PROBING INTERACTION MOTIFS FOR LIGAND BINDING PREDICTION FROM THREE PERSPECTIVES: ASSESSING PROTEIN SIMILARITY, LIGAND SIMILARITY AND COMPONENTS OF PROTEIN-LIGAND INTERACTIONS

By

Nan Liu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry – Doctor of Philosophy

2015

ABSTRACT

PROBING INTERACTION MOTIFS FOR LIGAND BINDING PREDICTION FROM THREE PERSPECTIVES: ASSESSING PROTEIN SIMILARITY, LIGAND SIMILARITY AND COMPONENTS OF PROTEIN-LIGAND INTERACTIONS

By

Nan Liu

The interactions between small molecules and diverse enzyme, membrane receptor and channel proteins are associated with important biological processes and diseases. This makes the study of binding motifs between proteins and ligands appealing to scientists. We use multiple computational techniques to unveil the protein-ligand interaction motifs from three perspectives. Firstly, from the perspective of proteins, by comparing the structure differences and common features of different binding sites for the same ligand, 3-dimensional motifs that represent the favorable interactions of the same ligands can be extracted. The goal is for such a motif to represent the shared features for binding a certain ligand in unrelated proteins, while discriminating from other ligands. The 3-dimensional motifs for cholesterol and cholate binding to non-homologous protein sites have been extracted, using SimSite3D alignment and analysis of the conserved interactions between these sites. The 3-dimensional protein motif for cholesterol binding can give about 80% accuracy of true positive sites with a low false positive rate. Furthermore, an online server CholMine was established so that the users can use this approach to predict cholesterol and cholate binding sites in proteins of interest. These motifs can help annotation of protein functions, drug discovery and the design of mutations.

Secondly, from the perspective of ligands, interaction motifs can be represented as

molecular features important for biological activities of ligands. Searching and summary of shared motifs from pretested series of ligand candidates can provide rational guidance to further drug improvement and screening. Here, we report a series of potential sea lamprey olfactory receptor 1 antagonists discovered from databases we designed of molecules that are similar to the native ligand, 3kPZS. Compounds with overall electrostatic and shape similarity to 3kPZS were assessed by using ROCS software, and their initial important feature matches to 3kPZS were analyzed, to prioritize compounds for biological testing. Then, the molecular features important to biological activities were summarized using SALI analysis and functional group matchprint analysis. By combining theses approaches, 12 compounds were discovered that suppress the detection of 3kPZS by at least 45%, and the most active compounds have entered field testing.

Thirdly, dissecting the components of protein-ligand binding energies is also important to define the key determinants of ligand interaction with a protein site. Through analyzing the correlation coefficient of interaction energies between a series of alpha-phenylalanine substitutes and PaPAM and biological activities of these compounds, the dominant factor that determine the activities of the compounds was revealed, which was steric effect between the binding site and these compounds. From the analysis, mutations at the residues of the binding site were suggested to change or improve the catalytic efficiency of the enzyme.

Given these three approaches, we envision a more integrated approach in the future that combines the analysis of shared protein-ligand interactions, shared interaction features from active ligands and shared features of protein binding sites to identify even more selective and tight-binding ligands.

This dissertation is dedicated to my beloved parents, younger brother and my husband.

ACKNOWLEDGEMENTS

Here I sincerely acknowledge all the people who offer me their kindness and helps during my graduate study. Without their support and guidance, I can’t finish my graduate study and dissertation.

First of all, with most gratitude, I thank to my advisor, Leslie Kuhn. She gives me a lot of support and technical guidance during my graduate study. She cultivates me with good research habits, such as writing weekly reports in scientific formats, always backing up the research data and so on. Furthermore, she encourages and inspires me to solve problems independently and innovatively. She is always approachable, patient and resourceful to me. She makes my stay at Michigan State University delightful, memorable and productive.

Secondly, I wound like to appreciate the helps from my committee members. I am grateful to Dr. Robert Cukier, Dr. Shelagh Ferguson-Miller and Dr. Kevin Walker. They give me many valuable feedbacks and suggestions on my research, based on which I can modify and improve my research skills. In addition, I learn a lot from the collaborations with Dr. Shelagh Ferguson-Miller on cholesterol prediction and the CholMine server building. Without Dr. Kevin Walker’s help, the PaPAM project cannot come to a cheerful end.

Thirdly, I owe my thanks to all of my collaborators. I need to thank to Dr. Fei Li and Dr. Jian Liu for their suggestions on CholMine software features to support experimental follow-up, and their feedbacks on this manuscript. I am grateful to Dr. Nishanka Dilini Ratnayake on the collaboration of the PaPAM project. I also appreciate Dr. Mar Huertas, Anne Scott (Graduate Research Assistant) and Dr. Weiming Li for their contributions to sea lamprey olfactory receptor 1 antagonist discovery project. Without their help and encouragement, I cannot produce the useful and instructive results for these projects.

Furthermore, I would like to appreciate my group members. John Johnson, the technique specialist, helped me to solve a lot of problems, such as implementation of CholMine server, various software environment setting problems and so on. Dr. Jeffrey Van Voorst, the developer of SimSite3D, gave me a lot of helps on how to use SimSite3D properly. Dr. Leann Buhrow helped me a lot when I started to work in the lab. Santosh Gunturu, a former undergraduate researcher in our lab, collaborated with me at the beginning of the CholMine and sea lamprey olfactory receptor 1 antagonist discovery projects. I also need to thank to Sebastian Raschka, Joe Bemister and Alex Wolf for their helps and suggestions on my research and presentations.

Lastly, I want to thank to my husband. He gave me unlimited love and support in my life and graduate study. Only under his support, understanding, and love, I can finish my graduate study and research. He gives me strength to overcome various problems in my life. I want to thanks to my parents. Without their unconditional love and support, I can’t overcome so many problems in my life. I owe my appreciations to all my families members.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

PDB: Protein Data Bank

vHTS: virtual high throughput screening

GPCRs: G-protein coupled receptors

SMILES: Simplified molecular-input line-entry system

SMARTS: SMiles ARbitrary Target Specification

RMSD: Root Mean Square Deviation

CcO: cytochrome c oxidase

CCM: cholesterol consensus motif

CRAC: cholesterol recognition amino acid consensus

TSPO: translocator protein

FXR: farnesoid X receptor

CLR: cholesterol

CHD: cholate

ATP: Adenosine triphosphate

*Pa*PAM: Phenylalanine aminomutases from the bacterium *Pantoea agglomerans*

MIO: 4-methylidene-1*H*-imidazol-5(4*H*)-one

vdW: van der Waals

SALI: structure-activity landscape index *ccoef*: correlation coefficient

3kPZS: 7a,12a,24-trihydroxy-3-one-5a-cholan-24-sulfate

SLOR1: sea lamprey olfactory receptor 1

GLL: GPCR ligand library

TAAR: trace amine-associated receptors

EOG: electro-olfactogram

TLC: taurolithocholic acid

Chapter 1 Introduction

1.1 Introduction

Understanding protein-ligand interaction motifs and the determinants for specific protein-ligand binding is the first step for scientists to uncover the secrets of protein-ligand recognition, and understand how small molecules regulate biological processes. There are about 68 000 protein–ligand complexes and 2 million ligand-binding sites found in all the protein-ligand 3-dimensional structures of the current Protein Data Bank (PDB, www.rcsb.org).[1] This enormous amount of structural data gives us the opportunity to mine protein-ligand binding motifs across different protein families to understand protein structures and their functions,[2] to modify enzyme functions,[3] and to discover novel drugs and pharmaceutical targets.[4]

1.1.1 Computer technologies used in biochemistry

Computer technologies have wide usage in biochemistry, including the study of protein-ligand interactions. In recent years, fast developments in computers, programming languages, and algorithms enable scientists to solve biochemical problems quantitatively and explain experimental data in sophisticated ways.

There are many resources on the Internet to retrieve protein data. For example, the Protein Data Bank (PDB, www.rcsb.org)[1] contains all the current protein and ligand 3D coordinates from X-ray crystal structures and NMR structures. PDBsum (http://www.ebi.ac.uk/pdbsum/)[5] allows users to analyze protein-ligand interactions using 2D LigPlot[6,7] figures. It also provides the binding cleft information for most protein structures[8].

In addition to providing protein data, there are also many online resources that can translate and store information regarding the small molecules that bind, or potentially can bind to proteins. ZINC 12 is an online accessible database (http://zinc.docking.org[9]) that provides structures of millions of chemical compounds, including many that are commercially available, which can be used for virtual high throughput screening (vHTS) in ligand discovery. Furthermore, computer graphic techniques also have broad usage in molecular modeling and structure comparisons. A graphic tool such as PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4, Schrödinger, LLC), allows scientists to visualize and manipulate 3D models of molecules on a graphical display device.

1.2 Representations of protein and small molecule structures

Learning about the different representations of protein and ligand structures is the beginning of utilizing that information to study protein-ligand interactions. In general, the information of protein and ligand structures can be stored in one-dimensional notation (1D) like SMILES strings or fingerprints, two-dimensional drawings (2D) and three-dimensional structures (3D), respectively, as shown in Figure 1.1. Different from 3D modeling, there is no quantitative data such as spatial coordinates of the structures in one-dimensional or two-dimensional representations.

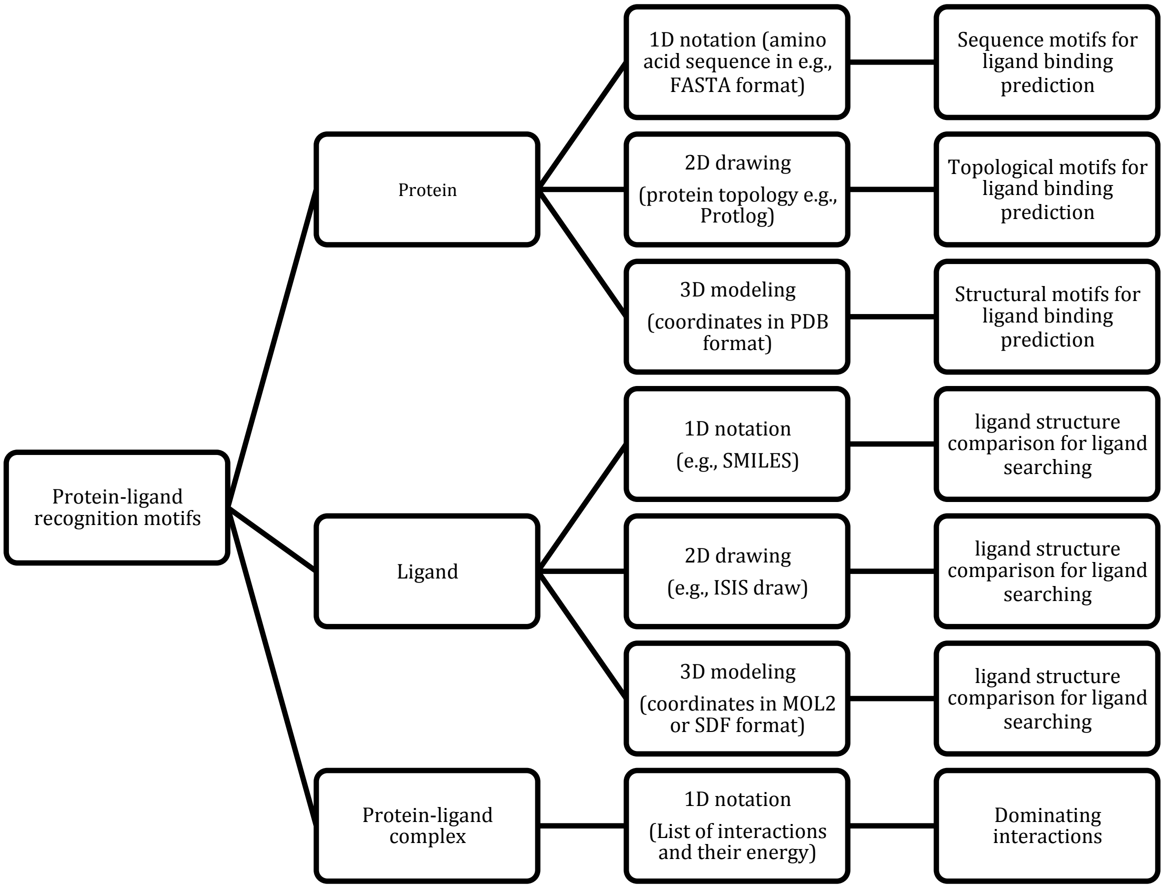

Protein-ligand recognition motifs
Protein
Ligand
Protein-ligand complex
1D notation (amino acid sequence in e.g., FASTA format)
2D drawing (protein topology e.g., Protlog)
3D modeling (coordinates in PDB format)
1D notation (e.g., SMILES)
2D drawing (e.g., ISIS draw)
3D modeling (coordinates in MOL2 or SDF format)
1D notation (List of interactions and their energy)
Sequence motifs for ligand binding prediction
Topological motifs for ligand binding prediction
Structural motifs for ligand binding prediction
ligand structure comparison for ligand searching
ligand structure comparison for ligand searching
ligand structure comparison for ligand searching
Dominating interactions

Figure 1.1 Representations of protein, ligand and protein-ligand interactions for structural and chemical similarity mining.

1.2.1 Protein structure representations and applications

An amino acid sequence or primary structure is a one-dimensional (1D) representation of protein structure. It can be written as a string of amino acids in one letter abbreviation, stored in various formats such as FASTA and is one of the most commonly used data types in bioinformatics. Through alignment between two or more amino acid sequences from different proteins, similarity scores between these sequences can be obtained, from which the structural, functional and evolutionary relationships of these proteins can be deduced. In addition, a commonly used drug discovery technique is to discover new ligands based on the structures of proteins using various docking tools in structure-based virtual screening.[10-13] However, sometimes the target protein structures are not available, especially for membrane proteins such as G-protein coupled receptors (GPCRs), which make up 40% of pharmaceutical targets in the pharmaceutical industry.[14,15] Under these circumstances, 3-dimensional models of these proteins can be built first if the degree of sequence similarity is high enough with an existing 3D structure, to enable structure-based virtual screening. Amino acid sequence alignment is the critical step of homology modeling.[16] The threshold of sequence identities to build a reliable homology model for the unknown protein must be above 25% over at least 80 aligned residues.[17]

Protein topology can be used to represent a protein's structure in 2D diagrams, describing the orientation and connection information of secondary structure elements (SSEs) of that protein.[18] Even though protein topology neglects the atomic information as shown in 3D structures, it can show SSEs in a way that helps scientists to analyze protein folds. This aids the annotation of protein families, domains, and functions, and the study of evolutionary relationships. However, there are limited applications of protein topology in drug discovery, due to the relatively a few types of protein topology found related to specific ligand binding.

3D structures of proteins allow the shape and chemical features of binding sites to be studied thoroughly. Unlike 1D and 2D representations, the advantage of 3D models of proteins is that they can be analyzed independent of specific protein residues or connections. Instead, the 3D models supply a spatial perspective to study the chemical interactions between proteins and ligands, from which the interaction details such as hydrogen-bonding can be analyzed. The details of the interactions can help us understand the determinants for specific ligand binding. In drug discovery, if the binding site residues that are important for ligand binding are known, it is possible to improve protein activities through modifying these residues.

1.2.2 Molecular structure representations and applications

Simplified molecular-input line-entry system (SMILES)[19,20] is a commonly used 1D representation of chemical compound structures in cheminformatics. It is a string of characters to describe molecular formulas, atomic connections, bond types and chiral atoms in molecules. A unique SMILES using a canonicalization algorithm is called canonical SMILES. SMILES containing isotopic and stereochemical information are called isomeric SMILES. SMILES strings are similar to condensed structural formulas while still having some differences. In SMILES, the lower-case letters are for aromatic atoms, and the capital letters are for aliphatic atoms. "@" is for anticlockwise chirality and "@@" is for clockwise chirality. SMiles ARbitrary Target Specification (SMARTS),[21, 22] an extension of generic SMILES, allows the representation of broader structural patterns for searching chemical compounds. Any valid SMILES expressions are valid SMARTS string, not vice versa. In SMARTS, "*" indicates that any atom can match,

“~” indicates that any bond can match. For example, SMARTS for a steroid ring are shown in Figure 1.2.

SMILES and SMARTS have broad usages in structure retrieving, substructure and similarity searching that enable identifying relevant sets of compounds to analyze. For example, based on isomeric SMILES, 3D structures of molecules with stereochemistry information can be generated. Based on SMARTS, Root Mean Square Deviation (RMSD) of substructures representing the closeness with which the molecules can be overlaid can be calculated. Online chemical resources, such as PubChem,[23] ZINC12,[9] and SCIFINDER (https://scifinder.cas.org/), provide interfaces to allow the user to enter SMILES strings for exact, substructure and similarity searches.

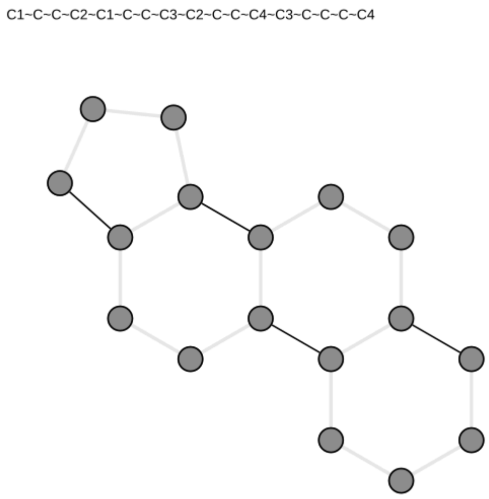


Figure 1.2 SMARTS and 2D sketching of steroid ring using SMARTSViewer.

2D drawings of chemical structures, such as ISIS drawing, provide direct views of molecular 2D structures, including atom types, bond connections and stereochemistry properties. 2D drawings, just like 1D notations, are often used for searching molecules with exact, similar or sub- structures. Current online resources, such as PubChem,[14] ZINC12 database,[9] provide the interface to let user draw the structures of molecules they are interested in, from which structure exact, substructure and similarity search can be performed.

In 3D models of molecules, not only the atom types, connectivity and stereochemistry information can be viewed directly, but also the molecule shape and electrostatic distribution can be depicted too. This allows molecular comparisons not only as ID strings and 2D connectivity diagrams, but also as 3-dimensional structures reflecting bioactive or other conformations. 3D modeling can go further by providing opportunity to calculate molecular similarity based on entire molecular shape and charge distribution, regardless of specific atom types and connections. After we know the presentations and applications of protein and ligand structures, how to translate the information into more general and applicable information to help study of protein-ligand interaction and further ligand discovery is still challenging.

1.3 Predicting ligand binding only given protein information

Starting from tools such as those described above, given only protein information, can we predict which ligands can bind in the given site of a protein? What is the relationship between protein motifs and specific ligand binding? To answer these questions, there is generally a two-step methodology. The first step is to obtain motifs through comparisons of protein information. Comparison of protein sequence information (1D) can provide potential protein interaction motifs, while comparison of protein structures in 2D can define topological motifs. Protein structural motifs in 3D can be defined by similar main-chain motifs and as we show in Chapter 2, can be generalized beyond residue correspondingly. Once a potential ligand binding motif has been defined, its predictive value can be evaluated statistically.

1.4 Combination of multiple techniques in drug discovery

All of the techniques to utilize the structures and properties of protein and small molecules can be integrated into virtual high throughput screening techniques from the ligand comparison, protein comparison or protein-ligand interaction perspectives. Virtual screening is a powerful and successful tool at the initial stage of drug or inhibitor discovery. It can aid scientists to discover lead compounds in one of the most efficient ways, not only because it integrates currently accessible information of structures, properties and functions of protein and small molecules, but also because of the fast and efficient screening speed and low cost.[13, 24, 25] There are two major common techniques in virtual screening techniques, one being structure-based virtual screening and the other being ligand-based virtual screening. In structure-based virtual screening, the structure of the target protein can be used for small ligand docking. In small ligand docking, millions of compounds are docked at the binding site of the target protein and evaluated by different scoring function.[10-13] Compounds with higher docking scores should have a higher probability to interact with the target protein. Secondly, by comparing a given potential binding cleft on a protein to all ligand-bound clefts in the Protein Data Bank, two kinds of information can be gained: which ligand(s) bind to similar sites, and which other proteins might be off-target hits, presenting specificity issues for a given designed inhibitor or agonist. However, docking results and site comparisons are influenced by the quality of protein structures, especially when target proteins have low-resolution crystal structures or homology models are used. Under these circumstances, ligand based virtual screening techniques can find compounds with similar structures to the native substrates or known ligands of the target proteins.[11, 26] The hypothesis of ligand-based screening is that the compounds with similar structures are likely to have similar biological activities.

1.5 Objectives of this dissertation

Given a binding pocket on membrane or soluble proteins, the goal of this research is to predict the most likely ligand or native lipid by comparing the site with established predictors, only using protein information. Because the native ligands or lipids binding to most membrane-exposed sites are undefined, due to the low resolution of structure determination or displacement by detergent or lack of crystal structures, a ligand binding predictor can provide good hypotheses as to the native ligand(s) that can be validated by experimental results. Our group has collaborated with three experimental groups working on bile acid and cholesterol binding: Professors Ferguson-Miller and Atshaves in the Biochemistry & Molecular Biology Department and Professor Li in the Fisheries & Wildlife Department. The initial focus lies on prediction of cholesterol (CLR) or cholate binding sites, and characterization of determinants that distinguish CLR or cholate binding from sites binding other molecules. This approach can elucidate whether the determinants of binding for cholesterol or cholate are the same in membrane proteins as in soluble proteins and the difference between cholesterol and cholate binding. This project is described in Chapter 2.

REFERENCES

REFERENCES

(1) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P.E., The Protein Data Bank *Nucleic Acids Research*, 2000, *28*, 235-242.

(2) Grishin, N. V. Fold Change in Evolution of Protein Structures. *Journal of Structural Biology*, 2001, *134*, 167–185.

(3) Gutteridge, A., Thornton, J.M. Understanding nature's catalytic toolkit. *Trends Biochem. Sci.* 2005, *30*, 622–629.

(4) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* 2007, *152*, 38–52.

(5) de Beer, T. A. P; Berka, K.; Thornton, J. M.; Laskowski, R. A. PDBsum additions. *Nucleic Acids Res.* 2014, *42*, 292-296.

(6) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *Protein Eng.* 1996, *8*, 127-134.

(7) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand−Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* 2011, *51*, 2778−2786.

(8) Laskowski, R. A. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics* 1995, *13*, 323-330.

(9) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* 2012, *52*, 1757−1768.

(10) Kroemer, R. T. Structure-Based Drug Design: Docking and Scoring. *Current Protein and Peptide Science*, 2007, *8*, 312-328.

(11) Reddy, A. S.; Pati, S. P.; Kumar, P. P.; Pradeep, H. N.; Sastry, G. N. Virtual Screening in Drug Discovery – A Computational Perspective. *Current Protein and Peptide Science*, 2007, *8*, 329-351.

(12) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins.* 2004, *56*, 235–249.

(13) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat Rev Drug Discov.* 2004, *3*, 935-949.

(14) Flower, D. R. Modelling G-protein-coupled receptors for drug design. *Biochim Biophys Acta* 1999, *1422*, 207–234.

(15) Robas, N.; O'Reilly, M.; Katugampola, S.; Fidock, M. Maximizing serendipity: strategies for identifying ligands for orphan G-protein-coupled receptors. *Curr Opin Pharmacol.* 2003, *3*, 121–126.

(16) Chothia, C.; Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 1986, *5*, 823–826.

(17) Sander C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins*, 1991, *9*, 56-68.

(18) Rawlings, C. J.; Taylor, W. R.; Nyakairu, J.; Fox, J.; Sternberg, M. J.E. Reasoning about protein topology using the logic programming language PROLOG. *Journal of Molecular Graphics*, 1985, *3*, 151-157.

(19)Weininger, D. *SMILES 1. Introduction and Encoding Rules", J. Chem. Inf. Comput. Sci.* 1988, *28*, 31.

(20)James, C. A.; Weininger, D. Daylight Theory Manual. Daylight Chemical Information Systems, Inc: 27401 Los Altos, 2006.

(21) Schomburg, K.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. From Structure Diagrams to Visual Chemical Patterns, *J. Chem. Inf. Model.*, 2010, *50*, 1529-1535.

(22) Schomburg, K.;Ehrlich, H.; Stierand, K.. Chemical pattern visualization in 2D–the SMARTSviewer. *Journal of Cheminformatics*, 2011, *3*, O12.

(23) Bolton, E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN *Annual Reports in Computational Chemistry*, Volume 4, American Chemical Society, Washington, DC, 2008 Apr.

(24) Doman, T. N.; McGovern SL; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D.T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* 2002, *45*, 2213–2221.

(25) Zarzycka, B.; Seijkens, T.; Nabuurs, S. B.; Ritschel, T.; Grommes, J.; Soehnlein, O.; Schrijver, R.; van Tiel, C. M.; Hackeng, T. M.; Weber, C.; Giehler, F.; Kieser, A.; Lutgens, E.; Vriend, G.; Nicolaes, G. A. F. Discovery of Small Molecule CD40−TRAF6 Inhibitors. *J. Chem. Inf. Model.*, 2015, *55*, 294–307.

(26) Krüger, D. M.; Evers, A. Comparison of structure- and ligand-based virtual screening protocols considering hit list complementarity and enrichment factors. *ChemMedChem.* 2010, *5*, 148-58.

Chapter 2 Decoding protein structural motifs for ligand binding prediction

Reprint (adapted) with permission from CholMine: Determinants and Prediction of Cholesterol and Cholate Binding Across Nonhomologous Protein Structures. Nan Liu, Jeffrey Van Voorst, John B. Johnston and Leslie A. Kuhn. *J. Chem. Inf. Model.* 2015, *55*, 747–759.

2.1 Introduction

Given a series of non-homologous proteins binding the same ligands, we show that binding motifs can be extracted from protein information alone. The lack of generalized binding motifs for certain ligands inspires us to use our local structure alignment tool SimSite3D to extract abstract and generalized motifs for the ligands we are particularly interested in, for example, cholesterol and cholate. By deciphering the determinants of binding for these important steroids, the CholMine tool we developed (which incorporates SimSite3D site alignment) may also aid in the design of selective inhibitors and detergents for targets such as G protein coupled receptors and bile acid receptors.

2.1.1 Conserved lipid binding sites in membrane proteins

Membrane proteins are surrounded by a complex mixture of lipids, including phospholipids, cholesterol and some bile salts (bile acids and alcohols). One of the bile salts, cholate, is often used as a detergent to solubilize membrane proteins.[1,2] Different types of lipids influence biological functions of membrane proteins in direct or indirect ways.[3,4,5] Conserved binding sites for certain lipids have been characterized on membrane proteins,[4,6,7] and these lipids can play an important role in structural stabilization and biological processes. For example, in bovine heart cytochrome c oxidase (CcO), the tails of two phosphatidylglycerol lipids regulate oxygen transfer to the active site, and phosphatidylethanolamine, cardiolipin, and phosphatidylglycerol are all associated with the dimer interface.[4,6] Detergents can occupy natural lipid sites under different experimental conditions.[7] For example, phosphatidylcholine in bovine CcO and the detergents decyl maltoside in *Rhodobacter sphaeroides* and lauryldimethylamine oxide in *Paracoccus denitrificans* CcO occupy the same crevices of the proteins in different crystal structures[7]. Defining the determinants of lipid binding can help scientists understand the structural basis for the specificity of these sites, and aid in the design of site-selective ligands and detergents for protein purification and structure determination.

2.1.2 Cholesterol, cholate and related sequence-based binding motifs

Cholesterol (Figure 1(A)) plays an important role in the function of many biological systems, including eukaryotic, viral and prokaryotic proteins. While cholesterol is often considered important because of its role in membrane organization, including lipid rafts,[8] cholesterol also exerts important regulatory effects via direct, specific binding to proteins. Through binding to the nicotinic acetylcholine receptor and many G protein-coupled receptors (GPCRs), cholesterol modifies the receptors' affinity for agonists.[9] Additionally, mutations in the cholesterol-binding sites of virus envelope proteins, such as the HIV protein gp41 and Semliki Forest virus E1 protein, inhibit virus invasion at the fusion and budding stages.[10] In addition, cholesterol binding by podocin and MEC-2, members of the prohibitin domain family, is essential for regulating the activity of their ion channel partners.[11]

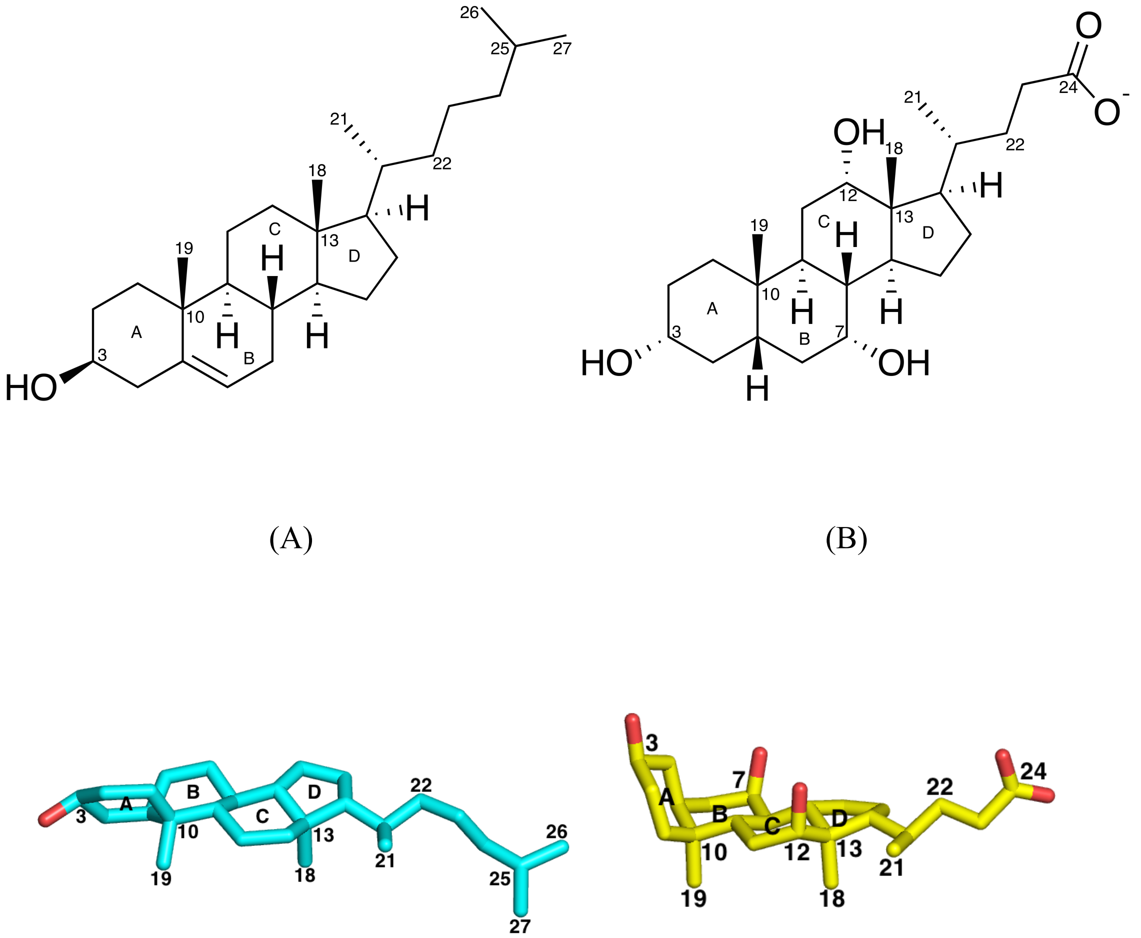


Figure 2.1 2D and 3D chemical structures of (A) cholesterol (blue) and (B) cholate (yellow), with the flexible tails from C21 to C24/C25 shown in arbitrary favorable conformations.

A recent proteomic study mapped cholesterol-protein interactions in mammalian cells with photoreactive sterol probes, followed by quantitative mass spectrometry.[12] Their work identified over 250 cholesterol binding proteins, including some known to biosynthesize, transport and regulate cholesterol, as well as others known to regulate sugars and glycerolipids or participate in vesicular transport and protein glycosylation and degradation.

Cholesterol-binding sequence motifs have been proposed for several protein families. For instance, a cholesterol consensus motif (CCM) has been identified in class A GPCRs as matching the amino acid sequence R/K-$(X)_{1-7}$-I/V/L-$(X)_{1-3}$-W/Y on one transmembrane alpha helix. The "strict CCM" also contains F/Y on a neighboring helix, based on residue conservation analysis between known cholesterol sites.[13] An expanded version of the CCM includes serine/glycine in one helix that forms an interhelical hydrogen bond with the CCM W/Y residue on an adjacent helix. The additional hydrogen bond is proposed to adjust the orientation of the aromatic side chain to enhance its stacking interactions with the steroid ring system.[14] A similar motif, the cholesterol recognition amino acid consensus or CRAC motif, has been defined in the outer mitochondrial membrane translocator protein (TSPO; also known as the peripheral benzodiazepine receptor). This consensus motif is L/V- $(X)_{1-5}$-Y- $(X)_{1-5}$-R/K, based on the loss of cholesterol uptake in TSPO Y153 and R156 mutants and alignment of this sequence region with other cholesterol binding proteins.[15,16] Recently, an enhanced version of the CRAC motif, LAF-CRAC, has been shown to be associated with nanomolar affinity for cholesterol in TSPO.[17] CARC, a cholesterol binding motif in the nicotinic acetylcholine receptor,[18] and a tilted peptide cholesterol binding motif have also been described.[19,20]

However, sequence motifs derived from one protein family often do not generalize well to predicting cholesterol-binding sites in other families, and these sequence motifs also match sites that do not bind cholesterol. For instance, analysis of 2,100 proteins in a bacterium that does not contain cholesterol found 5,000 matches to the CRAC motif.[21] Additional cholesterol binding sites are known that do not match any previously known motifs, for instance, the additional cholesterol sites known in some class A GPCRs. A GXXXG motif has been found to be critical for cholesterol binding to the β-amyloid precursor protein, as characterized by cholesterol titration and mutagenesis.[22] Cholesterol binding to this protein has been proposed to promote amyloidogenesis in Alzheimer's disease.[23] For cytolytic toxin recognition of cholesterol, a simple motif composed of a threonine-leucine pair in loop L1 has been identified by mutation analysis.[24] Thus, cholesterol binding sequence motifs appear to be fairly specific to protein families. Our aim is to uncover general features of cholesterol recognition that are shared by different protein families, and which discriminate cholesterol binding sites from other ligand sites. These features can then be tested for their ability to capture a broader range of cholesterol-binding sites via application of the resulting predictor, CholMine.

Prediction of cholate binding sites also attracts our attention for several reasons. Cholate (Figure 2.1(B)) is used extensively as a membrane protein solubilizing detergent.[1,2] Crystal structures show cholate occupying binding pockets on membrane proteins, and this molecule shares significant similarity with cholesterol in shape and steroidal chemistry, aside from its dissimilar polar tail. Cholate, a bile acid, functions in some cells as a steroid hormone that binds to nuclear receptors to modulate gene expression.[25] Several soluble nuclear receptors have been reported to bind bile acids, including farnesoid X receptor (FXR), liver X receptor alpha, and cyclopentyladenosine receptor. The resulting complexes stimulate or suppress gene transcription by binding to promoter regions.[25] Cholate is also one of the two major bile acids synthesized from cholesterol and plays an essential role in the absorption of fat and lipidic vitamins, by forming micelles to solubilize fat.[26,27] Cholate has been shown to be an agonist for the human bile acid G protein coupled receptor TGR5, involved in suppression of macrophage function.[28,29] Lastly, a relative of cholate, 3-keto petromyzonol sulfate, acts as a vertebrate pheromone through interaction with two other GPCRs.[30] Thus, understanding the determinants of cholate binding and identifying features that distinguish between cholate and cholesterol sites will be useful for designing site-selective ligands and detergents for stabilizing and purifying membrane proteins, and for interpreting ambiguous electron density in crystallography.

2.1.3 Determinants of lipid-membrane protein binding

What is known about the determinants for protein interaction with lipids, in general? Four important factors can be summarized from the literature. The first is the presence of aromatic residues such as tryptophan (W), tyrosine (Y) and phenylalanine (F). Tryptophan and tyrosine are preferred at membrane interfaces.[31] In the Ballesteros-Weinstein numbering scheme to facilitate comparison of G-protein coupled receptors (GPCRs), residues are labeled by two indices, X.Y, the first indexing the transmembrane helix number in which the residue occurs, and the second indicating the position within the helix. The position number 50 is assigned to the most highly conserved position in each helix, with numbers increasing towards the C-terminus.[32]

The Trp residue at position 4.50 in class A GPCRs, involved in cholesterol binding, is highly conserved (94%).[13] Aromatic residues contribute to cholesterol binding through favorable π and hydrophobic interactions with the steroid ring system of cholesterol.[13] The second class of residues contributing to lipid binding includes the positively charged residues lysine (K), arginine (R), and histidine (H), which form electrostatic interactions with the polar or negatively charged head groups of lipids.[3,33] Uncharged polar residues such as serine (S), threonine (T), and cysteine (C) also contribute by forming hydrogen bonds with lipids (where cysteine acts as a weak hydrogen-bond acceptor).[3,33] The last class of residues involved in lipid binding includes the moderately bulky hydrophobic residues isoleucine, leucine, and valine (I, L, V), as found in the CCM and CRAC motifs. Position 6.57 in GPCRs is conserved with isoleucine and valine in adenosine receptors.[34] These residues form van der Waals interactions with the hydrophobic part of lipids, participate in stacking interactions, and form hydrophobic grooves for binding.[3,31,34]

2.1.4 Previous prediction of lipid binding sites

Regions of lipid interaction have also been predicted using entire amino acid sequences, rather than motifs, along the lines of the transmembrane protein segment predictors that became popular in the 1980s. However, this type of prediction typically focuses on annotating membrane spanning regions of the protein sequence and does not provide information about pockets comprised of discontiguous parts of the protein that bind lipids tightly, the kind of lipid occupying each pocket, or the chemical and spatial determinants of lipid specificity. For example, different categories of lipid-interacting proteins have been predicted, according to lipid degradation, metabolism, synthesis, transport, and other functions, by using amino acid sequence information from the SwissProt database.[35] In addition, residues involved in lipid binding have been predicted based on amino acid sequence and residue conservation using a support vector machine.[36] However, this approach does not provide spatial or lipid-specificity information that extends to new protein classes. Lipid-binding sites in several key cytoskeletal proteins have been predicted using a matrix-based algorithm to identify highly hydrophobic or amphipathic amino acid segments,[37] again predicting transmembrane secondary structure segments rather than pockets where lipids bind tightly and specifically. The goal of the work presented here is to identify the shared chemical determinants of cholesterol and cholate binding across non-homologous protein sites, and develop a sensitive and specific predictor for these sites.

2.2 Methods

Our identification of the determinants for cholesterol and cholate binding employs SimSite3D to align and quantify the similarity between pairs of binding sites.[38] The predictive accuracy is enhanced by incorporating knowledge of conserved interaction hotspots shared by cholesterol or cholate binding sites. In developing the CholMine predictor, we test the hypothesis that cholesterol (or cholate) binding in different proteins involves a characteristic set of interactions that distinguish cholesterol/cholate binding from other ligands.

2.2.1 SimSite3D and site maps for aligning and comparing protein sites

To align pairs of non-homologous protein sites and find the relative orientation with maximum shape and chemical similarity in the absence of ligand information, we use SimSite3D.[38,41] This method aligns two protein sites based on their similarity in surface shape and chemical features, without requiring underlying sequence or structural similarity. For a given query site, the similarity to another site is measured in standard deviations relative to the query's mean score when aligned to all cases in a set of 140 ligand-binding sites (including one cholesterol site) chosen from proteins with undetectable sequence and structural homology to one another, representing a highly diverse set of ligand sites (Table A.1.1). This Z-score measures the statistical significance of a match. An alignment between two sites with a SimSite3D score less than -1.5 (in standard deviation units, where more negative values indicate greater similarity) results in 2 Å RMSD or better site alignment in 80% of cases, based on tests across pterin, adenine, peptide and xenobotic binding sites from which the ligand has been removed.[38,41] SimSite3D alignment and scoring can also discriminate binding sites with similar chemical features that do not bind the same ligand. By contrast, other ligand site prediction methods either use information for both the ligand and receptor,[39] or they only predict binding sites with high sequence similarity within certain protein families such as GPCRs.[40]

The site map representation used by SimSite3D is a set of chemically labeled points in 3-dimensional space derived from residues in a user-defined or known ligand binding site. The site map represents a negative chemical image of the protein, indicating ideal positions for ligand atoms of a given chemistry to interact favorably with the protein. Each site map point can be related back to the corresponding protein atom(s). Hydrophobic site map points are set down discretely in a hemispherical array around hydrophobic protein atoms based on internal protein coordinates, such that two perfectly overlaid identical side chains will have exactly matching hydrophobic points, regardless of their initial Cartesian coordinates. Similarly, polar points are generated according to the favored geometry of hydrogen bonds relative to donor or acceptor groups in the protein (as is done for SLIDE docking templates[42]), with hydrogen-bond donor-acceptor atom interactions in the range of 2.5-3.5 Å, and the angle between the donor, hydrogen and acceptor atoms falling between 120° and 180°. In SimSite3D, the matches of hydrogen-bonding groups are scaled according to the extent to which their hydrogen bonding vectors point in the same direction, based on the colinearity of (cosine of the angle between) their donor-acceptor vectors. Exact overlap (angle of 0°) yields a weight of 1 for the hydrogen bond match, and an angle of 90° yields a weight of 0. In the CholMine implementation, the boundaries of a site map are determined either by user specification of a set of residues comprising the cleft to be analyzed, or by a set of ligand atom coordinates (which can be based on an experimentally determined or hypothesized ligand position that the user would like to assess). The ligand coordinates are then used to define a volume for site map generation, by selecting the set of protein residues containing at least one atom within 4.5 Å of one or more ligand atoms. SimSite3D reads ligand coordinates in Tripos mol2 format for site map generation. Ligand coordinates are converted from PDB format to mol2 format, as needed, by using the molcharge utility in QuacPac v. 1.3.1, utilizing OEChem toolkit v. 1.6.1 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com).

2.2.2 Extraction of an interaction motif for binding the same ligand in non-homologous sites

The goal of this work is to identify a motif that characterizes the binding of cholesterol (or cholate) across non-homologous proteins. For moderately to highly polar ligand sites, the SimSite3D score, which calculates the degree of chemical match between two sets of aligned site map points and their degree of molecular surface shape similarity, is usually sufficient to filter out false positive site matches while aligning and detecting most of the true positive sites. However, cholesterol sites are unusually hydrophobic, and the degree of conservation of polar interactions between non-homologous cholesterol sites is low, particularly because crystal structures show that the cholesterol hydroxyl moiety is often exposed to bulk water rather than interacting directly with protein atoms. As a result, CholMine employs SimSite3D to align and score a pair of site maps, and then determines whether this alignment matches a majority of conserved points of hydrophobic interaction identified from known cholesterol binding sites. Table 2.1 lists protein structures containing the twelve low-homology cholesterol sites, which were divided into two sets: the first set for training to detect conserved points of cholesterol interaction, and the second set for unbiased testing of cholesterol site predictions on a series of unrelated proteins. The cholesterol sites from dogfish and pig sodium-potassium pump proteins (PDB entries 2ZXE and 3KDP) were both included in the training set because their cholesterol binding residues were in different conformations. The number of independently determined, well-resolved, non-homologous cholesterol binding sites in the Protein Data Bank is limited, likely due to the extreme difficulty in handling this ligand, which has extremely low aqueous solubility. However, including several cholesterol sites from the same protein family would bias towards identifying a family-specific motif, whereas the goal here is to discover the chemical determinants of cholesterol binding sites in general. Therefore, we tested the extent to which the cholesterol binding motif determined from the training set cases can predict cholesterol sites well in other proteins, including: the non-homologous cholesterol binding sites in the test set, a series of cholesterol-binding class A GPCR structures showing sequence and conformational diversity, a set of non-cholesterol steroid binding sites, a set of aliphatic lipid binding sites, a set of 109 bacterial membrane proteins that do not contain cholesterol binding sites, and 139 soluble protein sites known to bind ligands other than cholesterol. Including only membrane protein cholesterol binding sites in the training set and only soluble sites in the training set (and then inverting the sets) allowed us to further test whether cholesterol binding motifs are similar in these different cellular environments.

To determine the conserved cholesterol contacts shared by diverse binding sites, CholMine employs the binary string output of SimSite3D (Figure 2.2), representing spatially aligned SimSite3D interaction points. Once a set of known cholesterol or cholate training sites has been aligned by SimSite3D based on matching the 3-dimensional site map points and the surface shape derived from protein atom coordinates alone, the software determines which site map points overlay in 3-dimensional space and have the same chemical interaction type (are conserved between the sites). The most highly conserved interaction points can then serve as a fingerprint, or filter, that aids in recognizing cholesterol sites.

The determination of conserved interaction points can be conceptualized as a matrix of SimSite3D-aligned site map points (Figure 2.2) indexed relative to the points they match spatially in the representative site, which is the site with the highest degree of interaction point conservation with the other cholesterol sites. This procedure results in the unbiased detection of a 3-dimensional binding motif corresponding to shared interactions in non-homologous sites binding cholesterol, as indicated by the highlighted vertical green bars showing points of interaction common to 70% or more of the sites (Figure 2.2).

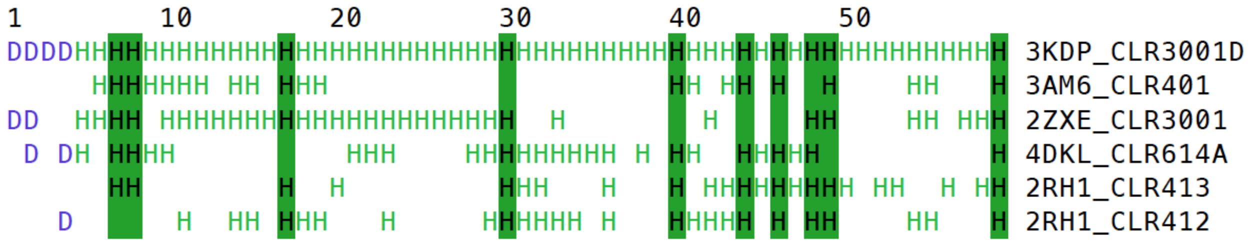
Table 2.1 Cholesterol binding proteins in the training and test sets.

Training set: membrane proteins

PDB code	Ligand	Source	Res.(Å)	R-factor	Protein Name
2RH1	Cholesterol	*H. sapiens*	2.4 Å	0.198	β2-adrenergic G protein-coupled receptor
3AM6	Cholesterol	*A. acetabulum*	3.2 Å	0.290	Proton-pumping rhodopsin II
2ZXE	Cholesterol	*S. acanthias*	2.4 Å	0.248	Sodium-potassium pump
3KDP	Cholesterol	*S. scrofa*	3.5 Å	0.243	Sodium-potassium pump
4DKL	Cholesterol	*M. musculus*	2.8 Å	0.235	μ-Opioid receptor

Test set: soluble proteins

PDB code	Ligand	Source	Res.(Å)	R-factor	Protein Name
1LRI	Cholesterol	*P. cryptogea*	1.45 Å	0.161	Beta-elicitin cryptogein
1N83	Cholesterol	*H. sapiens*	1.63 Å	0.202	Retinoic acid-related orphan receptor alpha
1ZHY	Cholesterol	*S. cerevisiae*	1.60 Å	0.216	KES1 protein
3GKI	Cholesterol	*H. sapiens*	1.80 Å	0.176	Niemann-pick c1 protein
3N9Y	Cholesterol	*H. sapiens*	2.10 Å	0.207	Cholesterol side-chain cleavage enzyme (Cyp11A1)



```
1        10        20        30        40        50
DDDDHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH 3KDP_CLR3001D
     HHHHHHH HH HHH                    HH HH H  H    HH   H 3AM6_CLR401
DD  HHHH HHHHHHHHHHHHHHHHHHHHH  H        H     HH    HH HHH 2ZXE_CLR3001
 D DH HHHH          HHH    HHHHHHHHH H HH  HHHHH          H 4DKL_CLR614A
      HH        H  H         HHH   H   H HHHHHHHHH HH  H HH 2RH1_CLR413
   D      H  HH HHH   H     HHHHHH H   HHHHH H HH    HH   H 2RH1_CLR412
```

Figure 2.2 Determining conserved site map points. Aligned site map points with matching chemical labels from the training set of cholesterol (CLR) sites are shown following SimSite3D spatial alignment. Hydrophobic (H) or hydrogen-bond donor (D) site map points are shown on lines 2-6 if they fall within 1.5 Å of a site map point of the same chemical type in the query site, 3KDP_CLR3001D, where the number and letter after the CLR residue code indicate its residue number and chain identifier in the PDB file. Hydrogen-bond acceptor (A) and donor and/or acceptor (N) points (e.g., hydroxyl interaction sites) also occur in cholesterol sites but are not found to be conserved between the sites. The 3KDP query site was chosen as the representative query site for cholesterol binding because it has the highest degree of site map point conservation with the other cholesterol sites. Highly conserved points (green backgrounds) comprising the conserved motif for cholesterol interation were identified based on occurring in at least 70% of these training cases aligned to the 3KDP query site.

2.2.3 Establishing a cholate site predictor

Creating a cholate site predictor for the CholMine software followed the same process as for cholesterol prediction. The first step was to set up the training and test databases. 20 cholate (PDB residue name CHD) binding sites in 12 non-redundant proteins were used to generate SimSite3D site maps representing points of favorable hydrophobic or hydrogen-bond interactions with cholate (Table 2.2). These 20 cholate binding sites were divided into two datasets of equal size. There were just four non-homologous membrane protein-bound cholate sites in the PDB, representing limited training power, with the 16 other cholate sites coming from soluble proteins. The training set thus included the 4 membrane protein cholate sites and 6 of the soluble cholate sites. There were no instances of cholate sites repeated (even with low homology) between the training and test sets, to guarantee that the test predictions would be unbiased. Due to the limited availability of unrelated cholate sites in the PDB, four bile acid binding proteins with moderate pairwise sequence identity (~60%) were included in the test set. Inverting the two sets in testing and training then allowed testing whether a more diverse set of cholate sites (the first set, with a mixture of unrelated membrane and soluble sites) or a set of sites with some similarity (from four diverse bile acid binding proteins and two unrelated proteins) provided greater cholate site detection power.

2.2.4 Summary of the steps for establishing a cholesterol (or cholate) site predictor

2.2.4.1 Step 1: Preparing the training and testing databases

The binding sites divided into training and test sets were processed by SimSite3D to create site maps. Sets of soluble and membrane proteins containing diverse or lipid ligands (as described in the section above, "SimSite3D and site maps for aligning and comparing protein sites" and in "Bacterial membrane proteins for evaluating false positive prediction rate", below) were also prepared as site maps for alignment and comparison as negative controls, to assess the rate of false positive predictions.

Table 2.2 Cholate binding proteins in the training and test sets.

Training set: mixture of membrane and soluble proteins

PDB ID [α]	Ligand	Source	Res.(Å)	R-factor	Protein Name
Δ 1EE2	Cholate	*E. caballus*	1.5Å	0.148	Alcohol dehydrogenase
Δ 1S9Q	Cholate	*M.musculus*	2.2Å	0.220	Estrogen-related receptor gamma
Δ 2AZY	Cholate	*S. scrofa*	1.9Å	0.167	Phospholipase A2
Δ 2DQY	Cholate	*H. sapiens*	3.0Å	0.226	Liver carboxylesterase 1
^ 2DYR	Cholate	*B. taurus*	1.8Å	0.202	Cytochrome c oxidase
Δ 2HRC	Cholate	*H. sapiens*	1.7Å	0.221	Ferrochelatase

Test set: soluble proteins

PDB ID	Ligand	Source	Res.(Å)	R-factor	Protein Name
Δ 1TW4	Cholate	*G. gallus*	2.0Å	0.216	Liver bile acid binding protein
Δ 2FT9	Cholate	*A. mexicanum*	2.5Å	0.260	Liver bile acid-binding protein
Δ 2QO4	Cholate	*D. rerio*	1.5Å	0.188	Liver bile acid-binding protein
Δ 2RLC	Cholate	*C. perfringens*	1.8Å	0.195	Choloylglycine hydrolase
Δ 3ELZ	Cholate	*D. rerio*	2.2Å	0.224	Ileal bile acid-binding protein
Δ 3QPS	Cholate	*C. jejuni*	2.4Å	0.204	CmeR

[α] Membrane proteins are indicated by ^ and soluble proteins by Δ. In PDB structures 2DYR, 2HRC, 1TW4, 2FT9, and 3ELZ, two or more independent cholate binding sites were included in training or testing.

2.2.4.2 Step 2: Choosing the most representative cholesterol (or cholate) binding site

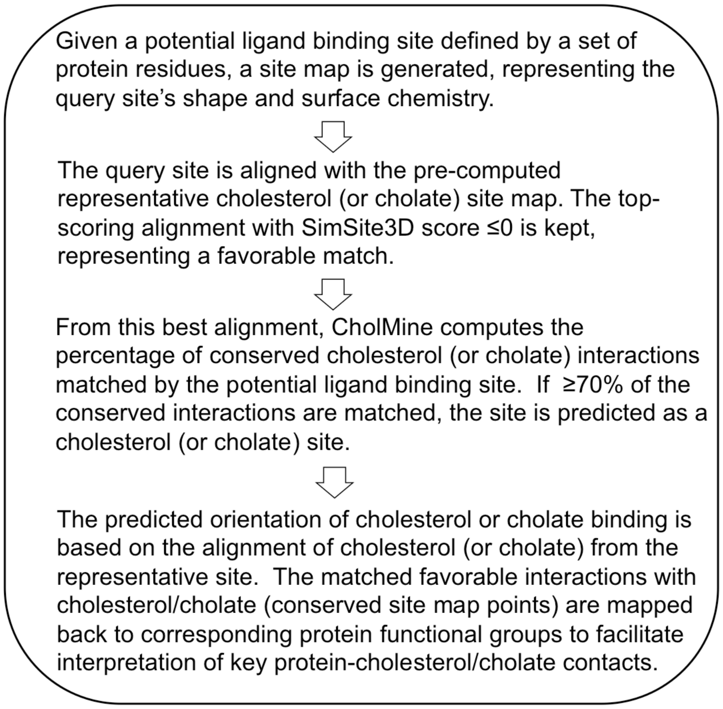
The goal of this step was to select the known site with the best SimSite3D scoring detection and quality of alignment with other cholesterol (or cholate) binding sites (as described for the site from PDB entry 3KDP in Figure 2.2). For cholesterol sites, the membrane set was initially assigned as the training set, the soluble set as a true positive test set, and the diverse ligand sites as a dataset with one true positive buried in many false positive cases. The SimSite3D normalized score threshold was set to 0.0 (keeping the best scoring orientation of any site that aligns favorably with the query site), and each of the 12 cholesterol sites was compared against all the others, and to the diverse set of 140 binding sites. The RMSD value representing the closeness of alignment (with 0 Å representing a perfect alignment) between the query site cholesterol atom positions and those in the aligned ligand sites was calculated by using the RMSD function in the OEchem toolkit v.1.6.1 (http://www.eyesopen.com; OpenEye Scientific Software, Santa Fe, NM). Assigning one query site from the training set and a separate query site from the test site allowed the two sets to be inverted for training and testing. The same procedure was followed for cholate sites.

2.2.4.3 Step 3: Extracting a fingerprint of conserved interactions from known cholesterol (or cholate) sites and applying it to predict on the test set

A high false positive rate results when SimSite3D alone is used to align hydrophobic sites with a generous scoring threshold, due to significant hydrophobic contact scores and the absence of directional hydrogen-bonding group matches (which are strong discriminants for polar sites binding the same ligand). This motivated our developing a way to pinpoint additional conserved features of cholesterol or cholate binding sites. Conserved hydrophobic interactions were identified between the cholesterol sites, based on site map points that overlaid in 3-dimensional space, as shown in Figure 2.2, for both the training and test sets. These points represent hydrophobic positions in the cholesterol sites that are ≥70% conserved with respect to the query site for the membrane (3KDP_CLR3001D) or soluble set (1ZHY_CLR1001A). The conserved points and their relative positions in space provide a shared recognition motif or fingerprint for cholesterol interaction that is implemented as a filter (following SimSite3D alignment) in the CholMine predictor. A test site is predicted to bind cholesterol or cholate if, upon 3-dimensional site map alignment with the query site, it matches at least 70% of the conserved points. The same procedure was followed for identifying and applying a conserved recognition motif for the cholate training and test sites.

2.2.5 Bacterial membrane proteins for evaluating false positive prediction rate

Bacteria contain no cholate or cholesterol, and are thus likely to provide a rigorous set of ligand sites to test for the rate of false positive cholesterol predictions because their membrane-exposed surfaces are hydrophobic and interact with other lipids. PDB codes of bacterial membrane proteins were extracted from the Membrane Proteins of Known 3D Structure Database (http://blanco.biomol.uci.edu/mpstruc/) and then entered in the Pisces server[43] (http://dunbrack.fccc.edu/Guoli/PISCES_InputB.php) to select a low-homology set of bacterial membrane proteins using default criteria: crystal structures with ≤ 25% pairwise sequence identity, ≤ 3.0 Å resolution, R-value ≤ 0.3, and chain length between 40 and 10,000 residues.



Given a potential ligand binding site defined by a set of protein residues, a site map is generated, representing the query site's shape and surface chemistry.

The query site is aligned with the pre-computed representative cholesterol (or cholate) site map. The top-scoring alignment with SimSite3D score ≤0 is kept, representing a favorable match.

From this best alignment, CholMine computes the percentage of conserved cholesterol (or cholate) interactions matched by the potential ligand binding site. If ≥70% of the conserved interactions are matched, the site is predicted as a cholesterol (or cholate) site.

The predicted orientation of cholesterol or cholate binding is based on the alignment of cholesterol (or cholate) from the representative site. The matched favorable interactions with cholesterol/cholate (conserved site map points) are mapped back to corresponding protein functional groups to facilitate interpretation of key protein-cholesterol/cholate contacts.

Figure 2.3 Steps in CholMine cholesterol and cholate site prediction.

2.2.6 CholMine server

The overall steps in cholesterol/cholate site prediction by CholMine are summarized in Figure 2.3. A web server implementation has been established to support automated prediction of cholesterol and cholate binding sites by users for their own protein structures (http://cholmine.bmb.msu.edu). Given a Protein Data Bank file and a ligand residue number and ligand chain ID for a placemarker ligand in the site, the server will provide the following information: a prediction of whether the site binds cholesterol or cholate; the predicted binding mode of the corresponding steroid; and the residues in the binding site forming conserved interactions with cholesterol or cholate. A prediction summary plus PDB files containing the ligand orientation and essential residues are e-mailed to the user, with an option to also provide a pre-formatted PyMOL molecular graphics file (Schrödinger, New York, NY; http://pymol.org) showing the predicted interactions. The set of key protein interactions can be used to design experiments that probe ligand binding, for instance by site-directed mutagenesis.

As well as supporting the use of a placeholder ligand (e.g., a crystallographic lipid or user-defined dummy residue) to define the binding site volume to analyze, the server also supports user uploading of a mini PDB file that contains up to 25 residues defining the protein region the user would like to assess for cholesterol or cholate binding. This set of residues is used to define the potential ligand binding site volume as a box bounded by the minimum and maximum x, y, and z coordinates of the residues provided. The volume for site map generation is then refined by placing probes on a 1.0 Å grid in the box and removing any probes within 3.5 Å (van der Waals contact distance) of protein atoms. The site map for CholMine analysis is generated within this volume for comparison to the conserved interaction points characteristic of cholesterol or cholate binding. 10,000 Å^3 was set as the maximum box volume in the server implementation.

2.3 Results

2.3.1 Cholesterol binding site training and testing

Of all the membrane cholesterol sites, 3KDP_CLR3001D gave the lowest average RMSD of alignment against the other membrane sites in the training set when used as the query (Figure 2.4(A)), so the site map and positions and chemistry of conserved interactions in this site were used as the basis to align and score the test cases. As shown in Figure 2.4(B), 1ZHY_CLR1001A gave the lowest average RMSD when used as the query for alignment of the set of soluble cholesterol sites. Thus, this site was chosen as the soluble site representative query when the training and test sets were inverted to determine which query had the greatest predictive power and lowest false positive rate.

As shown in Table 2.3, using the 3KDP_CLR3001D site as the query (where CLR is the residue name for cholesterol and 3001D is the ligand residue number), combined with requiring at least 70% of its conserved interactions to be matched for a site to be predicted as cholesterol binding resulted in prediction of 83% of the membrane protein cholesterol sites (training set) and 80% of the soluble protein cholesterol sites (true positives in the unbiased test set), with a relatively low rate (5%) of false positives in the 140-site diverse dataset. Self-prediction of a site (when used as both the query site and as a dataset entry) is not included in the calculation of the true positive rate, since self-prediction is guaranteed. In contrast, although the soluble cholesterol site 1ZHY_CLR1001A has a low false positive rate when at least 70% of its conserved interactions are matched, it fails to find any of the membrane protein cholesterol binding sites, while predicting 75% of the soluble sites. These results suggest that the membrane cholesterol sites share a conserved motif that is also part of the soluble site recognition of cholesterol. However, additional shared interactions within the soluble sites are not well-matched by the membrane sites, likely due to the fact that soluble proteins more fully surround and sequester cholesterol. Based on its superior performance on soluble as well as membrane cholesterol binding sites, the 3KDP query site and its conserved set of interactions were implemented in the CholMine server for cholesterol site detection.

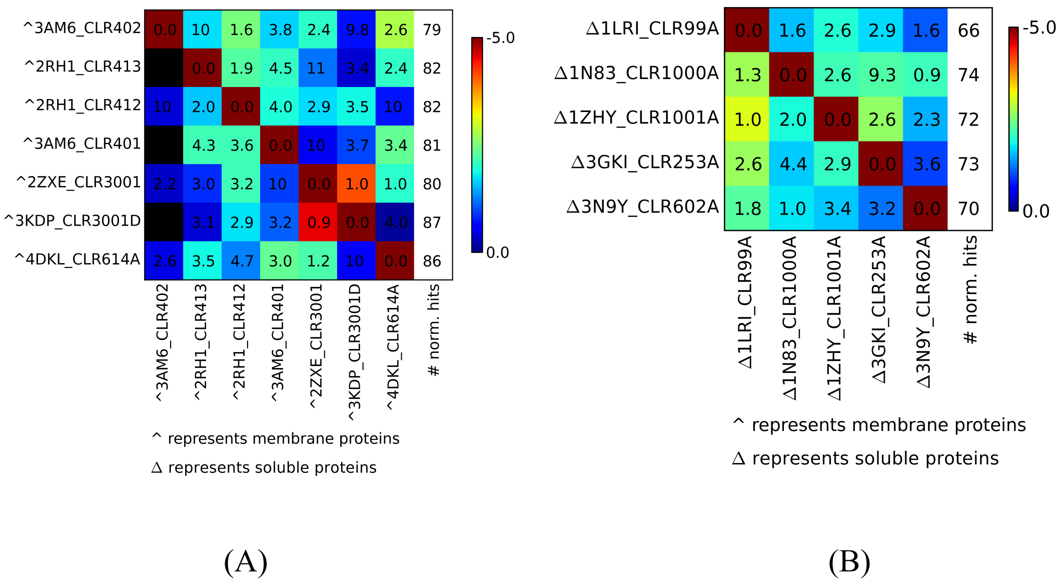


Figure 2.4 Pairwise alignment and similarity scoring. (A) All-against-all SimSite3D comparison for membrane protein cholesterol binding sites. (B) All-against-all comparison for soluble protein cholesterol binding sites. For the top-scoring alignment of each site pair, the SimSite3D similarity score values are colored from red (most similar) to dark blue (marginally similar) with corresponding score values ranging from -5 to 0 (in standard deviations above the mean score when the same query site is compared to the set of 140 diverse ligand binding sites, where more negative is more significant). Black indicates failure to meet the normalized score threshold of 0. Numbers reported in the grid are the RMSD values (Å) between cholesterol rings following SimSite3D site alignment. Lower RMSD indicates better alignment between sites. The "# norm. hits" column on the right side of each matrix reports the number of sites meeting the scoring threshold for similarity to the query site (labeled to the left in each row) when searching against the 140 sites in the diverse dataset (Table A.1.1), which includes one true positive cholesterol site. The high number of false positives is based on SimSite3D alignment score only, before the conserved interaction points for cholesterol sites have been considered.

Table 2.3 Prediction results for using cholesterol sites in 3KDP_CLR3001D (a membrane protein) and 1ZHY_CLR1001A (a soluble protein) for detecting cholesterol sites in other proteins, plus assessment of false positives in a set of 139 non-cholesterol ligand sites. When 1ZHY_CLR1001A was used as the query in the results below, the training and test sets were inverted relative to those listed in Table 2.1. Query self-matches were excluded from the statistics.

Query ID	True Positive Rate for Training Dataset	Unbiased True Positive Rate for Test Dataset	False Positive Rate for Diverse Dataset
3KDP_CLR3001D	5/6 (83%)	4/5(80%)	7/139 (5%)
1ZHY_CLR1001A	3/4 (75%)	0	2/139 (1.4%)

2.3.2 Cholate site training and testing

SimSite3D pairwise comparison of the cholate sites for the two datasets is shown in Figure 2.5, allowing the identification of the query site within each set that could best detect other cholate sites based on the lowest average RMSD of alignment over the most sites. The membrane protein site representative (2DYR_CHD525C) provided better predictive ability overall (Table 2.4). Predicting cholate sites as those matching at least 70% of the conserved interactions in this query site gave a true positive rate of 67% for cholate sites in the training set, a true positive rate of 70% for cholates in the unbiased test set, and a false positive rate of 12% on the set of 140 diverse ligand binding sites. 2QO4_CHD130A was identified as the best representative of the second, entirely soluble cholate site dataset. When this site was used as the query to find cholate sites matching its conserved interactions, a true positive rate of 67% was observed in the entirely soluble cholate site set, a true positive rate of only 10% in the mixed membrane/soluble protein set, and a false positive rate of 1.4% when applied to the set of 140 diverse cholate sites. The decreased generalization of the soluble site query and conserved points for predicting other cholate sites was expected, since a substantial number of sites in this set came from two sites in diverse members of the β-clamshell bile acid binding protein family. Similarly, by being a more family-specific motif, this query's lower false positive rate was expected on the diverse set of 140 non-cholate binding sites. The membrane cholate site query performed better as a cholate site predictor that generalizes across protein families, with almost twice the unbiased true positive rate (Table 2.4). Therefore, cholate site prediction in CholMine uses 2DYR_CHD525C as the query, combined with conserved interactions derived from the first dataset of mixed membrane and soluble protein cholate sites.

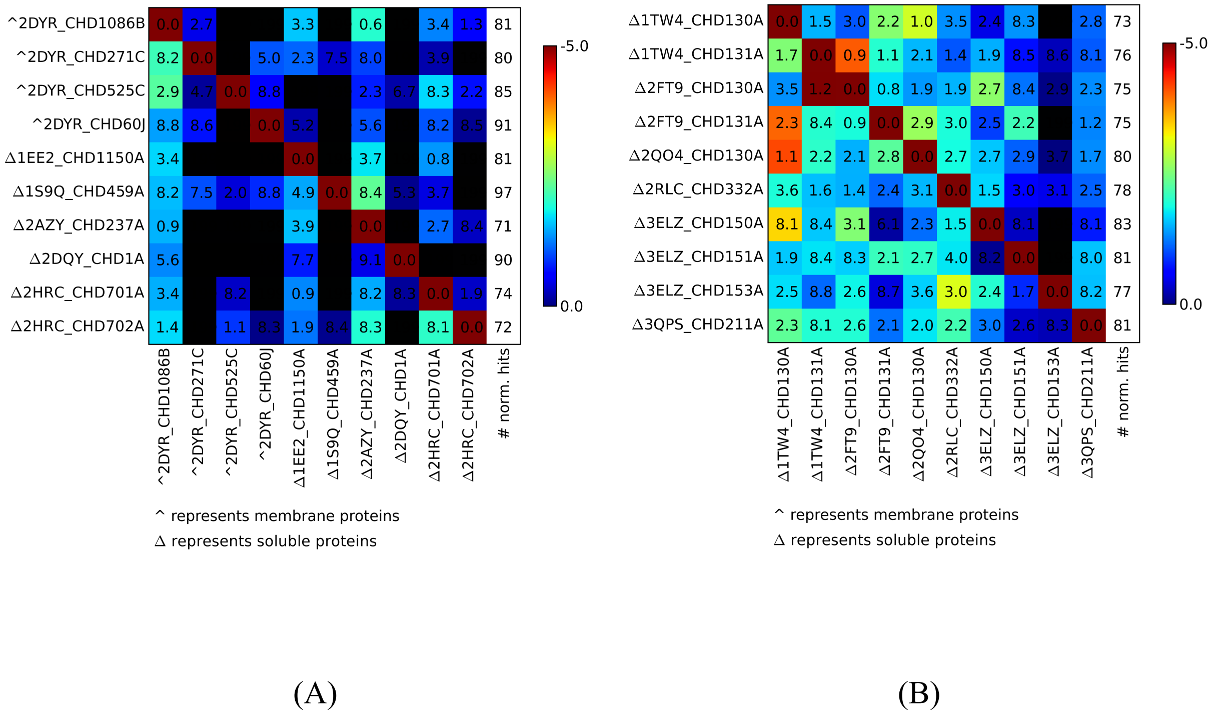


Figure 2.5 Pairwise alignment and similarity scoring. (A) All-against-all SimSite3D similarity comparison for the first dataset, which includes 4 membrane cholate binding sites and 6 soluble cholate binding sites. (B) All-against-all comparison for the second dataset, which includes another 10 soluble cholate binding sites unrelated to the first set. (See Figure 2.4 legend for additional details.)

Table 2.4 Prediction results from using cholate sites 2DYR_CHD525C (best representative from a membrane protein) and 2QO4_CHD130A (best representative from a soluble protein in the second set) for alignment and scoring to predict cholate binding sites in other proteins and assess false positive rate in a set of 140 non-cholate sites. Query self-matches were excluded from the results. The training and test sets were inverted relative to Table 2.2 when the 2QO4 query was used.

Query ID	True Positive Rate for Training Dataset	Unbiased True Positive Rate for Test Dataset	False Positive Rate for Diverse Dataset
2DYR_CHD525C	6/9 (67%)	7/10 (70%)	17/140 (12%)
2QO4_CHD130A	6/9 (67%)	1/10(10%)	2/140 (1.4%)

2.3.3 Evaluating the statistical significance of the cholesterol and cholate site predictors

The lift value is a common way to evaluate models in data mining, reflecting the enhancement in predictivity relative to random selection.[44] Suppose the predictor rule is that A implies B (e.g., a positive prediction by CholMine implies that the site binds cholesterol). The lift value for CholMine predictions can be calculated as:

$$Lift(A \Rightarrow B) = \frac{P(B \mid A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$$

$Lift(A \Rightarrow B) > 1$ means A and B have a positive relationship, and the numeric value reflects the n-fold enhancement of predictive rate (how many times higher?) relative to random prediction. $Lift(A \Rightarrow B) = 1$ indicates that A and B are independent, and $Lift(A \Rightarrow B) < 1$ means A and B have an inverse relationship. The chi-squared test can also be used to evaluate whether the correlation between A and B is statistically significant, by measuring the probability of there being a significant difference between the predicted versus actual result (e.g., the presence of a cholesterol binding site). For CholMine cholesterol site prediction, the lift value was 7.7, indicating CholMine is almost 8 times as effective as random prediction of cholate sites. The very small chi-squared P-value of 1.05e-13 indicates significant correlation between CholMine prediction and cholesterol binding. For CholMine prediction of cholate sites, the lift value is also significant (3.6), with a very small chi-squared P-value of 2.53e-08.

2.3.4 GPCR cholesterol binding site prediction

Putative cholesterol sites in class A GPCRs were analyzed as one way of testing the predictive ability of CholMine on additional cholesterol sites. The consensus motif (CCM) found in the cholesterol-binding site of human β2-adrenergic receptor (labeled as residue 412 in PDB code: 2RH1) is matched by the sequences in 44% of human class A G protein coupled receptors.[13] To assess the ability of CholMine to find sites matching the sequence-based consensus motif, prediction was performed on the structures available for 11 of these receptors (PDB codes: 3EML, 3PBL, 2KS9, 2Y00, 3RZE, 1U19, 2Z73, 3ODU, 3V2W, 3UON, and 4DJH; Table A.1.2). 82% of these proteins were predicted by CholMine to bind cholesterol in the region corresponding to cholesterol 412 in PDB entry 2RH1, in PDB entries 3EML, 2KS9, 2Y00, 3RZE, 1U19, 3ODU, 3V2W, and 3UON. In addition, for the 1.8Å resolution crystal structure of the human A2a adenosine receptor (PDB entry: 4EIY), which contains 3 cholesterol-bound sites unrelated to each other by symmetry or amino acid sequence, two of the three sites were predicted by CholMine (labeled as residues 404 and 405 in PDB entry 4EIY).

2.3.5 Comparison of CholMine structure-based predictions with sequence-based predictions using the CCM, CRAC, and GXXXG motifs

To compare the predictive ability of previously published cholesterol binding sequence motifs with that of CholMine, Sequery[45] was applied to identify sequences matching each motif in crystal structures of the same proteins used for CholMine prediction (Table 2.1 and Tables

A.2.1 and A.2.2). Matching the CCM, CRAC and GXXXG sequence motifs predicted the membrane protein cholesterol binding sites well (80-100% of these sites were predicted), predicted soluble sites less well (40-80%), and resulted in an unacceptable rate of false positives in the diverse dataset: 100 or more cholesterol sites were predicted in 139 sites known to bind a different ligand (Table 2.5).

Table 2.5 Comparison of cholesterol site prediction in true versus non-cholesterol binding sites by the CholMine conserved spatial motif versus sequence motif matching.

	Relaxed CCM[α]	CCM[α]	CCM + Surface accessibility	CRAC[α]	GXXXG[α]	CholMine predictor
Membrane set	5/5 (100%)	4/5 (80%)	2/5 (40%)	5/5 (100%)	4/5 (80%)	5/6 (83%)
Soluble Set	4/5 (80%)	2/5 (40%)	1/5 (20%)	3/5 (60%)	3/5 (60%)	4/5 (80%)
GPCRs	11/11 (100%)	10/11 (91%)	6/11 (54%)	11/11 (100%)	5/11 (45%)	9/11 (82%)
Diverse dataset (false positives)	130 /139 (94%)	105/139 (75%)	33/139 (24%)	116/139 (83%)	100/139 (72%)	7/139 (5%)

[α] Relaxed CCM: R/K- $(X)_{1\text{-}7}$-I/V/L- $(X)_{1\text{-}3}$-W/Y;[3,13] CCM: R/K -$(X)_{2\text{-}6}$-I/V/L-$(X)_3$-W/Y;[13] CRAC: L/V- $(X)_{1\text{-}5}$-Y- $(X)_{1\text{-}5}$-R/K;[15,16,17] $G(X)_3G$.[22]

One of the problems with sequence motif based prediction is that it does not assess the surface accessibility of the motif, which is required for cholesterol to access the site. To test whether including solvent accessibility as an additional criterion for sequence motif-based cholesterol site prediction can solve the overprediction problem, a solvent accessible surface threshold was set at 29 Å^2 for matching each residue in the CCM motif, corresponding to the minimum exposed surface area per residue in the cholesterol site of human β_2 adrenergic receptor (PDB entry: 2RH1). The results show that the true positive rate for membrane protein cholesterol sites decreased from 80% to 40%, for soluble protein sites from 40% to 20%, and for GPCRs from 91% to 54% (Table 2.5, CCM + Surface Accessibility column). The false positive rate decreased from 75% to 24%, while still resulting in 33 false positives in 139 proteins. Overall, even when surface accessibility is considered, sequence motif prediction has an unacceptably high false positive rate for cholesterol prediction (24%) and a moderate rate of true positive prediction (20-40%), whereas CholMine structure-based prediction results in few false positives (5%) and a high true positive rate (80-83%).

2.3.6 Deciphering the determinants of cholesterol binding

For cholesterol binding site prediction in membrane proteins, all the conserved site map points representing favorable cholesterol contacts derive from hydrophobic groups, more specifically, Ile D35, Leu D36, Tyr D39, Tyr D43, Glu C840, Ile C843, Tyr C847, and Met C852 in the representative query site, 3KDP_CLR3001D (Figures 2.2 and 2.6(A)). A smaller but similar set of interactions with cholesterol at this site is identified when the single 3KDP crystal structure is analyzed by LigPlot and LigPlot^{+}[46,47] (Figure 2.6(B,C)). Compared with the CCM (R/K-(X) $_{1\text{-}7}$-I/V/L-(X) $_{1\text{-}3}$-W/Y) and CRAC (L/V-(X)$_{1\text{-}5}$-Y-(X)$_{1\text{-}5}$-R/K) motifs, the CholMine spatially conserved binding motif exemplified by this site contains an I-L-(X) $_{2}$-Y motif, which matches the residues at the end of the CCM and the beginning of the CRAC motif. CholMine's conserved interaction points surround atoms on the steroid ring observed to have the highest frequency of protein interaction (Figure 2.6(A)). There may be several reasons for the observed lack of conserved polar interactions with cholesterol. First, there is only a single polar group, the A-ring hydroxyl substituent, in cholesterol. In seven cholesterol sites evaluated (two sites in 2RH1 and 3AM6, and one each in 2ZXE, 3KDP, and 4DKL), there was only a single direct protein hydrogen bond to the cholesterol hydroxyl group, with water-mediated interactions to cholesterol in another structure, and no protein hydrogen bonds to the cholesterol hydroxyl group observed in any of the other cases. This suggests that the hydroxyl group may help position cholesterol correctly at the interface between the lipid bilayer and bulk solvent, rather than being a recognition determinant for binding to proteins. Also supportive of a lesser role for polar group recognition is the observation that the arginine or lysine residue in the CCM is only 22% conserved in class A GPCRs; thus interactions of this residue with cholesterol are only mildly conserved.[13]

In soluble protein cholesterol binding sites, both faces of cholesterol are surrounded in the pocket, forming additional interactions with the protein. However, the conserved interaction points from soluble protein cholesterol binding sites perform less well than those from membrane proteins in predicting cholesterol sites in general (Table 2.3). The conserved membrane protein cholesterol interactions (Figure 2.6A) can predict and are characteristic of both membrane and soluble sites in unrelated proteins and are the basis for CholMine cholesterol site prediction.

2.3.7 CholMine distinguishes cholesterol sites from sites occupied by acyl chain lipids

CholMine was also applied to diverse lipid binding sites: the 22 independent acyl lipid sites in the adenosine receptor (PDB code: 4EIY) and five phosphatidylethanolamine and analog sites in PDB entries 3DDL, 2Z73, 3UTW, 3UTV (Table A.1.3). CholMine correctly predicted that 21 out of 22 sites in the adenosine receptor do not bind cholesterol, and the same for all five of the phosphatidylethanolamine sites.

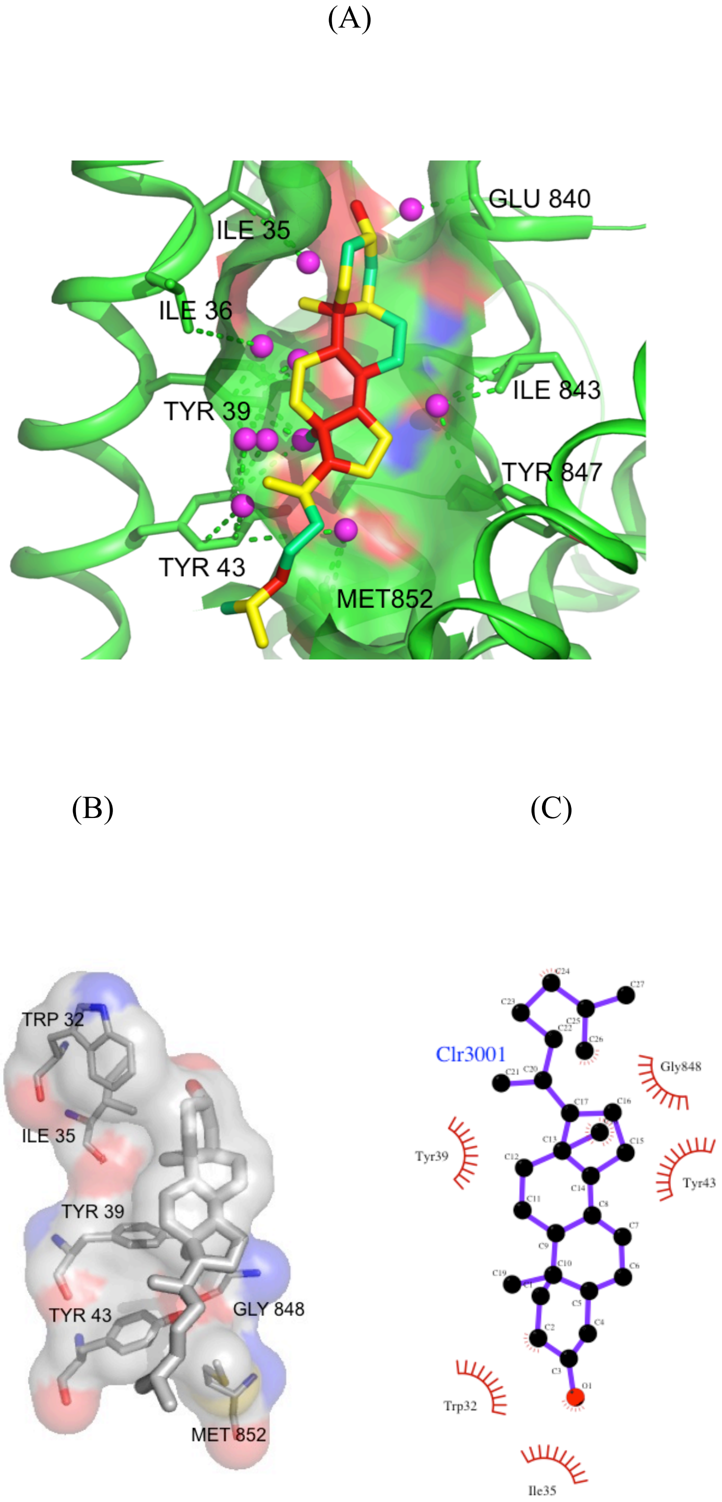


Figure 2.6 (A) Sodium/potassium-transporting ATPase cholesterol site (PDB entry 3KDP, residue D3001) used as the representative query for CholMine predictions. Purple spheres

Figure 2.6 (cont'd)
represent conserved interaction points in the membrane proteins binding cholesterol (from Figure 2.2), displayed in the context of the representative site from 3KDP. The green dashed lines connect the conserved interaction points to corresponding protein atoms. Cholesterol atoms colored in green contact a protein atom in 60% of the training set sites, atoms colored yellow have a 30-60% frequency of contact, and atoms colored in red contact the protein in <30% of the sites. (B) For comparison, LigPlot$^+$ 3-dimensional view (shown with PyMOL; Schrödinger, New York, NY; http://pymol.org) of key sodium/potassium-transporting ATPase cholesterol interactions identified in just the single structure of 3KDP. (C) Alternative LigPlot 2-dimensional view of these interactions.

2.3.8 Discriminating cholesterol and cholate sites from other steroid sites

To test whether CholMine can distinguish cholesterol sites from steroid binding sites in general, a variety of non-homologous crystal structures were tested: the progesterone sites in PDB entries 1A28, 2AA6, 2BAB, and 2HZQ, the estradiol sites in 1AQU, 1E6W, 1JGL, 1LHU, and 3OLL, and the testosterone sites in 2AM9, 1J96, and 3KDM (Table A.1.3). 10 out of the 12 sites were predicted as non-cholesterol sites, with two false positives, in 1AQU and 1J96. The cholesterol site predictor was also applied to the cholate training and test sets (Table 2.2) and vice versa (Table 2.1). The cholesterol site predictor predicts 30% of the training and 30% of the test set of cholate sites. The cholate site predictor predicts 57% of the membrane cholesterol sites and 80% of the soluble sites. Thus, cholesterol and cholate sites are harder to discriminate than cholesterol and steroid sites in general, and again we see a higher level of discrimination of cholesterol relative to cholate sites. Reasons for this are discussed below in the section below, "Comparison of cholesterol and cholate binding site conservation".

2.3.9 Bacterial membrane proteins for evaluating false positive predictions

Bacteria contain no cholate or cholesterol. Thus, known ligand sites, mostly lipid-binding, were analyzed in 109 low-homology bacterial membrane protein structures (Table A.1.4) as an additional stringent test of the false positive rate for cholesterol and cholate site prediction. Eleven of the 109 sites, or 10%, were falsely predicted as potential cholesterol sites. When analyzed as potential cholate sites, 14 (13%) sites were predicted. Though nominally these are false positives, eubacteria are known to contain sterol-like molecules including cyclic hopanoids, tetrahymanol, and squalene.[48,49] Thus, it remains possible that some sites that were occupied by unnatural molecules in the bacterial crystal structures may natively bind sterol-like molecules.

2.3.10 Cholate binding determinants

Cholate is an important detergent for membrane proteins and also a representative of bile acids that act as hormones, pheromones, and important metabolites of cholesterol. CholMine was trained for cholate site prediction similarly to the protocol for cholesterol, and the determinants for cholate binding in membrane proteins were found to differ somewhat from those in soluble proteins. For membrane protein cholate binding sites, the conserved interaction points were all hydrophobic. In the representative 2DYR_CHD525C (cytochrome c oxidase) site used for CholMine prediction, these interactions arise from TrpC99, HisA233, TrpA288, TyrA304A, and PheA305 (Figure 2.7). The latter trio of residues serve to anchor cholate in the binding pocket. Out of the 10 training set cholate molecules, half of the O3 hydroxyl groups (on the A ring of cholate) formed water-mediated and two formed direct hydrogen bonds to the protein. The O7 and O12 hydroxyls (on the B and C rings) formed fewer hydrogen bonds to protein: two O7 and four O12 water-mediated hydrogen bonds were observed, and 1 direct hydrogen bond was found in the 10 sites, with a low degree of conservation. The tail carboxylate oxygens formed 7 direct H-bonds overall, which were spatially varied in position.

2.3.11 Comparison of cholesterol and cholate binding site conservation

To understand why the number of conserved interaction points is greater for cholate sites (Figure 2.7) compared with cholesterol (Figure 2.6), the crystallographic mobility of atoms in these ligands was compared. In the training set of 10 cholate sites, the crystallographic B-factor average for cholate atoms was 48 Å^2, whereas in the training set of 7 cholesterol sites, the B-factor average for cholesterol atoms was 1.5 times as high (74 Å^2), reflecting significant mobility. Higher atomic mobility is thus likely the reason for fewer spatially conserved interactions in cholesterol sites.

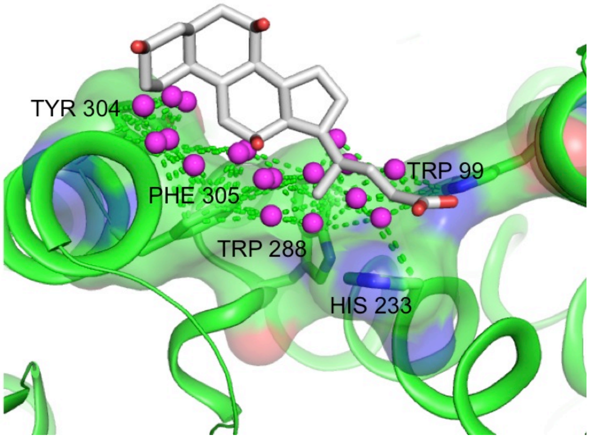


Figure 2.7 Conserved interaction points for CholMine cholate site prediction (purple spheres) are shown in the context of the interactions between the representative membrane protein query site 2DYR_CHD525C from cytochrome c oxidase, and its bound cholate molecule (white tubes with oxygen atoms in red). Essential residues contributing to the conserved interaction are labeled.

A generally similar pattern is seen in the edges and faces of cholate and cholesterol that predominate in forming conserved interactions with protein sites (Figure 2.8). Discrimination between cholesterol and cholate binding is not via polar interactions (which are not conserved across cholate or cholesterol sites), but by conserved interactions at the bend between the steroid A and B rings and near the center of the tail in cholate, versus a paucity of conserved interactions at the A-B ring junction or hydrophobic tail region in cholesterol. The conformational diversity of the tails when cholate and cholesterol bind to different sites results in their termini not being well conserved spatially whereas they still experience different chemical environments. Detecting differences in the general protein environments of the alpha face of the steroid ring (upper face in Figure 2.8) and the tail termini in cholate (polar) versus cholesterol (hydrophobic) sites will be a focus for enhancements in CholMine, as well as expanding the training data sets.

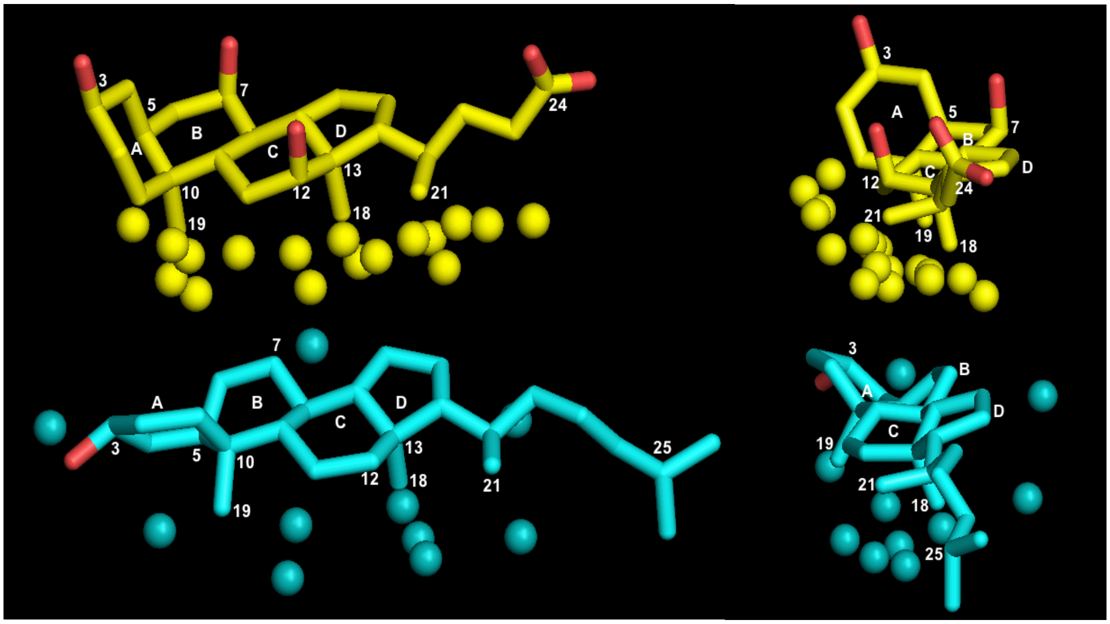


Figure 2.8 SimSite3D-identified conserved interactions for cholate (yellow) and cholesterol (blue) recognition abound along the groove formed between the row of C18, C19, and C21 methyl groups on the beta (lower) face of the steroid and the edge of the steroid ring system. The view on the right is rotated roughly 90 degrees about a vertical axis through the center of each molecule. Cholate sites are distinguished from cholesterol primarily based on interactions with the relatively conserved C22-C23 tail orientation in cholate, and numerous conserved interactions associated with the strongly bent (5-beta configuration) joint between the A and B rings of the cholate steroid ring system. Because the tail configurations are conformationally diverse in different binding sites, conserved interactions are absent in the C24-C25 region.

2.3.12 Computational efficiency of the CholMine server

For the 261 cholesterol, cholate, and other ligand sites analyzed here, the maximum protein volume for site map generation was <10,000 Å^3 (a box with edges of ~21 Å), and each prediction completed in less than 5 minutes (the time to exhaustively check and score all orientations of the user-defined cleft versus the representative site, then filter for conserved interaction matches). For the majority of cases, the server elapsed time was < 3 minutes per site.

2.4 Concluding discussion

CholMine, a predictor for cholesterol and cholate binding in protein 3-dimensional structures, has been established as a free web server at http://cholmine.bmb.msu.edu. This approach is based on the determination of conserved interactions for cholesterol and cholate binding to non-homologous membrane and soluble protein sites in PDB structures. SimSite3D alignment and scoring of site similarity serves as the first layer of prediction, considering the chemical interactions that can be made with the protein and their degree of surface match, independent of ligand information or protein structural conservation. This approach allows CholMine to focus on spatial conservation of chemical interactions rather than residue conservation. Requiring 70% match of the conserved spatial interactions of known cholesterol or cholate sites serves as the second layer of prediction, ruling out the vast majority of false positives in a dataset of diverse soluble ligand sites (resulting in a 5% false positive rate for cholesterol and 12% for cholate sites) and a slightly higher rate when applied to a dataset of diverse membrane proteins (10% for cholesterol and 13% for cholate sites). CholMine can predict 80% of known cholesterol and 70% of known cholate binding sites in diverse protein families including soluble and membrane proteins from different species, when applied to sites unrelated to those used in training. CholMine can discriminate ~75% of sites containing other steroids from cholesterol binding sites. Cholate site prediction is less steroid-selective; it also predicts two-thirds of the known cholesterol sites, likely due to the limited availability of non-homologous cholate sites for training the predictor. This problem can be addressed by periodic updating of the training set. However, the false positive rate of cholate site prediction on non-steroid sites is 5-fold lower, even for diverse lipid sites in membrane proteins.

Hydrophobic interactions focused along the groove between the steroid methyl group substituents and the ring system itself are found to be the major conserved determinants for the recognition of both cholesterol and cholate, with their polar groups not contributing to conserved interactions. Classical motifs for cholesterol site prediction have focused on amino acid residue conservation, and tend not to generalize well to other protein families, with particularly limited performance for predicting known binding sites in soluble proteins. Sequence motif-based prediction also results in many false positives (with 70% or more of 139 diverse non-cholesterol, non-cholate binding sites falsely predicted), which overwhelms the number of true positive predictions. The enhanced predictive specificity and selectivity of CholMine is based on inferring shared 3-dimensional shape and chemical information from non-homologous sites. This approach is now being generalized to create a LigPattern server that discovers the shared interaction determinants of other important regulatory ligands and substrates, including polar molecules such as adenosine.

APPENDIX

Table A.1.1 140 non-homologous protein sites binding diverse ligands, containing one cholesterol binding site (in PDB entry 1LRI) and no cholate sites.

PDB code	Ligand	Source	Res. (Å)	R-factor	Protein name
1R8S	GDP	*B. taurus*	1.46	0.159	ADP-ribosylation factor 1
1QXY	M2C	*S. aureus*	1.04	0.144	Methionyl aminopeptidase
1ECM	TSA	*E. coli*	2.2	0.192	Endo-oxabicyclic transition state analogue
1KYF	AAchain	*M. musculus*	1.22	0.154	Alpha-adaptin c
1I24	UPG	*A. thaliana*	1.2	0.192	Sulfolipid biosynthesis protein sqd1
1AWQ	His-Ala-Gly-Pro-Ile-Ala	*H. sapiens*	1.58	0.343	Cyclophilin A
1PUJ	GNP	*B. subtilis*	2.0	0.216	Conserved hypothetical protein ylqf
4UBP	HAE	*S. pasteurii*	1.55	0.151	Urease, chain A
1CHM	CMS	*P. putida*	1.9	0.177	Creatine amidinohydrolase
1KEK	HTL	*D. africanus*	1.9	0.178	Pyruvate-ferredoxin oxidoreductase
1EFY	BZC	*G. gallus*	2.2	0.194	Poly (ADP-ribose) polymerase
1MSK	SAM	*E. coli k12*	1.8	0.198	Cobalamin-dependent methionine synthase
1EVL	TSB	*E. coli*	1.55	0.215	Threonyl-trna synthetase
1JC9	NAG	*T. tridentatus*	2.01	0.183	Techylectin-5a
1DL5	SAH	*T. maritima*	1.8	0.182	Protein-l-isoaspartate o-methyltransferase
1GX5	GTP	*H. c virus (isolate bk)*	1.7	0.193	RNA-directed RNA polymerase
1GK8	CAP	*C. reinhardtii*	1.4	0.149	Ribulose-1,5 bisphosphate carboxylase larg
1FK5	OLA	*Z. mays*	1.3	0.135	Nonspecific lipid-transfer protein
1O7N	IND	*P. putida*	1.4	0.19	Naphthalene 1,2-dioxygenase alpha subunit
1M15	ARG	*L. polyphemus*	1.2	0.125	Arginine kinase
1KMV	LII	*H. sapiens*	1.05	0.13	Dihydrofolate reductase
1F20	NAP	*R. norvegicus*	1.9	0.186	Nitric-oxide synthase
1MXT	FAE	*S. sp.*	0.95	0.11	Cholesterol oxidase
1GS5	NLG	*E. coli*	1.5	0.2088	Acetylglutamate kinase

Table A.1.1 (cont'd)

1QD1	FON	*S. scrofa*	1.7	0.191	Formiminotransferase-cyclodeaminase
1C96	FLC	*B. taurus*	1.81	0.225	Mitochondrial aconitase
1K3Y	GTX	*H. sapiens*	1.3	0.148	Glutathione s-transferase a1
1T2D	NAD	*P. falciparum*	1.1	0.143	L-lactate dehydrogenase
1JET	Lys-Ala-Lys	*S. typhimurium*	1.2	0.229	Oligopeptide binding protein
1P7T	ACO	*E. coli str. k12 substr.*	1.95	0.197	Malate synthase G
1KGQ	NPI	*M. bovis*	2.0	0.179	Tetrahydrodipicolinate N-Succinyltransferase
1DMH	LIO	*A. sp.*	1.7	0.185	Catechol 1,2-dioxygenase
1XVA	SAM	*E. coli*	2.2	0.196	Glycine N-methyltransferase
1B37	FAD	*Z. mays*	1.9	0.199	Polyamine oxidase
1B5E	DCM	*E. phage t4*	1.6	0.189	Deoxycytidylate hydroxymethylase
1LTZ	HBL	*C. violaceum*	1.4	0.159	Phenylalanine-4-hydroxylase
1K5N	AAchain	*H. sapiens*	1.09	0.123	Major histocompatibility complex HLA-B*2709
1H16	DTL	*E. coli*	1.53	0.145	Formate acetyltransferase 1
1NKI	PPF	*P. aeruginosa*	0.95	0.148	Probable fosfomycin resistance protein
1G6S	S3P	*E. coli*	1.5	0.149	EPSP synthase
1LRI	CLR	*P. cryptogea*	1.45	0.161	Beta-elicitin cryptogein
1R1H	BIR	*H. sapiens*	1.95	0.211	Neprilysin
1AMU	PHE	*B. brevis*	1.9	0.213	Gramicidin synthetase 1
1L8B	MGP	*M. musculus*	1.8	0.224	Eukaryotic translation initiation factor 4E
1PFV	2FM	*E. coli*	1.7	0.186	Methionyl-tRNA synthetase
1M0K	RET	*H. salinarum*	1.43	0.134	Bacteriorhodopsin
1UZE	EAL	*H. sapiens*	1.82	0.188	Angiotensin converting enzyme
1AF7	SAH	*S. typhimurium*	2.0	0.2	Chemotaxis receptor methyltransferase CheR
1G72	PQQ	*M. methylotrophus*	1.9	0.161	Methanol dehydrogenase heavy subunit
1QZ5	KAB	*O. cuniculus*	1.45	0.17	Actin, alpha skeletal muscle
1DTD	Glu	*H. sapiens*	1.65	0.187	Carboxypeptidase A2
1JHG	TRP	*E. coli*	1.3	0.127	Trp operon repressor
1CCW	TAR	*C. cochlearium*	1.6	0.137	Glutamate mutase
1MQO	CIT	*B. cereus*	1.35	0.222	Beta-lactamase II

Table A.1.1 (cont'd)

1QMG	DMV	*S. oleracea*	1.6	0.196	Acetohydroxy-acid isomeroreductase
1UFY	MLI	*T. thermophilus*	0.96	0.11	Chorismate mutase
1KJQ	ADP	*E. coli*	1.05	0.19	Phosphoribosylglycinamide formyltransferase 2
1CIP	GNP	*R. norvegicus*	1.5	0.213	GI-alpha-1 subunit
1AYL	OXL	*E. coli*	1.8	0.195	Phosphoenolpyruvate carboxykinase
1GTE	IUR	*S. scrofa*	1.65	0.181	Dihydropyrimidine dehydrogenase
1MRJ	ADN	*T. kirilowii*	1.6	0.173	Alpha-trichosanthin
1PZ4	PLM	*A. aegypti*	1.35	0.187	Sterol carrier protein 2
1R4U	OXC	*A. flavus*	1.65	0.157	Uricase
1RQW	TAR	*T. daniellii*	1.05	0.127	Thaumatin I
2TCT	CTC	*E. coli*	2.1	0.18	Tetracycline repressor
1VJJ	GDP	*H. sapiens*	1.9	0.205	Glutamine glutamyltransferase
1PQ7	ARG	*F. oxysporum*	0.8	0.109	Trypsin
1CZA	G6P	*H. sapiens*	1.9	0.213	Hexokinase type I
1O2D	NAP	*T. maritima*	1.3	0.139	Alcohol dehydrogenase, iron-containing
1F0L	APU	*C. diphtheriae*	1.55	0.188	Diphtheria toxin
1TW6	AAchain	*H. sapiens*	1.71	0.156	Baculoviral IAP repeat-containing protein 7
2DPM	SAM	*S. pneumoniae*	1.8	0.238	Adenine-specific methyltransferase
1KA1	A3P	*S. cerevisiae*	1.3	0.134	Halotolerance protein Hal2
1F5N	GNP	*H. sapiens*	1.7	0.226	Interferon-induced guanylate-binding protein 1
1HQS	CIT	*B. subtilis*	1.55	0.202	Isocitrate dehydrogenase
1NVV	GNP	*H. sapiens*	2.18	0.208	Transforming protein p21/h-ras-1
1UNQ	ITS	*H. sapiens*	0.98	0.154	Rac-alpha serine/threonine kinase
1KRH	FAD	*A. sp.*	1.5	0.242	Benzoate 1,2-dioxygenase reductase
1M0W	3GC	*S. cerevisiae*	1.8	0.172	Glutathione synthetase
1UCD	URA	*M. charantia*	1.3	0.2	Ribonuclease MC
1HYO	HBU	*M. musculus*	1.3	0.181	Fumarylacetoacetate hydrolase
1DKX	AAchain	*E. coli*	2.0	0.206	Substrate binding domain of DNAK
1SOX	MTE	*G. gallus*	1.9	0.175	Sulfite oxidase
1LB6	AAchain	*H. sapiens*	1.8	0.203	TNF receptor-associated factor
1I1Q	TRP	*S. typhimurium*	1.9	0.219	Anthranilate synthase comp. I

Table A.1.1 (cont'd)

1ND4	KAN	*K. pneumoniae*	2.1	0.206	Aminoglycoside 3'-phosphotransferase
1EU1	MGD	*R. sphaeroides*	1.3	0.121	Dimethyl sulfoxide reductase
1BX4	ADN	*H. sapiens*	1.5	0.192	Protein (adenosine kinase)
1NOX	FMN	*T. thermophilus*	1.59	0.19	NADH oxidase
1HP1	ATP	*E. coli*	1.7	0.176	5'-nucleotidase
1LKK	AAchain	*H. sapiens*	1.0	0.133	Human p56 tyrosine kinase
1B4U	DHB	*S. paucimobilis*	2.2	0.161	Protocatechuate 4,5-dioxygenase
1GZ8	MBP	*H. sapiens*	1.3	0.153	Cell division protein kinase 2
1EYQ	NAR	*M. sativa*	1.85	0.237	Chalcone-flavonone isomerase
1TX4	GDP	*H. sapiens*	1.65	0.169	P50-rhogap
1US0	LDT	*H. sapiens*	0.66	0.0938	Aldose reductase
1UXY	EPU	*E. coli*	1.8	0.202	MURB
1J09	ATP	*T. thermophilus*	1.8	0.199	Glutamyl-tRNA synthetase
1D3V	ABH	*R. norvegicus*	1.7	0.157	Arginase
1KPF	AMP	*H. sapiens*	1.5	0.209	Protein kinase C interacting protein
1UUY	PPI	*A. thaliana*	1.45	0.163	Molybdopterin biosynthesis CNX1
1OUW	MLT	*C. sepium*	1.37	0.153	Lectin
1HFE	FCY	*D. vulgaris*	1.6	0.158	Fe-only hydrogenase
1JAK	IFG	*S. plicatus*	1.75	0.176	Beta-N-acetylhexosaminidase
1UIO	HPR	*M. musculus*	2.4	0.203	Adenosine deaminase
1P6O	HPY	*S. cerevisiae*	1.14	0.112	Cytosine deaminase
1KOL	NAD	*P. putida*	1.65	0.171	Formaldehyde dehydrogenase
1OAI	AAchain	*H. sapiens*	1.0	0.149	Nuclear RNA export factor
1FCY	564	*H. sapiens*	1.3	0.134	Retinoic acid receptor
1F3L	SAH	*R. norvegicus*	2.03	0.209	Protein arginine methyltransferase PRMT3
1N62	MCN	*O. carboxidovorans*	1.09	0.144	Carbon monoxide dehydrogenase small chain
1QJA	AAchain	*H. sapiens*	2.0	0.214	14-3-3 Protein zeta
1G2L	T87	*H. sapiens*	1.9	0.237	Coagulation factor X
2SLI	SKD	*M. decora*	1.8	0.185	Intramolecular trans-sialidase
1A9X	ORN	*E. coli*	1.8	0.191	Carbamoyl phosphate synthetase (large chain)
1TBB	ROL	*H. sapiens*	1.6	0.187	CAMP-specific 3',5'-cyclic phosphodiesterase 4D
1O7Q	UDP	*B. taurus*	1.3	0.1155	N-acetyllactosaminide
1RLZ	NAD	*H. sapiens*	2.15	0.199	Deoxyhypusine synthase
1U4G	HPI	*P. aeruginosa*	1.4	0.18	Elastase

Table A.1.1 (cont'd)

1TL2	NAG	*T. tridentatus*	2.0	0.162	Tachylectin-2
1RKD	RIB	*E. coli*	1.84	0.221	Ribokinase
1Q79	3AT	*B. taurus*	2.15	0.205	Poly(a) polymerase alpha
1PP9	SMA	*B. taurus*	2.1	0.25	Ubiquinol-cytochrome c reductase complex core protein
1E8G	FCR	*P. simplicissim.*	2.1	0.218	Vanillyl-alcohol oxidase
1L5O	2MP	*S. enterica*	1.6	0.174	CobT
1OEW	Ser-Thr	*C. parasitica*	0.9	0.121	Endothiapepsin
1H8E	ALF	*B. taurus*	2.0	0.201	Bovine mitochondrial F1-ATPase
1BGV	GLU	*C. symbiosum*	1.9	0.173	Glutamate dehydrogenase
1USC	FMN	*T. thermophilus*	1.24	0.203	Putative styrene monooxygenase small comp.
1MGP	PLM	*T. maritima*	2.0	0.202	Hypothetical protein tm841
1QNF	HDF	*S. elongatus*	1.8	0.197	Photolyase
1C1D	NAD	*R. sp.*	1.25	0.195	L-phenylalanine dehydrogenase
1UW6	NCT	*L. stagnalis*	2.2	0.22386	Acetylcholine-binding protein
1G55	SAH	*H. sapiens*	1.8	0.21	DNA cytosine methyltransferase DNMT2
1LUG	SUA	*H. sapiens*	0.95	0.119	Carbonic anhydrase II
1UF5	CDT	*A. sp.*	1.6	0.178	N-carbamyl-d-amino acid amidohydrolase
1V7R	CIT	*P. horikoshii*	1.4	0.202	Hypothetical protein ph1917
1D0C	INE	*B. taurus*	1.65	0.213	Bovine endothelial nitric oxide synthase heme domain
5CSM	TRP	*S. cerevisiae*	2.0	0.186	Chorismate mutase
1P5D	G1P	*P. aeruginosa*	1.6	0.157	Phosphomannomutase

Table A.1.2 Putative cholesterol binding sites in class A GPCRs[1].

PDB code (motif matched)	Source	Res. (Å)	Protein name	Alignment RMSD[β] with respect to PDB 2RH1
2RH1 (strict-CCM [α])	*H. sapiens*	2.4 Å	Beta adrenoceptor type 2 (ADRB2)	0.00 Å
3EML (strict-CCM [α])	*H. sapiens*	2.6 Å	Adenosine type 2A receptor (ADORA2A)	0.43 Å
3PBL (strict-CCM [α])	*H. sapiens*	2.9Å	Dopamine vertebrate type 3 receptor (DRD3)	0.48 Å
2KS9 (strict-CCM [α])	*H. sapiens*	NMR	Vertebrate tachykinin receptor (TACR1)	0.32 Å
2Y00 (CCM [α])	*M. gallopavo*	2.5 Å	Beta adrenoceptor type 1 (ADRB1)	0.29 Å
3RZE (CCM [α])	*H. sapiens*	3.1 Å	Histamine type 1 receptor	0.59 Å
1U19	*B. taurus*	2.2 Å	Rhodopsin	0.63 Å
2Z73	*T.pacificus*	2.5 Å	Rhodopsin	0.71 Å
3ODU	*H. sapiens*	2.5 Å	C-X-C chemokine receptor type 4 (CXCR4)	2.48 Å
3V2W	*H. sapiens*	3.35 Å	Sphingosine-1-phosphate receptor (EDG)	0.63 Å
3UON	*H. sapiens*	3.0 Å	M2 Human muscarinic acetylcholine receptor	0.44 Å
4DJH	*H. sapiens*	2.9 Å	κ-opioid receptor	0.67 Å

[α] strict-CCM: R/K -$(X)_{2\text{-}6}$-I/V/L-$(X)_3$-W/Y on one helix and F/Y on the neighboring helix[13]; CCM: R/K -$(X)_{2\text{-}6}$-I/V/L-$(X)_3$-W/Y[13]. Entries without motif notations belong to class A GPCRs but were not included in reference 12 or Table 2 predictions in the present manuscript.

[β] The alignment RMSD is based on relative positions of backbone atoms (N, C_α, C and O) of residues within 9 Å of cholesterol.

[1] Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V. P.; Chien, E. Y.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A Specific Cholesterol Binding Site is Established by the 2.8 Å Structure of the Human B2-Adrenergic Receptor. *Structure* 2008, *16*, 897–905.

Table A.1.3 Diverse non-cholesterol, non-cholate lipid binding sites.

PDB code	Ligand	Source	Res. (Å)	R-factor	Protein Name
1A28	Progesterone	*H. sapiens*	1.8 Å	0.191	Progesterone receptor
2AA6	Progesterone	*H. sapiens*	2.0 Å	0.197	Mineralocorticoid receptor
2ABA	Progesterone	*E. cloacae*	1.0 Å	0.129	Pentaerythritol tetranitrate reductase
2HZQ	Progesterone	*H. sapiens*	1.8 Å	0.189	Apolipoprotein D
1AQU	Estradiol	*M. musculus*	1.6 Å	0.218	Estrogen sulfotransferase
1E6W	Estradiol	*R. norvegicus*	1.7 Å	0.184	Short chain 3-hydroxyacyl-CoA dehydrogenase
1JGL	Estradiol	*M. musculus*	2.2 Å	0.199	Ig kappa-chain
1LHU	Estradiol	*H. sapiens*	1.8 Å	0.204	Sex hormone-binding globulin
3OLL	Estradiol	*H. sapiens*	1.5 Å	0.177	Estrogen receptor beta
2AM9	Testosterone	*H. sapiens*	1.6 Å	0.191	Androgen receptor
1J96	Testosterone	*H. sapiens*	1.2 Å	0.181	3-Alpha-hydroxysteroid dehydrogenase type 3
3KDM	Testosterone	*H. sapiens*	1.5 Å	0.181	Immunoglobulin light chain
4EIY	Oleic acid	*H. sapiens*	1.8 Å	0.176	Adenosine receptor A2a
3DDL	PX4	*S. ruber*	1.90 Å	0.247	Xanthorhodopsin
3DDL	PCW	*S. ruber*	1.90 Å	0.247	Xanthorhodopsin
2Z73	PC1	*T. pacificus*	2.50 Å	0.188	Rhodopsin
3UTW	MC3	*H. sp.*	2.40Å	0.206	Bacteriorhodopsin
3UTV	MC3	*H. sp.*	2.06Å	0.197	Bacteriorhodopsin

Table A.1.4 Sites in 109 low-homology bacterial membrane protein sites analyzed as potential false positive cases for cholesterol (CLR) or cholate (CHD) binding. Sites predicted to match the CholMine cholesterol or cholate site conserved interactions are noted in the third column. The last column indicates whether the crystallographic ligand at the prediction site (second column) was of lipid or lipid-like (L), drug-like (D), polar (P), or intermediate character (e.g., P/L for a polar lipid group). 73% of the sites contained lipids or partly lipidic molecules.

PDB entry	Ligand site analyzed	Prediction (CLR, CHD, or neither)	Crystal structure ligand type
1LGH	LYC A97	CLR,CHD	L
1M56	PEH A2009	CLR,CHD	L
1QD5	BOG A500	CLR	L
1U7G	BOG A400	CLR	L
1YC9	BOG A1001	CLR	L
2ERV	CXE A300	CLR	L
2YEV	5PL A900	CLR	L
3GP6	SDS A163	CLR	L
3RKO	LFA L614	CLR,CHD	L
4H44	7PH C303	CLR	L
4IL6	DGD C515	CLR,CHD	L
1B12	1PN B1001	---	D
1CWV	CIT A994	---	P
1EHK	BNG A901	---	L
1J79	NCD A950	---	P
1JB0	BCR A4001	---	L
1K4C	F09 A2001	---	L
1KMO	HTO A759	---	L
1KQF	MGD A1018	---	D/P
1LDF	GOL A476	---	P
1NKZ	RG1 A404	---	L
1Q16	MD1 A1300	---	D/P
1QFG	DDQ A1100	---	L
1QJP	C8E A1172	---	L
1UJW	GP1 A801	---	P
1UYN	CXE X2085	CHD	L
1XEZ	BOG A999	---	L
1XIO	RET A301	---	L
1XKW	LDA A2001	---	L
1Y4Z	MD1 A1800	---	D/P

Table A.1.4 (cont'd)

2A65	LEU A601	---	D
2BL2	UMQ A1162	---	L
2BS2	FAD A1656	---	D/P
2GSK	LDA A800	---	L
2GSM	DMU A5001	---	L
2GUF	MPG A701	---	L
2HDI	LDA A664	---	L
2IWV	TAM B1289	---	D
2J58	OCT A600	---	L
2NS1	BOG A601	---	L
2O4V	C8E A1295	---	L
2OQO	EPE A244	---	D
2POR	C8E A545	CHD	L
2QCU	TAM A805	---	D
2QI9	1PE C800	---	L
2SQC	C8E A632	---	L
2VDF	OCT A1254	---	L
2VPZ	MGD A1765	---	D/P
2VQG	MRD B1097	---	D
2WDQ	CBE C1130	CHD	D
2WIE	CVM A102	---	L/D
2WJN	MQ7 M1328	---	L
2WJR	EPE A1217	---	D
2WSW	CM5 A1505	---	L/D
2X27	C8E X1216	---	L
2X2V	DPV A200	---	L
2X55	C8E A1293	---	L
2XCI	PG4 A1353	---	D/P
2XOV	BNG A503	---	L
2YHC	URE A1234	---	P
2YNK	OCT A1001	---	L
2ZFG	C8E A342	---	L
3B9W	BOG A408	---	L
3BS0	C8E A501	---	L
3CSL	GOL A867	---	P
3DDL	UNL A1402	---	L
3DWN	LDA A502	CHD	L
3DWO	C8E X453	CHD	L
3DZM	C8E A209	---	L

Table A.1.4 (cont'd)

3FID	CXE A304	---	L
3HB3	LMT A568	---	L
3HYW	DCQ A500	CHD	L
3JQO	MPD D1	---	D
3KDS	NHX E998	CHD	D
3KLY	BOG A281	---	L
3L1L	BNG A447	---	L
3L7I	EDO B731	---	L/P
3M71	BOG A315	---	L
3OUF	MPD A501	---	L/P
3QE7	URA A430	---	D/P
3QRA	C8E A1	---	L
3RLB	VIB A191	---	D
3RLF	UMQ E5004	---	L
3RQW	ACH A323	---	P/D
3RVY	PX4 A4001	CHD	L
3SZV	C8E A385	---	L
3TIJ	URI A419	---	D/P
3USE	GOL L605	---	G
3V8X	C8E A1001	---	L
3WO6	OLC A302	---	L
4AFK	78M A1510	---	L
4DVE	BTN A201	---	D/L
4E1S	OLB A502	CHD	L
4EHW	MPD A402	---	D/L
4GBY	BNG A505	---	L
4GEY	DMU A510	---	L
4IKV	PG4 A613	---	L/P
4JR9	GYP A501	---	P
4MT4	3PK A1008	---	L
4N7W	MPG A402	---	L
4NHR	PEG A301	---	P/L
4NM9	FAD A2001	---	P
4NV5	U10 A501	CHD	L
4P1X	MPD A401	---	P/L
4PR7	OCT A301	---	L
4Q35	LDA A2004	---	L
4QNC	MYS A104	---	L
2J7A	LMT C1005	---	L

Table A.1.4 (cont’d)

3WU2	SQD A412	---	L

REFERENCES

REFERENCES

(1) Lund, S.; Orlowski, S.; Foresta, B. de; Champeil, P.; Maire M. Le; Møbller, J.V. Detergent Structure and Associated Lipid as Determinants in the Stabilization of Solubilized Ca^{2+}-Atpase from Sarcoplasmic Reticulum. *J. Biol. Chem.* 1989, *264*, 4907-4915.

(2) Seddon, A. M.; P. Curnow; Booth, P. J. Membrane Proteins, Lipids and Detergents: Not Just a Soap Opera. *Biochim. Biophys. Acta* 2004, *1666*, 105–117.

(3) Contreras, F.-X.; Ernst, A. M.; Wieland, F.; Brügger, B. Specificity of Intramembrane Protein-Lipid Interaction. *Cold Spring Harb. Perspect. Biol.* 2011, *3*, 1-18.

(4) Ernst. A. M.; Contreras. F.-X.; Brügger, B.; Wieland, F. Determinants of Specificity at the Protein–Lipid Interface in Membranes. *FEBS Lett.* 2010, *584*, 1713–1720.

(5) Hite, R. K.; Li, Z.; Walz, T. Principles of Membrane Protein Interactions with Annular Lipids Deduced from Aquaporin-0 2D Crystals. *EMBO J.* 2010, *29*, 1652-1658.

(6) Shinzawa-Itoh, K.; Aoyama, H.; Muramoto, K.; Terada, H.; Kurauchi, T.; Tadehara, Y.; Yamasaki, A.; Sugimura, T.; Kurono, S.; Tsujimoto, K.; Mizushima, T.; Yamashita, E.; Tsukihara, T.; Yoshikawa, S. Structures and Physiological Roles of 13 Integral Lipids of Bovine Heart Cytochrome C Oxidase. *EMBO J.* 2007, *26*, 1713–1725.

(7) Qin, L.; Hiser, C.; Mulichak, A.; Garavito, R. M.; Ferguson-Miller, S. Identification of Conserved Lipid Detergent-Binding Sites in a High-Resolution Structure of the Membrane Protein Cytochrome C Oxidase. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 16117–16122.

(8) Munro, S. Lipid Rafts: Elusive or Illusive? *Cell* 2003, *115*, 377–388.

(9) Burger, K.; Gimpl, G; Fahrenholz, F. Regulation of Receptor Function by Cholesterol. *Cell. Mol. Life Sci.* 2000, *57*, 1577-1592.

(10) Schroeder, C. Cholesterol-Binding Viral Proteins in Virus Entry and Morphogenesis. In *Cholesterol Binding and Cholesterol Transport Proteins: Structure and Function in Health and Disease*; Harris, J. R., Ed.; Springer: Dordrecht, 2010; Vol. 51, pp 77-108.

(11) Huber, T. B.; Schermer, B.; Müeller, R. U.; Höhne, M.; Bartram, M.; Calixto, A.; Hagmann, H.; Reinhardt, C.; Koos, F.; Kunzelmann, K.; Shirokova, E.; Krautwurst, D.;

Harteneck, C.; Simons, M.; Pavenstädt, H.; Kerjaschki, D.; Thiele, C.; Walz, G.; Chalfie, M.; Benzing, T. Podocin and MEC-2 Bind Cholesterol to Regulate the Activity of Associated Ion Channels. *Proc. Natl. Acad. Sci. USA* 2006, *103*, 17079–17086.

(12) Hulce, J. J; Cognetta, A. B.; Niphakis, M. J.; Tully, S. E.; Cravatt, B. F. Proteome-Wide Mapping of Cholesterol-Interacting Proteins in Mammalian Cells. *Nat. Methods* 2013, *10*, 259-64.

(13) Hanson, M. A.; Cherezov, V.; Griffith, M. T.; Roth, C. B.; Jaakola, V. P.; Chien, E. Y.; Velasquez, J.; Kuhn, P.; Stevens, R. C. A Specific Cholesterol Binding Site is Established by the 2.8 Å Structure of the Human B2-Adrenergic Receptor. *Structure* 2008, *16*, 897–905.

(14) Adamian, L.; Naveed, H.; Liang, J. Lipid-Binding Surface of Membrane Proteins: Evidence from Evolutionary and Structure Analysis. *Biochim. Biophys. Acta* 2011, *1808*, 1092–1102.

(15) Li, H; Papadopoulos, V. Peripheral-Type Benzodiazepine Receptor Function in Cholesterol Transport. Identification of a Putative Cholesterol Recognition/Interaction Amino Acid Sequence and Consensus Pattern. *Endocrinology* 1998, *139*, 4991-4997.

(16) Takeda, K.; Tonthat, N. K.; Glover, T.; Xu, W.; Koonin, E. V.; Yanagida, M.; Schumacher, M. A. Implications for Proteasome Nuclear Localization Revealed by the Structure of the Nuclear Proteasome Tether Protein Cut8. *Proc. Natl. Acad. Sci. USA* 2011, *108*, 16950–16955.

(17) Li, F.; Liu, J.; Valls, L.; Ferguson-Miller, S. Identification of a Key Cholesterol Binding Enhancement Motif in Translocator Protein 18 Kda (TSPO). *Biochemistry* 2015 *54*, 1441-1443.

(18) Baier, C. J.; Fantini, J.; Barrantes, F. J. Disclosure of Cholesterol Recognition Motifs in Transmembrane Domains of the Human Nicoticin Acetylcholine Receptor. *Sci. Rep.* 2011, *1*, 1-7.

(19) Fantini, J.; Yahi, N. Molecular Basis for the Glycosphingolipid-Binding Specificity of α-Synuclein: Key Role of Tyrosine 39 in Membrane Insertion. *J. Mol.Biol.* 2011, *408*, 654–669.

(20) Fantini, J; Barrantes, F. J.; How Cholesterol Interacts With Membrane Proteins: an Exploration of Cholesterol-Binding Sites Including CRAC, CARC, and Tilted Domains. *Front Physiol.* 2013, *4*, 1-9.

(21) Palmer, M. Cholesterol and the Activity of Bacterial Toxins. *FEMS Microbiol. Lett.* 2004, *238*, 281–289.

(22) Barrett, P. J.; Song, Y.; Van Horn, W. D.; Hustedt, E. J.; Schafer, J. M.; Hadziselimovic, A.; Beel, A. J.; Sanders, C. R. The Amyloid Precursor Protein Has a Flexible Transmembrane Domain and Binds Cholesterol. *Science* 2012, *336*, 1168-1171.

(23) Song, Y; Kenworthy, A. K.; Sanders, C. R. Cholesterol as a Co-Solvent and a Ligand for Membrane Proteins. *Protein Sci.* 2014, *23*, 1-22.

(24) Farrand, A. J.; LaChapelle, S.; Hotze, E. M.; Johnson, A. E.; Tweten, R. K. Only Two Amino Acids are Essential for Cytolytic Toxin Recognition of Cholesterol at the Membrane Surface. *Proc. Natl. Acad. Sci. USA* 2010, *107*, 4341-4346.

(25) Chiang, J. Y. Bile Acid Regulation of Gene Expression: Roles of Nuclear Hormone Receptors. *Endocr. Rev.* 2001, *23*, 443-463.

(26) Russell, D.W.; Setchell, K. D. R. Bile Acid Biosynthesis. *Biochemistry* 1992, *31*, 4737–4749.

(27) Hofmann, A.F. The Enterohepatic Circulation of Bile Acids in Man. *Clin. Gastroenterol.* 1977, *6*, 3–24.

(28) Maruyama, T.; Miyamoto, Y.; Nakamura, T.; Tamai, Y.; Okada, H.; Sugiyama, E.; Nakamura, T.; Itadani, H.; Tanaka, K. Identification of Membrane-Type Receptor for Bile Acids (M-BAR). *Biochem. Biophys. Res. Commun*. 2002, *298*, 714–719.

(29) Kawamata, Y.; Fujii, R.; Hosoya, M.; Harada, M.; Yoshida, H.; Miwa, M.; Fukusumi, S.; Habata, Y.; Itoh, T.; Shintani, Y.; Hinuma, S.; Fujisawa, Y.; Fujino, M. A G Protein-Coupled Receptor Responsive to Bile Acids. *J. Biol. Chem.* 2003, *278*, 9435-9440.

(30) Lischka, F.; Kuhn, L. A.; Libants, S.; Wu, H.; Yuan, Q.; Teeter, J.; Li, W. De-Orphanization of Two Vertebrate Pheromone Receptors. In preparation, 2015.

(31) Yau, W.-M.; Wimley, W. C.; Gawrisch, K.; White, S. H. The Preference of Tryptophan for Membrane Interfaces. *Biochemistry* 1998, *37*, 14713-14718.

(32) Ballesteros, J. A.; Weinstein, H. Integrated Methods for the Construction of Three-Dimensional Models and Computational Probing of Structure-Function Relations in G Protein-Coupled Receptors. *Methods Neurosci.* 1995, *25*, 366-428.

(33) Hunte, C. Specific Protein-Lipid Interactions in Membrane Proteins. *Biochem. Soc. Trans.* 2005, *33*, 938-942.

(34) Liu, W.; Chun, E.; Thompson, A. A.; Chubukov, P.; Xu, F.; Katritch, V.; Han, G. W.; Roth, C. B.; Heitman, L. H.; IJzerman, A. P.; Cherezov, V.; Stevens, R. C. Structural Basis for Allosteric Regulation of Gpcrs by Sodium Ions. *Science* 2012, *337*, 232-236.

(35) Lin, H. H.; Han, L. Y.; Zhang, H. L.; Zheng, C. J.; Xie, B.; Chen, Y. Z. Prediction of the Functional Class of Lipid Binding Proteins from Sequence-Derived Properties Irrespective of Sequence Similarity. *J. Lipid Res.* 2006, *47*, 824-831.

(36) Xiong, W.; Guo, Y.; Li, M. Prediction of Lipid-Binding Sites Based on Support Vector Machine and Position Specific Scoring Matrix. *Protein J.* 2010, *29*, 427-431.

(37) Scott, D. L.; Diez, G.; Goldmann, W. H. Prediction-Lipid Interactions: Correlation of a Predictive Algorithm for Lipid-Binding Sites with Three-Dimensional Structural Data. *Theor. Biol. Med. Model.* 2006, *3*, 1-14.

(38) Van Voorst, J. R.; Finzel, B. C.; Tonero, M. E.; Rai, B.; Narasimhan, L.; Howe, W. J.; Kuhn. L. A. Screening to Identify Similar Ligand-Binding Pockets in Diverse Proteins. In preparation, 2015.

(39) Weill, N.; Rognan, D. Development and Validation of a Novel Protein-Ligand Fingerprint to Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands. J. *Chem. Inf. Model.* 2009, *49*, 1049-1062.

(40) Madala, P. K.; Fairlie, D. P.; Boden, M. Matching Cavities in G Protein-Coupled Receptors to Infer Liand-Binding Sites. *J. Chem. Inf. Model.* 2012, *52*, 1401-1410.

(41) Van Voorst, J. R. Surface Matching and Chemical Scoring to Detect Unrelated Proteins Binding Similar Small Molecules. Ph.D. Thesis, Michigan State University, December 2011.

(42) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the Essential Features of a Protein Surface for Improving Protein-Ligand Docking, Scoring, and Virtual Screening. *J. Comp.-Aided Molecular Design* 2002, *16*, 883-902.

(43) Wang, G.; Dunbrack, R. L. Jr. PISCES: A Protein Sequence Culling Server. *Bioinformatics* 2003, *19*, 1589-1591.

(44) Tan, P.N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*; Addison-Wesley: Boston, 2006; pp 370-386.

(45) Craig, L.; Sanschagrin, P. C.; Rozek, A.; Lackie, S.; Kuhn, L. A.; Scott, J. K. The Role of Structure in Antibody Cross-Reactivity between Peptides and Folded Proteins. *J. Mol. Biol.* 1998, *281*, 183-201.

(46) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a Program to Generate Schematic Diagrams of Protein-Ligand Interactions. *Protein Eng.* 1996, *8*, 127-134.

(47) Laskowski, R. A.; Swindells, M. B. LigPlot+: Multiple Ligand−Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* 2011, *51*, 2778−2786.

(48) Barenholz, Y. Cholesterol and Other Membrane Active Sterols: From Membrane Evolution to Rafts. *Prog. Lipid Res.* 2002, *41*, 1-5.

(49) Majewska, M. D. Steroids and Ion Channels in Evolution: From Bacteria to Synapses and Mind. *Acta Neurobiol. Exp.* 2007, *67*, 219-233.

Chapter 3 Deciphering Substituent Effects of Ring-substituted α-Arylalanines on the Isomerization Reaction Catalyzed by an Aminomutase

Reprint (adapted) with permission from *Ring-Substituted α-Arylalanines for Probing Substituent Effects on the Isomerization Reaction Catalyzed by an Aminomutase*. Nishanka Dilini Ratnayake, Nan Liu, Leslie A. Kuhn, and Kevin D. Walker. *ACS Catal.*, 2014, 4, 3077–3090.

Here, we analyzed the substituent effects of ring-substituted α-arylalanines on the isomerization reaction catalyzed by an aminomutase. The goal was to determine how the protein-ligand interaction components of the binding energy dictate the relative biological activities of substrates.

3.1 Introduction

β-Amino acids are gaining use as building blocks for synthetic β-peptide oligomers that are used as biologically active antibiotics.[1] These β-peptides form ordered secondary structures similar to α-peptides, yet are less prone to cleavage than their α-peptide counterparts by most peptidases *in vivo*. In addition, biosynthesizing novel (*S*)-β-amino arylalanines, such as *o*-methyl-β-phenylalanine, has potential application in the synthesis of a pyrazole heterocycle compound that inhibits the function of a lysosomal serine protease cathepsin A (CatA). This inhibition of CatA was shown to prevent the development of salt-induced hypertension.[2] *m*-Fluoro-β-phenylalanine has also been used as an intermediate in the synthesis of potent chemokine receptor CCR5 antagonist.[3]

Enzymatic resolution and catalysis are described as elegant approaches to access enantiopure β-amino acids. Phenylalanine aminomutases from the bacterium *Pantoea agglomerans* (*Pa*PAM, EC 5.4.3.11) and an isozyme from *Taxus* plants (*Tc*PAM, EC 5.4.3.10) use a 4-methylidene-1*H*-imidazol-5(4*H*)-one (MIO) prosthetic group to isomerize

(2*S*)-α-phenylalanine to β-phenylalanine. *Tc*PAM makes the (3*R*)-β-amino acid, a precursor of the phenylisoserine side chain on the pathway to the antimitotic compound paclitaxel.[4] In an earlier study, *Tc*PAM was shown to convert several variously modified α-arylalanines to their cognate β-isomers.[5] In contrast, *Pa*PAM makes the (3*S*)-β-phenylalanine antipode on the biosynthetic pathway to the antibiotic andrimid (Figure 3.1).[6] Knowing the substrate scope of *Pa*PAM could increase the range of novel enantiopure β-arylalanines obtained biocatalytically.

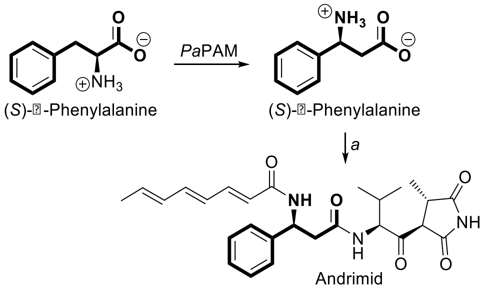


Figure 3.1 Partial andrimid biosynthetic pathway starting from (*S*)-β-phenylalanine *via* (*S*)-α-phenylalanine. *a*) Several steps.

Both PAMs belong to a class I lyase-like superfamily of catalysts,[6-9] along with other MIO-dependent aminomutases. Tyrosine aminomutases (*Cc*TAM and *Sg*TAM, respectively) are used on the biosynthetic pathways to the cytotoxic chondramides in *Chondromyces crocatus* and to the enediyne antitumor antibiotic C-1027, of the neocarzinostatin family, made by *Streptomyces globisporus*. A phenylalanine aminomutase from *Streptomyces maritimus* (*Sm*PAM) described earlier as a lysase was recently characterized.[7] A recently characterized aminomutase biosynthesizes (*R*)-2-aza-β-tyrosine from 2-aza-α-tyrosine found on the biosynthetic pathway to the enediyne kedarcidin in *Streptoalloteichus*.[10]

Recent structural characterization of *Pa*PAM supports the formation of an NH_2-MIO adduct, where the amino group of the substrate is covalently attached to the enzyme during the α/β isomerization (Figure 3.2).[11] A proton and the NH_2-MIO group are eliminated from the substrate to form a cinnamate intermediate (released occasionally as a minor by-product), followed by hydroamination of the intermediate from NH_2-MIO to form the β-amino acid.

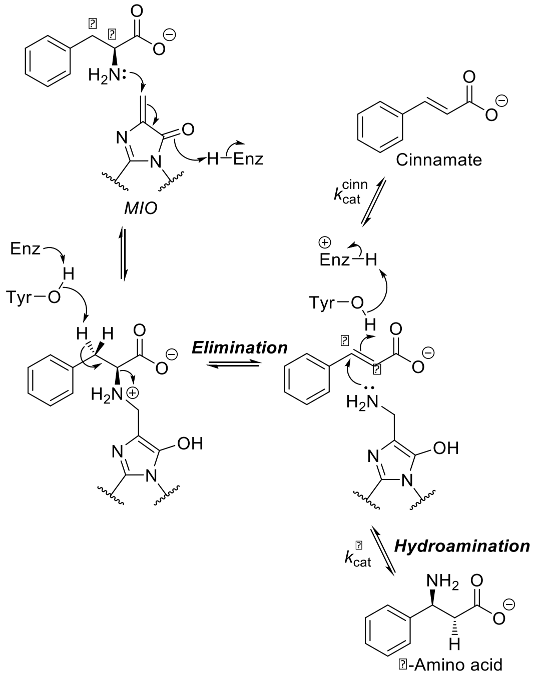


Figure 3.2 Mechanism of the MIO-dependent isomerization catalyzed by *Pa*PAM. MIO: 4-methylidene-1*H*-imidazol-5(4*H*)-one; $\boldsymbol{k_{\mathrm{cat}}^{\mathrm{cinn}}}$: the rate at which the cinnamate by-product is released; $\boldsymbol{k_{\mathrm{cat}}^{\beta}}$: the rate at which the β-amino acid product is released.

The broad substrate specificity of *Tc*PAM encouraged us to investigate, herein, the substrate specificity of the related MIO phenylalanine aminomutase. In addition, structural and mechanistic studies on MIO-based aminomutases are increasing our understanding of the reaction chemistry of the enzymes in this family.[9,13,15-19] Here, to gain further insights on these enzymes, we used computational chemistry to analyze how structural interaction energies relate to the *Pa*PAM isomerization kinetics of substrates with different aryl rings. We propose the *Pa*PAM reaction chemistry is influenced by different properties of the substrate, including sterics, and the magnitude and direction of electronic effects of the substituents on the aryl ring.

3.2 Materials and methods

3.2.1 Experiments

The experimental part of this work is done by Dr. Dilini, including expression and purification of *pa*PAM, assessment of the substrate specificity of *Pa*PAM for (2*S*)-α-phenylalanine analogs were accessed and measurement of kinetic parameters (K_{M} and $k_{\mathrm{cat}}^{\mathrm{total}}$) of *Pa*PAM for (2*S*)-α-phenylalanine analogs and inhibition assays for non-productive substrates. The experimental kinetic data is summarized in Table 3.1.

3.2.2 Modeling substrate-*Pa*PAM structural interactions to understand selectivity

To understand the differences in catalytic efficiency, which are largely dictated by differences in K_{M}, the substrates were modeled in the *Pa*PAM active site. Active configurations of the substrates were generated by overlaying their aryl rings onto the active conformation of α-phenylalanine in the crystal structure by using molecular editing in PyMOL 1.5.0.4 (Schrödinger, Inc., New York, NY) and fixed reference coordinates in OMEGA 2.4.6 (OpenEye Scientific Software).[12,13] Since the substrates form covalent bonds with binding site residues of *Pa*PAM, their orientation is highly restricted.

The position of the *ortho*- or *meta*-substituent breaks the C2 axis of symmetry in the phenyl ring of the substrates. Thus, the ring can adopt two configurations that are consistent with the orientation of α-phenylalanine in the crystal structure. In one configuration, called the "*NH_2-cis*," the substituent on the aryl ring is on the same side as the NH_2 group of the phenylalanine substrate. In the other configuration, the "*NH_2-trans*," obtained by a 180° rotation about the C_{β}-C_{ipso} bond, the substituent is oriented on the side opposite the NH_2 group. Alternative low-energy conformations of the substrates, in which the substrate orientation deviated from that of α-phenylalanine in the crystal structure, were sampled using OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com) and analyzed with respect to experimental K_M values. For energy calculations, AM1BCC charges were assigned to the substrates using molcharge 1.3.1 (Open Eye Scientific Software).[14]

3.2.3 Calculating substrate-*Pa*PAM interaction energies

The sum of protein-ligand interaction energy [$E_{(p\text{-}l)}$] and ligand internal energy [$E_{(l)}$] values for the 22 substrates was calculated using Szybki[15-17] 1.7.0 (OpenEye Scientific Software). The electrostatic Coulombic [$E_{C(p\text{-}l)}$] and steric van der Waals (vdW) interaction energy [$E_{V(p\text{-}l)}$] terms were extracted from the $E_{(p\text{-}l)}$ term for each conformer. Steric collisions between the substrates and the binding site residues were visualized pairwise by using a PyMOL script, show_bumps.py (created by Thomas Holder of Schrödinger, Inc.) showing vdW radius overlaps of 0.1 Å or more. The residues were then grouped according to which overlaps impacted the *o*-, *m*-, and *p*-positions of substrates. The component energy terms [$E_{(p\text{-}l)}$], [$E_{C(p\text{-}l)}$], [$E_{V(p\text{-}l)}$] and [$E_{(l)}$] were calculated with two protocols to evaluate which approach led to interaction energies that best correlated with the K_M values. First, a single-point energy calculation protocol employing a Poisson-Boltzmann electrostatics model was used when the substrate was placed in the *NH_2-cis* or *NH_2-trans* configuration. The *NH_2-cis* and *NH_2-trans* conformers were evaluated without energy minimization. The binding site of the protein was kept in its crystallographic conformation, to test the hypothesis that the active complex of the protein and substrate matches the crystallographic conformation observed with α-phenylalanine (PDB entry 3UNV). Second, a two-step protocol recommended by the OpenEye Scientific Software was used to explore whether energy minimization could improve the modeling of *Pa*PAM-substrate interactions by reducing any repulsive interactions. The backbone residues of *Pa*PAM were fixed, with the substrates in either the *NH_2-cis* or *NH_2-trans* configuration. Protein side chains within 4 Å of the substrates were then allowed to move towards an energy minimum, using the exact Coulomb electrostatics model. Because vdW clashes lead to large, unfavorable interaction energies, this energy minimization protocol reduces vdW overlap by small shifts in active site residues when possible. The energy estimate of each minimized configuration was then refined using the above single-point energy calculation with the Poisson-Boltzmann electrostatics model.

As an alternative approach, SLIDE (version 3.4) docking[18,19] was used to model potential conformational changes of the protein and substrate upon binding. SLIDE rotated active site residues to remove or reduce vdW overlap, while the phenylalanine ligands were fixed to maintain their initial *NH_2-cis* or *NH_2-trans* configuration.

3.2.4 Structure-activity landscape index analysis

To identify any additional steric or electrostatic factors important for the activity of *Pa*PAM substrates, structure-activity landscape index (SALI) analysis was used to identify "activity cliffs". These cliffs represent large changes in *Pa*PAM binding affinity among structurally-similar substrates.[20] For identifying activity cliffs, pairwise comparisons between substrates to measure structural similarity scores were performed using ROCS 2.4.2 software (OpenEye Scientific Software).[21] The SALI score was measured as $\mathrm{SALI}_{(i,j)} = |(K_{\mathrm{M}i} - K_{\mathrm{M}j})|/(2 - sim(i,j))$, in which the $sim(i,j)$ value (structural similarity between molecules i and j) was measured by the ROCS Tanimoto Combo score (with a maximum value of 2, reflecting equal contributions from shape and electrostatic match terms), and $K_{\mathrm{M}i}$ and $K_{\mathrm{M}j}$ were the experimental K_{M} values of molecules i and j.

3.3 Results and discussion

3.3.1 Overview of the PaPAM mechanism

The *Pa*PAM reaction goes through a cinnamate intermediate after elimination of the amino group and benzylic hydrogen from the α-amino acid substrate. Earlier deuterium isotope studies ($k_H/k_D > 2$) on a related aminomutase *Tc*PAM suggest the deprotonation step of the elimination reaction is rate-determining.[22] The coupling between the amine group of the substrate and the MIO is proposed to make a good alkyl ammonium leaving group. α,β-Elimination of the β-hydrogen and α-alkyl ammonium can advance through different routes. The concerted, one-step E2 (bimolecular elimination) mechanism proceeds through base-catalyzed removal of an acidic proton and a leaving group. By comparison, the two-step E1cB (unimolecular conjugate-base elimination) uses base-catalysis to remove a proton vicinal to a poor leaving group, yielding a carbanion intermediate. MIO-dependent aminomutase reactions likely follow an E2 or E1cB mechanism, where both depend on the rate of deprotonation of C_β, as proposed in an earlier work.[23] Thus, electron-withdrawing substituents on the aryl ring of the substrate that stabilize a δ^- charge on C_β should therefore increase the rate of the elimination step. In contrast, the two-step E1 (unimolecular elimination) reaction is not likely for MIO-dependent reactions. The attached, electron-withdrawing carboxylate of the substrate would destabilize the C_α carbocation formed after displacement of the NH_2-MIO adduct (Figure 3.3A).

The final reaction sequence of the MIO-dependent aminomutases involves an α,β-addition reaction, where the NH_2-MIO and a proton (H^+) add across the double bond of the acrylate intermediate. To obtain the β-amino acid in a concerted hydroamination, the polarity of the C_β (δ^+) needs to be opposite of that in the earlier elimination sequence. Here, the nucleophilic NH_2-MIO binds to C_β and the electrophilic H^+ attaches to C_α (Figure 3.3B).

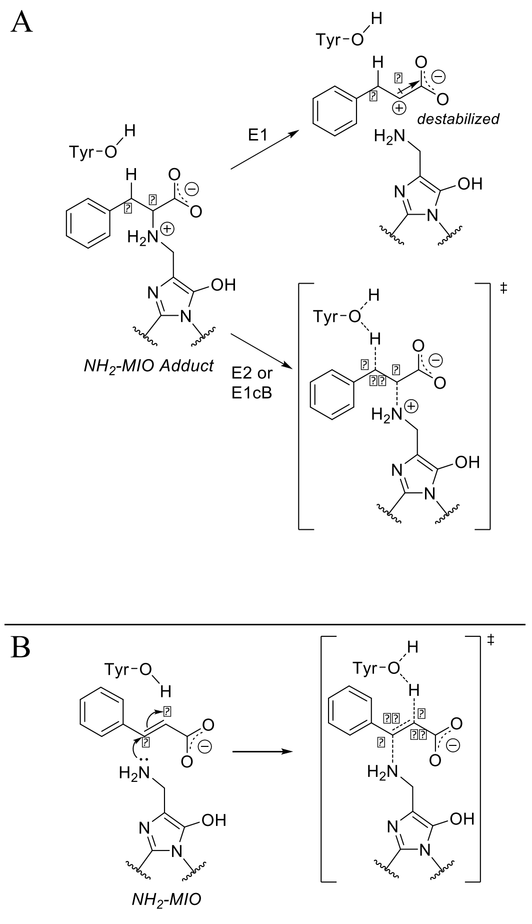


Figure 3.3 (A) Proposed elimination mechanisms for displacement of the NH_2-MIO adduct. E1: unimolecular, E2: bimolecular and E1cB: conjugate-base eliminations. (B) Concerted hydroamination of the acrylate intermediate. Shown is a transition state intermediate (right) highlighting the polarization of the π-bond in which the nucleophilic NH_2-MIO and the electrophilic H^+ approach C_β and C_α, respectively.

Alternatively, *Pa*PAM could use a stepwise addition sequence where the nucleophile (NH_2-MIO) couples to form a 1,4-Michael adduct. This conjugate addition route benefits from an electropositive (δ^+) C_β by delocalizing the π-electrons towards the carboxylate of the substrate. Theoretically, a substituent that places negative charge inductively within the ring or mesomerically the on C_{ipso} of the β-arylacrylate intermediate should also strengthen the formation of a δ^+ on C_β. These types of electrostatic considerations, along with binding affinity, were considered to explain the hydroamination reaction of *Tc*PAM for aryl acrylate substrates.[24,25]

In earlier accounts, the Michael addition mechanism was proposed,[26,27] but a presumed resonance structure has two repelling oxyanions on the carboxylate of the reactant that normally forms a monodentate salt bridge (Figure 3.4a), as evidenced in the *Pa*PAM crystal structure.[11] To alleviate build-up of this electrostatic repulsion, we propose that near-concerted protonation and amination of the π-bond likely minimizes formation of the unfavorable dianion (Figure 3.4b). A contrasting pathway is envisioned to first add a proton at C_α of the acrylate intermediate. The resulting intermediate has a positive charge (δ^+) on the benzylic C_β, which is resonance stabilized by the aryl ring and further stabilized by electron-releasing substituents (Figure 3.4c). Rapid, nucleophilic attack by the NH_2-MIO on the carbocation would ensue to complete the β-amino acid catalysis.

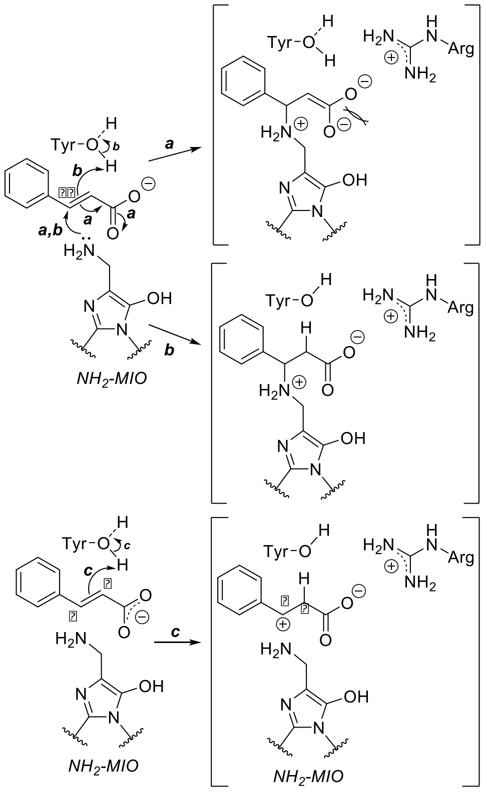


Figure 3.4 Route a) A stepwise Michael-addition pathway. Shown is an intermediate adduct (top right) with the π-electrons delocalized into the carboxylate group forming a repelling dianion prior to C_{α}-protonation. Route b) Concerted hydroamination of the acrylate π-bond. Shown is an intermediate (middle right) with maximal charge separation between repelling negative charges in the carboxylate group and the cation and anion. Route c) A stepwise hydroamination sequence. Shown is a proposed intermediate (bottom right) resulting from C_{α}-protonation as the first step, which places a positive charge at C_{β}. C_{β} is now primed for nucleophilic attack by the NH_2-MIO adduct.

Table 3.1 Kinetic Parameters[a] of *Pa*PAM for Various Substituted Aryl and Heteroaromatic Substrates.

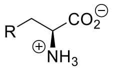
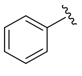
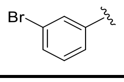
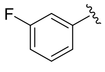
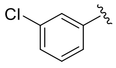
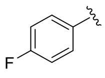
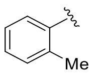
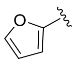
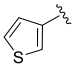
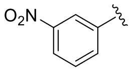
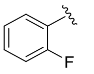
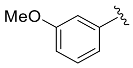
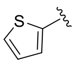
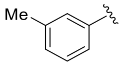
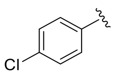
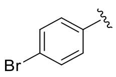
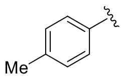
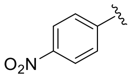
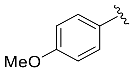
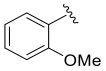
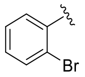
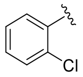
 R	K_{M}	k_{cat}^{β}	$k_{\mathrm{cat}}^{\mathrm{cinn}}$	$k_{\mathrm{cat}}^{\mathrm{total}}$	$k_{\mathrm{cat}}^{\mathrm{total}}/K_{\mathrm{M}}$
1 	168 (7)	0.301 92.8%	0.022 7.2%	0.323 (0.013)	1.93 (0.20)
2 	339 (15)	0.396 93.9%	0.024 6.1%	0.420 (0.014)	1.24 (0.12)
3 	27 (5)	0.027 85.2%	0.004 14.8%	0.031 (0.002)	1.2 (0.4)
4 	432 (26)	0.462 95.2%	0.022 4.8%	0.484 (0.02)	1.12 (0.14)
5 	29 (1)	0.020 85.7%	0.003 14.3%	0.023 (0.001)	0.79 (0.06)
6 	88 (6)	0.055 83.6%	0.009 16.4%	0.064 (0.002)	0.73 (0.09)
7 	415 (79)	0.143 34.8%	0.093 65.2%	0.236 (0.01)	0.588 (0.066)
8 	337 (27)	0.139 97.2%	0.004 2.8%	0.143 (0.004)	0.428 (0.063)
9 	430 (15)	0.136 92.6%	0.01 7.4%	0.146 (0.003)	0.340 (0.025)
10 	73 (6)	0.021 95.5%	0.001 4.5%	0.022 (0.001)	0.31 (0.04)
11 	990 (124)	0.201 99.0%	0.002 1.0%	0.203 (0.012)	0.209 (0.050)
12 	132 (5)	0.024 90.9%	0.002 9.1%	0.026 (0.001)	0.19 (0.02)
13 	204 (4)	0.048 78.3%	0.010 21.7%	0.058 (0.001)	0.19 (0.01)
14 	491 (82)	0.050 94.1%	0.003 5.9%	0.053 (0.003)	0.11 (0.03)
15 	525 (44)	0.043 95.6%	0.002 4.4%	0.045 (0.001)	0.09 (0.01)
16 	163 (9)	0.010 63.6%	0.003 36.4%	0.013 (0.001)	0.082 (0.010)
17 	752 (39)	0.025 48.0%	0.013 52.0%	0.038 $(<10^{-3})$	0.050 (0.005)

Table 3.1 (cont'd)

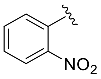
18		1187 (76)	0.022 97.7%	0.0005 2.3%	0.022 (<10^{-3})	0.019 (0.002)
19		164 (7)	0.002 70.0%	0.0007 30.0%	0.003 (<10^{-3})	0.02 (<10^{-2})



20



21



22

^{a}Standard error in parenthesis. Units: s^{-1} for k_{cat}, μM for K_{M}, and s^{-1}•M^{-1} × 10^3 for $k_{\mathrm{cat}}^{\mathrm{total}}/K_{\mathrm{M}}$. 20 – 22, not productive.

3.3.2 Comparing the effects of regioisomeric substituents on PaPAM catalysis and substrate affinity

The kinetic parameters of the *meta*/*para/ortho*-regioisomers (bromo-2/15/20; fluoro-3/5/10; chloro-4/14/21; nitro-9/17/22; methoxy-11/18/19; methyl-13/16/6) were compared. The binding affinities (estimated by K_{M}) for the fluoro- and methyl-substrate trifecta were approximately of the same order. However, the K_{M} of *Pa*PAM for the *o*-methoxy substrate 19 was nearly 10-times smaller than for its *meta*- and *para*-isomers (Table 3.1). The K_{I} values (μM) for *o*-bromo- (20), *o*-chloro- (21) and *o*-nitro- (21) substrates were 25-times smaller than the K_{M} values of *Pa*PAM for the corresponding *meta*- and *para*-isomers. This supported that the *ortho*-substituted substrates generally bound *Pa*PAM better than the *meta*- and *para*-isomers.

The relative binding affinity of each substrate was assessed as a function of the six substituents (of varying electronic and steric effects) in the *ortho*-, *meta*-, or *para*-position. The relative binding affinities predicted from the calculated energies of protein-ligand interactions and the internal energy of the ligand [$E_{(p\text{-}l)}$ + $E_{(l)}$] in the absence of energy minimization matched the trend (*m*~*p*>*o*) in the experimental K_M values for substrate isomers with halogens or nitro substituents (Tables A.3.1 and A.3.2). This supports the predictive value of the model in which the binding site residues and substrate maintain the positions found in the crystal structure with α-phenylalanine. The calculated vdW interaction energies ($E_{V(p\text{-}l)}$) also follow the "*m*~*p*>*o*" trend, except for chloro compounds, which bound less tightly to *Pa*PAM (i.e., had higher K_M) than predicted by $E_{V(p\text{-}l)}$ for chloro series, compared to other halogenated substrates (Tables A.3.1 and A.3.2). The chloro series will be discussed further in the activity cliff analysis section below.

Importantly, the binding affinity order for all substrates approximately corresponded to the vdW radii of the substituents. *Pa*PAM bound substrates with a fluoro group (~1.5 Å) the best, followed by methyl (~1.9 Å), then bromo and chloro groups (~1.8 Å). The least favorable substrate for binding to *Pa*PAM contained the bulkiest substituents: nitro (~3.1 Å; from the vdW radii of the C_{ar}–N bond length and the terminal O–N=O) and methoxy (~3.4 Å; from the vdW radii of the C_{ar}–O bond and the methyl C–H bonds of the methoxy).[29, 30] In general, *Pa*PAM was predicted by $E_{V(p\text{-}l)}$ to disfavor binding substrates with bulky groups at the *ortho*-position, which correlated well with the experimental K_M values. Surprisingly, substrates with *o*-methyl (6) (K_M = 88 μM) and *o*-methoxy (19) (K_M = 164 μM) groups bound *Pa*PAM better than expected from their calculated $E_{V(p\text{-}l)}$ (55 and 108 kcal/mol, respectively) (Tables A.3.1 and A.3.2). Binding of the *o*-methoxy group could become more energetically favorable if it rotated slightly from its crystallographic position to form hydrogen bonds with Tyr320 in *Pa*PAM (Figure A.2.1).

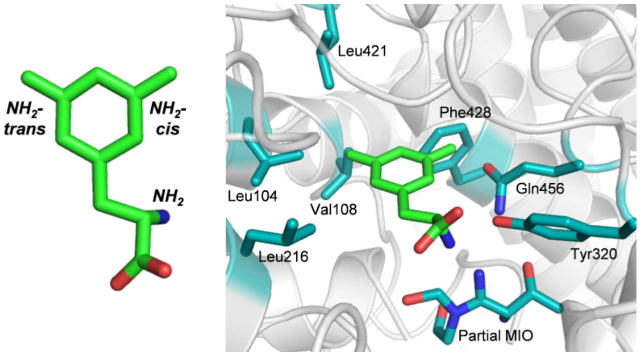


Figure 3.5 An overlay of the *NH$_2$-cis* and *NH$_2$-trans* configurations is illustrated, using the *m*-methyl-(*S*)-α-phenylalanine substrate (atoms are C, green; N, blue; O, red). The methyl group can be positioned on the same side (*NH$_2$-cis*) or the opposite side (*NH$_2$-trans*) as the reactive amino group of the chiral substrate (*left*). An overlay of the *NH$_2$-cis* and *NH$_2$-trans* active configurations of *m*-methyl-(*S*)-α-phenylalanine is modeled in the crystallographic position of α-phenylalanine in *Pa*PAM (PDB 3UNV). A partial MIO and the active site residues that cause van der Waals overlap with the ligands are shown (C, light blue; N, dark blue; O, red). SLIDE and other docking tools cannot model covalently bound ligands, which are interpreted as disallowed steric overlap (*right*). Thus, the alkene carbon atoms of the MIO were removed to dock the substrate.

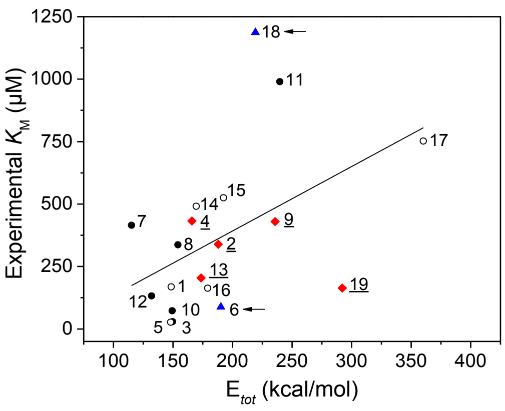


Figure 3.6 Plot of experimental K_M and $E_{tot} = E_{(p\text{-}l)}$ (protein-ligand interaction energy) + $E_{(l)}$ (the intra-ligand energy) calculated with Szybki. The substrates were modeled statically, according to the trajectory of α-phenylalanine in the *Pa*PAM crystal structure, without energy minimization. Substrates are labeled according to Table 3.1 and the lower energy of the two configurations [*NH_2-cis* (red ♦, underlined) and or *NH_2-trans* (blue ▲, arrowed)] is plotted for the substrates. Substrates with no significant difference in energy between the *NH_2-cis* and *NH_2-trans* (ΔE < 25 kcal/mol) are shown as filled dots (●). Substrates with *para*-substituents (except *p*-methoxy) without an *NH_2-cis* or *NH_2-trans* preference are open-circles (○). Non-productive substrates 20 – 22 (not shown) were predicted to prefer the *NH_2-trans* orientation in the *Pa*PAM active site.

3.3.3 Relationship between PaPAM-substrate interaction energies, flexibility, and K_M

The calculated interaction energies obtained from modeling provided insight into which energy terms correlated best with the K_M values of *Pa*PAM for each substrate. They also helped elucidate which substrate-docking model correlated best with experimental K_M. The static model placed the substrates identical to the trajectory of α-phenylalanine in the crystal structure. The flexible model, however, allowed bond-rotational motion for the protein side chains to relieve unfavorable interactions. The static modeling showed that the experimental K_M for each substrate (except for three unreactive *o*-bromo, *o*-chloro, and *o*-nitro substrates 20 – 22) increased with the total energy [$E_{(p\text{-}l)}$ + $E_{(l)}$], which approximated $\Delta G_{\text{binding}}$ and reflected unfavorable interactions (Figure 3.6). The linear correlation coefficient (*ccoef*) between [$E_{(p\text{-}l)} + E_{(l)}$] and K_M was 0.48 (Figure 3.6), while the *ccoef* between $E_{V(p\text{-}l)}$ and K_M was 0.54 (Figure A.2.4). Incidentally, the *ccoef* between the Coulombic energy [$E_{C(p\text{-}l)}$], a component of $E_{(p\text{-}l)}$, and K_M was lower (0.33; Figure A.2.3). These results suggested that the steric effects in the protein-ligand adduct and within the ligand are dominant over electrostatic interactions upon substrate binding. Moreover, when energy minimization was used to relieve vdW overlap between each substrate and the active site residues of *Pa*PAM (see Figure A.2.2), the *ccoef* between [$E_{(p\text{-}l)} + E_{(l)}$] and K_M decreased from 0.48 to 0.35. This result emphasizes the importance of vdW overlap-induced strain in affecting the binding affinity of *Pa*PAM for its substrates.

Another reason why energy minimization of the protein-ligand interaction likely affected the correlation between [$E_{(p\text{-}l)} + E_{(l)}$] and K_M is that, in some cases, groups were rotated that should have remained rigid. This may be due to inaccuracies in energy-minimization force field parameters for some functional groups, due to the prodigious challenge in deriving correct torsional energy barrier profiles for all bonds between all types of functional group that occur in organic molecules. For instance, the nitro substituent was rotated out-of-plane relative to the phenyl ring during energy minimization. However, our analysis of 200 nitrophenyl groups in small-molecule crystal structures in the Cambridge Structural Database 1.1.1 (http://webcsd.ccdc.cam.ac.uk) indicated that 87.5% of the nitrophenyl groups are entirely co-planar, regardless of other features in the structure.[31] The energy minimization-free protocol provided intermolecular energy values that correlated better with K_M. This observation suggests that the crystallographic placement of the substrates and *Pa*PAM was ideal for most substrates, and that modeling alternative, energy-minimized side group positions may reflect catalytically unproductive conformations.

Substrates were identified as either in the *NH_2-cis* or *NH_2-trans* configuration (Figure 3.5) if the difference (ΔE_{tot}) in the [$E_{(p\text{-}l)}$ + $E_{(l)}$] term for models of the two orientations was >25 kcal/mol (Tables A.3.3). Using this limit, *o*-methoxy- (19), *m*-methyl (13), *m*-bromo- (2), *m*-nitro- (9), *m*-chloro- (4) substrates were predicted to conform to the *NH_2-cis* configuration, while *p*-methoxy- (18), *o*-methyl- (6), *o*-chloro- (21), *o*-bromo- (20), and *o*-nitro- (22) substrates were predicted to favor the *NH_2-trans* configuration (Figure 3.6 and Table A.2.3). In substrate 18, the methyl of the methoxy group was predicted to adopt a quasi *NH_2-cis* configuration.

For *meta*-substituted substrates, the *NH_2-cis* is the preferred configuration because Leu104, Val108, and Leu421 sterically hinder the *NH_2-trans* conformers more than Gln456, Phe428, Gly85, Phe455, and Tyr320 hinder the *NH_2-cis* conformers (Figure 3.5). However, *m*-methoxy substrate 18 has no preference for the *NH_2-cis* or *NH_2-trans* configuration, as energy calculations suggest that the methoxy group interacts similarly with active sites resides on either side. It should be noted that Phe428, Val108, and Leu421 also sterically hinder substrates with *para*-substituted substrates. The *ortho*-substituted substrates (except for the *o*-methoxy substrate 19) are energetically more likely to adopt the *NH_2-trans* configuration. The *ortho*-substituted substrates have steric barriers created by residues Phe428, Gln456, and Tyr320 on the *NH_2-cis* side of *Pa*PAM (Figure 3.5). In addition, the *NH_2-trans* conformers of the *ortho*-substituted substrates encounter lower $E_{V(p\text{-}l)}$ between Leu216, Leu104 than between Tyr320, Gln456 of the *NH_2-cis* conformers (Figure 3.5). As mentioned previously, the *o*-methoxy substrate 19 bound to *Pa*PAM better than expected from its calculated vdW energy ($E_{V(p\text{-}l)}$) (Tables A.3.1 and A.3.2). The energy calculations predict that 19 favors the *NH_2-cis* conformer. This orientation is consistent with the hypothesis that the *o*-methoxy of 19 is near Tyr320 of *Pa*PAM and can potentially form an energetically favorable hydrogen bond. Of the nine substrates (1, 3, 5, 6, 10, 12, 13, 16, and 19) that bound *Pa*PAM the best ($K_M \lesssim 200$ µM, i.e., not >20% over the K_M of *Pa*PAM for 1), all except the *o*-methoxy substrate 19 ($E_{V(p\text{-}l)} = 108$ kcal/mol) had $E_{V(p\text{-}l)} \leq 55$ kcal/mol (designated as the energy threshold with low vdW overlap). On the other hand, the majority of poorest binding substrates, with $K_M > 500$ µM, and non-productive substrates had $E_{V(p\text{-}l)} \geq 80$ kcal/mol, with the *p*-nitro- (17), *o*-bromo- (20), and *o*-nitro- (22) substrates predicted to have comparatively higher vdW energy at $\gtrsim$190 kcal/mol (Table A.2.3). Relative binding energy, based on $E_{V(p\text{-}l)}$, is thus highly predictive of *Pa*PAM having a potentially high or low affinity for a substrate.

Generally, for productive substrates where the K_M of *Pa*PAM was ≤500 µM, the relative energy [$E_{(p\text{-}l)} + E_{(l)}$] of the *$NH_2$-cis* and *$NH_2$-trans* configurations tended to be ≤200 kcal/mol (see Table A.2.3). It was intriguing to find substrates that bind *Pa*PAM with the least affinity (highest K_M) (compound 18) or were non-productive (21, 20, 22) had differences of $\gtrsim$150 kcal/mol between the two orientations (see Table A.2.3). These results suggest that either the substituent on the substrate causes the enzyme to preferentially bind the substrate in one orientation over the other, or that low vdW barriers in the pocket enable the substrate to rotate to an active conformation for turnover.

The computational analyses identified residues that will help guide future mutational studies. Proposed mutations are envisioned to increase the binding affinity of *Pa*PAM for various substrates. The K_M of *Pa*PAM was higher for several substrates with *meta*- and *para*-substituents (except fluoro and methyl) than for 1. The presumed lower binding affinity was likely due to steric interactions between the substituents and the active site residues of *Pa*PAM. As mentioned herein, *meta*-substituted substrates were shown by modeling to prefer the *NH_2-cis* configuration to avoid steric clashes with branched hydrophobic residues. Mutation of Leu104, Val108, and Leu421 to alanines may improve the binding of *meta*-substituted substrates by providing flexibility to bind in the *NH_2-cis* or *NH_2-trans* configuration. Further, computational models predicted that *para*-substituents sterically clash with Phe428, Val108, and Leu421. Therefore, exchange of these residues for alanine may facilitate the binding of *para*-substituted substrates. Surprisingly, the computational analysis predicted that all *ortho*-substituted α-arylalanines bound well to *Pa*PAM ; however, relief of the active site sterics may enable these *ortho*-substituted α-arylalanines to better access a catalytically competent conformation and improve the turnover number for these substrates.

The flexible docking feature of SLIDE provided another approach to reduce vdW collisions between the crystallographic conformation of *Pa*PAM side chains and substituents on the arylalanine subtrates oriented in the *NH_2-cis* and *NH_2-trans* configuration. After application of the SLIDE flexibility modeling in the site, no significant correlation was found for SLIDE-calculated interaction energies and K_{M} values except for the unsatisfied polar interaction term: $\mathrm{E}_{(p\text{-}l)}$ (*ccoef* = 0.13), hydrophobic interaction energy, $\mathrm{E}_{H(p\text{-}l)}$ (*ccoef* = –0.19), and unfavorable energy of interaction due to unpaired or repulsive polar interactions, $\mathrm{E}_{UP(p\text{-}l)}$ (*ccoef* = 0.44). SLIDE also assessed the sum of unresolvable vdW overlaps in each complex, in Å, following flexibility modeling. The correlation of this value with K_{M}, *ccoef* = 0.27, was positive but somewhat lower than the correlation found between the Szybki intermolecular vdW energy and K_{M} in the absence of substrate or protein motion relative to the crystal structure (*ccoef* of 0.54). This is consistent with the decrease in correlation between Szybki intermolecular vdW energy and K_{M} (from 0.54 to 0.42) upon energy minimization, reflecting changes in the conformation of the complex. These results indicate that the favorability of vdW interactions and the absence of unsatisfied polar interactions when the substrate and protein are in their crystallographic conformation are the strongest predictors for favorable substrate K_{M}.

3.3.4 Activity cliff analysis

SALI values were used to identify "activity cliffs" that represent large changes in *Pa*PAM binding affinity among structurally-similar substrates.[20] The most obvious activity cliffs were found for substrates with fluoro-, methyl-, and chloro-substituents at the same positions (Figure 3.7). The chloro- and methyl-groups share similar vdW radii. When chloro is attached to an aryl ring carbon, its electron density delocalizes through resonance, placing a partial positive charge at the pole of the chloro atom furthest from the ring carbon.[32] The polarizability of the halogen atoms increases with atomic orbital size; therefore, the trend to form a halogen bond is in the order fluoro < chloro < bromo < iodo, where iodo normally forms the strongest interactions. Thus, the chloro- and bromo-substituents of substrates used in this study can act as electrophiles and can potentially form halogen bonds with nearby electron donor atoms, such as oxygen.

Favorable halogen-bonds between the halogen acceptor (X) and donor (O) have a C–X••••O angle of ~165° or a C–O••••X angle of ~120°, with a distance between X and O of ~3 Å.[32] However, the structure calculations and modeling revealed no evidence for chloro- or bromo-bonding between *Pa*PAM and the active orientation of the *o-, m-,* or *p*-chloro- or -bromo-substrates, based on searching for appropriate halogen-bond donors within 4 Å of the halogen. It is worth noting that the incompatibility between charged chloro groups and surrounding neutral carbon atoms in the binding pocket of *Pa*PAM may contribute to the higher K_M values for compounds with chloro-substituents relative to those with isosteric methyl-substituents. The *o*-, *m*-, *p*-fluoro substrates bound *Pa*PAM (K_M values between 27 and 73 μM) better than the natural substrate 1 (K_M = 168 μM), indicating a more favorable interaction between the fluoro group and surrounding hydrocarbon side chains.

In summary, vdW overlaps, estimated by the $E_{V(p\text{-}l)}$ in Szybki, and as the total sum (in Å) of vdW overlaps remaining following SLIDE docking, are most significant between the substrates and residues Phe428, Val108, Leu421, Leu104, Gln456 and Tyr320 of *Pa*PAM (Figure 3.5), which largely influence the binding affinity. Substrates without substituents on the aryl rings, the natural substrate 1, 2-furyl- (7), 2-thienyl- (12) and 3-thienyl- (8) alanine have no steric collisions with the binding site residues. This substrate specificity study was not exhaustive; there remain several arylalanine analogs to be tested in *Pa*PAM kinetics studies.

In the present study, the dependence of the reaction rate on the *Pa*PAM-catalyzed α/β-isomerization was probed with several arylalanine analogs. The influence of the substituents on the kcat of *Pa*PAM revealed a concave-down or a downward break in correlations with Hammett substituent constants (σ). The trend of these correlations[28] suggests that the rate-determining step changes from the elimination to the hydroamination step based on the direction and magnitude of the electronic properties of the substituent. In addition, the computational analyses provided a means to predict the docking conformation of substituted 22 arylalanine substrates. This information will guide future targeted amino acid mutagenesis of *Pa*PAM to increase the catalytic efficiency by improving the binding affinity for various other non-natural substrates.

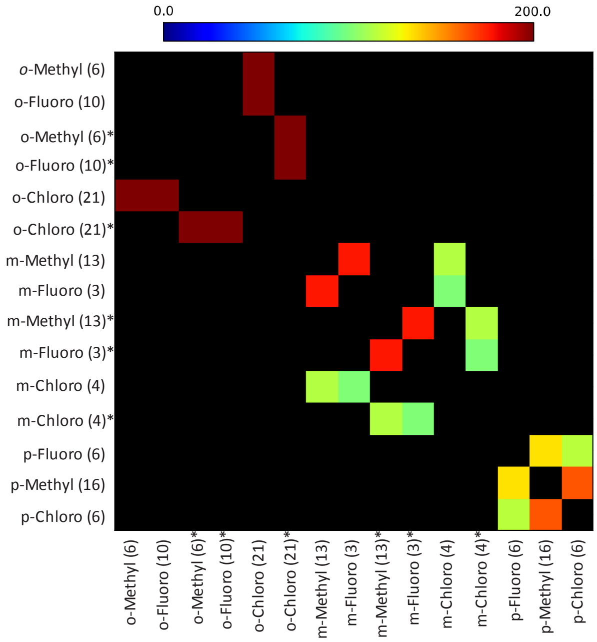


Figure 3.7 Structure-activity landscape index (SALI) analysis showing the subset of substrate pairs exhibiting a large change in K_M value upon a small change in structure. Substrate pairs with SALI scores near 200 (approaching red) indicate the most significant activity cliffs. Asterisks (*) indicate substrates in *NH*$_2$*-cis* configuration; all others are *NH*$_2$*-trans*.

APPENDIX

Table A.2.1 Comparison of the experimental K_M and predicted energetic order of each substituent at *ortho*-, *meta*-, *para*-positions.

	Fluoro-Substituents[a]			Chloro-Substituents[a]			Bromo-Substituents[a]		
	meta- (3)	*para*- (5)	*ortho*- (10)	*meta*- (4)	*para*- (14)	*ortho*- (21)	*meta*- (2)	*para*- (15)	*ortho*- (20)
K_M (μM)	27	29	73	432	491	-[c]	339	525	-
$E_{V(p\text{-}l)}$ (kcal/mol)	19	19	21	33	37	93	55	60	204
$(E_{(p\text{-}l)} + E_{(l)})$ (kcal/mol)	148	150	149	166	170	226	188	193	338
	Nitro-Substituents[a]			Methyl-Substituents[b]			Methoxy-Substituents[b]		
	meta- (9)	*para*- (17)	*ortho*- (22)	*ortho*- (6)	*para*- (16)	*meta*- (13)	*ortho*- (19)	*meta*- (11)	*para*- (18)
K_M (μM)	430	752	-	88 (I)	163 (II)	204 (III)	164 (I)	990 (II)	1187 (III)
$E_{V(p\text{-}l)}$ (kcal/mol)	48	186	205	55 (III)	46 (II)	40 (I)	108 (III)	86 (II)	81 (I)
$(E_{(p\text{-}l)} + E_{(l)})$ (kcal/mol)	236	360	393	190 (III)	179 (II)	174 (I)	292 (III)	240 (II)	219 (I)

[a]Computational approach correctly explained the trends in K_M values of substrate analogs.

[b]Trends in K_M did not correlate well with computationally predicted energy values, which fell within a relatively narrow range. Trends from most (I) to least (III) favorable are shown in (Roman numerals).

[c]Hyphens indicate non-productive substrates.

Table A.2.2 Comparison of the experimental K_{M} and predicted energetic order of each substituent at *ortho*-, *meta*-, *para*-positions. This data is the same as presented in Table A.2.1; here, it is organized according to substituent position rather than type.

	ortho-Substituents					
K_{M} (μM)	Fluoro	Methyl	Methoxy	Bromo	Chloro	Nitro
	73	88	164	-[a]	-	-
$E_{V(p\text{-}l)}$	Fluoro	Methyl	Chloro	Methoxy	Bromo	Nitro
(kcal/mol)	21	55	93	108	204	205
$(E_{(p\text{-}l)} + E_{(l)})$	Fluoro	Methyl	Chloro	Methoxy	Bromo	Nitro
(kcal/mol)	149	190	226	292	338	393
	meta-Substituents					
K_{M} (μM)	Fluoro	Methyl	Bromo	Nitro	Chloro	Methoxy
	27	204	339	430	432	990
$E_{V(p\text{-}l)}$	Fluoro	Chloro	Methyl	Nitro	Bromo	Methoxy
(kcal/mol)	19	33	40	48	55	86
$(E_{(p\text{-}l)} + E_{(l)})$	Fluoro	Chloro	Methyl	Bromo	Nitro	Methoxy
(kcal/mol)	148	166	174	188	236	240
	para-Substituents					
K_{M} (μM)	Fluoro	Methyl	Chloro	Bromo	Nitro	Methoxy
	29	163	491	525	752	1187
$E_{V(p\text{-}l)}$	Fluoro	Chloro	Methyl	Bromo	Methoxy	Nitro
(kcal/mol)	19	37	46	60	81	186
$(E_{(p\text{-}l)} + E_{(l)})$	Fluoro	Chloro	Methyl	Bromo	Methoxy	Nitro
(kcal/mol)	150	170	179	193	219	360

[a] Non-productive substrates are indicated by hyphens.

Table A.2.3 Evaluation of protein-ligand and ligand internal energy values and preference for *NH_2-cis* versus *NH_2-trans* configuration.

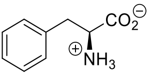
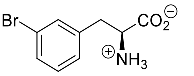
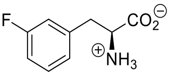
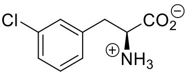
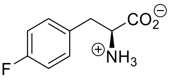
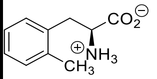
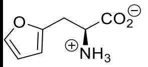
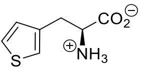
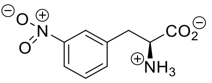
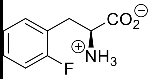
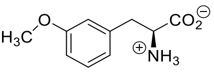
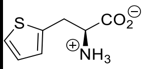
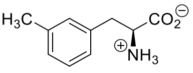
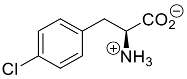
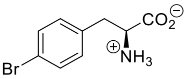
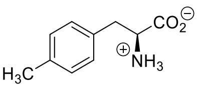
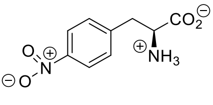
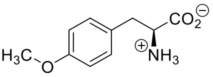
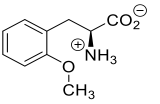
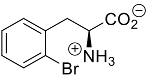
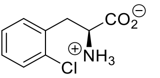
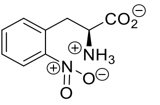
Substrate		*NH_2-trans* ($E_{(p-l)}$ + $E_{(l)}$)[a] (kcal/mol)	*NH_2-cis* ($E_{(p-l)}$ + $E_{(l)}$)[a] (kcal/mol)	$E_{V(p-l)}$[b] (kcal/mol)	K_M (μM)	Preferred Orientation[c]
1		149	149	19	168	Symmetrical[d]
2		429	188	55	339	*NH_2-cis*
3		153	148	19	27	NSD[e]
4		273	166	33	432	*NH_2-cis*
5		150	150	19	29	Symmetrical
6		190	489	55	88	*NH_2-trans*
7		133	115	21	415	NSD
8		156	154	21	337	NSD
9		1640	236	48	430	*NH_2-cis*
10		149	165	21	73	NSD
11		265	240	86	990	NSD
12		132	139	20	132	NSD
13		245	174	40	204	*NH_2-cis*
14		170	170	37	491	*Symmetrical*
15		193	193	60	525	*Symmetrical*
16		179	179	46	163	*Symmetrical*

Table A.2.3 (cont'd)

17		360	360	186	752	*Symmetrical*
18		219	947	81	1187	*NH2-trans*
19		409	292	108	164	*NH2-cis*
20		338	525	204	-f	*NH2-trans*
21		226	401	93	-	*NH2-trans*
22		393	2065	205	-	*NH2-trans*

[a]($E_{(p\text{-}l)}$ + $E_{(l)}$) is the sum of protein-ligand and ligand internal energy, where $E_{(p\text{-}l)}$ is the protein-ligand interaction energy and $E_{(l)}$ is the ligand internal energy. [b]$E_{V(p\text{-}l)}$ is the vdW energy of protein-ligand interaction, one of the terms contributing to $E_{(p\text{-}l)}$. The vdW energy is given for whichever orientation (*NH_2-cis* or *NH_2-trans*) had the lower, more favorable ($E_{(p\text{-}l)}$ + $E_{(l)}$) value. [c]Substrates were categorized as preferring an *NH_2-cis* or *NH_2-trans* configuration if the given orientation was at least 25 kcal/mol lower in ($E_{(p\text{-}l)}$ + $E_{(l)}$) value. [d]α-Phenylalanine and *para*-substituted substrates have symmetrical aryl rings with equal interaction energies for the *NH_2-cis* and *NH_2-trans* configurations. [e]Substrates observed to have no significant difference (NSD) in energy for the *NH_2-cis* or *NH_2-trans* configuration. [f]Non-productive substrates are indicated by hyphens. Note, all energies reported should be considered relative rather than absolute.

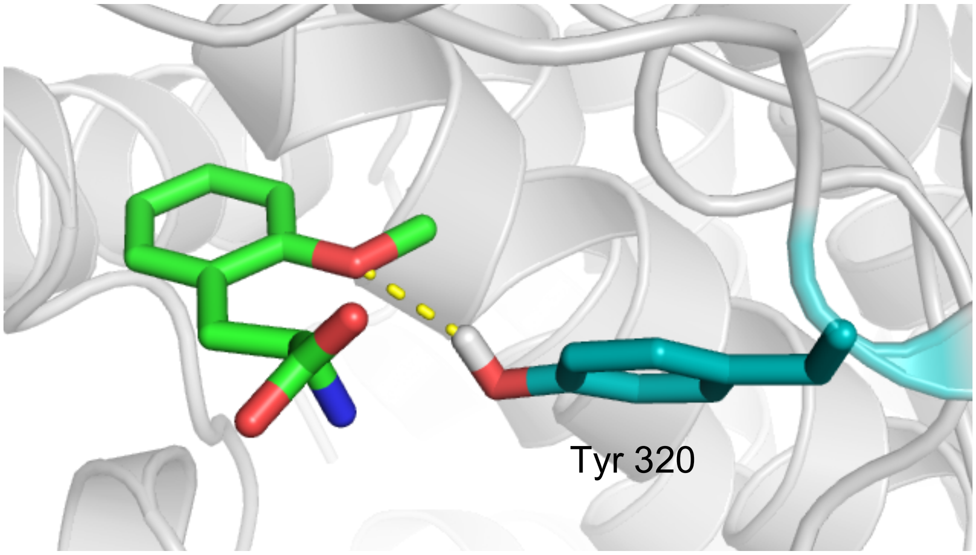


Figure A.2.1 H-bonding interaction of *ortho*-methoxy-α-phenylalanine (19) and active site Tyr320. *o*-Methoxy-α-phenylalanine atoms are colored as C, green; N, blue; O, red and Tyr320 atoms are colored as C, light blue; O, red; H, white.

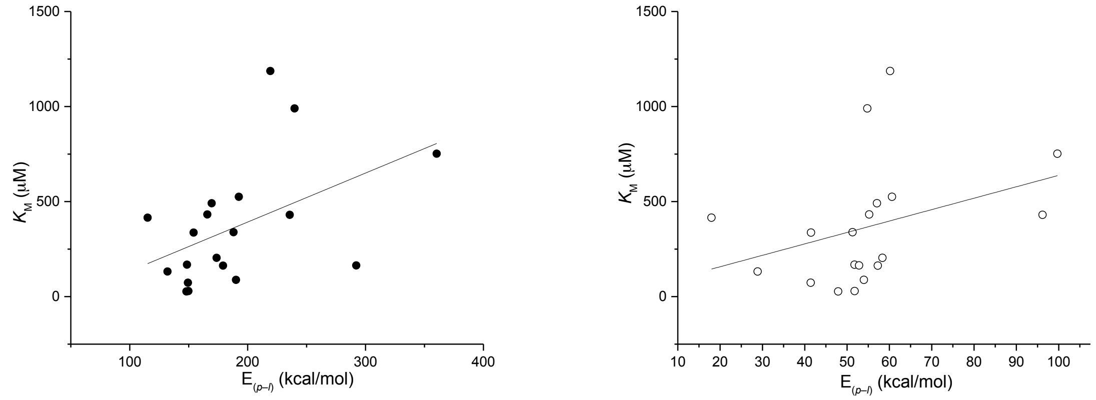


Figure A.2.2 Relationship between protein-ligand interaction energy $E_{(p\text{-}l)}$ and experimental K_M. Substrates were placed in the active site in *NH_2-cis* and *NH_2-trans* orientations overlaid with the crystallographic orientation of α-phenylalanine from PDB entry 3UNV, and the lower energy orientation was kept. Left panel: (●) Binding site residues of *Pa*PAM were maintained in their crystallographic orientation, yielding a linear correlation coefficient of 0.48 between $E_{(p\text{-}l)}$ and experimental $K_{M.}$ Right panel: (○) Energy minimization was used to reduce any repulsive interactions, leading to lower correlation between the resulting protein-ligand interaction energy and K_M value (correlation coefficient = 0.35).

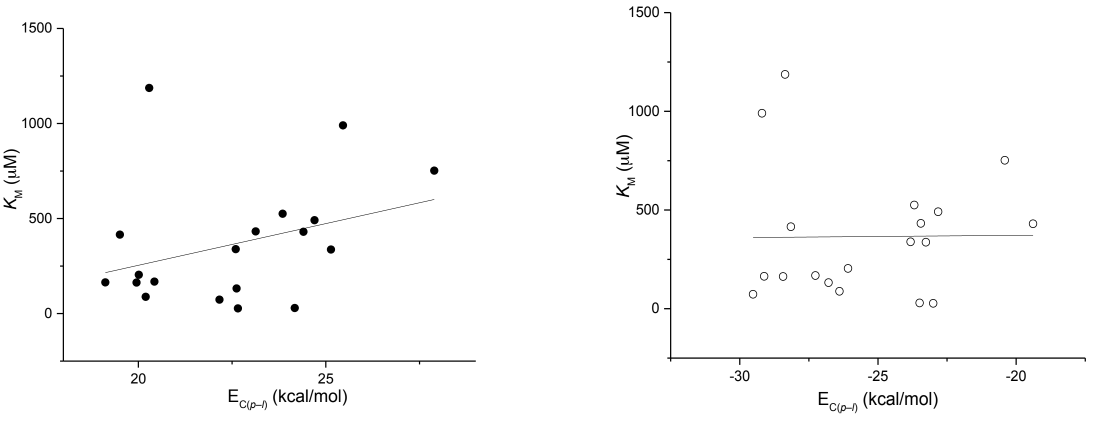


Figure A.2.3 Relationship between the electrostatic (Coulombic) component of the protein-ligand interaction energy $\mathrm{E}_{\mathrm{C}(p\text{-}l)}$ and experimental K_{M}. Substrates were placed in the active site in *NH$_2$-cis* and *NH$_2$-trans* configurations overlaid with the crystallographic orientation of α-phenylalanine, and the lower energy orientation was kept. Left panel: (●) Binding site of *Pa*PAM was kept in the crystallographic orientation (correlation coefficient = 0.33). Right panel: (○) Energy minimization was used to reduce any protein-ligand repulsive interactions (correlation coefficient = 0.011).

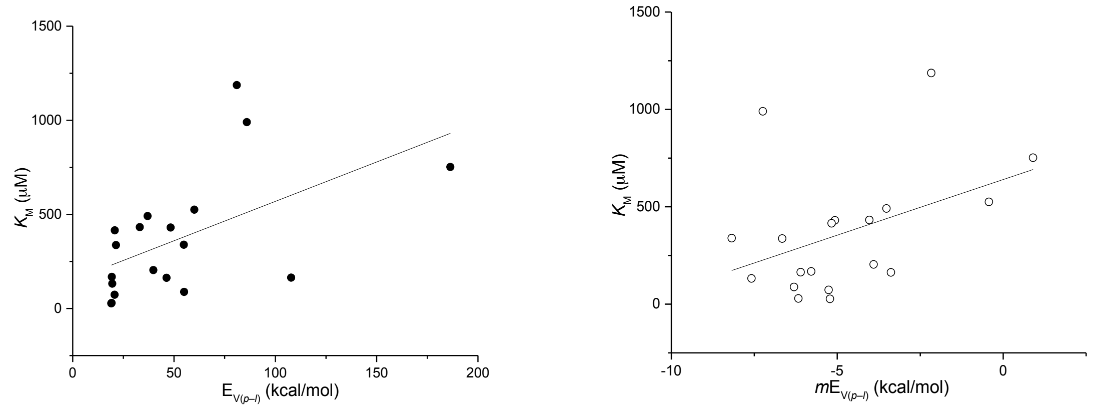


Figure A.2.4 Relationship between the van der Waals energy component of the protein-ligand energy $E_{V(p\text{-}l)}$ and experimental K_M. Substrates were again placed in *NH_2-cis* and *NH_2-trans* orientations overlaid with the crystallographic orientation of α-phenylalanine from PDB entry 3UNV, and the lower energy orientation was kept. Left panel: (●) Binding site residues of *Pa*PAM were kept in the crystallographic orientation (correlation coefficient = 0.54). Right panel: (○) Energy minimization was used to reduce any protein-ligand repulsive interactions (correlation coefficient = 0.42). These results indicate that the van der Waals interaction energy between the protein and each substrate overlaid with the α-phenylalanine-bound crystal structure is most predictive of the relative K_M values of the substrates.

REFERENCES

REFERENCES

(1) Horne, W. S. Peptide and peptoid foldamers in medicinal chemistry. *Expert Opin. Drug Discovery* 2011, *6*, 1247-1262.

(2) Ruf, S.; Buning, C.; Schreuder, H.; Horstick, G.; Linz, W.; Olpp, T.; Pernerstorfer, J.; Hiss, K.; Kroll, K.; Kannt, A.; Kohlmann, M.; Linz, D.; Hubschle, T.; Rutten, H.; Wirth, K.; Schmidt, T.; Sadowski, T. Novel β-Amino Acid Derivatives as Inhibitors of Cathepsin A. *J. Med. Chem.* 2012, *55*, 7636-7649.

(3) Huang, X.; O'Brien, E.; Thai, F.; Cooper, G. Practical Asymmetric Synthesis of RO5114436, a CCR5 Receptor Antagonist. *Org. Process Res. Dev.* 2010, *14*, 592-599.

(4) Jennewein, S.; Wildung, M. R.; Chau, M. D.; Walker, K.; Croteau, R. Random sequencing of an induced *Taxus* cell cDNA library for identification of clones involved in Taxol biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* 2004, *101*, 9149-9154.

(5) Klettke, K. L.; Sanyal, S.; Mutatu, W.; Walker, K. D. β-Styryl- and β-Aryl-β-alanine Products of Phenylalanine Aminomutase Catalysis. *J. Am. Chem. Soc.* 2007, *129*, 6988-6989.

(6) Magarvey, N. A.; Fortin, P. D.; Thomas, P. M.; Kelleher, N. L.; Walsh, C. T. Gatekeeping *versus* Promiscuity in the Early Stages of the Andrimid Biosynthetic Assembly Line. *ACS Chem. Biol.* 2008, *3*, 542-554.

(7) Chesters, C.; Wilding, M.; Goodall, M.; Micklefield, J. Thermal Bifunctionality of Bacterial Phenylalanine Aminomutase and Ammonia Lyase Enzymes. *Angew. Chem. Int. Ed.* 2012, *51*, 4344-4348.

(8) Feng, L.; Wanninayake, U.; Strom, S.; Geiger, J.; Walker, K. D. Mechanistic, Mutational, and Structural Evaluation of a *Taxus* Phenylalanine Aminomutase. *Biochemistry* 2011, *50*, 2919-2930.

(9) Röther, D.; Poppe, L.; Morlock, G.; Viergutz, S.; Rétey, An active site homology model of phenylalanine ammonia-lyase from *P. crispum*. J. *Eur. J. Biochem.* 2002, *269*, 3065-3075.

(10) Huang, S. X.; Lohman, J. R.; Huang, T.; Shen, B. A new member of the 4-methylideneimidazole-5-one–containing aminomutase family from the enediyne kedarcidin biosynthetic pathway. *Proc. Natl. Acad. Sci. U.S.A.* 2013, *110*, 8069-8074.

(11) Strom, S.; Wanninayake, U.; Ratnayake, N. D.; Walker, K. D.; Geiger, J. H. Insights into the Mechanistic Pathway of the *Pantoea agglomerans* Phenylalanine Aminomutase. *Angew. Chem. Int. Ed.* 2012, *51*, 2898–2902.

(12) Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* 2010, *50*, 572-584.

(13) Hawkins, P. C.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* 2012, *52*, 2919-2936.

(14) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* 2002, *23*, 1623-1641.

(15) Nicholls, A.; Wlodek, S.; Grant, J. A. *J. Comput. Aided Mol. Des.* SAMPL2 and continuum modeling. 2010, *24*, 293-306.

(16) Wlodek, S.; Skillman, A. G.; Nicholls, A. Ligand Entropy in Gas-Phase, Upon Solvation and Protein Complexation. Fast Estimation with Quasi-Newton Hessian. *J. Chem. Theory Comput.* 2010, *6*, 2140-2152.

(17) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* 1996, *17*, 490-519.

(18) Zavodszky, M. I.; Rohatgi, A.; Van Voorst, J. R.; Yan, H.; Kuhn, L. A. Scoring ligand similarity in structure-based virtual screening. *J. Mol. Recognit.* 2009, *22*, 280-292.

(19) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput. Aided Mol. Des.* 2002, *16*, 883-902.

(20) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* 2008, *48*, 646-658.

(21) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* 2007, *50*, 74-82.

(22) Mutatu, W.; Klettke, K. L.; Foster, C.; Walker, K. D. Unusual Mechanism for an Aminomutase Rearrangement: Retention of Configuration at the Migration Termini. *Biochemistry* 2007, *46*, 9785-9794.

(23) Schuster, B.; Rétey, J. The mechanism of action of phenylalanine ammonia-lyase: The role of prosthetic dehydroalanine. *Proc. Natl. Acad. Sci. U.S.A.* 1995, *92*, 8433-8437.

(24) Wanninayake, U.; DePorre, Y.; Ondari, M.; Walker, K. D. (*S*)-Styryl-α-alanine Used To Probe the Intermolecular Mechanism of an Intramolecular MIO-Aminomutase. *Biochemistry* 2011, *50*, 10082–10090.

(25) Weiner, B.; Szymanski, W.; Janssen, D. B.; Minnaard, A. J.; Feringa, B. L. Recent advances in the catalytic asymmetric synthesis of β-amino acids. *Chem. Soc. Rev.* 2010, *39*, 1656-1691.

(26) Szymanski, W.; Wu, B.; Weiner, B.; de Wildeman, S.; Feringa, B. L.; Janssen, D. B. Phenylalanine Aminomutase-Catalyzed Addition of Ammonia to Substituted Cinnamic Acids: a Route to Enantiopure α- and β-Amino Acids. *J. Org. Chem.* 2009, *74*, 9152-9157.

(27) Ratnayake, N. D.; Wanninayake, U.; Geiger, J. H.; Walker, K. D. Stereochemistry and Mechanism of a Microbial Phenylalanine Aminomutase. *J. Am. Chem. Soc.* 2011, *133*, 8531-8533.

(28) Hoffmann, J.; Klicnar, J.; Štěrba, V.; Večeřa, M. *Collect. Czech.* Kinetics of hydrolysis of substituted salicylideneanilines. *Chem. Commun.* 1970, *35*, 1387-1398.

(29) Batsanov, S. S. Van der Waals radii of elements from the data of structural inorganic chemistry. *Russ. Chem. Bull.* 1995, *44*, 18-23.

(30) Li, A. J.; Nussinov, R. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins* 1998, *32*, 111-127.

(31) Carpy, A. J. M.; Haasbroek, P. P.; Ouhabi, J.; Oliver, D. W. Keto/enol tautomerism in phenylpyruvic acids: structure of the o-nitrophenylpyruvic acid. *J. Mol. Struct.* 2000, *520*, 191-198.

(32) Metrangolo, P.; Meyer, F.; Pilati, T.; Resnati, G.; Terraneo, G. Halogen Bonding in Supramolecular Chemistry. *Angew. Chem. Int. Ed.* 2008, *47*, 6114-6127.

Chapter 4 Using multiple virtual screening techniques to bootstrap pheromone antagonist discovery

4.1 Introduction

4.1.1 Motivation

Virtual screening techniques have been used for human drug discovery successfully.[1-3] It is novel to apply these techniques in other fields such as agriculture and aquatic species. Here, we integrated these techniques into a screening pipeline to find antagonists to control an aquatic invasive species, the sea lamprey, in the Great Lakes (in collaboration with Prof. Weiming Li's lab in the Department of Fishery and Wildlife). We aimed to find an effective as well as environmentally friendly solution to control the population of sea lamprey in the Great Lakes. Sea lamprey is a well-known invasive species, which causes billions of dollars lost to the commercial fishery and threatens the survivals of large and medium fish.[4,5] Current strategies to control the sea lamprey population in the Great Lakes either trap the sea lamprey at a significant yet not a large rate 10% over the years[6] or employ chemical pesticides that unfortunately threaten the lake sturgeon, an endangered species in 19 out of the 20 states.[7]

The in-silico screening part of this pipeline focuses on hypothesis-driven ligand-based virtual screening due to its computational efficiency, assisted by structure-based virtual screening (docking ligand candidates into a protein structure). To manage and record data systematically and improve the screening efficiency to avoid repeatedly retrieving molecular data from the file containing giant (12 million molecules) small molecule dataset, a fellow graduate student in the Kuhn lab, Sebastian Raschka, designed a SQLite tool to store and retrieve small molecule data matching our criteria for inhibitor candidates.

4.1.2 Hypothesis

7a,12a,24-trihydroxy-3-one-5a-cholan-24-sulfate (3kPZS, Figure 4.1) is the pheromone released by male sea lamprey,[8] to attract the sexually mature females to the nesting grounds. Our hypothesis is that blocking the female detection of 3kPZS via blocking SLOR1 receptor is expected to halt the propagation of this invasive vertebrate species. 3kPZS specifically binds to the SLOR1 receptor at a concentration of 10^{-12} M.[8] Therefore, environmentally friendly inhibitors that mimic 3kPZS to compete for binding to the ligand binding site in SLOR1 and inhibit 3kPZS detection at small concentrations are a desirable approach for sea lamprey control. Success in such an approach has been shown for invasive insect control.[9]

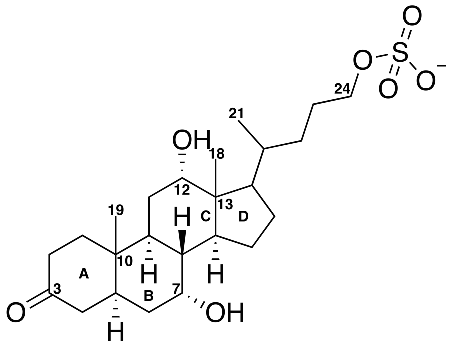


Figure 4.1 Structure of 3kPZS.

4.1.3 Significance

Discovery of antagonists of SLOR1 to mimic 3kPZS binding is based discovering mimics of the structure of this bile acid-like compound, which binds specifically to olfactory receptors, which are G-protein coupled receptors (GPCRs). Approximately 50%-60% of modern medicinal drugs are targeted at GPCRs.[10] GPCRs can interact with compounds like odor molecules, pheromones, hormones, and neurotransmitters, and also with other proteins and peptides.[11] Sea lamprey olfactory receptors are homologous to the rhodopsin or class A branch of the human GPCR family tree.[12] This provides a great case to study specificity of ligand recognition in the GPCRs, including our bile acid ligand, 3kPZS. Furthermore, it is a brand new application of virtual screening drug discovery techniques to control aquatic invasive species, which can be applied to other projects to discover agonists or antagonists based on known ligand structures.

4.2 Materials and Methods

Here, we design a complete pipeline (Figure 4.2) to integrate multiple screening techniques and experimental assays, to successfully discover inhibitors for the sea lamprey olfactory receptor 1 (SLOR1) by focusing on discovering mimics of its native ligand, 3kPZS. In the pipeline, we designed a database tool that annotates structure information of molecules from multiple compound libraries, which allows efficient identification of compounds matching functional groups we hypothesize are important to 3kPZS detection. The selected compounds are evaluated based on shape and electrostatic similarities with 3kPZS. The compounds after scoring for shape and electrostatic similarity are secondarily selected according to functional group matches. These selected compounds are then evaluated based on their docking into the SLOR1 receptor homology model. The candidates that have favorable interactions with the receptor binding site were then prioritized for experimental assays. Experimental results act as feedback to strengthen or refine the initial hypothesis and modify the screening strategy if required.

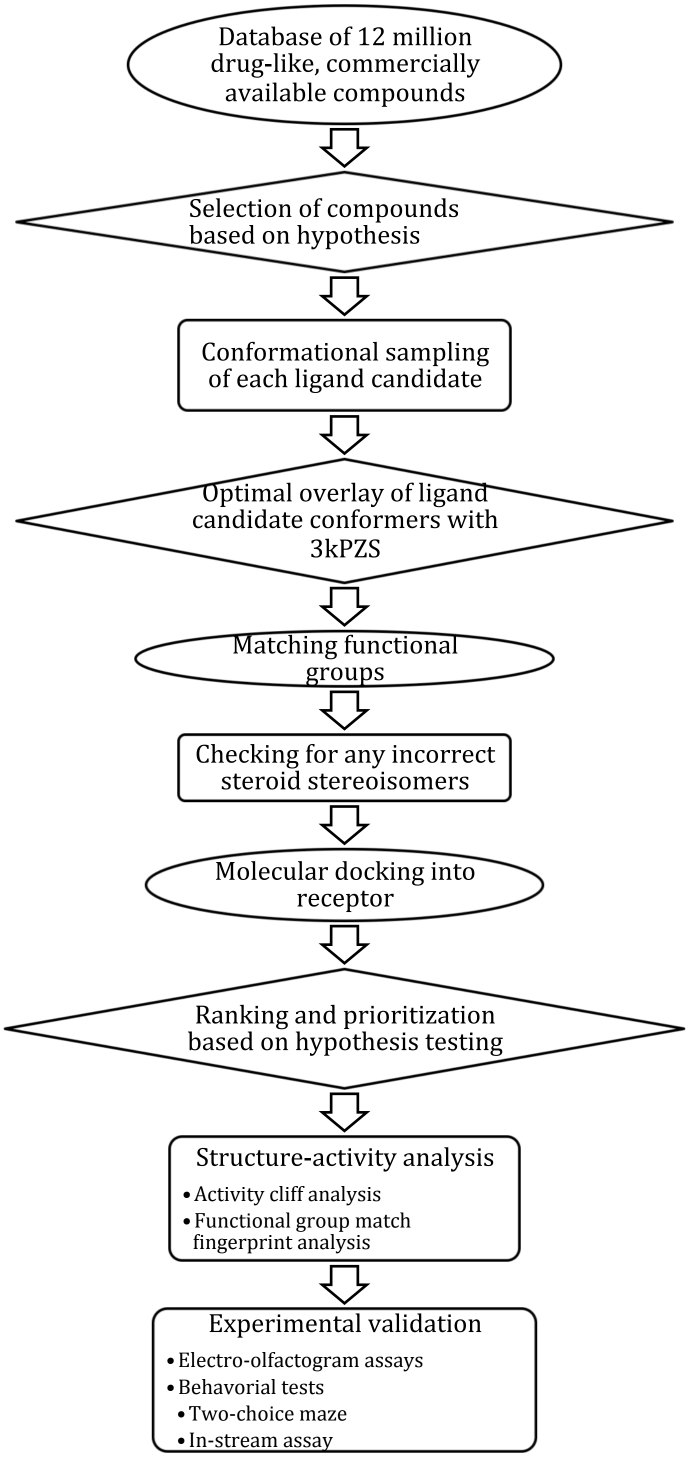

Database of 12 million drug-like, commercially available compounds
Selection of compounds based on hypothesis
Conformational sampling of each ligand candidate
Optimal overlay of ligand candidate conformers with 3kPZS
Matching functional groups
Checking for any incorrect steroid stereoisomers
Molecular docking into receptor
Ranking and prioritization based on hypothesis testing
Structure-activity analysis
• Activity cliff analysis
• Functional group match fingerprint analysis
Experimental validation
• Electro-olfactogram assays
• Behavorial tests
• Two-choice maze
• In-stream assay

Figure 4.2 Flowchart describing the pipeline for 3kPZS antagonist discovery. In step 2, an example of the potential hypothesis is that known GPCR ligands are likely to mimic 3kPZS and block SLOR1.

4.2.1 Virtual screening

4.2.1.1 3kPZS and SLOR1 structural model

The homology model of SLOR1 was constructed by Dr. Kuhn through the ModWeb implementation of Modeller[13] (version SVN.r972), based on the alignment of SLOR1 with the avian β1-adrenergic receptor crystal structure (Protein Data Bank entry 2vt4).[14] The sequence identity between the sequences of SLOR1 and avian β1-adrenergic receptor is 25.5% with an e-value of $6.4e^{-11}$, which indicates that there is a very low probability to obtain an alignment with this level of amino acid similarity at random. According to a previous statistical study, if two sequences have at least 24.8% identity over no fewer than 80 residues, then their corresponding main-chain structures are closely related.[15] This criterion is satisfied by the alignments between sequences of SLOR1 and β1-adrenergic receptor. The region corresponding to the orthosteric binding site in SLOR1 model was obtained by overlaying the volume of ligand binding in the main-chain structures of other related class A GPCRs such as rhodopsin, A2A and β1-adrenergic receptor.

The set of all favorable-energy conformations of 3kPZS was docked into the SLOR1 ligand binding cavity using SLIDE software with default settings[16] to predict the 3kPZS-SLOR1 mode of interaction.[17]

4.2.1.2 Preparation of the screening libraries based on hypothesis

The “Drug-Like" subset of the ZINC12 database containing 13.2 million compounds, is the largest screening library we used and processed by Santosh Gunturu and Sebastian Raschka.[18] All of the compounds in this subset satisfy "Lipinski's rule of 5",[19] as listed below:

- molecular mass ranging from 150 to 500D,
- octanol-water partition coefficient no greater than 5,
- no more than 5 hydrogen bond donors and no more than 10 hydrogen bond acceptors,
- no more than 7 rotatable bonds
- polar surface area less than 150 Å^2.

There are 7.4 million compounds labeled as “in stock” in the 13.2 million compound subset, which were the compounds of focus. The information on these compounds such as ZINC IDs, purchasability, number of rotatable bonds and functional groups were stored in an SQLite database for screening. The advantage to store the information in an SQLite database is that it allows fast selections of molecules based on an initial hypothesis (e.g., inhibitors must include a terminal sulfate group), as well as easy manipulation of data such as insertion, deletion, editing and storage of data. It enhanced the safety of record keeping, and reproducibility of screening results.

According to former results from Dr. Li's lab, a hypothesis has been proposed that an effective antagonist should contain a sulfate group matching the 3kPZS 24-sulfate and at least one oxygen (hydroxyl or keto) group matching the 3-oxygen position in 3kPZS. Based on this hypothesis, a subset of compounds containing a sulfate group and at least one oxygen (hydroxyl or keto) group were selected in the SQlite database. This results in fewer than 100,000 compounds, a much smaller database than the initial 7.4 million available compounds. This step reduces the cost dramatically in the following, computationally expensive steps.

I prepared the following databases that contain additional analogs of 3kPZS and known ligands for G protein-coupled receptors for screening, based on different hypotheses.

The combinatorial analog data set contains 332 variants of 3kPZS, sampling different combinations of functional groups at 3, 7 and 12 positions and steroid ring configurations that were designed as 2D structures by our collaborators Dr. Mar Huertus and Anne Scott in Dr. Li's lab. For example, hydroxyl groups and keto groups at the 3, 7 and 12 position (Figure 4.1) are substituted by keto oxygen, hydroxyl oxygen or hydrogen. The 5-beta configuration is replaced by a 5-alpha steroid ring, the configuration in 3kPZS. The carboxylate, sulfate or phosphate group were allowed at the C-24 position. According to the SMILES strings of the 332 variants, I generated 3-dimensional structures of these compounds by using OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com) and partial charges were assigned using molcharge 1.3.1 (Open Eye Scientific Software). The SMILES strings of the 332 compounds served as input to SciFinder (https://scifinder.cas.org) to find commercially available compounds using a similarity search with a threshold of 99%, corresponding to the subset of these compounds that have been synthesized and can be purchased. The corresponding CAS numbers (molecular database identifier) were exported and any duplicates were deleted. In the end, 84 unique compounds were found to be available.

To test the effect of configurations of steroid ring systems on biological activities, 5-beta steroids with bent configuration, different from the 5-alpha steroid with planar configuration as in 3kPZS, were identified in the ZINC12 database by substructure searching using the 5-beta steroid SMILES string “C1CCC3[C@@H(C1)CC[C@H]4[C@@H]2CCCC2CC[C@H]34”. Because 5-beta steroids are easier to synthesize, active compounds with this configuration would reduce experimental costs. In the end, a set of 690 compounds with 5-beta steroid configuration that had not been previously included in the ZINC12 screening library were extracted, with a subset of 200 compounds that were drug-like.

A final set of 2,995 steroid structures that are commercially available through other vendors while not already in the ZINC12 database was established, by searching steroid analogs in the CAS Registry database using SciFinder. Because the ZINC12 database does not cover all vendors of small organic molecules, the CAS Registry database containing 91 million compounds can serve as a complementary database to search steroid molecules with commercial vendors that are not present in the ZINC12 database. Using SciFinder, ~8000 CAS Registry steroids were exported. The SMILES strings of 2995 steroids could be determined from their CAS registry numbers by CACTUS (http://cactus.nci.nih.gov), which is a server to translate the compounds' information in different formats. Then, 3-dimensional structures of the 2995 steroids were built based on their SMILES strings by using OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com). Partial charges were assigned using molcharge 1.3.1 (Open Eye Scientific Software).

The GPCR ligand library (GLL) database, of ~25,000 known ligands for 147 GPCRs (http://cavasotto-lab.net/Databases/GDD/)[20] was also prepared for screening. The sdf files in the GLL data package downloaded from http://cavasotto-lab.net/Databases/GDD/Download/ were converted to 3D structures using OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com). In addition, the 3D structures of the ligands of the trace amine-associated receptors (TAAR) were generated by using their isomer SMILES strings by OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com), because these compounds are not included in the GLL database. Then partial charges were assigned to the atoms in all compounds by using molcharge 1.3.1 (Open Eye Scientific Software).

Using EOG assays, we identified 12 compounds from the above databases that suppress at least 45% of response to 3kPZS, based on measuring the blockage of the olfactory neurological response to 3kPZS by using a technique called an electro-olfactogram (EOG; assayed by Dr. Mar

Huertas & Anne Scott in the Li lab). We then identified ZINC compounds that share high similarity to the identified 12 compounds following the hypothesis that they are likely to have similar activity. Therefore, the compounds with >90% molecular similarities to the identified 12 compounds that are commercially available were extracted from the ZINC12 database for screening.

4.2.1.3 Sampling flexible compounds

Before virtual screening based on the 3D structures of molecules, multiple 3 dimensional conformations for each molecule need to be sampled to access all the possible low-energy conformations that the molecules can adopt in nature. The remaining compounds from the above databases were used to generate low energy conformers by using OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com). This is a tool that uses a knowledge-based approach to generate hundreds of low-energy conformers for each molecule. It can sample conformers with validated quality at a high speed of 2-2.5 sec/molecule on a machine with a standard computer configuration of 2.4 Ghz CPU, 4GB RAM. Overall, there are three steps to generate low energy conformers in OMEGA 2.4.6 (OpenEye Scientific Software, Santa Fe, NM; http://www.eyesopen.com)'s algorithm. The first step is to assemble an initial set of conformations of molecules from the pre-calculated chemical fragment library. The second step is to generate a large ensemble of conformations by applying all torsions in the molecules based on a pre-built torsion sampling dictionary. Lastly, a scoring function based on a modified

MMFF94 force field is used to eliminate conformers with internal clashes. To guarantee the uniqueness of sampled conformers, conformers with low pairwise RMSD threshold are eliminated as well.[21] There are on average 200 low energy conformers generated for each molecule in the above screening libraries. The distance between the 3-hydroxyl group and the 24-sulfate group in the potential active conformers of 3kPZS ranges from 13 to 20 Å. Therefore, conformers of database molecules with corresponding functional groups within this distance were selected to improve the efficiency and efficacy of screening.

4.2.1.4 Overlays of molecular structures using ROCS

The compounds in the above compound libraries after sampling and selection were then overlaid, one by one, with the 48 low-energy conformations of 3kPZS by using ROCS (OpenEye Scientific Software, Santa Fe, NM).[22] ROCS (OpenEye Scientific Software, Santa Fe, NM)[22] is a ligand-based approach to calculate the degree of similarity in shape and chemical properties of compounds compared with target ligands. It overlays the structures of molecules quickly and also supports multiprocessing, which makes it a feasible tool for ligand-based virtual screening. According to OpenEye reports, ROCS can overlay 20-40 molecules per second using a single CPU of a standard computer (2.4 Ghz CPU, 4 GB RAM). It uses a Gaussian function to represent the volume of each atom and a partial charge model to calculate chemical matches.[22] The ROCS structure similarity score, TanimotoCombo score, ranks the database compounds with value ranging from 0.0 for no match to 2.0 for a perfect match, equally weighting shape match and partial charge match. The distribution of ROCS scores across the ZINC12 database has a mean value of 0.64, with a standard deviation of 0.08. We considered compounds with scores greater than 2 standard deviations above the mean as significant matches. Compounds in this region of the score distribution were kept from all the 124 partitions of ZINC and evaluated for functional group matches to 3kPZS.

4.2.1.5 Matching functional groups in 3kPZS

A subset of the substituent groups in 3kZPS, including 3-keto, 7-OH, 12-OH, ester O conjugated to C-24, two methyl groups in the steroid ring and the terminal organosulfate group are hypothesized to be essential to the biological activity of 3kZPS. Tabulation of these functional groups for each compound in the screening libraries was performed and stored in the SQLite database (using code developed by Santosh Gunturu and Sebastian Raschka), according to suitable atomic charge threshold and hybridization state. If compounds in the screening libraries have corresponding atoms with proper charge and hybridization state within 1.0Å of these functional groups in the best-matching 3kPZS conformer, then these groups are labeled as matches.

4.2.1.6 Incorrect steroids

In order to check whether the stereochemistry of the compounds in the above database satisfy the natural stereochemistry of steroids, steroid checking by using SMILES representation of the canonical steroid ring system is performed using OpenEye toolkit. This was necessary because ZINC12 samples and includes unnatural isomers in cases where the vender did not provide complete stereo-chemical information for compounds. Also, duplicate compounds with the same chemical structures and vendors but different ZINC IDs are deleted based on the properties of their SMILE strings.

4.2.1.7 Molecular docking

Compounds with high ROCS TanimotoCombo score and functional group matches were docked into the binding site in the SLOR1 homology model by Santosh Gunturu and me using SLIDE. Then the compounds were evaluated by the ability to form a salt bridge with His 110 in the binding site, believed to be a crucial interaction for the pheromone, as well as the degree of isostericity of the steroid ring substructure. In addition, molecules with obvious steric clashes with the binding site were of lower priority experimental tests and molecules with favorable $\Delta G_{binding}$ values based on SLIDE scores were increased in priority.

4.2.1.8 Ranking and prioritization based on hypothesis testing

Using the datasets and screening toolkit described above, a series of hypotheses were defined to select subsets of compounds for EOG assays.

Hypothesis 1: Compounds containing the 3-keto and one or more sulfate oxygens in the *functional group matching with 3kPZS, and TanimotoCombo score above 0.85 will mimic 3kPZS.* This criterion tests the hypothesis that compounds that are highly similar to 3kZPS in overall shape and electrostatic properties, and contain the 3-keto and sulfate groups, can compete with and block detection of 3kPZS.

Hypothesis 2: Compounds containing a 3-hydroxy group, a sulfate oxygen, and at least one of the other functional groups present in 3kPZS, such as sulfate oxygen, hydroxyl or methyl, in the functional group matching with 3kPZS, and a TanimotoCombo score no less than 0.8 and ROCS electrostatic score no less than 0.25 will mimic 3kPZS. This hypothesis tests whether compounds that are highly similar to 3kZPS overall and match its electrostatic properties, and contain the 3-hydroxyl and sulfate groups, can inhibit detection of 3kPZS. In addition, as a secondary consideration, compounds that have a sulfate group that can dock close to His110 in the SLOR1 binding site with a docking energy of -7kcal/mol or less are more favorable and prioritized.

Hypothesis 3: Compounds that interact with the β1-adrenergic receptor will be active against the SLOR1 receptor. Because β1-adrenergic receptor is the known structure that has the highest overall and binding site sequence identity to SLOR1, the compounds that interact with β1-adrenergic receptor are selected for testing, including carvedilol (agonist; ZINC01530579), atenolol (selective antagonist; ZINC00014007), and dobutamine (partial agonist; ZINC00003911).

Hypothesis 4: Compounds with a 5-alpha steroid ring configuration and functional groups matching the 3-keto and sulfate oxygen in 3kPZS, and a ROCS TanimotoCombo score greater than 0.65 will mimic 3kPZS. We selected analogs with the same steroid ring configuration and which matched the oxygen-containing groups in 3kPZS, to test whether they mimic 3kPZS in activity

Hypothesis 5: Compounds with phosphate tail. Compounds with phosphate instead of sulfate tails at the terminus (C24) were selected to test whether phosphate can be a potential replacement to the sulfate moiety of 3kPZS and block detection of 3kPZS.

Hypothesis 6: Compounds with a 5-β steroid ring configuration and at least 2 sulfate oxygen matches or at least 5 functional group matches all together mimic 3kPZS. This tests whether bent steroids can simulate the interaction between planar steroids and SLOR1.

Hypothesis 7: Compounds with more negative charged sulfate group (with charges 0.3 units more negative than the sulfate oxygen charge in 3kPZS) will bind more tightly. It is postulated that compounds with a more negatively charged tail can form stronger salt bridges with His110 at the binding site of SLOR1 and more strongly compete with 3kZPS for binding.

Hypothesis 8: Compounds containing negatively charged, non-sulfate oxygen atoms matching at least one sulfate oxygen in 3kPZS would also compete with 3kPZS for binding

SLOR1. Additionally, compounds need to contain atoms matching the 3-keto and at least one of the other functional groups in 3kPZS, and have a ROCS TanimotoCombo score value of 0.8 or above, and a ROCS electrostatics complementarity value 0.25 or above. The compounds were further evaluated according to the distance between the sulfate tail group and His110 for the ability to make the salt bridge believed to confer SLOR1-3kPZS specificity and a docking score (<-7kcal/mol) assessing overall favorability of interaction. This test the hypothesis that compounds with negative, non-oxygen atoms at the tail can mimic the function of sulfate oxygen.

Hypothesis 9: Steroids containing epoxide. Epoxide containing compounds are reported to be able to form a covalent interaction with histidine in the binding site of at least one protein.[23] Therefore, steroids containing epoxide at the tail position or 3-O position were selected to test whether epoxide at these positions can form covalent bond with histidine or other residues nearby, to generate specific and permanent inhibition of SLOR1.

Hypothesis 10: Steroids with taurine tail. Because taurolithocholic acid has shown strong inhibition of 3kPZS detection in the EOG assays, we selected its analogs with taurine tail and high ROCS TanimotoCombo score with 3kZPS, to test the hypothesis that the steroids containing a taurine tail can block 3kPZS detection.

4.2.1.9 Activity cliff analysis

As mentioned in Chapter 3, activity cliff analysis[24] enables us to find the essential functional groups in molecules, in which slight chemical changes cause dramatic changes in the biological activities. Activity cliff analysis involves 2 matrices: structural similarity (which we can measure with ROCS and activity similarity (measured by EOG). To measure structural similarity for each pair of compounds that was tested by EOG, they were first overlaid with 3kPZS (using the best matching conformer of each relative to the docked conformer of 3kPZS), and the TanimotoCombo score of that ROCS overlay was reported. The SALI activity cliff score was then calculated using the following equation. The pairs with SALI score above 70 were analyzed.

$$\mathrm{SALI} = \frac{100 * |\text{activities difference}|}{(2 - \mathrm{RocsCombo})}$$

4.2.1.10 Functional group match fingerprint analysis

In order to find the relationship between structure and activity of the assayed 143 compounds as potential SLOR1 antagonists, functional group matchprint analysis was performed. There are 3 steps to perform this analysis. Step 1: multiple conformers of the 143 compounds were overlaid with 48 potentially active conformers of 3kPZS using ROCS. The top scored conformers of the 143 compounds were selected. Then functional group matchprints were generated by comparing the positions of the sulfate oxygens, sulfate ester oxygen, 3keto oxygen, 3-OH, 7-OH, 12-OH, 18-methyl and 19-methyl groups of the compounds with 3kPZS.

Generation of the matchprints was based on the code developed by Santosh Gunturu. Step 2: the matchprints of the top 6 compounds that suppressed EOG response of sea lamprey to 3kPZS were extracted as references (Table 4.1).

Table 4.1 The matchprints of the top 6 compounds that suppressed EOG response of sea lamprey to 3kPZS.

Zinc ID	(0-3) Sulfate Oxy	(0-1) Sulfate Ester	(0-1) 3-Keto Oxy	(0-1) 3-OH	(0-1) 7-OH	(0-1) 12-OH	18- Methyl	19- Methyl
ZINC72400307_28	3	0	0	1	1	1	1	1
ZINC35044325_22	3	0	0	1	0	0	1	1
ZINC04095893_61	2	0	0	0	0	0	2	1
ZINC72400309_95	3	1	0	0	0	0	1	1
ZINC12494532_16	3	1	0	0	0	0	0	0
ZINC01845398_1	3	1	0	0	0	0	0	0

Step 3: All matchprints were compared with the six matchprints, following the rationale that compounds that can match the presence/absence of functional groups in the six most active compounds are also likely to be active. By comparing the differences in the matchprints of other compounds with the top 6 compounds, we can determine the functional groups whose presence/absence results in enhance/reduction of biological activities. If an assayed compound differed in the presence/absence of these functional groups by at most one position, relative to the six most active compounds, then it was extracted for structure-activity analysis.

4.2.2 Experimental validation

Proposed antagonists were tested in Dr. Li's lab by three assays: (1) electro-olfactograms, which test the ability of sea lampreys to sense a compound in their olfactory epithelia, and whether this compound competes with 3kPZS sensing[25] and (2) behavorial tests including the two-choice maze with one channel containing the compound and the other channel containing a blank and (3) in-stream assay, which is set up like the two-choice maze but in an actual flowing stream with a more elaborate set of criteria for tracking sea lampreys' responses to a compound.

4.2.2.1 EOG assays

The protocol of EOG assays[25] was developed in Dr. Li's lab and carried out by Dr. Mar Huertas and Anne Scott. In EOG assays, female sea lampreys were anesthetized with 100 mg/L MS-222 and placed in a Plexiglas V-shaped stand. Continuous aerated water containing 50 mg/L MS-222 kept the gills irrigated and the fish anesthetized throughout the experiments. Then the surface skin in the nose was removed to expose the olfactory lamellae. A small capillary tube was used to deliver the chemical stimuli to the epithelial cells of olfactory rosette. The electrical potential changes upon detection of the tested stimulus were recorded through two Ag/AgCl electrodes (type EH-1S, World Precision Instruments, Sarasota, Florida, USA), which were placed between two lamellae and adjusted to maximize the ratio of signal to noise. Then, the signal was amplified and digitalized to analyze the ability of each stimulus to reduce the detection of 3kPZS.

To test whether the prioritized compounds affect the detection of 3kPZS, a mixture of 10^{-6} M 3kPZS and 10^{-6} M concentration of a potential antagonist was exposed to the olfactory epithelium for 4s. The EOG value of the charcoal filtered water was used as a blank control and subtracted from the EOG value upon exposure to the mixture, to normalize the response. Between different stimuli, there was a 2 min interval, in which the olfactory epithelium was flushed with charcoal filtered water. After measurement of three potential antagonists, the EOG value of the blank control (charcoal filtered water), 10^{-6} M 3kPZS, and 10^{-5} M L-arginine were recorded as well.

L-arginine was used as control to test whether the reduction of EOG response was caused by blocking of 3kPZS detection or by a general suppression of olfactory detection. L-arginine is a strong stimulus to sea lamprey,[26] and is known to interact with the olfactory epithelium through other mechanisms rather than by competing with 3kPZS.[27] Pre-exposure to one of the two stimuli will not influence the EOG response of the other.

4.3 Results and discussion

4.3.1 Binding mode of 3kPZS in SLOR1 structural model

Based on structural modeling of the SLOR1 receptor for 3kPZS done by Prof. Kuhn, the critical interactions between 3kPZS and residues at the binding sites are expected to include salt bridges, H-bond and hydrophobic interactions as shown in Figure 4.3. Salt bridges are formed between the sulfate tail of 3kPZS and protonated nitrogen atoms on the side chain of His110 in SLOR1. Tyr203 also forms a hydrogen bond with the sulfate tail, and there is an additional hydrogen bond between the Cys194 main chain and the 12-hydroxyl group in 3kPZS. The hydrophobic steroid ring system of 3kPZS forms favourable hydrophobic interactions with the hydrocarbon side chain groups from Phe87, Met106, Leu109, His110, Asp196, Pro277, Tyr280, and Thr284 in the binding site. The docked binding mode is consistent with the cholate binding mode predicted for this site in SLOR1 by using CholMine (Figure 4.4, http://cholmine.bmb.msu.edu, see Chapter 2).[28]

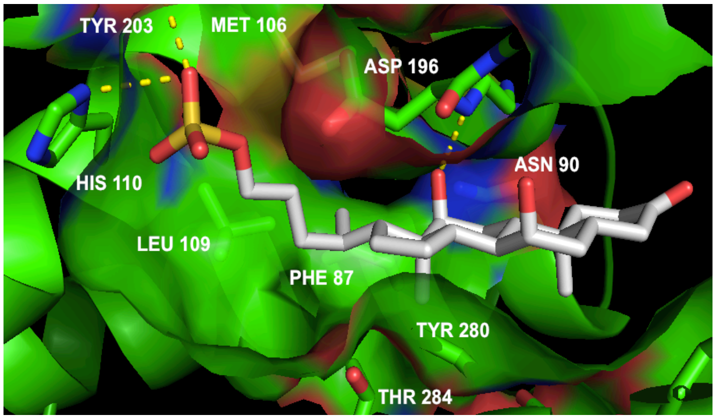


Figure 4.3 Interactions between 3kPZS and SLOR1 predicted by homology modeling and SLIDE docking performed by Dr. Leslie Kuhn and Qinghui Yuan. SLOR1 side-chain atoms and the binding site surface are colored green for carbon atoms, blue for nitrogen, red for oxygen, and yellow for sulfur. Carbon atoms of 3kPZS are shown in white tubes (center), with hydrogen bonds and salt bridges to the receptor shown as yellow dashed lines. The sulfate ester moiety is predicted to bind deep in the SLOR1 cleft (left), forming salt bridges with His110. The methyl-group face of the steroid ring (bottom-center) interacts with an entirely hydrophobic face of the cleft in SLOR1.

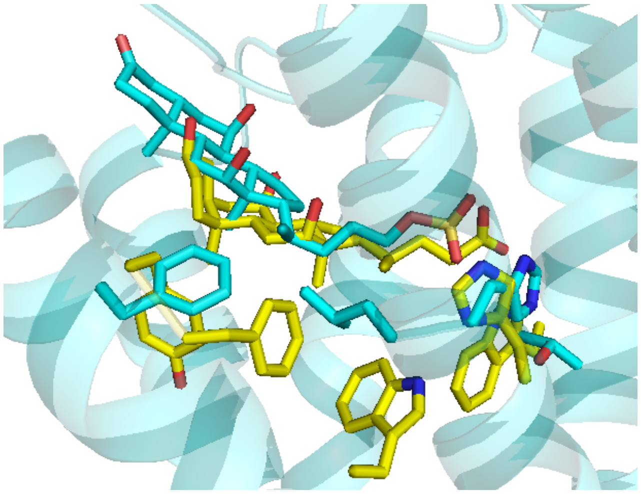


Figure 4.4 Bile acid binding motif in SLOR1 identified based on conserved features of cholate binding in a set of unrelated proteins (yellow), relative to the SLIDE docking orientation of 3kPZS (blue). The predicted binding orientation for cholate (horizontal molecule at center, with carbon atoms in yellow tubes) substantially overlays with the docked 3kPZS molecule (blue horizontal molecule), despite the bent (5-beta) cholate steroid ring in place of the relatively planar (5-alpha) 3kPZS steroid. Their negatively charged sulfate tail groups are predicted in highly similar positions (center-right). Side chains making key interactions with cholate in cytochrome C oxidase (PDB entry: 2DYR) are shown below in yellow (Tyr, Phe, Trp, and His), and SLOR1 side chains interacting with 3kPZS (Tyr, Leu, His) are shown in blue.

4.3.2 Electro-olfactograms (EOGs) assays identify antagonists for 3kPZS detection based on candidates from high-throughput computational screening

To provide feedback regarding our hypotheses on features important for small molecules to block the detection of 3kPZS, a histogram was created of the 143 assayed compounds as a function of their percentage reduction in sea lampreys' olfactory detection of 3kPZS (Figure 4.5). The structures of 8 out of the 11 most effective compounds that inhibit 3kPZS detections by at least 45% are noted on the histogram. The four most active compounds are sulfonated and have steroid backbones. In addition, there are two drug-like molecules without steroid ring structures and two alkyl tail analogs that apparently mimic the sulfate group in 3kPZS. In addition to the 8 most effective compounds in Figure 4.5, there are additional 3 compounds that inhibit detection of 3kPZS by at least 45%, including two steroid compounds, that is ZINC70666191 and 52205-73-9(CAS registry number), and one long alkyl chain compound ZINC1532179 with 12 carbon tail. However sulfonamides, such as the two non-steroidal compounds with ~0.5 activity, are known to be pain-assay interference compounds.[29]

Similar to the two drug-like compounds on the upper-right corner of Figure 4.5, three more recently assayed compounds, ZINC03531326 reduced 3kPZS detection by 43%, ZINC13790354 by 42%, and ZINC09227487 by 41%. These types of compounds all have alternative heterocyclic and hydrophobic rings with different linkers instead of steroid structures.

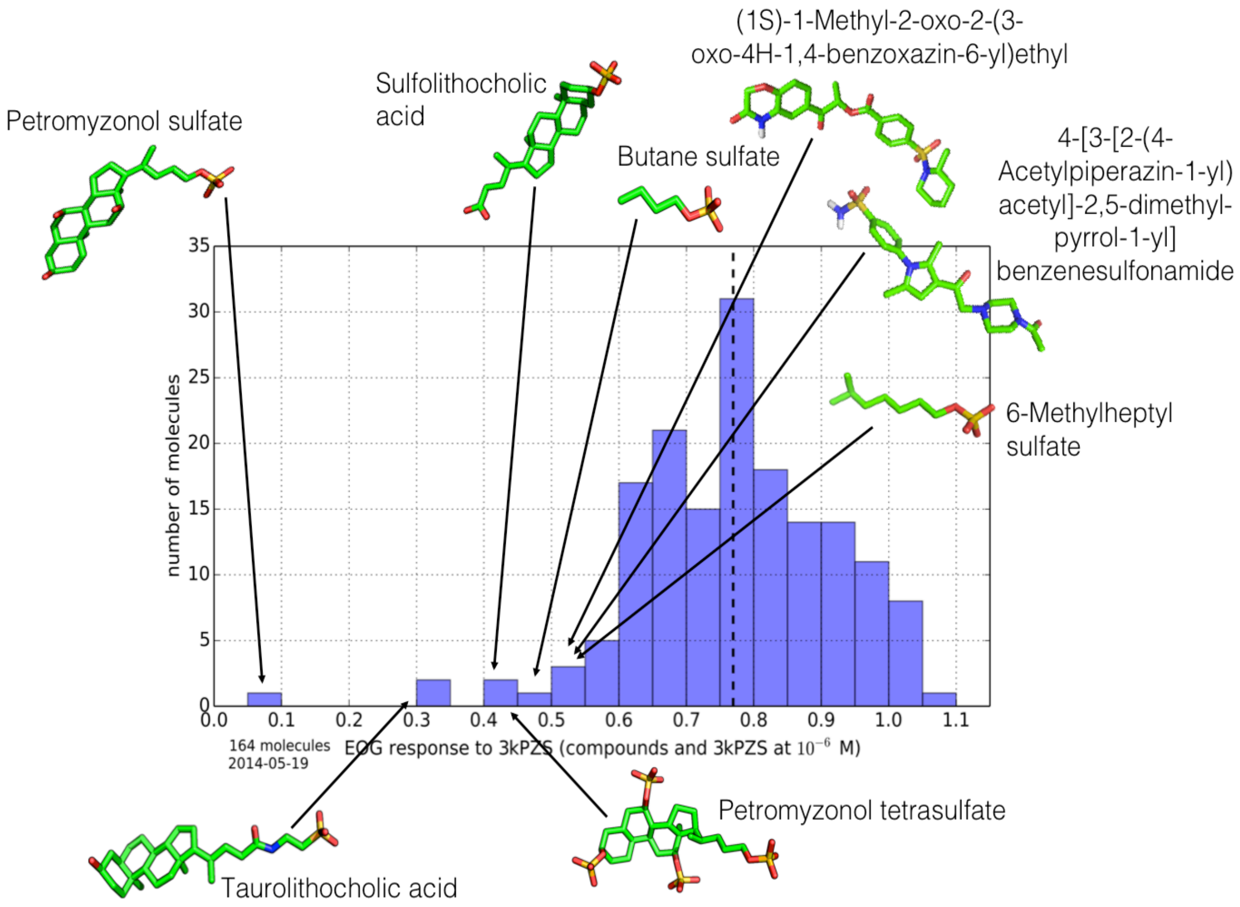


Figure 4.5 Histogram of the first 143 compounds according to their percent reduction in 3kPZS olfaction by sea lampreys. Chemical structures and names are shown for the eight most active compounds, which exhibit >45% reduction of 3kPZS response.

4.3.3 Structure-activity relationships analysis

4.3.3.1 SAR analysis based on SALI and functional group matchprint

We analyzed the structure-activity relationships for the EOG-assayed 143 compounds using structure-activity landscape index (SALI), as mentioned in Chapter 3. The higher the SALI score, the more significant an activity cliff there is. In the above equation, ROCS TanimotoCombo score is used to evaluate the similarities between the assayed compound, and to generate a heatmap of a SALI landscape. The pairs of compounds with SALI scores above 70 were selected (Figure 4.6), in which the pairs with only one functional group difference were analyzed (Figure 4.7). As shown in Figure 4.7(A), a pair of taurolithocholate analogs only differs in the presence or absence of the 7-OH group. The compound without 7-OH was twice as active as the compound with 7-OH. This phenomenon is consistent with 3kPZS docking results, in which there are no obvious favorable interactions between SLOR1 and the 7-OH of 3kPZS (Figure 4.3, with -OH group appearing near the Asn90 label). As shown in Figure 4.6 (B), a pair sulfate of tail analogs with the same carbon chain length but different tail functional groups show that the sulfate tail compound is 17% more active than the phosphate tail compounds. The more negatively charged sulfate group may have stronger interactions with HIS 110 in the binding site than the less negative phosphate group.

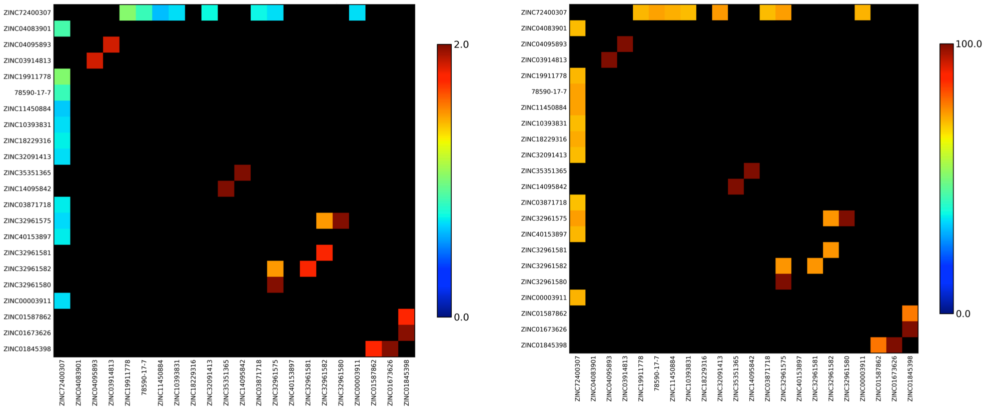


Figure 4.6 (A) ROCS TanimotoCombo scores for the pairwise compounds with significant activity cliffs with SALI score ≥ 70. (B) SALI scores for the pairwise compounds with significant activity cliffs with SALI score ≥ 70.

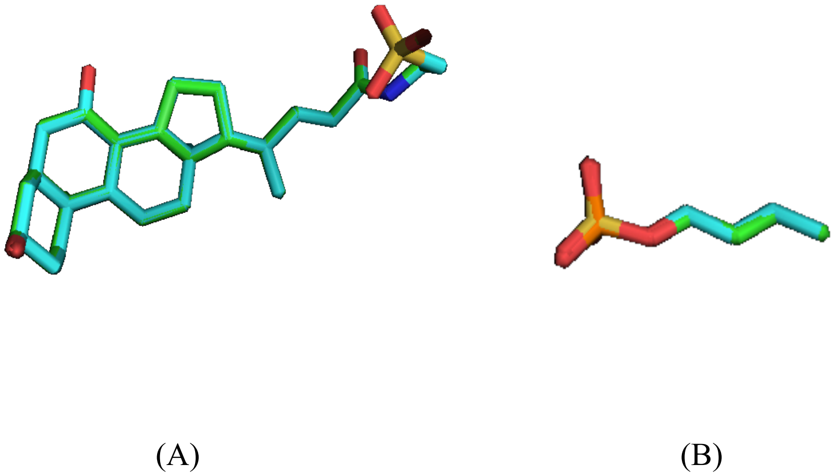


Figure 4.7 (A) Compound without 7-OH group (in green) is twice as active (70% reduction in 3kPZS response) as compound with 7-OH (in blue; 35% reduction). Tail structure is same in both. (B) Butane sulfate is 16% more active than butane phosphate.

In functional group matchprint analysis, we analyzed the pairs with only one functional group difference at the positions of the sulfate oxygens, sulfate ester oxygen, 3-keto, 3-OH, 7-OH, 12-OH, 18-methyl and 19-methyl groups in 3kPZS. A series of tail analogs with carbon chains of various lengths (Figure 4.8) that differ by one functional group relative to another compound were assayed by EOG and their structure-activity relationships are analyzed. The tail analog with 4 carbons in the chain has the highest activity and the analogs with 8 and 12 carbons have similar activities to the analogs with 4 carbons. The analogs with 5 and 6 carbons have low activities, in which the one with branches has the worst performance. It is possible that aliphatic chain can be used to substitute the steroid backbone. The angatonist' activities fluctuate according to the length of the carbon tail.

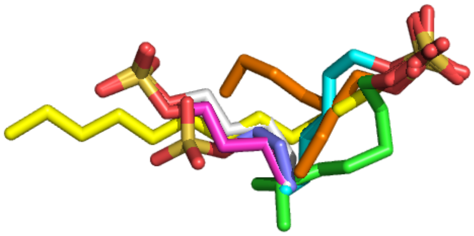


Figure 4.8 Assayed compounds with aliphatic tails. Shown in purple is the 3 carbon compound (ZINC01587861) with 38% inhibition of EOG response of 3kPZS; Shown in red is the 4 carbon compound (ZINC01845398) with 50% inhibition of EOG response; Shown in gray is the 5 carbon compound (ZINC01587862) with 32% inhibition of EOG response; Shown in cyan is the 6 carbon compound (ZINC01841381) with 31% inhibition of EOG response; Shown in orange is 6 carbon compound (ZINC01680379) with ethyl group, which inhibits EOG response by 18%; Shown in green in the 8 carbon compound (ZINC14591952) with 48% inhibition of EOG response. 0.52; Shown in yellow is the 12 carbon compound (ZINC01532179) with 46% inhibition of EOG response.

4.3.3.2 Other structure-relationship analysis

Six of the 11 most active compounds, which reduced the response to 3kPZS by 45-100%, had steroidal substructures. Both 3kPZS and the antagonist candidates were tested at 10^{-6} M concentration in the initial EOG assays. Surprisingly, several of these antagonists had none of the canonical hydroxyl groups on the steroid ring system. However, the three most active compounds, including PZS (the 3-OH analog of 3kPZS), which nullified the response to 3kPZS by 92%, all had 3-hydroxyl groups in place of the 3-keto group present in 3kPZS. This was a valuable discovery, because previous data from the Li lab[17] indicated that only 3kPZS could activate the SLOR1 receptor, not PZS, suggesting that PZS did not bind to SLOR1. Our hypothesis is that PZS successfully competes with 3kPZS for binding to SLOR1, which leads to its antagonist activity. We aim to test this by developing a receptor-based ligand-binding assay in collaboration with Prof. Rick Neubig (Chair, Pharmacology & Toxicology, MSU). Such an assay will also facilitate structure-activity relationship analysis (how antagonist side groups influence SLOR1 activation or inhibition) and structure-based antagonist optimization for this pheromone receptor.

Six of the 11 most active compounds mimic 3kPZS by matching the C and D steroid rings and C24-sulfate group. This result shows that the 3-keto oxygen and steroid ring system in 3kPZS can be removed or substituted by other functional groups. Because sulfate tails exist in both steroidal and non-steroidal compounds, it is considered as an indispensable functional group in effective antagonists. As shown in Figure 4.8, one of the simplest compounds with sulfate tail, 1-butane sulfonate can reduces EOG response of 3kPZS by 51%.

4.4 Conclusion

Antagonists that inhibit 3kPZS detection can potentially hinder the mating process of sea lamprey by blocking the ability of female sea lamprey to detect this pheromone, and aid in controlling lamprey population. Based on this rationale, we developed an effective and efficient antagonist discovery pipeline based on the hypothesis of overall volumetric and electrostatic mimicry of 3kPZS and its important functional groups for binding to SLOR1. Through this pipeline, ~300 potential antagonists were prioritized from a screening library of compounds, of which 143 compounds were tested in EOG assays. Of the 143 compounds, 11 compounds that inhibit 3kPZS at least 45% were identified. Three compounds, including PZS, taurolithocholic acid (TLC) and tetrasulfonated-PZ (tetra-PZS), were shown to be behavioral antagonists in the two-choice maze and PZS was shown to be behavioral antagonist in both maze and stream tests. It is most interesting that PZS, whose structure differs from 3kPZS only at the 3-position in the steroid backbone, acts as an effective antagonist to neutralize or repel the attraction of 3kPZS to female sea lamprey in the behavioral tests. The other two steroid compounds including TLC and tetra-PZS, are shown to repel female sea lamprey significantly at low concentration as well. The three repellents or neutralizers are being considered in combination with other strategies for effective sea lamprey control.

The importance of the sulfate tail, 3-keto group and 7-OH group were revealed through structure-activity analysis. The compounds with more negatively charged tail groups have higher activities than less negatively charged compounds. Substitution of the 3-keto group with a 3-hydroxyl group switches the activities of compounds from agonist to antagonist. The 7-OH group attenuates the inhibition activities of compounds, which is consistent with the fact that 7-OH is predicted to have no direct favorable interactions with the binding site of SLOR1 shown in the docking results. In the 11 most active compounds, almost half of the compounds are non-steroidal hydrophobic structures with sulfate group, which block detection of 3kPZS by at least 45%. The presence of a terminal sulfate group in all the active compounds suggests that it is an important determinant for activity. The non-steroidal hydrophobic backbones can replace the steroid ring while still keeping the antagonist activities of these compounds; however, optimization of these compounds to attain activity similar to steroids would be needed.

In addition to 3kPZS, there are additional two male sea lamprey mating pheromones discovered, including DKPES and PAMS-24. DKPES has a similar behavioral effect as 3kPZS, which can guide females at close range to the nesting area. PAMS-24 serves as a male territorial pheromone, which repels mature males from nest boundaries. In the future, we will apply the antagonist discovery-screening pipeline to identify potential antagonists that mimic DKPES and PAMS-24. The identified compounds can be combined with 3kPZS antagonist to reach the highest effect of repelling or causing sea lamprey to not locate spawning grounds.

REFERENCES

REFERENCES

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat Rev Drug Discov.* 2004, *3*, 935-949.

(2) Doman, T. N.; McGovern SL; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R.; Stallings, W. C.; Connolly, D.T.; Shoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* 2002, *45*, 2213–2221.

(3) Zarzycka, B.; Seijkens, T.; Nabuurs, S. B.; Ritschel, T.; Grommes, J.; Soehnlein, O.; Schrijver, R.; van Tiel, C. M.; Hackeng, T. M.; Weber, C.; Giehler, F.; Kieser, A.; Lutgens, E.; Vriend, G.; Nicolaes, G. A. F. Discovery of Small Molecule CD40−TRAF6 Inhibitors. *J. Chem. Inf. Model.*, 2015, *55*, 294–307.

(4) Kitchell, J. F. The Scope for Mortality Caused by Sea Lamprey. *Transactions of the American Fisheries Society* 1990, *119*, 642-648.

(5) Bergstedt, R. A.; Schneider, C.P. Assessment of Sea Lamprey (Petromyzon-Marinus) Predation by Recovery of Dead Lake Trout (Salvelinus-Namaycush) from Lake-Ontario, 1982-85. *Canadian Journal of Fisheries and Aquatic Sciences* 1988, *45*, 1406-1410.

(6) Buchinger, T.J.; Wang, H.; Li, W.; Johnson, N.S. Evidence for a receiver bias underlying female preference for a male mating pheromone in sea lamprey. *Proceedings of the Royal Society B* 2013, *280*, 1771.

(7) Boogaard, M. A.; Bills, T. D.; Johnson, D. A. Acute toxicity of TFM and a TFM/niclosamide mixture to selected species of fish, including lake sturgeon (Acipenser fulvescens) and mudpuppies (Necturus maculosus), in laboratory and field exposures. *J. Great Lakes Research* 2003, *29*, 529-541.

(8) Li, W.; Scott, A.P.; Siefkes, M.J.; Yan, H.G.; Liu, Q.; Yun, S.S.; Gage, D.A. Bile acid secreted by male sea lamprey that acts as a sex pheromone. *Science* 2002, *296*, 138-141.

(9) Kain, P.; Boyle, S.M.; Tharadra, S.K.; Guda, T.; Pham, C.; Dahanukar, A.; Ray, A. Odour receptors and neurons for DEET and new insect repellents. *Nature* 2013, *502*, 507–512.

(10) Lundstrom, K. An overview on GPCRs and drug discovery: structure-based drug design

and structural biology on GPCRs. *Methods Mol Biol.* 2009, *552*, 51-66.

(11)Lin, S. H.; Civelli, O. Orphan G protein-coupled receptors: Targets for new therapeutic interventions. *Ann. Med.* 2004, *36*, 204-214.

(12) Katritch, V.; Cherezov, V.; Stevens, R. C. Diversity and modularity of G protein-coupled receptor structures. *Trends Pharmaco. Sci.* 2012, *33*, 17-27.

(13) Eswar, N., John, B.; Mirkovic, N.; Fiser, A.; Ilyin, V. A.; Pieper, U.; Stuart, A. C.; Marti-Renom, M. A.; Madhusudhan, M. S.; Yerkovich, B.; Sali, A. Tools for Comparative Protein Structure Modeling and Analysis, *Nucleic Acids Research*, 2003, *31*, 3375–3380.

(14) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of the Beta1-Adrenergic G Protein-Coupled Receptor, *Nature* 2008, *454*, 486-491.

(15) Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment, *Proteins* 1991, *9*, 56-68.

(16) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput. Aided Mol. Des.* 2002, *16*, 883-902.

(17) Lischka, F.; Kuhn, L. A.; Libants, S.; Wu, H.; Yuan, Q.; Teeter, J.; Li, W. Deorphanization of Olfactory and Vomeronasal Receptors that Respond Potently to a Vertebrate Pheromone, 2014, submitted.

(18) Irwin, J. J.; Shoichet, B.K. ZINC - A free database of commercially available compounds for virtual screening, *J. Chem. Inf. Model.* 2005, *45*, 177-182.

(19) Lipinski, C.A. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods*, 2000, *44*, 235-249.

(20) Gatica, E.A.; Cavasotto, C.N. Ligand and Decoy Sets for Docking to G Protein-Coupled Receptors, *J. Chem Inf. Model.* 2012, *52*, 1-6.

(21)Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* 2010, *50*, 572-584.

(22) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking

as Virtual Screening Tools. *J. Med. Chem.* 2007, *50*, 74-82.

(23) Chen, G.; Heim, A.; Riether, D.; Yee, D.; Milgrom, Y.; Gawinowicz, M.A.; Sames, D. Reactivity of Functional Groups on the Protein Surface: Development of Epoxide Probes for Protein Labeling, *J. Am. Chem. Soc.* 2003, *125*, 8130-8133.

(24) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* 2008, *48*, 646-658.

(25) Siefkes, M. J.; Scott, A. P.; Zielinski, B.; Yun, S. S.; Li, W. M., Male sea lampreys, Petromyzon marinus L., excrete a sex pheromone from gill epithelia. *Biology of Reproduction*, 2003, *69*, 125-132.

(26) Li, W.; Sorensen, P. W.; Gallaher, D. D. The Olfactory System of Migratory Adult Sea Lamprey (Petromyzon-Marinus) Is Specifically and Acutely Sensitive to Unique Bile-Acids Released by conspecific larvae. *J Gen Physiol.* 1995, *105*, 569-587.

(27) Li, W.; Sorensen, P. W. Highly independent olfactory receptor sites for naturally occurring bile acids in the sea lamprey, Petromyzon marinus. *Journal of Comparative Physiology a-Sensory Neural and Behavioral Physiology.* 1997, *180*, 429-438.

(28) Liu, N; Van Voorst, J; Johnston, J. B.; Kuhn, L. A. CholMine: Determinants and Prediction of Cholesterol and Cholate Binding Across Nonhomologous Protein Structures. *J. Chem. Inf. Model.* 2015, *55*, 747–759.

(29) Baell, J.B.; Holloway, G.A. New substructure filters for removal of pan assay interference compounds (PAINs) from screening libraries and for their exclusion in bioassays, *J. Med. Chem.* 2010, *53*, 2719–2740.

Chapter 5 Conclusions and future directions

In this thesis, three aspects to predict ligand binding were presented, including from aspects of protein similarity, ligand similarity and protein-ligand interaction energy.

Given only protein information, three-dimensional ligand binding motifs, particularly for cholesterol and cholate, were extracted from non-homologous proteins and CholMine, an online server was built for public usage purpose. Three-dimensional motifs generalize the characteristic of specific ligand binding across diverse protein families and show stronger prediction ability than sequence motifs. This method deciphers the determinants of specific ligand binding only from protein information across different protein families, which has advantages over the other 3-dimensional ligand binding prediction methods which need to incorporate ligand information. This method can be used to find off-target proteins that are likely to bind to cholesterol/cholate and provide guidance on the design of compounds that mimic the biological activities of cholesterol/cholate. Since this method has shown good performance on the prediction of sites for hydrophobic ligand such as cholesterol and cholate, in the future we can apply this method to prediction of binding sites of hydrophilic compounds such as compounds containing adenine and pteridine and show its generality. Preliminary results already suggest that this method can automatically detect interaction submotifs for a ligand with distinct binding motifs to different proteins. From the submotifs detected, the evolutionary relationship of the proteins binding to the same ligand could be analyzed.

Given the protein structure and a series of substituted compounds, the differences in biological activities of the compounds with same backbone but different substituents can be explained partially through protein-ligand interaction energy analysis, as we have shown in the analysis of PaPAM interactions with α-arylalanines. The residues at the binding site that contribute to the differences in biological activities were identified, and mutations at these sites were suggested to improve catalytic efficiency of the enzyme.

Finally, given only protein sequence and a native ligand structure, compounds that mimic native ligands for inhibition of a target protein can be identified using a hypothesis-driven inhibitor discovery screening pipeline. The hypothesis of the pipeline is that compounds that mimic the volumetric and electrostatic properties of the native ligand and match the functional group side chains of the native ligands that are important for the specific binding can compete with the native ligand for binding. In the future, this pipeline can be used for inhibitor discovery based on additional known pheromone ligand structures such as DKPES and PAMS-24 and facilitate the inhibitor discovery process.