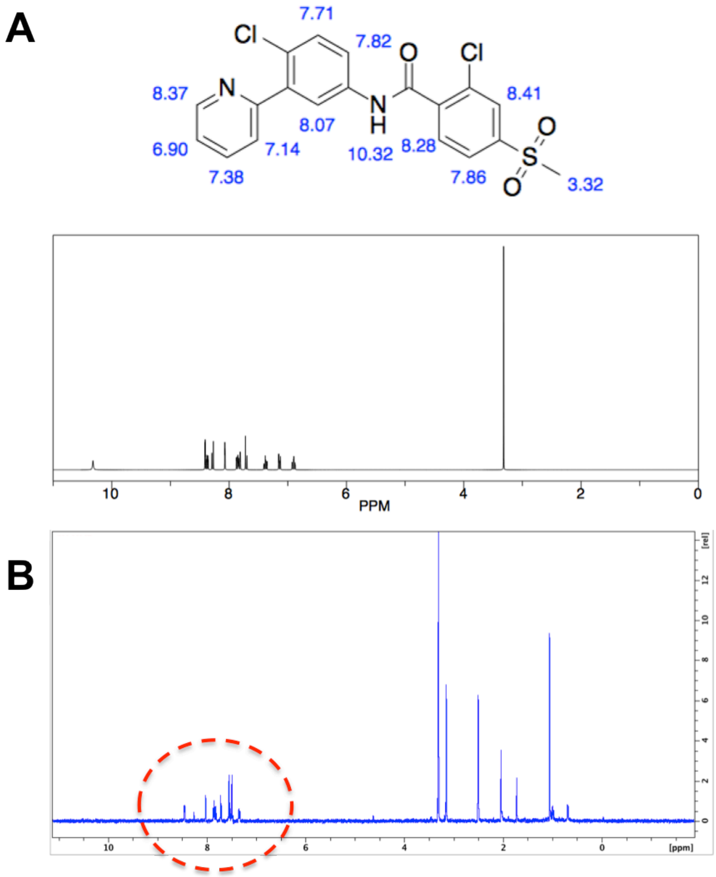


FIGURE 4-18: ^{1}H NMR spectra of Vismodegib

A) Predicted ^{1}H NMR spectra for vismodegib (ChemDraw)

B) Actual ^{1}H NMR spectra for 200 uM vismodegib. Aromatic region of interest is circled in red.

`

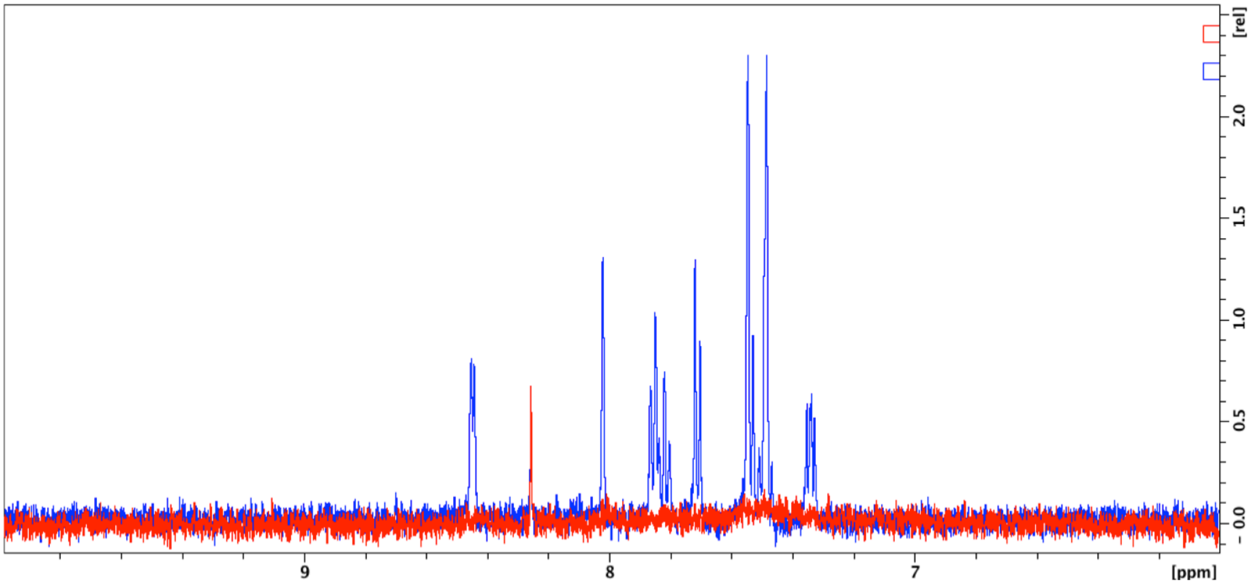


FIGURE 4-19: ^{1}H NMR spectra of Vismodegib incubated with MBP-wsSMO Protein

^{1}H NMR spectra of **200 uM vismodegib (blue)** and **200 uM incubated with 10 uM MBP-wsSMO**. Upon incubation with the drug, there is a loss of signal of the aromatic region, suggesting that the aromatic region is bound within the pocket of MBP-wsSMO.

`

^{19}F binding studies with Taladegib

The biggest drawback of the ^{1}H NMR binding experiment is the buffer exchange step where MBP-wsSMO is placed in deuterated-Tris/deuterated-H_20 solvent. At the end of the seven centrifuge cycles needed to get H_20 percentage down to 0.5%, we experienced a protein loss of ~50%. This was problematic because (1) it took several time-consuming FPLC steps to obtain purified MBP-wsSMO from one bacterial pellet, and (2) one would ideally want to have enough MBP-wsSMO protein available to test several compounds at several different concentrations.

One way to avoid these buffer exchange steps would be to perform a ^{19}F NMR binding experiment instead of a ^{1}H NMR binding experiment. These experiments do not require the use of deuterated solvents, which would allow for ligand-binding studies to be performed in the regular storage buffer and prevent the dramatic loss of MBP-wsSMO protein that we experienced. As with ^{1}H NMR binding studies, binding events with ^{19}F-NMR are indicated by broadening of peaks, decreased peak intensity, or peak shifts[93]. For this experiment, we used the SMO inhibitor Taladegib (**FIG 4-20**) (currently in clinical trials), which has two aromatic-ring fluorine substituents (a -CF3 and an -F). The ^{19}F NMR spectra of 200 uM Taladegib was obtained, demonstrating two peaks that correspond to the two fluorinated groups. To determine if non-specific binding was occurring, 200 uM of Taladegib was incubated with 10 uM of purified MBP, resulting in a subtle decrease in peak intensity and a slight broadening of peak width; this suggests that there is a degree of non-specific binding to Taladegib to the MBP tag. However, when 200 uM of Taladegib was incubated with 10 uM MBP-wsSMO, a larger degree of peak broadening and peak intensity decrease was observed compared to the spectra with Taladegib-MBP, in addition to a shift of these peaks. These data suggest that

`

Taladegib is binding to MBP-wsSMO, and that these observed NMR changes are not due to non-specific binding to the MBP tag.

`

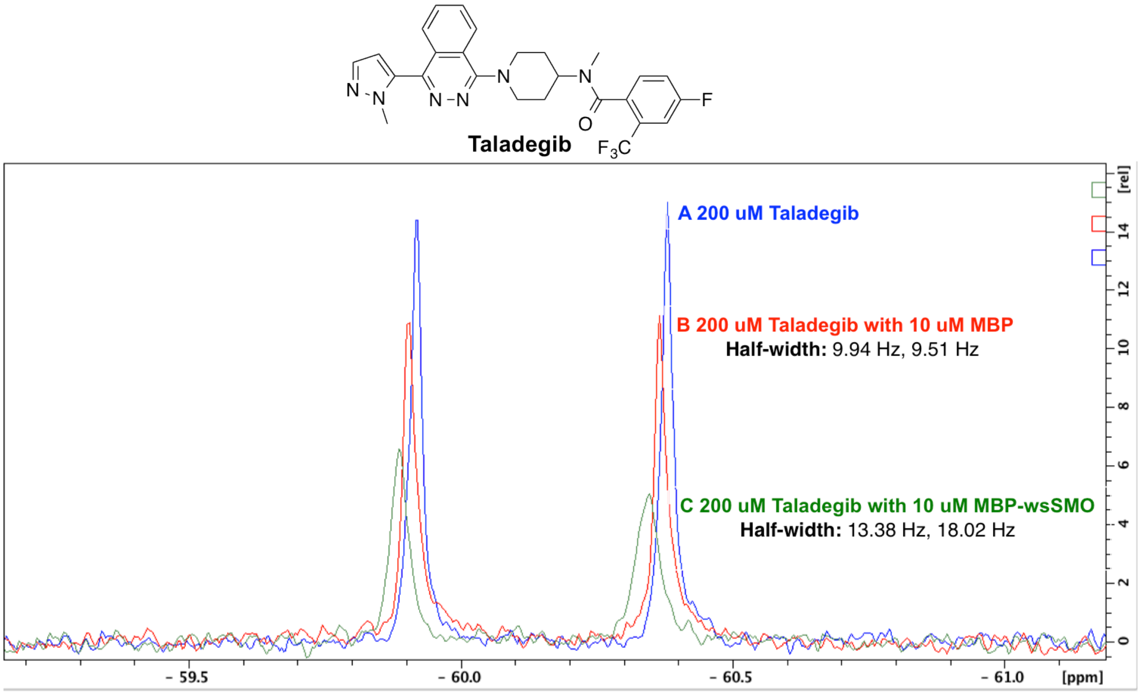


FIGURE 4-20: ^{19}F NMR spectra of Taladegib incubated with MBP-wsSMO Protein

200 uM Taladegib was incubated with **10 uM of MBP (negative control)** and **10 uM of MBP-wsSMO**. A slight peak broadening, as well as a slight decrease in peak intensity, is observed for the Taladegib/MBP sample. However, the Taldegib/MBP-wsSMO sample shows a larger amount of peak broadening (as indicated the peak half-width) and peak intensity. These data suggest that Taladegib is making transient interactions with MBP (likely non-specific binding), but making stronger and deeper interactions with MBP-wsSMO.

`

d. Summary of findings thus far

Based on the experiments described in sections c-f, I conclude the following about the MBP-wsSMO fusion protein:

- it is expressed in a soluble form in *E. coli.* We did not have to isolate it from the insoluble fraction, and did not have to refold it using a variety of buffer conditions.
- It exhibits α-helical character upon a CD wavelength scan, and does not exhibit random-coil character that is consistent with an unfolded protein
- It has been sequence-verified by M/S protein identification.
- It has a CD wavelength and thermal melt profile that differ from isolated MBP tag, indicating that this is not the case that all of the signal is coming from the MBP portion of the fusion protein
- It is more thermal stable than the isolated MBP tag, and has a longer period of unfolding (over a longer temperature range) than the MBP tag
- It appears to be stabilized upon incubation with the known SMO ligand 200 uM SANT, over a temperature range of 70^0C to 90^0C. The MBP tag alone, however, exhibits no change in helicity or stabilization upon incubation with the same drug.
- It appears to bind to compounds that are known to bind to the wild-type SMO, as assessed by ^{1}H and ^{19}F NMR binding studies

e. Experiments in progress

As of writing, we are performing the following experiments:

Negative control experiment: The above experiments suggest that 1) MBP-wsSMO is binding to the known SMO inhibitors taladegib, vismodegib and SANT-1, and 2) As demonstrated by ^{19}F NMR and circular dichroism thermal shift assay, binding of MBP-wsSMO to the these inhibitors is not due to non-specific binding of these inhibitors to the MBP tag. A negative control would determine if the NMR binding assay is specific for binding to SMO inhibitors versus molecules that bind to other GPCRS. The binding study with drug maraviroc, which binds to the GPCR CCR5 and which has a ^{19}F-group, will be used for this negative control experiment. First, the ^{19}F NMR spectra will be determined for maraviroc, and for maraviroc incubated with 10 uM MBP-wsSMO and 10 uM MBP. If this binding assay is specific, there should be no (or minimal) change in the NMR spectra between the maraviroc ^{19}F spectra alone and maraviroc incubated with MBP or MBP-wsSMO.

`

CONCLUSIONS

In this thesis, I addressed three key problems in medicinal chemistry and drug discovery using a combination of computational and experimental approaches. I will conclude my thesis by providing a summary of the methodologies that I used to address these problems, as well discussing how my methodologies provide a practical improvement over the approaches that are currently used in the field.

PART I. DESIGNING A DRUG

Problem 1. Designing brain penetrant inhibitors

Many protein targets that are important in the pathophysiology of neurological disorders are expressed in peripheral tissues. Many of these targets are also involved in other diseases, and already have effective drugs that inhibit them. However, the majority of these drugs have a mechanism of action that involve very polar or charged functional groups (i.e. electrostatic interactions, multiple hydrogen bonds) that prevent them from passing through the blood brain barrier. Because of this problem, developing brain penetrant inhibitors has been a huge challenge in the medicinal chemistry community. In **Chapter I** of my thesis, I posed the following question: can one use computational approaches to determine the ways to modify an inhibitor of a known protein target so that the modified compound can not only get into the brain, but can still inhibit its target once there?

The case that I focused on is the carboxylic acid group, which is found in a plethora of drugs yet which forms carboxylate at physiological pH. The traditional method that has been used in medicinal chemistry over the past ~50 years (which is still being used today) has been to randomly replace the carboxylic acid group with a multitude of

`

derivatives, and to test them all individually in assays for brain penetration (i.e. partioning of the compounds between water and octanol, or through a PAMPA artitifial membrane). However, the chemical space around which the carboxylic acid group can be modified is vast (see **FIG 2.6** for a subset of isosteres). For chemistry academic labs, it would take one post-doc working for close to a decade to synthesize the all of the common carboxylic acid isosteres (and sub-derivatives of these compounds) necessary to make a screening library. In larger drug development programs in the pharmaceutical industry, teams of organic chemists can work for years a decade produce thousands of derivatives. There are two main problems with this approach. First, even with an x-ray crystal structure of the target, there are an immense number of changes that can be made at several sites of the inhibitor that can influence its ability to bind the target. Second, the approach of randomly making compound analogs and screening them against the target to find a hit has historically had a very low success rate, akin to finding a needle in a haystack. Third, even if it is the case that an analog turns out to be a hit from the screen, it must now have the physiochemical properties that are amenable to brain penetration. Hence, there does not exist a straightforward procedure for predicting whether a given carboxylic acid analog can 1. still inhibit the target and 2. has the physiochemical properties that would allow for brain penetration.

In **Chapter I**, I addressed these problems using tools from computational chemistry and synthetic organic chemistry. A brief summary of the approach I used is as follows. Our group made several model compounds that, historically in the medicinal chemistry field, have been used to as replacements of the carboxylic acid group. We then experimentally determined different physiochemical properties of these isosteres, and came up with a subset of isosteres that had a high probability of penetrating the brain (based on their $\log P_{app}$ and $\log D_{7.4}$ values). Next, I used a chemical enumeration

`

algorithm to place these isosteres onto the scaffolds of drugs that were known to bind to my target, resulting in ~24,000 candidate compounds. Afterwards, I used a molecular docking algorithm to screen these compounds against the x-ray crystal structure of my drug target, in order to determine those candidate compounds that had a high likelihood of inhibiting the target even after the carboxylic acid group was replaced. Lastly, to demonstrate that my computational strategy worked, I synthesized a number of these screening hits. A collaborator experimentally determined that these synthesized compounds did indeed inhibit the target in an LC-MS based assay that is the standard for testing drug activity.

The isosteric replacement/molecular docking/organic synthesis strategy I put forth in Chapter I provides a practical improvement over the approach that is currently used in the field. For brain penetrant inhibitors, chemists can use the physiochemical property data that we determined to guide the types of compounds that they want to make for their screening libraries. The molecular docking approach can be used to further limit the number of compounds to make and test, which would not only save a substantial amount of time and money, but increase the likelihood that a synthesized compound will be a true hit. This strategy does not need to be limited to carboxylic acid isosteres: there are several functional group classes that are widespread in medicinal chemistry (i.e. amide bond isosteres), and this approach can easily be translated into drug discovery programs involving these different classes.

Problem 2. Designing drugs to inhibit two targets simultaneously

Many pathophysiological processes are driven by multiple protein targets that work together synergistically to promote a disease state. These proteins are often in different biological pathways and have different functions. One therapeutic strategy would be to

`

administer multiple drugs that inhibit these protein targets: this is a common strategy used in infectious diseases (i.e. H.A.A.R.T therapy in HIV, and antibiotic combinations for serious drug-resistant bacterial strains such as methicillin-resistant *Staph aureus*). This strategy is useful when all of the protein targets are known, and there already exist inhibitors for each of them. However, it is not feasible when one or more of the targets does not already have an inhibitor. A different strategy would be to take a known inhibitor of one of these targets, and to modify it so that it now not only still inhibits its targets, but now inhibits a *second target*. This strategy is difficult for many of the reasons explained for **Problem 1. Designing brain penetrant inhibitors**: since there are an immense number of chemical changes that can be made at several sites of an inhibitor, a very large number of analogs will need to be made and screened against both targets. In **Chapter II** of my thesis, I posed the following question: can one use computational approaches predict which changes can lead to compounds that can inhibit both targets?

I illustrated my methodology using the example of Alzheimer's Disease. Experimental evidence from our group demonstrated that both the neuroinflammation pathway (driven by the proteins COX and LOX) and microtubule destabilization are involved in Alzheimer's Disease onset and progression. Hence, COX, LOX and microtubules represent potential drug targets. For my approach, I first identified known microtubule-stabilizing agents that structurally resembled the COX inhibitor celecoxib. Next, I took these compounds and performed molecular docking using the x-ray structure of COX-1 with celecoxib. Celecoxib does not have the classic carboxylic acid-ARG120 interaction that the other COX inhibitors have; because of this, other molecular interactions had to be identified (i.e. GLN 192 hydrogen bond with the amide of the inhibitor that is present deeper in the binding cavity). I used these alternative interactions to determine if the candidate microtubule-stabilizing agents were binding in the COX-1 site in the proper

`

orientation, and that they were making similar interactions to that of celecoxib with COX-1. In the end, 11 candidate inhibitors were from the screen were initially synthesized. 7 of these inhibitors came directly from the docking screen (without modification), while 3 were subtle modifications of these hits. In the end, 10/11 of these compounds were hits in both the LC-MS COX assay and the LC-MS LOX assays. These results provide a starting point for tri-active Alzheimer's Disease drugs.

The strategy that I put forth in this chapter represents an advancement in drug development programs for not only Alzheimer's Disease, but with other diseases that are driven by multiple protein targets. If it is the case that one such protein target has an x-ray crystal structure, molecular docking can be used to identify compounds that have a higher probability than binding to other targets.

PART II. (RE)DESIGNING A PROTEIN TARGET

Problem 3. Membrane proteins are involved in a plethora of disease processes spanning multiple organ systems. However, developing drugs against them is difficult because there does not exist a straightforward method for testing whether a ligand can bind to a membrane protein of interest. This is due to the plethora of hydrophobic residues that exist on the exterior of membrane proteins. Because of these residues, membrane proteins are difficult to isolate and incorporate into ligand-binding studies. The goal of Part II of my thesis was to develop a computational approach for designing a water-soluble variant of a membrane protein that retains the wild-type's key structural and ligand-binding properties. Such a variant could then be incorporated into a ligand-binding assay that does not require detergents or labeling (either of the protein or of the drug).

`

How my method can advance our understanding of a specific membrane protein: the cancer-associated GPCR SMO

Aberrant Hedgehog (Hh) pathway activation been implicated in a variety of cancer types. Attempts to target this pathway via SMO inhibition have yielded promising clinical trial results. Over the past several years, key structural features of SMO have been elucidated. Various synthesized compounds have been used to identify extracellular binding sites, and x-ray structures have been obtained of SMO in different conformations. Despite this progress, many of SMO's direct extracellular and intracellular interactions are not known. For example, it has been proposed that the membrane protein Patched represses SMO by actively pumping out its endogenous agonist; however, this agonist has not been discovered. Furthermore, it has not known whether SMO activates oncogenic transcription factors through a direct interaction or indirectly through other proteins. The lack of a simple experimental system for testing SMO binding has prevented progress in our understanding of these key molecular interactions. One commonly used SMO binding assay tests whether a candidate molecule is able to displace binding of radioactive or fluorescently-labeled cyclopamine (a SMO antagonist) from cells that are over-expressing SMO. However, this method is unable to identify molecules that bind outside of the cyclopamine binding site. Another method involves over-expressing SMO in a cell line (i.e. HEK 293T) and performing binding studies on isolated membrane fractions. In addition to being a technically challenging and low-yield experiment, this approach is problematic due to the presence of non-SMO proteins in the fraction that could bind to the candidate molecule.

The development of wsSMO was incorporated into a cell-free NMR-based binding assay. This assay was used to test whether known candidate molecules were binding to

`

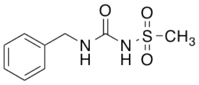
wsSMO via a changes in the NMR spectra of the compounds. In the future, I foresee wsSMO itself serving as a useful experimental tool for large- scale discovery experiments such as the proteomics assay SILAC, which can be used to discover proteins and peptide ligands that bind to wsSMO via a mass spectrometry read-out. Hence, both wsSMO and the wsSMO-NMR assay will provide researchers with much-needed experimental tools for obtaining new biological insights into how activated SMO leads to tumorigenesis. This information can potentially lead to novel methods for therapeutically attacking the Hh pathway, via inhibiting newly discovered SMO binding sites or the proteins it interacts with.

How my method can advance drug development efforts for membrane proteins in general

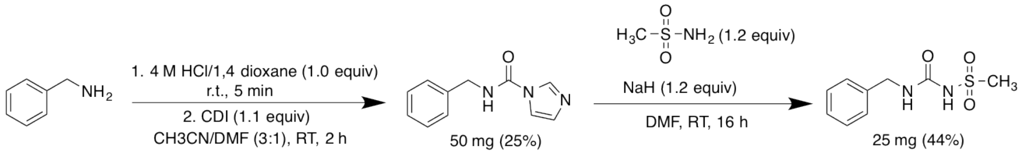
Membrane proteins are the targets of over 60% of all pharmaceutical drugs on the market. Drug resistance due to membrane protein binding-site mutations is a widespread challenge in medicine and has been identified in other cancer types (i.e. EGFR in lung cancer) and in a variety of infectious diseases such as HIV (CCR5), Influenza (M2 channel) and methicillin-resistant *Staph Aureus* (penicillin-binding site). Though this proposal is focused on SMO drug-resistance, the technologies I am developing can be applied generally to the study of disease-associated membrane proteins. For a membrane protein that is difficult to purify in quantities needed for structural or experimental studies, the membrane protein solubilization approach can be used to develop a water-soluble variant that can be isolated in large amounts from a protein expression system and used in a variety of experiments. Incorporating a water-soluble variant into an NMR assay can provide a quick and simple method for testing hypotheses about molecular interactions and drug binding.

`

APPENDIX A. SYNTHESIS OF SULFONYLUREA ISOSTERE

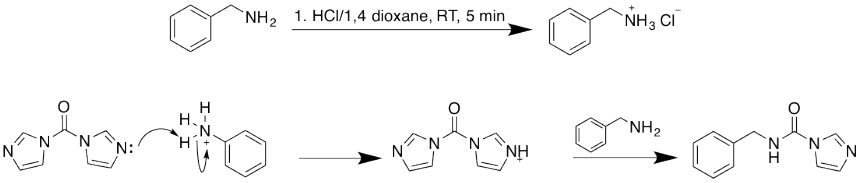


Unknown

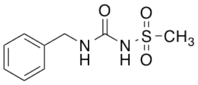


^1H NMR, ^{13}C NMR, HR-MS, IR, MP, x-ray

Proposed mechanism of first step:

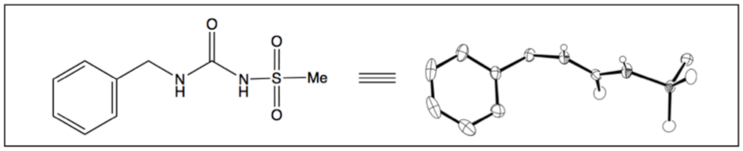


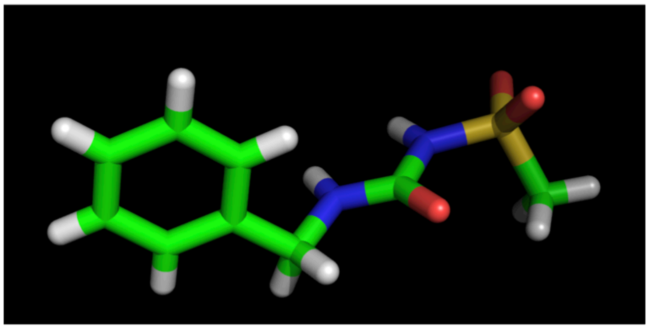
J. Org Chem. **2012**, *77*, 10362-10368.



Unknown

Crystals were grown (m.p = 163-164⁰C)





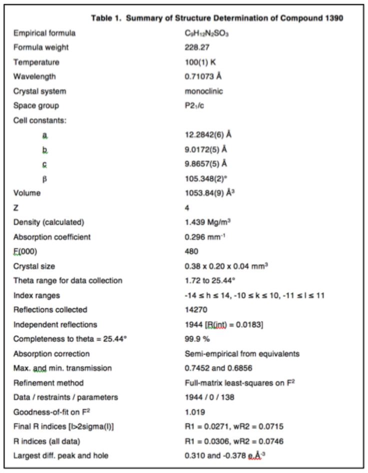
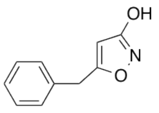


Table 1. Summary of Structure Determination of Compound 1390

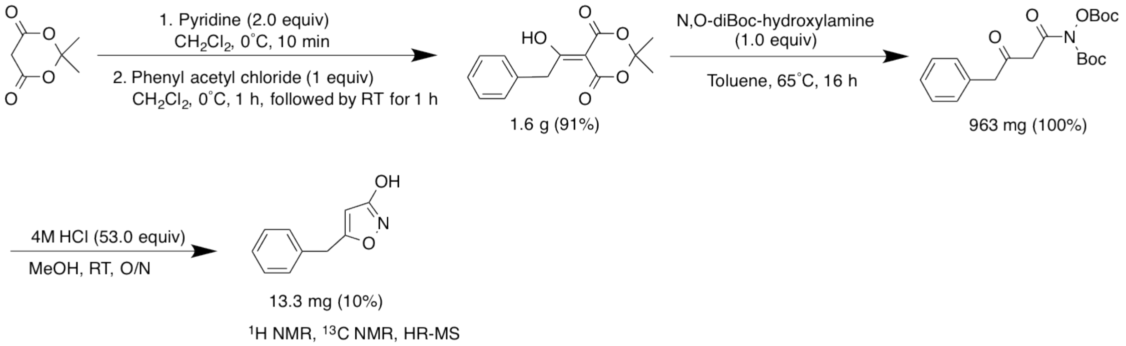
Empirical formula	$C_9H_{12}N_2SO_3$
Formula weight	228.27
Temperature	100(1) K
Wavelength	0.71073 Å
Crystal system	monoclinic
Space group	$P2_1/c$
Cell constants:	
a	12.2842(6) Å
b	9.0172(5) Å
c	9.8657(5) Å
β	105.348(2)°
Volume	1053.84(9) $Å^3$
Z	4
Density (calculated)	1.439 Mg/m^3
Absorption coefficient	0.296 mm^{-1}
F(000)	480
Crystal size	0.38 x 0.20 x 0.04 mm^3
Theta range for data collection	1.72 to 25.44°
Index ranges	$-14 \le h \le 14, -10 \le k \le 10, -11 \le l \le 11$
Reflections collected	14270
Independent reflections	1944 [R(int) = 0.0183]
Completeness to theta = 25.44°	99.9 %
Absorption correction	Semi-empirical from equivalents
Max. and min. transmission	0.7452 and 0.6856
Refinement method	Full-matrix least-squares on F^2
Data / restraints / parameters	1944 / 0 / 138
Goodness-of-fit on F^2	1.019
Final R indices [I>2sigma(I)]	R1 = 0.0271, wR2 = 0.0715
R indices (all data)	R1 = 0.0306, wR2 = 0.0746
Largest diff. peak and hole	0.310 and -0.378 e.$Å^{-3}$

`

APPENDIX B. SYNTHESIS OF 3-ISOXAZOLOL ISOSTERE

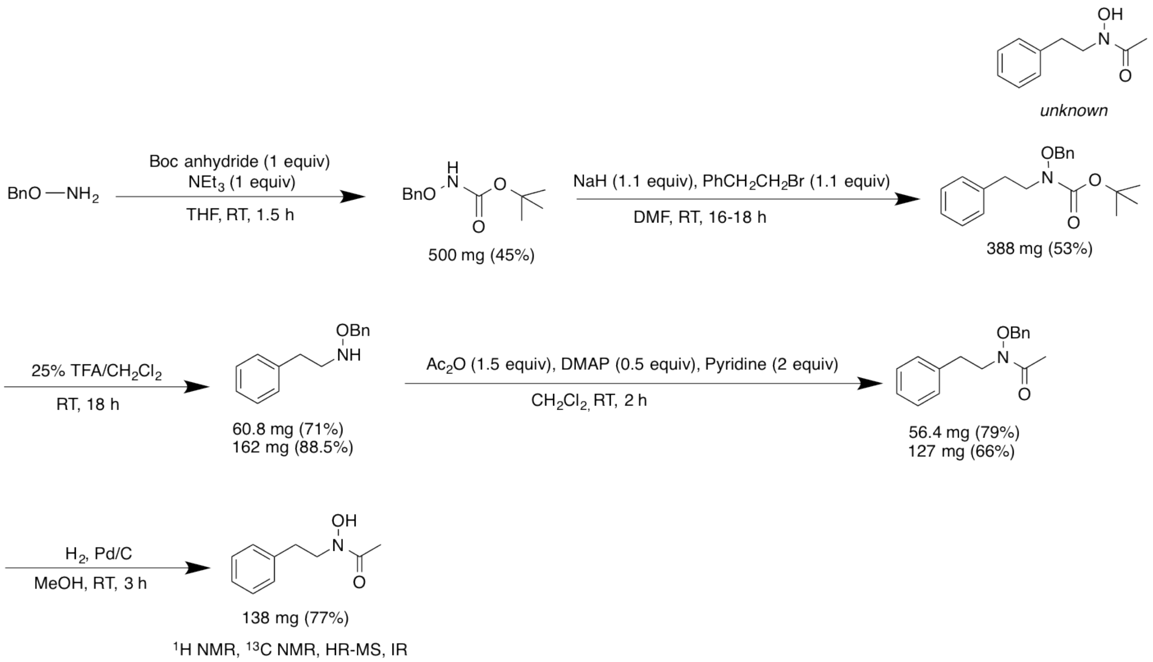


Known



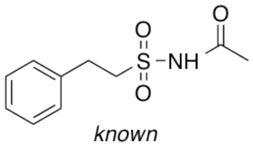
`

APPENDIX C. SYNTHESIS OF HYDROXAMIC ACID ISOTERE

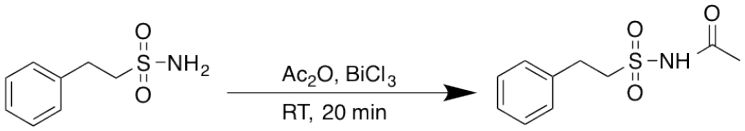


`

APPENDIX D. SYNTHESIS OF ACYLSULFONAMIDE ISOSTERE



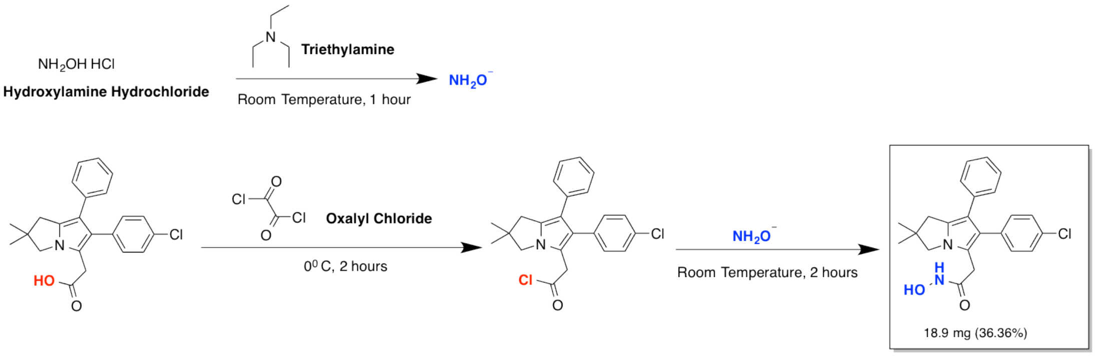
known



32.6 mg (crude)

`

APPENDIX E. SYNTHESIS OF LICOFELONE HYDROXAMIC ACID ISOSTERE



`

BIBLIOGRAPHY

1. Schames, J.R., R.H. Henchman, J.S. Siegel, C.A. Sotriffer, H. Ni, and J.A. McCammon, Discovery of a novel binding trench in HIV integrase. J Med Chem, 2004. 47(8): 1879-81.

2. Vonitzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Phan, T. V.; Smythe, M. L.; White, H. F.; Oliver, S. W.; Colman, P. M.; Varghese, J. N.; Ryan, D. M.; Woods, J. M.; Bethell, R. C.; Hotham, V. J.; Cameron, J. M.; Penn, C. R., Rational design of potent sialidase-based inhibitors of influenza-virus replication. Nature 1993, 363 (6428), 418-423.

3. Varghese, J. N.; Laver, W. G.; Colman, P. M. (1983). "Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 a resolution". Nature. 303 (5912)

4. Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W., Computational Methods in Drug Discovery. Pharmacological Reviews 2014, 66 (1), 334-395.

5. Goodsell, D. S.; Olson, A. J., Automated docking of substrates to proteins by simulated annealing. Proteins-Structure Function and Genetics 1990, 8 (3), 195-202.

6. Wang C, Wu H, Katritch V, et al. Structure of the human smoothened receptor bound to an antitumour agent. Nature. 2013 May 16;497(7449):338-43.

7. Wang C, Wu H, Evron T, Vardy E, Han GW, Huang XP, Hufeisen SJ, Mangano TJ, Urban DJ,Katritch V, Cherezov V, Caron MG, Roth BL, Stevens RC. Structural basis for Smoothened receptor modulation and chemoresistance to anticancer drugs. Nature Communications. 2014 Jul 10;5:4355.

8. Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J., Docking and scoring in virtual screening for drug discovery: Methods and applications. Nature Reviews Drug Discovery 2004, 3 (11), 935-949.

9. Zhu, K.; Cordeiro, M. L.; Atienza, J.; Robinson, W. E.; Chow, S. A., Irreversible inhibition of human immunodeficiency virus type 1 integrase by dicaffeoylquinic acids. Journal of Virology 1999, 73 (4), 3309-3316.

10. Bissantz, C.; Kuhn, B.; Stahl, M., A Medicinal Chemist's Guide to Molecular Interactions. Journal of Medicinal Chemistry 2010, 53 (14), 5061-5084.

11. Friere E. Do enthalpy and entropy distinguish first in class from best in class? Drug Discovery Today. 2008 Oct;13(19-20):869-74.

12. Chodera, J. D.; Mobley, D. L., Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. In Annual Review of Biophysics, Vol 42, Dill, K. A., Ed. Annual Reviews: Palo Alto, 2013; Vol. 42, pp 121-142.

13. Biela, A.; Nasief, N. N.; Betz, M.; Heine, A.; Hangauer, D.; Klebe, G., Dissecting the Hydrophobic Effect on the Molecular Level: The Role of Water, Enthalpy, and Entropy in

Ligand Binding to Thermolysin. Angewandte Chemie-International Edition 2013, 52 (6), 1822-1828.

14. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 1998, 19 (14), 1639-1662.

15. Bohm, H. J., The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. Journal of Computer-Aided Molecular Design 1994, 8 (3), 243-256.

17. Goodford, P. J., A computational-procedure for determining energetically favorable binding-sites on biologically important macromolecules. Journal of Medicinal Chemistry 1985, 28 (7), 849-857.

18. "Loll, P. J.; Picot, D.; Ekabo, O.; Garavito, R. M., Synthesis and use of iodinated nonsteroidal antiinflammatory drug analogs as crystallographic probes of the prostaglandin H-2 synthase cyclooxygenase active site. Biochemistry 1996, 35 (23), 7330-7340.

19. Huey, R.; Morris, G. M.; Olson, A. J.; Goodsell, D. S., A semiempirical free energy force field with charge-based desolvation. Journal of Computational Chemistry 2007, 28 (6), 1145-1152.

20. Stouten, P. F. W.; Frommel, C.; Nakamura, H.; Sander, C. An effective solvation term based on atomic occupancies for use in protein simulations. Molecular Simulation 1993, 10 (2-6), 97-&.

21. Goodsell, D. S.; Morris, G. M.; Olson, A. J., Automated docking of flexible ligands: Applications of AutoDock. Journal of Molecular Recognition 1996, 9 (1), 1-5.

22. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. Journal of Computational Chemistry 1998, 19 (14), 1639-1662.

23. Goodford, P. J., A computational-procedure for determining energetically favorable binding-sites on biologically important macromolecules. Journal of Medicinal Chemistry 1985, 28 (7), 849-857.

24. Goodsell, D. S.; Olson, A. J., Automated docking of substrates to proteins by simulated annealing. Proteins-Structure Function and Genetics 1990, 8 (3), 195-202.

26. Ravindranath, P. A.; Forli, S.; Goodsell, D. S.; Olson, A. J.; Sanner, M. F., AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. Plos Computational Biology 2015, 11 (12), 28.

27. Tubert-Brohman, I.; Sherman, W.; Repasky, M.; Beuming, T., Improved Docking of Polypeptides with Glide. Journal of Chemical Information and Modeling 2013, 53 (7), 1689-1699.

28. Forli, S.; Olson, A. J., A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking. *Journal of Medicinal Chemistry* **2012,** *55* (2), 623-638.

29. Leeson, P. D.; Springthorpe, B., The influence of drug-like concepts on decision-making in medicinal chemistry. Nature Reviews Drug Discovery 2007, 6 (11), 881-890.

30. Keseru, G. M.; Makara, G. M., The influence of lead discovery strategies on the properties of drug candidates. Nature Reviews Drug Discovery 2009, 8 (3), 203-212.

31.Silveran and Holladay. The organic chemistry of drug design and drug action Third Edition. Page 82. 2014 Elsevier.

32. Silveran and Holladay. The organic chemistry of drug design and drug action Third Edition. Page 75. 2014 Elsevier.

33. Silveran and Holladay. The organic chemistry of drug design and drug action Third Edition. Page 76. 2014 Elsevier.

34. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews 2012, 64, 4-17.

`

35. Veber, D. F., Molecular properties that influence the oral bioavailability of drug candidates. Abstracts of Papers of the American Chemical Society 2003, 225, U208-U208.

36. Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J., Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. Chemistry & Biology 2014, 21 (9), 1115-1142.

37. Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L., Quantifying the chemical beauty of drugs. Nature Chemistry 2012, 4 (2), 90-98.

38. Zhang, M. Q.; Wilkinson, B., Drug discovery beyond the 'rule-of-five'. Current Opinion in Biotechnology 2007, 18 (6), 478-488.

39. Hopkins, A. L.; Keseru, G. M.; Leeson, P. D.; Rees, D. C.; Reynolds, C. H., The role of ligand efficiency metrics in drug discovery. Nature Reviews Drug Discovery 2014, 13 (2), 105-121\

40- Heppner, F. L.; Ransohoff, R. M.; Becher, B., Immune attack: the role of inflammation in Alzheimer disease. Nature Reviews Neuroscience 2015, 16 (6), 358-372.

41- Manev, H.; Chen, H.; Dzitoyeva, S.; Manev, R., Cyclooxygenases and 5-lipoxygenase in Alzheimer's disease. Progress in Neuro-Psychopharmacology & Biological Psychiatry 2011, 35 (2), 315-319.

`

42- 24. Herbst-Robinson, K. J.; Liu, L.; James, M.; Yao, Y. M.; Xie, S. X.; Brunden, K. R., Inflammatory Eicosanoids Increase Amyloid Precursor Protein Expression via Activation of Multiple Neuronal Receptors. Scientific Reports 2015, 5, 16

43- Yao, Y. M.; Chinnici, C.; Tang, H. G.; Trojanowski, J. Q.; Lee, V. M. Y.; Pratico, D., Brain inflammation and oxidative stress in a transgenic mouse model of Alzheimer-like brain amyloidosis. Journal of Neuroinflammation 2004, 1, 9.

44- Cicero, A. F. G.; Laghi, L., Activity and potential role of licofelone in the management of osteoarthritis. Clinical Interventions in Aging 2007, 2 (1), 73-79.

45- Mancini, J. A.; Riendeau, D.; Falgueyret, J. P.; Vickers, P. J.; Oneill, G. P., ARGININE-120 OF PROSTAGLANDIN G/H SYNTHASE-1 IS REQUIRED FOR THE INHIBITION BY NONSTEROIDAL ANTIINFLAMMATORY DRUGS CONTAINING A CARBOXYLIC-ACID MOIETY. Journal of Biological Chemistry 1995, 270 (49), 29372-29377.

46- Ballatore, C.; Huryn, D. M.; Smith, A. B., Carboxylic Acid (Bio)Isosteres in Drug Design. Chemmedchem 2013, 8 (3), 385-395.

47- Carini, D. J.; Duncia, J. V.; Aldrich, P. E.; Chiu, A. T.; Johnson, A. L.; Pierce, M. E.; Price, W. A.; Santella, J. B.; Wells, G. J.; Wexler, R. R.; Wong, P. C.; Yoo, S. E.; Timmermans, P., Nonpeptide angiotensin-ii receptor antagonists - the discovery of a series of n-(biphenylylmethyl)imidazoles as potent, orally active antihypertensives. Journal of Medicinal Chemistry 1991, 34 (8), 2525-2547.

`

48- Lassalas, P.; Gay, B.; Lasfargeas, C.; James, M. J.; Tran, V.; Vijayendran, K. G.; Brunden, K. R.; Kozlowski, M. C.; Thomas, C. J.; Smith, A. B.; Huryn, D. M.; Ballatore, C., Structure Property Relationships of Carboxylic Acid Isosteres. *Journal of Medicinal Chemistry* **2016,** *59* (7), 3183-3203.

49- Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions: Marvin 14.9.1, 2014 , ChemAxon
(http://www.chemaxon.com)

50- Aparoy, P.; Reddy, K. K.; Reddanna, P., Structure and Ligand Based Drug Design Strategies in the Development of Novel 5-LOX Inhibitors. Current Medicinal Chemistry 2012, 19 (22), 3763-3778.

51-Martel-Pelletier, J.; Lajeunesse, D.; Reboul, P.; Pelletier, J. P., Therapeutic role of dual inhibitors of 5-LOX and COX, selective and non-selective non-steroidal anti-inflammatory drugs. Annals of the Rheumatic Diseases 2003, 62 (6), 501-509.

52- Rimon, G.; Sidhu, R. S.; Lauver, D. A.; Lee, J. Y.; Sharma, N. P.; Yuan, C.; Frieler, R. A.; Trievel, R. C.; Lucchesi, B. R.; Smith, W. L., Coxibs interfere with the action of aspirin by binding tightly to one monomer of cyclooxygenase-1. Proceedings of the National Academy of Sciences of the United States of America 2010, 107 (1), 28-33.

53- RBL assay; Bingham B.R., Monk P.N. and Helm B.A. (1994) Defective Protein Phosphorylation and Ca2+ Mobilization Linowa SecretingVariant of the Rat Basophilic Leukemia Cell Line. The Journal of Biological Chemistry: 269(30), pp. 19300-19306.

54. Iqbal, K.; Liu, F.; Gong, C. X., Tau and neurodegenerative disease: the story so far. Nature Reviews Neurology 2016, 12 (1), 13.

55. Wang, Y. P.; Mandelkow, E., Tau in physiology and pathology. Nature Reviews Neuroscience 2016, 17 (1), 5-21.

56. Neugroschl, J.; Sano, M., Current Treatment and Recent Clinical Research in Alzheimer's Disease. Mount Sinai Journal of Medicine 2010, 77 (1), 3-16.

57. Rimon, G.; Sidhu, R. S.; Lauver, D. A.; Lee, J. Y.; Sharma, N. P.; Yuan, C.; Frieler, R. A.; Trievel, R. C.; Lucchesi, B. R.; Smith, W. L., Coxibs interfere with the action of aspirin by binding tightly to one monomer of cyclooxygenase-1. Proceedings of the National Academy of Sciences of the United States of America 2010, 107 (1), 28-33.

58. Lampiasi, N.; Fodera, D.; D'Alessandro, N.; Cusimano, A.; Azzolina, A.; Tripodo, C.; Florena, A. M.; Minervini, M. I.; Notarbartolo, M.; Montalto, G.; Cervello, M., The selective cyclooxygenase-1 inhibitor SC-560 suppresses cell proliferation and induces apoptosis in human hepatocellular carcinoma cells. International Journal of Molecular Medicine 2006, 17 (2), 245-252.

59. Lamberth, C.; Dumeunier, R.; Trah, S.; Wendeborn, S.; Godwin, J.; Schneiter, P.; Corran, A., Synthesis and fungicidal activity of tubulin polymerisation promoters. Part 3: Imidazoles. Bioorganic & Medicinal Chemistry 2013, 21 (1), 127-134.

60. Slovic AM, Kono H, Lear JD, Saven JG, DeGrado WF. Computational design of water-soluble analogues of the potassium channel KcsA. Proc Natl Acad Sci U S A. 2004 Feb 17;101(7):1828-33.

61. Perez-Aguilar JM, Xi J, Matsunaga F, Cui X, Selling B, Saven JG, Liu R. A computationally designed water-soluble variant of a G-protein-coupled receptor: the human mu opioid receptor. PLoS One. 2013 Jun 14;8(6):e66009.

62. Cui T, Mowrey D, Bondarenko V et al. NMR structure and dynamics of a designed water-soluble transmembrane domain of nicotinic acetylcholine receptor. Biochim Biophys Acta. 2012 Mar;1818(3):617-26.

63. Sharpe, H. J.; Wang, W. R.; Hannoush, R. N.; de Sauvage, F. J., Regulation of the oncoprotein Smoothened by small molecules. Nature Chemical Biology 2015, 11 (4), 246-255.

64. Rubin LL, de Sauvage FJ. Targeting the Hedgehog pathway in cancer. Nature Reviews Drug Discovery. 2006 Dec;5(12):1026-33.

65. Amakye D, Jagani Z, Dorsch M. Unraveling the therapeutic potential of the Hedgehog pathway in cancer. Nature Medicine. 2013 Nov;19(11):1410-22.

66. Kelleher FC. Hedgehog signaling and therapeutics in pancreatic cancer. Carcinogenesis. 2011 Apr;32(4):445-51.

67. Xie J, Bartels CM, Barton SW, Gu D. Targeting hedgehog signaling in cancer: research and clinical developments. Onco Targets Ther. 2013 Oct 10;6:1425-1435.

`

68. Von Hoff DD, LoRusso PM, Rudin CM, et al. Inhibition of the hedgehog pathway in advanced basal-cell carcinoma. N Engl J Med. 2009 Sep 17;361(12):1164-72.

69. Metcalfe C, de Sauvage FJ. Hedgehog fights back: mechanisms of acquired resistance against Smoothened antagonists. Cancer Res. 2011 Aug 1;71(15):5057-61.

70. Kim J, Aftab BT, Tang JY, Kim D, Lee AH, Rezaee M, Kim J, Chen B, King EM, Borodovsky A, Riggins GJ, Epstein EH Jr, Beachy PA, Rudin CM. Itraconazole and arsenic trioxide inhibit Hedgehog pathway activation and tumor growth associated with acquired resistance to smoothened antagonists. Cancer Cell. 2013 Jan 14;23(1):23-34.

71. Frank-Kamenetsky M1, Zhang XM, Bottega S, Guicherit O, Wichterle H, Dudek H, Bumcrot D, Wang FY,Jones S, Shulok J, Rubin LL, Porter JA. Small molecule modulators of Hedgehog signaling: identification and characterization of Smoothened agonists and antagonists. Journal of Biology. 2002 Nov 6;1(2):10.

72. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. (2011). Theoretical and computational protein design. Annu Rev Phys Chem. 62:129-49.

73. Saven JG. (2011). Computational protein design: engineering molecular diversity, nonnatural enzymes, nonbiological cofactor complexes, and membrane proteins. Curr Opin Chem Biol. 15(3):452-7.

74. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D.

`

(2008). De novo computational design of retro-aldol enzymes. Science 319(5868):1387-91.

75. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. (2008). Kemp elimination catalysts by computational enzyme design. Nature 453(7192):190-5.

76. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D. (2010). Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329(5989):309-13.

77. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJ Jr, Stoddard BL, Baker D. Computational redesign of endonuclease DNA binding and cleavage specificity. (2006). Nature 441(7093):656-9.

78. Zou J, Saven JG. Statistical theory of combinatorial libraries of folding proteins: energetic discrimination of a target structure. J Mol Biol. 2000 Feb 11;296(1):281-94.

79. Kono H, Saven JG. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. J Mol Biol. 2001 Feb 23;306(3):607-28.

80. Richard F. M (1977). Annual Reviews of Biophysics and Bioengineering. 6, 151-176

`

81. Lee, B; Richards, FM. (1971). "The interpretation of protein structures: estimation of static accessibility". J Mol Biol. 55 (3): 379–400.

82. Rose, G. D.; Geselowitz, A. R.; Lesser, G. J.; Lee, R. H.; Zehfus, M. H., HYDROPHOBICITY OF AMINO-ACID RESIDUES IN GLOBULAR-PROTEINS. Science 1985, 229 (4716), 834-838.

83. Fauchere, J. L.; Pliska, V., Hydrophobic parameters-pi of amino-acid side-chains from the partitioning of n-acetyl-amino-acid amides. European Journal of Medicinal Chemistry 1983, 18 (4), 369-375.

84. Fraczkiewicz R, Braun W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. J Comput Chem. 1998 19: 319–333.

85. Reuten, R.; Nikodemus, D.; Oliveira, M. B.; Patel, T. R.; Brachvogel, B.; Breloy, I.; Stetefeld, J.; Koch, M., Maltose-Binding Protein (MBP), a Secretion-Enhancing Tag for Mammalian Protein Expression Systems. Plos One 2016, 11 (3), 15.

86. Greenfield, N. J., Using circular dichroism spectra to estimate protein secondary structure. Nature Protocols 2006, 1 (6), 2876-2890.

88. Cottrell, J. S., Protein identification using MS/MS data. Journal of Proteomics 2011, 74 (10), 1842-1851.

`

89. Zhang, R. M.; Monsma, F., Fluorescence-based thermal shift assays. Current Opinion in Drug Discovery & Development 2010, 13 (4), 389-402.

90. Greenfield, N. J., Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nature Protocols 2006, 1 (6), 2527-2535.

91. Niesen, F. H.; Berglund, H.; Vedadi, M., The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. Nature Protocols 2007, 2 (9), 2212-2221.

92. Rumin Zhang (Merck Research Laboratories), personal communication

93. Jordan et al. Fragment-based Drug Discovery: Practical Implication Based on ^{19}F Spectroscopy. Journal of Medicinal Chemistry 2012, 55, 678-687

94. Meyer and Peters. NMR Spectroscopy Techniques for Screening and Identifying Ligand Binding to Protein Receptors. Angew. Chem. 2003, 42, No.8

95. Diercks, Coles, Kessler. Applications of NMR to Drug Discovery. Current Opinion in Chemical Biology June 2001; 5(3): 285-291