# IDENTIFICATION AND SCORING OF PARTIAL COVALENT INTERACTIONS IN PROTEINS AND PROTEIN LIGAND COMPLEXES

By

Steven Anthony Combs

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

December, 2013

Nashville, Tennessee

Approved:

Professor Jens Meiler

Professor Brian Bachmann

Professor Richard Armstrong

Professor Martin Egli

# ACKNOWLEDGEMENTS

The work presented here was made possible by the Training Grant award 3 T32 GM 8320-21 S1 from the National Institute of Health for the Center of Structural Biology Training Program. Computational resources were provided by the Advanced Computing Center for Research and Education and the Center for Structural Biology at Vanderbilt University.

I would first and foremost like to thank my advisor, Dr. Jens Meiler, who provided me with guidance throughout my graduate degree. Through both my personal and professional struggles, Jens has provided input that has molded me into the scientist that I am today. His unending support for me to investigate problems through my own studies has led me to be an independent and strong scientist. In addition to Jens support, I would like to thank Dr. Brian Bachmann and Dr. Martin Egli for being flexible and showing interest in my research and training.

The Vanderbilt Chemistry department has provided me with unparalleled support through its staff and faculty. In particular, I would like to thank Leigh Clayton for her help in setting up the Rosetta Workshop and quickly reimbursing me of any expenses because I am a poor graduate student. I'd also like to thank Sabuj, Roy, Jarrod and Jonathan from the Center of Structural Biology for their computational expertise.

Current and former Meiler Lab members have been invaluable as friends and colleagues. I owe a special thanks to Dr. Ralf Mueller for sound advice and guiding me through one of the hardest times in my personal life. Without his help, I would not have been able to continue with graduate school and not be in my current state of self-worth and happiness. I would also like to thank Dr. Kristian Kaufmann for his guidance during my rotation in the Meiler lab and his continuing advice concerning my career. Without the help from Sandeep Kothiwale, I would weigh fifty pounds more than I currently do, be depressed, and not have a post-doctoral fellowship. Long conversations with him provided me with inspiration and motivation to complete my graduate work. I would like to thank Elizabeth Dong, and Sam and Stephanie Deluca for their support throughout my entire graduate career and for providing me a place to sleep when I had none.

Evan Burner has been a life-long friend who has been by my side regardless of where our lives have taken us. Whether it is three in the afternoon or three in the morning, I know that I can count on him to answer the phone and provide endless entertainment and advice. Although he is not family, he is the closest thing to a brother that I have. In addition to Evan, I would like to thank Veronica Combs who, regardless of our relationship, has always believed that I can achieve anything that I set my mind to. Finally, my parents, Randy and Kristine Combs and sister, Kathryn Combs, have always been there for me. Without their advice and support, I would not have made it this far into my life.

# SUMMARY

The overall focus of this dissertation was to create a robust scoring function within the Rosetta software suite that accurately described and scored partial covalent interactions (PCIs) such as hydrogen bonds, salt bridges, cation-π, and π-π interactions found in macromolecular structures and complexes.

Chapter I introduces the importance of the Rosetta software suite and the impact that it has had on the field of macromolecular modeling. The introduction describes the most commonly used algorithms in Rosetta, comparative modeling, protein, ligand interface, and enzyme design, rotamer recovery analysis, loop building, all atom refinement, and finally the features reporter. In addition to commonly used applications, analysis methods for models output by Rosetta protocols are discussed. Finally, the Rosetta score function is described in detail and compared to molecular mechanics force fields. Portions of Chapter I came from a protocols article titled “Small-molecule ligand docking into comparative models with Rosetta” by Steven Combs, Samuel Deluca, Stephanie Deluca, Gordon Lemmon, David Nannemann, Elizabeth Nguyen, Jordan Willis, Jonathon Sheehan, and Jens Meiler and a journal article titled “Partial Covalent Interactions in Rosetta” by Steven Combs and Jens Meiler. The author of this dissertation contributed to all portions of the articles through writing, performing experiments, or editing.

Chapter II introduces advances in docking small-molecule antidepressants into comparative models of the human serotonin transporter (hSERT) and the *drosophila* serotonin transporter (dSERT). Through this work, the apparent deficiencies in the Rosetta score function are noted. The work is largely based on the publication entitled “E444 Interaction Required for High-Affinity S-Citalopram Binding in the Human Serotonin Transporter” by Steven Combs, Kristian Kaufmann, Julie R. Field, Randy D. Blakely, and Jens Meiler. While the work on this project was highly collaborative, the computational experiments, data analysis, and results were work of the author of this dissertation. The chapter concludes with a set of experiments used to determine suitable ligands and alkyl spacer lengths to conjugate quantum dots for tracking of the hSERT within a cell membrane. Methods and procedures described in the beginning of the chapter are used to guide the study; however, because of deficiencies in the Rosetta score function, a separate computational approach is used. The work is based on a manuscript in preparation: “Guided Design of Conjugated Ligand Quantum Dots for the Human Serotonin Transporter” by Steven Combs, Emily Jones, Zack Glazer, Ian Tomlinson, Sandra Rosenthal and Jens Meiler. Significant portions of the computational experiments, analysis of results, and writing were done by the author of this dissertation.

As a direct result of the deficiencies in the Rosetta score function described in chapter II, Chapter III focuses on the computational methods used to improve the Rosetta score function. First, a brief summary of interactions intended to model are summarized followed by a discussion on errors in the original Rosetta score function.

Approaches used to model partial covalent interactions, assumptions made in scoring interactions, and computational procedures to implement a new score function are discussed. The chapter ends with a comparison of the versions for the score function. The work presented in this chapter is over an unpublished manuscript by Steven Combs and Jens Meier entitled “Partial Covalent Interactions in Rosetta”. All work done was by the author of this dissertation.

Chapter IV implements the new score function into common Rosetta applications that were discussed in Chapter I of this dissertation. Improvements and deficiencies in all applications are detailed along with the protocol used to run the applications. A summary of the results are presented at the end of the chapter which shows larges improvements in scoring and identification of partial covalent interactions in Rosetta applications.

Chapter V provides the major conclusions of this dissertation and how they relate to the current field of research. Additionally, future directions on improving scoring and better programming practices are included.

## TABLE OF CONTENTS

Page

# LIST OF FIGURES

## LIST OF EQUATIONS

## LIST OF TABLES

# CHAPTER I

# INTRODUCTION

Part of the work presented in this chapter was published in (Combs, DeLuca et al. 2013)

## Molecular Modeling in Rosetta

Molecular modeling is a broad term used to describe computational approaches to determine properties and behaviors of molecules, macromolecules, and large macromolecule complexes. While substantial progress has been made in X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, the availability of high-resolution structures is limited due to the frequent inability to crystallize or obtain sufficient NMR restraints for large or flexible proteins. Computational methods can be used to both predict unknown protein structures and model ligand interactions when experimental data is unavailable in a process known as small-molecule docking into comparative models.

Small-molecule docking into comparative models allows for structure-based drug design and hypothesis generation for protein/ligand systems for which there is no high-resolution structure. Often, a homologous structure has been structurally characterized at sufficient resolution for ligand docking into a constructed comparative model. Many software packages exist for the specific task of comparative modeling and ligand docking. The Rosetta software suite includes algorithms for both of these tasks and is developed for computational modeling and analysis of protein structures; further, it is free for non-commercial users. It has enabled notable scientific advances in computational biology, including *de novo* protein design, enzyme design, ligand docking, and structure prediction of biological macromolecules and macromolecular complexes(Kuhlman, Dantas et al. 2003; Rohl, Strauss et al. 2004; Misura, Chivian et al. 2006; Das and Baker 2008; Davis and Baker 2009; Davis, Raha et al. 2009; Siegel, Zanghellini et al. 2010). The broad spectrum of applications available through Rosetta allows for multiple computational problems to be addressed in one software framework. A series of algorithms important to this thesis are summarized below.

### *Comparative modeling with Rosetta*

One of the most commonly used applications of Rosetta is protein structure prediction via *de novo* folding and comparative modeling(Rohl, Strauss et al. 2004; Kaufmann, Lemmon et al. 2010). When only the primary sequence of a protein is known, *de novo* folding can be used to predict the protein's tertiary structure. However, to date, Rosetta has been shown to successfully fold only small, soluble proteins (fewer than 150 amino acids) and performs best if the proteins are mainly composed of secondary structural elements (alpha-helices and beta-strands)(Meiler and Baker 2003). Structures of helical membrane proteins between 51-145 residues were predicted to within 4 Å of the native structure,(Yarov-Yarovoy, Schonbrun et al. 2006) but only very small proteins (up to 80 residues) have been predicted to atomic-detail accuracy(Bradley, Malmström et al. 2005; Bradley, Misura et al. 2005; Das, Qian et al. 2007). Accurate prediction of larger and/or more complex proteins can be achieved with the addition of experimental data, such as NMR chemical shifts and distance data(Rohl 2005; Lange and Baker 2012; Lange, Rossi et al. 2012). Given these limitations, whenever an experimental structure of a related protein is available, comparative modeling is preferred to *de novo* folding.

Comparative modeling refers to the elucidation of the tertiary fold of a protein, guided by a known structure of another, often homologous, protein. The unknown structure is commonly called the “target,” while the protein of known structure, upon which the primary sequence of the target is threaded, is termed the “template.” The known template structure reduces the conformational search space by providing a protein backbone scaffold; areas where the template and target sequences diverge significantly are typically remodeled and refined via the loop building application. The application is known as “loop building” because it is most commonly applied to flexible loop regions between secondary structure elements. However, a “loop” is defined here as any area where the backbone needs to be rebuilt *de novo*, which most often occurs in flexible regions but can also include secondary structural elements in cases of insertions/deletions or low sequence identity. Comparative models have played a major role in aiding experimental design and the interpretation of experimental results. They can be employed to help predict structure-function relationships(Kaufmann, Dawson et al. 2009), predict binding pockets for ligands during structure-based drug design(Lees-Miller, Subbotina et al. 2009), and aid in the determination of target residues for site-directed mutagenesis(Keeble, Joachimiak et al. 2008; Fortenberry, Bowman et al. 2011).

### *Ligand docking with RosettaLigand*

Small-molecule ligand docking applications attempt to predict the protein/small-molecule binding free-energy, as well as critical binding interactions(Perola, Walters et al. 2004; Mooers and Matthews 2006; Mobley, Graves et al. 2007). These predictions can provide structural information of a ligand binding site(Davis, Raha et al. 2009), filter high-throughput screening libraries for likely hits(Ballester, Westwood et al. 2010; Carlsson, Coleman et al. 2011), or guide *de novo* drug design(Schneider and Fechner 2005; Schneider, Hartenfeller et al. 2009).

RosettaLigand requires input structures of a receptor (protein) and a ligand (small molecule)(Davis and Baker 2009; Davis, Raha et al. 2009). Because it does not perform binding pocket detection, the user must have prior knowledge of the location of the binding site. Other programs, such as SURFNET(Laskowski 1995), LIGSITE(Huang and Schroeder 2006), and PocketDepth(Kalidas and Chandra 2008), can be used to identify the ligand binding site before using RosettaLigand for small-molecule docking. Ligand and receptor side chain conformations are explored through Monte Carlo sampling of rotamers(Dunbrack and Karplus 1993). Predicted protein-ligand interactions are deemed favorable and are accepted if they improve the Rosetta energy score (see subsection Rosetta Score Function)(Davis and Baker 2009). Backbone flexibility of the protein is modeled via a gradient-based minimization of phi and psi torsion angles(Li and Scheraga 1987). Performing ligand docking on an ensemble of ligand conformations and protein backbones can be used to increase the conformational space sampled if the protein-ligand interaction does not fit the simple lock-and-key paradigm(Siegel, Zanghellini et al. 2010).

The accuracy of RosettaLigand was assessed by Davis, et al. via both retrospective and prospective benchmark studies(Davis and Baker 2009). In 54 of 85 cases (64%), RosettaLigand’s top scoring model was within 2.0 Å root mean square deviation (RMSD) from the experimentally determined structure. These results were achieved by including backbone and side chain flexibility, as well as ligand flexibility through conformer selection and torsion angle adjustments.

Ligand docking algorithms can be categorized based on their scoring function and search methodology. RosettaLigand uses a knowledge-based scoring function derived from statistical analysis of the Protein Data Bank (PDB)(Simons, Kooperberg et al. 1997). The conformational search of the binding site is accomplished using a Metropolis Monte Carlo algorithm (Metropolis, Rosenbluth et al. 1953; Kuhlman, Dantas et al. 2003; Rohl, Strauss et al. 2004; Misura, Chivian et al. 2006; Das and Baker 2008; Davis and Baker 2009; Davis, Raha et al. 2009; Siegel, Zanghellini et al. 2010). Other search strategies include geometric hashing (FlexX)(Rarey, Kramer et al. 1996), genetic algorithms (GOLD)(Verdonk, Cole et al. 2003), and systematic sampling (Glide)(Friesner, Banks et al. 2004). Different scoring functions include physics-based force fields (Dock)(Ewing, Makino et al. 2001), chemical descriptor models (FlexX(Rarey, Kramer et al. 1996)), and knowledge-based potentials (RosettaLigand(Meiler and Baker 2006), DrugScore(Gohlke, Hendlich et al. 2000)).

A 2009 study compared the performance of the RosettaLigand docking method to nine other commonly used ligand docking programs (Dock, Dockit, FlexX, Flo, Fred, Glide, GOLD, LigandFit, MOE, and MVP). Ligand docking algorithm performance was compared using a benchmark set of 136 ligands and eight target receptors provided by Glaxo-Smith-Kline. This study demonstrated that RosettaLigand performance was comparable or better than the other ligand docking algorithms considered. The study used crystallographic protein structures as input rather than comparative models. Kaufmann, *et. al*. demonstrated the predictive power of Rosetta ligand docking into Rosetta-built comparative models(Kaufmann, Dawson et al. 2009). In another study, RosettaLigand and AutoDock were used to dock twenty protein-ligand complexes(Davis and Baker 2009). In ten cases, RosettaLigand's flexible backbone docking protocol found top-scoring models under 2.0 Å RMSD. In contrast, AutoDock identified only four such structures. However, the authors note that RosettaLigand consumed significantly more computational resources (40-80 CPU hours per input) than AutoDock (5-22 CPU Fhours per input).

### *Protein, Ligand Interface, and Enzyme Design in Rosetta*

Protein design maximizes the total free energy of a tertiary structure through optimization of the primary sequence(Kuhlman and Baker 2000; Kuhlman and Baker 2004). On a fixed backbone, amino acid side chain identities and conformations are stochastically swapped and scored with the Rosetta all-atom scoring function. The sequence with the lowest free energy is chosen as the optimal sequence for the given backbone. Because design is done on a rigid backbone from experimental structures, small local frustrations/clashes exist that – when designing the protein – are relived through replacing a large amino acid with a smaller one thereby producing an artificially low energy(Kuhlman, Dantas et al. 2003). This is in particular true for the larger proteins. In order to relieve small clashes, an all-atom refinement of experimental structures is suggested.

Similar to protein design, ligand interface design attempts to optimize the local sequence around a specific ligand in order to maximize affinity to the small molecule. The small-molecule is allowed to sample conformational space through small perturbations around the binding pocket in a series of small translational (0.1 Å – 0.5 Å) and rotational (5-10°) steps while the side chain conformations are allowed to accommodate the ligand through side chain repacking. At the end of the ligand movement, side chain identities are optimized to create binding contacts with the ligand using the algorithm described in the fixed backbone application.

Theozymes are theoretical enzymes that have an ideal arrangement of the transition state structure surrounded by catalytic functional groups(Tantillo, Chen et al. 1998). In general, a theozyme is created computationally via quantum mechanical studies. In this process, a transition state is modeled and geometrically optimized. Once the transition state is geometrically optimized, catalytic functional groups are added to stabilize the transition state. Instead of using a full atom protein, the amino acid side chain functional group is used: ammonium for lysine, acetamide for aspargine, etc. Placement of functional groups is done through approximation of ideal distances/angles. Then, quantum mechanical studies are used to find the optimal geometric constraints. Once a set of geometrical constraints are derived, they are input into Rosetta Matcher(Zanghellini, Jiang et al. 2006), which searches a set of proteins that can accommodate the ligand and geometric constraints for the side chains. Once a set of proteins are identified, the algorithm Rosetta EnzDes is used to optimize binding ligand interface contacts through the same algorithm used for interface design.

### *Rotamer Recovery*

Conformational sampling for proteins side chains is a combinatorial problem that produces a large search space. Experimentally determined structures contain side chains in an energetically favorable conformation. Unlike protein design, rotamer recovery attempts to recover the correct conformation of a protein side chain in context of all other side chains in a protein. A stringent test for the score function is to see if the energetic minimum for side chain conformations can be recovered through sampling multiple rotamers (conformations) for each side chain(Petrella, Lazaridis et al. 1998; Liang and Grishin 2002). Side chain rotamer recovery is measured by systematically swapping out each amino acid rotamer with rotamers from the dunbrack library(Shapovalov and Dunbrack 2011) while keeping the conformations of all other side chains fixed. After a rotamer is picked, the side chain is allowed to minimize. The lowest scoring rotamer/minimization is compared by chi angles to the original side chain conformation. If chi angles are equal, then the residue is considered recovered. Residues with a neighbor count of 16 or less were considered surface residues while residues with more than 16 neighbors were considered to be in the core of the protein.

Neighbor counts are measured by the amount of residues with a CB (CA for GLY) distance within 10Å of the residue being repacked.

### *Loop Building in Rosetta*

Rosetta includes two loop-building algorithms. Cyclic coordinate descent (CCD), inspired by inverse kinematic applications in robotics, adjusts residue dihedral angles to minimize the sum of the squared distances between three backbone atoms of the moving N-terminal anchor and the three backbone atoms of the fixed C-terminal anchor(Canutescu and Dunbrack 2003). The advantages of CCD are its speed and its ability to close a loop over 99% of the time. Conversely, kinematic loop closure (KIC) analytically determines all mechanically accessible conformations for torsion angles of a peptide chain using polynomial resultants(Coutsias, Seok et al. 2004; Mandell, Coutsias et al. 2009). While KIC has been shown to recover loops from experimentally determined structures more accurately, it relies heavily on the location of the N- and C-terminal anchors and may not be an ideal choice for comparative modeling.

Rosetta loop building by CCD uses fragment libraries for generating loop coordinates for missing density in the threaded model. The fragment file is comprised of the target sequence divided into 3- and 9-amino acid overlapping sequence windows. There are 200 peptide fragments for each sequence window. After dividing the target primary sequence into 3- and 9-amino acid sequence windows, both Robetta and the fragment picker(Gront, Kulp et al. 2011) application query a structural database of non-redundant proteins(Wang and Dunbrack 2003) for each peptide sequence and store the corresponding Cartesian coordinates and secondary structure information in fragment files. For more detailed background and information on this application, see Gront, *et al*.(Gront, Kulp et al. 2011) or go to www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/dc/d10/app_fragment_p icker.html. Fragments can also be generated using NMR data using RosettaNMR(Rohl 2005). For details on the procedure, please visit spin.niddk.nih.gov/bax/software/CSROSETTA/.

Loop building takes place in two stages. In the first stage, a fast, low-resolution remodeling step with CCD consisting of broad sampling of backbone conformations is performed. In the second stage, the model is represented in all-atom detail and evaluated by Rosetta's all-atom scoring function. It has been suggested by Kaufmann *e*t. *al.(Kaufmann and Meiler 2012)* and others that ligands in the binding site of the template structure be carried into the comparative modeling process. Although not done here, it is anticipated that using such an approach would pre-arrange and maintain the pocket shape for small-molecule binding and result in higher quality homology models.

### *All Atom Refinement in Rosetta*

The protocol used for structural refinement in Rosetta, visually described in **Figure 1**, is often referred to as “relax.” The goal of the relax protocol is to explore the local conformational space and to energetically minimize the protein. During this process, local interactions are improved by iterative side chain repacking, in which new side chain conformations, or “rotamers,” are selected from the Dunbrack library(Leaver-Fay, Kuhlman et al. 2005), and gradient-based minimization of the entire model, where the energy of the model is minimized as a function of the score. These small structural changes are evaluated according to the all-atom scoring function and are sampled in a Metropolis Monte Carlo(Metropolis, Rosenbluth et al. 1953) method. The relax protocol has been shown to dramatically lower the overall energy of the Rosetta model and is essential to achieving atomic detail accuracy(Rohl, Strauss et al. 2004; Bradley, Misura et al. 2005).

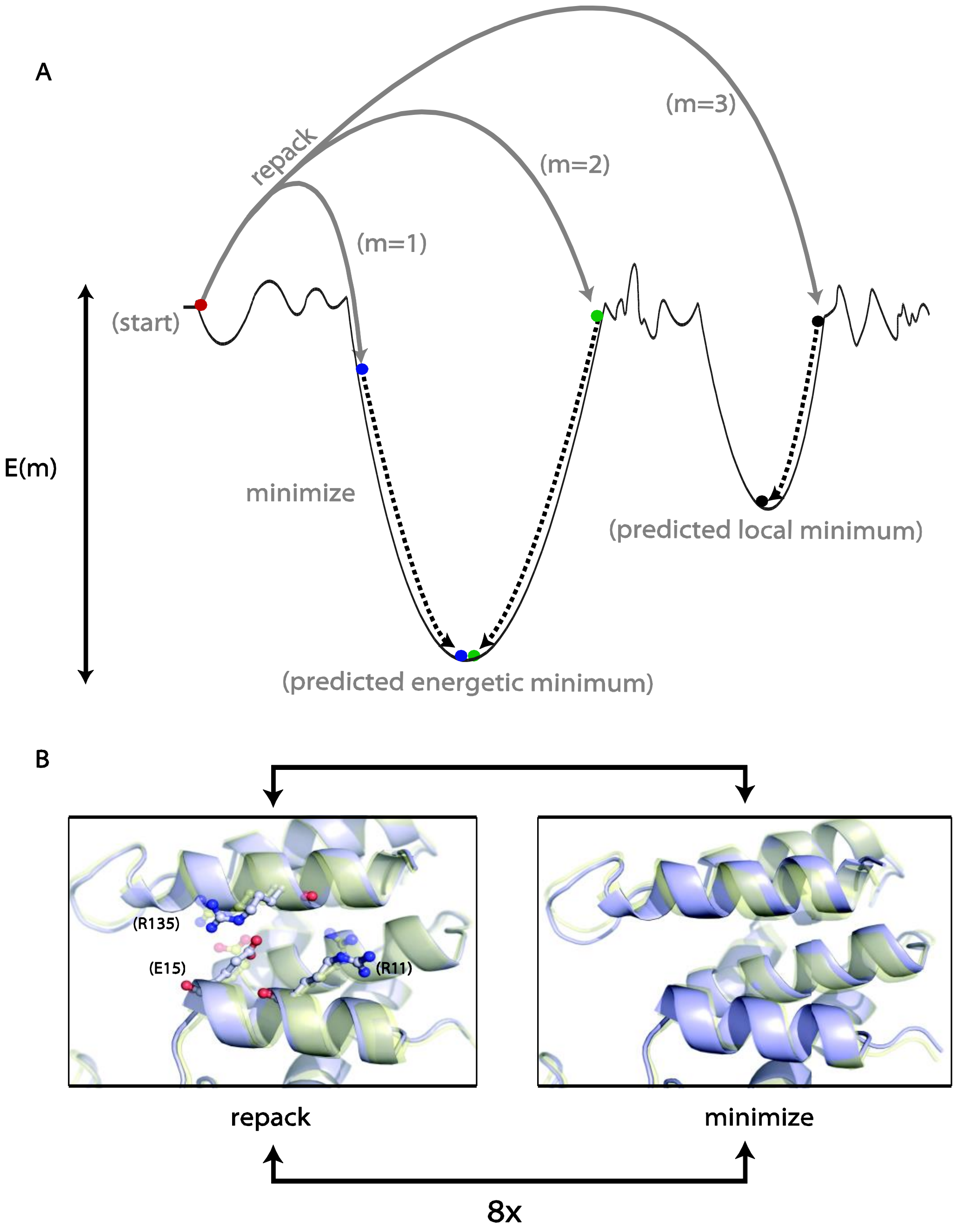


**Figure 1: Overview of Rosetta energetic minimization and all-atom refinement via relax protocol.**

(**a**) Simplified energy landscape of a protein structure. The relax protocol combines small backbone perturbations with side-chain repacking. The coupling of Monte Carlo sampling with the Metropolis selection criterion allows for sampling of diverse conformations on the energy landscape. The final step is a gradient-based minimization of all torsion angles to move the model into the closest local energy minimum. (**b**) Comparison of structural perturbations introduced by the repack and minimization steps. During repacking, the backbone of the input model is fixed, whereas side-chain conformations from the rotamer library are sampled. Comparison of the initial (transparent yellow) and final (light blue) models reveals conservation of the R135 rotamer but changes to the R11 and E15 rotamers. Minimization affects all angles and changes the backbone conformation.

## *Rosetta Features Framework*

RosettaFeatures(Leaver-Fay, O'Meara et al. 2013) measures a series of properties from a dataset of protein structures and stores the properties in a relational database (SQL, MYSQL, POSTGRES). Data stored in the database can be extracted and summarized through a set of Analyses scripts written in the functional language R. Specifically, geometric parameters used to define specific score function terms such as hydrogen bond distances and angles (Figure 2A) and orbital distances and angles (Figure 2C) can be analyzed and summarized in a graphical format. A powerful tool of the analysis scripts and the RosettaFeatures framework is the direct comparison of crystal structure features compared to models generated from a Rosetta protocol. In addition to storing score function data, metadata used in generating the models which allows for reproducibility of the protocol, generated models, and protein metrics (SASA, packing density, number of buried unsatisfied hydrogen bond acceptors, etc) are stored.

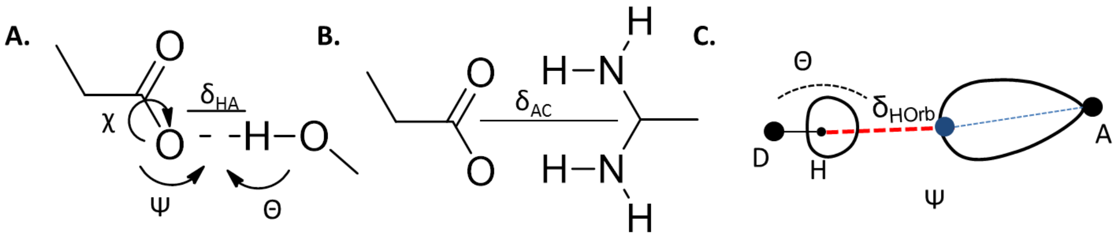


**Figure 2: Geometric parameters for partial covalent interactions in the Rosetta score function.** Schematic representation of geometric definitions for derivation of KBP for (a) hydrogen bonds, (b) pair potential, and (c) orbitals. A) δHA distance between the acceptor and hydrogen atom, χ torsion angle between the base acceptor-acceptor-hydorgen-donor atom, Ψ angle between the base acceptor-acceptor-hydrogen, and Θ angle between the acceptor – hydrogen –donor. B) δAC distance between the action center for two polar residues. C) δHOrb, distance between the orbital and hydrogen atom, Ψ angle between the acceptor – orbital – hydrogen, and Θ the angle between the donor – hydrogen – orbital.

### *Analysis of Protein Structures in Rosetta*

After each Rosetta protocol, several models are produced; however, score alone does not always predict biochemically relevant models. It is important, therefor, to analyze models based on several metrics. In the following section, model evaluation methods are described.

Sequence recovery is calculated by (n_recovered)/(n_designed), where n_recovered were the number of residues that where designed and matched native residues for the given position and n_designed is the total number of residues designed. The average is taken for all proteins in a design benchmark dataset. Surface and core residue designation is determined by the number of residues with c-beta (c-alpha for gly) atoms within 10.0 Å of the designed residue c-beta where counts of residues fewer than 16 residues are considered to be on the surface and counts larger than 16 to be in the core.

Position Specific Scoring Matrix (PSSM) recovery measures the fraction of amino acids that is converted to an identity that has been seen in evolution in this position. The measure seeks to circumvent one limitation of the sequence recovery measure: often multiple amino acids will be acceptable in a certain position; therefore, failure to recover the native amino acid is not indicative of a poor energy function. PSSM recovery addresses this limitation in part as it accepts all amino acid sampled by evolution in a given position as acceptable. Computing PSSM recovery requires i) creating a PSSM for each individual protein using PSI-BLAST and ii) calculating percent recovery counting all amino acid types as correctly recovered that have a positive value in the log-odds PSSM (n_pssm_recovered/n_designed).

An aspect of highly stable proteins, proteins resistant to heat denaturation and enzymatic degradation is how well the protein core excludes water. An indirect measurement of this feature is packing density, a measure of how well a protein is packed(Sheffler and Baker 2009). To determine the packing density, empty spaces (cavities) within the core of a protein are determined through determining the surface of the protein. All cavities not directly open to the surface of the protein are considered cavities. Contact surface area around each cavity ball is then calculated with a probe radii between 0.1Å and 3.0Å with 0.1 Å steps. A support vector machine (SVM) is used to create a summary statistic of the contact surface areas for the cavities and is reported in an interval between 0 and 1 which can be interpreted roughly as probabilities of how well the protein is packed.

Predicted structures generated by comparative modeling, *de novo* folding, and ligand docking are often clustered to help identify structurally similar models. Clustering is performed with the assumption that the deepest energy well, and hence the global energy minimum, will also be the widest(Shortle, Simons et al. 1998) Figure 3.

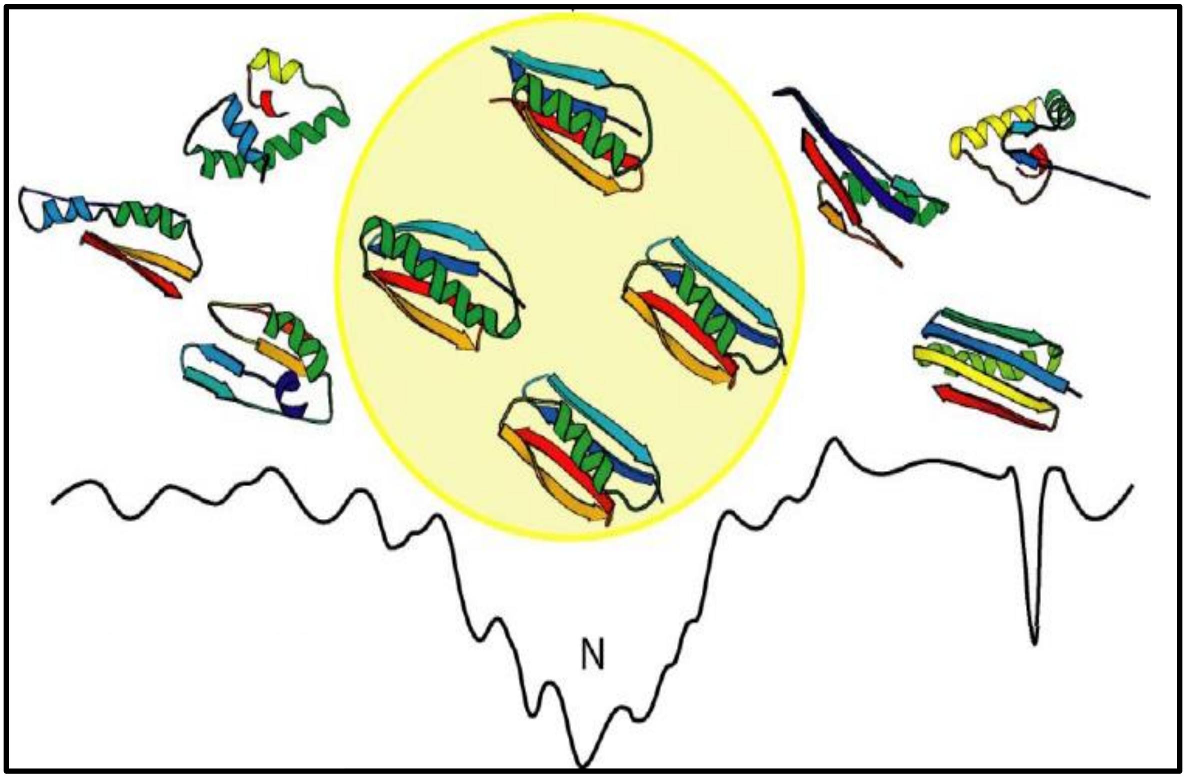


**Figure 3: Clustering within Rosetta identifies energetic minimum.**
Clustering models with similar topology through pairwise root mean square deviation (RMSD) distances identifies set of clusters that correspond to energy minima. Because the Rosetta energy function is coarse, the width of the energy well, or cluster centers with large number of models, correspond to the depth of an energy well. In both ligand docking and *ab initio* folding, the largest cluster center is accepted as a representation of the experimental model.

As a result, it is expected that the largest clusters will potentially contain the predicted model that is closest to the native structure. Rosetta includes a tool for clustering protein models. The cluster application avoids the memory requirements associated with computing a complete distance matrix for large numbers of models. The Rosetta clustering method starts by computing a distance matrix for the first 400 input models. Each model in the distance matrix is assigned to the cluster to which it is nearest (typically in terms of RMSD). If the model is not within a specified radius of any cluster, it is assigned to a new cluster.

Because the Rosetta clustering application outputs most of its statistical information in its log file, a script can be used on the data to produce a clear summary of the results. Given a set of PDB or Rosetta silent files and a Rosetta options file, `clustering.py` will produce a set of clustered PDBs, a histogram file showing the distribution of pairwise RMSDs between models, and a summary file showing which models are in which clusters. The Rosetta options file can contain a number of options that control the behavior of the cluster application.

## Force Fields in Molecular Modeling

Free energy calculations are used for model discrimination in molecular modeling. During model generation, free energy calculations are used to drive the simulation towards an energy minimum or equilibrium; however, final free energy calculations on completed simulations are used to discriminate and rank native-like models. In context of molecular modeling, several methods exist to calculate the energy of the system: molecular mechanics force fields and knowledge-based force fields.

### *Molecular Mechanics Force Fields*

The underlying functional forms for free energy calculation in molecular mechanics force fields is $E_p = E_{covalent} + E_{noncovalent}$ where $E_{covalent}$ is the energy associated between bonded atoms and $E_{noncovalent}$ is the energy between non-bonded atoms. $E_{covalent}$ attempts to capture vibration of bonds, bending of valence angles, and rotation of dihedrals. Therefore, $E_{covalent}$ is further separated through:

**Equation 1:**

$$\boldsymbol{E_{covalent}} = \sum_{\boldsymbol{bonds}} \boldsymbol{K_b(b - b_0)^2} + \sum_{\boldsymbol{angles}} \boldsymbol{K_\theta(\theta - \theta_0)^2} + \sum_{\boldsymbol{dihedrals}} \boldsymbol{K_\chi[1 + \cos(n\chi - \sigma)]}$$

The first term in Equation 1 sums over all bonded pairs of atoms where *b* is the bond length (interatom distance), $K_b$ and $b_0$ are parameters that describe the stiffness and the equilibrium length of the bond, respectively. Bending of angles is described through the connections of triplet atoms forming an angle, *A*, *B*, and *C* where *A* is bonded to *B* and *B* is bonded to *C*. $\theta$ is the angle formed by the two bond vectors and similar to the bonds, $K_\theta$ and $\theta_0$ describe the stiffness and equilibrium geometry of the angle. Finally, the third term describes the energy associated with dihedral angles for *A*, *B*, *C,* and *D* where *A* is bonded to *B* and *B* is bonded to *C* and C is bonded to *D*. $\chi$ is the value of the dihedral, $K_\chi$ is the energetic parameter that determines barrier heights, $n$ is the periodicity, and $\sigma$ is the phase. Because rotation is periodic in nature, the *cos* of the terms are taken.

$E_{noncovalent}$ is an energetic evaluation for all non-bonded atoms with a distance separation of at least three bonds. The expanded form is:

**Equation 2:**

$$E_{noncovalent} = \sum_{\substack{nonbonded \\ pairs}} (\varepsilon_{ij}[(\frac{R_{min,ij}}{r_{ij}})^{12} - 2 * (\frac{R_{min,ij}}{r_{ij}})^{6}] + \frac{q_i q_j}{r_{ij}})$$

The set of terms encompassed in the brackets is the formulation of the Lennard-Jones equation and models attractive dispersion and repulsive Pauli-exclusion interactions over atoms *i* and *j* with a prefactor $\varepsilon_{ij}$ which is a parameter based on the type of interacting atoms. $R_{min,ij}$ describes the minimum of the LJ energy and is based on the type of interacting atoms and $r_{ij}$ is the interatomic distance. The second portion of the equation is the formulation of Coulomb's Law where $r_{ij}$is the interatomic distance and $q_i q_j$ is the charges on the two interacting atoms.

### *Knowledge Based Force Fields*

In contrast to molecular mechanics, KBPs are based upon the assumption that frequently observed molecular geometries correspond to low energy states(Sippl 1995). They require availability of a large databank of experimentally determined structures – a prerequisite fulfilled for proteins through the protein databank (PDB)(Berman, Westbrook et al. 2000) and small molecules through the Cambridge Structural Database(Allen 2002). A set of measurable geometric variables can be extracted from the PDB and compiled into a statistical distribution. According to Boltzmann's principle, energies and probability densities are related quantities and can be used to rapidly assess protein models. The general definition for the energy of a KBP is expressed via $E(r) = -kTln[(\int(r))]$ where *r* is a measured quantity between two atoms, *E(r)* is the energy associated from the measured quantity, *k* is Boltzmann's constant, *T* is the absolute temperature, and $\int(r)$ is the probability density.

### *The Rosetta Energy Function*

Specific to this dissertation is the Rosetta score function. The energy, or scoring, function in Rosetta is derived empirically through analysis of observed geometries of a subset of proteins in the PDB. The measurements include, but are not limited to: radius of gyration, packing density, distance/angle between hydrogen bonds, and distance between two polar atoms. The measurements are converted into an energy function through Bayesian statistics(Simons, Kooperberg et al. 1997).

The scoring function in Rosetta can be separated into two main categories: centroid-based scoring and all-atom scoring. The former is used for *de novo* folding and initial rounds of loop building(Simons, Kooperberg et al. 1997; Simons, Ruczinski et al. 1999; Rohl, Strauss et al. 2004). The side chains are represented as "super-atoms," or "centroids," which limit the degrees of freedom to be sampled while preserving some of the chemical and physical properties of the side chain. This centroid-based scoring function is important for *de novo* folding and not all-atom modeling.

The all-atom scoring function represents side chains in atomic detail. Like the centroid-based scoring function, the all-atom scoring function is comprised of weighted individual terms that are summed to create a total energy for a protein. Most of the scoring terms are derived from knowledge-based potentials. The scoring function contains Newtonian physics-based terms, including a 6-12 Lennard-Jones potential and an implicit solvation potential(Lazaridis and Karplus 1999). The 6-12 Lennard-Jones potential is split into two terms, an attractive term (`fa_atr`) and a repulsive term (`fa_rep`) for all van der Waals interactions(Neria, Fischer et al. 1996; Kuhlman and Baker 2000). The solvation potential (`fa_sol`) models water implicitly and penalizes the burial of polar atoms(Lazaridis and Karplus 1999). Inter-atomic electrostatic interactions are captured through a pair potential (`fa_pair`)(Simons, Ruczinski et al.

1999), and an orientation-dependent hydrogen bond potential for long range and short range hydrogen bonding (hbond_sc, hbond_lr_bb, hbond_sr_bb, hbond_bb_sc, respectively)(Gordon, Marshall et al. 1999; Wedemeyer and Baker 2003). In addition to the electrostatic terms, the Rosetta all-atom scoring function contains terms that dictate frequently observed side chain conformations in the PDB through the Dunbrack rotamer library (fa_dun)(Dunbrack and Karplus 1993; Dunbrack and Cohen 1997), preference for a specific amino acid given a pair of phi/psi angles (p_aa_pp), and preference for the phi/psi angles in a Ramachandran plot (rama)(RAMACHANDRAN, RAMAKRISHNAN et al. 1963; Wedemeyer and Baker 2003; Rohl, Strauss et al. 2004). Finally, specific for design, the reference term (ref) attempts to correlate the free energy of an amino acid in the unfolded state. The total energy of a protein or residue is the summation of all weighted terms:

**Equation 3:**

$$E = W_{\text{atr}}E_{\text{atr}} + W_{\text{rep}}E_{\text{rep}} + W_{\text{intra\_rep}}E_{\text{intra\_rep}} + W_{\text{sol}}E_{\text{sol}} + W_{\text{hbond\_sc}}E_{\text{hbond\_sc}} + W_{\text{hbond\_sr\_bb}}E_{\text{hbond\_sr\_bb}} + W_{\text{hbond\_lr\_bb}}E_{\text{hbond\_lr\_bb}} + W_{\text{hbond\_bb\_sc}}E_{\text{hbond\_bb\_sc}} + W_{\text{dun}}E_{\text{dun}} + W_{\text{p\_aa\_pp}}E_{\text{p\_aa\_pp}} + W_{\text{pair}}E_{\text{pair}} + W_{\text{ref}}E_{\text{ref}}$$

# CHAPTER II

# SMALL MOLECULE DOCKING INTO THE HUMAN SEROTONIN TRANSPORTER

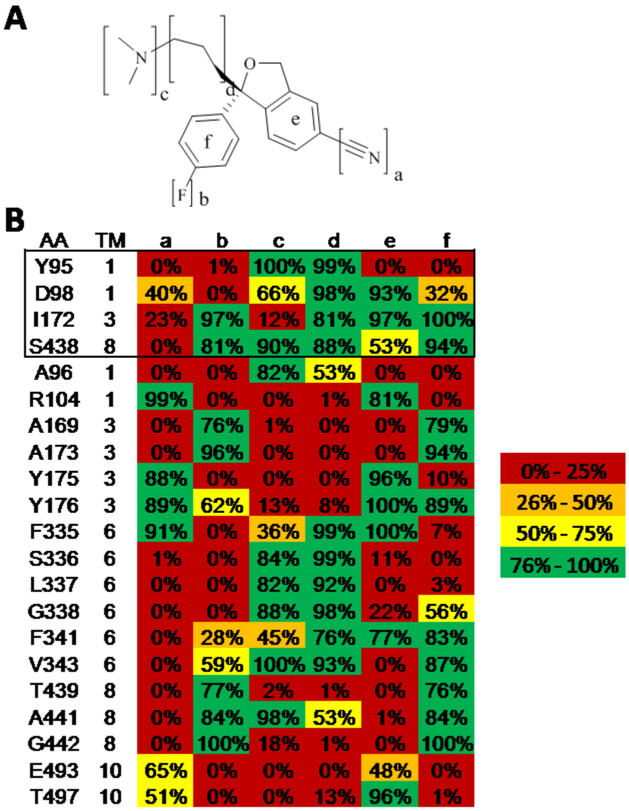
Depression affects close to 20 million (6.5% of the total population) Americans each year(Kessler, Chiu et al. 2005) and an estimated 4.5% of the world population(Organization 2009). The monoamine theory postulates that disruption or malfunction of CNS serotonergic and noradrenergic systems drives risk and/or symptoms of depression(Hirschfeld 2000). The human 5-HT transporter (hSERT, *SLC6A4*) is an integral membrane protein localized to serotonergic terminals. SERT is responsible for the uptake of 5-HT, Na+ ions, and Cl- ions across the presynaptic cell membrane, thereby limiting 5-HT actions in space and time(Barker, Burris et al. 1991). A variety of brain disorders, including depression, autism, attention deficit hyperactivity disorder (ADHD), obsessive-compulsive disorder (OCD), and addiction(Hirschfeld 2000; Prasad, Steiner et al. 2009), are linked to alterations in uptake of 5-HT from the presynaptic cell membrane via hSERT(Blakely, Berson et al. 1991). For treatment of these diseases, selective 5-HT reuptake inhibitors (SSRIs) were developed that elevate 5-HT levels at the synapse. One such SSRI, citalopram, alleviates the symptoms of depression and is the focus of our current study(Hyttel 1994).

## Importance and Binding Determinants of Citalopram in hSERT

The stereochemistry of citalopram determines its activity. The active S-configuration binds with high-affinity to hSERT, whereas the R- configuration exhibits reduced binding affinity at hSERT but may also modulate interactions of the S- isomer with the transporter(Henry, Field et al. 2006). Although the effects of citalopram are well studied, the structural determinants of interactions of S- and R-citalopram with hSERT remain a focus of current investigations. We have shown that S- and R-citalopram exhibit potency differences between hSERT and dSERT that largely derive from interactions with two residues (Y95 and I172 in hSERT)(Henry, Field et al. 2006). These differences provide an important, suitable test of small molecule docking methods to homology models of SERT proteins, an approach we have recently used for docking of 5-HT(Kaufmann, Dawson et al. 2009). Here, we dock both stereoisomers of citalopram computationally into models of human and fly SERT. The predictive power of the models is tested by introducing known mutations that disrupt S- and R-citalopram interactions with hSERT. Our models recapitulate the species-selective binding affinity differences between hSERT and *Drosphila* SERT (dSERT), as well as the distinct binding affinities of S- and R-citalopram isomers to wildtype transporters.

### *Construction of hSERT homology model and SAR of Citalopram*

SERT homology models were constructed based on multiple sequence alignments(Beuming, Shi et al. 2006) of SLC6 family members with the crystal structure of $\mathrm{LeuT_{Aa}}$ as a template(Singh, Yamashita et al. 2007). A series of structure-activity relationships (SAR) and mutational studies guide analysis of structural determinants of



| AA | TM | a | b | c | d | e | f |
|---|---|---|---|---|---|---|---|
| Y95 | 1 | 0% | 1% | 100% | 99% | 0% | 0% |
| D98 | 1 | 40% | 0% | 66% | 98% | 93% | 32% |
| I172 | 3 | 23% | 97% | 12% | 81% | 97% | 100% |
| S438 | 8 | 0% | 81% | 90% | 88% | 53% | 94% |
| A96 | 1 | 0% | 0% | 82% | 53% | 0% | 0% |
| R104 | 1 | 99% | 0% | 0% | 1% | 81% | 0% |
| A169 | 3 | 0% | 76% | 1% | 0% | 0% | 79% |
| A173 | 3 | 0% | 96% | 0% | 0% | 0% | 94% |
| Y175 | 3 | 88% | 0% | 0% | 0% | 96% | 10% |
| Y176 | 3 | 89% | 62% | 13% | 8% | 100% | 89% |
| F335 | 6 | 91% | 0% | 36% | 99% | 100% | 7% |
| S336 | 6 | 1% | 0% | 84% | 99% | 11% | 0% |
| L337 | 6 | 0% | 0% | 82% | 92% | 0% | 3% |
| G338 | 6 | 0% | 0% | 88% | 98% | 22% | 56% |
| F341 | 6 | 0% | 28% | 45% | 76% | 77% | 83% |
| V343 | 6 | 0% | 59% | 100% | 93% | 0% | 87% |
| T439 | 8 | 0% | 77% | 2% | 1% | 0% | 76% |
| A441 | 8 | 0% | 84% | 98% | 53% | 1% | 84% |
| G442 | 8 | 0% | 100% | 18% | 1% | 0% | 100% |
| E493 | 10 | 65% | 0% | 0% | 0% | 48% | 0% |
| T497 | 10 | 51% | 0% | 0% | 13% | 96% | 1% |

🟥 0% - 25%
🟧 26% - 50%
🟨 50% - 75%
🟩 76% - 100%

**Figure 4: S-citalopram and heatmap of S-citalopram contacts with backbone side chains.**
**A)** S-citalopram functional groups involved in binding are labeled [a]-[f]. **B)** Contacts within the S-citalopram/hSERT high-affinity binding mode are displayed as a heatmap. Given is the fraction of models that display a contact within the largest cluster of models. Residues known to affect S-citalopram binding are shown within the black box.

S-citalopram binding (Figure 4a). The cyano group [a], the dimethyl-amine [c], and the fluorine of the flouro-phenyl substituted ring [b] of S-citalopram were determined as key functional groups for inhibitor potency through substitution with hydrogen(Eildal, Andersen et al. 2008). Through mutational studies, several amino acids have been identified as dictating high affinity binding for S-citalopram. D98 is known as a conserved residue found in all monoamine transporters and has been shown to stabilize a protein-ligand salt bridge(Barker, Moore et al. 1999). Henry (Henry, Field et al. 2006) et al. produced point mutations to convert two residues in hSERT to the corresponding dSERT residues (Y95F and I172M), resulting together in a ~6,000 fold loss of potency for S-citalopram(Henry, Field et al. 2006). The single mutations of Y95F and I172M resulted in a 19 and 344 fold loss, respectively. In contrast, R-citalopram displays a 79 and 5 fold loss of potency for the same mutations and only a 117 fold loss if the mutations are combined. In this study we propose that the dimethyl-amine group [c] of S-citalopram interacts with Y95 while the cyanophthalane ring [e] and the flouro-phenyl ring [f] interacts with I172(Henry, Field et al. 2006). More recently, Andersen(Andersen, Taboureau et al. 2009) et al. has demonstrated the importance of S438 in binding the dimethyl-amine group through a mutation S438T that results in a 300 fold loss of potency for S-citalopram; however, the S438T mutant only results in a loss of potency of 15-20 fold for R-citalopram (personal communication with Andersen).

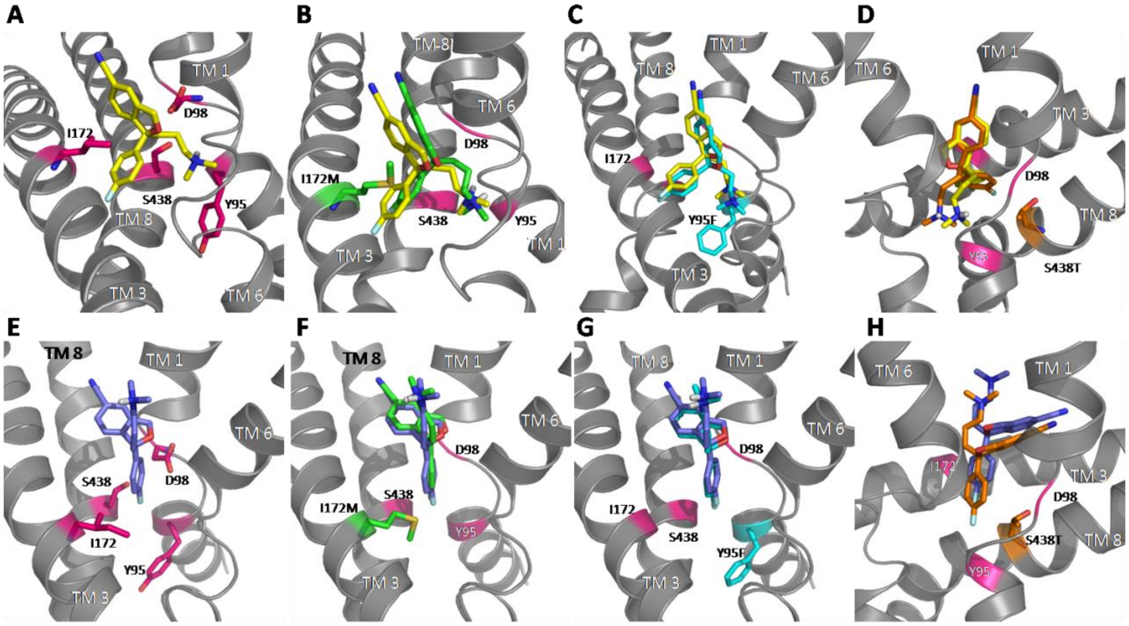
### *Experimental Procedure for Mutants Proposed from Computational Methods*

Site-directed mutagenesis of hSERT and dSERT in pcDNA3.1 was performed using the QuikChange mutagenesis kit and protocol (Stratagene). Sense and antisense oligonucleotides, purchased from Invitrogen, were designed to generate the single amino acid substitutions. Oligonucleotide sequences used for mutagenesis are available upon request. Sequencing of all mutants was performed at the DNA Sequencing Facility of the Division of Genetic Medicine at Vanderbilt University Medical Center. Successful mutants were transformed into DH5α *Escherichia coli* cells for amplification and purified using the Qiafilter Maxiprep kit (Qiagen). HEK-293 cells, maintained at 37 °C in a 5% $CO_2$ humidified incubator, were grown in complete medium (Dulbecco's modified Eagle's medium, 10% fetal bovine serum, 2 mm l-glutamine, 100 units/ml penicillin, and 100 µg/ml streptomycin). For initial evaluation of mutant transporter activity, cells were plated at a density of 50,000 cells per well in 24-well culture plates. Cells were transfected with hSERT or dSERT constructs using TransIT transfection reagent (Mirus Inc., 6 µl per µg of DNA), in Opti-MEM medium. Following transfection (24–48 h), cells were washed with KRH assay buffer (120 mm NaCl, 4.7 mm KCl, 2.2 mm $CaCl_2$, 1.2 mm $MgSO_4$, 1.2 mm $KH_2PO_4$, 10 mm HEPES, pH 7.4) and incubated for 10 minutes with increasing concentrations of nonradiolabeled competitor, followed by the addition of a constant concentration of $[^3H]$5-HT (5-hydroxy$[^3H]$tryptamine-trifluoroacetate (100-110 Ci/mmol); Amersham Biosciences) for 15 minutes. Transport assays were terminated by washing two times with assay buffer, and cells were dissolved in MicroScint 20 (PerkinElmer Life Sciences) scintillation fluid. The extent of $[^3H]$5-HT accumulation was determined by liquid scintillation counting on a TopCount System (PerkinElmer Life Sciences). Uptake in mock-transfected cells was subtracted from transporter-transfected cells to determine specific uptake. Specific uptake was normalized to percent uptake of control wells that lacked competitor and plotted versus the log of the molar concentration of competitor. The data were fit to a nonlinear one-site competition curve and apparent $K_I$ values were derived using the Cheng-Prusoff equation in Prism 4 for Mac (Graphpad Software). All experiments were performed in triplicate and repeated in three or more separate assays.

### *Computational Method to Dock Citalopram into hSERT/dSERT Models*

RosettaLigand(Davis and Baker 2009) was used to dock S- and R-citalopram separately into the hSERT homology models while accounting for full protein and ligand flexibility. Specifically, the protein backbone conformation underwent repeated energy minimization from the initial homology model which resulted in a large conformational ensemble of backbones. An ensemble of the top ten energy minimized backbones was used in the docking steps. Amino acid side chain conformations were chosen from a rotamer library(Dunbrack 2002) and further optimized through gradient minimization. Ligand flexibility was modeled through knowledge-based torsion angle potentials derived from the chemical Cambridge Structural Database (CSD)(Kaufmann 2008) yielding a total of 1,000 S- and R-citalopram conformations. The generation of 1,000 conformations for S- and R-citalopram allows for thorough sampling of non-clashing conformations for the five rotatable bonds. The 1,000 conformations cluster into 238 rotamers. The molecule has five rotatable bonds. A total of 238 non-clashing conformations allows an average of three states for each of the bonds to be independently sampled. This number is reasonable to ensure thorough sampling of conformational space. The protocol for construction of these homology models, the creation of conformational ensembles, and docking are described in detail elsewhere(Kaufmann, Dawson et al. 2009). A total of 23,500 models for the S-citalopram/hSERT and R-citalopram/hSERT complex were created.

Our models were first filtered according to their RosettaLigand binding energy selecting the top 10%. The resulting 2,350 models were clustered into separate binding modes based upon a 3Å RMSD threshold. This resulted in multiple clusters of varying size. The models presented here are the lowest energy models from the largest clusters, respectively. After a binding mode had been established for S-citalopram/hSERT and R-citalopram/hSERT complex, mutations were introduced into the hSERT backbone while maintaining the putative binding mode of the S-citalopram/hSERT and R-citalopram/hSERT complex. The mutations analyzed were Y95F, I172M, S438T, and the double mutant Y95F/I172M. After introduction of the mutations each model was refined through Monte Carlo moves of up to 2Å and 10° for the ligand while the protein backbone and side-chains were minimized. For each mutant the models with the lowest RosettaLigand binding energy were analyzed. To explore the proposed conformation of S-citalopram in relation to dSERT, the putative binding mode of S-citalopram/hSERT complex was placed into nine different backbones of dSERT(Kaufmann 2008). The binding mode underwent Monte Carlo refinement and optimization. The lowest energy was chosen for analysis.



| WT | I172M | Y95F | S438T |
|---|---|---|---|
| S-citalopram 1 nm, -8.06 REU | 1,000 nm, -6.2 REU | Not available, -8.0 REU | 7,924 nm, -6. REU |
| R-citalopram 180 nm, -6.54 REU | 22,0 00 nm, -6.1 REU | Not available, -6.2 REU | Not available, -6.1 REU |

**Figure 5: Binding energies, experimental inhibitory constants, and models of S- and R-Citalopram in the human serotonin transporter.**

S- and R-citalopram in complex with mutants of hSERT. The extracellular side of the protein is shown on the top of all images whereas the cytoplasm is at the bottom of all images. S-citalopram/hSERT complex is shown on top, R-citalopram/hSERT complex is shown below. Experimental binding affinities, $K_i$, and computationally predicted binding energies are given below each image. A and E) WT hSERT in complex with S- and R-citalopram. Experimentally verified residues involved in binding are shown as sticks and highlighted in red and labeled. B and F) S- and R-citalopram putative wild type binding mode docked into I172M mutant of hSERT (green). The original wild type binding mode is displayed in yellow and blue (S-citalopram and R-citalopram respectively). The mutation I172M is shown in green and stick format with the experimentally verified residues shown in cartoon and highlighted with red. C and G) S- and R-citalopram docked into Y95F mutant. The putative wild type binding is colored cyan with the mutant Y95F colored in cyan and shown as a stick. D and H) S- and R-citalopram docked into S438T. The putative wild type binding is colored in orange with the mutant S438T shown in sticks and colored orange.

### *Binding Determinants of Citalopram*

A single cluster of interactions for the S-citalopram /hSERT complex stood out by both its size (400 members out of a total of 2,350 models) and RosettaLigand binding energy of -8.1 Rosetta Energy Unit (REU) (Figure 5a). This model preserves the hypothesized D98[12] dimethyl-amine [c] contact along with a hydrogen bond between the dimethyl-amine group [c] and the backbone carbonyl group of Y95(Henry, Field et al. 2006). Additionally, the dimethyl-amine group [c] is in contact with S438. The model is further characterized by a hydrophobic clamp that is formed by the flouro-phenyl ring [f] and the cyanophthalane ring [e] around I172(Henry, Field et al. 2006).

S-citalopram/hSERT contacts were compiled as a heat map to illustrate amino acid residues in and around the S-citalopram binding pocket of our model (Figure 4b). Novel contacts proposed by the present model include the methylene groups of the 3-(dimethylamino)propyl tail [d] which form VDW contacts with residues S336, L337, G338 in TM 6 and A441 in TM 8. The cyano group [a] is pointed towards the extracellular space, out of the membrane and binding pocket, and towards residue R104 in TM 1 and E494 in TM 10 and has VDW contacts with G100 in TM 1, F335 in TM 6, and Y175 and Y176 in TM 3. Additionally, the flouro-phenyl ring [f] sits in a hydrophobic pocket created by residue A169(Larsen, Elfving et al. 2004) in TM 1 and A441 and G442 in TM 8.

Docking of R-citalopram results in two clusters of similar size and energy. The largest cluster (124 members out of 2,350) has an energy of -6.5 REU and is in a different conformation found in the S-citalopram/hSERT complex. Interestingly, the second largest cluster has R-citalopram in the similar conformation as the S-citalopram/hSERT complex; however, the dimethyl-amine tail is distant from S438 and pointed in the direction between TM 1 and TM 6. To remain consistent with the analysis of S-citalopram, analysis of the largest cluster was performed for R-citalopram.

Our model indicates that the R-citalopram/hSERT complex has a significantly different binding conformation than S-citalopram/hSERT complex. The cyanopthalane ring [e] and dimethyl-amine [c] of R-citalopram is shifted out of the binding pocket and positioned towards the extracellular face of the transporter. In particular, R104 in TM 1 and E493 in TM 10 form hydrogen bonds with the dimethyl-amine group [c], whereas the cyano group [a] is directed towards W103 in TM 1. The flouro-phenyl ring [f] sits in the binding pocket and is pointed downward towards Y95 and D98 in TM 1. Additionally, Y175 in TM 3 is π -stacked against the flouro-phenyl fing [f].

### **In silico** ***Mutational Analysis Supports Citalopram Binding Mode***

To test our models, the putative binding modes for S- and R-citalopram were examined via *in silico* mutations that have been previously studied experimentally: I172M, Y95F, S438T, and an I172M/Y95F double mutant(Henry, Field et al. 2006; Andersen, Taboureau et al. 2009). Experimentally, the I172M mutation results in a 344 fold loss of potency for S-citalopram(Henry, Field et al. 2006). Accordingly, docking into the I172M mutant results in a reduced binding energy of -6.2 REU (Figure 5). The conformation of S-citalopram within the binding pocket is shifted away from the site of mutation I172M. The reduction in score is attributed to compromised hydrophobic packing at site 172 and loss of a π-stacking interaction with F355.

Experimentally, S-citalopram interaction with a Y95F mutant of hSERT results in a 19 fold loss of potency(Henry, Field et al. 2006). This experimental finding as well as studies with 5-HT suggest that Y95 is directly involved in hSERT interaction with

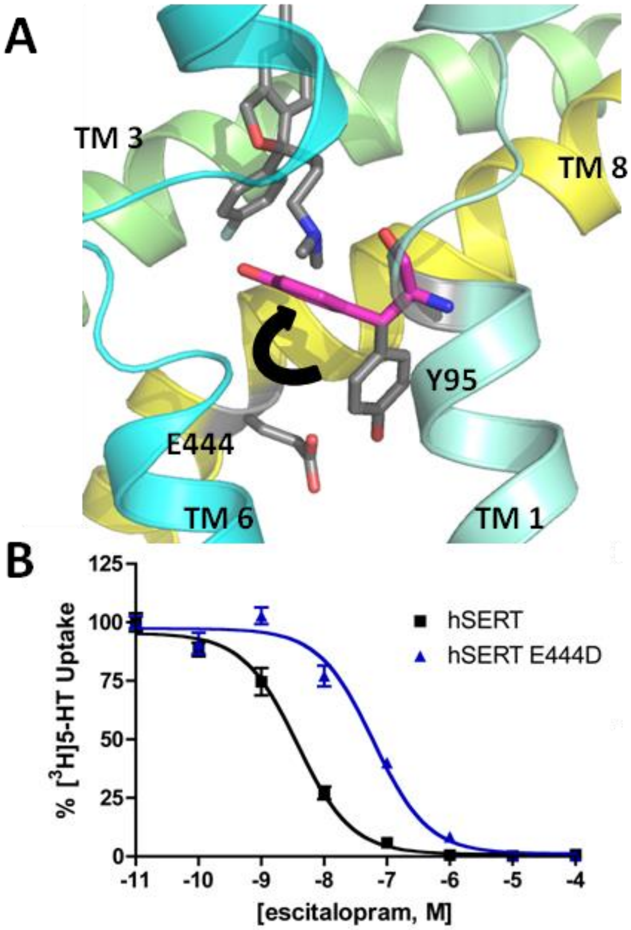


**Figure 6: Proposed binding mode of Y95 and E444 gate in the human serotonin transporter.**
**A)** Binding mode of S-citalopram/hSERT complex with depiction of the Y95 (grey) in a downward position. Mutation of E444D (not shown) results in Y95 populated in two positions, a downward position (grey) and an upward position (pink). B) Mutation of E444D results in a 10 fold loss of potency for S-citalopram suggesting that the Y95 switches between two different conformations, an upward (pink) and downward conformation (grey). Wild Type is shown in by black squares and line and E44D is shown in circles and spheres.

ligands(Henry, Field et al. 2006). However, our comparative model has Y95 pointing away from the binding pocket and engaged in a hydrogen bond with E444. A conformational change of Y95 into an upward pointing conformation requires substantial rearrangement of the protein backbone from the template structure LeuT$_{Aa}$ and breaking of the Y95-E444 hydrogen bond. To experimentally test the effects of breaking the hypothesized Y95-E444 hydrogen bond with minimal impact on other areas of the transporter a hSERT E444D mutant was created. We found that this mutant displays a modest, 10 fold loss of potency for S-citalopram (Figure 6a,b). This observation supports the contention that E444 is indirectly involved in citalopram binding. Our data does not conclusively prove that E444 locks Y95 into a downward orientation required for high-affinity binding of citalopram. We speculate that an upward pointing conformation of Y95 may be engaged in earlier stages of S-citalopram binding (e.g. an outward facing, open conformation of hSERT) followed by the downward pointing conformation by binding of S-citalopram thereby explaining the experimental findings (Fig 3). Regardless, docking studies of S-citalopram/Y95F hSERT resulted in a slightly reduced binding energy of -8.0 REU coupled with a shift of S-citalopram away from the extracellular face of hSERT to deeper inside the binding pocket towards F95.

Similar to the I172M mutation, our model of the Y95F/I172M double mutant predicts a movement of S-citalopram away from I172M and a reduced binding energy of -7.6 REU. The loss in binding affinity is largely attributed to the I172M mutation.

Recently, Andersen(Andersen, Taboureau et al. 2009) and colleagues mutated S438T and tested the mutant against binding of dimethylated inhibitors. A mutation of S438T resulted in a loss of potency for dimethylated inhibitors, however, a 300 fold loss for S-citalopram potency was reported. In our model the S-citalopram complex with the S438T mutant results in a significant reduction of energy to -6.0 REU. The dimethyl-amine group shifts away from T438 which results in repulsive interaction with residues on TM 1. The influence of Na+ ions on the S438T mutant was also tested. However, no significant change in the model and energy was observed (see supplementary data).

For R-citalopram introduction of bulk via the I172M mutation results in a slight loss of binding energy to -6.1 REU, in agreement with experimental findings(Henry, Field et al. 2006). The flouro-phenyl ring [f] is shifted away from the mutation and towards TM 1 and TM 6. The Y95F hSERT mutation results in a slightly lower binding energy than wild type (-6.2 REU). The double mutant I172M/Y95F (-6.1 REU) displays an additional shift of the flouro-phenyl ring [f] away from M172, consistent with the contention that I172 is a critical contact site for the SSRI(Henry, Field et al. 2006). Andersen reported that a mutation of S438T resulted in a 15 fold loss of potency for R-citalopram/hSERT complex (personal communication). In agreement with his findings, the R-citalopram/S438T hSERT complex displays a reduced binding energy of -6.1 REU coupled with a shift away from the site of mutation.

### *Validation of Citalopram Models through Comparative Docking into dSERT*

To further validate the S-citalopram/hSERT model, dSERT homology models were used to dock the putative binding mode of S-citalopram. dSERT differs in selectivity for racemic citalopram compared to hSERT. Racemic citalopram exhibits a $K_i$ value of 400 nM in dSERT, however, a point mutation of M167I (I172 hSERT) results in an increase of binding affinity to 4 nm(Henry, Field et al. 2006). To determine if our docking paradigm could reproduce the experimental finding, S-citalopram/dSERT models were constructed. The S-citalopram/dSERT binding energy for the lowest-energy pose (Figure

7) was significantly worse when compared to the S-citalopram/hSERT complex (-6.6 REU and -8.1 REU respectively). This loss is attributed to two energy contributions when comparing the two models, the hydrogen bonding term and the solvation term (+0.8 REU and +0.7 REU respectively). Specifically, a network of hydrogen bonding interactions between sidechains and the cyano group [a] in S-citalopram/hSERT are absent in the S-citalopram/dSERT model. The hSERT residues that are hydrogen bonded to the cyano group [a] are R104 and E493. Although R104 is preserved in dSERT, E493 is substituted to N, which is predicted to alter the hydrogen bond network. To test the significance of the hydrogen bond network, reciprocal mutants of hSERT E493N and dSERT N481E were constructed and tested for S-citalopram potency. We observed a ~3 fold average loss of potency in the hSERT E493N mutant that did not reach statistical significance (data not shown). The reciprocal mutation dSERT N481E was without effect. We had hypothesized that the E493-R104 salt-bridge in hSERT positions R104 for a constructive interaction with the cyano group [a]. Therefore a N481E mutation in dSERT was expected to increase S-citalopram affinity. However, the interaction between R99 (residue R104 in hSERT) and the cyano group [a] in WT dSERT is not strengthened through the N481E mutation. A possible explanation is that the positive charge of R104 is compensated through E481 weakening the interaction with the partial negative charge of the cyano group [a]. Additional experiments are needed to verify this hypothesis.

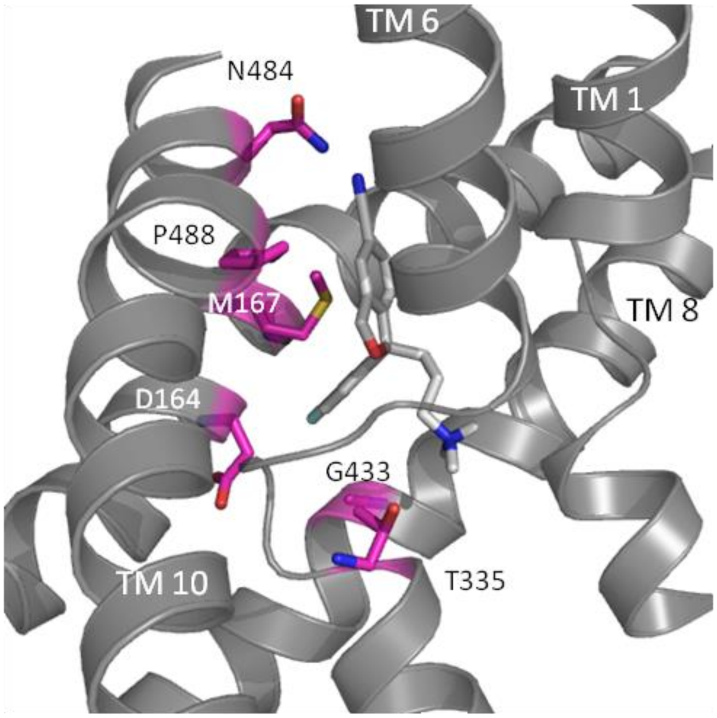


**Figure 7: Binding site of S-citalopram/dSERT.**
The putative binding mode of S-citalopram/dSERT complex. S-citalopram is shown in white. Residues that contribute to lowering the energy of the complex are highlighted in red. Substitution of A169 to D164 dSERT results in a higher solvation score for the complex. Additionally, M167 dSERT results in an increase in the VDW potential.

The change of solvation energy of S-citalopram in our models can be attributed to the burial of the flouro-phenyl ring [f] in a hydrophobic environment. hSERT and dSERT residues are homologous in this region, with the exception of hSERT residues A169 and I172 which corresponds to dSERT residues D164 and M167. Of these two residues, the solvation energy for hSERT A169 and dSERT D164 are drastically different (+1.1 REU in dSERT). This finding matches experimental studies that affiliate the A169D point mutation with a loss of S-citalopram potency in hSERT(Larsen, Elfving et al. 2004). Further, the I172M mutation in hSERT causes a binding energy increase in the VDW potential of S-citalopram/hSERT complex by +2.8 REU. As discussed above, the hSERT I172M mutant has been tested experimentally and displayed a significant loss of S-citalopram potency. However, when the reverse mutant is expressed, dSERT M167I, dSERT regains hSERT-like potency for S-citalopram(Henry, Field et al. 2006). When the computational mutant is tested, a similar energy to S-citalopram/hSERT is observed. S-citalopram/dSERT M167I results in an energy of -8.1 REU and recapitulates the docked S-citalopram/hSERT pose.

### *Comparison of Presented Models to Previous Publications and Conclusions*

Recently, Andersen[1], et al published an independent computational model for S-citalopram in hSERT. The authors created 64 point mutants to validate their binding mode. The proposed model by Andersen places the cyano group [a] sandwiched between TM 6 and 8 whereas our model positions the cyano group [a] outward towards the extracellular face of the protein. This results in an RMSD of ~5.1Å difference to our model. The ligand occupies the same binding pocket but experiences a ~30° rotation. A qualitative and quantitative comparison of mutation data presented by Andersen displays a similar level of agreement between the two models suggesting that further experiments are needed to more precisely define the position of S-citalopram (see supplementary data). One noticeable difference between the models and reason for the slight shift and rotation in the S-citalopram is a manual positioning of the Y95 residue in an upward configuration, providing new contacts with S-citalopram. Experiments are needed to validate this positioning of Y95 during the transport cycle and antagonist binding. They also identify unexpected interactions between TM1 and 8 that may contribute to critical conformational transitions in the dynamic organization of other SLC6 transporters.

In addition to Andersen's model, Koldsø[19] et al created models of S-citalopram and R-citalopram complexed in hSERT. Unfortunately the coordinates of these models are not available to us for quantitative comparison. Qualitatively, the S-citalopram/hSERT complex proposed compares to the model presented here. In Koldsø's model, the flouro-phenyl ring [f] is in the hydrophilic pocket lined by A169, A13, N177, S438, T439, and L443. Our model positions the flouro-phenyl ring [f] in this pocket as well. The cyano group [a] in Koldsø's model is positioned next to T497, F335, F341, and V501. In contrast, our model positions the cyano group [a] pointed outside of the binding pocket and is next to F335, R104, and E493.

The R-citalopram conformations are noticeably different. Koldsø's R-citalopram/hSERT complex differs from the S-citalopram/hSERT complex by the reverse placement of the cyanophthalane ring [e] and flouro-phenyl ring [f]. In contrast, the presented R-citalopram/hSERT in this paper is bound higher in the binding pocket in a completely different conformation. The contrasting results might suggest multiple low affinity binding sites for R-citalopram and require further experimental investigation.

In summary, our study confirms aspects of the S-citalopram binding mode using orthogonal computational techniques, providing a detailed analysis of the impact of citalopram isomerization and further elucidates the species selectivity for SSRI recognition.

## Quantum Dot Docking into hSERT Models

The human serotonin transporter (hSERT) has been implicated in depression, autism, attention deficit hyperactivity disorder (ADHD), obsessive-compulsive disorder (OCD), and addiction(Hirschfeld 2000; Prasad, Steiner et al. 2009). A large emphasis has been placed to optimize selective serotonin reuptake inhibitors (SSRI) to block the reuptake of the substrate, serotonin, and alleviate the symptoms of the diseases/syndromes(Gether, Andersen et al. 2006; Apparsundaram, Stockdale et al. 2008); however, a major challenge associated with the discovery of new, more efficacious antidepressants is the lack of a resolved crystal structure of the serotonin transporter and efficient screening protocols.

To investigate hSERT activity, protocols include conventional biochemical and radiolabeling approaches such as phosphorylation assays, electrophysiology, or radio-isotope substrate uptake assays(Ramsey and DeFelice 2002). However, these approaches include time-consuming, labor-intensive experimental work, poor spatiotemporal resolution, and safety concerns associated with isotope handling. Alternatively, the rapid and high resolution method of fluorescent probes can be used for targeted selective drug screening of hSERT. Typically, an antibody-based labeling approach is utilized to detect single membrane proteins(Dahan, Levi et al. 2003; Fichter, Flajolet et al. 2010); however, many proteins lack suitable extracellular domains that can be targeted by antibodies without functional disruption. Furthermore, the use of popular fusion tags to label membrane proteins, usually consisting of a short peptide sequence with a recognition epitope for a complementary binding partner, such as hemaglutinin (HA), requires genetic perturbation of the protein target(Kovtun, Ross et al. 2012) and thus does not allow direct visualization of endogenous protein of interest.

Organic ligands conjugated to quantum dots can achieve specific targeting of transporter proteins, which allows visualization of endogenous protein without the functional disruption. Quantum dots (QDs), nanometer-sized semiconductor nanocrystals, exhibit excellent photo stability and brightness enabling long-term imaging of biological systems(Rosenthal, Chang et al. 2011). Their broad absorption spectra and size-dependent, narrow, symmetric emission spectra of QDs considerably simplify multiplexed, molecular imaging experiments. Based on the molecular determinants for binding and distance within a membrane protein from the cell surface, the organic ligand can be tailored with a saturated linker arm to optimize QD conjugate binding to the target protein(Kovtun, Ross et al. 2012). Traditional approaches for optimizing binding of conjugated ligands to QDs require time consuming synthesis with all iterations of linker size. A rapid approach to find the structural requirements for optimal binding activity can be achieved through a series of guided synthesis approaches from molecular modeling.

Here, we present a method for modeling, identification, and synthesis of small-molecules to target the hSERT with QD conjugated ligands. RosettaLigand(Lemmon and Meiler 2012) is used to identify small-molecules that bind hSERT and are easily used to attach a linker for quantum dot conjugation. After a small molecule is identified, RosettaLigand is used to identify the number of carbons needed to build an alkyl spacer from the small-molecule to the outer membrane. Finally, we show that using RosettaLigand along with experimental techniques can rapidly identify the correct ligand and spacer length without too much experimental work. The approach is widely applicable to any membrane protein and eliminates problems seen in traditional docking methods with scoring and ranking models by using a distance to the outer membrane as criteria to determine which ligands to synthesize.

### *Computational Methods for Quantum Dot Labeling*

Rapid conformational sampling for the ligand is done through pregenerated rotamers created with the Molecular Operating Environment (MOE) 2009 application using the LowModeMD approach(Boyd 2005). For the parent drugs without any alkyl spacer length, 3-(1,2,3,6-tetrahydropyrindin-4-yl)-1H-indole and 3-(1,2,3,6-tetrahydropyridin-4-yl)-1H-indole-5-carbonitrile, a root mean square distance (RMSD) cutoff of 0.5 Å to each reoccurring conformation was used as a filter to remove similar conformations. The resulting sets of conformations were filtered based on energy and the top 10 lowest scored conformers were kept as rotamers for ligand docking in RosettaLigand.

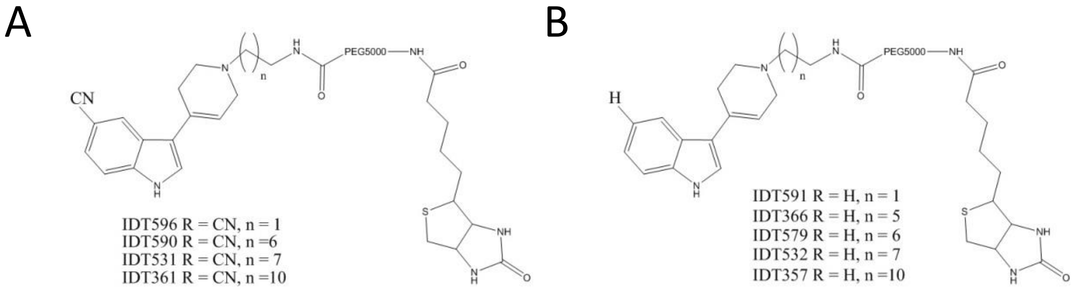


**Figure 8: Base structures for conjugation of quantum dots in the human serotonin transporter. A)** 3-(1,2,3,6-tetrahydropyridin-4-yl)-1H-indole-5-carbonitrile and **B)** 3-(1,2,3,6-tetrahydropyrindin-4-yl)-1H-indole ligands.

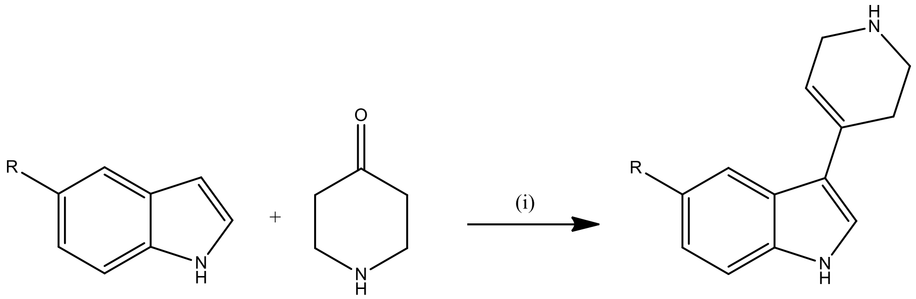
A separate method was used to generate rotamers for the parent drug with the varying alkyl spacer lengths. For alkyl spacer length one to three, the linker arm attached to the parent drug was allowed to fully sample conformational space while the lowest scored conformation resultant from initial docking studies of the parent drug was kept static. For alkyl spacer lengths of three to twelve, the resultant conformation from the previous docking study was kept constant, including linker arm, while the last three carbons of the linker arm were allowed to sample conformational space. The same protocol for ligand conformation generation of the parent drug was used to sample the alkly spacer conformations.

RosettaLigand within the RosettaScripts(Lemmon and Meiler 2012) framework was used in all docking steps. An initial docking step was performed on the small-molecules IDTX and IDTY into an experimentally verified homology model of hSERT(Kaufmann, Dawson et al. 2009) to ensure that the substituent pyrrole group was positioned towards the extracellular face for conjugation with the alkyl spacer. After initial placement within the binding site of hSERT, the ligands were allowed to translate with a Gaussian distribution 5.0 Å and rotate 360.0°. A total of 5,000 models were generated for each compound and the top ten lowest scoring models by Rosetta Energy Units (REU) were examined for three-dimensional orientation of substituent pyrrole group positioned towards the extracellular matrix. The lowest scoring model with substituent pyrrole group pointed towards the extracellular matrix was chosen as the starting structure for building the alkyl chain.

A separate method was used to dock the parent drug with alkyl spacers. In an iterative process, the previously best scoring complexes were used as the initial position docking. For example, the starting position for a six carbon spacer length ligand was the ending position for the five carbon spacer length ligand. At each step, the ligand was allowed to rotate 5.0$^{o}$ and translate 1.5 Å allowing for induced fit.

### *Experimental Methods*

The synthesis of the conformationally restricted homotryptamine derivatives **(10a)** and **(10b)** used in this study is outlined in scheme 1. These were obtained by refluxing 4-piperidone mono hydrate monochloride with the appropriate indole in a methanolic solution of potassium hydroxide as previously described,(Tomlinson, Mason et al. 2005) giving 3-(1,2,3,6-tetrahydropyridin-4-yl )-1H-indole-5-carbonitrile **(10a)** in a 22.8% yield and 3-(1,2,3,6-tetrahydropyridin-4-yl)-1H-indole **(10b)** in a 72.5% yield.

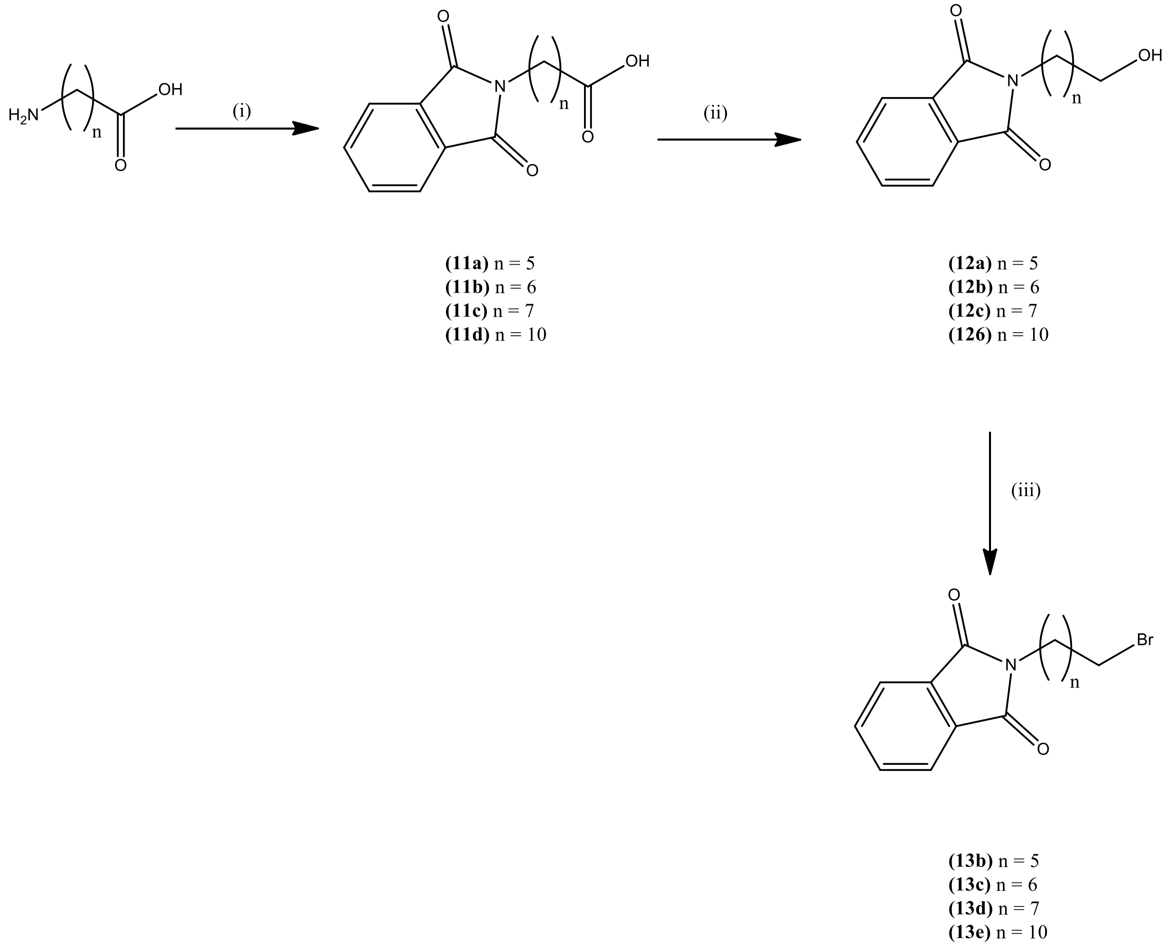


**(10a)** R = CN
**(10b)** R = H

**Scheme 1.** (i) (a) KOH, Methanol, reflux 18 hours; (b) $H_2O$, ambient temperature, 72 hours.

Once the parent drug had been synthesized the protected alkyl spacer was attached to the nitrogen atom of the tetrahydropyridyl ring. A phthalimide protected alkyl bromide was selected as our reagent of choice. The 2 carbon alkyl spacer intermediate 2-(2-bromethyl)-1*H*-isoindole-1,3-(2H)-dione **(13a)** was commercially available and purchased from Sigma Aldrich (Milwake, WI). The other phthalimide protected alkyl spacers were synthesized by converting the phthalimide protected amino acid derivatives to the phthalimide protected bromo alkane as outlined in scheme 2.

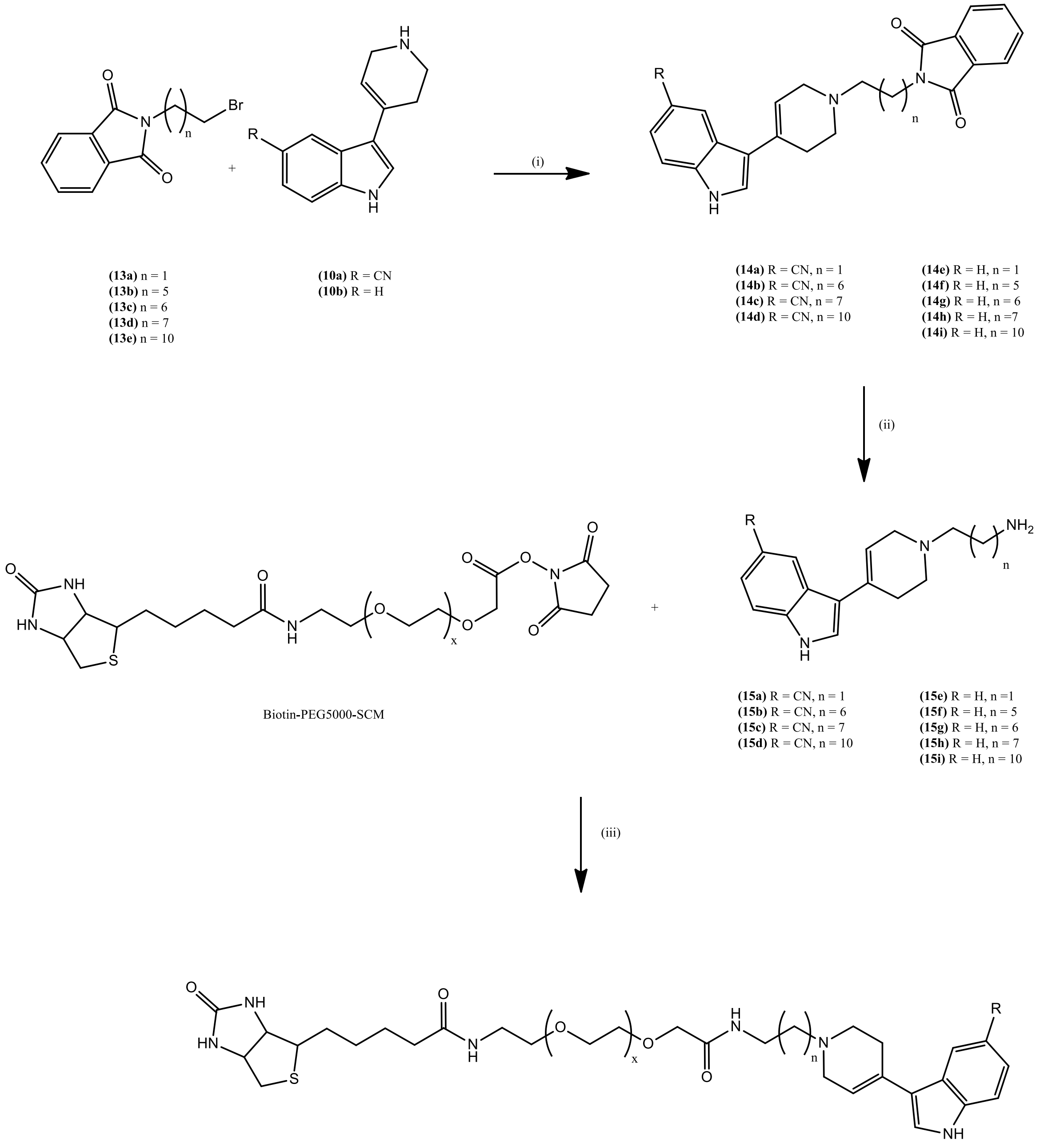
Initially the amino acids were protected by reacting the appropriate amino acid with N-carbethoxy phthalimide in an aqueous solution of potassium carbonate(Gademann, Kimmerlin et al. 2001) to give **(11a)**-**(11d)**. After which the carboxylic acids were reduced to an alcohol by treatment with ethyl chloroformate in dry THF to form the mixed anhydride which was subsequently reduced to the alcohols **(12a)**-**(12d)** by treatment with sodium borohydride and methanol in one pot(Tomlinson, Warnerment et al. 2007). Finally the alcohol was converted to a bromide using a variation of the Appel reaction by treatment of the alcohols **(12a)**-**(12d)** with N-bromo succinimide and triphenyl phosphine in methylene chloride at ambient temperature to give **(13b)**-**(13e)**. When the amino acid reagent was not commercially available or prohibitively expensive (i.e. n = 6 and 7) they were synthesized using the method described by Bell et. al.(Bell, Faggiani et al. 1992) and characterized by nmr which was consistent with previously reported nmr data(Bell, Faggiani et al. 1992).



**Scheme 2**

(i) N-carbethoxyphthalimide, $Na_2CO_3$ (aq); (ii) (a) Ethyl chloroformate, $Et_3N$, THF; (b) $NaBH_4$, MeOH, 0°C; (iii) NBS, $PPh_3$, $CH_2Cl_2$, ambient temperature, 18 hours.

The phthalimide protected alkyl spacers **(13a)**-**(13e)** were attached to the homotryptamine derivatives **(10a)** and **(10b)** by refluxing the appropriate length spacer with the homotryptamine derivatives in the presence of triethylamine as outlined in scheme 3 to give the intermediates **(14a)**-**(14i)**. Then the phthalimide protecting group was removed by treatment with hydrazine hydrate in ethanol to give the intermediate amines **(15a)**-**(15i)** and the amino terminated intermediates were attached to the end of the biotinylated polyethylene glycol chain. This was achieved by stirring the appropriate amine with Biotin-PEG5000-SCM in methylene chloride at ambient temperature for 18 hours to give the crude ligands **(1)**-**(9)** and these were purified by chromatography on a sephadex G10 column followed by preparative HPLC. All of the ligands **(1)-(9)** were characterized by LC-MS and MALDI-TOF mass spectroscopy to determine their purity and average molecular weights. Since **(1)**-**(9)** contain a poly dispersed PEG the average molecular weight of the ligands had to be calculated. This was done by integrating the peak areas for each signal in the MALDI spectrum and then normalizing the peak area to the total area as a percentage. The normalized areas were then multiplied by the *m/z* value for that peak and these values were summed to obtain the average molecular weight.



**(13a)** n = 1
**(13b)** n = 5
**(13c)** n = 6
**(13d)** n = 7
**(13e)** n = 10

**(10a)** R = CN
**(10b)** R = H

**(14a)** R = CN, n = 1
**(14b)** R = CN, n = 6
**(14c)** R = CN, n = 7
**(14d)** R = CN, n = 10

**(14e)** R = H, n = 1
**(14f)** R = H, n = 5
**(14g)** R = H, n = 6
**(14h)** R = H, n =7
**(14i)** R = H, n = 10

**(15a)** R = CN, n = 1
**(15b)** R = CN, n = 6
**(15c)** R = CN, n = 7
**(15d)** R = CN, n = 10

**(15e)** R = H, n =1
**(15f)** R = H, n = 5
**(15g)** R = H, n = 6
**(15h)** R = H, n = 7
**(15i)** R = H, n = 10

IDT596 **(1)** R = CN, n = 1
IDT590 **(2)** R = CN, n = 6
IDT531 **(3)** R = CN, n = 7
IDT361 **(4)** R = CN , n = 10

IDT591 **(5)** R = H, n = 1
IDT366 **(6)** R = H, n = 5
IDT579 **(7)** R = H, n = 6
IDT532 **(8)** R = H, n = 7
IDT357 **(9)** R = H, n = 10

**Scheme 3.** (i) acetonitrile, triethylamine, reflux 18 hours; (ii) (a) ethanol, hydrazine mono hydrate, ambient temperature, 1 hour. (b) $CH_2Cl_2$, ambient temperature, 18 hours; (iii) $CH_2Cl_2$, ambient temperature, 18 hours.

5-cyanoindole, indole, piperidine-4-one hydrochloride hydrate, 11-aminoundecanioic acid, 6-amino hexanoic acid, cycloheptanone, cyclooctanone, N-carbethoxyphthalimide, ethyl chloroformate, sodium borohydride, N-Bromosuccinimide, triphenylphosphine, hydroxylamine hydrochloride, 2-(2-bromethyl)-1*H*-isoindole-1,3-(2H)-dione **(13a)**, methylamine (2M in THF), deuterated dimethyl sulfoxide, deuturated methanol and deuterated chloroform were obtained from the Sigma Aldrich (Milwake, WI). Biotin-PEG5000-SCM was obtained from Laysan Bio. Inc. (Arab, Al) this was a poly dispersed PEG derivative with an average molecular weight of 5000 Daltons (determined by nmr). Sodium carbonate, magnesium sulfate, hydrochloric acid, and sulfuric acid were obtained from EMD Millipore (Darmstadt, Germany). Hydrazine solution in water (85%, w/w) was obtained from VWR (West Chester, PA). All reagents were used as supplied without further purification. Proton and 13C studies were carried out using a Brucker 300 MHz nmr spectrophotometer. MALDI-TOF mass spectroscopy was performed on a Voyager DE-STR MALDI-TOF instrument.

The hSERT stably transfected HEK293T host cell line was provided by Dr. Randy Blakely’s lab (Vanderbilt University) and was grown in complete DMEM, supplemented with 10% dialyzed FBS and incubated in humidified atmosphere with 5% CO2 at 37°C. The SERT-expressing HEK293T cells were selected in the presence of 400 μg/mL G418.

The SERT protein activity in living HEK293T cells was examined before each experiment by using IDT307, a fluorescent neurotransmitter substrate.15 IDT307 compound is nonfluorescent in solution but fluoresces as the substrate is accumulated into the nucleus membrane, affording real-time evaluation of SERT uptake activity.15 The assay to verify successful SERT expression in HEK293T cells involves the addition of IDT307 directly to the culture media at a final concentration of 5 µM and incubating at 37°C for 10 minutes. Fluorescent images were then acquired immediately after IDT307 addition, and successful transporter expression was evident by an observable increase in intracellular fluorescence (see Figure 3).

The stably transfected hSERT-expressing HEK293T cells were treated using a two-step quantum dot, Qdot® 655 streptavidin conjugate (InvitrogenTM), labeling protocol. Previously before flow cytometry experiments, SERT-expressing HEK293T cells were seeded MatTek (MetTek Corporation) polylysine-coated plates and were allowed to grow for approximately 48hours at 37oC and 5%CO2.Adherent SERT-expressing (1) competition control cells were pre-blocked by exposure to 10 µM of Paroxetine and incubated for 10 minutes at 5% CO2 / 37°C, (2) competition control, wild type, and positive cells were incubated with 0.5µM SERT ligands with inhibitor mixture for 10 min for all sample types, (3) washed several times with DMEM free media and incubated with 1 nM SavQD655/1% BSA (bovine serum albumin) mixture, (4) After the SavQD655 labeling step, HEK cells were washed with DMEM free media to remove any unbound SavQD655 and cells images were acquired on the Zeiss Laser Scannin Microscope 510 inverted confocal microscope. Representative images of hSERT-HEK293T or HEK293T cells exposed to two-step QD labeling protocol using the ligand, IDT357 can be seen in Figure 10.

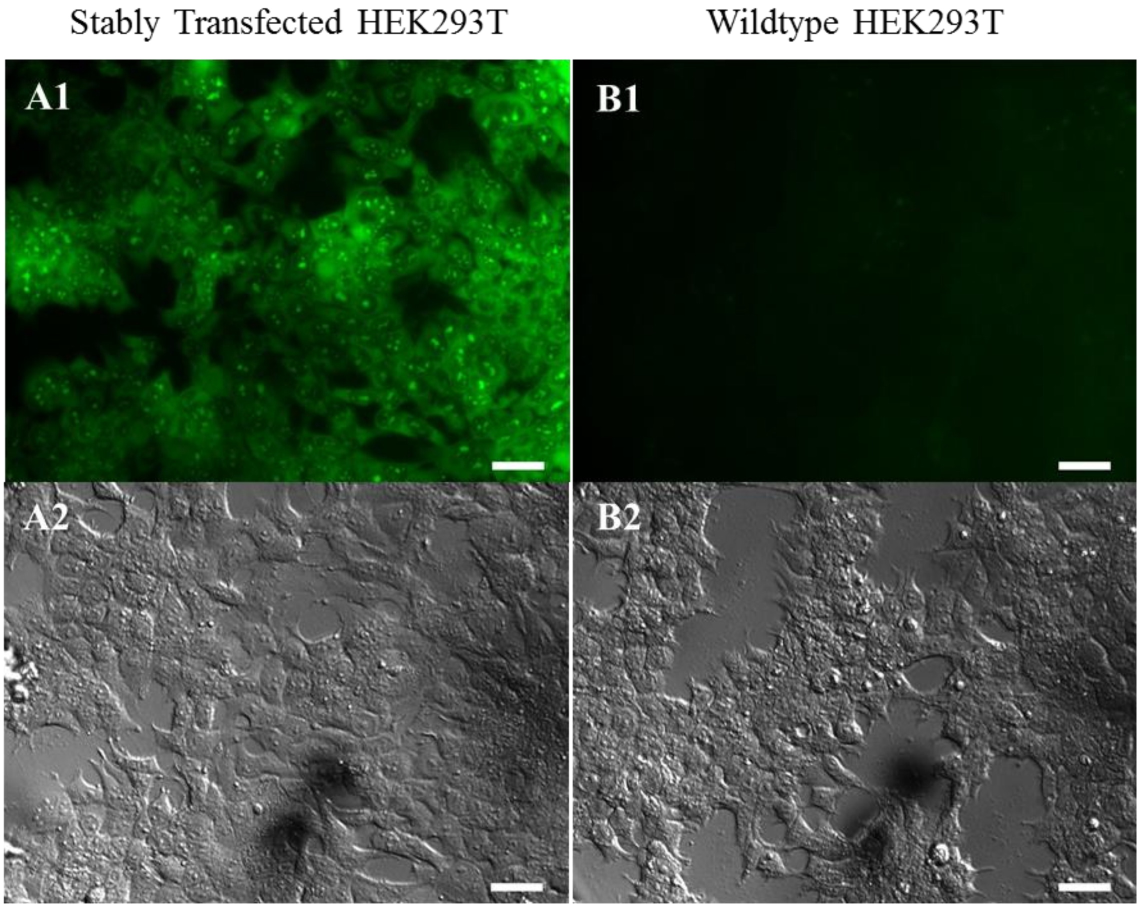


**Figure 9: Images of hSERT-HEK293T or HEK293T cells exposed to flourscent compound, IDT307.** Represenative images are shown for the stably transfected hSERT-HEK293T cells (A1, A2) and HEK293T (wildtype) Cells (B1, B2). Al samples were treated with 5 ɥM IDT307, a fluorescent neurotransmitter substrate, and incubated at 37° for 10 minutes. Flourscent (top and DIC (bottom) images are shown. Calibration Bar = 40um.

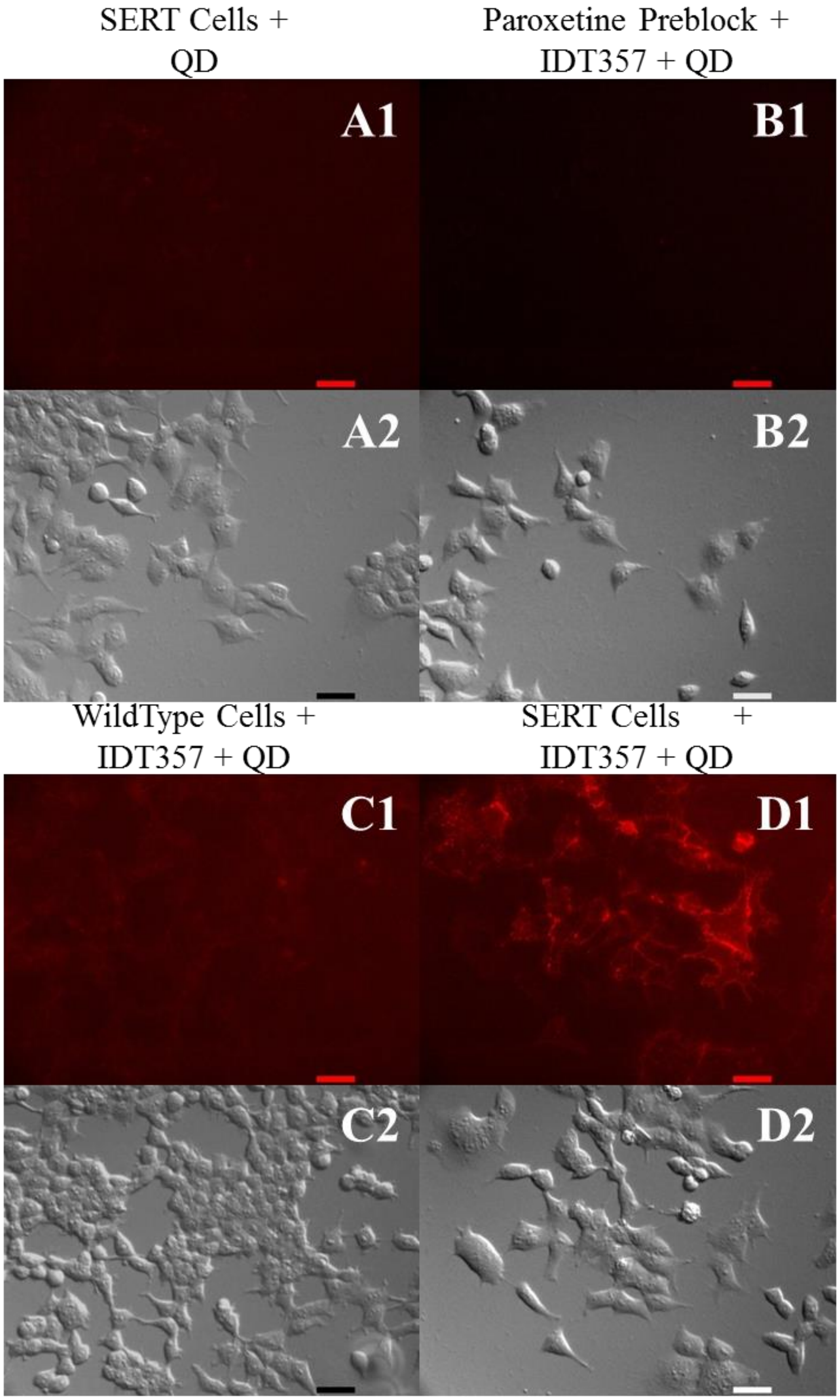


**Figure 10: Representative images of hSERT-HEK293T or HEK293T cells.**
Representative images of hSERT-HEK293T or HEK293T cells exposed to two-step QD labeling protocol using the ligand, IDT357, SavQD655 labeling of hSERT stably expressed in HEK293T cells (A1-D2) and HEK293T (Wildtype) cells (C1, C2). Compared to QD only control (A1, A2), Paroxetine pre-block competition (B1, B2), and wild type cells (C1, C2), the fluorescent intensity increased with hSERT-HEK294T IDT357 positively labeld cells. Flourescent (top) and DIC (bottom) images are shown. Calibration Bar = 40 um.

The stably transfected hSERT-expressing HEK293T cells were treated using a two-step quantum dot, Qdot® 655 streptavidin conjugate (InvitrogenTM), labeling protocol. Previously before flow cytometry experiments, SERT-expressing HEK293T cells were seeded in 24-well polylysine-coated plates (BD Bioscience®) and were allowed to grow for approximately 48hours at 5% CO2 / 37°C.

Next, a flow cytometry-based probe screening protocol was established that would allow multi-well plate screening of both adherent and suspension cell cultures using our ligand-conjugated SavQDs. The screening platform is depicted in Figure 11. Adherent SERT-expressing (1) competition control cells were pre-blocked by exposure to 10 µM of Paroxetine, a potent and selective serotonin reuptake inhibitor (SSRI), and incubated for 10 minutes at 5% CO2 / 37°C, (2) competition control, wildtype, and positive cells were incubated with 0.5µM SERT ligands with inhibitor mixture for 10 min for all sample types, (3) washed several times with DMEM free media and incubated with 1 nM SavQD655/1% BSA (bovine serum albumin) mixture, (4) nonenzymatically dissociated (Cellstripper™), and (5) assayed by flow cytometry. Before each fluorescence measurement are obtained, the voltage of the 5-laser BD LSRII cytometer are adjusted by using Attune™ Performance tracking or calibration beads (Applied Biosystems®).

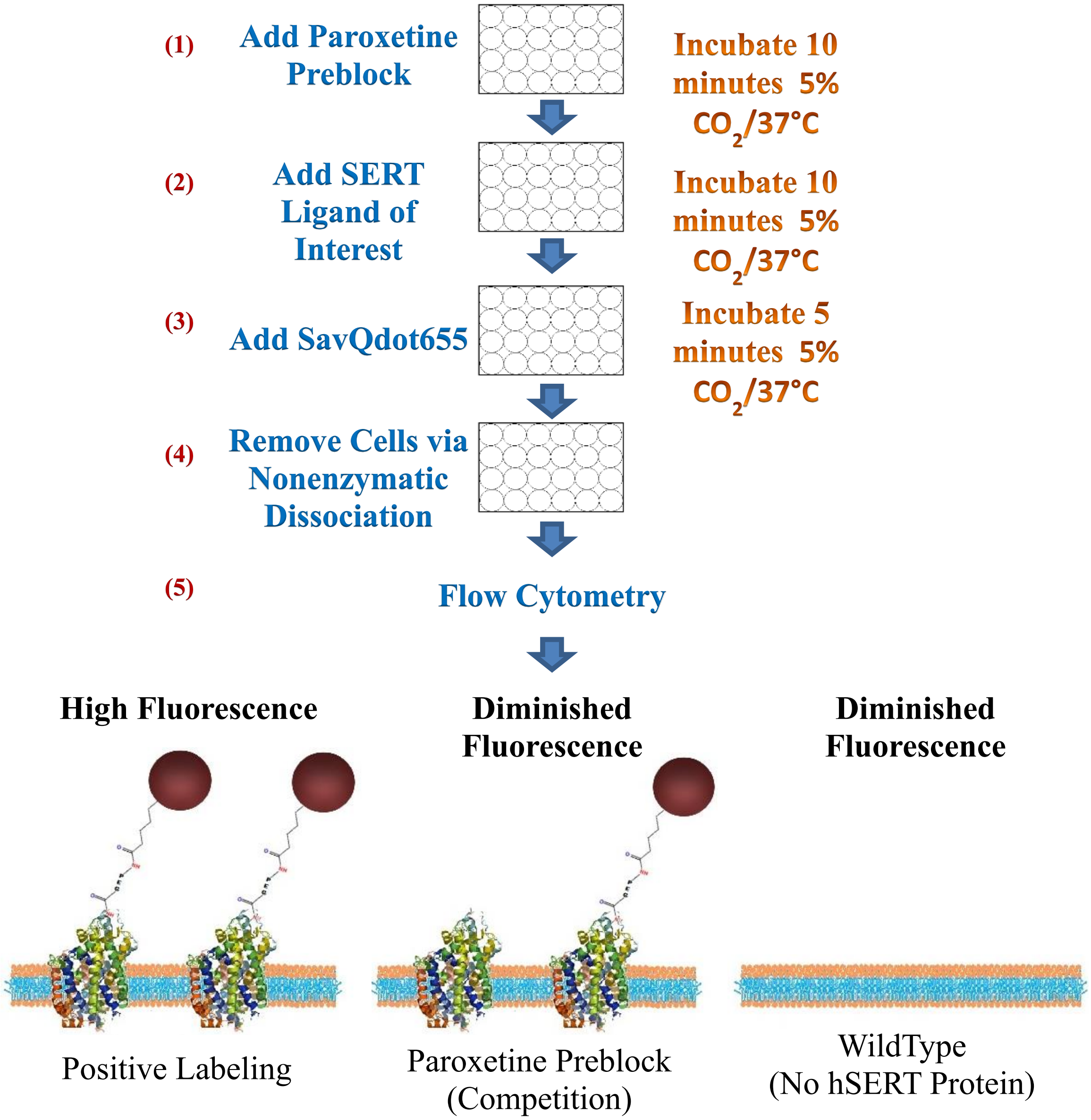

(1)
Add Paroxetine Preblock
Incubate 10 minutes 5% CO2/37°C
(2)
Add SERT Ligand of Interest
Incubate 10 minutes 5% CO2/37°C
(3)
Add SavQdot655
Incubate 5 minutes 5% CO2/37°C
(4)
Remove Cells via Nonenzymatic Dissociation
(5)
Flow Cytometry
High Fluorescence
Diminished Fluorescence
Diminished Fluorescence
Positive Labeling
Paroxetine Preblock (Competition)
WildType (No hSERT Protein)

**Figure 11: Schematic of a QD-based assay**

Schematic of a QD-based assay that investigates the length of alkyl spacer in the linker arm of the organic ligands, using 2, 6, 7, 8, or 11 alkyl spacer lengths, and the effect of binding in ligand models. Adherent SERT-expressing (1) competition control cells were exposed to 10 µM of Paroxetine, a potent and selective serotonin reuptake inhibitor (SSRI), and incubated for 10 minutes at 5% CO2 / 37°C, (2) competition control, wildtype, and positive cells were incubated with 0.5µM SERT probes with inhibitor mixture for 10 min for all sample types, (3) washed several times with DMEM free media and incubated with 1 nM SavQD655/1% BSA (bovine serum albumin) mixture, (4) nonenzymatically dissociated (Cellstripper™), and (5) assayed by flow cytometry.

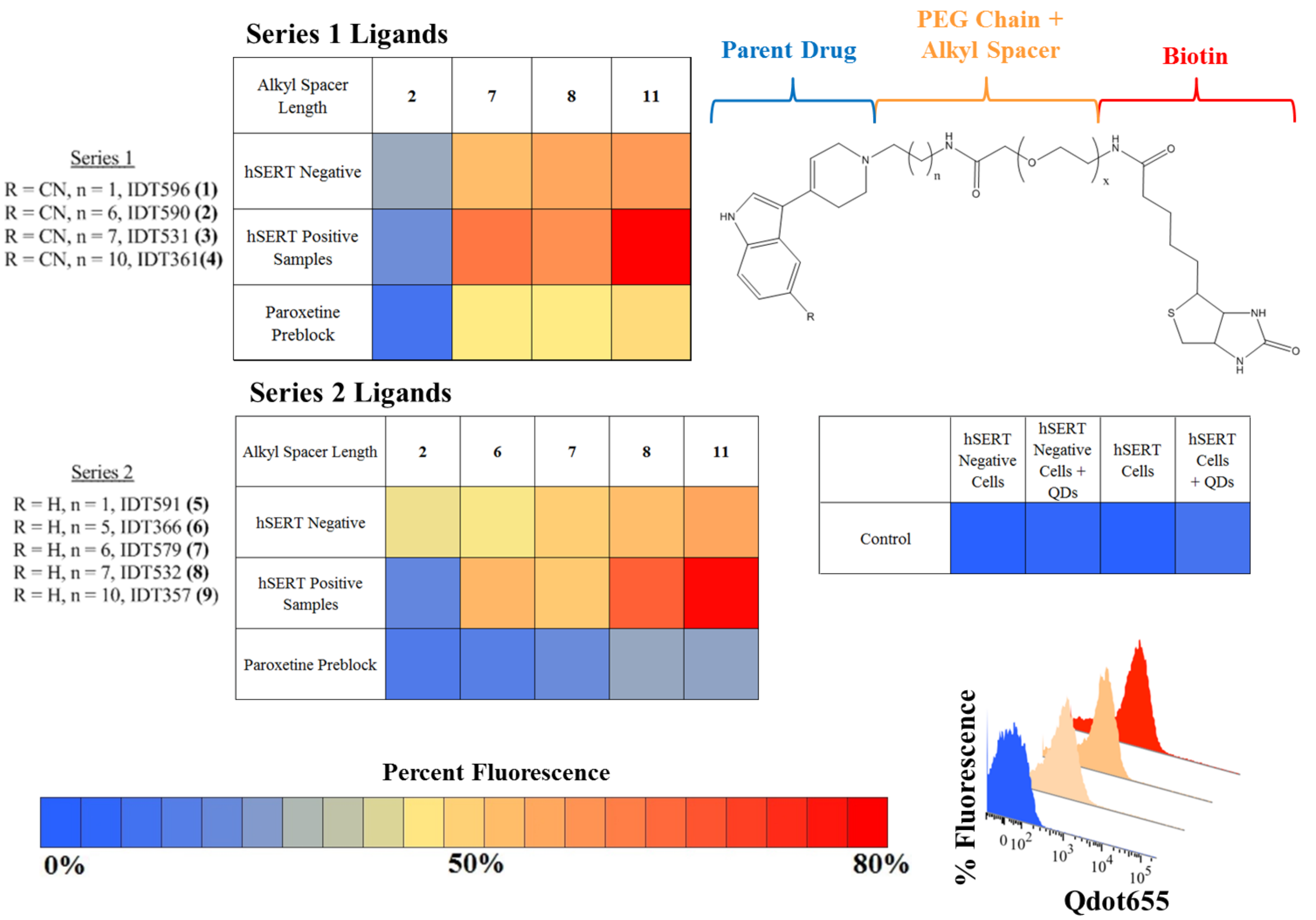
.

### *Results*

The stably transfected hSERT-expressing HEK293T cell line was used as a model system to investigate the effect of the length of alkyl spacer in the linker arm on binding, using 2, 6, 7, 8, or 11 alkyl spacer lengths. The effect length of alkyl spacer in the linker arm on ligand binding were compared in both ligand models containing the 3-(1,2,3,6-tetrahydropyrindin-4-yl)-1H-indole drug end and its cyano derivative ligand, 3-(1,2,3,6-tetrahydropyridin-4-yl)-1H-indole-5-carbonitrile.

The serotonin transporter expression level in living HEK293T cells was examined before each flow cytometry experiment by using IDT307, a fluorescent monoamine neurotransmitter transporter substrate. Representative images of hSERT-HEK293T or HEK293T cells after exposure to IDT307 are shown in Figure 10. Intracellular accumulation of IDT307 by functional membrane SERTs (Figure 10:A1, A2) resulted in a characteristic mitochondria- and nucleoli-associated fluorescence. In contrast, the HEK293T (Wildtype) cells (Figure 10:B1, B2) demonstrated no such fluorescence.

Both hSERT-expressing HEK293T and wildtype cells were subjected to a two-step QD labeling protocol and SERT QD labeling was demonstrated by techniques microscopy and flow cytometry. The representative histogram plots and heat map of the effects of increasing the alkyl spacer length on QD conjugate binding are shown in Figure 6. Comparison of control samples (QD only treated cells, Paroxetine pre-block and wildtype cells); the increase of average percent fluorescent intensity is closely correlated with the alkyl spacer length.



| Alkyl Spacer Length | 2 | 7 | 8 | 11 |
|---|---|---|---|---|
| hSERT Negative | 🟦 | 🟧 | 🟧 | 🟧 |
| hSERT Positive Samples | 🟦 | 🟧 | 🟧 | 🟥 |
| Paroxetine Preblock | 🟦 | 🟨 | 🟨 | 🟨 |

| Alkyl Spacer Length | 2 | 6 | 7 | 8 | 11 |
|---|---|---|---|---|---|
| hSERT Negative | 🟨 | 🟨 | 🟧 | 🟧 | 🟧 |
| hSERT Positive Samples | 🟦 | 🟧 | 🟧 | 🟥 | 🟥 |
| Paroxetine Preblock | 🟦 | 🟦 | 🟦 | 🟦 | 🟦 |

|  | hSERT Negative Cells | hSERT Negative Cells + QDs | hSERT Cells | hSERT Cells + QDs |
|---|---|---|---|---|
| Control | 🟦 | 🟦 | 🟦 | 🟦 |

**Figure 12: Screening of stably transfected hSERT-expressing HEK293T cell line.**
Screening of stably transfected hSERT-expressing HEK293T cell line to investigate the effect of length of alkyl spacer in the linker arm, using 2, 6, 7, 8, or 11 alkyl spacers. The heat map and representative histogram plots of the effects of increasing the alkyl spacer length on QD conjugate binding are shown. The heat map and representative histogram plots were generated using average percent fluorescent intensity values pooled from five independent experiments with triplicate samples of each ligand subty

### 3-(1,2,3,6-tetrahydropyridin-4-yl)-1H-indole-5-carbonitrile Ligands

| | Average | Standard Error of the Mean (SEM) |
|---|---|---|
| SERT Cells | 0.147 | ± 0.026 |
| SERT Cells + QD | 3.086 | ± 0.657 |
| WildType + Cells | 0.165 | ± 0.060 |
| WildType + QD | 1.096 | ± 0.314 |
| SERT Cells + IDT596 | 7.384 | ± 2.207 |
| Paroxetine Preblock + IDT596 | 3.060 | ± 0.565 |
| WildType + IDT596 | 12.061 | ± 2.702 |
| SERT Cells + IDT590 | 48.538 | ± 8.610 |
| Paroxetine Preblock + IDT590 | 22.557 | ± 11.487 |
| WildType + IDT590 | 32.887 | ± 6.449 |
| SERT Cells + IDT531 | 42.866 | ± 8.487 |
| Paroxetine Preblock + IDT531 | 22.386 | ± 8.508 |
| WildType + IDT531 | 37.810 | ± 7.645 |
| SERT Cells + IDT361 | 75.090 | ± 3.783 |
| Paroxetine Preblock + IDT361 | 26.035 | ± 10.383 |
| WildType + IDT361 | 40.731 | ± 7.888 |

### 3-(1,2,3,6-tetrahydropyrindin-4-yl)-1H-indole Ligands

| | Average | Standard Error of the Mean (SEM) |
|---|---|---|
| SERT Cells | 0.147 | ± 0.026 |
| SERT Cells + QD | 3.086 | ± 0.657 |
| WildType + Cells | 0.165 | ± 0.060 |
| WildType + QD | 1.096 | ± 0.314 |
| SERT Cells + IDT591 | 6.230 | ± 0.790 |
| Paroxetine Preblock + IDT591 | 4.295 | ± 0.375 |
| WildType + IDT591 | 20.750 | ± 0.563 |
| SERT Cells + IDT366 | 34.100 | ± 5.000 |
| Paroxetine Preblock + IDT366 | 5.015 | ± 0.455 |
| WildType + IDT366 | 21.825 | ± 0.891 |
| SERT Cells + IDT579 | 30.150 | ± 4.050 |
| Paroxetine Preblock + IDT579 | 6.960 | ± 0.330 |
| WildType + IDT579 | 29.100 | ± 1.578 |
| SERT Cells + IDT532 | 53.750 | ± 4.846 |
| Paroxetine Preblock + IDT532 | 11.555 | ± 4.272 |
| WildType + IDT532 | 32.671 | ± 9.156 |
| SERT Cells + IDT357 | 73.625 | ± 5.951 |
| Paroxetine Preblock + IDT357 | 10.898 | ± 3.165 |
| WildType + IDT357 | 37.936 | ± 9.778 |

**Figure 13: Percent fluorescent intensity values for screened ligands.**
The figure details all of the average percent fluorescent intensity values of each of the ligands screened. SavQD655 labeling of hSERT stably expressed in HEK293T cells and HEK293T (Wildtype) cells.

This correlation can also been seen in the presence of membrane-associated QD fluorescent labeling on the acquired microscopy images (Figure104). SavQD655 labeling of hSERT is demonstrated in stably expressing HEK293T cells (Figure 10:A1 – D2) and HEK293T (Wildtype) cells (Figure 10:C1, C2). Compared to QD only control (Figure 10:A1, A2), Paroxetine pre-block competition (Figure 10:B1, B2), and Wild Type cells (Figure 4:C1, C2), the fluorescent intensity increased with hSERT-HEK293T IDT357 positively labeled cells. A detailed depiction of all of the average percent fluorescent intensity values can be seen in Table and in Figure 14. These data demonstrate that our QD-based approach can be used to investigate the length of alkyl spacer in the linker arm in a living cell assay.

Visualization of the trend between the average percent fluorescent intensity values and the carbon alkyl spacer length can be seen in Figure 8. As the carbon alkyl spacer length increased in both ligand models containing the 3-(1,2,3,6-tetrahydropyrindin-4-yl)-1H-indole drug end and its cyano derivative ligand, 3-(1,2,3,6-tetrahydropyridin-4-yl)-1H-indole-5-carbonitrile, the average percent fluorescent intensities increased. The ideal binding length from RosettaLigand modeling predicted that an alkyl spacer greater than six would provide sufficient labeling.

### *Conclusions and future directions*

These flow cytometry fluorescent intensity results are in good agreement with the trend obtained by the predictions with RosettaLigand and the RosettaScore software. According to the RosettaLigand calculations, as the alkyl spacer length increased in both ligand models, the binding affinity for the SERT protein increased, with the 6-carbon alkyl spacer being the minimum length to maximize binding affinity of the SERT ligand. Based on these findings, one can combine molecular modeling and fluorescence-based assays to improve ligand design and development.

# CHAPTER III

## IMPLEMENTATION OF PARTIAL COVALENT INTERACTIONS SCORING IN ROSETTA

Partial covalent interactions (PCIs) such as hydrogen bonds(Rose and Wolfenden 1993) (Figure 15a), salt bridges (Figure 15a)(Kumar and Nussinov 2002), cation-π (Figure 15b,d)(Dougherty 1996) and π-π (Figure 15c,e)(McGaughey, Gagne et al. 1998) interactions are abundant in proteins and contribute to their stability (Jeffrey and Huang 1991; Nations, Eaton et al. 1991; Kumar, Tsai et al. 2000). PCIs can be defined as an electron deficient hydrogen with an anti-bonding σ* orbital that interacts with the p- and π-orbitals of an electron rich atom. Because of their abundance and contribution to protein stability, computational algorithms have included PCIs in their energy calculations(Mayo, Olafson et al. 1990; Cornell, Cieplak et al. 1995; Kortemme, Morozov et al. 2003; Brooks, Brooks et al. 2009). However, typically to reduce complexity the interacting orbitals are not explicitly modeled but interactions are approximated using atom coordinates. Further, while almost all algorithms include hydrogen bonding interactions, many disregard salt bridges, cation-π, or π-π interactions. In result, PCIs are treated inconsistently in many force fields.

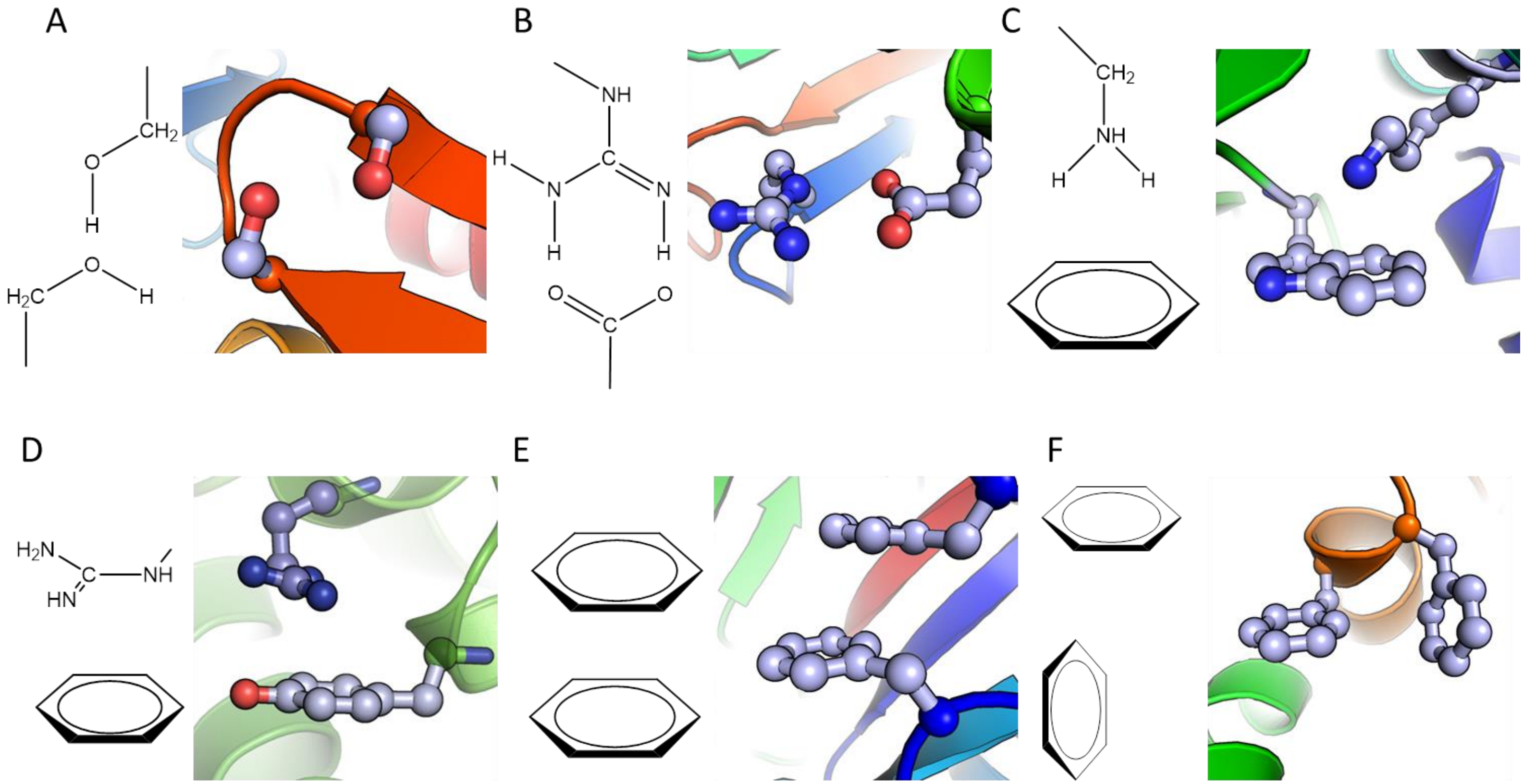


**Figure 14: Schematic representations for partial covalent interactions.**

A) (left) schematic representation of a salt bridge interaction and (right) a salt bridge interaction between R98 and E131 in crystal structure 1wr8. B) (left) schematic of T-stacked cation-π interaction and (right) T-stacked cation-π interaction between W289 and R298 in crystal structure 2oiz. C) (left) schematic of an offset parallel cation-π interaction and (right) an offset parallel cation-π interaction between Y177 and R183 in crystal structure 2bo4. D) (left) schematic of T-stacked π-π interaction and (right) a T-stacked π-π interaction between F71 and F95 in crystal structure 1vph. E) (left) schematic of offset parallel π-π interaction and (right) a π-π interaction between F285 and F348 in crystal structure 1pam.

In order to provide a rapid approximation of PCIs, molecular mechanics or knowledge-based potentials can be utilized. Common applications that contain molecular mechanics potentials are CHARMM(Brooks, III et al. 2009), AMBER(Case, Cheatham et al. 2005), and GROMACS(Hess, Kutzner et al. 2008). These applications utilize physical based laws which describe the motion of atoms under a given force. The force field (energy potential) for molecular mechanics is calculated through $E_p = E_{covalent} + E_{noncovalent}$ where $E_{covalent}$ is the energy associated between bonded atoms and $E_{noncovalent}$ is the energy between non-bonded atoms. A major component of $E_{noncovalent}$, $E_{electrostatic}$, determines the energy contribution for PCIs.

Robust physical based methods for calculating $E_{electrostatic}$, such as the Generalized Born (GB)(Still, Tempczyk et al. 1990) and the Tanford-Kirkwood(Havranek and Harbury 1999) models, approximate PCIs by treating the protein as a low dielectric charged solute embedded in a high dielectric medium. Free energy calculations are performed using a set of analytical series solutions to the Poisson–Boltzmann equation. Although approximations to the Poisson–Boltzmann equation are robust and have been made pair-wise decomposable(Marshall, Vizcarra et al. 2005), they are still computationally expensive to calculate making application difficult for protein design, protein-protein and protein-ligand docking.

In contrast to molecular mechanics, KBPs are based upon the assumption that frequently observed molecular geometries correspond to low energy states(Sippl 1995). They require availability of a large databank of experimentally determined structures – a prerequisite fulfilled for proteins through the protein databank (PDB)(Berman, Westbrook et al. 2000). A set of measurable geometric variables can be extracted from the PDB and compiled into a statistical distribution. According to Boltzmann’s principle, energies and probability densities are related quantities and can be used to rapidly assess protein models. The general definition for the energy of a KBP is expressed via $E(r) = -kTln[(\int(r))]$ where *r* is a measured quantity between two atoms, *E(r)* is the energy associated from the measured quantity, *k* is Boltzmann’s constant, *T* is the absolute temperature, and $\int(r)$ is the probability density.

The Rosetta scoring function utilizes a series of score terms derived from KBP(Rohl, Strauss et al. 2004). The diverse nature of the Rosetta Software Suite spans many biological applications and includes RosettaLigand(Lemmon and Meiler 2012), RosettaDesign(Kuhlman, Dantas et al. 2003), RosettaFolding(Simons, Kooperberg et al. 1997), RosettaRNA(Das and Baker 2007), and RosettaDNA(Morozov, Havranek et al. 2005). The success of Rosetta is contributed to a robust conformational sampling algorithm and to the common shared scoring function between applications. PCIs are determined in Rosetta through a hydrogen bond KBP and a ‘pair potential’.

Kortemme(Kortemme, Morozov et al. 2003) et al., created a hydrogen bond KBP based upon frequently observed geometries between a polar hydrogen and an acceptor atom. Geometric parameters measured were a distance between the hydrogen of the donor atom and the acceptor atom ($\delta_{HA}$), an angle at the acceptor atom ($\psi$), an angle at the hydrogen atom ($\Theta$), and a dihedral angle between the acceptor-acceptor base bond

(*X*) (Figure 16A). The energy potential was then incorporated into the Rosetta Software Suite. With addition of the hydrogen bond KBP, protein design, protein-protein docking, and protein interface design showed significant improvements.

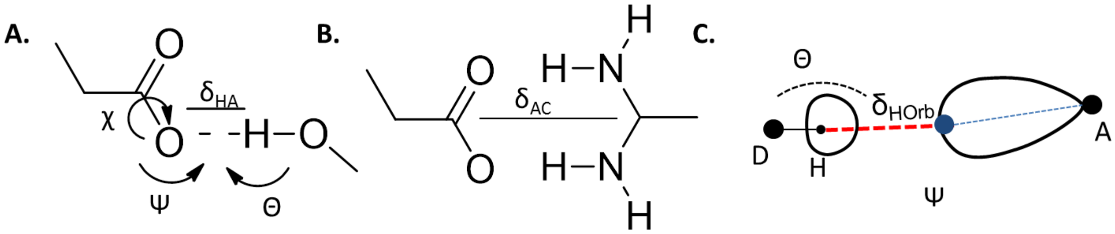


**Figure 15: Geometric parameters for partial covalent interactions in the Rosetta score function.**

Schematic representation of geometric definitions for derivation of KBP for (a) hydrogen bonds, (b) pair potential, and (c) orbitals. A) δHA distance between the acceptor and hydrogen atom, χ torsion angle between the base acceptor-acceptor-hydorgen-donor atom, Ψ angle between the base acceptor-acceptor-hydrogen, and Θ angle between the acceptor – hydrogen –donor. B) δAC distance between the action center for two polar residues. C) δHOrb, distance between the orbital and hydrogen atom, Ψ angle between the acceptor – orbital – hydrogen, and Θ the angle between the donor – hydrogen – orbital.

To better capture non-hydrogen bond PCIs in the Rosetta score function, Kuhlman(Kuhlman, Dantas et al. 2003) et al. introduced a score term which aims to capture the interaction between a pair of polar residues. The 'pair potential' measures the distance between the center of charge for two polar residues (Figure 16B). Unlike the hydrogen bond KBP, the pair potential is not environment dependent. Taken together the pair potential and hydrogen bond KBP are the primary components of the standard Rosetta score function for scoring PCIs. These terms do not, however, describe π-π or cation-π PCIS, although attempts within Rosetta have been developed(Havranek, Duarte et al. 2004; Misura, Morozov et al. 2004).

We argue that the current implementation of PCIs in Rosetta is 1) incomplete as important interactions such as cation- π and π-π interactions are currently excluded from the score function, 2) inconsistent as the level of detail between the pair and hydrogen bond KBP is vastly different (simple distance of charged-atoms for the first, complex geometrical arrangements of four atoms for the second), and 3) convoluted as salt bridges are evaluated by a combination of the pair and hydrogen bond KBP.

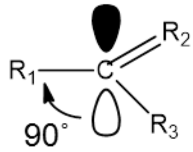
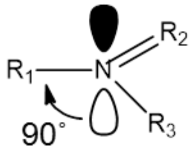
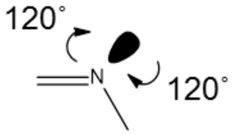
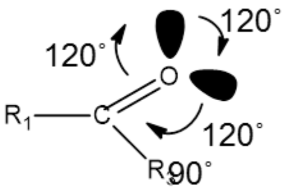
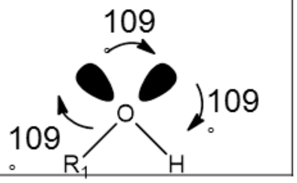
We propose a holistic, orbital-centric, extendable KBP that captures all PCIs. We demonstrate that this new scoring function replaces the current model of hydrogen bonds and the pair potential in Rosetta without a reduction in performance. In fact, the new scoring function recapitulates more accurate geometries for hydrogen bonds, salt bridges, cation-π and π-π interactions and improves performance of several basic modeling tasks within Rosetta such as energy minimization and protein design.

### *Orbital Placement on Rosetta Atom Types Results in Five Orbital Classes*

For generation of the KBP, explicit placement of orbitals is needed. Molecular orbital theory provides a robust approach for modeling orbitals on atoms. However, calculation of molecular orbitals is computationally expensive as each molecular orbital is influenced by the geometry of the overall molecule. Further, geometric constraints required to derive a knowledge based potential from molecular orbitals are difficult to define as there is not a single point to measure angles and distances. A rapid approximation of orbital geometry is available using the Valence Shell Electron Pair Repulsion (VSEPR) theory(Gillespi.Rj 1970). VSEPR theory states that valence shell electron pairs around an atom repel each other and adopt a geometrical arrangement that minimizes this repulsion. Arrangement of orbitals and bonds are determined through the steric number of an atom. The steric number for each atom is identified by the number of bonded atoms and lone pair electrons to that atom. VSEPR theory is leveraged for placement of orbitals in Rosetta by constraining the possible arrangement of orbitals on an atom through the atom's steric number.

Because the distance between the nuclei of an atom and its accompanying electron is computationally expensive to calculate, the most probable distance of an electron to the nucleus of an atom was used for placement of the orbital, the Bohr radius. The geometric angle of the orbital in regards to the surrounding bonded atoms is determined through VSEPR theory in accordance to the steric number. For this purpose, each atom type in Rosetta that is capable of containing a π- or a lone pair p-orbital was assigned a steric number. A total of five orbital classes were created based on the element, C, N, O, the occupancy of electrons, and the hybridization state of the atom, bonded atoms, and orbitals: tetrahedral (Te), Triganol (Tr) (Table 1). Through creation of orbital classes, a precise definition of the atom and occupying orbitals is defined, thus allowing for ease of extendibility to non-standard atom types seen in small-molecules, DNA/RNA nucleotides, and non-canonical amino acids. Specific to hydrogen atoms orbital placement, the Bohr radius is negligible, therefor, the anti-bonding σ* orbital from a hydrogen was not modeled. Instead, the nuclease atomic coordinates of the hydrogen atom was used as the σ* orbital location.

**Table 1: List of orbital classes.**
List of orbital classes, the driving interaction, interaction type, and score term associated with the PCI score function.

| Orbital Class | Rosetta Atom Type | Driving Interaction | Score Term | Pictorial |
|---|---|---|---|---|
| C_TrTrTrPi | aroC | hpol | pci_cation_pi |  |
| | | haro | pci_pi_pi | |
| | | C_TrTrTrPi | pci_pi_pi | |
| N_TrTrTrPi2 | Narg, Ntrp | C_TrTrTrPi | pci_cation_pi |  |
| N_Tr2TrTrPi | Nhis | hpol | pci_salt_bridge |  |
| | | haro | pci_cation_pi | |
| O_Tr2Tr2TrPi | OOC, ONH2, Obb | hpol | pci_salt_bridge |  |
| O_Te2Te2TeTe | OH | hpol | pci_hbond |  |

## *Geometric Parameters for PCIs Include One Distance and Two Angles*

In selecting the geometrical parameters for describing PCIs we assumed that 1) there is an optimal, short orbital-hydrogen or orbital-orbital distance that drives the interaction and 2) that a straight line acceptor – orbital – hydrogen – donor is formed. Figure 2C illustrates the three geometric measurements used in creating the energy function. 1) a distance ($\delta_{HOrb}$) between the orbital and hydrogen, 2) the angle Ψ between the acceptor – orbital – hydrogen (AOH), and Θ angle between the donor – hydrogen – orbital (DHO) angle. When the distance between two π-orbitals were shorter than an orbital-hydrogen pair (specific for weight sets *pci_cation_pi, pci_pi_pi*), a distance ($\delta_{OrbOrb}$) along with two angles, Ψ, the acceptor-orbital-orbital (AOO) and Θ, the donor-orbital-orbital (DOO) were measured (Figure 16D). Inclusion of a direct measurement between angles AOH, AOO, DHO, and DOO involving the orbital removes the need to indirectly calculate the relationship between the acceptor, hydrogen, and donor using torsion angles between four atoms as seen with the hydrogen bond potential(Kortemme, Morozov et al. 2003).

### *Derivation of Knowledge-Based Potential*

For derivation of the knowledge-based potential, the RosettaFeatures reporter(Leaver-Fay, O&apos;Meara et al. 2013) was used to obtain distances and angles representative of PCIs in the top8000 dataset (see supplemental for command lines). The inverse Boltzmann relation was used to convert the propensity of $\delta_{HOrb}$, cos(Ψ) and $\delta_{HOrb}$, cos(Θ) into an energy: $E(X) = -RTln(P_{observed}(X)/P_{background}(X))$ where E(X) is the energy for X, the feature observed, R the gas constant, T the temperature and $P_{observed}(X)$ the probability of the feature observed and $P_{background}(X)$ is the probability of the given observation seen by chance. The total energy for a given PCI is determined by the summation of $E(PCI|\delta_{HOrb}, cos(\Psi))$ and $E(PCI \mid \delta_{HOrb}, cos(\Theta))$ where PCI is the partial covalent interaction being modeled, $\delta_{HOrb}$, cos(Ψ) is the distance and acceptor – orbital – hydrogen (AOH) angle and $\delta_{HOrb}$, cos(Θ)) is the distance and donor – hydrogen – orbital (DHO) angle. For interactions between two π-orbitals, the summation of

E(PCI|$\delta_{OrbOrb}$, cos(Ψ)) and E(PCI | $\delta_{OrbOrb}$, cos(Θ)) where E is the energy for the partial covalent interaction (PCI) for cation-π or π-π interactions modeled and $\delta_{OrbOrb}$, cos(Ψ) is the distance between two π-orbitals and cos(Ψ) is the acceptor – orbital – orbital angle and cos(Θ) is the donor – orbital – orbital angle.

PCI distributions were determined by the shortest distance ($\delta_{HOrb}$ or $\delta_{OrbOrb}$) between two separate participating residues' hydrogen atom or orbital and orbital pairs. Once the shortest distance was determined, the cosine of both Ψ and Θ were determined. By taking the cosine Ψ and Θ, a normalization is applied to each angle to account for bias in observing a given angle by chance. Because the strength of PCIs is reliant upon both angle and distance geometric parameters, two-dimensional histograms were created for all pairs of interactions ($\delta_{HOrb}$, cos(Ψ) and $\delta_{HOrb}$, cos(Θ) and $\delta_{OrbOrb}$, cos(Ψ) and $\delta_{OrbOrb}$, cos(Θ))) with bin fractions set to 0.1 Å for distances $\delta_{HOrb}$ and $\delta_{OrbOrb}$ 0.05 for cos(Ψ) and cos(Θ) with the exception of $\delta_{HOrb}$, cos(Ψ) and $\delta_{HOrb}$, cos(Θ) for aromatic hydrogens (haro) where the bin fractions for cos(Ψ) and cos(Θ) was set to 0.025. The expected background probabilities for $\delta_{HOrb}$ and $\delta_{OrbOrb}$ were determined by dividing each bin fraction by the squared distance (r^2) for each observed bin fraction. Further, pseudo counts were added to each bin fraction to ensure that all non-observed geometries are given an equally high penalty.

Although the shortest distance for PCIs was used to determine bin fractions, a bicubic interpolation of all distance/angle pairs for every PCI was used to determine the energy associated with a PCI between two residues. This has two direct effects, i) the energy function becomes a continuous, differentiable function and ii) bicubic interpolation ensures that $\delta_{HOrb}$, cos(Ψ) and $\delta_{HOrb}$, cos(Θ) remain tightly coupled during minimization.

### *Optimization of the Orbital Score Function*

The overall energy score *E* computed by Rosetta is a linear combination of weighted scoring terms. The base score function in Rosetta is composed of van der Waals interactions (*fa_atr*), an inter side chain (*fa_rep*) and intra side chain repulsive term (*fa_intra_rep*), an implicit solvation model (*fa_sol*), a hydrogen bond term for side chain – side chain (*hbond_sc*), backbone – backbone (*hbond_sr_bb, hbond_lr_bb*) and backbone – side chain (*hbond_bb_sc*), a backbone – dependent rotamer probability (*fa_dun*), the probability of an amino acid given *phi* and *psi* angles (*p_aa_pp*), the probability of two polar residues being within a certain distance of each other (*fa_pair),* and reference energies to resemble the quantity of residues seen in any given protein (*ref*):

**Equation 4:**

$$E = W_{atr}E_{atr} + W_{rep}E_{rep} + W_{intra\_rep}E_{intra\_rep} + W_{sol}E_{sol} + W_{hbond\_sc}E_{hbond\_sc} + W_{hbond\_sr\_bb}E_{hbond\_sr\_bb} + W_{hbond\_lr\_bb}E_{hbond\_lr\_bb} + W_{hbond\_bb\_sc}E_{hbond\_bb\_sc} + W_{dun}E_{dun} + W_{p\_aa\_pp}E_{p\_aa\_pp} + W_{pair}E_{pair} + W_{ref}E_{ref}$$

The relative weights for all scoring terms were optimized by redesigning proteins in a dataset of high resolution experimental structures to maximize the probability of recovering the native amino acid at each position in the protein(Kuhlman and Baker 2000; Kortemme, Morozov et al. 2003). Modification, addition, or removal of scoring terms therefore requires adjustment of the individual weights.

To allow different weights for different PCIs, the orbital score function was split into four distinct score terms for weight optimization (see previous sections). An advantage of KBP is the ability to implicitly capture interactions that are difficult to model. A danger is to double-count certain interactions – for example the pair potential and the hydrogen bond potential in the traditional energy function collectively apply to salt bridges. Consequently, with the introduction of the PCI score terms we removed all side chain hydrogen bonding interactions and the pair potential:

**Equation 5**

$$E = W_{\text{atr}}E_{\text{atr}} + W_{\text{rep}}E_{\text{rep}} + W_{\text{intra\_rep}}E_{\text{intra\_rep}} + W_{\text{sol}}E_{\text{sol}} + W_{\text{hbond\_sr\_bb}}E_{\text{hbond\_sr\_bb}} + W_{\text{hbond\_lr\_bb}}E_{\text{hbond\_lr\_bb}} + W_{\text{dun}}E_{\text{dun}} + W_{\text{p\_aa\_pp}}E_{\text{p\_aa\_pp}} + + W_{\text{ref}}E_{\text{ref}} + W_{\text{pci\_pi\_pi}}E_{\text{pci\_pi\_pi}} + W_{\text{pci\_salt\_bridge}}E_{\text{pci\_salt\_bridge}} + W_{\text{pci\_hbond}}E_{\text{pci\_hbond}} + W_{\text{pci\_cation\_pi}}E_{\text{pci\_cation\_pi}} + W_{\textit{orbitals\_hpol\_bb}}\mathit{E}_{\textit{orbitals\_hpol\_bb}}$$

An iterative approach was used to optimize weights for the new PCI score function. Because PCIs counter-balance the cost of de-solvating polar residues, the weight for the solvation potential needed to be adjusted. The particle swarm optimization algorithm, OptE, was used to optimize the weight for all orbital score terms, the solvation term, and the reference energies. Initial weights for the optimized terms were then varied by 0.05 to 0.3 while the reference energies were allowed to be optimized by OptE. Resulting weight sets were benchmark on a series of tasks described below to arrive at a single weight set with optimal results.

## C++ class structure

An important aspect of all computational projects is use of an object oriented programming language to perform a given task. To this extent, C++ was used within the Rosetta Software Suite framework to generate statistics for the partial covalent interactions KBP, creation of properties and placement of orbitals, and scoring of residues with the new partial covalent interactions score function. In the next sections, the data structures and general layout of the written code is described.

### *OrbitalType Class Data Structure*

Initially, to reduce the amount of code added to Rosetta, orbitals were placed onto residues as virtual atoms. By doing so, orbitals were devoid of properties and were updated whenever the residue conformation changed; however, the overall speed of Rosetta drastically slowed, regardless of which score function used, standard or the partial covalent interactions score function. In order to keep the standard Rosetta score function from experiencing a decrease in speed and allow for extendibility of orbitals by inclusion of orbital properties, a separate class modeled after the AtomType and MMAtomType classes for orbitals were created.

The OrbitalType class contains the "chemical" information for orbitals and not the xyz coordinates which are managed and updated by conformation/Residue.hh. This does not contain the actual. Orbital properties are assigned by a separate class, OrbitalTypeSet, which reads in a flat text file located in rosetta database (chemical/orbital_type_sets/fa_standard/orbital_properties.txt) which contains the properties and types of orbitals. The orbital type properties include the steric number, the hybridization state of the atom the orbital is attached to, the distance between the orbital and the base atom (Bohr radius), and the orbital class type name. Because the properties are read in through a flat text file, addition, removal, and changing orbital properties requires no programming.

### *Orbital Initial Placement and Update during Conformational Changes*

Placement of orbitals onto residues is done through the AssignOrbitals class. At initialization of Rosetta, the OrbitalType class is initiated which creates the orbital properties that are used for placement of orbitals onto residues. After all residue types, atom types, and orbital types are created, the AssignOrbitals class is initialized if the option –add_orbitals is present. This ensures that all residues with atoms that contain orbitals created through the ResidueType class will be assigned orbitals. A set of logic based upon the orbital type, distance from atom, hybridization state of atom, and bonded atoms to the base atom are used to place orbitals onto each atom that requires an orbital. Atom types which contain orbitals are predetermined through the atom type property, which is assigned through a flat text file (chemical/atom_type_set/fa_standard/atom_properties.txt). This allows for addition and removal of orbitals to atoms based upon the atom type. Once orbital placement onto atoms is determined, the ResidueType class is modified, updated, and finalized to contain a set of internal coordinates for orbitals. In addition, a separate class, conformation/Orbitals.cc is initialized which contains the Cartesian coordinates of the orbitals.

After orbitals have been added, any change to the conformation of a residue or atom results in a change of the Cartesian coordinates of the orbitals. First an observer design pattern is triggered within the class core/conformation/Conformation.cc notifying the conformation/Orbitals.cc class that a change in conformation has occurred. The class is then responsible for updating the Cartesian coordinates for orbitals by retrieving information from the internal coordinates class for Orbitals. Coupling orbital conformation update through the observer design pattern results in updating orbitals whenever a change in conformation is observed.

### *Scoring Function Structure*

Partial covalent interactions can be defined by an anti-bonding σ* orbital from a hydrogen that engages p- and π-orbitals: hydrogen bonds (n→σ*), salt bridges (n→σ*) and cation-π (π→σ*) interactions. Further, PCIs can be classified as two π-orbitals that interact with each other to form a quadropole moment: parallel cation-π and π-π interactions. Both cation-π and π-π interactions can occur in three separate conformations, parallel, offset parallel, and T-stacked(McGaughey, Gagne et al. 1998; Gallivan and Dougherty 1999). For parallel interactions, two π- orbitals interact together while in T-stacked and offset parallel interactions, the σ* orbital interacts with the π-orbital.

Given the above definitions for PCIs, PCIs were separated into four common weight terms based on the driving interactions of the orbitals (Table 1). 1) For PCIs that occur between the σ* orbital of a polar hydrogen (Rosetta atom type 'hpol') and a p- and π-orbital (hydrogen bonds, salt bridges, and cation-π), the weight term *pci_hbond, pci_salt_bridge, and pci_cation_pi (*both offset parallel and T-stacked cation-π interactions) were assigned. 2) Interactions between the polar hydrogen of the polypeptide backbone and a side chain orbital were assigned *orbitals_hpol_bb*. 3) *pci_pi_pi* includes the interaction between the σ* orbital of an aromatic hydrogen (Rosetta atom type 'haro') and a π – orbital (offset parallel and T-stacked π-π). 4) Finally, for interactions between two π- orbitals, the score terms pci_cation_pi (parallel cation – π) and pci_pi_pi(parallel π – π) were assigned. Because cation-π and π-π interactions can occur between both a σ* orbital and two π orbitals, scores were divided by the total set of orbital interactions, thus keeping geometries between residues balanced (parallel, offset parallel, and T-stacked ). Separation of PCI interactions into types of participating interactions allows for explicit control over the strength of the interaction through adjustment of the weight.

Two C++ object oriented classes are responsible for scoring residues that contain orbitals: OrbitalsLookup.cc and OrbitalsScore.cc. When Rosetta is initialized, OrbitalsLookup.cc loads all the KBP data from the folder "database/scoring/score_funcitons/orbitals" and creates a bicubic spline over the data which is stored in a two-dimensional array. When a residue is scored, the OrbitalsScore.cc class is called. Scoring is done on a pair of residue basis. The class first checks that each residue's orbitals Cartesian coordinates are updated. Then, all polar hydrogen-orbital pair, backbone polar hydrogen-orbital pair, aromatic hydrogen-orbital pair, and orbital-orbital pair are identified. To reduce calculation time, only pairs that are within 4.0 Å to each other are evaluated. The cosine of the pairs of angles is evaluated along with the distance of the pairs. The distance and angle are then passed onto the OrbitalsLookup.cc class to identify the energy associated with the pair. The energy for all interacting pairs is then summed for the total energy of that residue. Minimization for each interacting hydrogen-orbital and orbital-orbital pair is done in a similar fashion as described above; however, with one caveat. Because orbitals are represented as virtual anchor points, derivatives obtained for orbitals are placed on the atom bonded to the orbitals. This ensures that the forces dictated by orbitals are still measured.

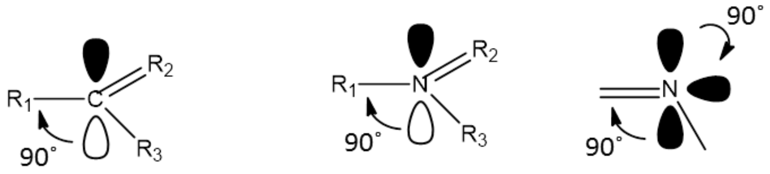
## Versions of the Partial Covalent Interactions Score Function

As with all code development, several iterations of the Partial Covalent Interactions score function were created. As time progressed, the code base expanded and became more object oriented in nature. In the following sections, changes to the score function are explained.

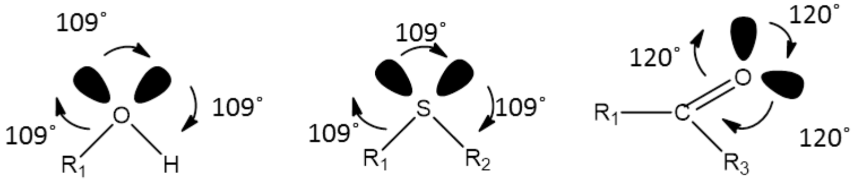
### *Version 1*

In the first version of the PCI score function, orbital location was calculated on the fly as virtual points in space next to an atom which contained orbitals. A total of six classes for orbitals were created for elements carbon, sulfur, nitrogen, and oxygen. Each class was named through a non-descriptive classification: Class I-VI (Table 2).

**Table 2: Old version of orbital classes.**
Old version of orbital classes with hybridization of atom, rosetta atom type, and residue associated with each orbital class.



| | Class I | Class II | Class III |
|---|---|---|---|
| hybridization | sp1 | sp1 | sp2 |
| Rosetta atom types | CNH2, COO, aroC, Cobb | Ntrp, NH2O, Narg, Npro, Nbb | Nhis |
| Residue | N, Q, E, D, Y, H, W, F | W, N, Q, R, P | H |



| | Class IV | Class V | Class VI |
|---|---|---|---|
| hybridization | sp3 | sp3 | sp2 |
| Rosetta atom types | OH | S | ONH2, OOC, OCbb |
| Residue | S, T, Y | C, M | Q, N, D, E |

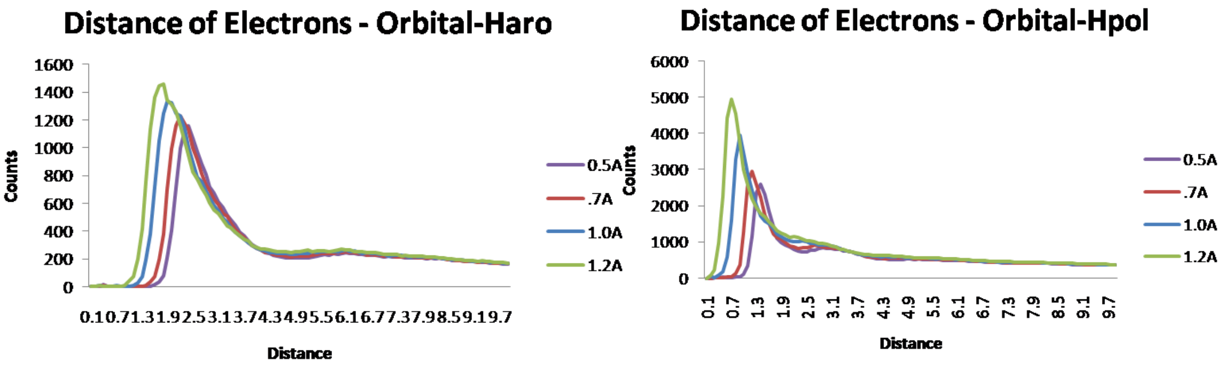


**Figure 16: Distances tested for the orbital - atom bond**

A set of distances measured between the bonded atom and orbital were tested. As seen in Figure17, the distance between the bonded atom to orbital of 0.70 Å showed the most features and was used for all orbital-atom bonded distances. Interactions were measured between a polar hydrogen, an aromatic hydrogen, and an orbital to an orbital with the shortest distance within each group of interactions was used to derive potentials. A dataset from the protein databank (PDB) was used to generate a set of three statistics based solely on a distance measurement between interacting pairs: all polar hydrogen – orbital interactions, all aromatic hydrogen – orbital interactions, and all orbital – orbital interactions regardless of the type of orbital class interacting. Bin fractions for each count were set at 0.10 Å. To account for noise in the dataset, counts were normalized in each bin fraction by dividing each bin fraction by the squared distance. A summary of the statistics are found(Figure 18) and individual peaks are described. Because the counts from each class were summed together before normalization, several inconsistencies occurred. First, specifically for orbitals-hpol interactions (figure of the breakdown of counts, hydrogen bond minimum occurs at 2.8 Å while a salt bridge minimum is longer in distance, ~3.0 Å. Further, cation-π interactions occur at an even longer distance of ~3.5 Å. Summing all counts together followed by normalization results in an imprecise score function that removes important details.

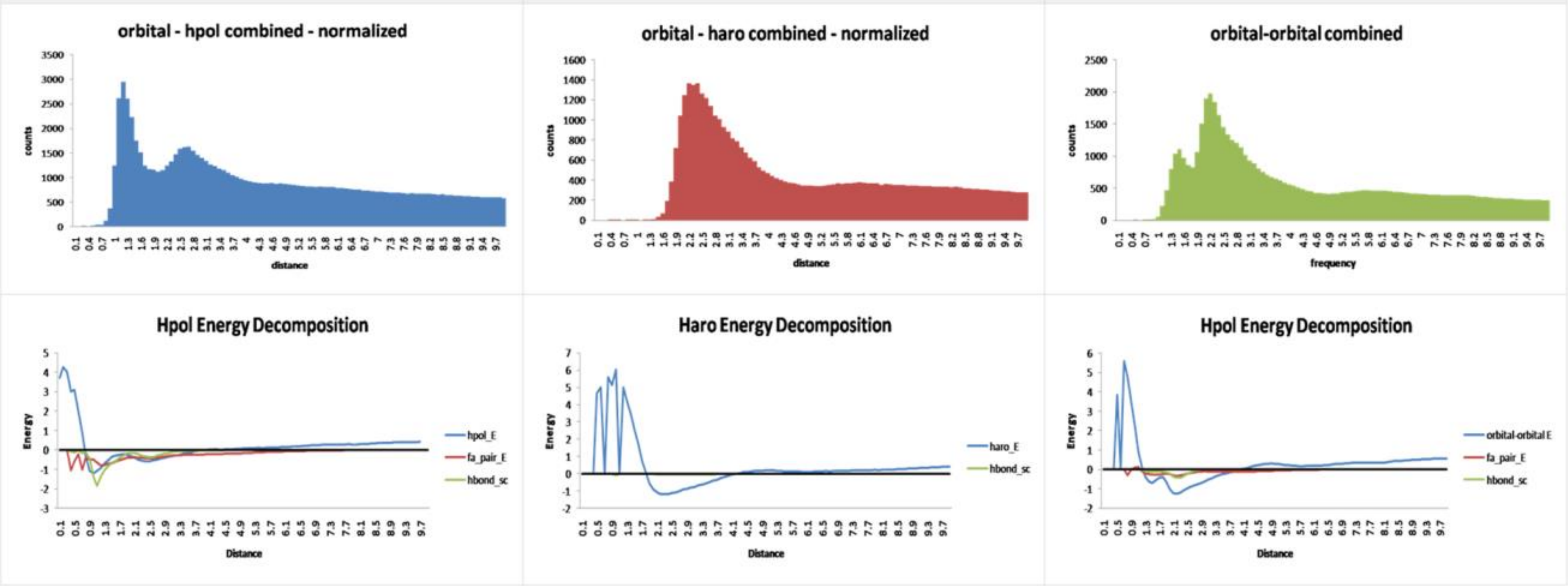


**Figure 17: Normalized counts for orbital-hydrogen and orbital-orbital interactions (top) and energy contributions from score terms (bottom)** The first two peaks in the orbital-hpol graph, 1.1 Å, are attributed to hydrogen bonds and salt bridges. The second peak at 2.5 Å are cation-π interactions between a polar hydrogen (arg/lys) and the Class I orbital (Table 2). For aromatic hydrogens and orbital interactions, the peak at 2.2 Å corresponds to t-stack and offset stack π-π interactions. For orbital-orbital interactions, the biomodal distribution is attributed to both π-π and cation-π interactions.

### *Version 2*

To increase the speed of the score function, each residue parameter file was edited to include information that orbitals should be present on the residue. Further, to account for the improper creation of the KBP, all orbitals were added at once to a residue followed by taking the shortest distance between the orbital and hydrogen atom. The same procedure to normalize counts were used as discussed in Version 1.The resultant KBP is shown in Figure 19. By adding all orbitals at once and taking the shortest distance between an orbital and hydrogen atom, the second peak.

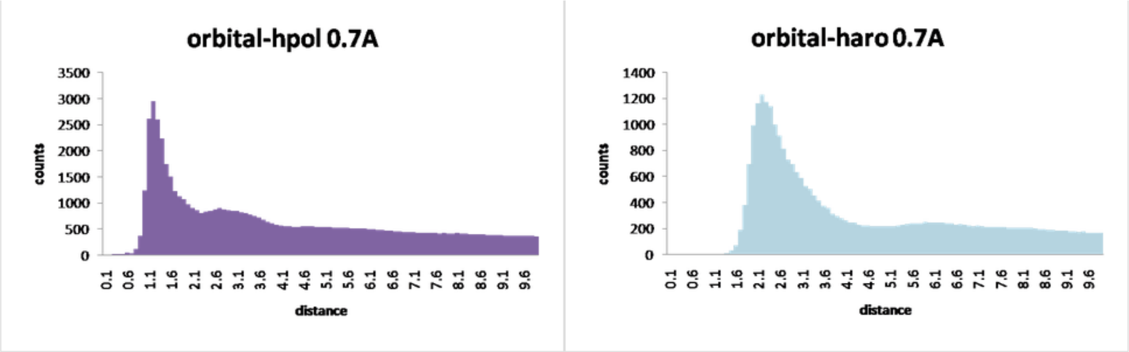


**Figure 18: Renormalization of orbital-hpol and orbital-haro knowledge based potentials.** Removing double counts results in a smoothing of the orbital-hpol knowledge based potential, removing the second peak seen in the first version of the score function (Figure 18)

### *Version 3*

Version 3 of the score function contained the most radical change in both how the orbitals were calculated and generation of the KBP statistics. Orbital distance to the bonded atom switched from a constant 0.70 Å to the Bohr radius, the most probable distance of an electron to an atom, of the bonded atom. Further, to reduce time for calculating orbital placement, code was written to store orbitals in an internal coordinate system. For each canonical residue params file, internal coordinates were added for orbitals along with a specific function in core/chemical/residue_io.cc to parse the internal coordinates. Orbitals were defined as virtual atoms and were updated when the atom tree was changed.

Scoring of residues along with generation of the KBP underwent drastic changes. In addition to a distance, the acceptor – orbital – hydrogen angle were also used to derive the KBP making the potential two-dimensional. Additionally, instead of summing all classes into one potential, each individual class was used, thus making a total of six new KBPs (five for orbitals-hpol and one for orbitals-haro). A summary of Version 3's KBPs are shown in Figure20.

Version 3.1 further changed the way orbitals were added and modified for the score function. Because orbital placement onto residue required modification of params file, non-canonical amino acids and small-molecules used in Rosetta did not have orbitals. A class specifically designed to add orbitals onto all residues was created. From the previous section in this chapter, Orbital Initial Placement and Update during Conformational Changes: After all residue types, atom types, and orbital types are created, the AssignOrbitals class is initialized if the option `–add_orbitals` is present. This ensures that all residues with atoms that contain orbitals created through the ResidueType class will be assigned orbitals. A set of logic based upon the orbital type, distance from atom, hybridization state of atom, and bonded atoms to the base atom are used to place orbitals onto each atom that requires an orbital.

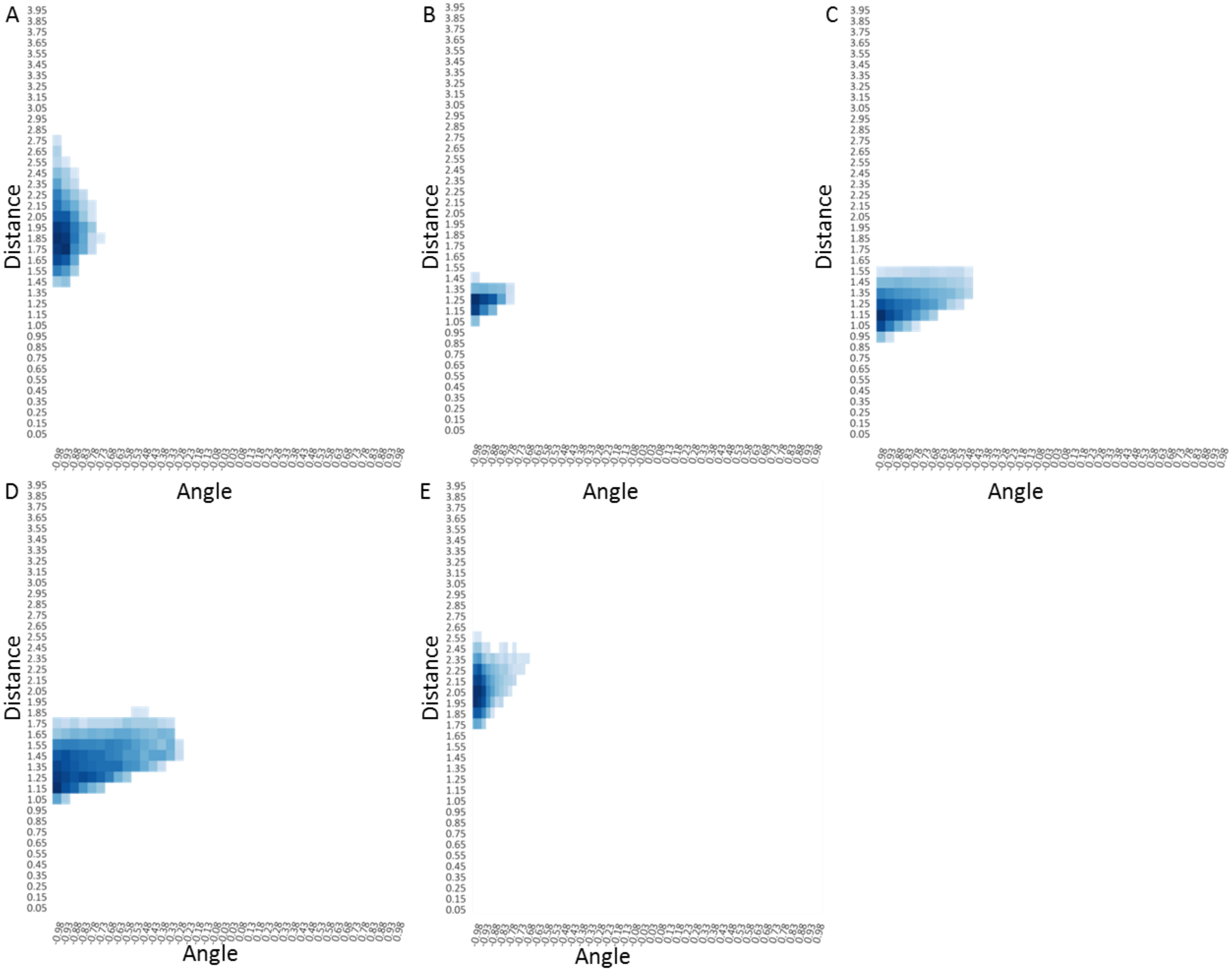


**Figure 19: Knowledge Based potentials with angle and distance component for orbital-hydrogen interactions.**
A) C_TrTrTrPi – hpol B) N_Tr2TrTrPi – hpol C) O_Tr2Tr2TrPi – hpol D) )_Te2Te2TeT – hpol and D) C_TrTrTrPi – hpol. Each minimum occurs at a perfect 180.0 ° angle with varying distances based on orbital type.

### *Version 4*

Version 4 of the score function changes the angles considered during scoring. Initially, only the acceptor – orbital – hydrogen angle was considered for derivation; however, during scoring and minimization, incorrect geometries for sp3 donors to sp2 acceptors were observed in Lambert-Azimuthal equal area projection maps (Figure 21A-E). Minimization occurred at off angle points because the donor – hydrogen – orbital angle was not constrained during minimization. By removing the acceptor – orbital – hydrogen angle and only minimizing with consideration to the donor – hydrogen – orbital angle (Figure 21E), resultant relaxed geometries were similar to the crystal structures. Inclusion of both the acceptor – orbital – hydrogen and the donor – hydrogen – orbital angle during scoring and minimization recapitulated closely the geometries seen in crystal structures (Figure 21F); however, bandings throughout the plot were observed.

As seen in figure 21F, banding located at 30.0° occurred during minimization. This was a direct result from including a degenerative orbital for sp3 and sp2 hybridized atoms Table 2. Removal of the degenerative orbital during minimization creates minimized structures with geometries close to crystal structures; however, banding around the angle of 90.0° remained. Removal of the π-orbital located on the carboxylic acids (OOC atom types) resulted in figure 21E, which is even better recapitulation of crystal structure geometry.

Finally, corrections to the spline removed improper minimization to 180.0° seen in Figure 22F. Once the spline was corrected, the score function properly minimized to 180.0°.

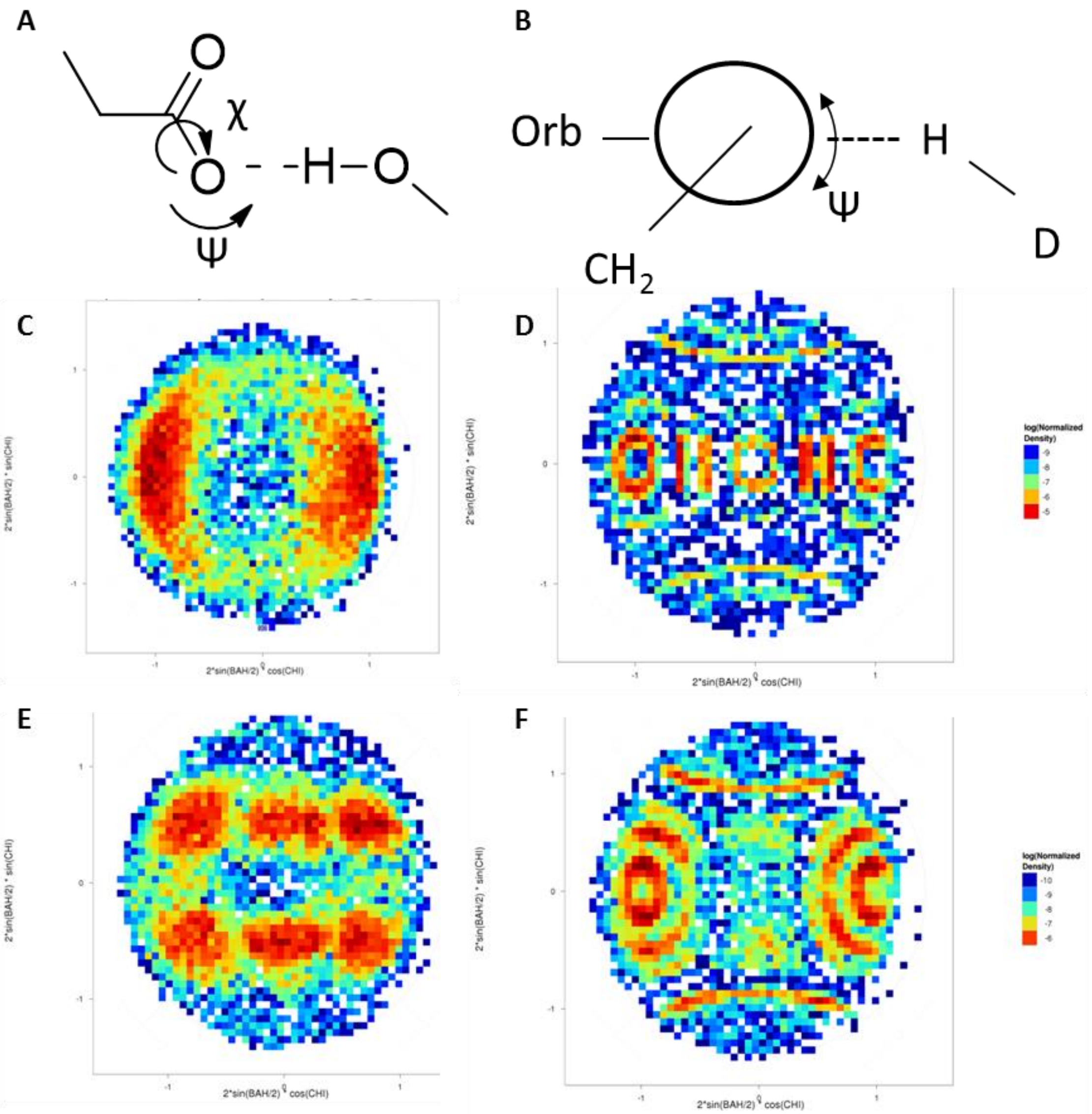


**Figure 20: Geometric parameters used to create the Lambert-Azithumul plot for Version 4 of the PCI score function and resulting plots.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-hydrogen-donor atoms and Ψ the base acceptor – acceptor – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal hydrogen bond structure, the hydrogen placement will be directly located at the orbital. C) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for crystal structures. High density is located directly where the orbital of an sp2 acceptor would be present. D) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using Version 3 of the PCI score function. Density distribution is located near the degenerative orbital of the sp2 oxygen and the π-orbital on the oxygen. E) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using only the donor-hydrogen-orbital angle as constraints. F) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using both the acceptor-orbital-hydrogen and donor-hyrogen-orbital angles as constraints.

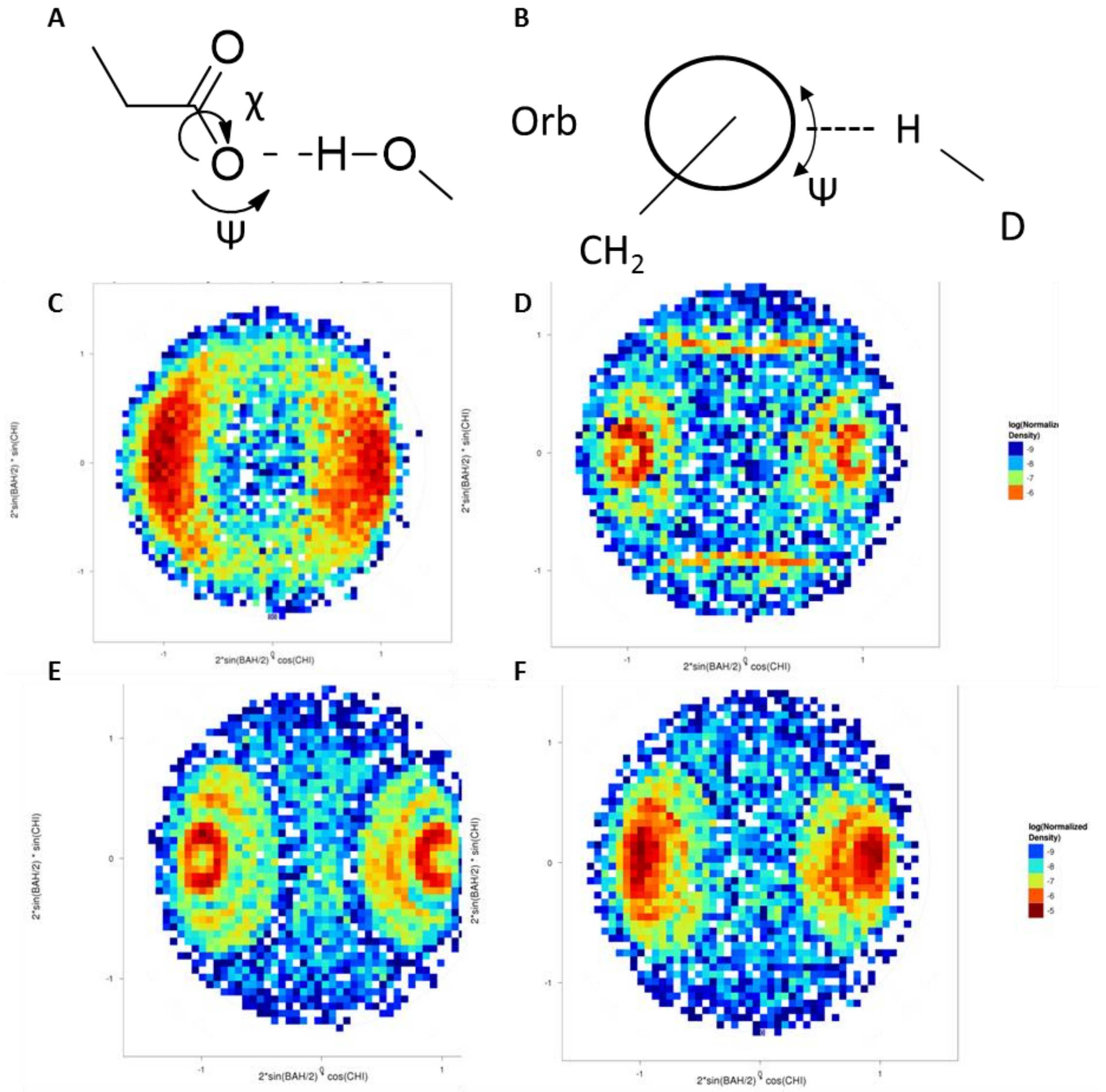
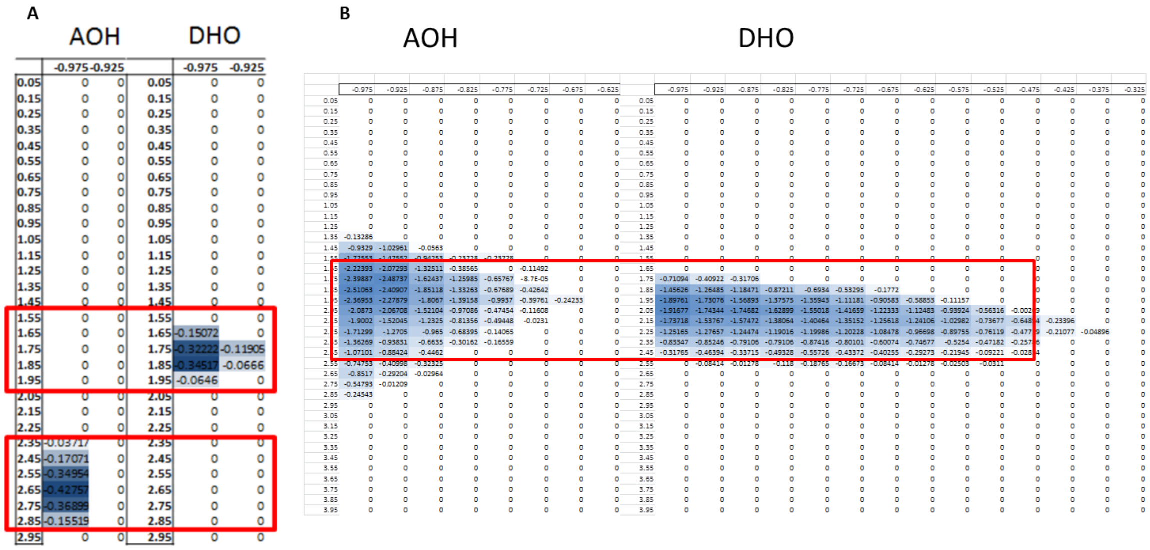


**Figure 21: Further corrections for Version 4 of the PCI score function with geometric parameters used to create the Lambert-Azithumul plot and resulting plots.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-hydrogen-donor atoms and Ψ the base acceptor – acceptor – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal hydrogen bond structure, the hydrogen placement will be directly located at the orbital. C) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for crystal structures. High density is located directly where the orbital of an sp2 acceptor would be present. D) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using both AOH and DHO constraints and removal of the degenerative orbital for sp2 and sp3 atoms. Removal of the degenerative orbital removes off angle minimization seen at 30.0° angles. E) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using both AOH and DHO constraints. Removal of the π-orbital on the sp2 oxygens results in close recapitulation of crystal structures. F) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using both AOH and DHO constraints and fixed bicubic spline interpolation. Crystal like geometries were obtained.

Version 4.1 includes better normalization of the KBP and correcting orbital placement during minimization of the protein. Initially, KBP were constructed without pseudo counts; however, this resulted in an inconsistent distance minimum between the acceptor – orbital – hydrogen and donor – hydrogen – orbital angles for each orbital class (Figure 23A). Adding pseudo counts to each bin fraction resulted in the same distance minimum between the acceptor – orbital –hydrogen and donor – hydrogen – orbital angle for each orbital class Figure 23B.


A

AOH | DHO

|  | -0.975 | -0.925 |  | -0.975 | -0.925 |
|---|---|---|---|---|---|
| 0.05 | 0 | 0 | 0.05 | 0 | 0 |
| 0.15 | 0 | 0 | 0.15 | 0 | 0 |
| 0.25 | 0 | 0 | 0.25 | 0 | 0 |
| 0.35 | 0 | 0 | 0.35 | 0 | 0 |
| 0.45 | 0 | 0 | 0.45 | 0 | 0 |
| 0.55 | 0 | 0 | 0.55 | 0 | 0 |
| 0.65 | 0 | 0 | 0.65 | 0 | 0 |
| 0.75 | 0 | 0 | 0.75 | 0 | 0 |
| 0.85 | 0 | 0 | 0.85 | 0 | 0 |
| 0.95 | 0 | 0 | 0.95 | 0 | 0 |
| 1.05 | 0 | 0 | 1.05 | 0 | 0 |
| 1.15 | 0 | 0 | 1.15 | 0 | 0 |
| 1.25 | 0 | 0 | 1.25 | 0 | 0 |
| 1.35 | 0 | 0 | 1.35 | 0 | 0 |
| 1.45 | 0 | 0 | 1.45 | 0 | 0 |
| 1.55 | 0 | 0 | 1.55 | 0 | 0 |
| 1.65 | 0 | 0 | 1.65 | -0.15072 | 0 |
| 1.75 | 0 | 0 | 1.75 | -0.32222 | -0.11905 |
| 1.85 | 0 | 0 | 1.85 | -0.34517 | -0.0666 |
| 1.95 | 0 | 0 | 1.95 | -0.0646 | 0 |
| 2.05 | 0 | 0 | 2.05 | 0 | 0 |
| 2.15 | 0 | 0 | 2.15 | 0 | 0 |
| 2.25 | 0 | 0 | 2.25 | 0 | 0 |
| 2.35 | -0.03717 | 0 | 2.35 | 0 | 0 |
| 2.45 | -0.17071 | 0 | 2.45 | 0 | 0 |
| 2.55 | -0.34954 | 0 | 2.55 | 0 | 0 |
| 2.65 | -0.42757 | 0 | 2.65 | 0 | 0 |
| 2.75 | -0.36899 | 0 | 2.75 | 0 | 0 |
| 2.85 | -0.15519 | 0 | 2.85 | 0 | 0 |
| 2.95 | 0 | 0 | 2.95 | 0 | 0 |

B

AOH

|  | -0.975 | -0.925 | -0.875 | -0.825 | -0.775 | -0.725 | -0.675 | -0.625 |
|---|---|---|---|---|---|---|---|---|
| 0.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.35 | -0.13286 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.45 | -0.9329 | -1.02961 | -0.0563 | 0 | 0 | 0 | 0 | 0 |
| 1.55 | -1.72553 | -1.47552 | -0.94753 | -0.73728 | -0.73728 | 0 | 0 | 0 |
| 1.65 | -2.22393 | -2.07293 | -1.32511 | -0.38565 | 0 | -0.11492 | 0 | 0 |
| 1.75 | -2.39887 | -2.48737 | -1.62437 | -1.25985 | -0.65767 | -8.7E-05 | 0 | 0 |
| 1.85 | -2.51063 | -2.40907 | -1.85118 | -1.33263 | -0.67689 | -0.42642 | 0 | 0 |
| 1.95 | -2.36953 | -2.27879 | -1.8067 | -1.39158 | -0.9937 | -0.39761 | -0.24233 | 0 |
| 2.05 | -2.0873 | -2.06708 | -1.52104 | -0.97086 | -0.47454 | -0.11608 | 0 | 0 |
| 2.15 | -1.9002 | -1.52045 | -1.2325 | -0.81356 | -0.49448 | -0.0231 | 0 | 0 |
| 2.25 | -1.71299 | -1.2705 | -0.965 | -0.68395 | -0.14065 | 0 | 0 | 0 |
| 2.35 | -1.36269 | -0.93831 | -0.6635 | -0.30162 | -0.16559 | 0 | 0 | 0 |
| 2.45 | -1.07101 | -0.88424 | -0.4462 | 0 | 0 | 0 | 0 | 0 |
| 2.55 | -0.74753 | -0.40998 | -0.32325 | 0 | 0 | 0 | 0 | 0 |
| 2.65 | -0.8517 | -0.29204 | -0.02964 | 0 | 0 | 0 | 0 | 0 |
| 2.75 | -0.54793 | -0.01209 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.85 | -0.24543 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

DHO

|  | -0.975 | -0.925 | -0.875 | -0.825 | -0.775 | -0.725 | -0.675 | -0.625 | -0.575 | -0.525 | -0.475 | -0.425 | -0.375 | -0.325 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.75 | -0.71094 | -0.40922 | -0.31706 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.85 | -1.45626 | -1.26485 | -1.18471 | -0.87211 | -0.6934 | -0.53295 | -0.1772 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1.95 | -1.89761 | -1.73076 | -1.56893 | -1.37575 | -1.35943 | -1.11181 | -0.90583 | -0.58853 | -0.11157 | 0 | 0 | 0 | 0 | 0 |
| 2.05 | -1.91677 | -1.74344 | -1.74682 | -1.62899 | -1.55018 | -1.41659 | -1.22333 | -1.12483 | -0.93924 | -0.56316 | -0.002[illegible]9 | 0 | 0 | 0 |
| 2.15 | -1.73718 | -1.53767 | -1.57472 | -1.38064 | -1.40464 | -1.35152 | -1.25618 | -1.24106 | -1.02982 | -0.73677 | -0.648[illegible]4 | -0.23396 | 0 | 0 |
| 2.25 | -1.25165 | -1.27657 | -1.24474 | -1.19016 | -1.19986 | -1.20228 | -1.08478 | -0.96698 | -0.89755 | -0.76119 | -0.477[illegible]9 | -0.21077 | -0.04896 | 0 |
| 2.35 | -0.83347 | -0.85246 | -0.79106 | -0.79106 | -0.87416 | -0.80101 | -0.60074 | -0.74677 | -0.5254 | -0.47182 | -0.257[illegible]6 | 0 | 0 | 0 |
| 2.45 | -0.31765 | -0.46394 | -0.33715 | -0.49328 | -0.55726 | -0.43372 | -0.40255 | -0.29273 | -0.21945 | -0.09221 | -0.028[illegible]4 | 0 | 0 | 0 |
| 2.55 | 0 | -0.08414 | -0.01278 | -0.118 | -0.18765 | -0.16673 | -0.08414 | -0.01278 | -0.02503 | -0.0311 | 0 | 0 | 0 | 0 |
| 2.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2.95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.75 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.85 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3.95 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Figure 22: Version 4.1 of the PCI score function fixes incorrect normalization.**
A) Incorrect normalization without pseudo counts. The red box indicates a deviation in the distance between the acceptor – orbital – hydrogen (AOH) and donor – hydrogen – orbital (DHO) angle. B) Correct implementation for normalizing with pseudo counts. The distance minimum for both AOH and DHO knowledge based potentials match.

During minimization, orbital locations must be updated after derivative calculations. Previous implementations of the code did not update orbital coordinates correctly Figure 24; inclusion of a set of logic placed in the virtual function, setup_for_scoring() corrected the problem.

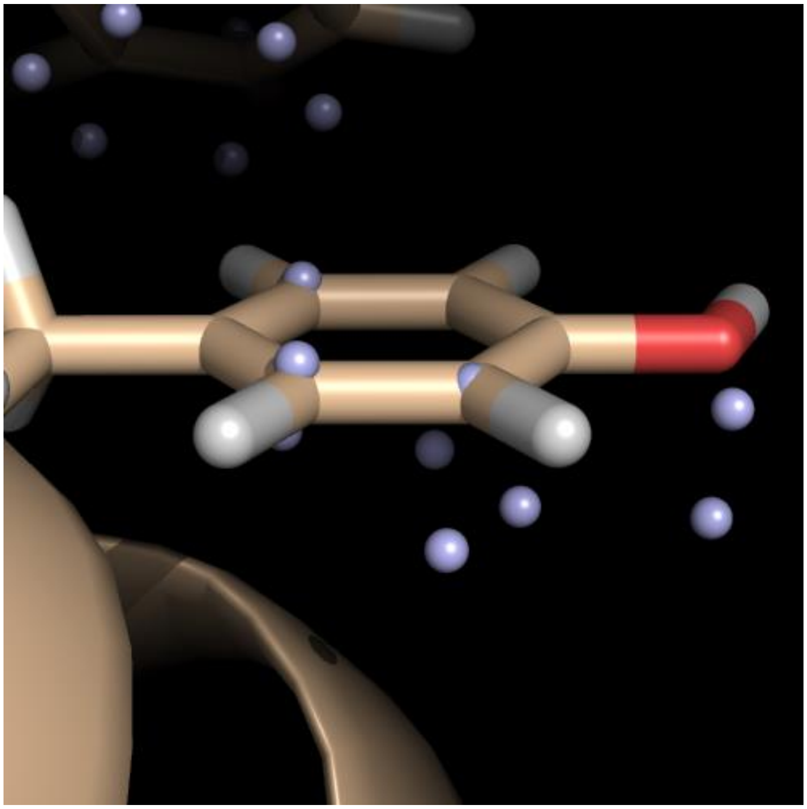


**Figure 23: Incorrect placement of orbitals after minimization.**
During minimization, the conformation of both the orbitals and atoms must be updated. Previous implementations of the partial covalent interactions score function incorrectly setup the residues for minimization. Functionality placed into the setup_for_scoring() function correctly updates orbital placement during minimization.

### *Version 5*

Version 5 fixes errors for scoring and minimizing aromatic residues. Initially, an orbital was positioned at the center of aromatic residues (Version 1); however, because there was no chemical relevance for this orbital, it was removed in Version 3. Examination of crystal structure Lambert-Azimuthal equal area projection graphs for π-π interactions for hydrogen placement relative to the orbital show that placement of the hydrogen is approximately 30.0° off the plane of the orbital (Figure 25). It is apparent in examining crystal structures that the aromatic hydrogen in π-π interactions is often placed above the center of the participating residue. Further, when structures are minimized, the resultant models reproduce structures at differing minimum than seen in crystal structures. To correct for errors during minimization, a separate orbital to the center of aromatic rings was added. Minimization of π-π interactions results in structures with two defined minimum (Figure 26). This is because the bin width for the KBP for π-π interactions is too large and can find a minimum between two orbitals, thus receiving double the energy. Decreasing the bin width size for the angle component of π-π interactions and including the center orbital results in recapitulation of crystal structures geometries for π-π interactions.

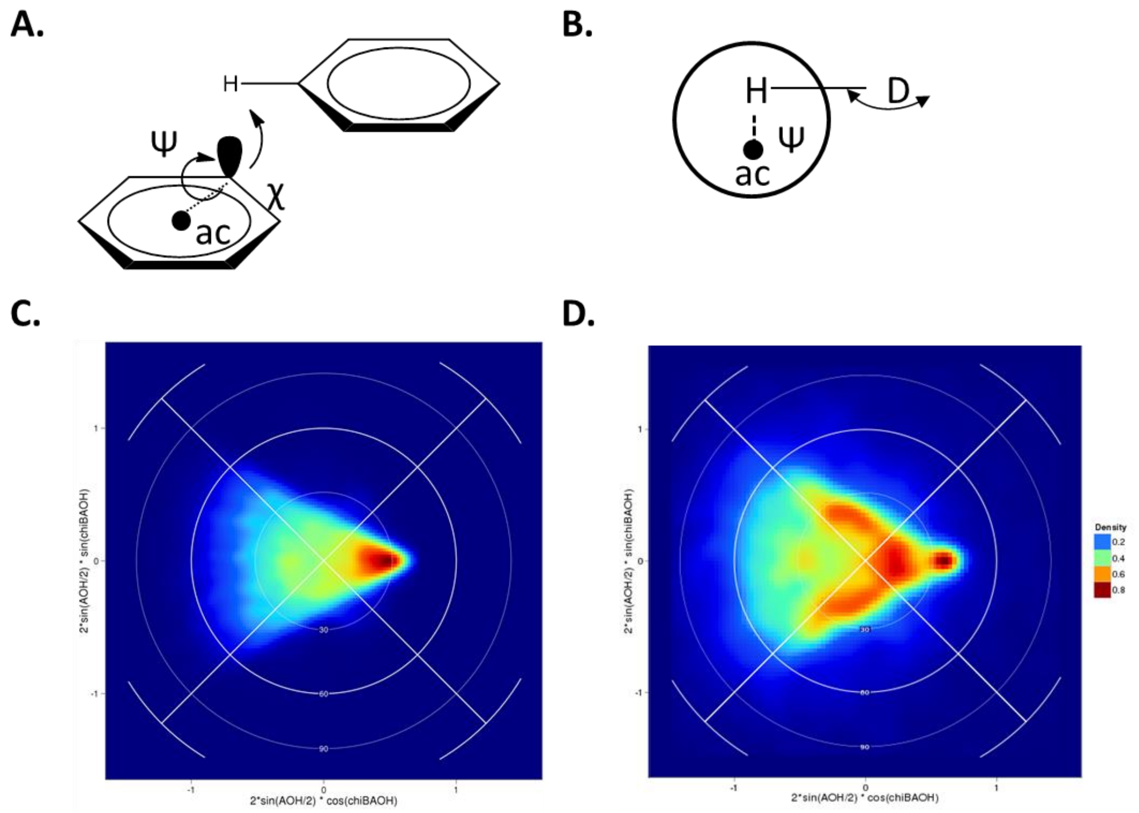


**Figure 24: Geometric parameters used to create the Lambert-Azithumul plot and Lambert-Azithmul equal area projection plots for side chain π-π interactions between an aromatic ring with an aromatic Hydrogen side chain group.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-orbital-hydrogen atoms. The action center (ac) was used as the base acceptor so that the X angle remained consistent between all interactions. Ψ is the acceptor – orbital – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal π-π interaction, the hydrogen placement will be directly located at the orbital. The Newman projection plot depicts an offset of the hydrogen for clarification of the plot. C) Lambert-Azimuthal projection of an orbital to an aromatic Hydrogen for crystal structures. High density is located at 30.0° indicative that wrong geometric parameters were used to create the plots. D) Lambert-Azimuthal projection of an sp2 acceptor and an aromatic hydrogen for relax models using the partial covalent scoring function. Specific minimum of each orbital placement above each carbon is observed.

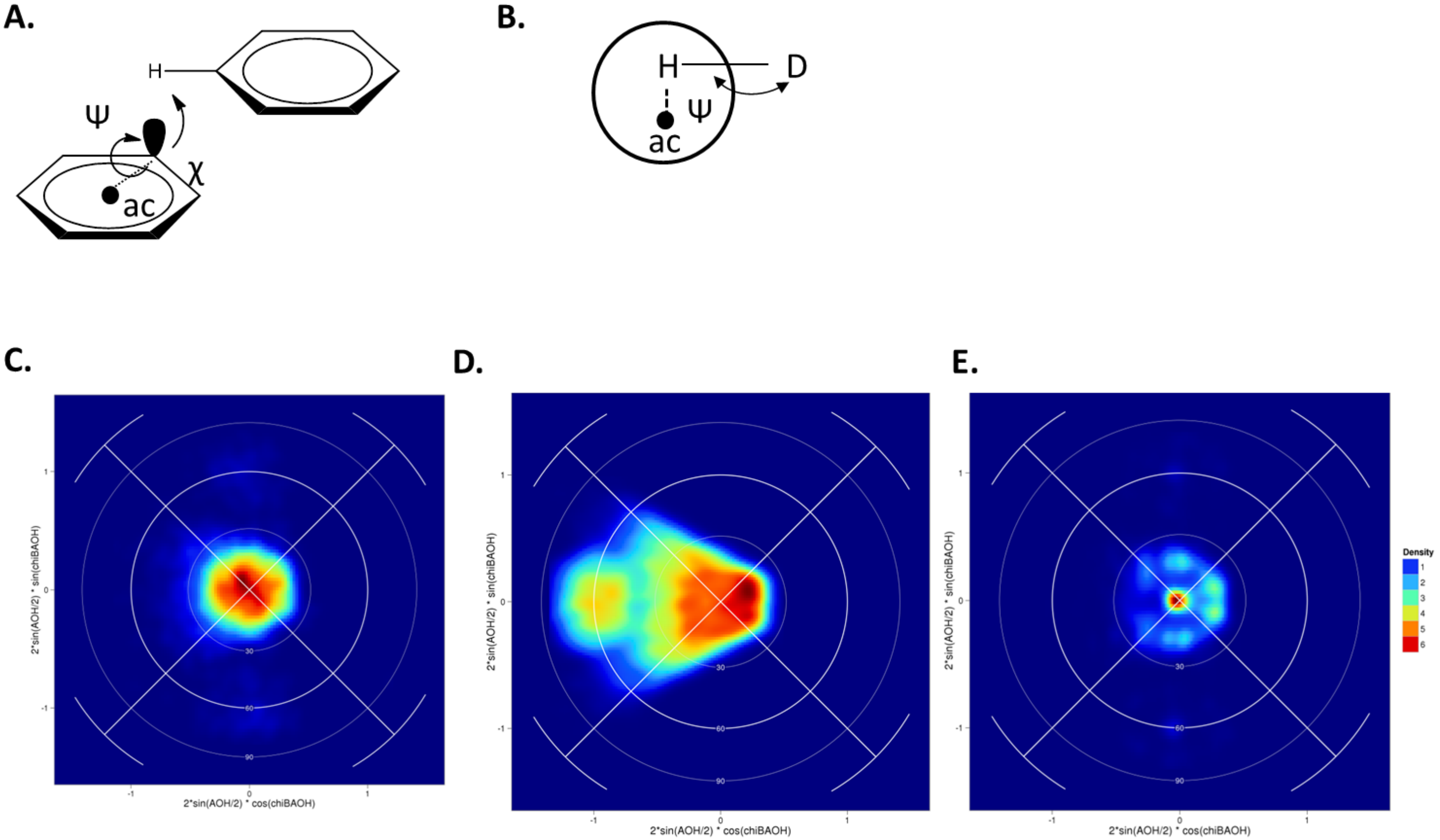


**Figure 25: Geometric parameters used to create the Lambert-Azithumul plot and Lambert-Azithmul equal area projection plots for side chain π-π interactions between an aromatic ring with an aromatic Hydrogen side chain group with middle orbital in the aromatic ring.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-orbital-hydrogen atoms. The action center (ac) was used as the base acceptor so that the X angle remained consistent between all interactions. Ψ is the acceptor – orbital – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal π-π interaction, the hydrogen placement will be directly located at the orbital. The Newman projection plot depicts an offset of the hydrogen for clarification of the plot. C) Lambert-Azimuthal projection of an orbital to an aromatic hydrogen for crystal structures. High density is located at 0.0° indicative that correct parameters were used to generate the plots and score functions.  D) Relaxed structures using the partial covalent score function. A biomodal distribution is shown due to the large binning used to create the knowledge based potential. E) Relaxed structures using the partial covalent score function. Biomodal distrubition is removed by including smaller angle bins in creation of the KBP

# CHAPTER IV

## BENCHMARK RESULTS FOR THE PARTIAL COVALENT INTERACTIONS SCORE FUNCTION

### *Datasets of High Resolution Crystal Structures Used to Benchmark Rosetta Protocols*

To derive the probability distribution for geometrical features describing PCIs, the top8000 dataset of crystal structures was used. The top8000 dataset(Keedy 2012) contains monomeric proteins with at least 25% of side chains present, greater than 38 residues, and has a MolProbity score of < 2.0 (Chen, Arendall et al. 2010). Missing hydrogen atoms were added to all crystal structures using Reduce(Word, Lovell et al. 1999) and then converted to Rosetta hydrogen atom types via a python script.

The dataset of proteins used for the protein design benchmark was created through the protein sequence culling server PISCES(Wang and Dunbrack 2003). X-ray structures with a sequence identity limit of 25%, resolutions better than 1.5 Å, and sequence lengths between 175 and 250 residues were created. A total of 415 of crystal structures were obtained given this criteria. For the first time we benchmark protein design on a dataset of large proteins – typically smaller protein domains with even higher resolution are used to accelerate the calculation and emphasize quality of the structures. However, as focus is shifting towards designing large proteins(Fortenberry, Bowman et al. 2011), it is critical to adjust benchmarks accordingly.

Accurate analysis of rotamer recovery requires that all side chain atoms be present in the dataset. A set of 29 proteins with all side chain atoms present and a resolution of 1.8Å or better was obtained(Liang and Grishin 2002). The dataset ensures accurate rotamer recovery analysis by being diverse (all α,all β, and α/β mix) and high resolution with all side chain atoms present.

## Analysis of Orbitals in Experimentally Determined Crystal Structures

Knowledge-based potentials (KBP) were created for each orbital class and driving interaction (see methods) (Figure 27). For all side chain – side chain PCIs, the most energetically favorable angle bin occurred at a straight line angle between the acceptor – orbital – hydrogen angle (Ψ) and donor – hydrogen – orbital (Θ) angle (Figure 3, 27cos(Ψ) and cos(Θ)). The assumption that the orbital is a single distance equal to the Bohr radius of an atom is reflected by the distance component's, $\delta_{HOrb}$, energetic minimum, which varies between each orbital class. For the most energetically favorable angle/distance bin, an example from each PCI interaction depicted is shown in experimentally validated models (Figure 27). Orbitals are displayed as a single grey sphere while hydrogen atoms are depicted as a white sphere. For brevity, only side chain – side chain PCIs KBPs are shown while side chain – backbone interactions are summarized in the supplementary information.

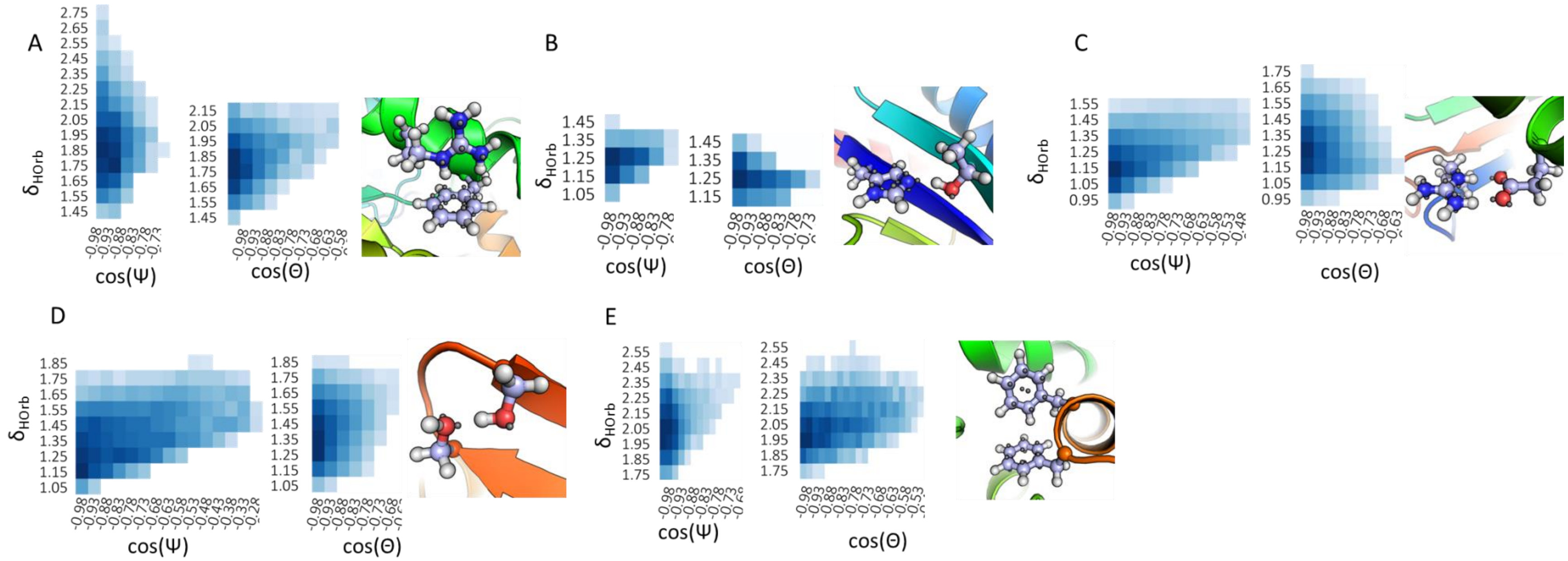


**Figure 26: The energy potential for each side chain side chain PCI and a cartoon representation of the energetic minimum for both cos(Ψ) and cos(Θ) favorable.**

A) Energy potential for orbital class C.pi.sp2 interacting with polar hydrogen (hpol). Left, E($\delta_{HOrb}$, cos(Ψ)), middle E($\delta_{HOrb}$,cos(Θ)), right, a cartoon representation for the energetic minimum of cos(Ψ) and cos(Θ). B) Energy potential for orbital class N.p.sp2 interacting with polar hydrogen (hpol). Left, E($\delta_{HOrb}$, cos(Ψ)), middle E($\delta_{HOrb}$,cos(Θ)), right, a cartoon representation for the energetic minimum of cos(Ψ) and cos(Θ). C) Energy potential for orbital class O.p.sp2 interacting with polar hydrogen (hpol). Left, E($\delta_{HOrb}$, cos(Ψ)), middle E($\delta_{HOrb}$,cos(Θ)), right, a cartoon representation for the energetic minimum of cos(Ψ) and cos(Θ). D) Energy potential for orbital class O.p.sp3 interacting with polar hydrogen (hpol). Left, E($\delta_{HOrb}$, cos(Ψ)), middle E($\delta_{HOrb}$,cos(Θ)), right, a cartoon representation for the energetic minimum of cos(Ψ) and cos(Θ). E) Energy potential for orbital class C.pi.sp2 interacting with an aromatic hydrogen (haro). Left, E($\delta_{HOrb}$, cos(Ψ)), middle E($\delta_{HOrb}$,cos(Θ)), right, a cartoon representation for the energetic minimum of cos(Ψ) and cos(Θ).

A vigorous test of the Rosetta score function is recapitulation of crystal structure features after a Rosetta protocol has modified the input structure. We utilize a series of benchmarks designed to test the new PCI score function against original crystal structure features. The standard Rosetta score function, score12’(Leaver-Fay, O&apos;Meara et al.), that had been previously optimized with OptE was also tested. The first two tests exam recapitulation of crystal structure geometric parameters for cation-π, π - π, salt bridges, and hydrogen bond interactions that were used to derive the knowledge-based potential after perturbation of side chains and backbone atoms. The third test measures how well the score function recovers conformations of side chains through rotamer recovery. Finally, the fourth test is based on the assumption that native protein sequences are close to optimal for their fold (Kuhlman and Baker) and measure the score functions ability to recover native and evolutionary observed amino acid identities.

The relax protocol in Rosetta is used to obtain low energy and native-like conformations of macromolecular models(Tyka, Keedy et al. 2011). During the relax protocol, models undergo a series of Monte Carlo perturbations to the backbone followed by rotamer optimization and a gradient based minimization. At each step in the relax protocol, the energy function is engaged to drive the simulation to an energetic minimum. A first test of the newly introduced energy function is therefore whether after relaxation PCIs are maintained in native-like conformations. As a control, the standard Rosetta score function is also tested. Two analysis of the resultant models were used to examine if PCIs were reproduced: comparison of geometric features through probability distributions and comparison of hydrogen population around an orbital in equal area projection plots.

### *Recapitulation of geometric parameters of PCIs*

To assess quantitatively the score functions ability to recapitulate angles and distances seen in PCIs, the probability distributions for PCIs were derived for the relaxed models for each score function, PCI and standard Rosetta (score12'), and compared to the ones used for creation of the PCI KBP in the first place. The top8000 dataset was relaxed using the resultant weight set from OptE (see previous section). The Root Mean Square Deviation (RMSD) in each bin fraction for each PCI driving interaction between relaxed structures and native crystal structures was measured. The RMSD for both score functions remained low confirming that there is no bias towards the destruction or creation of PCIs. Further, A PCI probability distribution was created for each orbital class and driving interaction. Each probability distribution created was normalized by dividing each bin fraction by the sum of all counts and subtracted from the normalized probability distributions found in crystal structures (Figure 28). Therefore, a positive number indicates that more counts for a given bin fraction occurred in the relax model than the crystal structures. Further, to facilitate comparison, the probabilities were also converted into a pseudo (Figure 28). The expected result for each resultant probability distribution and pseudo KBP should be a convergence towards the minimum of the input KBP. Only significant changes in geometric parameters between the standard and PCI score function with respect to experimental structures are described below and are shown in the figures.

Overall, the PCI geometries from the relax models for both $\delta_{HOrb}$, cos(Ψ) and $\delta_{HOrb}$, cos(Θ) in the PCI score function match closely the distribution derived from crystal structures. For offset parallel, T-stacked cation-π and π-π interactions and hydrogen bonds, negligible differences between $\delta_{HOrb}$, cos(Ψ) and $\delta_{HOrb}$, cos(Θ) probability distributions and pseudo KBP were observed for both the standard Rosetta score function and PCI score function. However, significant differences for salt bridges between the orbital class N.p.sp2 and O.p.sp2 were observed for the standard Rosetta score function (Figure 28) for $\delta_{HOrb}$, cos(Ψ).

Salt bridges between the orbital class N.p.sp2 and polar hydrogen atoms in the standard Rosetta score function when minimized converged to an off angle bin fraction for cos(Ψ) of 0.875 or 151°. Further, counts in each bin fraction were reduced for the straight angle of 180° for Ψ by the normalized count of 0.02 to 0.06. Additionally, the distance distribution became bimodal with normalized count increases of 0.02 in bin fractions 1.45Å and 1.85Å. For $\delta_{HOrb}$, cos(Θ), little difference between angle and distance distributions were observed. Conversely, the PCI score function for salt bridges involved with the orbital class N.p.sp2 and polar hydrogen atoms converged to a straight angle of 180° for Ψ and did not diverge in the distance distributions (Figure 28A,C).

For salt bridges that involve the orbital class O.p.sp2 and a polar hydrogen, significant difference were observed for density distributions of $\delta_{HOrb}$, cos(Ψ) for the standard Rosetta score function when compared to the PCI score function. An increase in off angle distributions for cos(Ψ) was observed for the standard Rosetta score function. Angle distributions did not converge upon a straight acceptor – orbital –

hydrogen angle, instead, an increase in normalized counts for a wide range of angles was observed over 90° to 180°. Further, the converged minimum occurred at an incorrect distance of 1.35Å as opposed to experimentally determined structures at 1.05Å. In contrast to the standard Rosetta score function, the PCI score function resultant distributions converged upon the bin fractions seen in experimental structures. An increase in normalized counts for a bin fraction of 0.975Å and 1.05Å was observed indicating that the PCI score function converged upon the energy minimum seen in crystal structures (Figure 28B,D).

Interestingly, similar difference for salt bridges between the orbital class N.p.sp2 and O.p.sp2 and polar hydrogen atoms are observed in the standard Rosetta score function. The resultant density distributions indicate that the distance and angle component of the hydrogen bond potential for salt bridges have difficulty controlling the acceptor – orbital – hydrogen angle as compared to the donor – hydrogen – orbital angle which showed little to no difference. In both instances, the PCI score function outperformed the standard Rosetta score function by converging to experimental determined geometries.

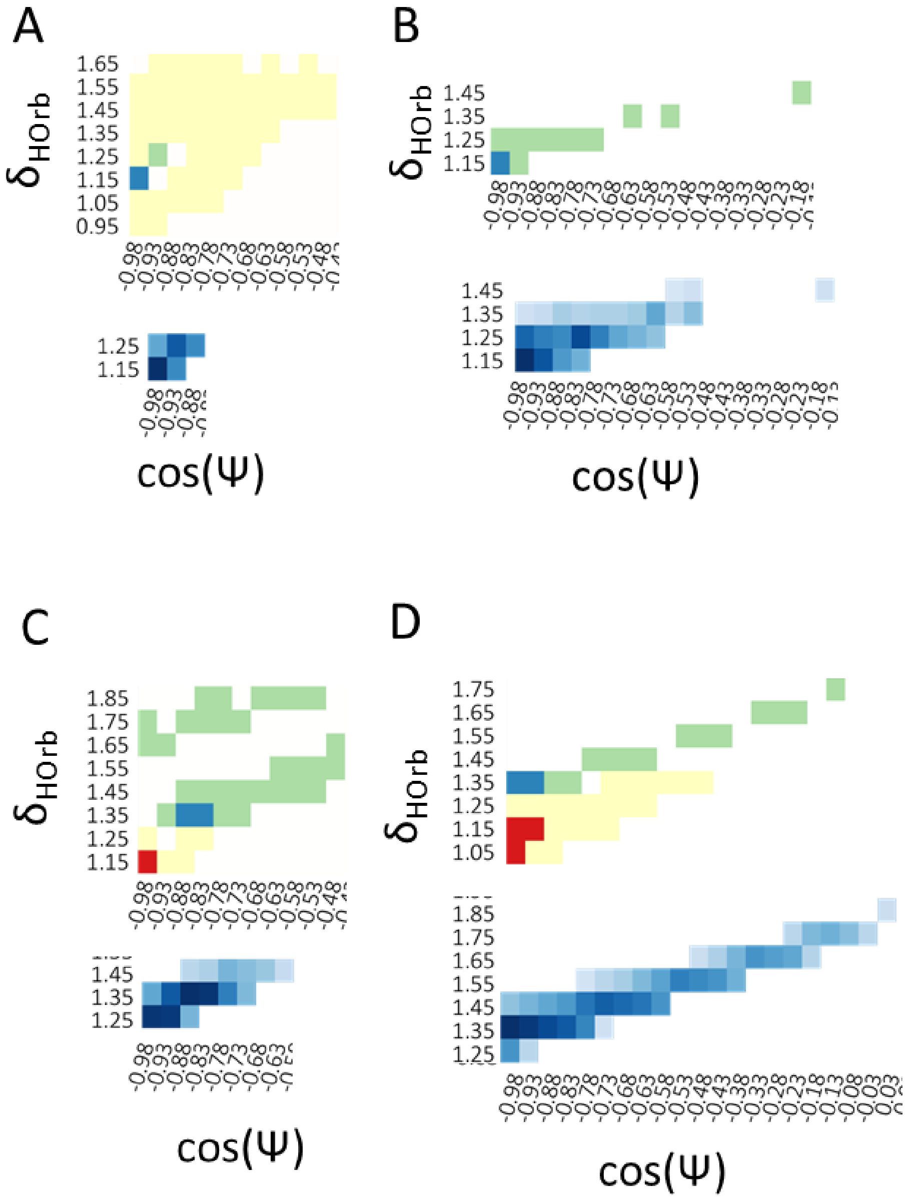


**Figure 27: Probability densities from relaxed structures measured against the native probability densities (top) and pseudo KBPs created from relaxed structures.**

A) Probability densities (top) and pseudo KBP (bottom) for orbital class N.p.sp2 for experimental structures relaxed in the PCI score function. B) Probability densities (top) and pseudo KBP (bottom) for orbital class O.p.sp2 for experimental structures relaxed in the PCI score function. C) Probability densities (top) and pseudo KBP (bottom) for orbital class N.p.sp2 for experimental structures relaxed in the standard Rosetta score function. D) Probability densities (top) and pseudo KBP (bottom) for orbital class O.p.sp2 for experimental structures relaxed in the standard Rosetta score function.

### *Lambert-Azimuthal Equal Area Projection Plots from Relax Models*

The previous analysis determines if the KBPs recover the geometric features in terms of distance and angles to orbitals that were input. However, we also wanted to confirm that the PCI KBPs recover native-like geometries for angles and torsion angles between sets of atoms similar to the ones used to derive the original hydrogen-bonding potential. Although torsion angles were not used in derivation of the PCI potential, examination of torsion angles involved in PCI ensures that the PCI potential does not induce non-native geometries for surrounding atoms. To examine the area occupied, a commonly used cartography tool, the Lambert-Azimuthal equal area projection plots, can be used to plot the distribution of hydrogen atoms around an acceptor orbital or atom. The distribution should remain consistent between the relaxed models and the original crystal structures. The RosettaFeatures reporter (Leaver-Fay, O&apos;Meara et al.) in conjunction with R scripts was used to create Lambert-Azimuthal plots for each major partial covalent interaction, cation-π, π- π, salt bridges, and hydrogen bonds. Multiple plots for all geometric parameters were created; however, only a representative subset of the plots are discussed below. The overall distribution of hydrogen atoms to orbitals changed little between relaxed models and crystal structures for the PCI potential, indicating that the weights and energy function are correctly recapitulating PCI geometries. However, there were significant changes in the distributions of relaxed models with the standard Rosetta score function.

To compare the torsion angles measured by the hydrogen bond potential in the standard Rosetta score function, the area occupied by hydrogen atoms in relationship between the base – acceptor – hydrogen – donor (Figure 1a, χ) and the acceptor – hydrogen – donor (Figure 1a, Ө) in crystal structures for hydrogen bonds between charged acidic residues with a hydroxyl donor are plotted via a Lambert-Azimuthal equal area plot (Figure 29C). Hydrogen bond distributions correspond to the measurements reported by Kortemme (Kortemme, Morozov et al. 2003) et al. of 120° and 180° for the Ө and χ angle respectively which results in high density at (-1,0) and (1,0) in the Lambert-Azimuthal plots (Figure 29C). The hydrogen atom distributions coincide with the orbitals for the sp2 acceptor atoms.

After performing a relax protocol with the standard Rosetta score function, a new Lambert-Azimuthal plot was created (Figure 29D,E). In contrast to the crystal structure plots, the standard score function inaccurately minimizes the torsion (X) and Ө angle. The hydrogen atoms are dispersed around the acceptor atom uniformly with a Ө angle of 120° and a varying X angle. In contrast to the standard Rosetta score function, the PCI score function results in models' distribution of hydrogen atoms focused into an ideal hydrogen bond structure of 120° and 180° for the Ө and χ angle respectively. Because the straight line geometry for the acceptor – orbital – hydrogen – donor is controlled by the PCI score function via the Ψ and Ө angles (Figure 28), a tight distribution of hydrogen atoms around the orbital results from the relax protocol. The equal area plots are indicative that the weight set for *pci_hbond* are correctly weighted and the geometrical parameters used to define the PCI score function result in native like conformations.

In contrast to measuring χ and Θ angles in terms of atoms, a direct test of the PCI score function is to define the geometric parameters with orbitals. Salt bridge geometries specific to the PCI function were measured through definition of the χ angle as the acceptor – orbital – hydrogen – donor and Θ angle as acceptor – orbital – hydrogen. In crystal structures (Figure 30C), ideal salt bridges occur with an undefined χ (at (0,0) in the Lambert-Azimuthal plots) angle and a Θ angle at 180° (0,0)(Kortemme, Morozov et al. 2003). As the Θ angle moves from 180° to 120° and 60°, the density shifts from (0,0) to (0,0.5) off straight line angle between the acceptor – orbital – hydrogen. After the relax protocol with the standard Rosetta score function, hydrogen atom distributions were dispersed around the orbital at non-native angles, when compared to crystal structures (Figure 30D). Conversely, a clear preference for optimal salt bridge geometries with an undefined χ angle and an 180° Θ angle (0,0) is recapitulated with the PCI score function. The distribution of hydrogen atoms around the orbitals are more focused when compared to the experimentally determined structures, indicating convergence in the relax protocol. However, bifurcation of the hydrogen atom between the two carboxylic acid p-orbitals becomes pronounced after the relax protocol ( (-0.5,0) in the Lambert-Azimuthal plots) with the PCI score function.

T-stacked and offset parallel cation-π interactions are defined in equal area plots with a χ angle as the acceptor – orbital – hydrogen – donor and Θ angle as acceptor – orbital – hydrogen (Figure 31). The acceptor atom is defined as an atom in an aromatic ring whereas the orbital is the π-orbital belonging to the acceptor atom (Figure 31). The distribution for hydrogen atoms are centered around the π-orbital of the aromatic ring with a χ that is undefined and ϴ angle of 180°. In a report on energetically favorable cation-π interactions, Gallivan(Gallivan and Dougherty 1999) et al., described the majority of favorable cation-π interactions occur with the N-atom above the π-orbital. With the orientation of the N-atom above the π-orbital, the acceptor – orbital – hydrogen angle is 180° as seen in the crystal structure Lambert-Azimuthal plots. The standard Rosetta score function does not account for cation-π interactions, therefore, after relax with the standard score function hydrogen atom distributions are dispersed in a wide area around the orbital (Figure 31E). However, with the PCI score function, the hydrogen atom distribution converges to the orbital (Figure 31D).

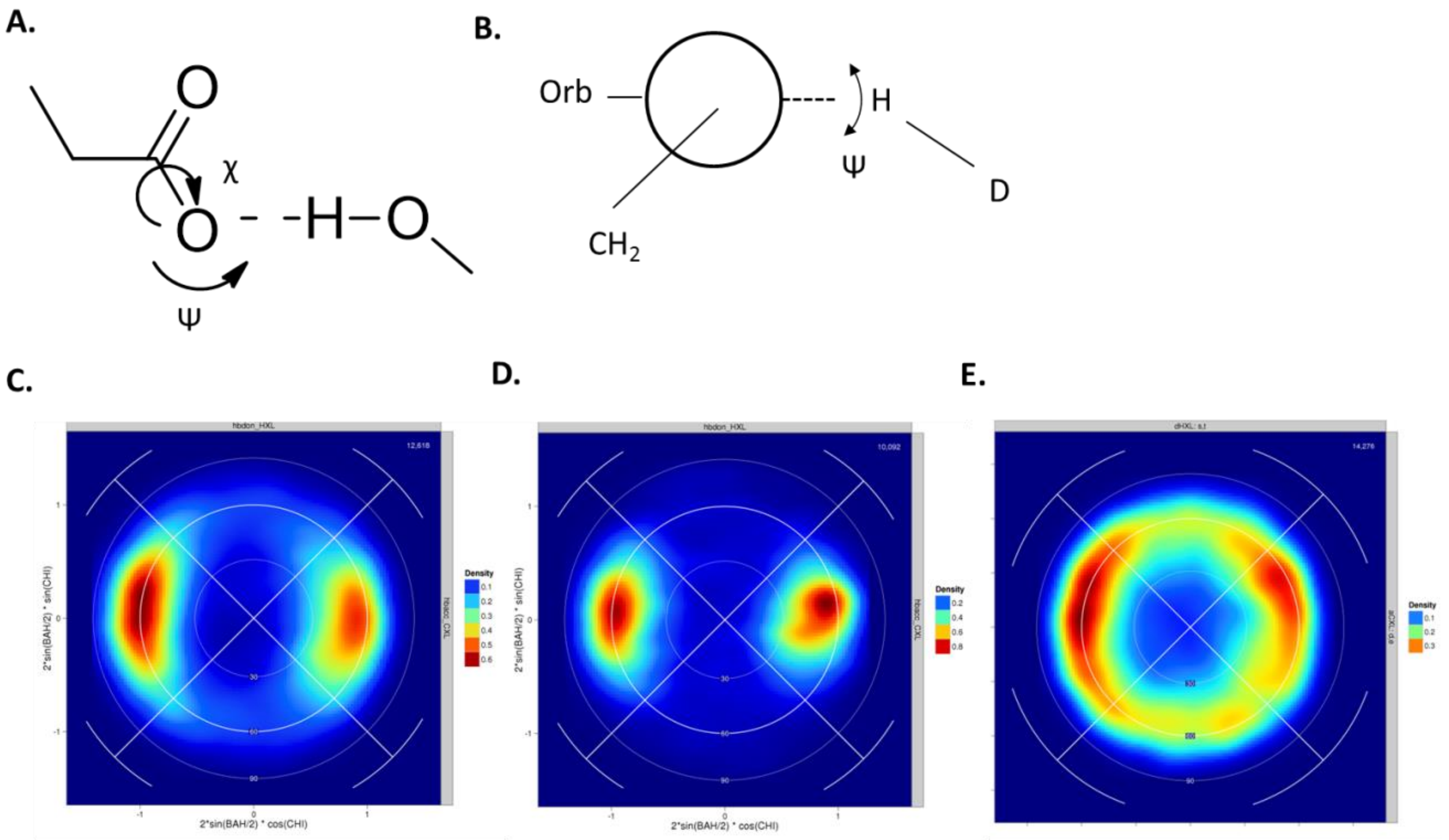


**Figure 28: Geometric parameters used to create the Lambert-Azithumul plot (A,B) and Lambert-Azithmul equal area projection plots (C,D) for side chain hydrogen bonds with an sp2 acceptor and a hydroxyl donor.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-hydrogen-donor atoms and Ψ the base acceptor – acceptor – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal hydrogen bond structure, the hydrogen placement will be directly located at the orbital. C) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for crystal structures. High density is located directly where the orbital of an sp2 acceptor would be present. D) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using the orbital score function. Density distribution is consistent between the crystal structure plot (C). E) Lambert-Azimuthal projection of an sp2 acceptor and a hydroxyl donor for relax models using Rosetta’s standard score function. The density distribution is not centered where the orbitals are located on the acceptor atom.

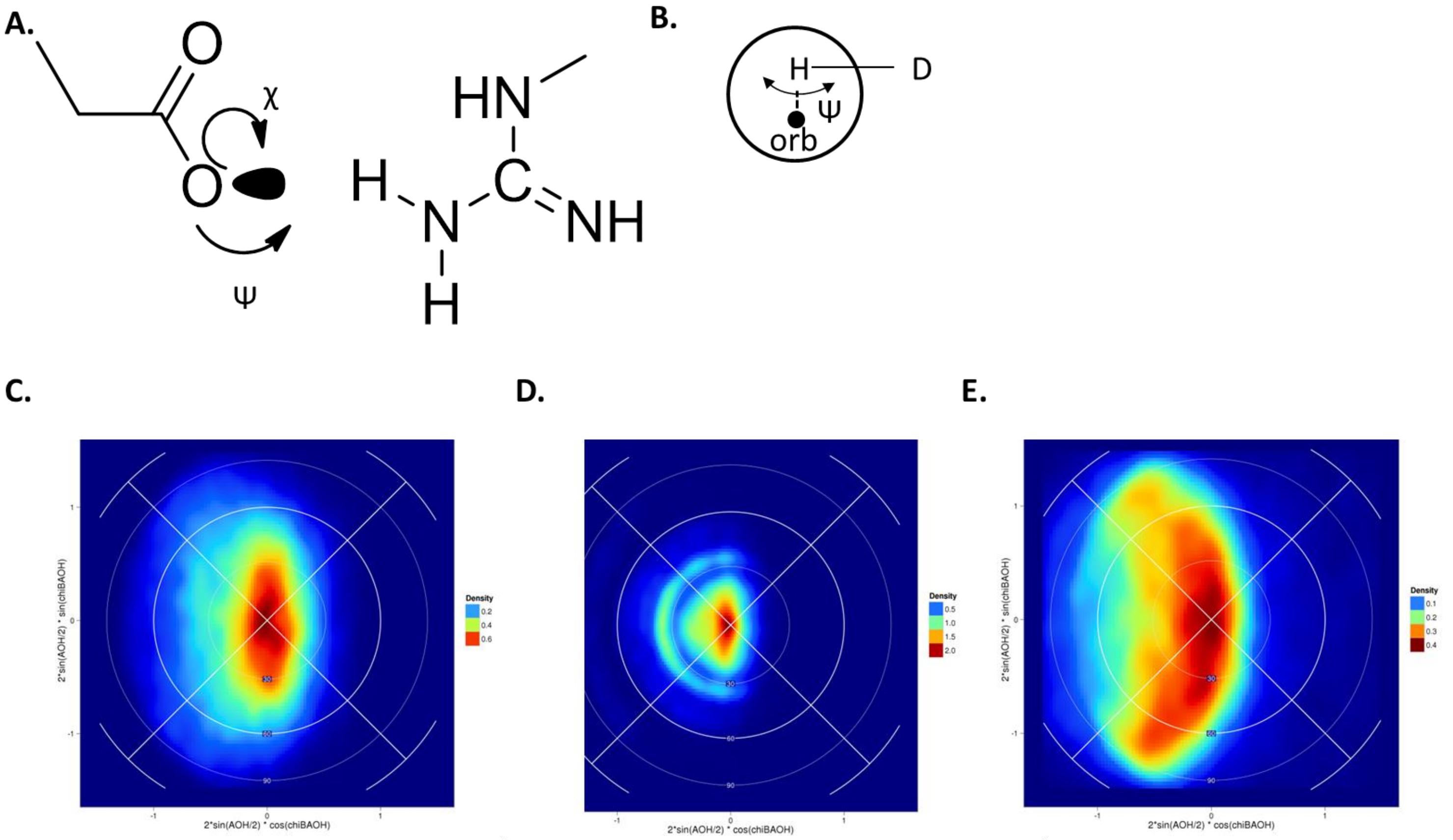


**Figure 29: Geometric parameters used to create the Lambert-Azithumul plot (A,B) and Lambert-Azithmul equal area projection plots (C,D) for side chain salt bridge interactions between an acidic residue with an amine or guanidine Hydrogen side chain group.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-orbital-hydrogen atoms. Ψ is the acceptor – orbital – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal salt bridge interaction, the hydrogen placement will be directly located at the orbital. The Newman projection plot depicts an offset of the hydrogen-orbital for clarification of the plot. C) Lambert-Azimuthal projection of an O.p.sp2 orbital to a polar Hydrogen for crystal structures. High density is located directly where the orbital of an sp2 acceptor would be present. D) Lambert-Azimuthal projection of an sp2 acceptor and a polar Hydrogen for relax models using the orbital score function. Density distribution is consistent between the crystal structure plot (C). E) Lambert-Azimuthal projection of an sp2 acceptor and a polar Hydrogen for relax models using the Rosetta standard score function. Density distribution broadened and not as focused around the orbital as seen with the orbital score function (D) and crystal structures (C).

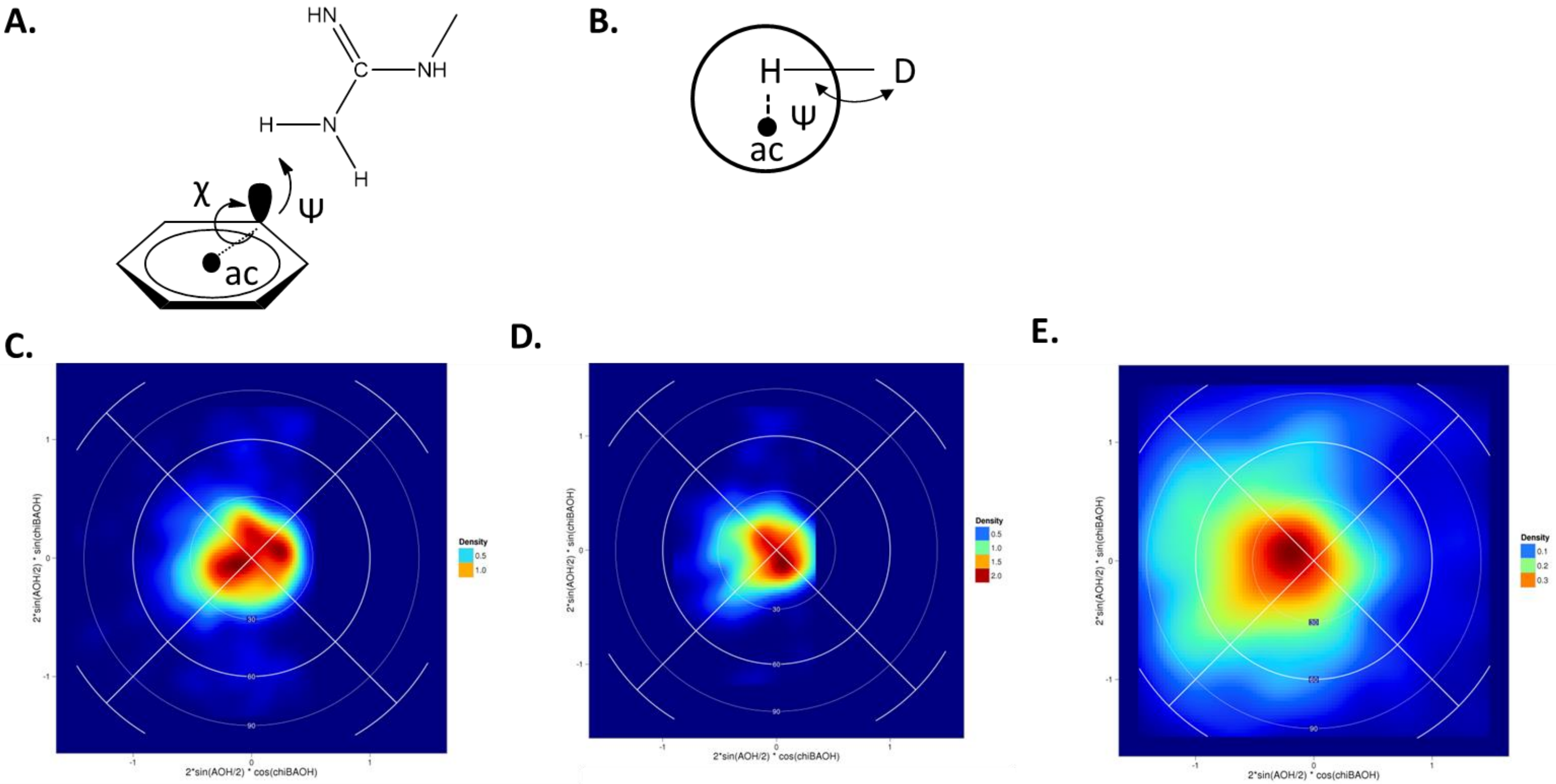


**Figure 30: Geometric parameters used to create the Lambert-Azithumul plot (A,B) and Lambert-Azithmul equal area projection plots (C,D) for side chain cation-π interactions between an aromatic ring with an amine or guanidine side chain group.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-orbital-hydrogen atoms. The action center (ac) was used as the base acceptor so that the X angle remained consistent between all interactions. Ψ is the acceptor – orbital – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal cation-π interaction, the hydrogen placement will be directly located at the orbital. The Newman projection plot depicts an offset of the hydrogen for clarification of the plot. C) Lambert-Azimuthal projection of an C.pi.sp2 orbital to a amine or guanidine Hydrogen for crystal structures. High density is located directly where the orbital of an sp2 acceptor would be present. D) Lambert-Azimuthal projection of an sp2 acceptor and an amine or guanidine Hydrogen for relax models using the orbital score function. Density distribution is consistent between the crystal structure plot (C). E) Lambert-Azimuthal projection of an sp2 acceptor and an amine or guanidine Hydrogen for relax models using the Rosetta standard score function. Density distribution is off center and more widely distributed around the orbital in contrast to the crystal structures and orbital score function.

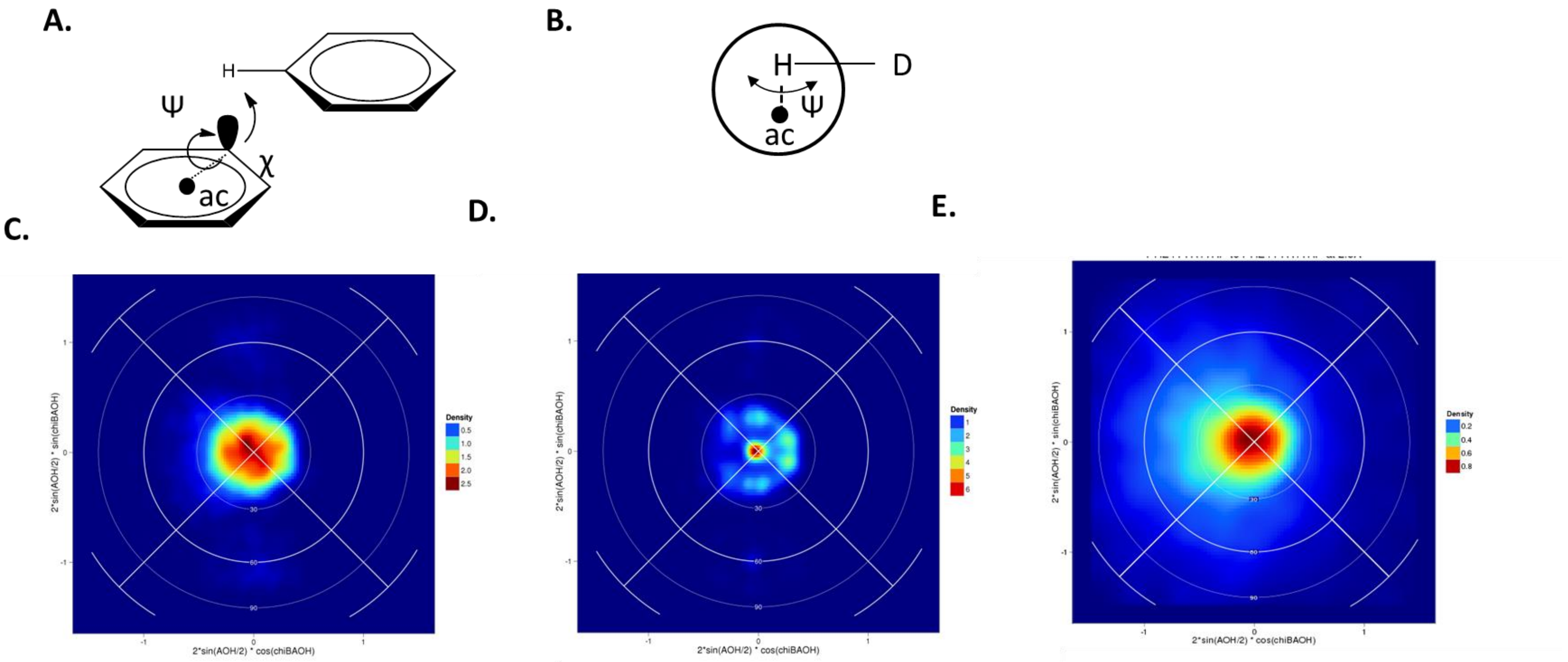


**Figure 31: Geometric parameters used to create the Lambert-Azithumul plot (A,B) and Lambert-Azithmul equal area projection plots (C,D) for side chain π-π interactions between an aromatic ring with an aromatic Hydrogen side chain group.**

A) Schematic representation of parameters used to create the equal area plot. X, the torsion angle between the base acceptor- acceptor-orbital-hydrogen atoms. The action center (ac) was used as the base acceptor so that the X angle remained consistent between all interactions. Ψ is the acceptor – orbital – hydrogen angle. B) Newman projection depicting the X angle and the Ψ angle used to create Lambert-Azimuthal plots. In an ideal π-π interaction, the hydrogen placement will be directly located at the orbital. The Newman projection plot depicts an offset of the hydrogen for clarification of the plot. C) Lambert-Azimuthal projection of an C.pi.sp2 orbital to an aromatic Hydrogen for crystal structures. High density is located directly where the orbital of an sp2 acceptor would be present. D) Lambert-Azimuthal projection of an sp2 acceptor and an aromatic Hydrogen for relax models using the orbital score function. Density distribution is consistent between the crystal structure plot (C). D) Lambert-Azimuthal projection of an sp2 acceptor and an aromatic Hydrogen for relax models using the Rosetta standard score function. Density distribution is consistent between the crystal structure plot (C), but is slightly less focused.

### *Rotamer recovery*

Conformational sampling for proteins side chains is a combinatorial problem that produces a large search space. Experimentally determined structures contain side chains in an energetically favorable conformation. Unlike protein design, rotamer recovery attempts to recover the correct conformation of a protein side chain in context of all other side chains in a protein. A stringent test for the score function is to see if the energetic minimum for side chain conformations can be recovered through sampling multiple rotamers (conformations) for each side chain(Petrella, Lazaridis et al. 1998; Liang and Grishin 2002). To this end, rotamer recovery was measured for all twenty natural occurring amino acids on a dataset of residues that contained all side chain atoms (Figure 33). Overall, rotamer recovery improved when using the PCI score function (74.4%) as compared to the standard score12' score function (73.1%) (Figure 9). Residues involved in hydrogen bonds, salt bridges, and π-interactions showed the most improvement in rotamer recovery; however, both His and Phe showed a slight loss for rotamer recovery. Interestingly, a large improvement(3.0%-10.0%) in rotamer recovery for both Asn and Gln was seen. This improvement is a direct result from correctly parameterizing the sp2 hyrbidization of the carbonyl functional group. A similar improvement is seen in both Asp and Glu residues. Rotamer recovery improved greater overall for surface residues (2.0% improvement) as compared to a slight increase in overall core residue rotamer recovery (0.5% improvement). It is unclear if the improvement in surface residues is a result from differing weights for the solvation potential or improved parameterization of residues containing orbitals.

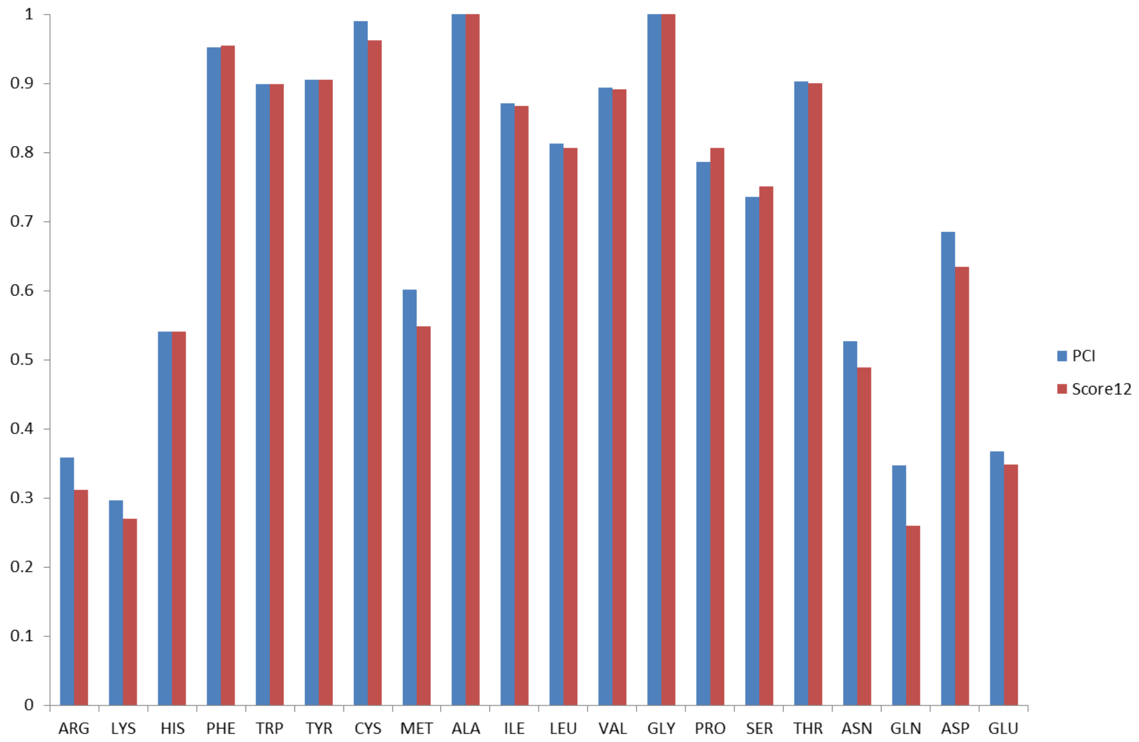


**Figure 32: Percent rotamer recovery for each amino acid.**
Blue, PCI score function, Red, standard rosetta score function.

### *Side Chain Identity Recovery and PSSM Recovery*

*In silico* protein design is benchmarked typically on recovering side chain identity in crystal structures assuming that native sequences of proteins are close to optimal(Kuhlman and Baker 2000). Here we measure sequence recovery to the native amino acid as well as recovery to evolutionary sampled amino acid identities (PSSM recovery, see methods). Further we analyze amino acid identity distribution in designed proteins and substitution profiles from designed models to the original crystal structures. The number and geometry of PCIs after design is evaluated as well as the packing density for the core of the protein.

A large dataset of 414 monomeric proteins (see methods) were designed using two different methods. The first method is design on a fixed backbone while the second method is design on an energy minimized fixed backbone. RosettaScripts was used for both methods. After complete design of the protein, the recovery of the naturally occurring amino acid was measured (Table 3). For proteins that were designed before relax, the overall sequence recovery was 37% while sequence recovery for standard Rosetta was 36%. For both score functions, the sequence recovery for the core of the protein was higher than the surface of the protein. This is in part contributed to the restriction in the degrees of freedom in the core of the protein compared to the surface of the protein. The core of the protein has an increase chance for clashes between residues whereas the surface conformational sampling can result in little to no clashes between residues. In a second test of design, potential crystallographic errors were accounted for through relax of the protein dataset with both score functions. Sequence recovery was 51% and 48% for the PCI and standard score function, respectively. Interestingly, the PCI score function outperforms the standard score function by 3% whereas with no-relax design, the PCI score function only recovers 1% more of residues. We hypothesize that a significant increase in recovery is due to the PCI score function correctly recapitulating side chain geometries (see Results,*Recapitulation of PCIs in energy-minimized models*) whereas the standard score function incorrectly minimizes PCI.

| | Sequence Recovery | | | PSSM Recovery | Rotamer Recovery | | | Partial Covalent Interactions | | | Packing |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Buried | Surface | Overall | Overall | Buried | Surface | Salt Bridges | Cation-π | π-π | Overall |
| Native | NA | NA | NA | NA | NA | NA | NA | 12.6 | 4.5 | 7.8 | 0.72 |
| Score12’ | 0.36 | 0.44 | 0.26 | 0.66 | 73 | 82 | 58 | 21.1 | 2.8 | 6.3 | 0.5 |
| Score12’ relaxed | 0.49 | 0.58 | 0.37 | 0.72 | NA | NA | NA | 21.6 | 3.6 | 8.6 | 0.6 |
| PCI | 0.35 | 0.42 | 0.35 | 0.68 | 74 | 83 | 61 | | | | |
| PCI relaxed | 0.48 | 0.68 | 0.51 | 0.73 | NA | NA | NA | 12.8 | 4.6 | 9.3 | 0.7 |

**Table 3: Summary of statistics from design and rotamer recovery runs.**

A test of whether the optimization algorithm OptE recovers native amino acid composition can be measured through the ratio of naturally occurring residues in the dataset compared to residues designed. Native occurring amino acids were divided by the number of residues designed for both scoring functions. Overall, amino acid composition remained consistent. For the PCI score function, a slight increase in the design of GLU was observed while for the standard score function, CYS was slightly over designed. For the PCI score function, an increase in recovery for residues involved in hydrogen bonds, cation-π and π-π interactions was observed when compared to the standard Rosetta score function. However, charged negative residues involved in salt bridges had poor sequence recovery. To examine this observation substitution profiles for the designed proteins were created. For both the non-relaxed and pre-relaxed datasets, the native amino acid was favored over any other designed residue with the exception of glutamine which was equally substituted for glutamic acid. Substitution of amino acid remained consistent for the amino acid biochemical profile; basic residues were substituted for other basic residues. This general trend is indicative that the original interactions the residues were creating (van der Waals packing, PCIs) were further optimized after design. Specifically, residues that comprise salt bridge interactions (K,R,D,E) were substituted from smaller residues (D,K) to larger residues (E,R) or vice versa. Overall, the polar (N,Q,S,T) and charged residues (R,K,D,E) were substituted more frequently than the aromatic residues and apolar residues. This could be because the polar residues and charged residues can participate in more diverse interactions such as hydrogen bonds, salt bridges, and cation-pi interactions.

### *PSSM Recovery*

As a final test, the percent recovery for a position specific scoring matrix (PSSM) can be measured (see methods). Sequence recovery for a PSSM addresses the limitation that sequence recovery and substitution profiles assume that the lowest free energy amino acid is the native residue. This is not necessarily true as multiple residues may have evolved for functionality. Both the standard score function and the PCI performed similar (Table 3). The overall PSSM recovery for unrelaxed structures was 0.68 and 0.66 for the PCI and standard score function, respectively. Although a 2% difference is seen between the unrelaxed designed structures, the difference for relaxed designed structures between the two scoring functions is negligible (1%).

### *PCI type recovery*

One limitation of substitution analysis is the inability to identify the type of interactions created. To this end, PCIs were measured between the native crystal structure dataset and the designed models (Table 3). For the PCI score function, the number of salt bridges were reduced by half, post relax/design, while cation-π and π-π content remained consistent with the native structures. In contrast, nearly twice as many salt bridges were designed with the standard score function while cation-π and π-π interactions were designed out. The over evaluation of salt bridges in the standard score function is attributed to two score terms that double count sat bridge interactions, *fa_pair* and *hbond_sc*. Both terms within the PCI score function are removed and replaced with a single term to evaluate salt bridges *pci_salt_bridge* (see methods).

### *Packing Metrics*

Sheffler(Sheffler and Baker 2009) et al. demonstrated that packing metrics could be used to identify Rosetta designed models and native structures. Native structures where packed densely when compared to the designed models. To this end, packing metrics were measured for the designed dataset. The overall packing density of pre-relax designed proteins were 0.56 and 0.53 for the PCI and standard score function, respectively, while for relaxed designed proteins were 0.64 and 0.59 (Table 3A). Although packing within the core of the protein is directly related to hydrophobic residues', the PCI score function after design packs the core better than the standard score function. A possible explanation is that arrangement of outer core residues, which are typically amphipatic residues, are better geometrically arranged into PCI interactions which influences core packing.

# CHAPTER V

# CONCLUSIONS

Recent advances in hardware and algorithm development have increased the speed at which programs run; however, robust and rapid algorithms for sampling and scoring macromolecular complexes are still needed. Rosetta attempts to provide rapid sampling through Monte Carlo moves and scoring through knowledge-based potentials. Although Rosetta has enjoyed many successes, improvements in the scoring function are needed (Chapter II and Chapter III).

Specifically, the second chapter focused on using Rosetta to dock ligands into a comparative model. Through both computational and experimental analysis, a putative binding mode for citalopram was proposed. Additionally, a lower gate through which serotonin and ions can transfer was identified; however, deficiencies in the Rosetta score function resulted in a poor correlation between all point mutations and the corresponding Rosetta scores. Because of this poor correlation, a second approach was used to identify spacer lengths that span between the binding pocket and the outer membrane in the human serotonin transporter for conjugation to quantum dots. This approach utilized a simple distance metric to the outer membrane to identify which small molecules to be synthesized. Once the small molecules had been synthesized, they were tested experimentally for binding and fluorescence in HEK cells.

In the third and fourth chapters, improvements to the scoring function in Rosetta were introduced through an orbital centric knowledge based potential. By including orbitals in the energy calculation, important interactions which were ignored in the Rosetta score function such as cation-π and π- π interactions are now evaluated. Orbital placement on atoms was done as a virtual point at a distance equivalent to the Bohr radius of the atom. The geometric angles of the orbital placement were based on the valence shell electron pair repulsion (VSEPR) theory. Geometric parameters for scoring paired residues was based upon a distance between the orbitals and hydrogen/orbitals, the acceptor – orbital – hydrogen, and the donor – hydrogen - orbital angles. Several iterations to the partial covalent interactions scoring function were created with a final version that closely recapitulates crystal structure geometries and metrics (Chapter III and Chapter IV). The improvements were made publically available to the Rosetta community and continue to be tested.

## Future Directions

The biogenic monoamine neurotransmitters, serotonin transporter (SERT), norepinephrine transporter (NET), and to a lesser extent the dopamine transporter (DAT), are targets for antidepressant drugs. Selective serotonin reuptake inhibitors (SSRI), serotonin norepinephrine reuptake inhibitors (SNRI), and norepinephrine reuptake inhibitors (NRI), contain structurally similar scaffolds that target SERT with high potency (SSRI), SERT and NET with similar potency (SNRI), or exclusively NET (NRI). Extensive structure activity relationship (SAR) studies have identified important ligand fragments that selectively inhibit NET or SERT; however, little is known about the antidepressant complex with SERT/NET. Crystal structures of SERT/NET have not been resolved but a crystal structure homolog from Aquifex aeolicus LeuT provides a template for comparative modeling of SERT/NET. An interesting avenue of research would be to identify important residues of SERT/NET in binding of antidepressants using a comparative model. Because SERT and NET transporters share an evolutionary background, these studies can help trace the evolutionary path in addition to identifying important resideus for binding antidepressants. A protocol has been established with RosettaLigand making the computational experiments easy to perform. Evaluation of docked poses can be guided by known $K_i$'s of antidepressant/transporters complex. SAR studies provide an added layer of evaluation.

Scoring functions evolve throughout the time of a software package. New parameters for atoms, changing solvation models, and new paradigms to scoring proteins are implemented. Recently, an unpublished set of improvements, coined Talaris2013, to the Rosetta score function were set as the default score function. The set of improvements included corrected atomic coordinates for amino acids, an analytical evaluation of Lennard-Jones and solvation terms, a new rotamer library, bicubic interpolation of knowledge-based terms, improved recognition of sp2 hydrogen bonds, expanded hydroxyl-chi sampling, an explicit Columbic electrostatics term, removal of the *fa_pair* term, improved disulfide geometries, and a new set of reference energies. These improvements were implemented into the partial covalent interactions score function; however, they have yet to be fully tested against an extensive set of benchmarks such as ligand docking, protein docking, loop remodeling, RNA/DNA folding, *ab initio* folding, and enzyme design.

Although the partial covalent interactions score function replaces the hydrogen bonding scoring terms in the current Rosetta score function, the Rosetta score function was updated to included improved hydrogen bonds with the Talaris2013 update; however, calculations for cation-π and π-π interactions are still not implemented. The partial covalent interactions score function is separated into individual terms, *pci_cation_pi, pci_pi_pi, pci_salt_bridge, pci_hbond* which allows for testing combinations of the partial covalent interactions score functions with the improved hydrogen bond potential. It would be beneficial to include the cation-π and π-π score terms with the Talaris2013 improvements. In order to accomplish this, several options need to be added to split the functionality of the score function.

In addition to combing Talaris2013 with the partial covalent interactions cation-π and π-π term, minimization of the orbitals needs to be improved. Currently, derivative calculations during minimization for orbitals are placed onto the bonded atom or the hydrogen atom instead of the orbital. This causes "Inaccurate G steps!" when minimizing because the force vectors f1() and f2() are incorrectly assigned to atoms, not orbitals. In order to completely minimize the structure, orbitals should be represented as virtual atoms and inserted into the atom trei. This will allow for f1() and f2() to be correctly calculated and minimize the structure. Through this implementation, code responsible for updating orbital coordinates (Chapter III Version 3) can be removed. Further, scoring of proteins will be more memory efficient as additional vectors needed to hold orbital conformational data will be removed and faster as additional function calls are no longer needed.

# APPENDIX

## Appendix I – Protein Docking

**command line:**

```
docking_protocol.static.linuxgccrelease     @flags.txt          -
score:weights   <weight>   -out:prefix   <prefix>   -in:file:native
<native> -s <structure> -ignore_unrecognized_res
```

**flags.txt:**

```
-database <database>
-use_input_sc
-ex1
-ex2
-ex2aro
-extrachi_cutoff 0
-nstruct 100
-dock_pert 3 8
-spin
-multiple_processes_writing_to_one_directory
-out:path:score outputs
```

## Appendix II – Enzyme Design

**command line:**

```
EnzdesFixBB.linuxgccrelease -database <database> @flags.txt
```

**flags.txt:**

```
-l ./inputs/pdb.list
-enzdes::detect_design_interface
-enzdes::cut1 6.0
-enzdes::cut2 8.0
-enzdes::cut3 10.0
-enzdes::cut4 12.0
-enzdes::cst_design
-enzdes::design_min_cycles 2
```

```
-enzdes::cst_min
-enzdes::chi_min
-out::file::o ./outputs/enz_score.out
-out::path::pdb ./outputs/
-ex1
-ex2
-ex1aro
-ex2aro
-extrachi_cutoff 1
-soft_rep_design
-flip_HNQ
-nstruct 1
-enzdes::no_unconstrained_repack
-enzdes::lig_packer_weight 1.8
-docking::ligand::old_estat
-linmem_ig 10
-extra_res_fa        inputs/2b3b.params        inputs/2ifb.params
inputs/1sw1.params      inputs/2FQX.params     inputs/2p0d.params
inputs/2DRI.params      inputs/1fby.params     inputs/1ZHX.params
inputs/2RDE.params      inputs/1db1.params     inputs/2h6b.params
inputs/1z17.params      inputs/2FME.params     inputs/1y3n.params
inputs/1urg.params      inputs/1FZQ.params     inputs/1y52.params
inputs/1POT.params      inputs/1XT8.params     inputs/2FR3.params
inputs/2UYI.params      inputs/1USK.params     inputs/1n4h.params
inputs/2qo4.params      inputs/2GM1.params     inputs/2rct.params
inputs/2HZQ.params      inputs/1hsl.params     inputs/1A99.params
inputs/1uw1.params      inputs/1l8b.params     inputs/3B50.params
inputs/1H6H.params      inputs/2Q2Y.params     inputs/1hmr.params
inputs/1OPB.params      inputs/1x7r.params     inputs/2Q89.params
inputs/1nl5.params      inputs/1TYR.params     inputs/2e2r.params
inputs/1LKE.params      inputs/2PFY.params     inputs/1wdn.params
inputs/1nq7.params      inputs/1y2u.params     inputs/2ioy.params
inputs/1J6Z.params      inputs/1RBP.params     inputs/1XZX.params
inputs/2f5t.params
```

## Appendix III – Features Reporter

**command line:**

```
rosetta_scripts.linuxgccrelease   @flags.txt   -l   <list>      -
parser:script_vars  output_db=<output_db>  source_sample=<source>
id=<id> struct_id=<struct>  weight=<weight> -database <database>
```

**flags.txt:**

```
-ignore_unrecognized_res
-add_orbitals
-output_orbitals
-out:file:renumber_pdb
-jd2
 -delete_old_poses
-parser
 -protocol relax_features.xml
-mute protocols.jd2
-mute core.io.pdb.file_data
-mute core.scoring.etable
-mute core.io.database
-mute core.scoring.ScoreFunctionFactory
-mute core.pack.task
-mute protocols.ProteinInterfaceDesign.DockDesign
```

**relax_features.xml**

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <s weights="%%weight%%"/>
      </SCOREFXNS>
      <TASKOPERATIONS>
            <ReadResfileFromDB                 name=relevant_chain
database_name="rosetta_inputs.db3" table="resfiles"/>
      </TASKOPERATIONS>
      <MOVERS>
            <SavePoseMover                          name=init_struct
reference_name=init_struct/>
            <FastRelax name=fast_relax scorefxn=s/>
            <ReportToDB                       name=features_reporter
database_name="%%output_db%%"      sample_source=%%source_sample%%
protocol_id=%%id%% task_operations=relevant_chain>
                  <feature name=ResidueFeatures/>
                  <feature name=HBondFeatures scorefxn=s/>
                  <feature name=HBondParameterFeatures scorefxn=s/>
                  <feature name=PairFeatures/>
                  <feature name=PdbDataFeatures/>
                  <feature name=PoseCommentsFeatures/>
                  <feature name=PoseConformationFeatures/>
                  <feature
name=ProteinBackboneAtomAtomPairFeatures/>
```

```
                <feature
name=ProteinBackboneTorsionAngleFeatures/>
                <feature
name=ProteinResidueConformationFeatures/>
                <feature                 name=ProteinRMSDFeatures
reference_name=init_struct/>
                <feature name=RadiusOfGyrationFeatures/>
                <feature name=ResidueBurialFeatures/>
                <feature name=ResidueSecondaryStructureFeatures/>
                <feature name=ResidueTypesFeatures/>
                <feature name=SaltBridgeFeatures/>
                <feature name=OrbitalsFeatures/>
           </ReportToDB>
      </MOVERS>
      <PROTOCOLS>
           <Add mover_name=init_struct/>
           <Add mover_name=fast_relax/>
           <Add mover_name=features_reporter/>
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

## Appendix IV – Interface Design

**command line:**

```
rosetta_scripts.linuxgccrelease -database <database> @flags.txt -
s "'<structure ligand>'" -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10
-extra_res_fa   <params>   -parser:script_vars   weight=<weight>
soft_rep=<soft_weight>
```

**flags.txt:**

```
-add_orbitals
-parser
 -protocol mv_design_scfxn.xml
-mute protocols.jd2
-mute core.io.pdb.file_data
-mute core.scoring.etable
-mute core.io.database
-mute core.scoring.ScoreFunctionFactory
-mute core.pack.task
-mute protocols.ProteinInterfaceDesign.DockDesign
```

**mv_design_scfxn.xml:**

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <ligand_soft_rep weights="%%soft_rep%%"/>
            <hard_rep weights="%%weight%%"/>
      </SCOREFXNS>
      <TASKOPERATIONS>
            <DetectProteinLigandInterface name=design_interface
cut1=6.0 cut2=8.0 cut3=10.0 cut4=12.0 design=1
resfile="/blue/meilerlab/home/combss/orbitals/new_work/interface_
design/options/Resfile"/>
            <ExtraRotamersGeneric name=extra_chi ex1=1 ex2=1
extrachi_cutoff=0/>
      </TASKOPERATIONS>
      <LIGAND_AREAS>
            <docking_sidechain chain=X cutoff=6.0
add_nbr_radius=true all_atom_mode=true minimize_ligand=10/>
            <final_sidechain chain=X cutoff=6.0
add_nbr_radius=true all_atom_mode=true/>
            <final_backbone chain=X cutoff=7.0
add_nbr_radius=false all_atom_mode=true Calpha_restraints=0.3/>
      </LIGAND_AREAS>
      <INTERFACE_BUILDERS>
            <side_chain_for_docking
ligand_areas=docking_sidechain/>
            <side_chain_for_final ligand_areas=final_sidechain/>
            <backbone ligand_areas=final_backbone
extension_window=3/>
      </INTERFACE_BUILDERS>
      <MOVEMAP_BUILDERS>
            <docking sc_interface=side_chain_for_docking
minimize_water=true/>
            <final sc_interface=side_chain_for_final
bb_interface=backbone minimize_water=true/>
      </MOVEMAP_BUILDERS>
      <MOVERS>
      single movers
            <ddG name=calculateDDG jump=1 per_residue_ddg=1
repack=0 scorefxn=hard_rep/>
            StartFrom chain=X>
                  Coordinates x=5 y=7 z=-2>
                  Coordinates x=-4 y=10 z=-6>
            /StartFrom>
            <Translate name=translate chain=X distribution=uniform
angstroms=.0001 cycles=50/>
```

```
           <Rotate name=rotate chain=X distribution=uniform
degrees=360 cycles=1000/>
           <SlideTogether name=slide_together chains=X/>
           <HighResDocker name=high_res_docker cycles=6
repack_every_Nth=3 scorefxn=ligand_soft_rep
movemap_builder=docking/>
           <PackRotamersMover name=designinterface
scorefxn=hard_rep task_operations=design_interface,extra_chi/>
           <FinalMinimizer name=final scorefxn=hard_rep
movemap_builder=final/>
           <InterfaceScoreCalculator name=add_scores chains=X
scorefxn=hard_rep/>
      compound movers
           <ParsedProtocol name=low_res_dock>
                <Add mover_name=translate/>
                <Add mover_name=rotate/>
                <Add mover_name=slide_together/>
           </ParsedProtocol>
           <ParsedProtocol name=high_res_dock>
                <Add mover_name=high_res_docker/>
                <Add mover_name=final/>
           </ParsedProtocol>
      </MOVERS>
      <PROTOCOLS>
           <Add mover_name=low_res_dock/>
           <Add mover_name=designinterface/>
           <Add mover_name=high_res_dock/>
           <Add mover_name=calculateDDG/>
           <Add mover_name=add_scores/>
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

**resfile.txt:**

```
ALLAA                # allow all amino acids
AUTO
#EX 1 EX 2         # allow extra chi rotameters at chi-id 1 and
#USE_INPUT_SC        # allow the use of the input side chain
conformation     ( see below for more detailed description of
commands)
start
```

## Appendix V – Relax/Design

**command line:**

```
rosetta_scripts.linuxgccrelease -database <database> @flags.txt -
s <structure>  -parser:script_vars -ex1 -ex2 -ex1aro -linmem_ig
10 -packing:extrachi_cutoff 0
```

**flags.txt:**

```
-ignore_unrecognized_res
-add_orbitals
-parser
 -protocol relax_design_features.xml
```

**relax_design_features.xml:**

```
<ROSETTASCRIPTS>
	<SCOREFXNS>
		<s weights="%%weight%%"/>
	</SCOREFXNS>
	<TASKOPERATIONS>
		<InitializeFromCommandline name=ifcl/>
	</TASKOPERATIONS>
	<MOVERS>
		<SavePoseMover name=init_struct
reference_name=init_struct/>
		<FastRelax name=fast_relax scorefxn=s
task_operations=ifcl/>
		<PackRotamersMover name=design scorefxn=s
task_operations=ifcl/>
	</MOVERS>
	<PROTOCOLS>
		<Add mover_name=init_struct/>
		<Add mover_name=fast_relax/>
		<Add mover_name=design/>
	</PROTOCOLS>
</ROSETTASCRIPTS>
```

## Appendix VI – Ligand Docking

**command line:**

```
rosetta_scripts.linuxgccrelease -database <database> @flags.txt -
s "'<structure ligand>'" -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10
-extra_res_fa <params> -parser:script_vars hard_rep=<hard_scfxn>
soft_rep=<soft_scfxn> -add_orbitals -out:pdb_gz -nstruct 500 -
in:file:native <native>
```

**flags.txt:**

```
-add_orbitals
-parser
 -protocol full_sample.xml
-mute protocols.jd2
-mute core.io.pdb.file_data
-mute core.scoring.etable
-mute core.io.database
-mute core.scoring.ScoreFunctionFactory
-mute core.pack.task
-mute protocols.ProteinInterfaceDesign.DockDesign
```

**full_sample.xml:**

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ligand_soft_rep weights="%%soft_rep%%"/>
        <hard_rep weights="%%hard_rep%%"/>
    </SCOREFXNS>
    <TASKOPERATIONS>
        <ExtraRotamersGeneric name=extra_chi ex1=1 ex2=1
extrachi_cutoff=0/>
    </TASKOPERATIONS>
    <LIGAND_AREAS>
        <docking_sidechain chain=X cutoff=6.0
add_nbr_radius=true all_atom_mode=true minimize_ligand=10/>
        <final_sidechain chain=X cutoff=6.0
add_nbr_radius=true all_atom_mode=true/>
        <final_backbone chain=X cutoff=7.0
add_nbr_radius=false all_atom_mode=true Calpha_restraints=0.3/>
    </LIGAND_AREAS>
    <INTERFACE_BUILDERS>
        <side_chain_for_docking
ligand_areas=docking_sidechain/>
        <side_chain_for_final ligand_areas=final_sidechain/>
        <backbone ligand_areas=final_backbone
extension_window=3/>
    </INTERFACE_BUILDERS>
    <MOVEMAP_BUILDERS>
        <docking sc_interface=side_chain_for_docking
minimize_water=true/>
```

```
        <final sc_interface=side_chain_for_final
bb_interface=backbone minimize_water=true/>
    </MOVEMAP_BUILDERS>
    <MOVERS>
        <ddG name=calculateDDG jump=1 per_residue_ddg=1
repack=0 scorefxn=hard_rep/>
        <Translate name=translate chain=X distribution=uniform
angstroms=2.5 cycles=50/>
        <Rotate name=rotate chain=X distribution=uniform
degrees=360 cycles=1000/>
        <SlideTogether name=slide_together chains=X/>
        <HighResDocker name=high_res_docker cycles=6
repack_every_Nth=3 scorefxn=ligand_soft_rep
movemap_builder=docking/>
        <FinalMinimizer name=final scorefxn=hard_rep
movemap_builder=final/>
        <InterfaceScoreCalculator name=add_scores chains=X
scorefxn=hard_rep/>
        <ParsedProtocol name=low_res_dock>
            <Add mover_name=translate/>
            <Add mover_name=rotate/>
            <Add mover_name=slide_together/>
        </ParsedProtocol>
        <ParsedProtocol name=high_res_dock>
            <Add mover_name=high_res_docker/>
            <Add mover_name=final/>
        </ParsedProtocol>
    </MOVERS>
    <PROTOCOLS>
        <Add mover_name=low_res_dock/>
        <Add mover_name=high_res_dock/>
        Add mover_name=calculateDDG/>
        <Add mover_name=add_scores/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

## Appendix VII – Loop Remodeling

**command line:**

```
loopmodel.linuxgccrelease @flags.txt  -score:weights <weight> -
database <database> -out:prefix <prefix>-in:file:native <native>
```

```
-s <structure> -loops:loop_file <loop> -add_orbitals -mute all
```

**flags.txt:**

```
-in:file:fullatom
-loops:remodel perturb_kic
-loops:refine refine_kic
-loops:outer_cycles 5
-nstruct 20
-ex1
-ex2
-extrachi_cutoff 0
```

## Appendix VIII – OptE

**command line:**

```
mpiexec optE_parallel.mpistatic.linuxgccrelease -database
<database> @flags.txt -s <list> -add_orbitals -
ignore_unrecognized_res
```

**flags.txt:**

```
-mute all
-optE
 -fixed fixed.txt
 -free free.txt
 -optimize_nat_aa
 -n_design_cycles 30
 -optimize_starting_free_weights true
 -mpi_weight_minimization
 -fit_reference_energies_to_aa_profile_recovery true
-skip_set_reasonable_fold_tree
-no_his_his_pairE
-no_optH true
-packing
 -ex1
 -ex2
 -ex1aro
 -linmem_ig 10
 -multi_cool_annealer 10
-add_orbitals
```

**fixed.txt:**

```
empty
fixed:
fa_atr 0.8
fa_rep 0.4
fa_sol 0.7
fa_intra_rep 0.004
hack_elec 0.6
pro_close 1
orbitals_hpol_bb 0.1
pci_cation_pi 0.5
pci_pi_pi 0.065
pci_salt_bridge 0.1
pci_hbond 0.15
hbond_sr_bb 0.585
hbond_lr_bb 1.17
dslf_fa13 1.0
rama 0.2
omega 0.5
fa_dun 0.56
p_aa_pp 0.32
```

## Appendix IX – RNA

**command line:**

```
rna_denovo.linuxgccrelease -database <database> -nstruct 2500 -
vall_torsions 1jj2_exciseSRL_exciseKT.torsions -minimize_rna -
cycles 5000 -mute all -filter_lores_base_pairs -vary_geometry -
score:weights <weight> -native <native> -fasta <fasta> -
params_file <params> -add_orbitals
```

## Appendix X – Rotamer Recovery

**command line:**

```
rosetta_scripts.linuxgccrelease -database <database> -l
<pdb.list> @flags.txt
```

**flags.txt:**

```
-ignore_unrecognized_res
-no_optH
-out:nooutput
-jd2:delete_old_poses
```

-add_orbitals
-parser:protocol rr_RTMin_ChiDiff.xml

**rr_RTMin_ChiDiff.xml:**

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <s weights="%%weight%%"/>
      </SCOREFXNS>
      <TASKOPERATIONS>
            <ExtraRotamersGeneric     name=extra_chi     ex1=1     ex2=1
extrachi_cutoff=0/>
      </TASKOPERATIONS>
        <MOVERS>
            <ReportToDB  name=features_reporter  database_mode="sqlite3"
database_name="%%db_name%%" batch_description="t"> batch_description
                  <feature name=ResidueFeatures/>
                  <feature name=ResidueBurialFeatures/>
                  <feature name=ResidueSecondaryStructureFeatures/>
                  <feature name=OrbitalsFeatures/>
                  <feature    name=RotamerRecoveryFeatures    scorefxn=s
protocol=RRProtocolRTMin                      comparer=RRComparerChiDiff
task_operations=extra_chi/>
        </ReportToDB>
        </MOVERS>
        <PROTOCOLS>
                  <Add mover_name=features_reporter/>
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

# BIBLIOGRAPHY

Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." Acta Crystallographica Section B-Structural Science **58**: 380-388.

Andersen, J., O. Taboureau, et al. (2009). "Location of the Antidepressant Binding Site in the Serotonin Transporter: IMPORTANCE OF SER-438 IN RECOGNITION OF CITALOPRAM AND TRICYCLIC ANTIDEPRESSANTS." J Biol Chem **284**(15): 10276-10284.

Apparsundaram, S., D. J. Stockdale, et al. (2008). "Antidepressants targeting the serotonin reuptake transporter act via a competitive mechanism." J Pharmacol Exp Ther **327**(3): 982-990.

Ballester, P. J., I. Westwood, et al. (2010). "Prospective virtual screening with Ultrafast Shape Recognition: the identification of novel inhibitors of arylamine N-acetyltransferases." Journal of the Royal Society, Interface / the Royal Society **7**(43): 335-342.

Barker, E. L., K. D. Burris, et al. (1991). "Phosphoinositide hydrolysis linked 5-HT2 receptors in fibroblasts from choroid plexus." Brain Res **552**(2): 330-332.

Barker, E. L., K. R. Moore, et al. (1999). "Transmembrane domain I contributes to the permeation pathway for serotonin and ions in the serotonin transporter." J Neurosci **19**(12): 4705-4717.

Bell, R. A., R. Faggiani, et al. (1992). "Synthesis, C-13 Nmr, and X-Ray Crystal-Structure of N6,N9-Octamethylenepurinecyclophane." Canadian Journal of Chemistry-Revue Canadienne De Chimie **70**(1): 186-196.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Research **28**(1): 235-242.

Beuming, T., L. Shi, et al. (2006). "A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na+ symporters (NSS) aids in the use of the LeuT structure to probe NSS structure and function." Mol Pharmacol **70**(5): 1630-1642.

Blakely, R. D., H. E. Berson, et al. (1991). "Cloning and expression of a functional serotonin transporter from rat brain." Nature **354**(6348): 66-70.

Boyd, S. (2005). "Molecular operating environment." Chemistry World **2**(9): 66-66.

Bradley, P., L. Malmström, et al. (2005). "Free modeling with Rosetta in CASP6." Proteins **61 Suppl 7**: 128-134.

Bradley, P., K. M. S. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science (New York, NY) **309**(5742): 1868-1871.

Brooks, B. R., C. L. Brooks, 3rd, et al. (2009). "CHARMM: the biomolecular simulation program." J Comput Chem **30**(10): 1545-1614.

Brooks, B. R., C. L. B. III, et al. (2009). "CHARMM: The biomolecular simulation program." Journal of Computational Chemistry **30**(10): 1545-1614.

Canutescu, A. A. and R. L. Dunbrack (2003). "Cyclic coordinate descent: A robotics algorithm for protein loop closure." Protein science : a publication of the Protein Society **12**(5): 963-972.

Carlsson, J., R. G. Coleman, et al. (2011). "Ligand discovery from a dopamine D3 receptor homology model and crystal structure." Nature chemical biology **7**(11): 769-778.

Case, D. A., T. E. Cheatham, 3rd, et al. (2005). "The Amber biomolecular simulation programs." J Comput Chem **26**(16): 1668-1688.

Chen, V. B., W. B. Arendall, 3rd, et al. (2010). "MolProbity: all-atom structure validation for macromolecular crystallography." Acta Crystallogr D Biol Crystallogr **66**(Pt 1): 12-21.

Combs, S. A., S. L. DeLuca, et al. (2013). "Small-molecule ligand docking into comparative models with Rosetta." Nature protocols **8**(7): 1277-1298.

Cornell, W. D., P. Cieplak, et al. (1995). "A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules." Journal of the American Chemical Society **117**(19): 5179-5197.

Coutsias, E. A., C. Seok, et al. (2004). "A kinematic view of loop closure." Journal of computational chemistry **25**(4): 510-528.

Dahan, M., S. Levi, et al. (2003). "Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking." Science **302**(5644): 442-445.

Das, R. and D. Baker (2007). "Automated de novo prediction of native-like RNA tertiary structures." Proc Natl Acad Sci U S A **104**(37): 14664-14669.

Das, R. and D. Baker (2008). "Macromolecular modeling with rosetta." Annual review of biochemistry **77**: 363-382.

Das, R., B. Qian, et al. (2007). "Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home." Proteins **69 Suppl 8**: 118-128.

Davis, I. W. and D. Baker (2009). "RosettaLigand docking with full ligand and receptor flexibility." J Mol Biol **385**(2): 381-392.

Davis, I. W. and D. Baker (2009). "RosettaLigand docking with full ligand and receptor flexibility." Journal of molecular biology **385**(2): 381-392.

Davis, I. W., K. Raha, et al. (2009). "Blind docking of pharmaceutically relevant compounds using RosettaLigand." Protein science : a publication of the Protein Society **18**(9): 1998-2002.

Dougherty, D. A. (1996). "Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp." Science **271**(5246): 163-168.

Dunbrack, R. L. and F. E. Cohen (1997). "Bayesian statistical analysis of protein side-chain rotamer preferences." Protein science : a publication of the Protein Society **6**(8): 1661-1681.

Dunbrack, R. L., Jr. (2002). "Rotamer libraries in the 21st century." Curr Opin Struct Biol **12**(4): 431-440.

Dunbrack, R. L. and M. Karplus (1993). "Backbone-dependent rotamer library for proteins. Application to side-chain prediction." Journal of molecular biology **230**(2): 543-574.

Eildal, J. N., J. Andersen, et al. (2008). "From the selective serotonin transporter inhibitor citalopram to the selective norepinephrine transporter inhibitor talopram: synthesis and structure-activity relationship studies." J Med Chem **51**(10): 3045-3048.

Ewing, T. J., S. Makino, et al. (2001). "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases." Journal of computer-aided molecular design **15**(5): 411-428.

Fichter, K. M., M. Flajolet, et al. (2010). "Kinetics of G-protein-coupled receptor endosomal trafficking pathways revealed by single quantum dots." Proceedings of the National Academy of Sciences of the United States of America **107**(43): 18658-18663.

Fortenberry, C., E. A. Bowman, et al. (2011). "Exploring Symmetry as an Avenue to the Computational Design of Large Protein Domains (vol 45, pg 18026, 2011)." Journal of the American Chemical Society **133**(51): 21028-21028.

Fortenberry, C., E. A. Bowman, et al. (2011). "Exploring symmetry as an avenue to the computational design of large protein domains." Journal of the American Chemical Society **133**(45): 18026-18029.

Friesner, R. A., J. L. Banks, et al. (2004). "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy." Journal of Medicinal Chemistry **47**(7): 1739-1749.

Gademann, K., T. Kimmerlin, et al. (2001). "Peptide folding induces high and selective affinity of a linear and small beta-peptide to the human somatostatin receptor 4." Journal of Medicinal Chemistry **44**(15): 2460-2468.

Gallivan, J. P. and D. A. Dougherty (1999). "Cation-pi interactions in structural biology." Proc Natl Acad Sci U S A **96**(17): 9459-9464.

Gether, U., P. H. Andersen, et al. (2006). "Neurotransmitter transporters: molecular function of important drug targets." Trends Pharmacol Sci **27**(7): 375-383.

Gillespi.Rj (1970). "Electron-Pair Repulsion Model for Molecular Geometry." Journal of Chemical Education **47**(1): 18-&.

Gohlke, H., M. Hendlich, et al. (2000). "Knowledge-based scoring function to predict protein-ligand interactions." Journal of molecular biology **295**(2): 337-356.

Gordon, D. B., S. A. Marshall, et al. (1999). "Energy functions for protein design." Current opinion in structural biology **9**(4): 509-513.

Gront, D., D. W. Kulp, et al. (2011). "Generalized fragment picking in Rosetta: design, protocols and applications." PloS one **6**(8): e23294.

Havranek, J. J., C. M. Duarte, et al. (2004). "A simple physical model for the prediction and design of protein-DNA interactions." Journal of Molecular Biology **344**(1): 59-70.

Havranek, J. J. and P. B. Harbury (1999). "Tanford-Kirkwood electrostatics for protein modeling." Proc Natl Acad Sci U S A **96**(20): 11145-11150.

Henry, L. K., J. R. Field, et al. (2006). "Tyr-95 and Ile-172 in transmembrane segments 1 and 3 of human serotonin transporters interact to establish high affinity recognition of antidepressants." J Biol Chem **281**(4): 2012-2023.

Hess, B., C. Kutzner, et al. (2008). "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation." Journal of Chemical Theory and Computation **4**(3): 435-447.

Hirschfeld, R. M. (2000). "History and evolution of the monoamine hypothesis of depression." J Clin Psychiatry **61 Suppl 6**: 4-6.

Huang, B. and M. Schroeder (2006). "LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation." BMC structural biology **6**: 19.

Hyttel, J. (1994). "Pharmacological characterization of selective serotonin reuptake inhibitors (SSRIs)." Int Clin Psychopharmacol **9 Suppl 1**: 19-26.

Jeffrey, G. A. and D. B. Huang (1991). "Hydrogen bonding in the crystal structure of the tetrasaccharide stachyose hydrate: a 1:1 complex of two conformers." Carbohydr Res **210**: 89-104.

Kalidas, Y. and N. Chandra (2008). "PocketDepth: a new depth based algorithm for identification of ligand binding sites in proteins." Journal of structural biology **161**(1): 31-42.

Kaufmann, K. G., K.; Mueller, R.; Meiler, J. (2008). Small Molecule Rotamers Enable Simultaneous Optimization of Small Molecule and Protein Degrees of Freedom in ROSETTALIGAND Docking. German Conference on Bioinformatic. S. A. Beyer, M. Dresden, Germany**:** 148-157.

Kaufmann, K. W., E. S. Dawson, et al. (2009). "Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies." Proteins **74**(3): 630-642.

Kaufmann, K. W., E. S. Dawson, et al. (2009). "Structural determinants of species-selective substrate recognition in human and Drosophila serotonin transporters revealed through computational docking studies." Proteins **74**(3): 630-642.

Kaufmann, K. W., G. H. Lemmon, et al. (2010). "Practically useful: what the Rosetta protein modeling suite can do for you." Biochemistry **49**(14): 2987-2998.

Kaufmann, K. W. and J. Meiler (2012). "Using RosettaLigand for Small Molecule Docking into Comparative Models." PloS one **7**(12): e50769.

Keeble, A. H., L. A. Joachimiak, et al. (2008). "Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases." Journal of molecular biology **379**(4): 745-759.

Keedy, D. A., Arendall III, W. B., Chen, V. B., Williams, C. J., Headd, J. J., Echols, N., Richardson, J. S., and Richardson, D. C. (2012). "Torsional bioinformatics: 1.5 million quality-filtered residues for better Ramachandran validation." In Preparation.

Kessler, R. C., W. T. Chiu, et al. (2005). "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication." Arch Gen Psychiatry **62**(6): 617-627.

Kortemme, T., A. V. Morozov, et al. (2003). "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes." J Mol Biol **326**(4): 1239-1259.

Kovtun, O., E. J. Ross, et al. (2012). "A flow cytometry-based dopamine transporter binding assay using antagonist-conjugated quantum dots." Chemical Communications **48**(44): 5428-5430.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proc Natl Acad Sci U S A **97**(19): 10383-10388.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proceedings of the National Academy of Sciences of the United States of America **97**(19): 10383-10388.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proceedings of the National Academy of Sciences of the United States of America **97**(19): 10383-10388.

Kuhlman, B. and D. Baker (2004). "Exploring folding free energy landscapes using computational protein design." Current opinion in structural biology **14**(1): 89-95.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science **302**(5649): 1364-1368.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science (New York, NY) **302**(5649): 1364-1368.

Kumar, S. and R. Nussinov (2002). "Relationship between ion pair geometries and electrostatic strengths in proteins." Biophys J **83**(3): 1595-1612.

Kumar, S., C. J. Tsai, et al. (2000). "Factors enhancing protein thermostability." Protein Eng **13**(3): 179-191.

Lange, O. F. and D. Baker (2012). "Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation." Proteins **80**(3): 884-895.

Lange, O. F., P. Rossi, et al. (2012). "Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples." Proceedings of the National Academy of Sciences of the United States of America **109**(27): 10873-10878.

Larsen, M. B., B. Elfving, et al. (2004). "The chicken serotonin transporter discriminates between serotonin-selective reuptake inhibitors. A species-scanning mutagenesis study." J Biol Chem **279**(40): 42147-42156.

Laskowski, R. A. (1995). "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions." Journal of molecular graphics **13**(5): 323-330- 307-328.

Lazaridis, T. and M. Karplus (1999). "Effective energy function for proteins in solution." Proteins **35**(2): 133-152.

Leaver-Fay, A., B. Kuhlman, et al. (2005). Rotamer-pair energy calculations using a trie data structure. Algorithms in Bioinformatics, Proceedings. **3692:** 389-400.

Leaver-Fay, A., M. J. O'Meara, et al. (2013). "Scientific Benchmarks for Guiding Macromolecular Energy Function Improvement." Methods in Protein Design **523**: 109-143.

Leaver-Fay, A., M. J. O&apos;Meara, et al. (2013). "Scientific benchmarks for guiding macromolecular energy function improvement." Methods in enzymology **523**: 109-143.

Lees-Miller, J. P., J. O. Subbotina, et al. (2009). "Interactions of H562 in the S5 helix with T618 and S621 in the pore helix are important determinants of hERG1 potassium channel structure and function." Biophysical journal **96**(9): 3600-3610.

Lemmon, G. and J. Meiler (2012). "Rosetta Ligand docking with flexible XML protocols." Methods Mol Biol **819**: 143-155.

Lemmon, G. and J. Meiler (2012). "Rosetta Ligand docking with flexible XML protocols." Methods in molecular biology (Clifton, NJ) **819**: 143-155.

Li, Z. and H. A. Scheraga (1987). "Monte Carlo-minimization approach to the multiple-minima problem in protein folding." Proceedings of the National Academy of Sciences of the United States of America **84**(19): 6611-6615.

Liang, S. and N. V. Grishin (2002). "Side-chain modeling with an optimized scoring function." Protein Sci **11**(2): 322-331.

Mandell, D. J., E. A. Coutsias, et al. (2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." Nature methods **6**(8): 551-552.

Marshall, S. A., C. L. Vizcarra, et al. (2005). "One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations." Protein Sci **14**(5): 1293-1304.

Mayo, S. L., B. D. Olafson, et al. (1990). "Dreiding - a Generic Force-Field for Molecular Simulations." Journal of Physical Chemistry **94**(26): 8897-8909.

McGaughey, G. B., M. Gagne, et al. (1998). "pi-Stacking interactions. Alive and well in proteins." J Biol Chem **273**(25): 15458-15463.

Meiler, J. and D. Baker (2003). "Coupled prediction of protein secondary and tertiary structure." Proceedings of the National Academy of Sciences of the United States of America **100**(21): 12105-12110.

Meiler, J. and D. Baker (2006). "ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility." Proteins **65**(3): 538-548.

Metropolis, N., A. W. Rosenbluth, et al. (1953). "Equation of state calculations by fast computing machines." The Journal of chemical physics **21**: 1087.

Misura, K., D. Chivian, et al. (2006). "Physically realistic homology models built with ROSETTA can be more accurate than their templates." Proceedings of the National Academy of Sciences of the United States of America **103**(14): 5361-5366.

Misura, K. M. S., A. V. Morozov, et al. (2004). "Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction\." Journal of Molecular Biology **342**(2): 651-664.

Mobley, D. L., A. P. Graves, et al. (2007). "Predicting absolute ligand binding free energies to a simple model site." Journal of molecular biology **371**(4): 1118-1134.

Mooers, B. H. M. and B. W. Matthews (2006). "Extension to 2268 atoms of direct methods in the ab initio determination of the unknown structure of bacteriophage P22 lysozyme." Acta crystallographica Section D, Biological crystallography **62**(Pt 2): 165-176.

Morozov, A. V., J. J. Havranek, et al. (2005). "Protein-DNA binding specificity predictions with structural models." Nucleic Acids Res **33**(18): 5781-5798.

Nations, D., J. G. Eaton, et al. (1991). Stratigraphy, depositional environments, and sedimentary tectonics of the western margin, Cretaceous Western Interior Seaway. Boulder, Colo., Geological Society of America.

Neria, E., S. Fischer, et al. (1996). "Simulation of activation free energies in molecular systems." The Journal of chemical physics **105**(5): 1902-1921.

Organization, W. H. (2009). "Revised global burden of disease (GBD) 2002 estimates."

Perola, E., W. P. Walters, et al. (2004). "A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance." Proteins **56**(2): 235-249.

Petrella, R. J., T. Lazaridis, et al. (1998). "Protein sidechain conformer prediction: a test of the energy function." Fold Des **3**(5): 353-377.

Prasad, H. C., J. A. Steiner, et al. (2009). "Enhanced activity of human serotonin transporter variants associated with autism." Philos Trans R Soc Lond B Biol Sci **364**(1514): 163-173.

RAMACHANDRAN, G. N., C. RAMAKRISHNAN, et al. (1963). "Stereochemistry of polypeptide chain configurations." Journal of molecular biology **7**: 95-99.

Ramsey, I. S. and L. J. DeFelice (2002). "Serotonin transporter function and pharmacology are sensitive to expression level - Evidence for an endogenous regulatory factor." Journal of Biological Chemistry **277**(17): 14475-14482.

Rarey, M., B. Kramer, et al. (1996). "A fast flexible docking method using an incremental construction algorithm." Journal of molecular biology **261**(3): 470-489.

Rohl, C. A. (2005). "Protein structure estimation from minimal restraints using Rosetta." Methods in enzymology **394**: 244-260.

Rohl, C. A., C. E. M. Strauss, et al. (2004). "Modeling structurally variable regions in homologous proteins with rosetta." Proteins **55**(3): 656-677.

Rohl, C. A., C. E. M. Strauss, et al. (2004). "Protein structure prediction using rosetta." Numerical Computer Methods, Pt D **383**: 66-+.

Rohl, C. A., C. E. M. Strauss, et al. (2004). "Protein structure prediction using Rosetta." Methods in enzymology **383**: 66-93.

Rose, G. D. and R. Wolfenden (1993). "Hydrogen bonding, hydrophobicity, packing, and protein folding." Annu Rev Biophys Biomol Struct **22**: 381-415.

Rosenthal, S. J., J. C. Chang, et al. (2011). "Biocompatible Quantum Dots for Biological Applications." Chemistry & Biology **18**(1): 10-24.

Schneider, G. and U. Fechner (2005). "Computer-based de novo design of drug-like molecules." Nature reviews. Drug discovery **4**(8): 649-663.

Schneider, G., M. Hartenfeller, et al. (2009). "Voyages to the (un)known: adaptive design of bioactive compounds." Trends in biotechnology **27**(1): 18-26.

Shapovalov, M. V. and R. L. Dunbrack, Jr. (2011). "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions." Structure **19**(6): 844-858.

Sheffler, W. and D. Baker (2009). "RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation." Protein Sci **18**(1): 229-239.

Shortle, D., K. Simons, et al. (1998). "Clustering of low-energy conformations near the native structures of small proteins." Proceedings of the National Academy of Sciences of the United States of America **95**(19): 11158-11162.

Siegel, J. B., A. Zanghellini, et al. (2010). "Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction." Science (New York, NY) **329**(5989): 309-313.

Simons, K. T., C. Kooperberg, et al. (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." J Mol Biol **268**(1): 209-225.

Simons, K. T., C. Kooperberg, et al. (1997). "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." Journal of molecular biology **268**(1): 209-225.

Simons, K. T., I. Ruczinski, et al. (1999). "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins." Proteins-Structure Function and Genetics **34**(1): 82-95.

Simons, K. T., I. Ruczinski, et al. (1999). "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins." Proteins **34**(1): 82-95.

Singh, S. K., A. Yamashita, et al. (2007). "Antidepressant binding site in a bacterial homologue of neurotransmitter transporters." Nature **448**(7156): 952-956.

Sippl, M. J. (1995). "Knowledge-based potentials for proteins." Curr Opin Struct Biol **5**(2): 229-235.

Still, W. C., A. Tempczyk, et al. (1990). "Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics." Journal of the American Chemical Society **112**(16): 6127-6129.

Tantillo, D. J., J. G. Chen, et al. (1998). "Theozymes and compuzymes: theoretical models for biological catalysis." Current Opinion in Chemical Biology **2**(6): 743-750.

Tomlinson, I. D., J. N. Mason, et al. (2005). "Inhibitors of the serotonin transporter protein (SERT): The design and synthesis of biotinylated derivatives of 3-(1,2,3,6-tetrahydro-pyridin-4-yl)-1H-indoles. High-affinity serotonergic ligands for conjugation with quantum dots." Bioorganic & Medicinal Chemistry Letters **15**(23): 5307-5310.

Tomlinson, I. D., M. R. Warnerment, et al. (2007). "Synthesis and characterization of a pegylated derivative of 3-(1,2,3,6-tetrahydro-pyridin-4yl)-1H-indole (IDT199): A high affinity SERT ligand for conjugation to quantum dots." Bioorganic & Medicinal Chemistry Letters **17**(20): 5656-5660.

Tyka, M. D., D. A. Keedy, et al. (2011). "Alternate states of proteins revealed by detailed energy landscape mapping." J Mol Biol **405**(2): 607-618.

Verdonk, M. L., J. C. Cole, et al. (2003). "Improved protein-ligand docking using GOLD." Proteins **52**(4): 609-623.

Wang, G. and R. L. Dunbrack (2003). "PISCES: a protein sequence culling server." Bioinformatics (Oxford, England) **19**(12): 1589-1591.

Wang, G. and R. L. Dunbrack, Jr. (2003). "PISCES: a protein sequence culling server." Bioinformatics **19**(12): 1589-1591.

Wedemeyer, W. J. and D. Baker (2003). "Efficient minimization of angle-dependent potentials for polypeptides in internal coordinates." Proteins **53**(2): 262-272.

Word, J. M., S. C. Lovell, et al. (1999). "Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation." Journal of Molecular Biology **285**(4): 1735-1747.

Yarov-Yarovoy, V., J. Schonbrun, et al. (2006). "Multipass membrane protein structure prediction using Rosetta." Proteins-Structure Function and Bioinformatics **62**(4): 1010-1025.

Zanghellini, A., L. Jiang, et al. (2006). "New algorithms and an in silico benchmark for computational enzyme design." Protein Science **15**(12): 2785-2794.