The Road to Virtual Chemistry: Computer-Aided Molecular Design Skirting the Boundary between Structure and Ligand-Based Approaches

By

Joshua Mark Pottel

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Department of Chemistry, McGill University
Montréal, Québec, Canada
September 2015

I dedicate this thesis to my family: my partner in crime Lara Reichman, my father Ronnie Pottel, my mother Rona Edelman, my step-parents Frances Paspaliaris and David Sprintis, my sister Jillian Rosenberg and my brother Ryan Pottel, for always supporting my dreams and pursuits and to my friend and mentor, Nicolas Moitessier, for inspiring, motivating and guiding me for the past 4 years.

Abstract

Novel synthetic methodologies that are effective, environmentally friendly and efficient are becoming ever increasingly difficult to design. In fact, the development of such techniques is often labor-intensive, wasteful and costly. Several years ago, instrumental techniques, such as nuclear magnetic resonance, high performance liquid chromatography and mass spectrometry, were integrated into the chemistry toolbox and their maturity significantly accelerated the process of synthetic discovery. Surprisingly and contrastingly, computational advances have not yet been incorporated as far as the imagination can take them. Fifty years after Gordon E. Moore, the co-founder of Intel, first predicted a yearly two-fold expansion of computational power, we are attaining its peak, and yet information technologies are still under-utilized in chemical settings.

Until now, computational techniques employed in designing chemical and biochemical synthesis have been merely a tease. Expanding the abilities of computational molecular discovery methods is an attractive solution to exploring a vast amount of unknown synthetic approaches. Furthermore, making these virtual methodologies accessible to the organic and medicinal chemistry communities will allow them to reach their full potential. Currently, only a handful of research groups in Canada truly blend computational and organic chemistry and often in a rationalization capacity rather than as a design strategy.

This thesis describes efforts to develop new computational design approaches for small molecules and biological structures and apply them to organo- and biocatalytic research programs. A contemporary perspective on software programs is required for their inclusion in the chemistry toolbox for several reasons. Currently, hundreds, if not thousands, of computational chemistry software packages exist; however, in most cases, it is only the developers that make use of these tools, the ones that simulate chemical phenomena, as opposed to visualization software suites. A significant lack of usability – simplicity in running routine experiments – is likely one of the largest causes for this disappointing reality. Accurate results are also necessary to build trust from the experimental chemistry community. These issues were the focus of this work to

demonstrate that the integration of computational tools within organic chemistry is not only plausible, but increasingly valuable when no advanced, expert training is necessary.

More specifically, we have created and developed several computational tools, FINDERS, REACT2D, CONSTRUCTS and ACE 2.2, in order to automate and guide the design, discovery and synthesis of asymmetric organocatalysts. Their validation and full integration into one easy-to-use software platform, the VIRTUAL CHEMIST, establishes the potential significance of such an innovation. In addition, these concepts have been applied to biological systems, mainly cytochrome P450 metabolic enzymes and zinc-containing metalloenzymes involved in several disease-related pathways. The transition state modeling approach that is implemented in ACE was integrated into our docking software, FITTED, and automated to yield IMPACTS, a site of metabolism prediction software that includes a trivalent approach. Moreover, a proton-shuttling mechanism was modeled into FITTED, significantly improved results of zinc-binding and demonstrated the positive effect of accurately modeling biochemical phenomena. The final goal was to successfully create a single point mutation protocol to simulate protein engineering in an efficient manner and to effectively model biocatalysis. By combining this procedure with IMPACTS it would be possible to virtually, and accurately, mutate cytochrome P450s, craft new transition states and eventually design new biochemical reactions. In all, the work in this thesis represents efficient approaches to improve a wide range of computational chemistry applications and demonstrate their viability, value and rightful place in the chemistry toolbox.

Résumé

La conception et le développement de nouvelles méthodes de synthèse efficaces et respectueuses de l'environnement deviennent de plus en plus difficiles. En fait, le développement de ces techniques est souvent dispendieux, requiert beaucoup de travail de laboratoire et produit une grande quantité de déchets. Depuis plusieurs années, des techniques instrumentales telles que la résonance magnétique nucléaire, la chromatographie liquide à haute performance et la spectrométrie de masse ont été intégrées dans la boîte à outils de la chimie et leur maturité a significativement accéléré le procédé de découverte en chimie organique. Cependant, il est surprenant que les avancées informatiques n'aient pas encore été incorporées aussi loin que l'on pourrait l'imaginer. Cinquante ans après que Gordon E. Moore, le co-fondateur d'Intel, ait prédit pour la première fois le doublement annuel de la puissance de calcul des ordinateurs, nous atteignons présentement son sommet, et pourtant les technologies de l'information demeurent sous-utilisées en chimie.

Jusqu'à maintenant, les techniques informatiques employées dans la synthèse chimique et biochimique n'ont été qu'un préambule. Améliorer la découverte en chimie guidée par des méthodes informatiques représente une solution attrayante pour explorer une vaste quantité d'approches synthétiques inconnues. De plus, la possibilité de rendre ces technologies virtuelles accessibles aux communautés de la chimie organique et des sciences pharmaceutiques leur permettra d'atteindre leur plein potentiel. En ce moment, seulement une poignée de groupes de recherche au Canada combinent véritablement la chimie informatique et la chimie organique et ce, en utilisant souvent une approche de rationalisation plutôt qu'en tant que stratégie de conception.

Cette thèse décrit les efforts visant à développer de nouvelles approches de conception assistée par ordinateur pour la recherche de petites molécules et structures biologiques ayant des propriétés organo- ou biocatalytiques. Une vision plus actuelle sur les logiciels de programmation et leur inclusion dans le jeu d'outils des chimistes organiciens est nécessaire pour plusieurs raisons. Actuellement, des centaines, sinon des milliers de logiciels de chimie informatique existent. Toutefois, dans la majorité des cas, ce sont seulement les développeurs qui font usage de ces outils (ceux qui modélisent les

phénomènes chimiques et non les interfaces graphiques). Un manque important de convivialité et de simplicité pour des calculs de routineest l'une des plus grandes causes d'une réalité qui n'est pas à la hauteur des possibilités offertes. Des résultats fiables sont également nécessaires pour bâtir la confiance de la communauté de chimie expérimentale. Ces préoccupations ont été au cœur de nos travaux qui visaient à démontrer que l'intégration d'outils informatiques au sein de la chimie organique est non seulement possible, mais aussi de plus en plus précieuse lorsqu'une formation avancée d'expert n'est pas requise.

Plus précisément, nous avons créé et développé plusieurs outils informatiques, FINDERS, REACT2D, CONSTRUCTS et ACE 2.2, dans le but d'automatiser et guider la conception, la découverte et la synthèse d'organocatalyseurs asymétriques. Leur validation et leur pleine intégration dans une plate-forme de logiciels facile à utiliser, le VIRTUAL CHEMIST (Chimiste virtuel), pourra établir l'importance d'une telle innovation. De plus, ces concepts ont été appliqués à des systèmes biologiques, principalement les enzymes métaboliques cytochrome P450 et les métalloenzymes au zinc qui sont impliquées dans plusieurs voies liées à des maladies. La modélisation des états de transition à partir D'ACE a été intégrée dans notre logiciel de « docking », FITTED, et automatisée menant au logiciel de prédiction de sites de métabolisme, IMPACTS. En outre, la modélisation d'un mécanisme de transport de protons a aussi été implémenté dans FITTED et a amélioré les résultats de liaison du zinc de façon significative. Notre dernier objectif était de créer un protocole de mutation en un seul point afin de simuler l'ingénierie de protéines efficace et de modéliser avec précision les processus de biocatalyse. En combinant cette procédure avec IMPACTS, il est possible de muter les cytochromes P450 de façon efficace et précise, de créer de nouveaux états de transition et par conséquent, de concevoir de nouvelles réactions biochimiques et de nouveaux biocatalyseurs. En résumé, le travail décrit dans cette thèse présente des approches efficaces qui contribuent à un large éventail d'applications en chimie informatique et démontre leur viabilité, leur valeur et leur place légitime dans le coffre à outils de la chimie.

Traduit de l'anglais par Katherine Bujold

Acknowledgments

There are many people that deserve credit for helping me through my time as a graduate student at McGill. First, my supervisor, Nicolas Moitessier, for guiding me through the past 4 years and always supporting my research ideas. Our countless brainstorming sessions, that weren't arguments, led to a lot of success and he never held me back from reaching my goals. He taught me what it meant to be a scientist, a mentor and most importantly, how to treat people with respect.

I would also like to thank Jim Gleason for leading great collaborations and being patient and attentive in meetings. He always took the necessary time to teach when I didn't understand and made himself available for advice and support.

When I joined the lab, Eric Therrien was there from day one offering me his time and knowledge whenever I came asking. He taught me a lot of the fundamentals of computational chemistry and even after he left the lab, still helps me with whatever I need. We missed him for a few program naming sessions.

Michelle Bezanson, Paolo Schiavini and Sylvain Rocheleau were patient with me during our various collaborations. They were relatively patient while teaching me "how to" chemistry and biochemistry. They taught me all I need to know about NMR and the horrors of Halifax, assays and the Italian Mafia, and sugars and insects.

Anna Tomberg and Leo Liu, the CCG team, were great venting partners and always offered support with the cluster and software management issues. We always had great debates about force fields and travelling with Anna was fun, even if I'll never admit it. Additionally, I'd like to thank all past and present members (Moeed, Gaëlle, Jessica, Sébastien...) of the Moitessier group that I had the pleasure of meeting and working with, including Jerry, even if he's a Toronto fan

The chemistry department and specifically Chantal Marotte offered a great support system; the department offered me several scholarships and Chantal made navigating complicated applications and graduate school in general a breeze.

Stephane DeCesco, although we never collaborated explicitly, never turned down Timmies runs or conversations about anything science, Canadiens or Impact. Pleky and

Sadie's friendship would probably be like ours; we really see eye-to-eye even though everyone thinks we're always sniping at each other. I think I still owe him a conformational search. Vanja was my seat buddy and I could not have asked for anyone more friendly or supportive.

Outside of the lab and into my honorary group, David Polcari proofread any document I sent him and never shied away from telling me his honest opinion about my work or my decisions. From sports to BBQ-ing, we did most McGill things together and became great friends outside of both our labs. Maybe someday he'll join Philippe Dauphin-Ducharme and me on the west coast. He might skate faster than I do, but I won't forget the time he challenged me to a running race.

My decision to study at McGill was made easier by my friends in the Holmes Room Crew. Katherine never passed up an opportunity to tell me about the awesomeness of Magog. Nicole (dun, dun, dun, Verdun), Janet (drinking buddy), and Matt (reluctantly listed you by name...) always kept the group hangout going. Eventually we will all be able to get together at the same time.

I owe a big thanks to all the sports crews. This kept me sane and also relatively healthy since I sat at a computer all day. Degenerate Orbitals, H2Owned, Nice Snatch, Bone Patrol, Huckel's Rule, To Kill a Blocking Nerd and Nash Potatoes & McGrady were like family and everyone on those teams put up with my pregame speeches and often let my desire to win go unchecked, especially Lana Greene. Special shout-out to Graham Hamblin, Jack Cheong and Mitch Huot for Dyslexia United and whatever happened in 2v2 soccer, I've mostly blocked it from my mind. To everyone else that I haven't mentioned here, thanks so much for wonderful teams, outings and strategic planning, especially Andrew Danis, those are times that I will never forget and it was a pleasure being on the field/court/pool/ice with you.

I am grateful to my long-time Montreal friends and essentially family, Zack Goldig, Geoff Crampton, Alex Leibner and Adam Gordon. You guys were never really sure what I did, but I did my best to explain it, even on the bathroom floor. Our history is endless and I owe you guys more than you can imagine. That being said, I'm going to beat you all at fantasy football and I hope you'll come visit us in San Francisco.

My family has always been there to support me and any success I achieve will forever be shared with them. From Montreal to Florida to San Jose, I could always count on a phone call to vent my frustrations, share my concerns and brag when things were going well. The final thank you goes to Sadie, I mean Lara, my best friend who I can always count on to take care of me and be there for me whenever I ask. Without her, none of this would have been possible. I can never thank this group of people enough.

Table of Contents

List of Figures

List of Tables

List of Equations

List of abbreviations and terms

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
6-31G*	Basis set name
A	Constant in energy functions
ACE	Angiotensin-converting enzyme
ARG	Arginine
ASN	Asparagine
ASP	Aspartic acid
AU-ROC	Area under receiver operating characteristic
AVG	Average
B	Constant in energy functions
B3LYP	Becke, 3-parameter, Lee-Yang-Parr (density functional)
BFS	Breadth-first search
BO	Bond order
c_n	Weight constants
CA	Carbonic anhydrase
CADD	Computer-aided drug design
CD	Computational development
Compound I	Reactive state of the iron-oxo heme system
CPU	Central processing unit
CYP	Cytochrome P450
CYS	Cysteine
δ	Equilibrium torsion
$\Delta G^{\ddagger}$, $\Delta H^{\ddagger}$, $\Delta S^{\ddagger}$	Gibbs, enthalpy and entropy energies of activation respectively
ΔG_{SASA}	Solvent accessible surface area
$\Delta G_{solvation}$	Generalized Born surface area solvation free energy
D_e	Dissociation energy
d.e.	Diastereomeric excess
DFT	Density functional theory
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DUD	Directory of useful decoys
ε	Energy-well depth

E_a	Energy of activation
E_x	Potential energy term representing "x" (bond, angle, docking …)
FF	Force field
GA	Genetic algorithm
GAFF	Generalized amber force field
GLN	Glutamine
GLU	Glutamic acid
GSFF	Ground state force field
h	Planck's constant
HDAC	Histone deacetylase
HIS	Histidine
ILE	Isoleucine
κ	Transmission coefficient
K_a	Acid dissociation constant
K_B	Boltzmann constant
k_b, k_s, k_t	Force constants
L	Ligand
LD	Laboratory development
LEU	Leucine
LJ	Lennard-Jones
LYS	Lysine
MA	Matching algorithm
MD	Molecular dynamics
MET	Methionine
MM	Molecular mechanics
MMP	Matrix metalloproteinase
MP2	Møller–Plesset perturbation theory to the second order
Mol2	A molecular file format
n	Rotational periodicity
N	Number of ligands/substrates
N_{rot}	Number of rotatable bonds
N_{water}	Number of bridging water molecules
NADPH	Nicotinamide adenine dinucleotide phosphate
PDB	Protein Data Bank
PES	Potential energy surface
pH	$-\log([H^+])$

PHE	Phenylalanine
pK_a	$-\log(K_a)$
PM3	Parameterized model 3 (semi-empirical method)
POP	Prolyl oligopeptidase
PRO	Proline
q	Point-charge
QM	Quantum mechanics
QSAR	Quantitative structure-activity relationship
r_0	Equilibrium bond distance
R	Gas constant
RHF	Restricted Hartree-Fock
RMSD	Root mean square deviation
RNA	Ribonucleic acid
σ	Zero-energy point
S_{dock}	Non covalent FITTED scoring function RankScore
$Score_{TS}$	IMPACTS scoring function
SER	Serine
SoM	Site of metabolism
θ_0	Equilibrium angle
T	Temperature
THR	Threonine
TRP	Tryptophan
TS	Transition state
TSFF	Transition state force field
TST	Transition state theory
TYR	Tyrosine
UHF	Unrestricted Hartree-Fock
VAL	Valine
VBT	Valence bond theory
vdW	Van der Waals
VS	Virtual screening
WT	Wild type of an enzyme
χ	Torsion value
ZBG	Zinc binding group

Contributions to Original Knowledge

1) The review written by the author of this thesis in chapter 1 is a summary of available techniques in the field of transition state modeling. The invitation to contribute our own expertise and original knowledge and opinions are found in this chapter.

2) The computational platform, VIRTUAL CHEMIST, reported in chapter 2, is unique and never has such an innovation been reported to the knowledge of the author. The automated protocols use a novel molecular representation that differs from the standard SMILES format. The copyright and report of invention belonging to the author further demonstrate the originality of the software.

3) Chapter 3 contains further expertise and original knowledge in the form of another invited review. The description of FITTED, its development and applications depict its usefulness and originality in the field of molecular docking.

4) Simulation of the biochemical phenomena reported in chapter 4 was the first successful depiction of the dynamic proton-shuttling mechanism in a static docking software program. The enormous improvement seen in the docking to zinc metalloenzymes validated the new implementation. Several reports from other research groups followed this original publication.

5) The trivalent approach to site of metabolism prediction reported in chapter 5 is unique to this published work. The combination of ligand reactivity, molecular docking and transition state modeling was demonstrated to be effective and accurate. The validation set used was heavily curated and is often referenced in reviews and other published works as being appropriate and well crafted.

6) The statistical libraries reported in chapter 6 are extensive and cover more data than in previously reported conformation libraries. More importantly, the developed technique to generate new libraries was validated and reported.

7) All software reported in this thesis is made available to academic groups worldwide free of charge and is already in use at several institutions.

Author Contributions

Nicolas Moitessier provided funding, research objectives, and intellectual guidance for all of the projects described in this thesis.

Chapter 2

This chapter consists of software design and several validation experiments to demonstrate its accuracy. ACE was previously designed by **Nicolas Moitessier** and **Christopher R. Corbeil**, and was modified in this work. The remaining software developments and applications were contributions of the author of this thesis.

Chapter 3

Eric Therrien contributed to the knowledge in sections 3.3 and 3.4.7. **Pablo Englebienne** contributed to the knowledge found in sections 3.4.1 and 3.4.2. **Zhaomin Liu** contributed to G-Quadruplex knowledge in section 3.4.2. **Anna Tomberg** contributed to the knowledge in section 3.4.6 and **Christopher R. Corbeil** contributed to the knowledge of sections 3.4.1, 3.4.2 and 3.4.7. All other sections, figures and overall editing were contributions of the author of this thesis.

Chapter 4

Eric Therrien initiated the idea of having docking to metalloenzymes within FITTED with **James L. Gleason** as part of a team grant and they contributed to the knowledge of using the FORECASTER/FITTED suites of programs and the knowledge of HDACs and their inhibitors respectively. All coding and application experiments were contributions of the author of this thesis.

Chapter 5

Valérie Campagna-Slater encoded the transition-state modeling into FITTED to create the foundation for IMPACTS. **Eric Therrien** contributed to the knowledge of using the FORECASTER/FITTED suites of programs and **Louis-David Cantin** was a co-P.I. on a grant and contributed to the knowledge in medicinal chemistry and metabolism. Other coding and validation experiments were contributions of the author of this thesis.

Chapter 6

This chapter consists of software design and statistical evaluation of its accuracy. The reported developments and applications were contributions of the author of this thesis.

The work in Chapters 1, 4, and 5 has been previously published.

Introduction to Chapter 1

The goals of this thesis are two-fold. First, the development of accurate computational tools for asymmetric organocatalysis and biocatalysis was the primary focus. The general strategy for achieving this objective was to first identify unattained computational modeling milestones in these domains, an automated asymmetric catalyst discovery, for example. Next, the relevant chemical or biochemical phenomena were determined and were often the missing piece from existing software packages. Frequently, modeling necessitates shortcuts in order to minimize the time requirements, but certain omissions can be impactful. Finally, accurate modeling of these phenomena was a main focus which led to improved results seen throughout validation experiments.

The second aim of this thesis was to consider the accessibility to such software by organic chemists without any necessary expertise or advanced training in computational techniques or programming. The keys to creating handy simulation packages are to automate, simplify input and output, and be fast. The general procedure to develop such tools was to recognize, with the help of organic chemists, the unmet needs or support for laboratory experiments and synthetic plans. Then, as with chemical phenomena, these steps must be accurately modeled and encoded. Specifically, in a given synthetic methodology, a chemical reaction is identified, reagents are purchased, catalysts are prepared and then applied in a second chemical reaction so that those resulting in the best enantioselectivities can be labeled and kept for future reactions. This can take months or even years for proper screening while software can reduce this to days.

In summary, this thesis aims at developing novel computational chemistry protocols that are both accurate and accessible by correctly modeling chemical reactions that could be ongoing in a flask or a Petri dish. The primary linking topic is transition state modeling since the majority of chemical reactions must pass through an activation barrier to break bonds. This chapter presents an overview of transition state modeling techniques available to the modeling community and reviews their development, validation, and availability.

(This page was left blank intentionally)

Chapter 1:

Introduction – Efficient Transition-State Modeling using Molecular Mechanics Force Fields for the Everyday Chemist

This chapter is currently in press and is reproduced from the invited book chapter: "Efficient Transition-State Modeling using Molecular Mechanics Force Fields for the Everyday Chemist", Pottel, J.; Moitessier, N.; *Reviews in Computational Chemistry*, **2015**, *29*, in press. Wiley (2015).

1.1 Introduction

Computational chemistry and computer-assisted molecular modeling have advanced tremendously due to their efficiency in clarifying chemical problems and offering insights that may otherwise be missed. While the throughput of quantum mechanical (QM) methods remains low, molecular mechanics (MM) computations are significantly faster and can, nowadays, be applied to the study of large systems (e.g., molecular dynamics simulations of proteins in aqueous medium) and/or large libraries of small molecules (e.g., screening of thousands of small molecules in drug discovery). In fact, in computational chemistry, there is always a struggle between obtaining the most accurate calculations and the time needed to perform these computations. Over the past few decades, both sides of this conflict have been improved by the development of better force fields (this concept will be explained further in this chapter) and faster computers. Empirical force fields (FFs) can be described as sets of mathematical equations and parameters (derived empirically) used in MM to describe the potential energy of a molecule. FFs can be further used to describe atomistic movements (e.g., molecular dynamics) and molecular properties (bond vibrations) and to predict the outcome of an experiment (e.g., IR spectra). FFs are often embedded in software and, along with the visualization software that exists, can be used without much expertise or extensive knowledge of how they were built or developed (although, as described below, many FFs exist for different purposes and selecting the correct one can be crucial to the success of a project). The progress in this field is directed towards many different molecular studies

such as molecular dynamics, conformational analysis of small to very large systems (small molecule catalysts up to proteins), as well as equilibrium and transition state (TS) structure modeling. In turn, all of these areas cover applications in the fields of chemical biology, medicine, materials and more. The evolution of FFs from methods used for simple potential energy minima identification to those used for complex TS modeling has followed the trends in experimental chemistry, the needs of the chemical and pharmaceutical industries, and, importantly, the availability of computing power.

In the sub-disciplines of synthetic, medicinal and process chemistry, having more efficient and greener catalysts is required as a response to increasing economic and environmental pressures. Consequently, there has been a drive to better understand chemical reactions and transformations and, furthermore, to develop new, efficient and cleaner reactions. Computational chemists have reacted to these needs and developed methods to guide experimentalists in the development of these reactions. A highly accurate but time-consuming option is to use QM techniques. They are based on non-intuitive concepts and can be difficult to comprehend for some bench chemists who are trained in valence-bond views of bond making and bond breaking; this option is often avoided by those with little expertise in quantum chemistry and molecular modeling because meaningful results can only be obtained with advanced knowledge of the underlying techniques and methods (e.g., many of the existing density functionals have severe limitations and, while they may work well for one class of materials, they do not work well for others). In contrast, MM techniques offer more intuitive, user-friendly, black-box methods for organic and medicinal chemists to use for molecular modeling – assuming they are accurate enough for the problem under investigation.

In the field of synthesis, TS modeling with MM is very important but far less developed than the MM modeling of energy minima (i.e., ground states). This chapter is designed to first educate the non-expert, whether industrial or academic, in the subject of MM and FFs, assuming a basic knowledge of chemical principles, and second, to show how to apply this theory to TS modeling. An overview of current MM-derived techniques used in TS modeling will be given, discussing the theory, advantages, potential drawbacks and availability of software packages. The aim is to offer students, researchers and teachers a foundation of understanding in order to be comfortable with using the

available procedures while being aware of the concerns and potential drawbacks of different methodologies. For a more detailed review of the theory and methods, see a review from Corbeil and Moitessier.[1]

This review focuses on methods that are well documented. Other, less well documented, methods are available, but due to lack of information, these will not be detailed herein. Among them is a TS modeling technique implemented in PCMODEL available at http://www.serenasoft.com/pcm8.html and a TS searching technique that uses QM for refinement known as AFIR, artificial force induced reaction.[2]

1.2 Molecular Mechanics and Transition State Basics

1.2.1 Molecular Mechanics

MM is often used to calculate equilibrium structures (ground state geometries in three dimensional space with minimum potential energy often referred to as "stable conformations"), energies associated with these structures and an assortment of other properties using classical mechanics as mentioned in the introduction.[3-7] MM is usually taught in secondary and post-secondary education as a ball and spring model where atoms "feel" classical forces between them. Explicit electron considerations are omitted in this model (in more advanced versions, polarization may be considered[8] and electron lone pairs may be introduced for directionality in hydrogen bonding[9]) and the potential energy surface (PES) is described by functions that characterize spring stretching, bending, dihedral (torsional) angles and more. These functions are parameterized and critiqued based on experimental data such as X-ray diffraction and NMR data[10-18] or the more accurate computational techniques that do take into consideration electronic effects and more complex phenomena. The functions represent a PES (Figure 1.1) with a dimensionality that depends on the equations used. A very basic FF is the combination of a set of functions (Eqs. 1.1-1.6) and associated parameters, assuming simple additivity of these energy terms (i.e., bond energy is independent of angle energy).

$$E_{\text{total}} = \sum E_{\text{bond}} + \sum E_{\text{angle}} + \sum E_{\text{torsion}} + \sum E_{\text{van der Waals}} + \sum E_{\text{electrostatic}} \quad \textbf{(1.1)}$$

where, for example,

$$E_{\text{bond}} = k_{\text{s}} \times (r - r_0)^2 \quad \textbf{(1.2)}$$

$$E_{\text{angle}} = k_{\text{b}} \times (\theta - \theta_0)^2 \quad \textbf{(1.3)}$$

$$E_{\text{torsion}} = k_{\text{t}} \times (1 \pm \cos(n\varphi - \delta)) \quad \textbf{(1.4)}$$

$$E_{\text{van der Waals}} = \frac{A}{r^{12}} - \frac{B}{r^6} \quad \textbf{(1.5)}$$

$$E_{\text{electrostatic}} = \frac{q_i \times q_j}{r_{ij}} \quad \textbf{(1.6)}$$

In these equations k_s, k_b and k_t are force constants, r_0, θ_0 and δ are equilibrium bond length, angle and torsion values, n is the rotational periodicity, A and B are constants based on the interaction between two atoms and q is a point-charge. All of these values, known as parameters, can be different depending on the potential functions comprising the FF. The sum of all of the energies for the aforementioned terms equals the total energy of the system. What is described here is a very general set and only a simple example of a FF; a more elaborate look at possible functions for each of the terms in Eq. 1.1 can be found in a review from Pettersson and Liljefors.[19]

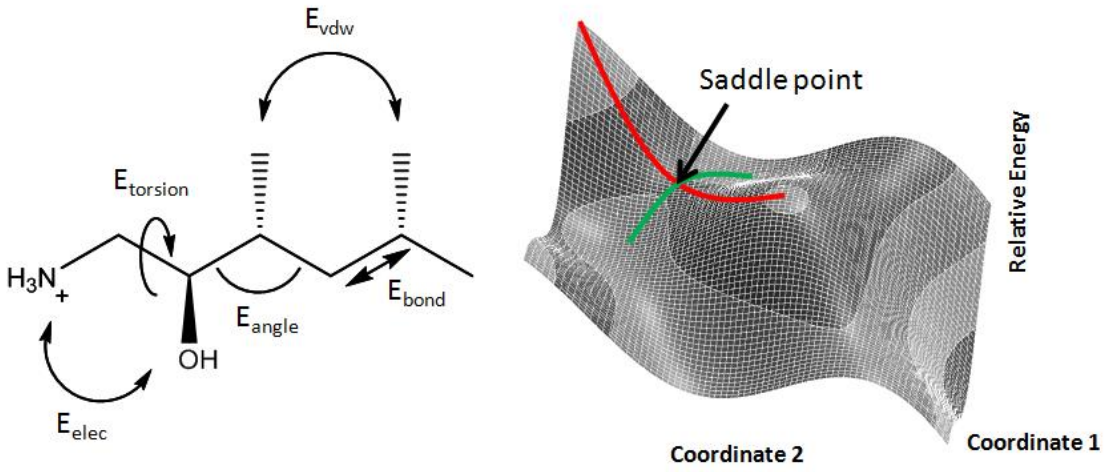


Figure 1.1. Some of the different energy functions accounted for in a FF and a hypothetical 3-dimensional potential energy surface (only 2 components of E_{tot} considered) that a particular FF is intended to reproduce.

For comparison purposes, consider the range in complexity, accuracy and computation time for two different functions. The CHARMm FF[20] employs a bond stretching term similar to Eq. 1.2 while the MM3 FF[21-23] uses more terms (Eq. 1.7) in order to better simulate the anharmonicity that is described by the Morse Potential (Figure 1.2).[24] Eq. 1.2 is the harmonic approximation of the Morse Potential.

$$E_{\text{bond}} = k_s \times (r - r_0)^2 \times \left(1 - 2.55 \times (r - r_0) + \frac{7}{12} \times 2.55^2 \times (r - r_0)^2\right) \quad \textbf{(1.7)}$$

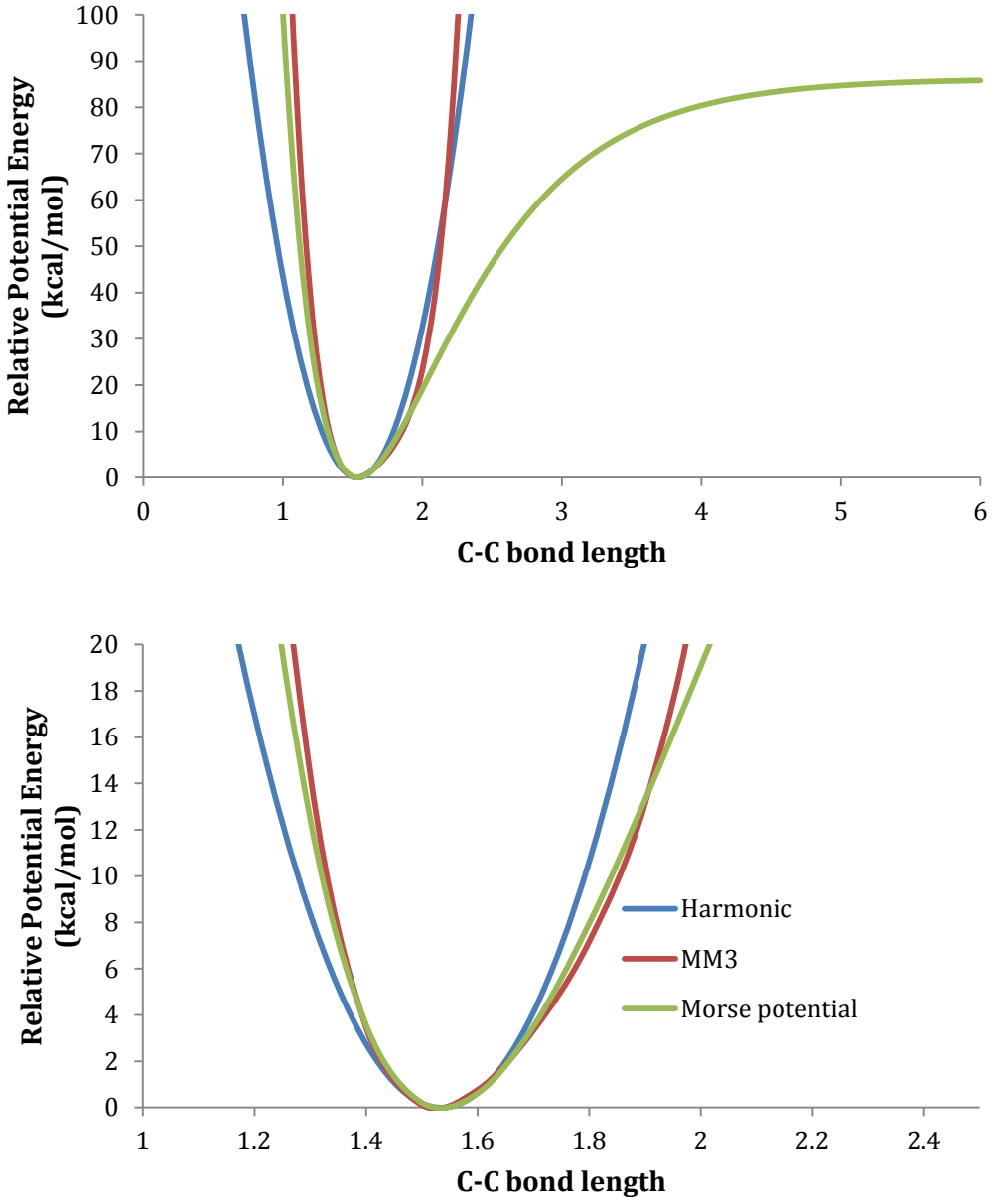


Figure 1.2. Harmonic, MM3 and Morse potentials. Top graph: overall curve; bottom graph: expansion of the equilibrium bond length within 20 kcal/mol distortion energy.

The harmonic approximation can often model the equilibrium energy adequately and the computing speed is fast while functions closer to the Morse Potential can model the equilibrium bond lengths correctly as well as those lengths that are far from equilibrium (Figure 1.2). Note that the variable r is a squared term in Eq. 1.2 whereas in the more complex, accurate Eq. 1.7, the r terms range from squared to the fourth power. Also note that Eq. 1.2 is significantly faster to calculate because there are fewer CPU operations to execute. This trade-off between speed and accuracy must be taken into consideration when contemplating the size of the system to be studied or the needs of the user. For many small molecule studies, the computation time will not be large and the accuracy of a non-equilibrium bond length may be vital, suggesting the use of Eq. 1.7, whereas for proteins the time requirements can be significant (many more atoms and bonds) and the accuracy of predicting a bond length using Eq. 1.2 can be satisfactory since long-range effects often dominate. One must keep in mind that this example pertains only to E_{bond} and other energy terms have their respective computational cost and accuracy.

It is important to note that while the absolute value of E_{tot} and the other energy terms are not significant, the relative values are useful because they can indicate preferred energy minima (e.g., conformers), or potential TSs (i.e., energy maxima) which will be described in more detail below. As shown in Figure 1.2, the harmonic approximation (red curve) has a minimum at potential energy = 0, while a C-C bond has a dissociation energy of ca. -80 kcal/mol which means E_{bond} = -80 kcal/mol at optimal distance. Thus, the harmonic approximation provides relative bond energy values and not absolute values. Consequently, comparisons are only meaningful when comparing molecular systems with the same set of parameters such as conformers, stereoisomers and some regioisomers. When necessary, comparing FF-derived energies of molecules with different atom connectivity (including regioisomers) and bond structures must be done cautiously. To solve this issue, MM3 provides information on heats of formation, although, once more, the provided data should be taken with caution.

For a lengthy list of FF parameters, consider the review from Jalaie and Lipkowitz.[25] Advances have been made to most of these FFs since then and new ones have been developed with varying applications.[26-29] One area that has been explored is that of TS modeling.[30] Lowering activation barriers with catalysts, discovering mechanisms and

predicting reaction outcomes accurately are crucial for efficient, environmentally-friendly and cost-reducing chemistry.

1.2.2 Transition States

A TS is characterized as a geometry representing a saddle point (first-order) on the PES, which is a maximum on the reaction coordinate, but minimum in all other directions (Figure 1.1). TSs are either states between two conformations of the same molecule (a conformational change described by bond rotation, stretching, etc.) or states between reactants and products in a chemical transformation (configuration changes are observed between reactants A and B forming product C, for example). This review focuses on the latter since analyzing TSs resulting from conformational changes does not involve bond breaking/making and because these energy barriers are often included in the parameterization process[31] and they have been studied for over 50 years.[32-33] Thus, throughout this chapter, whenever we refer to a TS, the connotation is that of a chemical reaction and not a conformational transformation.

The most common and simple way to describe chemical reactivity relies on transition-state theory (TST).[34] The energy difference between reactant(s) and TS structure(s) is the activation energy (Figure 1.3) which defines the reaction rate as shown in the Arrhenius and Eyring equations (Eqs. 1.8-1.9),

$$k = Ae^{\frac{-E_a}{RT}} \qquad (\mathbf{1.8})$$

$$k = \kappa \frac{K_B T}{h} e^{\frac{-\Delta G^{\ddagger}}{RT}} = \kappa \frac{K_B T}{h} e^{\frac{\Delta S^{\ddagger}}{R}} e^{\frac{-\Delta H^{\ddagger}}{RT}} \qquad (\mathbf{1.9})$$

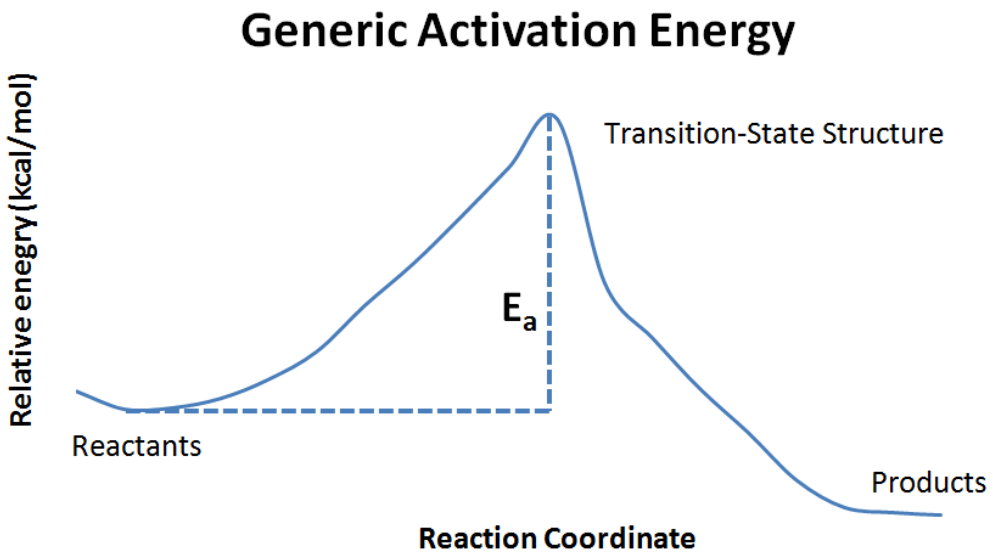


Figure 1.3. An example of a 1-D PES illustrating the energy of activation (E_a) and TS structure, which is to be located in TS determination.

Here, k is the reaction rate, A is the pre-exponential constant, E_a is the activation energy, R is the gas constant, T is the temperature, κ is the transmission coefficient, K_B is Boltzmann's constant, h is Planck's constant, $\Delta G^{\ddagger}$ is the Gibbs energy of activation, $\Delta S^{\ddagger}$ is the entropy of activation and $\Delta H^{\ddagger}$ is the enthalpy of activation.

As presented in Eq. 1.9, the rate should be defined by the total free energy difference between reactant and transition structure. Thus all contributors, such as all motions (vibrations, rotations, translations), should be considered but these properties can be difficult to obtain. Generally, it is assumed that these values for vibration, rotation and translation are similar for different TS structures of a given reaction (e.g., diastereomeric TS structures) and thus can be neglected when computing a relative energy of activation (ΔE_a). The actual energy barrier (E_a) is of course more intricate than this approximation and will not be covered going forward since it requires explicit, accurate treatments of the above mentioned contributors to the energy function.[35]

A recurrent problem in computational chemistry is the time-accuracy trade-off. Using MM is much faster than using QM; however, the accuracy is not expected to be comparable, unless advanced training of the FF is carried out on the chemical system under investigation.[36] Nevertheless, there have been many attempts at computing or

modelling the PES or finding other creative ways to put together the MM principles with the goal of finding TS structures. Some generalized FF techniques can offer a promising starting point for refinement by more accurate, time-consuming computational methods, although other well-trained, highly specific FFs can be nearly as accurate as QM techniques. Below, we outline several available FF methods for TS modeling and focus more on their scientific foundation and less on their successes. Very often, the users themselves can tip the scales from failure to success, but this is only possible by understanding the underlying principles of the software being implemented. There is a fine line, however, between bias and expertise.

1.3 Ground State Force Field Techniques

1.3.1 Introduction

In 2003, Jensen and Norrby noted that many MM applications are treated as black-boxes that are available in common modeling software packages; however, TS modeling was not yet one of those techniques.[35] Over ten years later, TS modelling remains a field requiring some level of expertise and here we present an overview of the available TS modeling approaches. They are classified into two general groups: ground state force field (GSFF) techniques and transition state force field (TSFF) techniques. For each method, we provide a general explanation followed by the motivation behind the technique and the principles upon which the method was founded. Some reported performance data and the accessibility for industry and/or academia is also given when available.

GSFF techniques often involve a modification of the core MM infrastructure developed for standard FF applications (i.e., GSFF parameters with additional functions developed for TS modeling). Knowledge of ground state reactants and products and their individual potential energy surfaces is used to locate the TS.[30]

1.3.2 ReaxFF

1.3.2.1 Theory

The first GSFF developed for the study of chemical transformations that we will discuss is ReaxFF, developed by van Duin, Dasgupta, Lorant and Goddard.[37] This is an example of what is termed a "reactive force field" because it accounts for bond making/breaking. This FF was originally developed to model TSs for hydrocarbon systems[38] but has, over the past decade, been extended to an array of organometallic species,[39-49] other organics,[50-52] and a variety of different reactions/applications.[53-56] It has also been implemented into a complete package including a user interface. Here we present some "under-the-hood" information that is important for understanding the software/method.

The major advantage of ReaxFF is the treatment of all bonding terms (included in E_{Bond}, E_{Angle}, E_{Torsion}, etc.) including TS forming/breaking bonds in the energy function (Eq. 1.10). To account for the longer-range covalent interactions that are signatures of TS structures, the bonding terms here are bond-order dependent (Figure 1.4). For example, a breaking/forming bond will have a bond order less than 1, while a double bond being converted to a single bond would typically have a bond order between 1 and 2. Thus, since the designation of a bond is no longer binary as in traditional non-reactive FFs (where the bond either exists or it does not) there is instead a bonding spectrum. Accordingly, no explicit bond labeling is required in the input (topological list of bonds set at the beginning of the calculations, which becomes problematic when bonds must break or form during a simulation) as it is for most MM applications. The bond-order depends on the inter-atomic distances and can thus be easily calculated "on-the-fly." It is then corrected based on the valence to ensure that the bond order does not exceed the valencies of the atoms as defined by the E_{bond} term of Eqs. 1.11-1.13. Similar relationships are made for E_{angle} and E_{torsion}.[37]

$$E_{\text{Total}} = E_{\text{bond}} + E_{\text{Over}} + E_{\text{Under}} + E_{\text{Angle}} + E_{\text{Penalty}} + E_{\text{Torsion}} + E_{\text{Conjugation}} + E_{\text{van der Waals}} + E_{\text{Coulomb}} \quad (\mathbf{1.10})$$

$$BO'_{ij} = BO'^{\sigma}_{ij} + BO'^{\pi}_{ij} + BO'^{\pi\pi}_{ij}$$

$$= exp\left(a_1\left(\frac{r_{ij}}{r_0^{\sigma}}\right)^{a_2}\right) + exp\left(a_3\left(\frac{r_{ij}}{r_0^{\pi}}\right)^{a_4}\right) + exp\left(a_5\left(\frac{r_{ij}}{r_0^{\pi\pi}}\right)^{a_6}\right) \quad \textbf{(1.11)}$$

$$BO_{ij} = f\left(BO'_{ij}, Val_i, Val_j\right) \quad \textbf{(1.12)}$$

$$E_{\text{bond}} = -D_{\text{e}} \times BO_{ij} \times exp\left(a_6 \times \left(1 - BO_{ij}^{a_6}\right)\right) \quad \textbf{(1.13)}$$

Here, BO'_{ij} is the bond-order between atoms i and j, σ, π, and $\pi\pi$ denote the bond-character (single, double, triple), r_{ij} is the inter-atomic distance, r_0 is the equilibrium distance, Val_i is the valence of atom i, BO_{ij} is the corrected bond-order (corrected for over-estimation or underestimation of the valence), D_{e} is the dissociation energy and the a-values, trained for each bond pair, are empirical parameters (a description of the required parameterization can be found at the end of this section).

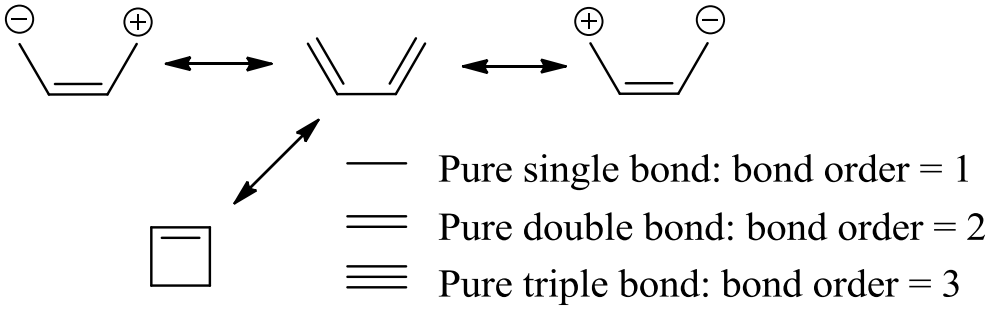


Figure 1.4. An example of bond order using 1,3-butadiene and cyclobutene. None of the bonds in this system are pure single or pure double in character and thus the bond-order will be calculated to be somewhere between 1 and 2.

Eq. 1.11 is continuous with no issues in transitions between bond characters, but requires correction factors for over-coordination. The empirical parameters denoted "*a*" have a maximum value of 1 and then drop to 0 depending on the inter-atomic distance. For example, for C-C bonds, there is a maximum BO of 3 and C-H bonds a maximum of 1. In the same vein, a carbon atom should not exceed a total BO of 4 due to its valence, based on valence bond theory,[57] yet this would occur without an atomic over/under-coordination penalty term (E_{Over}, E_{Under}) within this methodology. The angle and torsion terms (E_{Angle}, E_{Torsion}) are fairly standard functional forms with the exception of their being based on BO. Once again, a penalty function (E_{Penalty}) is invoked on the angle term due to the boundary conditions necessary for bond orders as mentioned above. Conjugation effects ($E_{\text{Conjugation}}$) are considered when successive bond orders are at

approximately 1.5. The non-bond and electrostatic factors ($E_{\text{van der Waals}}$, E_{Coulomb}) are represented by a modified Morse potential and expanded Eq. 1.6 respectively.

1.3.2.2 Validation

To train and validate their FF, van Duin *et al*. used heats of formation as criteria of performance.[37] They tested ReaxFF on a variety of hydrocarbons including conjugated and non-conjugated systems, radicals, various conformations and crystal structures, and compared all values to quantum chemical data as well as to MM3 (a high quality FF for small molecules) results. The rationale for deviations between this FF and MM3 values for heats of formation (the error for ReaxFF is about double that of MM3) is that MM3 employs empirical corrections specifically developed for these classes of molecules (different ring systems for example). Experimental values for heats of formation are given, however no discussion is offered to explain the observed differences for conjugated/non-conjugated or radical systems. Importantly, the developers recognized that some of their validation data was biased because they used the same (or similar) systems for training their set of parameters. They required 93 parameters to describe hydrocarbons and took satisfaction in the generality of their energy descriptions, noting specifically that no special treatment for sp^3-sp^2 C-C bonds is needed as compared to MM3. They demonstrated this FF can be used as a non-reactive FF (for ground states, not for TS modeling). Their examination of dissociation curves compared to DFT data then demonstrated ReaxFF's potential as a reactive FF. The developers are aware of the limitations involving quantum chemical phenomena, as, for example, the orbital overlap/symmetries involved in the ring-opening of cyclobutene to form butadiene (Figure 1.4). Because this FF is based on empirical data, it is incapable of modeling complex reactivity for which data does not exist for parameter development, nor can it discover completely new reactions. This limitation, however, is offset by the immense speed-up for structure/reactivity prediction compared to semi-empirical and DFT QM techniques (2 orders of magnitude faster than PM3 (semi-empirical) and 5 orders of magnitude faster than DFT for a system of about 80 atoms).[37]

1.3.2.3 Availability

ReaxFF is offered with a free 30-day trial and many different licenses can be purchased including regional discounts, varying core allocations and teaching-only licenses. The pricing is dependent upon the length of the contract, region and purpose. Their website (www.scm.com) offers tutorials, pricing information, references and manuals. As mentioned previously, it is worth noting that the parameterization process is not necessary if the system under investigation incorporates only atoms/reactions that have already been parameterized; however, as described in many of the references, deriving the empirical values in each term of Eq. 1.10 may be necessary. It is also of value to know that this GSFF is one of the most widely-referenced methods, the original publication having been cited over 800 times according to scopus.com.

1.3.3 RFF

1.3.3.1 Theory

In 1997, Rappé and co-workers proposed the reaction force field (RFF) in order to model the bond breaking/forming process.[58] For their purpose, the reaction $A-B+C \rightarrow A+B-C$ was partitioned into two components $A-B \rightarrow A+B$ and $B-C \rightarrow B+C$ and the crucial aspect is that the TS can be found somewhere on the two PESs. To model these surfaces, the authors developed a dissociative bond term that accounts for homolytic cleavage, polar bond cleavage, changes in hybridization and 1,3-interactions between atoms on fragments (Figure 1.5).

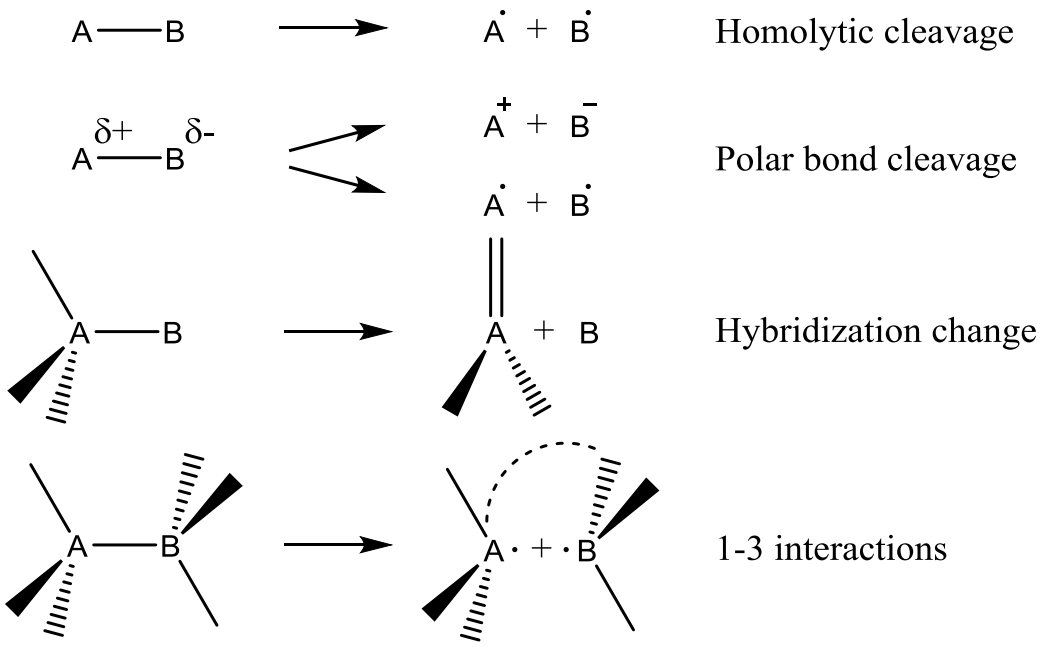


Figure 1.5. Dissociative bond considerations included in RFF.

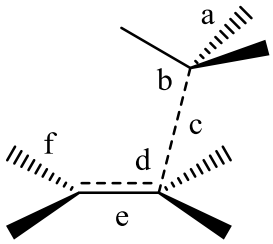
Apart from the dissociative bond term, the remaining terms of the FF are standard as in Eq. 1.1 along with an inversion term (a Fourier expansion for trivalent atoms – keeping the cone-like shape of trivalent nitrogen and phosphorous but allowing the centre to invert while keeping sp^2 atoms flat). The terms, which can be found in the original RFF publication, resemble those of the Universal force field (UFF), developed by the same group,[59] and rely on bond orders as in ReaxFF (Eqs. 1.11-1.13) albeit not exactly in the same way. Key differences are that the electronegativity is no longer considered in the bond stretch term since the polarity is handled in the dissociation of bonds, and the van der Waals term has been modified to better model the 1-3 interactions depicted in Figure 1.5.

1.3.3.2 Validation

The first generation RFF was used to calculate vibrational frequencies as well as TS geometries for the Diels-Alder cycloaddition, the Cope rearrangement and the chemical reaction of a methyl radical with ethylene. The activation energies were well estimated for these three reactions as were the geometries when compared to experiment and QM

methods (Unrestricted Hartree Fock – UHF for example). An example is summarized in Table 1.1.

Table 1.1. RFF and QM comparison for the reaction of a methyl radical with ethylene.



Term	Property	RFF Prediction	UHF Value
a	Length	1.08 Å	1.08 Å
b	Angle	106.3 °	101.0 °
c	Length	2.19 Å	2.25 Å
d	Angle	123.3 °	109.1 °
e	Length	1.36 Å	1.38 Å
f	Length	1.08 Å	1.08 Å
E_{TS}	Energy	8.1 kcal/mol	Exp. Value: 7.9 kcal/mol

Only these hydrocarbon systems were investigated and any user should be wary about the use of this FF with heteroatoms unless suitable changes are made. RFF was applied to zirconium systems by Dunn *et al.*[60] but only as a means to obtain a starting geometry (conformational search) while DFT was used as a refinement technique. The RFF energies were not deemed accurate when compared to experiment; this is not surprising since the goal of MM techniques is often to obtain geometries and not absolute energies, especially in the case of metals that exhibit many electronic effects. RFF is still considered a valuable FF technique; wide coverage of the periodic table has been reported,[61-62] much like UFF, although no validation is presented with systems featuring heteroatoms.

1.3.3.3 Availability

The authors must be contacted in order to obtain RFF. No further information is found on the author's website regarding the equations, parameters or availability.

1.3.4 SEAM

1.3.4.1 Theory

SEAM is another GSFF method. It was developed by Jensen in 1992[63] in response to criticism by Houk and co-workers[64-67] about gaps that needed to be filled[68-69] with respect to the TS location problem (discussed in the TSFF section below). While the work prior to that of Jensen was criticized for its over-parameterization and small validation set, Jensen attempted to generalize the approach to modeling TSs and eliminate the need for specific parameters for each and every TS. Jensen postulated that because the TS is the apex of the lowest-energy path from the reactants to the products along the reaction coordinate, the TS could then be described as the seam of the intersecting PESs describing the reactants and the products (Figure 1.6).

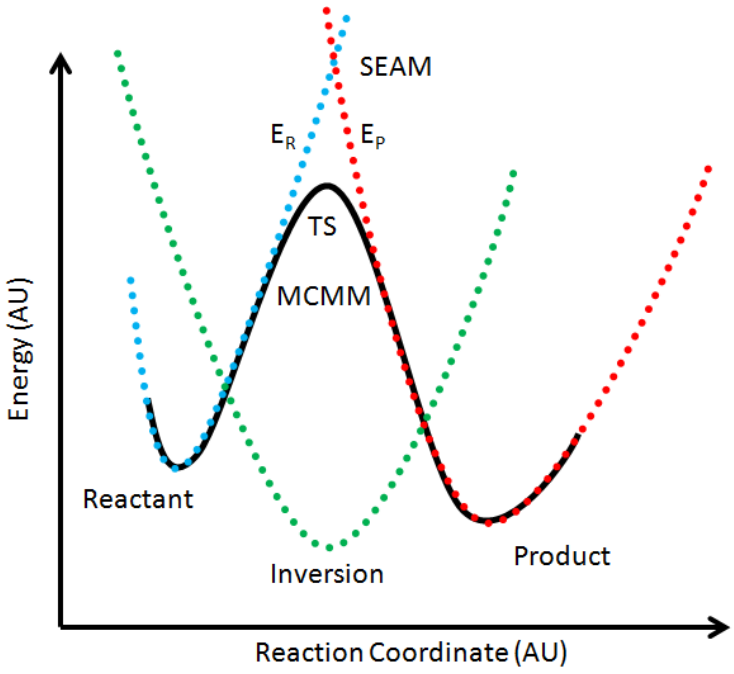


Figure 1.6. Illustration of intersecting PESs for reactant and product yielding the SEAM, the inversion technique proposed by Houk, and MCMM methods advocated by Truhlar.

The GSFF used to develop SEAM was MM2, the precursor to MM3, but the SEAM approach can be used with almost any available ground state force field, with accuracies dependent on the FF selected. One parameter, c in Eq. 1.14, was added in the description of the bond stretching (Eq. 1.2) to allow for greater deviation from the equilibrium bond distance - something that one might expect to see in a TS. Additionally, a constant value is required for each reaction so that each PES is set to the same scale; because the connectivity is different for reactant and product, the FFs cannot be compared directly and an offset is required.[70] No other changes were made to the FF except for adding missing parameters for certain atoms/fragments, regardless of TS modeling.

$$E_{\text{bond}} = k_s \times (r - r_0)^2 \times \left(1 - c \times (r - r_0)\right) \qquad \textbf{(1.14)}$$

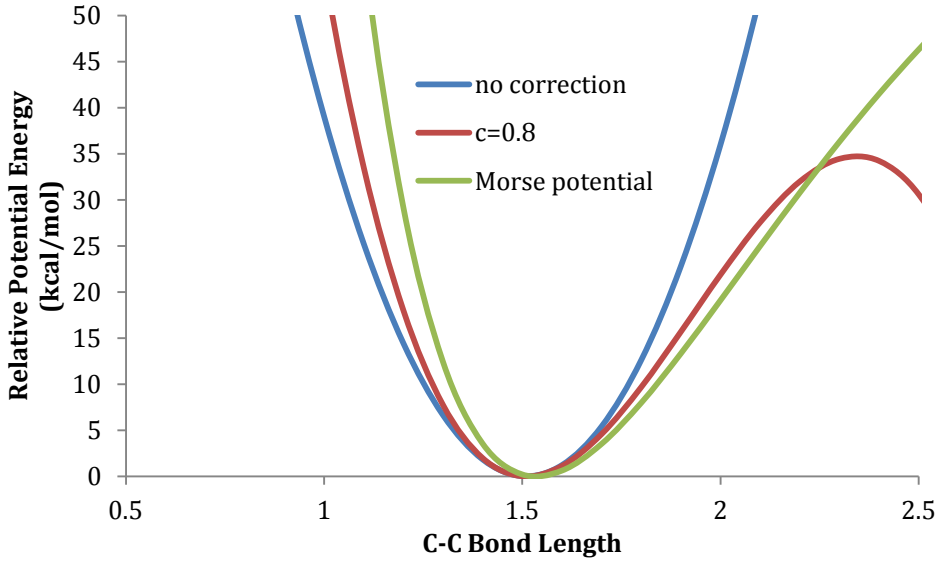


Figure 1.7. Corrected bond energy term according to Eq. 1.14. The inversion of energy at highly stretched bond length is an artefact of the method and must be considered upon selecting the value of c.

The notion of locating the intersection of reactant and product PESs means determining where along the coordinate the energy of the initial state, E_R, is equal to that of the final state, E_P (Figure 1.6). Because the sum of the two energies should be at a minimum, this converts the search for the location of the TS into an energy minimum search that can be carried out by most optimizers embedded in MM programs. This is solved in this instance using Lagrange multipliers and is improved iteratively until convergence. The technical details are described in the original publications[35,63,70] and the

idea was later expanded into mixing three PESs, the third of which models resonance energy terms.[71] It is important to note that the starting geometries in the validation test were those of the reactant and/or product and it is believed the global minimum in the optimization is found in each instance for the constrained systems that were tested.

1.3.4.2 Validation

Validation was done by matching the TSs generated using the SEAM approach (MM2 FF) with those generated using *ab initio* techniques. Jensen probed small structural changes, distant from the reaction centre, in order to verify that his method could reproduce trends in reactivity. The small changes allowed him to neglect solvent effects and entropic considerations. The reactions considered (compared to *ab initio* structures) in the first publication[63] are summarized in Figure 1.8.

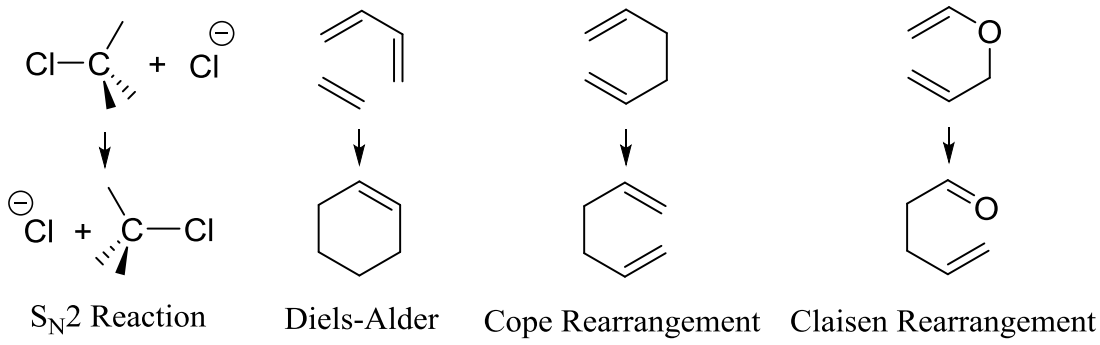
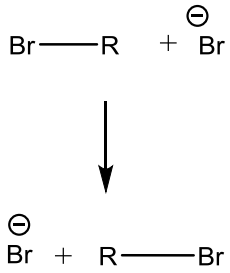


Figure 1.8. Initial reactions tested for SEAM that were compared to *ab initio* structures

SEAM was able to find secondary (i.e., higher in energy) TS geometries as well as the energetically favored *ab initio* ones based on MCSCF – Multi-configurational self-consistent field – a QM method. For example, both the chair-like and boat-like TSs were found for the Cope rearrangement with an energy difference similar to that of MCSCF. The Claisen rearrangement proved more difficult because the reaction is not thermoneutral, i.e., it does not depend solely on steric effects, and heats of formation must be considered. The resulting breaking/forming bonds were too short compared to QM data[72] (RHF/6-31G*): 1.748 and 1.974 Å compared to 1.917 and 2.264 Å for breaking and forming bonds respectively. When heats of formation were considered, the derived TS had a longer breaking bond (1.765 Å) and a shorter forming bond (1.943 Å). This

agrees with the Hammond postulate that states an earlier TS is observed in an exothermic reaction.[73] Similar issues pertaining to non-thermoneutral reactions were encountered with the Diels-Alder reaction for which special attention was required. Additionally, the c term in Eq. 1.14 may require fine-tuning depending on the reaction being considered, especially whether it depends primarily upon steric factors or other factors influencing the TS. The trends in steric effects were generally well predicted especially when compared to experimental data when steric repulsions are the dominant driving force and when the substrates did not differ significantly (Table 1.2).

Table 1.2. Activation energy trends for steric effects.



R-Group	C-Br (Å) (SEAM)	C-Cl[(a)] (Å) (HF/MINI)[(b)]	ΔE_A (kcal/mol) (SEAM)	ΔE_A (kcal/mol) (experimental, in solution)
methyl	2.422	2.418	0.00	0.00
ethyl	2.451	2.442	3.93	1.56
propyl	2.453		3.52	1.41
i-propyl	2.494	2.410	8.24	3.77
i-butyl	2.461		5.84	3.07
n-pentyl	2.474		10.91	5.55
t-butyl	2.657		20.63	

(a) Data obtained from secondary source with chlorine atom demonstrated that the lack of variation in the TS bond distance was not abnormal.[74]

(b) Hartree-Fock (MINI is a basis set, not covered in this chapter).

Jensen noted that minimal limitations of the SEAM method exist if certain conditions are applied. For example, if the FF used for both reactants and products is known to be accurate for distorted geometry (i.e., well-described outside favorable geometries as shown in Figure 1.7), it should produce reasonable TS geometries. If this condition does

not hold, the same TS could be found for multiple starting geometries or optimizations may not converge. One drawback, as seen in Figure 1.6, is that the absolute energy of activation is over-estimated. However this effect is nullified when investigating relative energies, as, for example, when comparing diastereomeric TSs in asymmetric reactions. Jensen also noted that these GSFF methods are only effective when steric concerns are the driving forces of the reaction since the FFs used do not account for electronic effects either directly at the reaction centre or indirectly from other molecular fragments in the reactant or product. He proposed that an efficient method to find the TS structure is to use SEAM for geometry optimization and then to perform a single point DFT calculation for electronic structure data. That approach was successfully carried out by Anglada *et al*. on, for example, the ring opening of the cyclopropyl radical.[71] SEAM was later applied to enzymatic TSs, more specifically the decarboxylation of orotidine by the decarboxylase enzyme orotidine-5'-monophosphate decarboxylase.[75] This protocol was attractive because it provided significant advantages in time over QM and QM/MM techniques (a hybrid of QM and MM not covered in this chapter) although variations in the energy between the different TS structures found with the same enzyme were significant (attributed to large structural differences between the 20 TS structures examined).

1.3.4.3 Availability

SEAM for AMBER[76] is available free of charge upon request. The website (http://www.teokem.lu.se/~ulf/Methods/seam.html) contains useful references and a guide for preparing input. An external MM program is required. Examples are offered when combining SEAM with AMBER (ambermd.org) and a brief outline of how to run the program and find the output is provided.

1.3.5 EVB/MCMM

1.3.5.1 Theory

The empirical valence bond (EVB) technique was proposed in 1980 by the Nobel laureate, Warshel, and his co-worker, Weiss[77] and then further developed, modified and

applied to TS modelling by Truhlar and co-workers, denoted multiconfiguration molecular mechanics (MCMM) in the year 2000.[78] In this section, we focus on the applications of MCMM, keeping in mind that Warshel pioneered the VB theory/application to modelling. Some differences between EVB and MCMM are outlined in a letter published by Truhlar.[79] Those distinctions will not be described here but they are important and the novice is urged to read this.

EVB/MCMM focuses on mixing potential energy surfaces of both reactants and products, whether they are monatomic or polyatomic molecules. MCMM is an application of the valence bond theory (VBT) that originated from London, Eyring and Polanyi.[80] In contrast to the SEAM method that searches for the crossing of the two PESs (Figure 1.6), the MCMM method uses an energy term to describe the mixing of the two PESs in order to calculate the TS. More specifically, a mixing term (E_{RP}) is used to convert the two diabatic states (E_R and E_P) into the proper adiabatic states (the actual TS) and not the crossing point as with SEAM. The adiabatic states (E) are determined by solving the following matrix which boils down to Eq. 1.15 in the case that two states, E_P and E_R, are considered.

$$\begin{vmatrix} E_R - E & E_{RP} \\ E_{RP} & E_P - E \end{vmatrix} = 0$$

$$E = \frac{\left((E_R + E_P) - \sqrt{(E_R - E_P)^2 + 4 * {E_{RP}}^2} \right)}{2} \qquad \textbf{(1.15)}$$

For the minima (in the reactant or product states, before or after the chemical reaction), the difference between E_R and E_P is so large that E_{RP} becomes negligible and the equation reduces to the minimum energy, either E_R, representing the reactants PES or E_P, representing the products PES. If we consider the S_N2 reaction that Jensen investigated with the SEAM method, our hypothetical potential energy surface could be illustrated by Figure 1.9. Another example (nucleophilic addition) can be found in an excellent review from Jensen and Norrby.[35]

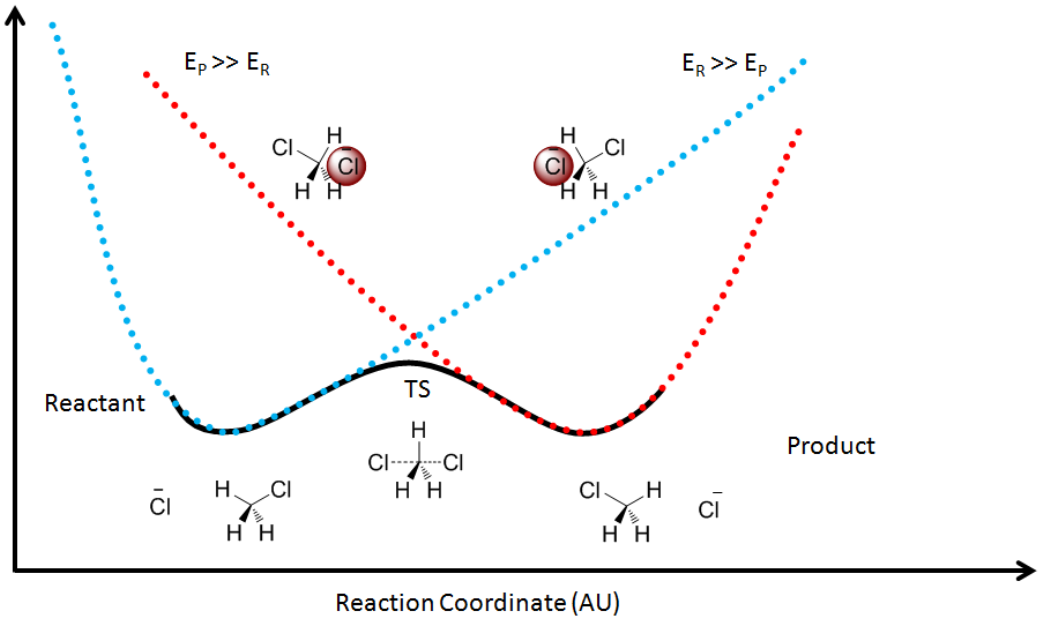


Figure 1.9. Illustration of mixing PESs of an S_N2 reaction.

Again, and it cannot be overstated, the most important criterion for this method to be applicable, similar to SEAM, is that the FF(s) used to describe E_R and E_P should model the structures correctly at large distances from the minimum along the reaction coordinate. This usually requires a Morse-like potential for the bonding term[24] at long distances and a modified Lennard-Jones potential (efficient for calculating van der Waals interactions) for short distances[23] since the repulsion is inaccurate at short distance in its original form. Additionally, the angles can be highly distorted and should be modelled appropriately. Basically all FF terms that are required to describe energetics far from equilibrium structural values must be valid at these geometries. Furthermore, the FF should also be able to calculate the relative energy of reactants and products accurately, something not expected from most FFs. While some corrections can be made (normally heats of formation), others require external calculations or experimental data. The problem arises from the changed connectivity between reactants and products along the reaction pathway; most FFs are/were designed to calculate relative energies for different conformations, but not for changes in configurations.

The mixing term, E_{RP}, can be either a constant reaction-specific value or a function that depends on the reaction coordinate. Initial work by Chang and Miller[81] was used to

solve E_{RP} in order to match QM data for geometries, energies and frequencies for the TS. Subsequently, Truhlar and co-workers proposed a modified version of the Chang and Miller technique by fitting the mixing term at multiple points on the PES and then interpolating additional points. Their method depends on internal coordinates and is restricted to only the reactive center. Determining E_{RP} requires additional input data, either barrier height and approximate TS geometry or electronic structure data from high-level computations prior to the use of the actual MCMM method, regardless of the method used. This technique can be perceived by bench chemists as complicated and less user friendly than the aforementioned methods, but this is a very powerful tool for modeling bond making/breaking processes. More details can be found in the original publication.[78]

1.3.5.2 Validation

The validation was performed using the MM3 FF (with some modifications) and AM1/PM3 for semi-empirical dynamics simulations[82-83] on three test reactions (Figure 1.10): the isomerisation of 1,3-*cis*-pentadiene, the hydrogen transfer between a hydroxyl group and methane and the hydrogen transfer between CH_2F and CH_3Cl, chosen for their sensitivity to the shape of the PES.

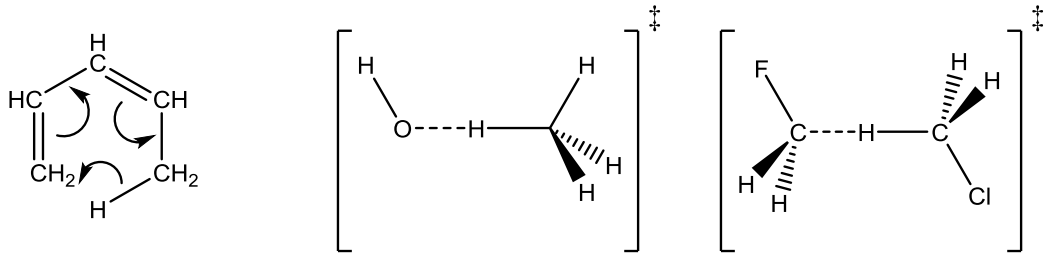


Figure 1.10. Initial reactions tested for MCMM.[78]

The validation was done by comparing computed results to direct dynamics calculations using gaussrate.[84] By tuning of some parameters and carefully selecting the number of non-stationary points (the additional interpolated points discussed above) to be modeled on (or off) the reaction coordinate, the semi-empirical dynamics results were

reproduced successfully using MCMM. The authors stressed that the major advantages of this technique are the need for only a small amount of QM electronic structure information (reactants and products) as input information, the internal, automatic generation of surface information, and, that the reaction coordinate can be more than one-dimensional. A major attribute is the ability to calculate rate constants and vibrational frequencies along the reaction path. Truhlar and co-workers noted that their weighting function is very sensitive. Although they assessed a variety of different forms, the chosen form of the weighting function may not be applicable in all cases. This method and its closely related VBT method have been reviewed extensively[35,85-91] and they have been used in multiple studies in order to expand the capabilities and functionalities.[92-106]

In addition to small molecule reactions, there is a great need for accurate methods that can model biocatalysed transformations. One of the most widely studied classes of proteins is the cytochromes P450 (CYPs). CYPs are metabolic enzymes that oxidize a large fraction of the drugs currently on the market. Predicting the oxidation of drugs is of great value for drug design studies. As a result, this approach has been applied to CYPs and mode specifically to the modeling of testosterone oxidation.[102]

1.3.5.3 Availability

A software package that carries out MCMM calculations is named MCSI; it is distributed by Professor Truhlar at the University of Minnesota. Prior to 2010, it was known as MC-TINKER (it uses TINKER, developed by Professor Jay Ponder[107]). The website (comp.chem.umn.edu/mcsi) offers information on how to obtain a license and download the software packages. There is no cost associated with obtaining the program package; only a license form is required. The website also includes an extensive, well-written manual and revision history. For EVB-specific software, Professor Warshel's website (laetro.usc.edu/software.html) refers to a few programs, including their in-house software MOLARIS-XG, which can be downloaded with permission from their executive (free of charge). Detailed manuals with examples are also provided. There also exists a web-based platform for EVB in AMBER (http://ambermd.org/evb_pmf.html) among others.

1.3.6 ACE

1.3.6.1 Theory

ACE (Asymmetric Catalyst Evaluation) was developed by Moitessier and co-workers and first reported in 2008.[108] ACE was originally designed from chemical principles such as the Hammond–Leffler postulate stating that the TS is most similar to the species – reactants or products – to which it is closest in energy. This led the developers to consider the TS as being a linear combination of reactants and products. In practice, it is similar to the SEAM approach but with a tunable factor mimicking the Hammond–Leffler postulate. Conceptually, the forming bonds are considered as a combination of covalent bonds (products) and non-bond interactions (reactants) with λ defining the product character of the TS (i.e., the position of the TS on the PES) (Eq. 1.16). For example, the TS of a Diels-Alder reaction (Figure 1.11) is the combination of two partial bonds (λ ranges from 0 to 1) from the product as defined in a FF and of non-bonds from the reactants (e.g., van der Waals and electrostatic interaction). Specifically, if in Figure 1.11a we set λ to 0.25, the bond between atoms 1 and 2 has 75% double bond character (in the reactant) and 25% single bond character (of the product). Similarly, atoms 7 and 5 would be 75% in non-bonded interaction and 25% in an angle. This approach is fully automated and combined with MM routines and a genetic algorithm into a single independent program (ACE). The MM3* FF was used in the original version. The genetic algorithm takes care of the conformational search and is necessary to optimize the complete TS complexes including the atoms involved in the breaking/forming bonds (e.g., 4, 7, 1 and 8 in Figure 1.11) as well as all the other atoms.

$$E_{\mathrm{TS}} = (1-\lambda) \times E_{\mathrm{reactant}} + \lambda \times E_{\mathrm{product}} \qquad \textbf{(1.16)}$$

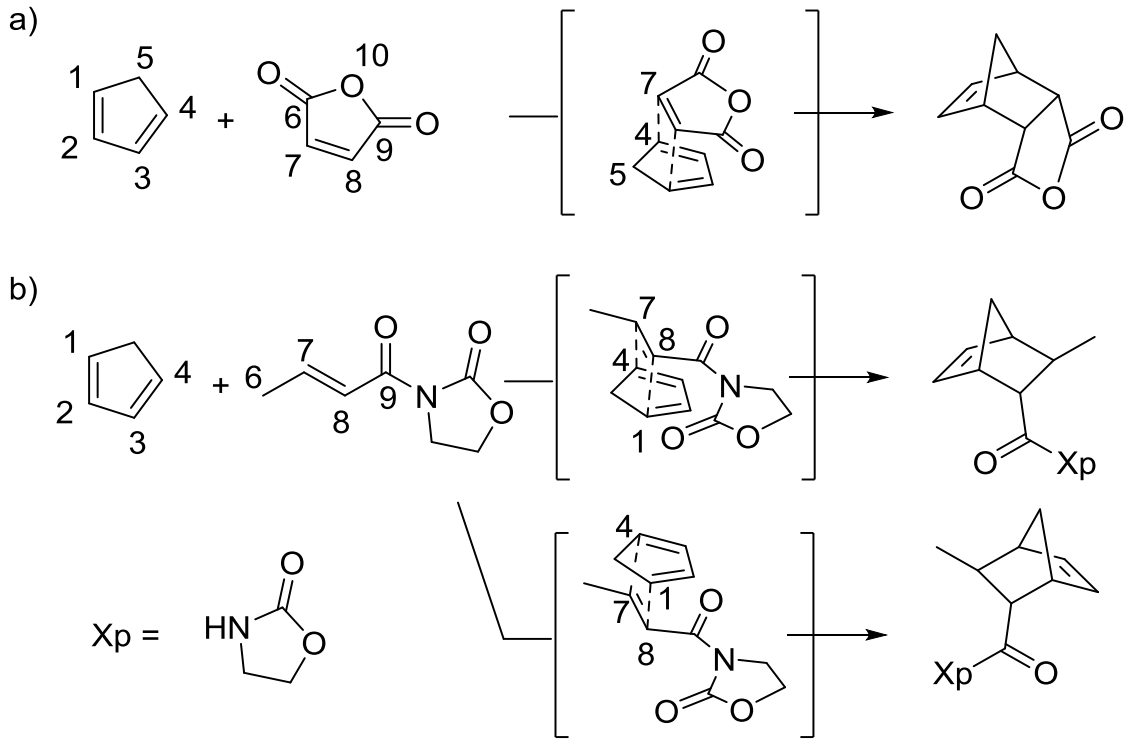


Figure 1.11. Diels-Alder reactions and transition states.

In a second version of ACE, solvent effects in the form of implicit solvent models were implemented (the GB/SA method was used, see original paper[109]). As Diels-Alder reactions with dissymmetrical reagents have been found to be highly asynchronous (the bond between carbons 4 and 7 is much shorter than that between atoms 1 and 8 in the TS in Figure 1.11b), modeling of asynchronous TSs was made possible in this second version of the program. ACE 2.0 also approximates a Boltzmann population, designed to improve the prediction of temperature-dependent phenomena.[110] In a more recent version (ACE 2.2, unpublished), the focus was on automating the entire computation protocol. In ACE 2.0, each TS (two diastereomeric TSs in Figure 1.11b) was computed separately and the predicted diastereomeric excesses (d.e.) were computed manually by the user. ACE 2.2 can now take all the potential TSs and compute the d.e. in a single, automated run. It can also take a library of catalysts and compute a set of d.e. in a single run. ACE 2.2 was then incorporated into a fully automated interface for asymmetric catalyst design that will be discussed in the coming chapter.

The Moitessier group developed a docking program (FITTED[111]) that was combined with the ACE TS modeling method leading to the IMPACTS program.[112] ACE was developed to compare energies and structures of diastereomeric TSs; because the TSs' connectivity is identical (TSs in Figure 1.11b), $\Delta\Delta G^{\ddagger}$ (the related d.e.) could be computed accurately with force fields. Combining some of the FITTED and ACE features, IMPACTS was developed to identify the most likely site of metabolism (SoM) of drugs and includes a variety of mechanisms like hydrogen abstraction, aromatic oxidation and sulfide oxidation. To compare the *different* mechanisms accurately (i.e., TS structures having different connectivity and sets of parameters), a reaction-dependent correction was necessary. Similarly, different carbon atoms reacting via the *same* mechanism (e.g., aromatic oxidation of multiple aromatic carbons) cannot be compared with fidelity using traditional MM approaches. To complete the computation of the relative TS potential energies, an empirical ligand reactivity factor was introduced. To do that, the energy of activation for a large set of fragments was computed using DFT and tabulated. Within the IMPACTS package, each possible SoM is assigned a fragment and then an energy of activation, should that SoM be the one selected. This program enables the study of TSs within a protein binding site and was developed specifically for CYP-mediated drug metabolism.

1.3.6.2 Validation

ACE 2.0 was validated with approximately 150 examples of asymmetric Diels-Alder cycloadditions, organocatalyzed aldol reactions and epoxidation reactions. The generated TS structures were compared to those reported previously using DFT methods (Figure 1.12).[110]

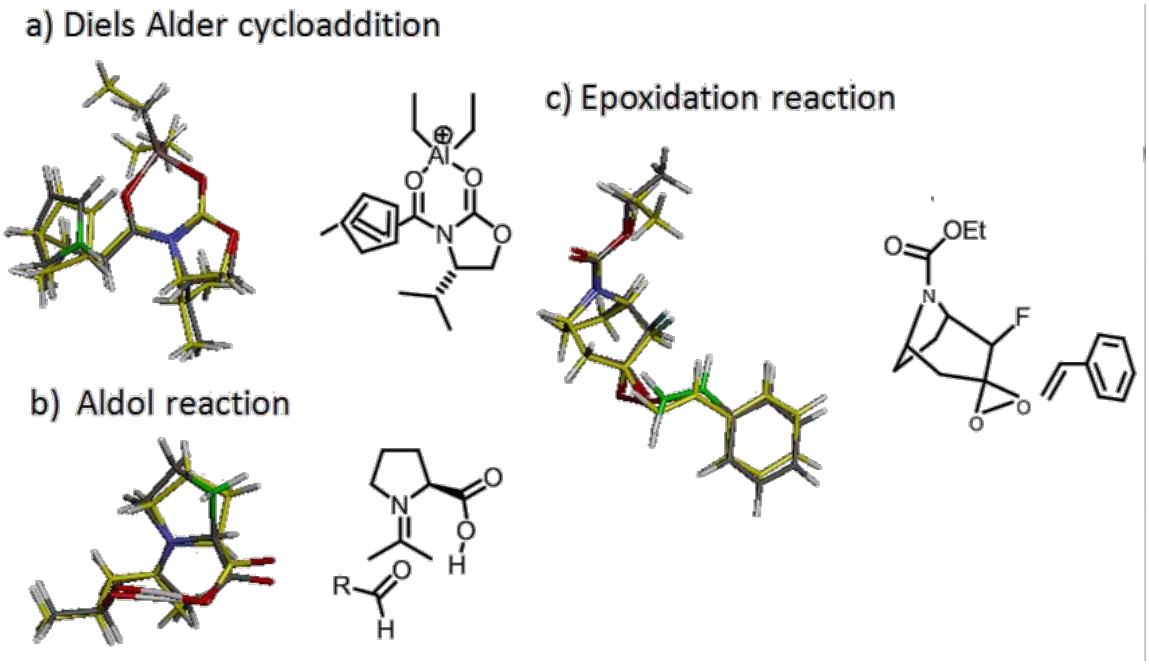


Figure 1.12. ACE 2.0 (yellow) vs. DFT (grey) generated TSs. The largest deviations between ACE and DFT predictions are highlighted in green. a) Diels Alder reaction between butenoyl - chiral auxiliary (in the back) and cyclopentadiene (in the front); b) Epoxidation reaction using a chiral dioxirane reagent; c) Proline-catalyzed aldol reaction.

The predicted selectivities were compared to experiment and also to DFT-generated predictions. A good correlation was determined and a mean unsigned error in the range of 1 kcal/mol was obtained. An example is given in Figure 1.13 where DFT was compared to ACE 2.0 predictions on a set of epoxidation catalysts for a set of alkenes.

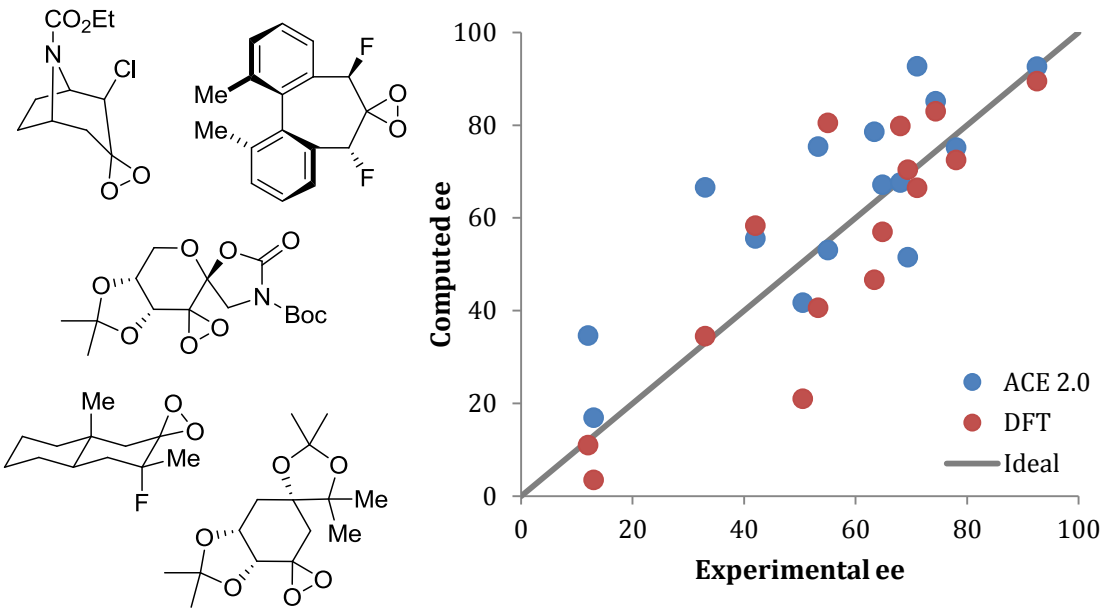


Figure 1.13. Representative catalysts screened (left) on multiple alkenes and the average predicted and observed ee's for each catalyst.

IMPACTS has also been validated extensively using sets of P450 substrates. The predicted metabolites were compared to those predicted using other state-of-the-art programs (i.e., MetaSite) and to predictions made by experts in the field.[112] As can be seen in Table 1.3, this tool proved to be accurate in modeling TSs in cytochromes P450 (CYPs). This work also demonstrated the applicability of the ACE approach to biocatalysis.

Table 1.3. Accuracy[a] of IMPACTS in predicting the correct SoMs for respective datasets.

CYP isoform	N[b]	Random selection	IMPACTS[c]	Experts[d]
1A2	137	31	77	69 (5)
2C9	129	29	79-82	71 (7)
2D6	157	27	76	64 (4)
3A4	293	28	72-75	61 (6)
All 4	716	28	77	65 (5)

[a] % of molecules with an observed SoM in the predicted two SoMs. [b] Number of substrates. [c] Multiple crystal structures were assessed. [d] Average predictions by medicinal chemists and biotransformation experts with standard deviation given in brackets.

1.3.6.3 Availability

The ACE and IMPACTS programs are available with all the necessary accessories as part of the VIRTUAL CHEMIST and FORECASTER platforms respectively upon request at http://www.molecularforecaster.com/.

1.4 Transition State Force Field Techniques

1.4.1 Introduction

Decades ago, FFs were used to evaluate reactivity of simple substrates. Although MM had been used to rationalize a variety of reactions, scientists were mostly looking at strain energies using GSFFs. Of note are the reports from Garbisch in 1965 who studied the equilibrium between enol forms using FF-like potential energy computations on experimentally determined (NMR and EPR) structures[113] as well as the reactivity of alkenes in diimine reduction[114] (Figure 1.14), and, the reports from DeTar and Tenpas who used parameters developed for alkanes and *ortho*-esters to model the TS of the acid-catalyzed hydrolysis of esters in 1976.[115]

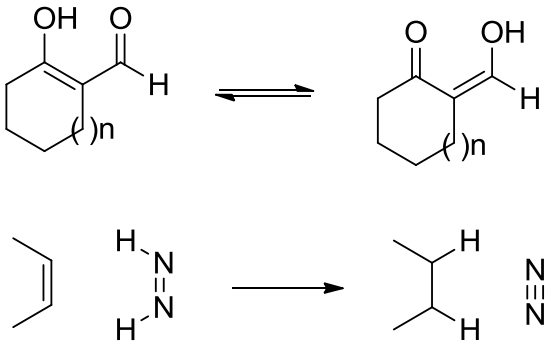


Figure 1.14. Systems investigated by Garbisch in 1965.

An alternative to GSFFs are reaction-specific transition state force fields (TSFFs). One of the main advantages of this approach is its transferability to most MM packages. The user assigns new atom types and develops the corresponding parameters for the TS, and then adds them to the force field, which is often in the form of a text file. The major disadvantage is the need to develop FF parameters for any new reaction, a step that requires expertise. As mentioned earlier in this chapter, we consider only TSs that are

saddle points (maximum on the reaction coordinate, but minimum in all other directions, Figure 1.1). However, to be able to use the traditional MM optimization (minimization) routines, the TS must be modeled as a minimum on that PES and not a true TS ("inversion" in Figure 1.6). In principle, a proper set of parameters can be developed that considers the TS as being a minimum, an artefact that has its own limitations. Nevertheless, approaches for doing such calculations have been reported and validated for a number of reactions.

In this context, a QM-derived TSFF was first introduced over 30 years ago and applied to asymmetric hydroboration of alkenes.[116-117] In most of those early applications, the TS was developed using QM methods and simply kept frozen in MM applications. This approximation came from the belief that the stereochemical outcomes of reactions could be attributed primarily to steric interactions between the atoms not directly involved in the bond breaking/bond forming reactions. Thus no optimization of the position/interaction of the atoms directly involved in the bond breaking/forming was deemed essential and no specific parameters were required for these reacting atoms and interactions between them. As an example, Houk investigated the asymmetric hydroboration of alkenes (Figure 1.15).[116] TSs were first computed using *ab initio* calculations (3-21G basis set) for simple systems like that shown in Figure 1.15b. Then the four atoms forming/breaking bonds in the TS (C, C, B, H) were frozen in space while the atoms directly connected to those reacting atoms were free to move with non-reactive MM2 parameters. This purely MM approach was accurate in identifying major contributors (sterics and electronics) to the stereoselectivity of a number of hydroboration reactions. Although a number of other reactions have been investigated using this approach, they were reviewed by Eksterowicz and Houk[117] and will not be covered in this chapter.

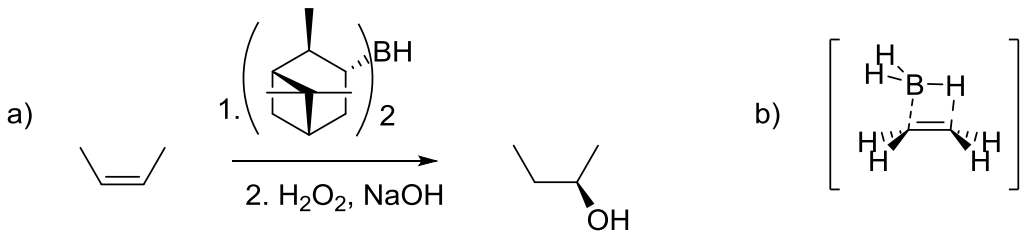


Figure 1.15. a) Asymmetric hydroboration of *cis*-butene and b) TS for hydroboration of ethylene with borane.

While using frozen QM-generated TSs was successful, allowing for transition state flexibility is expected to be more accurate. Houk and co-workers developed parameters defining the TSs as energy minima rather than as saddle points (Figure 1.6).[118] This can be viewed as being too gross an approximation, and one might anticipate that distorted TSs would be poorly defined. However it has been found that TS geometries are often well-defined regardless of the substrates and reagents, unless large steric clashes are imposed. More importantly, this approximation enabled the use of optimization routines such as those in MacroModel[119] (now available from Schrödinger, http://www.schrodinger.com/MacroModel/) using regular GSFFs like MM2. For example, MM2 parameters for the TS of hydroboration of alkenes were developed using *ab initio* calculations (MP2/6-31G*/3-21G) and were applied successfully to the chemistry of Masamune chiral borane.[120] All of these studies used QM-derived TS structures to derive TS parameters. They were very successful and demonstrated the accuracy of TSFFs using TSs as minima. Unfortunately, such approaches are not easy for non-experts to use, are difficult to automate, and, accordingly, will not be described further herein. With this in mind, more user-friendly methods including Q2MM were developed.

1.4.2 Q2MM

1.4.2.1 Theory

Q2MM (Quantum to Molecular Mechanics) was developed by Norrby and co-workers to generate accurate reaction-specific TSFFs from QM-derived TSs. In this approach, as in TSFFs described above, the TS is considered to be a minimum on the PES

(Figure 1.1 and Figure 1.6). In contrast to the development of other TSFFs, Q2MM makes use of the Hessian of the TS as computed by QM methods. The Hessian is a matrix of the second derivatives of the energy with respect to all atomic coordinates. It provides information about the curvature of the PES (narrow or large energy well) and relates to the vibrational frequencies on each atom (which relate to the normal modes of vibrations observed by IR spectroscopy). As shown in Figure 1.16, the Hessian matrix is diagonalized into a new matrix D (this is achieved by finding a matrix P and its inverse P^{-1} which relates the Hessian matrix to its diagonalized matrix D as given in Figure 1.16). The b_{ii} are the eigenvalues of the Hessian and *P* the matrix made of the eigenvectors. If a given structure has only positive eigenvalues, it is at a minimum on the PES, but if it has a single negative eigenvalue, the structure is a TS. Thus, if a TS is identified (i.e., a single eigenvalue is negative), converting it into a minimum can be achieved by replacing this value (b_{22} in Figure 1.16) by any positive value. Then the Hessian is reconstructed using the new eigenvalues and the eigenvectors.

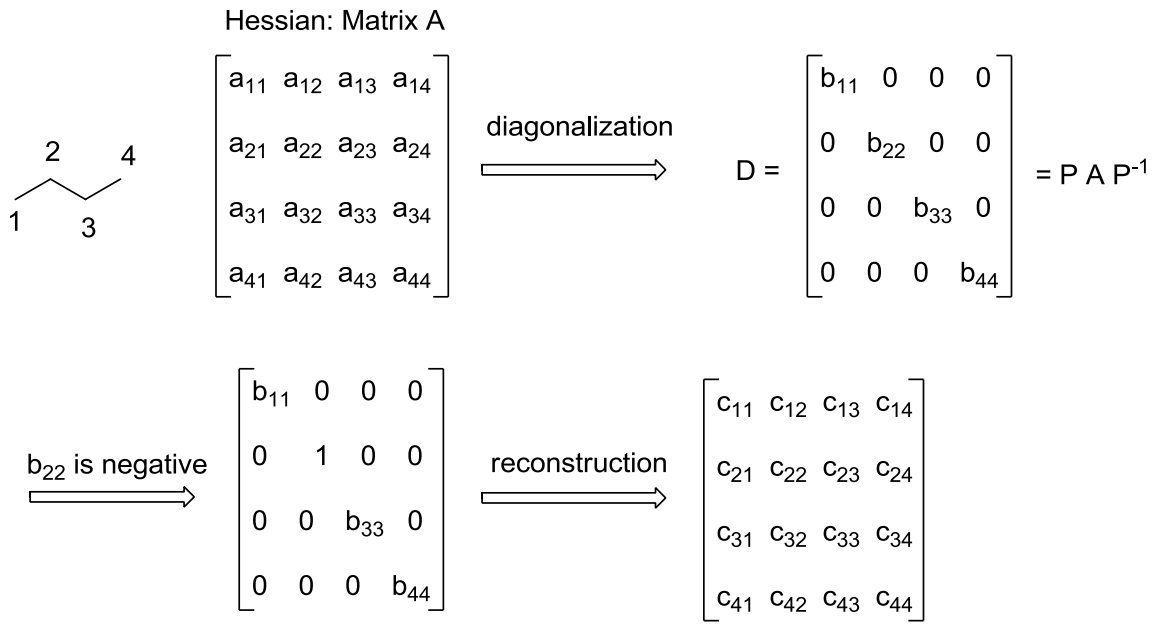


Figure 1.16. Conversion of a saddle point into a minimum within Q2MM matrix.

At this stage, the eigenvalues are all positive and the Hessian represents a minimum. In other words, the TS becomes a minimum on the PES. Through an iterative procedure the FF is next trained to reproduce this new PES, where the TS is a minimum, leading to a

TSFF for this specific reaction. In this process, the Hessian is computed using the FF and the FF is modified to minimize the difference between the reference QM-derived Hessian and the FF-derived Hessian. For more details, the reader is referred to a very detailed protocol reported in 2010.[121]

1.4.2.2 Validation

The Q2MM method has been successfully applied to a variety of asymmetric reactions including the dihydroxylation of alkenes,[122-123] Horner-Wadworth-Emmons (HWE) reactions[124] and diethylzinc addition to aldehydes[125] and more recently, the asymmetric hydrogenation of ketones using Noyori's catalysts (Figure 1.17).[126]

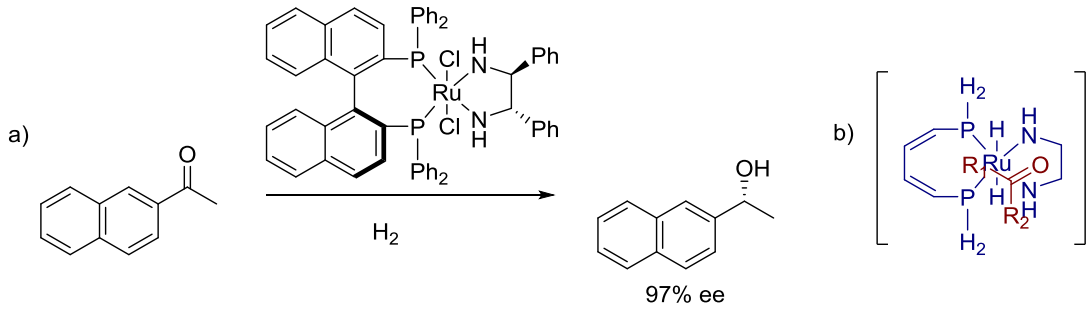


Figure 1.17. a) Asymmetric reduction of ketones with a Ru complex that was used to test the derived TSFF. b) An example of a system used to locate the TS.

Using B3LYP/LACVP* (a density functional and a basis set appropriate for transition metals), Norrby and co-workers located the TSs for a small set of substrates and catalysts containing small diamine ligands (Figure 1.17b). The corresponding Hessian matrices were next used to derive the TSFF parameters needed for MM3*, a force field implemented in MacroModel. The electrostatic parameters were also refined to fit the DFT-derived atomic charges.

This TSFF was then used to study 13 systems including that depicted in Figure 1.17a. The correlation between experimental data and computed ("predicted") data in the form of $\Delta\Delta G^{\ddagger}$, the free energy difference between the two diastereomeric TSs leading to either enantiomer of the product, had $R^2 = 0.83$ with a mean unsigned error well below 0.7

kcal/mol. This high level of accuracy is in line with the aforementioned studies (dihydroxylation, HWE and diethylzinc addition) using Q2MM.

As MCMM and ACE (IMPACTS), Q2MM has been applied to CYP-mediated oxidation.[127] In this study, Norrby, Rydberg and co-workers developed a TSFF for hydrogen abstraction, which is the major oxidation reaction carried out by CYPs, leading to N/O-demethylation and hydroxylation. Other reactions including sulfur oxidation and aromatic oxidation were not considered by this TSFF. Because the heme present in CYPs is responsible for the oxidation, the rest of the protein was removed. Using the DZP/6-31G* basis set (a special basis was necessary for iron), the TSs were located for a number of substrates using a truncated (i.e., smaller) heme only (Figure 1.18).

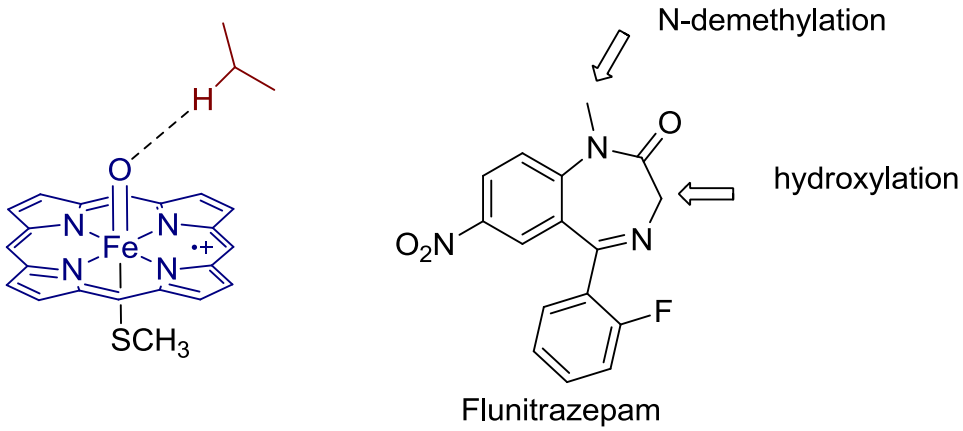


Figure 1.18. CYP-mediated oxidation of sp^3 carbon atoms - model used to derive TSs.

Q2MM was used to develop TS parameters compatible with other GAFF (General Amber FF) parameters,[128] a force field frequently used in protein/ligand modeling. The derived TSFF was first demonstrated to reproduce heme structures derived by DFT accurately, and was then used to predict the most stable TSs (of many possible docking poses) with two different drugs binding to CYP2C9 and CYP3A4.[129] As mentioned throughout this chapter, FFs like GAFF cannot compute relative energies between molecules having different connectivity, including different metabolites (with different site of metabolism, SoM). Consequently, this TSFF (modified GAFF) was not expected to identify the preferred SoM accurately. As a result, the energies of the TSs leading to different SoMs were estimated by combining QM-derived energies and binding energies. This approach proved to be satisfactory in identifying the metabolites for Flunitrazepam

and progesterone with CYP2C9 and CYP3A4. Because this TSFF has already been developed and incorporated into GAFF, any other drugs reacting with any heme-containing enzymes through hydrogen abstraction can now be modeled.

1.4.2.3 Availability

The Q2MM package is available upon request from Per-Ola Norrby. A suitable quantum program (Jaguar or Gaussian in the studies above) is required to generate the TSs and the necessary Hessian(s) and a MM package (MacroModel in the studies above) is required to use the TSFF and combine it with other MM routines such as those used to optimize the structures. The P450 parameters are available for use with AMBER (http://www.teokem.lu.se/~ulf/Methods/ponparm.html).

1.5 Conclusion and Prospects

Over the last fifty years, researchers have used computational methods to rationalize the outcome of organic chemistry transformations and to help in the design of improved catalysts or reagents. Modeling of TS structures has been dominated by QM methods with relatively few efforts dedicated to developing and using faster and more intuitive MM approaches.

In this chapter, we described the available MM methods, as well as their theory and validation studies. Despite these being available to experimentalists, the number of both users and reports of their application to catalyst design has been limited. This is in great contrast to the field of computer-aided drug design (CADD) where many drug companies and academic groups are utilizing easy-to-use software for discovering new research avenues. We believe that the methods described herein will remain relatively unknown and underused by experimentalists until the dependence on QM data is alleviated. That being said, these software packages are based on well-established theory described in this chapter and they are often useable with limited training. Consequently, we recommend contacting any developers of the software you wish to explore. Moreover, reactions already investigated/parameterized will not require any additional QM manipulation.

To further advance the field of TS modeling using MM methods, we propose to integrate computational chemistry into organic synthesis laboratories as well as create an environment at the educational level where using software becomes routine and is not feared by those without expertise in the development process. It is our belief that once modeling techniques are more established at the fundamental learning stages for any chemist, the novelty in chemical research will greatly expand; computational methods will improve and consequently accounts of their use will grow exponentially.

1.6 References

1. Corbeil, C. R.; Moitessier, N., Theory and Application of Medium to High Throughput Prediction Methods Techniques for Asymmetric Catalyst Design. *J*. *Mol*. *Cat*. *A* **2010,** *324*, 146-155
2. Hatanaka, M.; Maeda, S.; Morokuma, K., Sampling of transition states for predicting diastereoselectivity using automated search method - Aqueous lanthanide-catalyzed Mukaiyama aldol reaction. *J*. *Chem*. *Theor*. *Comp*. **2013,** *9* (7), 2882-2886.
3. Friesner, R. A.; Abel, R.; Goldfeld, D. A.; Miller, E. B.; Murrett, C. S., Computational methods for high resolution prediction and refinement of protein structures. *Curr*. *Opin Struct*. *Biol*. **2013,** *23* (2), 177-184.
4. Hill, J. R.; Freeman, C. M.; Subramanian, L., Use of force fields in materials modeling. In *Reviews in Computational Chemistry*, 2000; Vol. 16, pp 141-216.
5. Canongia Lopes, J. N.; Pádua, A. A. H., CL&P: A generic and systematic force field for ionic liquids modeling. *Theor*. *Chem*. *Acc*. **2012,** *131* (3), 1-11.
6. Ditzler, M. A.; Otyepka, M.; Šponer, J.; Walter, N. G., Molecular dynamics and quantum mechanics of RNA: Conformational and chemical change we can believe in. *Acc*. *Chem*. *Res*. **2010,** *43* (1), 40-47.
7. Farah, K.; Müller-Plathe, F.; Böhm, M. C., Classical reactive molecular dynamics implementations: State of the art. *ChemPhysChem* **2012,** *13* (5), 1127-1151.

8. Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A., Development of a polarizable force field for proteins via ab initio quantum chemistry: First generation model and gas phase tests. *J*. *Comp*. *Chem*. **2002,** *23* (16), 1515-1531.

9. Lii, J. H.; Allinger, N. L., The important role of lone-pairs in force field (MM4) calculations on hydrogen bonding in alcohols. *J*. *Phys*. *Chem*. *A* **2008,** *112* (46), 11903-11913.

10. Lin, E.; Shell, M. S., Convergence and Heterogeneity in Peptide Folding with Replica Exchange Molecular Dynamics. *J*. *Chem*. *Theo*. *Comput*. **2009,** *5* (8), 2062-2073.

11. Liu, Z.; Ensing, B.; Moore, P. B., Quantitative Assessment of Force Fields on Both Low-Energy Conformational Basins and Transition-State Regions of the (ϕ−ψ) Space. *J*. *Chem*. *Theo*. *Comput*. **2010,** *7* (2), 402-419.

12. Mu; Kosov, D. S.; Stock, G., Conformational Dynamics of Trialanine in Water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS Force Fields to NMR and Infrared Experiments. *J*. *Phys*. *Chem*. *B*. **2003,** *107* (21), 5064-5073.

13. Zaman, M. H.; Shen, M.-Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R., Investigations into Sequence and Conformational Dependence of Backbone Entropy, Inter-basin Dynamics and the Flory Isolated-pair Hypothesis for Peptides. *J*. *Mol*. *Biol*. **2003,** *331* (3), 693-711.

14. Shi, Z.; Chen, K.; Liu, Z.; Kallenbach, N. R., Conformation of the Backbone in Unfolded Proteins. *Chem*. *Rev*. **2006,** *106* (5), 1877-1897.

15. Wickstrom, L.; Okur, A.; Simmerling, C., Evaluating the performance of the FF99SB force field based on NMR scalar coupling data. *Biophysical Journal* **2009,** *97* (3), 853-856.

16. Hu, H.; Elstner, M.; Hermans, J., Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine “dipeptides” (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Prot*. *Struct*. *Func*. *Bioinf*. **2003,** *50* (3), 451-463.

17. Zaman, M. H.; Shen, M. Y.; Berry, R. S.; Freed, K. F.; Sosnick, T. R., Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the flory isolated-pair hypothesis for peptides. *Journal of Molecular Biology* **2003,** *331* (3), 693-711.

18. Liu, Z.; Ensing, B.; Moore, P. B., Quantitative assessment of force fields on both low-energy conformational basins and transition-state regions of the (ø-ψ) space. *Journal of Chemical Theory and Computation* **2011,** *7* (2), 402-419.

19. Pettersson, I.; Liljefors, T., Molecular mechanics calculated conformational energies of organic molecules: A comparison of force fields. In *Reviews in Computational Chemistry*, 1996; Vol. 9, pp 167-189.

20. Bernard R. Brooks, R. E. B., Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus, CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J*. *Comp*. *Chem*. **1983,** *4*, 187-217.

21. Allinger, N. L.; Yuh, Y. H.; Lii, J. H., Molecular mechanics. The MM3 force field for hydrocarbons. 1. *J*. *Am*. *Chem*. *Soc*. **1989,** *111* (23), 8551-8566.

22. Lii, J. H.; Allinger, N. L., Molecular mechanics. The MM3 force field for hydrocarbons. 2. Vibrational frequencies and thermodynamics. *J*. *Am*. *Chem*. *Soc*. **1989,** *111* (23), 8566-8575.

23. Lii, J. H.; Allinger, N. L., Molecular Mechanics. The MM3 force field for hydrocarbons. 3. The van der Waals' potentials and crystal data for aliphatic and aromatic hydrocarbons. *J*. *Am*. *Chem*. *Soc*. **1989,** *111* (23), 8576-8582.

24. Morse, P. M., Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys*. *Rev*. **1929,** *34* (1), 57-64.

25. Jalaie, M.; Lipkowitz, K. B., Published force field parameters for molecular mechanics, molecular dynamics, and Monte Carlo simulations. In *Reviews in Computational Chemistry*, 2000; Vol. 14, pp 441-486.

26. Mackerell Jr, A. D., Empirical force fields for biological macromolecules: Overview and issues. *J*. *Comp*. *Chem*. **2004,** *25* (13), 1584-1604.

27. Zhu, X.; Lopes, P. E. M.; Mackerell, A. D., Recent developments and applications of the CHARMM force fields. *Comput*. *Mol*. *Sci*. **2012,** *2* (1), 167-185.

28. Yang, Z. Z.; Wang, J. J.; Zhao, D. X., Valence state parameters of all transition metal atoms in metalloproteins - Development of ABEEMσπ fluctuating charge force field. *J*. *Comput*. *Chem*. **2014,** *35* (23), 1690-1706.

29. Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; Distasio Jr, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T., Current status of the AMOEBA polarizable force field. *J*. *Phys*. *Chem*. *B* **2010,** *114* (8), 2549-2564.

30. Eksterowicz, J. E.; Houk, K. N., Transition-state modeling with empirical force fields. *Chem*. *Rev*. **1993,** *93* (7), 2439-2461.

31. Ewig, C. S.; Berry, R.; Dinur, U.; Hill, J. R.; Hwang, M. J.; Li, H.; Liang, C.; Maple, J.; Peng, Z.; Stockfisch, T. P.; Thacher, T. S.; Yan, L.; Ni, X.; Hagler, A. T., Derivation of class II force fields. VIII. Derivation of a general quantum mechanical force field for organic compounds. *J*. *Comp*. *Chem*. **2001,** *22* (15), 1782-1800.

32. Westheimer, F. H., A calculation of the energy of activation for the racemization of 2,2'-dibromo-4,4'-dicarboxydiphenyl. *J*. *Chem*. *Phys*. **1947,** *15* (5), 252-260.

33. Dahlgren, M. K.; Schyman, P.; Tirado-Rives, J.; Jorgensen, W. L., Characterization of biaryl torsional energetics and its treatment in OPLS all-atom force fields. *J*. *Chem*. *Inf*. *Model*. **2013,** *53* (5), 1191-1199.

34. Eyring, H., The activated complex in chemical reactions. *J*. *Chem*. *Phys*. **1935,** *3* (2), 63-71.

35. Jensen, F.; Norrby, P. O., Transition states from empirical force fields. *Theo*. *Chem*. *Acc*. **2003,** *109* (1), 1-7.

36. Gundertofte, K.; Liljefors, T.; Norrby, P. O.; Pettersson, I., A comparison of conformational energies calculated by several molecular mechanics methods. *J*. *Comp*. *Chem*. **1996,** *17* (4), 429-449.

37. Van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard III, W. A., ReaxFF: A reactive force field for hydrocarbons. *J*. *Phys*. *Chem*. *A* **2001,** *105* (41), 9396-9409.

38. Chenoweth, K.; Van Duin, A. C. T.; Goddard Iii, W. A., ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *J*. *Phys*. *Chem*. *A* **2008,** *112* (5), 1040-1053.

39. Kim, S. Y.; Kumar, N.; Persson, P.; Sofo, J.; Van Duin, A. C. T.; Kubicki, J. D., Development of a ReaxFF reactive force field for titanium dioxide/water systems. *Langmuir* **2013,** *29* (25), 7838-7846.

40. Song, W. X.; Zhao, S. J., Development of the ReaxFF reactive force field for aluminum-molybdenum alloy. *J*. *Mat*. *Res*. **2013,** *28* (9), 1155-1164.

41. Bae, G. T.; Aikens, C. M., Improved reaxFF force field parameters for Au-S-C-H systems. *J*. *Phys*. *Chem*. *A* **2013,** *117* (40), 10438-10446.

42. Van Duin, A. C. T.; Bryantsev, V. S.; Diallo, M. S.; Goddard, W. A.; Rahaman, O.; Doren, D. J.; Raymand, D.; Hermansson, K., Development and validation of a ReaxFF reactive force field for Cu cation/water interactions and copper metal/metal oxide/metal hydroxide condensed phases. *J*. *Phys*. *Chem*. *A* **2010,** *114* (35), 9507-9514.

43. Aryanpour, M.; Van Duin, A. C. T.; Kubicki, J. D., Development of a reactive force field for iron-oxyhydroxide systems. *J*. *Phys*. *Chem*. *A* **2010,** *114* (21), 6298-6307.

44. Labrosse, M. R.; Johnson, J. K.; Van Duin, A. C. T., Development of a transferable reactive force field for cobalt. *J*. *Phys*. *Chem*. *A* **2010,** *114* (18), 5855-5861.

45. Rahaman, O.; Van Duin, A. C. T.; Bryantsev, V. S.; Mueller, J. E.; Solares, S. D.; Goddard III, W. A.; Doren, D. J., Development of a ReaxFF reactive force field for aqueous chloride and copper chloride. *J*. *Phys*. *Chem*. *A* **2010,** *114* (10), 3556-3568.

46. Järvi, T. T.; Kuronen, A.; Hakala, M.; Nordlund, K.; Van Duin, A. C. T.; Goddard Iii, W. A.; Jacob, T., Development of a ReaxFF description for gold. *Eur*. *Phys*. *J*. *B* **2008,** *66* (1), 75-79.

47. Goddard III, W. A.; Van Duin, A.; Chenoweth, K.; Cheng, M. J.; Pudar, S.; Oxgaard, J.; Merinov, B.; Jang, Y. H.; Persson, P., Development of the ReaxFF reactive force field for mechanistic studies of catalytic selective oxidation processes on BiMoOx. *Top*. *Cat*. **2006,** *38* (1-3), 93-103.

48. Han, S. S.; Van Duin, A. C. T.; Goddard III, W. A.; Lee, H. M., Optimization and application of lithium parameters for the reactive force field, ReaxFF. *J*. *Phys*. *Chem*. *A* **2005,** *109* (20), 4575-4582.

49. Senftle, T. P.; Meyer, R. J.; Janik, M. J.; Van Duin, A. C. T., Development of a ReaxFF potential for Pd/O and application to palladium oxide formation. *J*. *Chem*. *Phys*. **2013,** *139* (4).

50. Shan, T. R.; Van Duin, A. C. T.; Thompson, A. P., Development of a ReaxFF reactive force field for ammonium nitrate and application to shock compression and thermal decomposition. *J*. *Phys*. *Chem*. *A* **2014,** *118* (8), 1469-1478.

51. Larsson, H. R.; Van Duin, A. C. T.; Hartke, B., Global optimization of parameters in the reactive force field ReaxFF for SiOH. *J*. *Comp*. *Chem*. **2013,** *34* (25), 2178-2189.

52. Van Duin, A. C. T.; Strachan, A.; Stewman, S.; Zhang, Q.; Xu, X.; Goddard Iii, W. A., ReaxFFSiO reactive force field for silicon and silicon oxide systems. *J*. *Phys*. *Chem*. *A* **2003,** *107* (19), 3803-3811.

53. Schönfelder, T.; Friedrich, J.; Prehl, J.; Seeger, S.; Spange, S.; Hoffmann, K. H., Reactive force field for electrophilic substitution at an aromatic system in twin polymerization. *Chem*. *Phys*. **2014,** *440*, 119-126.

54. Raiteri, P.; Demichelis, R.; Gale, J. D., Development of accurate force fields for the simulation of biomineralization. In *Methods in Enzymology*, 2013; Vol. 532, pp 3-23.

55. Qi, T.; Bauschlicher, C. W.; Lawson, J. W.; Desai, T. G.; Reed, E. J., Comparison of ReaxFF, DFTB, and DFT for phenolic pyrolysis. 1. Molecular dynamics simulations. *J*. *Phys*. *Chem*. *A* **2013,** *117* (44), 11115-11125.

56. van Duin, A.; Verners, O.; Shin, Y. K., Reactive force fields: Concepts of reaxff and applications to high-energy materials. *Int*. *J*. *Energ*. *Mat*. *Chem*. *Prop*. **2013,** *12* (2), 95-118.

57. Lewis, G. N., The atom and the molecule. *J*. *Am*. *Chem*. *Soc*. **1916,** *38* (4), 762-785.

58. A. K. Rappe, M. A. P., D.C. Wiser, J. R. Hart, L.M. Bormann, W. M. Skiff, RFF, Conceptual Development of a Full Periodic Table Force Field for Studying Reaction Potential Surfaces. *J*. *Mol*. *Eng*. **1997,** *7*, 385-400.

59. Rappé, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard Iii, W. A.; Skiff, W. M., UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J*. *Amer*. *Chem*. *Soc*. **1992,** *114* (25), 10024-10035.

60. Dunn, A. R.; Sweet, L. E.; Wiser, D. C.; LoCoco, M. D.; Jordan, R. F., Computational modeling of ansa-zirconocene amide complexes. *Organometallics* **2004,** *23* (24), 5671-5680.

61. Santiso, E. E.; Gubbins, K. E., Multi-scale molecular modeling of chemical reactivity. *Mol*. *Sim*. **2004,** *30* (11-12), 699-748.

62. Ziegler, T.; Autschbach, J., Theoretical methods of potential use for studies of inorganic reaction mechanisms. *Chemical Reviews* **2005,** *105* (6), 2695-2722.

63. Jensen, F., Locating minima on seams of intersecting potential energy surfaces. An application to transition structure modeling. *J*. *Amer*. *Chem*. *Soc*. **1992,** *114* (5), 1596-1603.

64. Houk, K. N.; Duh, H. Y.; Wu, Y. D.; Moses, S. R., Steric models for stereoselectivity of nitrile oxide cycloadditions to chiral alkenes [6]. *J*. *Am*. *Chem*. *Soc*. **1986,** *108* (10), 2754-2755.

65. Dorigo, A. E.; Houk, K. N., Transition structures for intramolecular hydrogen atom transfers: The energetic advantage of seven-membered over six-membered transition structures. *J*. *Am*. *Chem*. *Soc*. **1987,** *109* (7), 2195-2197.

66. Wu, Y. D.; Houk, K. N.; Trost, B. M., Origin of enhanced axial attack by sterically undemanding nucleophiles on cyclohexenones [28]. *J*. *Amer*. *Chem*. *Soc*. **1987,** *109* (18), 5560-5561.

67. Mukherjee, D.; Wu, Y. D.; Fronczek, F. R.; Houk, K. N., Experimental tests of models to predict nucleophilic addition stereochemistries. *J. Am. Chem. Soc.* **1988,** *110* (10), 3328-3330.

68. Sherrod, M. J.; Menger, F. M., "Transition-state modeling" does not always model transition states. *J. Amer. Chem. Soc.* **1989,** *111* (7), 2611-2613.

69. Menger, F. M.; Sherrod, M. J., Origin of high predictive capabilities in transition-state modeling. *J. Am. Chem. Soc.* **1990,** *112* (22), 8071-8075.

70. Olsen, P. T.; Jensen, F., Modeling chemical reactions for conformationally mobile systems with force field methods. *J. Chem. Phys* **2003,** *118* (8), 3523-3531.

71. Anglada, J. M.; Besalú, E.; Bofill, J. M.; Crehuet, R., Prediction of approximate transition states by Bell-Evans-Polanyi principle: I. *J. Comp. Chem.* **1999,** *20* (11), 1112-1129.

72. Vance, R. L.; Rondan, N. G.; Houk, K. N.; Jensen, F.; Borden, W. T.; Komornicki, A.; Wimmer, E., Transition structures for the Claisen rearrangement. *J. Am. Chem. Soc.* **1988,** *110* (7), 2314-2315.

73. Hammond, G. S., A correlation of reaction rates. *J. Am. Chem. Soc.* **1955,** *77* (2), 334-338.

74. Hirao, K.; Kebarle, P., SN2 reactions in the gas phase. Transition states for the reaction: Cl−+RBr=ClR+Br−, where R=CH3, C2H5, and iso-C3H7, from abinitio calculations and comparison with experiment. Solvent effects. *Can. J. Chem.* **1989,** *67* (8), 1262-1267.

75. Hansen, M. B.; Jensen, H. J. A.; Jensen, F., Modeling enzymatic transition states by force field methods. *Int. J. Quant. Chem.* **2009,** *109* (2), 373-383.

76. Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M. J.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005,** *26* (16), 1668-1688

77. Warshel, A.; Weiss, R. M., An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J*. *Amer*. *Chem*. *Soc*. **1980,** *102* (20), 6218-6226.

78. Kim, Y.; Corchado, J. C.; Villà, J.; Xing, J.; Truhlar, D. G., Multiconfiguration molecular mechanics algorithm for potential energy surfaces of chemical reactions. *J*. *Chem*. *Phys*. **2000,** *112* (6), 2718-2735.

79. Truhlar, D. G., Reply to comment on molecular mechanics for chemical reactions. *J*. *Phys*. *Chem*. *A* **2002,** *106* (19), 5048-5050.

80. Eyring, H.; Polanyi, M., Zur Berechnung der Aktivierungswärme. *Die Naturwissenschaften* **1930,** *18* (44), 914-915.

81. Chang, Y. T.; Miller, W. H., An empirical valence bond model for constructing global potential energy surfaces for chemical reactions of polyatomic molecular systems. *J*. *Phys*. *Chem*. **1990,** *94* (15), 5884-5888.

82. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P., AM1: A new general purpose quantum mechanical molecular model. *J*. *Am*. *Chem*. *Soc*. **1985,** *107* (13), 3902-3909.

83. Stewart, J. J. P., Optimization of parameters for semiempirical methods II. Applications. *J*. *Comp*. *Chem*. **1989,** *10* (2), 221-264.

84. J. C. Corchado, Y.-Y. C., E. L. Coitino, and D. G. Truhlar *Gaussrate 8.1*, University of Minnesota, Minneapolis, Minnesota, 1999.

85. Shaik, S.; Hiberty, P. C., Valence bond theory, its history, fundamentals, and applications: A primer. In *Reviews in Computational Chemistry*, 2004; Vol. 20, pp 1-100.

86. Shaik, S.; Hiberty, P. C., The valence bond diagram approach: A paradigm for chemical reactivity. In *Theory and Applications of Computational Chemistry*, 2005; pp 635-668.

87. Nakamura, H., Roles of electrostatic interaction in proteins. *Quart*. *Rev*. *Biophys*. **1996,** *29* (1), 1-90.

88. Warshel, A., *Computer Modeling of Chemical Reactions in Enzymes and Solutions*. Wiley-Interscience: New York, 1991.

89. Shaik, S.; Shurki, A., Valence bond diagrams and chemical reactivity. *Angew*. *Chem*. *Int*. *Ed*. **1999,** *38* (5), 586-625.

90. Åqvist, J.; Warshel, A., Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches. *Chem*. *Rev*. **1993,** *93* (7), 2523-2544.

91. Shurki, A.; Warshel, A., Structure/function correlations of proteins using MM, QM/MM, and related approaches: Methods, concepts, pitfalls, and current progress. In *Advances in Protein Chemistry*, 2003; Vol. 66, pp 249-313.

92. Sauer, J.; Sierka, M., Combining Quantum Mechanics and Interatomic Potential Functions in Ab Initio Studies of Extended Systems. *J*. *Comp*. *Chem*. **2000,** *21* (16), 1470-1493.

93. Albu, T. V.; Corchado, J. C.; Truhlar, D. G., Molecular mechanics for chemical reactions: A standard strategy for using multiconfiguration molecular mechanics for variational transition state theory with optimized multidimensional tunneling. *J*. *Phys*. *Chem*. *A* **2001,** *105* (37), 8465-8487.

94. Kim, K. H.; Kim, Y., Variational transition state theory calculations for the rate constants of the hydrogen scrambling and the dissociation of BH5 using the multiconfiguration molecular mechanics algorithm. *J*. *Chem*. *Phys*. **2004,** *120* (2), 623-630.

95. Lin, H.; Pu, J.; Albu, T. V.; Truhlar, D. G., Efficient molecular mechanics for chemical reactions: Multiconfiguration molecular mechanics using partial electronic structure hessians. *J*. *Phys*. *Chem*. *A* **2004,** *108* (18), 4112-4124.

96. Lin, H.; Zhao, Y.; Tishchenko, O.; Truhlar, D. G., Multiconfiguration molecular mechanics based on combined quantum mechanical and molecular mechanical calculations. *J*. *Chem*. *Theor*. *Comp*. **2006,** *2* (5), 1237-1254.

97. Tishchenko, O.; Truhlar, D. G., Optimizing the performance of the multiconfiguration molecular mechanics method. *J*. *Phys*. *Chem*. *A* **2006,** *110* (50), 13530-13536.

98. Tishchenko, O.; Truhlar, D. G., Global potential energy surfaces with correct permutation symmetry by multiconfiguration molecular mechanics. *J*. *Chem*. *Theor*. *Comp*. **2007,** *3* (3), 938-948.

99. Higashi, M.; Truhlar, D. G., Electrostatically embedded multiconfiguration molecular mechanics based on the combined density functional and molecular mechanical method. *J*. *Chem*. *Theor*. *Comp*. **2008,** *4* (5), 790-803.

100. Tishchenko, O.; Truhlar, D. G., Efficient global representations of potential energy functions: Trajectory calculations of bimolecular gas-phase reactions by multiconfiguration molecular mechanics. *J*. *Chem*. *Phys*. **2009,** *130* (2).

101. Tishchenko, O.; Truhlar, D. G., Non-hermitian multiconfiguration molecular mechanics. *J*. *Chem*. *Theor*. *Comp*. **2009,** *5* (6), 1454-1461.

102. Zhang, Y.; Lin, H., Quantum tunneling in testosterone 6β-hydroxylation by cytochrome P450: Reaction dynamics calculations employing multiconfiguration molecular - mechanical potential energy surfaces. *J*. *Phys*. *Chem*. *A* **2009,** *113* (43), 11501-11508.

103. Han, J. A.; Kim, Y., Multiconfiguration molecular mechanics studies for the potential energy surfaces of the excited state double proton transfer in the 1:1 7-azaindole:H2O complex. *Bull*. *Kor*. *Chem*. *Soc*. **2010,** *31* (2), 365-371.

104. Hwang, J. K.; Warshel, A., How important are quantum mechanical nuclear motions in enzyme catalysis? *J*. *Am*. *Chem*. *Soc*. **1996,** *118* (47), 11745-11751.

105. Sagnella, D. E.; Tuckerman, M. E., An empirical valence bond model for proton transfer in water. *J*. *Chem*. *Phys*. **1998,** *108* (5), 2073-2083.

106. Čuma, M.; Schmitt, U. W.; Voth, G. A., A multi-state empirical valence bond model for weak acid dissociation in aqueous solution. *J*. *Phys*. *Chem*. *A* **2001,** *105* (12), 2814-2823.

107. Ponder, J. W. *TINKER 3.5*, Washington University School of Medicine: St. Louis, MO, 1997.

108. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development and application of a rapid and accurate program based on organic principles. *Angew. Chem. Int. Ed.* **2008,** *47* (14), 2635-2638.

109. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C., The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* **1997,** *101* (16), 3005-3014

110. Weill, N.; Corbeil, C. R.; De Schutter, J. W.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development of the molecular mechanics-based program ACE and application to asymmetric epoxidation reactions. *J. Comp. Chem.* **2011,** *32* (13), 2878-2889.

111. Corbeil, C. R.; Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007,** *47* (2), 435-449

112. Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.-D.; Moitessier, N., Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s. *J. Chem. Inf. Model.* **2012,** *52* (9), 2471-2483.

113. Garbisch, E. W., Strain Effects. I. The Hydroxymethylene Ketone-Aldo Enol Equilibrium. *J. Am. Chem. Soc.* **1965,** *87* (3), 505-510.

114. Garbisch, E. W.; Schildcrout, S. M.; Patterson, D. B.; Sprecher, C. M., Strain Effects. II. Diimide Reductions of Olefins. *J. Am. Chem. Soc.* **1965,** *87* (13), 2932-2944.

115. DeTar, D. F.; Tenpas, C. J., Calculations of steric hindrance in ester hydrolysis based on estimation of van der Waals strain energies of alkanes. *J. Am. Chem. Soc.* **1976,** *98* (15), 4567-4571.

116. Houk, K. N.; Rondan, N. G.; Wu, Y. D.; Metz, J. T.; Paddon-Row, M. N., Theoretical studies of stereoselective hydroborations. *Tetrahedron* **1984,** *40* (12), 2257-2274

117. Eksterowicz, E.; Houk, K. N., Transition-state modeling with empirical force fields. *Chem*. *Rev*. **1993,** *93* (7), 2439-2461.

118. Houk, K. N.; Paddon-Row, M. N.; Rondan, N. G.; Wu, Y. D.; Brown, F. K.; Spellmeyer, D. C.; Metz, J. T.; Li, Y.; Loncharich, R. J., Theory and modeling of stereoselective organic reactions. *Science* **1986,** *231* (4742), 1108-1117.

119. Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendrickson, T.; Still, W. C., Macromodel - an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J*. *Comp*. *Chem*. **1990,** *11* (4), 440-467.

120. Masamune, S.; Kennedy, R. M.; Petersen, J. S.; Houk, K. N.; Wu, Y. D., Organoboron compounds in organic synthesis. 3. Mechanism of asymmetric reduction of dialkyl ketones with (R,R)-2,5-dimethylborolane. *J*. *Am*. *Chem*. *Soc*. **1986,** *108* (23), 7404-7405.

121. Lill, S. O. N.; Forbes, A.; Donoghue, P.; Verdolino, V.; Wiest, O.; Rydberg, P.; Norrby, P.-O., Application of Q2MM to Stereoselective Reactions. *Curr*. *Org*. *Chem*. **2010,** *14* (15), 1629-1645.

122. Norrby, P. O.; Rasmussen, T.; Haller, J.; Strassner, T.; Houk, K. N., Rationalizing the stereoselectivity of osmium tetroxide asymmetric dihydroxylations with transition state modeling using quantum mechanics- guided molecular mechanics. *J*. *Am*. *Chem*. *Soc*. **1999,** *121* (43), 10186-10192

123. Fristrup, P.; Tanner, D.; Norrby, P. O., Updating the asymmetric osmium-catalyzed dihydroxylation (AD) mnemonic: Q2MM modeling and new kinetic measurements. *Chirality* **2003,** *15* (4), 360-368.

124. Norrby, P. O.; Brandt, P.; Rein, T., Rationalization of Product Selectivities in Asymmetric Horner-Wadsworth-Emmons Reactions by Use of a New Method for Transition-State Modeling. *J*. *Org*. *Chem*. **1999,** *64* (16), 5845-5852

125. Rasmussen, T.; Norrby, P. O., Modeling the stereoselectivity of the beta-amino alcohol-promoted addition of dialkylzinc to aldehydes. *J*. *Am*. *Chem*. *Soc*. **2003,** *125* (17), 5130-5138

126. Limé, E.; Lundholm, M. D.; Forbes, A.; Wiest, O.; Helquist, P.; Norrby, P.-O., Stereoselectivity in Asymmetric Catalysis: The Case of Ruthenium-Catalyzed Ketone Hydrogenation. *J*. *Chem*. *Theor*. *Comp*. **2014,** *10* (6), 2427-2435.

127. Rydberg, P.; Olsen, L.; Norrby, P. O.; Ryde, U., General Transition-State Force Field for Cytochrome P450 Hydroxylation. *J*. *Chem*. *Theor*. *Comp*. **2007,** *3* (5), 1765-1773

128. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general Amber force field. *J*. *Comp*. *Chem*. **2004,** *25* (9), 1157-1174

129. Rydberg, P.; Hansen, S. M.; Kongsted, J.; Norrby, P.-O.; Olsen, L.; Ryde, U., Transition-State Docking of Flunitrazepam and Progesterone in Cytochrome P450. *J*. *Chem*. *Theory Comput*. **2008,** *4* (4), 673-681

Introduction to Chapter 2

Many computational medicinal chemistry tools, like docking and supporting protocols, have been developed in our group for many years. It was only in 2008 and later in 2011 that ACE, a tool for predicting the outcome of asymmetric catalysis, was designed and added to the computational arsenal. Unfortunately, the software required programming knowledge and several independent scripts in order to be operated. As a consequence, it was not used to its full potential, although not for a lack of trying by organic chemists in our lab. This chapter presents, in the context of an ongoing synthesis research project, the re-imagining of ACE and its supporting protocols for a greater purpose. No longer would this software only be used by computational experts as automated, usable and user-friendly tools were created to carry out an entire synthetic plan, virtually. Several new algorithms were developed on the basis of a new molecular representation in order to enable the computation/examination of thousands of compounds in the timeframe of one night, or even a lunch break.

(This page was left blank intentionally)

Chapter 2:

VIRTUAL CHEMIST: A Computational Toolbox for Chemists

This chapter describes software design and several validation experiments to demonstrate its accuracy. ACE was previously designed by **Nicolas Moitessier** and **Christopher R. Corbeil**, and was modified in this work.

2.1 Abstract

As soon as a hypothesis is set and a research plan is designed, several tedious, time-consuming and repetitive experiments will often be necessary to develop novel organic chemistry methodologies. Can all these steps and experiments be performed virtually and automatically? To date, although computational chemistry has been instrumental in drug discovery, it has seldom reached the synthetic chemistry laboratories. We proposed to design a fully automated protocol for virtual synthetic methodology covering all the steps, beginning from the selection and ordering of chemicals to the organic reaction simulation and testing of the outcome. The design and development of several algorithms, within the framework of the VIRTUAL CHEMIST, will be presented that recreate: searching chemical catalogs for reagents, reacting these reagents to virtually synthesize catalysts, evaluating their promise for enantioselectivity and all the computational chemistry intricacies in between. Application to the development of asymmetric catalysts for given chemical reactions demonstrated its effectiveness and ease-of-use.

2.2 Introduction

The field of asymmetric catalysis is growing exponentially and is especially valuable to the fields of medicinal chemistry and advanced materials. An organic chemist can consider a specific chemical reaction and often use molecular scaffolds to imagine more complex asymmetric catalysts with the goal of developing a more efficient and stereoselective catalyst. However, the exploration of the vast chemical space and infinite

possibilities is lengthy and tedious (and impossible) if performed using solely traditional experimental techniques.[1-2] Computer simulations have been used with great success in drug discovery to reduce time requirements and labor needs[3-4] so how come similar principles cannot be applied to facilitate the discovery of catalysts for well-known reactions in synthetic organic chemistry. Searching the chemical space would undoubtedly be more efficient using computational power. Furthermore, computationally applying the identified and selected catalysts to a specific reaction is within reach. Therefore, it would appear that the foundation exists for virtual discovery of asymmetric catalysts.

A major difference between computer-aided drug design (CADD) and asymmetric catalyst design is the required accuracy. Drug discovery often investigates molecules hitting a target with reasonable binding affinity; CADD requires accuracy of a few kcal/mol in order to differentiate strong binders from weak or non-binders. In contrast, as little as 1.0 kcal/mol between stereomeric transition states (TSs) can distinguish highly from poorly stereoselective asymmetric catalysts. As a result, a "chemical accuracy" of 1.0 kcal/mol or less is needed.

Nowadays accurately computing the stereoselectivity, or at least rationalizing the stereochemical outcome of an asymmetric chemical transformation, is often possible through the use of quantum mechanical (QM) techniques such as density functional theory (DFT)[5-11] However, despite the demonstration of its feasibility, it is not until 2009 that the first use of DFT on sets of asymmetric catalysts and substrates was reported,[12] with little communicated thereafter.[13] We believe that identifying stereoselective catalysts *in silico* must be significantly faster than discovering them experimentally, or computational methods will only be used to rationalize observations rather than be exploited in the planning stages of synthetic chemistry. An effective, rapid screening of a large number of catalysts therefore precludes the use of QM.

Alternatively, significantly more time-efficient methods have been devised using molecular mechanics (MM) as described in the previous chapter.[14] Since most force fields were designed for ground state modeling, their transferability to diastereomeric TSs had to be addressed. In this context, the Q2MM method, which relies on QM-derived TS

force fields, has demonstrated high accuracy on a number of asymmetric reactions.[15-18] Similarly, our group reported ACE (Asymmetric Catalyst Evaluation), a program accurately predicting the stereochemical outcome of asymmetric reactions defining the TS structure as a linear combination of reactant and product structures.[19-20] Although Q2MM and ACE were good candidates for computer-guided development of asymmetric catalysts, their full integration into experiments had yet to reach its full potential. On one hand, Q2MM's reliance on QM-derived TS force fields limits its use to previously parameterized reactions for non-QM experts.[14] On the other hand, ACE required the construction of starting structures using a combination of manual TS building and independent Linux-based programs: a tedious and unintuitive process.

Currently, the crucial step is to port this technology to the hands of organic chemists as was successfully achieved with NMR spectroscopy, mass spectrometry and HPLC to name a few. Organic chemists run their own routine NMR experiments without the necessity of an in-depth understanding of the technique, but using computational methods requires extensive training in programming and QM or MM. Removing the need for training entails that the input and output be simple, familiar and readable, and that all the simulated experimental stages are fully automated and integrated. It is important to clarify, at this stage, the definition of usable and user-friendly in the context of this research. Usable is meant to refer to routine experiments which an organic chemist would be interested in running. As with NMR, an organic chemist will rarely attempt to investigate new pulse sequences; this is the domain of the experts in the field. Similarly, this software is not meant to replace QM or the modeling experts worldwide. User-friendly is meant to refer to the accessibility and ease-of-use of a software package: the color pictures, automated workflows, checkboxes and intuitive steps that are required to run a virtual experiment. Herein we present our efforts that led to the improvement of ACE 2.2 and the development of novel computational tools fully integrated into an intuitive, usable and user-friendly platform, VIRTUAL CHEMIST.

Overall, an asymmetric catalyst development project can be designed as shown in Figure 2.2. As an example, Gerosa *et al*. recently reported the computer-aided design and development of novel asymmetric organocatalysts for Diels Alder reactions (Figure 2.2).[13] In this excellent piece of work, they first tested catalysts prepared following the

synthetic scheme shown in Figure 2.2a. ONIOM (B3LYP/6-31G*:AM1), a QM-based software package, was then demonstrated to be highly predictive to model the stereoselectivity of the reaction in Figure 2.2b and used to subsequently virtually screen 62 potential catalysts. The most promising one was selected for synthesis and turned out to be the most stereoselective of the series. Unfortunately, no temporal information is reported although it can be estimated that these predictions required time on the order of weeks or even months as well as an expertise in programming. Could this work be reproduced and a good catalyst identified in just a few mouse-clicks at a desktop computer? Several technically challenging, and technologically advanced tools have been developed in this context.

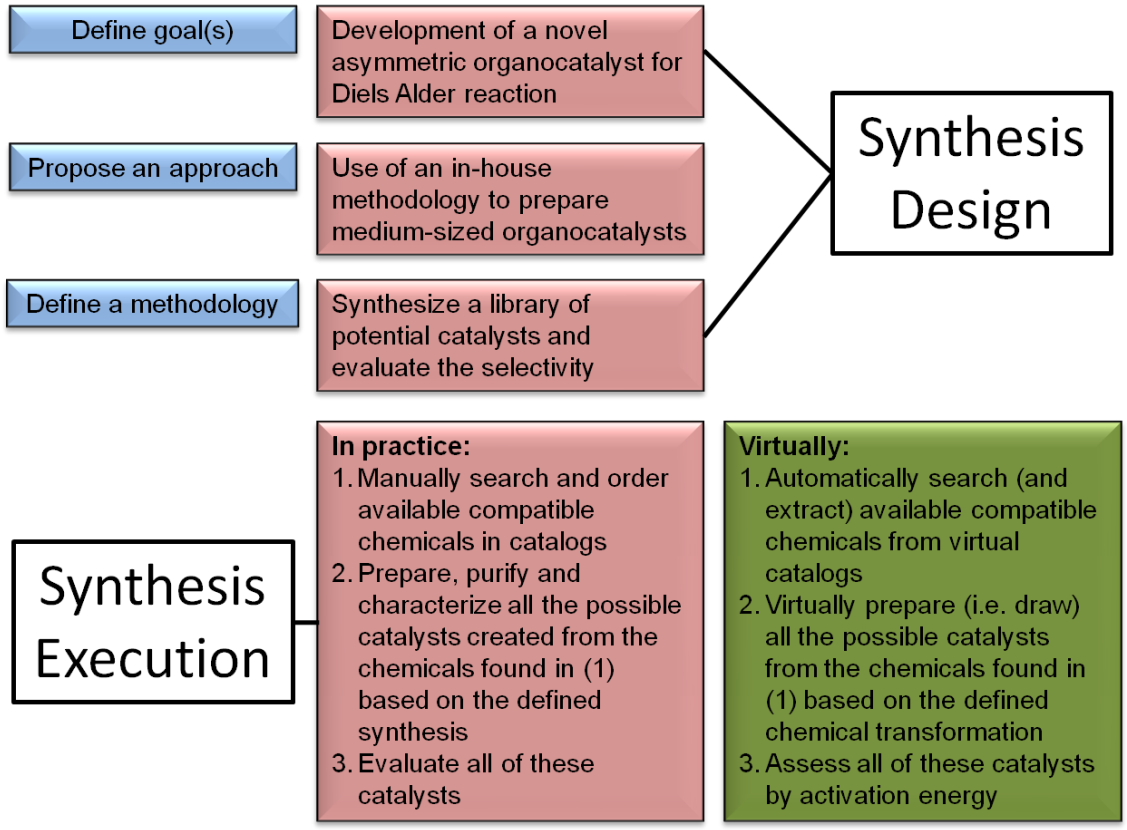


Figure 2.1. General synthesis planning and execution stages with a virtual comparison

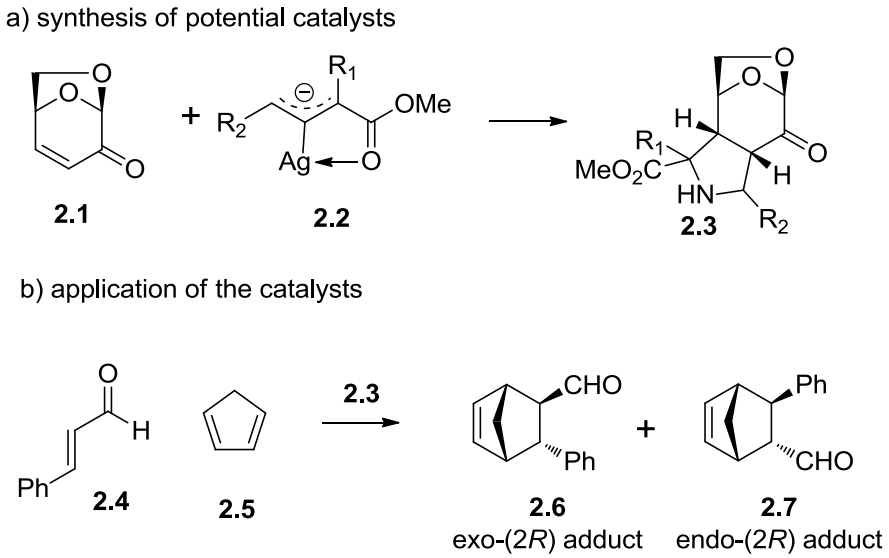


Figure 2.2. a) Reaction scheme defining the synthetic route to an asymmetric organocatalyst. b) The application of this catalyst in a common Diels Alder reaction.

2.3 Theory

2.3.1 Comparing chemical structures

An organic chemist usually wants to find most, if not all, matches of a chemical scaffold with other variable features, referred to as R groups. Usually this is drawn in the form of a 2D structure as shown in Figure 2.2a. This is stored on a computer as a text file of atom names, coordinates and connections. Some transformations are required from this representation in order to establish a comparison. Generally, from an algorithmic perspective, finding a substructure within a structure is known as sub-graph isomorphism. A graph, in computer science, is very similar to a molecule in the sense that it has vertices (atoms) and edges (bonds) and all together these describe the graph (molecule). There are several types of graph-matching: exact, the simplest, substructure and similarity, the most difficult. The three endeavors each have an intrinsic complexity, regardless of relative simplicity, due to the encoding of molecules. SMILES[21] is a method of keeping all the molecular information in one text string, something useful for exact matching, however there can be inconsistencies from one output source to another. Further technical difficulties arise when adding the ambiguities and variability for substructure and

similarity searching. We thought to develop a minimalistic approach to encode molecular structures into a string-like format that could then be compared from one to another using a breadth-first search (BFS) algorithm. This problem is known to be NP-Complete[22] meaning no polynomial-time solution is known. Much of the sub-graph isomorphism problem applied to chemistry is made simpler by the constraints imposed by the nature of molecules.[23] Certain rules exist that, for example, enforce a limit on edges that touch a given vertex (i.e., a carbon atom can have a maximum of 4 bonds) which allows us to take shortcuts and greatly reduce the running time.

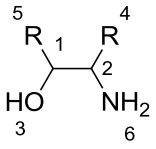
This 'genotypic' representation could then encode the necessary information from the actual molecule (phenotype). We planned to only perform manipulations/computations based on this genotypic representation as opposed to accessing data structures (phenotypes) during each computation. The text-based encoding will then be used to identify scaffold matches, exact matches and largest common substructures between molecules, processes necessary to search for chemicals and encode chemical transformations.

Many of the shortcuts to filtering non-matches and attempts to identify exactitudes as efficiently as possible are based on the approaches of the human brain. A chemist can look at two molecule drawings and almost immediately label them as identical or not. A computer should replicate these efficiencies.

First, before computationally considering the specific properties of the genotypic representations, conceptually, a chemist would look for easily identified, superficial features.[24] If well-selected, this should reduce the run-time of the software significantly. The key notion is to select independent properties that will cover the widest possible range and that will be computationally easy to compare. For example, if the number of atoms and bonds differ from one molecule to the next, then they cannot be exactly the same. Similarly, if they do not contain the same number of each element then they are not identical. Terminal atoms, those that are bound to only one other atom are also an easy target to distinguish dissimilar molecules; like a comparison of picture *frames* before looking at the pictures themselves. These quick filtering techniques allow the presented software to accurately search through 23 million molecular structures for one scaffold in a

matter of hours (ca. 1000 per second), something that would take a human being years or decades and is prone to errors.

If the first filters are passed, the genotypic string is created to represent the molecular structure. When considering two images, one technique to eliminate possible similarity is to start in a distinct region that is likely to differ from one to the other. In this context, the string is rooted, or anchored, at one of the rarest elements appearing in the template and is then expanded in a breadth-first-search manner (Figure 2.3). Then for the second layer, the neighbors of each atom in layer 1, except those appearing in the previous layer (layer 0) will be appended, and so on to eventually create the complete graph/molecule. Within the string, only specific information for defining any atom is kept: the parent (previously visited, connected atom), the element, the atom number, the degree of un-saturation and a flag for cycles (the connecting atom that forms a ring). While more information may be required to describe more complex systems, our current testing does not demonstrate a need for any additions. With this genotype in hand, structures can now be compared with various goals: substructure, exactitude and common substructure identification.



a) anchor atom:

Element	Atom number	Unsaturation	Ring bridge	Layer end
N ;	6 ;	0 ;	0	:

b) following atoms

Parent	Element	Atom number	Unsaturation	Ring bridge	Layer end
6 ;	C ;	2 ;	0 ;	0	:

c) complete string ',' if in same layer

N;6;0;0:6;C;2;0;0:2;R;4;0;0,2;C;1;0;0:1;R;5;0;0:1;O;3;0;0:

Figure 2.3. Example of indexing an amino-alcohol into its genotypic representation.

2.3.2 Substructure search

Finding a substructure (e.g., **template** with R groups) within a structure (e.g., actual chemical from a **library**) is crucial for chemical diversity. In this context, an R-group will

be allowed to be any functional group with no restrictions on element or bond-type. In the case of R=H, some explicit rules were required, as the string representation described above does not include hydrogens. The library phenotypes must now be converted, one at a time, to their genotypes as was demonstrated above for the template. As the string is built, at each layer, the two strings will be compared (Figure 2.4). If, in this comparison, discrepancies arise, the following atom is used as the anchor and a new string is built. If potential matches are not discovered for every single atom in the substructure (after examining all possible genotypic representations) then this library molecule does not contain the template. If the library string is completed and all template atoms have a potential match, it progresses to the next stage of evaluation. To clarify, it cannot yet be concluded that the library molecule contains the template since the strings are built in a forward manner – meaning information of the past is lost; each atom "knows" only to whom they are connected and nothing beyond (Figure 2.4). There are consequently instances that require a more rigorous, extensive investigation; consider the number of times one must go back and forth between two handsomely dressed men to determine if they are wearing exactly the same clothing.

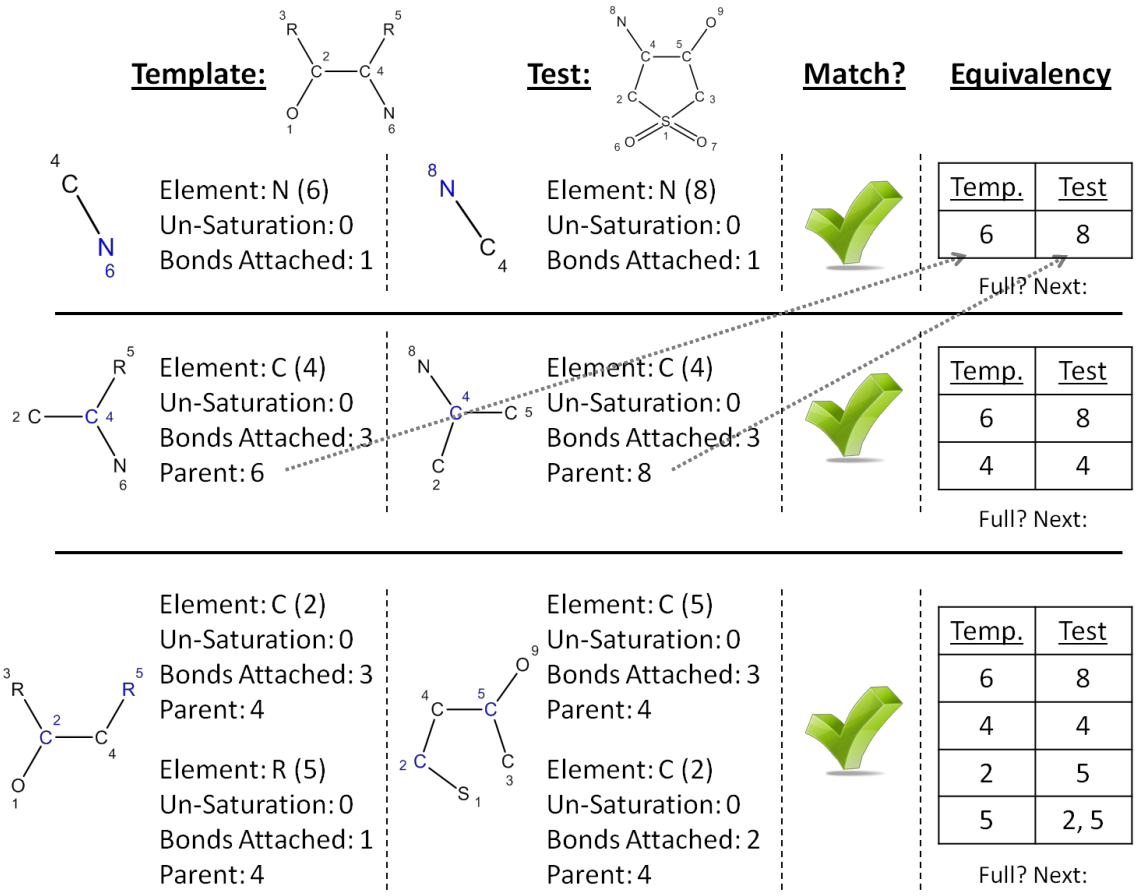


Figure 2.4. A sample run-through of the BFS matching algorithm (only 3 iterations)

A series of deductive tests are carried out to establish the unique corresponding atoms between template and library molecules. Each template atom must have only one counterpart in the library molecule to ensure correctness as well as to properly label the equivalencies. This is achieved by creating new string genotypes anchored at atoms in the template that have more than one potential equivalent atom in the library molecule. If the new genotype does not match the corresponding genotype created from the library molecule, anchored at the potential matching atom, then these atoms cannot be counterparts. As a result, the molecules are evaluated forwards and backwards when these uncertainties arise. If, after any iteration, one template atom is no longer matched to a library atom, the entire molecule is skipped and does not match. A proper substructure is identified if and only if each and every template atom is labeled with one counterpart. This approach accounts for ring junctions (special treatment for 3-member rings) as well as local symmetries found within many molecular structures.

2.3.3 Exact structure matching

All of the components presented in section 2.3.2 are pertinent to exact structure matching except no R-groups are present and every atom in the template must match exactly one in the library molecule. It is in fact much simpler algorithmically to determine if two molecular structures are identical. No consideration of tautomers was included in the development at this stage.

2.3.4 Largest common substructure identification

Establishing the largest common substructure found in two molecules (Figure 2.5) was a challenging task. Unlike the previous two search goals, an anchor is not obvious, the number of corresponding atoms in both structures is variable and the criteria for being labeled an equivalency is much less strict. To put this in perspective, as opposed to examining the picture frames in the case of dissimilarity of one substructure, in this instance, we are searching to cut out the largest piece of a picture that is identical to a piece of the same size in the second picture.

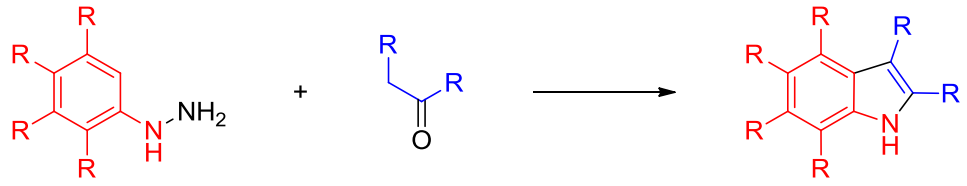


Figure 2.5. Largest common substructures identified for two reagents within a product. Atoms or bonds in black are not part of the largest common substructure.

Utilizing the purpose of this tool (discussed later in this chapter), the only known atom that must be found in both structures is at least one R-group. Therefore, this can be established as the anchor. Upon building and comparing genotypes, many leniencies are allowed when searching for equivalencies. The atoms connected to a given atom, the level of un-saturation of bonds and the ring junctions can all differ between two molecular substructures yet they can still be common substructures. To circumvent these variables, the genotype construction is slightly modified to consider number of bonds as opposed to un-saturation and if this and the element match, it is preliminarily set as a potential counterpart. Using the above-mentioned extensive graph matching anchored at

"promiscuous" atoms (with multiple equivalencies), the result is narrowed to the maximum number of atoms that are labeled with one single counterpart and the number of atoms is stored. The entire largest substructure procedure is repeated from each R atom in order to achieve the largest possible match.

2.3.5 3D substructure matching

The 3D aspect of structure matching adds a geometry element to section 2.3.2, but is otherwise the same. A substructure is identified using the genotypic representations and now this set of equivalent atoms must be superimposed in 3D space and verified for a good geometry overlap (Figure 2.6).

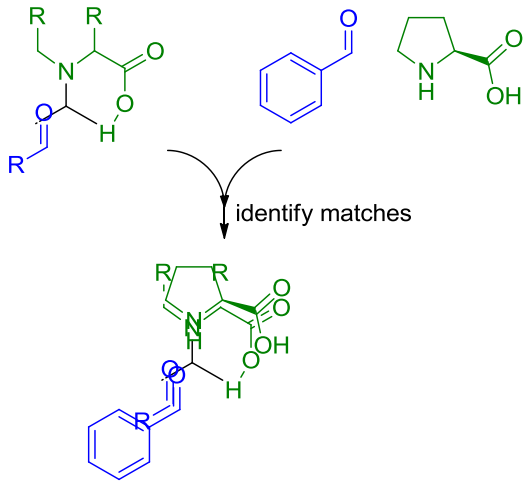


Figure 2.6. Sample 2D superposition (right) of a catalyst onto a TS (left). Extra bonds defined for reactant and product stages and only exist for technical reasons.

This becomes crucial when considering a carbon atom with 3 R-groups. Previously, in 2D, any assignment of R-group would be acceptable. Now, with 3D coordinates existing in both the template and library molecules, the correct R-groups must be paired otherwise a potentially distorted geometry conformation can arise and/or an inversion of configuration can result. The way this is accomplished, similarly to section 2.3.4, is by keeping record of the best match thus far and attempting all possible matching sequences, i.e. checking all genotypic representations that could lead to an acceptable equivalency assignment. In this instance, it is not the number of matching atoms which determines the

best assignment; rather a score is given to the torsion angle distortions and bond stretching upon attempted superposition. Further considerations were required for local symmetry, 3-member rings and preservation of the input stereochemistry (Figure 2.7).

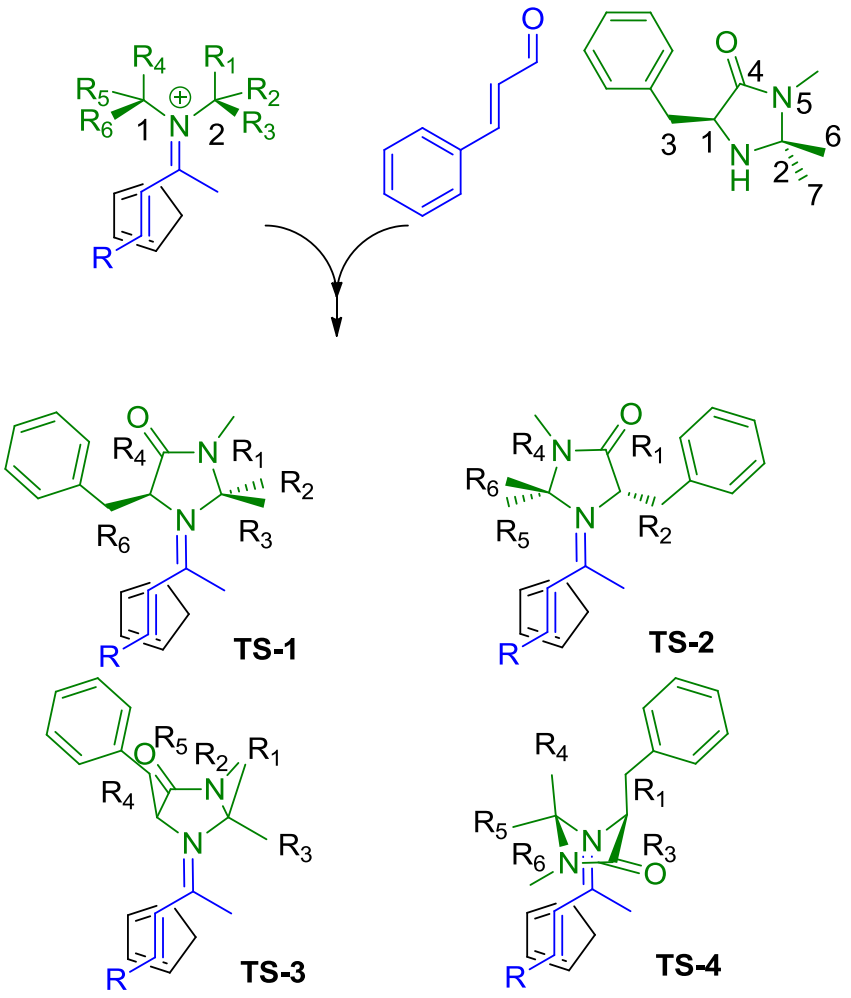


Figure 2.7. The crucial role of geometry in 3D superposition of substructures. The 4 TSs shown represent the potential equivalency labeling, but only TS-1 and TS-2 offer reasonable geometrical conformations.

2.4 Implementation

2.4.1 Searching for chemical scaffolds

An automated sequence of algorithms has been implemented to search through user-defined catalogs based on an input reaction scheme; an organic chemist can draw a 2D scheme in ChemDraw, A+B→C (A→B also acceptable), and run the software package in order to extract libraries of reagents matching scaffolds A and B from the Aldrich catalog, for example, all in a matter of seconds or minutes depending on the generality of the scaffolds. The software package, FINDERS (Filtering, Identifying, Negating Duplicates and Evaluating Reaction Substructures), functions as follows. First, a quick filter is run based on the superficial properties of the reagent(s), mentioned previously, to eliminate "obvious" molecules from the catalog. Next, the given scaffold is expanded for protecting/leaving groups by the developed algorithm: CREATE (Chemical Reagent Expansion After Template Evaluation). Up to 4 "X-groups" are allowed and a list of available options is given in Appendix 1 (Figure 2.8). These groups need to be explicitly defined since they differ from R-groups which can be any functional group.

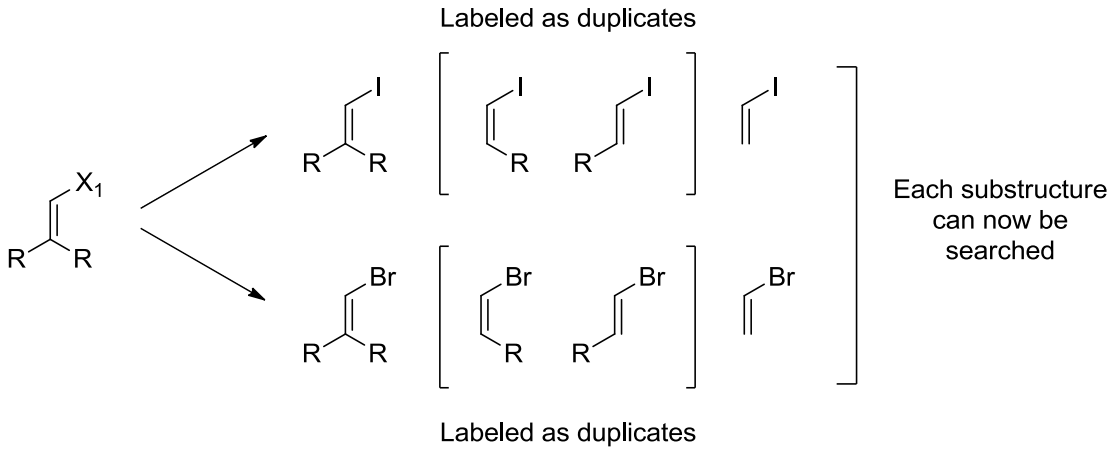


Figure 2.8. An example of CREATE expanding templates for X-groups and R being a functional group or H. Cis and trans is not implemented in the current version of the software and thus these will be labeled as duplicate as described below.

Each generated scaffold is now queried amongst the filtered libraries for substructure matches using the theory in section 2.3.2 and output to a new file. Then, several in-house packages are used to automatically: add hydrogens (CONVERT – Conformational Optimization of Necessary Virtual Enantiomers, Rotamers and Tautomers), label

functional groups (SMART – Small Molecule Atom typing and Rotatable Torsions assignment), and remove user-defined incompatible chemical groups outside the input scaffold (REDUCE – Recognition and Elimination by Descriptors of Undesired Chemical Entities). An organic chemist would not select a molecule containing a second aldehyde if the desired scaffold contained one and this was the reactive site. The final step is to remove duplicate structures since vendors may have multiple salt versions, concentrations, quantities or physical states of the same compound; in molecular files, all of these copies are stored, however only one instance should be kept in the final library. The developed algorithm DIVERSE (Duplicate Identification Validated by Evaluation of Regio- and Stereochemical Exactitudes), uses the theory in section 2.3.3 to evaluate all remaining molecules and eliminate duplicated structures. This is often the most time-consuming step since an exponential number of molecular comparisons are required in its current state.

2.4.2 Performing combinatorial chemistry

The following software package that was developed was to use two chemical libraries to carry out an A+B→C type of reaction and generate all the possible combinations of A and B forming C (or carry out an A→B reaction) again, in a matter of seconds. REACT2D (Rapid Enumeration by an Automated Combinatorial Tool in 2D) works as follows. It is not simply a follow-up to FINDERS and needed to be designed to be an independent set of algorithms if perhaps properly focused libraries were already in-hand. For this reason, many of the algorithms are re-run with some minor modifications. First, CREATE expands the scaffolds once more, but in this instance, leaving/protecting groups are labeled as X after being matched since these groups would not appear in the final product C and would have the same combinatorial outcome. Thus, when DIVERSE is carried out, the same scaffolds differing only by a leaving/protecting group would be labeled as identical so as to not unnecessarily grow the number of combinations. Furthermore, it is applied at this stage and not after the combinatorial chemistry since the number of molecular comparisons is far fewer in separate libraries rather than library of products. While most

of the current available software requires the user to explicitly label atoms belonging to the different reactants in the product, our implementation precludes human interference.

The substructure matching has, therefore, again identified the corresponding atoms between template and library molecules, the exact matching has removed ones that would result in duplicate products and now the reaction can be carried out. Using the theory in section 2.4.4, the reacting templates, A and B, can be matched to the atoms found in C. The leniencies defined regarding largest common substructure are due to bonds forming/breaking and rings being closed from reactant to product. Finally, using matrix algebra, library molecules can be templated onto the products C and be joined in a combinatorial fashion. An overall summary of both FINDERS and REACT2D can be seen in Figure 2.9.

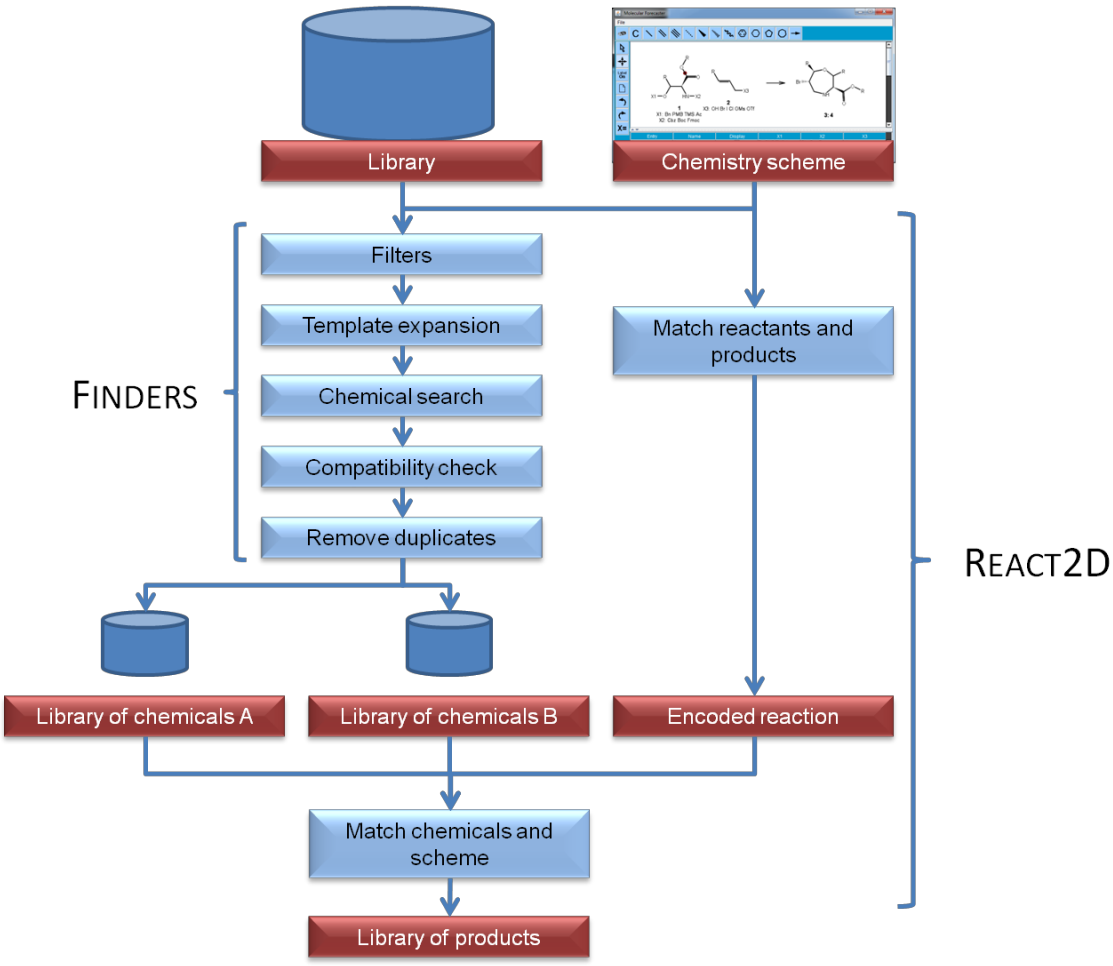


Figure 2.9. Schematic representation of FINDERS and REACT2D going from a 2D reaction scheme and a chemical catalog to combinatorial chemistry.

2.4.3 Building 3D transition states

A software package was created to generate 3D TSs as input for the in-house enantioselectivity evaluation software, ACE (Asymmetric Catalyst Evaluation). ACE requires 3D TSs in a reasonable geometry with two different bond configurations: one at the reactant stage and one at the product stage. Additionally, the input must include the TS leading to both enantiomers in order to output a predicted enantioselectivity for each catalyst. CONSTRUCTS (Converting and Orienting Native Structures on Templates of Rotatable and Unoptimized Chemical Transition States) is fully automated and prepares all the necessary input for ACE, previously a labor-intensive and knowledge-requiring step. The required input is a library of 2D catalysts (and substrates, if desired) and a TS template (can be available from literature or generated using QM techniques). First, using CONVERT, the catalysts are transformed into 3D using a series of MM equations. Then, SMART is applied in order to run an in-house geometry optimization algorithm, MINIMIZE resulting in acceptable 3D conformations. Now, the TS is expanded, similarly to the expansion of templates with CREATE, but with no X-groups, simply Rs can be Hs. The theory in section 2.3.5 is applied and the library of catalysts will match the correct TS and be superimposed in 3D space. Subsequently, the substrates are added (also matched if from a library) and the TS is thus built piecewise and all the combinations of catalysts and substrates are generated (Figure 2.10). The output also includes the two configurations required for ACE as "reactant" and "product" files.

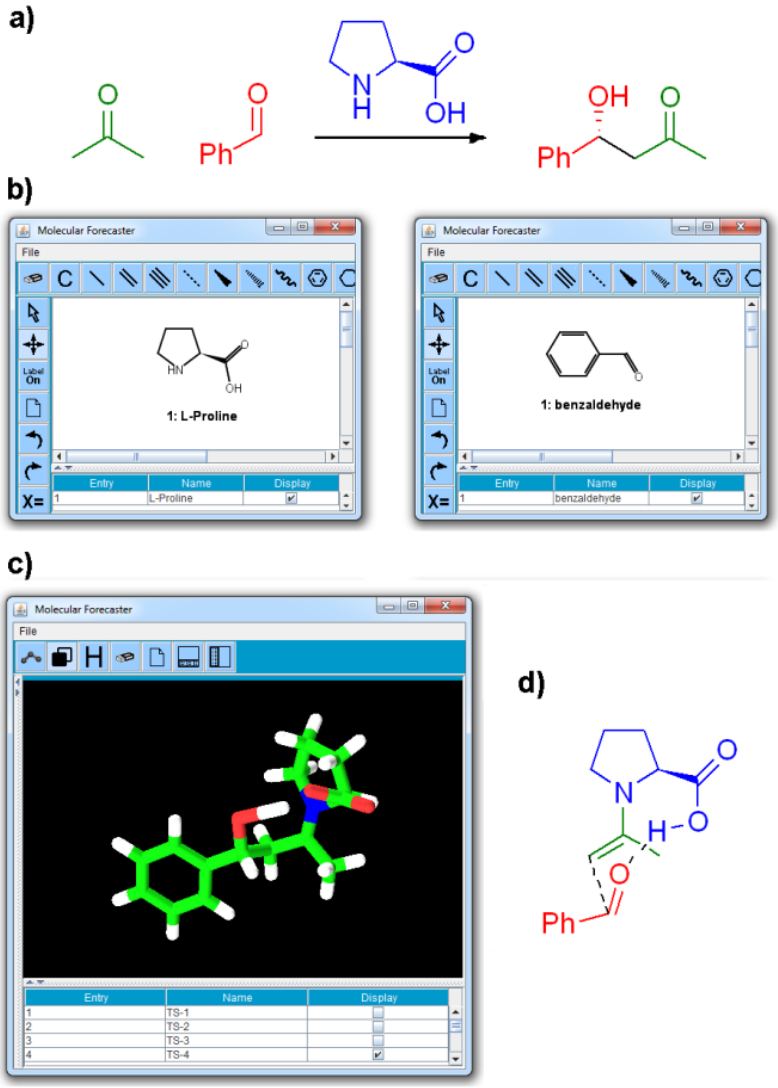


Figure 2.10. From 2D sketches (a) of catalysts and (b) substrates to (c) usable TS structures for proline-catalyzed aldol reaction. (d) The scheme of the TS is given in 2D for clarity.

2.4.4 Evaluating enantioselectivity

ACE underwent minor changes to make it more accessible to users by implementing a parallelization that allows the software to run quicker, and a routine that enables to read and run multiple substrate/catalyst systems. Furthermore, the output was simplified to enable a more efficient analysis of the results. The technical details of ACE[19-20] will not be described in this thesis since it is not the work of the author.

2.5 Validation

2.5.1 FINDERS/REACT2D

In order to validate these virtual combinatorial chemistry tools, a set of chemical transformations among the most widely used in medicinal chemistry were selected as reported by Hartenfeller *et al.*[25] and are shown in Appendix 1 and some examples in Figure 2.11.

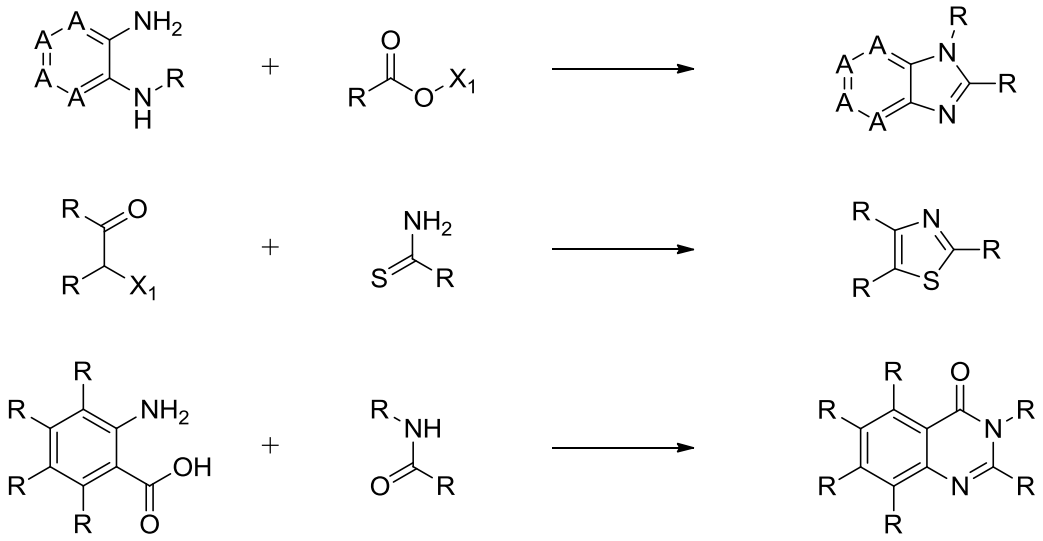


Figure 2.11. Examples of chemical reactions investigated with FINDERS/REACT2D – benzimidazole synthesis (top), thiazole synthesis (middle) and Niementowski quinazoline synthesis (bottom).

Reagent libraries were created using the Aldrich catalog downloaded from the ZINC database (~59000 compounds), the drawn reaction schemes and FINDERS (Table 2.1). It is difficult to exhaustively validate a substructure search algorithm since it requires a known library of matches and decoys as well as a wide diversity of chemical groups, configurations and even atom numbering. The work presented here represents a significant exploration of possibilities, and constant investigation of inconsistencies, but is likely not exhaustive.

Table 2.1. FINDERS validation with a set of 13 diverse chemical reactions. The name of the reaction is given with the number of the reaction from Hartenfeller.[25] MW and N are the maximum molecular weight and number of atoms allowed in the search. Time is the total runtime for FINDERS to search the Aldrich catalog and output the results.

				FINDERS		REDUCE		DIVERSE		Time (min:s)
Entry	**Name**	**MW**	**N**	**A**	**B**	**A**	**B**	**A**	**B**	
1	Pictet-Spengler (1)	150	30	105	3250	20	269	15	257	3:46
2	Benzimidazole synthesis (2)	150	30	16	14309	6	571	6	549	23:11
3	Thiazole synthesis (7)	200	30	398	193	131	94	115	94	2:10
4	Niementowski quinazoline (8)	200	30	86	9011	7	307	7	286	11:51
5	Tetrazole terminal (9)	150	30	2051	n/a	268	N/A	267	N/A	2:30
6	Tetrazole connect (10)	150	30	2051	4489	268	18	267	17	5:10
7	3-Nitrile-Pyridine (17)	150	30	250	n/a	37	N/A	37	N/A	0:20
8	Spiro-Chromanone (18a)	150	30	49	46	7	17	7	10	0:30
9	Paal-Knorr pyrrole (21)	120	18	176	12308	13	450	11	421	13:56
10	Fischer indole (23)	120	18	69	11179	4	283	3	250	12:57
11	Reductive amination (30.1)	100	15	42146	3250	425	30	251	30	38:08
12	Stille coupling (43)	150	30	12929	38	128	4	71	3	29:00
13	Grignard reaction (44)	120	18	5072	2051	176	97	56	96	9:30

REACT2D was then validated (Table 2.2) by visually inspecting a subset of structures and confirming the expected number of products from two reaction types: A+B→C (coupling) or A→B type (transformation). The combinations are indeed exhaustive except for one instance which is a very special case (Figure 2.12). In this situation, 2 R-atoms become 1 atom and technically match the same one. In the current protocol, this is disallowed and the molecule is said to not be a match in REACT2D. In reality, in this specific reaction, this would be disallowed regardless due to geometric constraints.

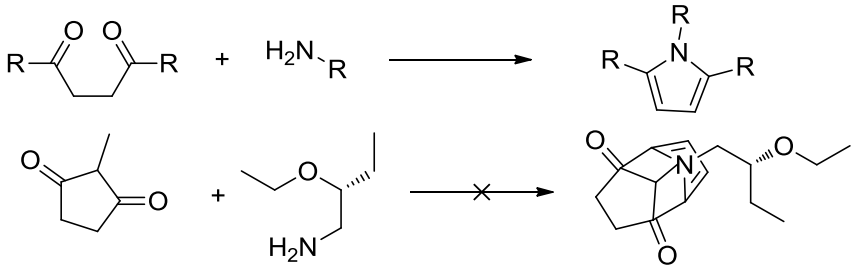


Figure 2.12. Paal-Knorr pyrrole synthesis reaction scheme (top). A demonstration of the disallowed 2 R-group/1 equivalent case (bottom)

Table 2.2. REACT2D validation with a set of 13 diverse chemical reactions. The name of the reaction is given with the number of the reaction from Hartenfeller.[25] Input libraries for A and B are resultant from FINDERS in Table 2.1. Time is the total runtime for REACT2D to execute exhaustive combinatorial chemistry and output the results.

		FINDERS		DIVERSE (X)		REACT2D	Time (min:s)
Entry	**Name**	**A**	**B**	**A**	**B**	**C**	
1	Pictet-Spengler (1)	19	257	15	257	3855	1:15
2	Benzimidazole synthesis (2)	6	550	6	550	3300	9:06
3	Thiazole synthesis (7)	115	94	89	94	8366	0:27
4	Niementowski quinazoline (8)	7	301	7	286	2002	1:41
5	Tetrazole terminal (9)	267	N/A	267	N/A	267	1:10
6	Tetrazole connect (10)	267	17	267	17	4539	1:18
7	3-Nitrile-Pyridine (17)	37	N/A	37	N/A	37	0:01
8	Spiro-Chromanone (18a)	7	17	7	10	70	0:12
9	Paal-Knorr pyrrole (21)	13	421	11 (8)	421	3368	4:54
10	Fischer indole (23)	4	272	3	250	750	1:15
11	Reductive amination (30.1)	425	30	251	30	7530	3:22
12	Stille coupling (43)	128	4	71	3	213	0:14
13	Grignard reaction (44)	164	96	55	96	5280	0:19

2.5.2 CONSTRUCTS/ACE2.2

As a validation of the efficiency and accuracy of the software, we focused on the organocatalyzed aldol and Diels Alder reactions. In 2000, the proline-catalyzed intermolecular aldol reaction was reported by List, Barbas and co-workers[26-27] while the first Diels-Alder "organocatalyst" was reported by MacMillan and co-workers.[28] A number of other organocatalysts have since been reported for these two reactions, including several proline analogues, and chiral binaphthyl diamine derivatives.[29] Interestingly, the reported enantioselectivity rarely exceeds 95% ee despite the large amount of work dedicated to the development of organocatalysts.[30-31] Consequently, these two reactions were good candidates for this study: there was sufficient literature data for validation as well as space for computer-guided improvement.

First, a set of 36 aldol organocatalysts was assembled from the literature with most of them previously experimentally reacted with aldehydes **2.8a-c**. TS templates needed by CONSTRUCTS were built from those reported by Bahmanyar and Houk.[32-33] Ultimately,

108 (36 x 3) catalyst-substrate systems were built for the aldol reaction (examples in Figure 2.13) within under 5 minutes. Since ACE is applied to TS templates, it implicitly assumes a specific reaction mechanism and appropriate catalytic activity. Pyrrolidine is an efficient monofunctional catalyst known to catalyze the aldol reaction by activating the ketone,[34] while the additional carboxylic acid of proline activates the aldehyde making proline a bifunctional catalyst. However, the 9-membered cyclic TS proposed for proline may be competing with an open TS as observed with pyrrolidine and **2.9h**.[35] Additionally, reported stereoselectivities for aldol reactions vary significantly with catalyst loading, solvent,[36] additives and temperature.[37] Unfortunately, the collected data includes stereoselectivities collected under different conditions and therefore includes noise when compared to our automated process which simulates screening under a single set of user-defined conditions. The set consists of two types of catalysts: first, proline and other carbocylic acid or thioamide-containing catalysts[37-38] (e.g., **2.9a-c**) feature acidic protons and assumed to work primarily through the cyclic TS and second, amide-containing catalysts[35] which may provide both cyclic and open TSs (Figure 2.14). These alternate mechanisms are exemplified by **2.9d-f**. It was observed that the yield and enantioselectivity dropped when going from medium-sized and slightly acidic (electronwithdrawing CF_3 group, 88% yield, 45% ee) **2.9e** to less acidic and larger tertbutylamide **2.9f** (55% yield, 15% ee). Experimentally, reports have shown that the acidity of this proton was critical likely due to these two co-existing mechanisms.[38]

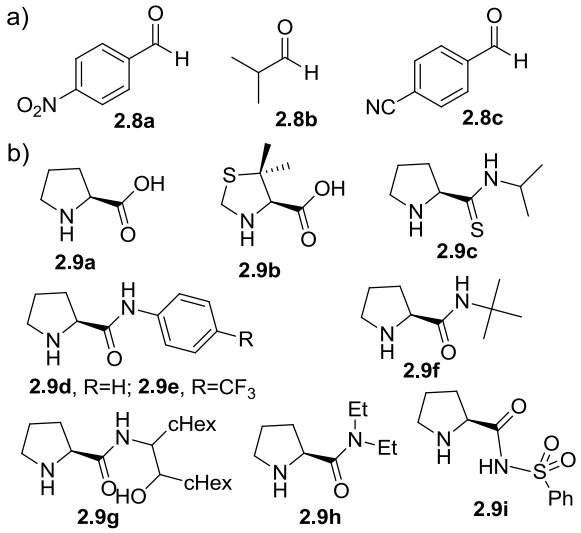


Figure 2.13. Known substrates (a) and catalysts (b) for organocatalyzed aldol reactions.

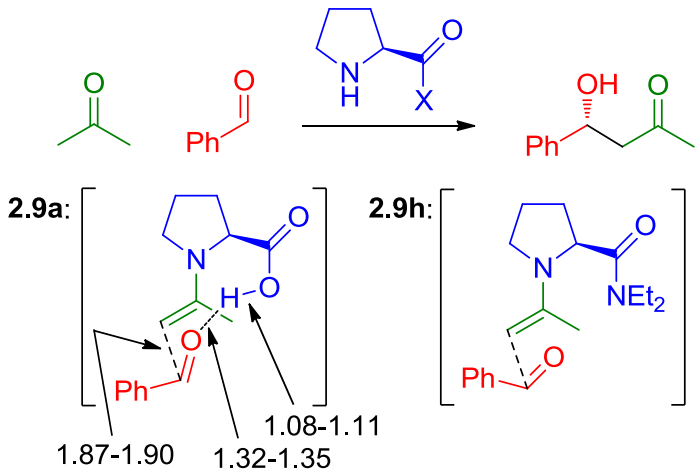


Figure 2.14. Cyclic and open TSs for organocatalyzed aldol reaction. Simple systematic tests revealed that $\lambda_{C\text{-}C1} = 0.4$, $\lambda_{COOH\text{-}O} = 0.1$ and $\lambda_{CONH\text{-}O} = 0.0001$ reproduced the DFT-derived TSs for the aldol reaction catalyzed with proline and prolinamide derivatives.[39]

To test this hypothesis the data was extracted for the first class of catalysts (black in Figure 2.15). This data confirmed that when a single mechanism is in play (i.e., cyclic TS only), ACE can predict enantioselectivity with accuracy high enough to be used to screen organocatalysts for aldol reactions. For the second class (green and red in Figure 2.15), when two pathways may co-exist, the predicted major isomers produced with four catalysts are minor isomers experimentally (top left quadrant in Figure 2.15). In red are highlighted the results with three catalysts including **2.9f** and **2.9g** which are bulky amide-containing systems. Thus, in the cases where the mechanism of action may vary, ACE is capable of predicting the correct isomer in most cases but with poor enantioselectivity accuracy. As an additional limitation sulfonamide derivatives such as **2.9i** were not fully parameterized in MM3 and had to be removed.

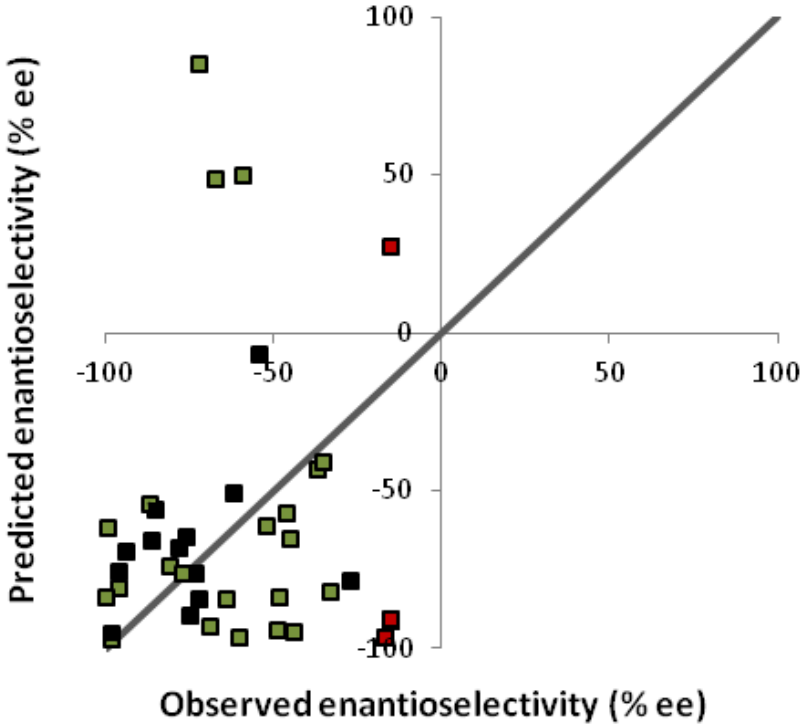


Figure 2.15. Predicted vs. experimental enantioselectivities (the sign indicates one isomer versus the other).

Next, a set of 17 known Diels-Alder organocatalysts and 3 substrates were compiled (examples in Figure 2.16) for both the *endo* and *exo* TSs in ca. 15 minutes (102 systems). TS templates were built from those reported by Gordillo and Houk[8] (Figure 2.17). 48 of these systems have been previously evaluated experimentally (21 *endo* and 27 *exo*).

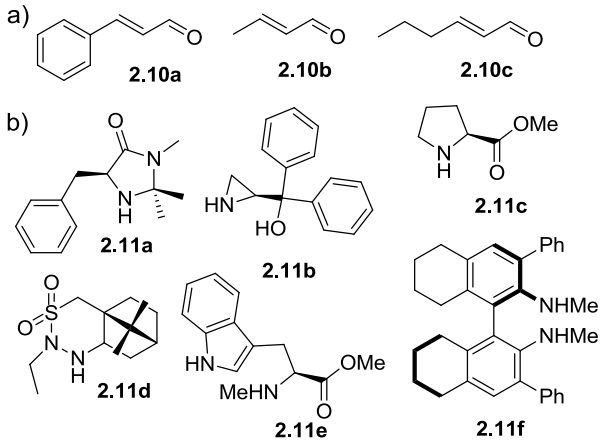


Figure 2.16. Previously reported substrates (a) and catalysts (b) for organocatalyzed Diels-Alder reaction.

The TS systems of catalysts and substrates were built using the fully automated CONSTRUCTS/ACE protocol. The predicted enantioselectivies were then compared to those reported. Poor sulfonhydrazine (**2.11d**) parameters led to unexpectedly high values of ΔΔG* likely due to the lack of parameters in the MM3 force field used by ACE. These were discarded. As can be seen in Figure 2.18, the correlation between predictions and experimental data is apparent. As a ΔΔG* of only 1 kcal can result in an increase of enantioselectivity from 0% ee to ca. 70% ee at room temperature due to the log-scale in the conversion, we looked at the ranking rather than the absolute values. It is important to evaluate whether our program could discriminate the catalysts inducing high or low stereoselectivity since, in practice, a chemist would use this tool to identify those most likely to be successful. As can be seen in Table 2.3, the top of the ranked list (computational prediction) is overall inducing (experimentally) significantly greater stereoselectivity.

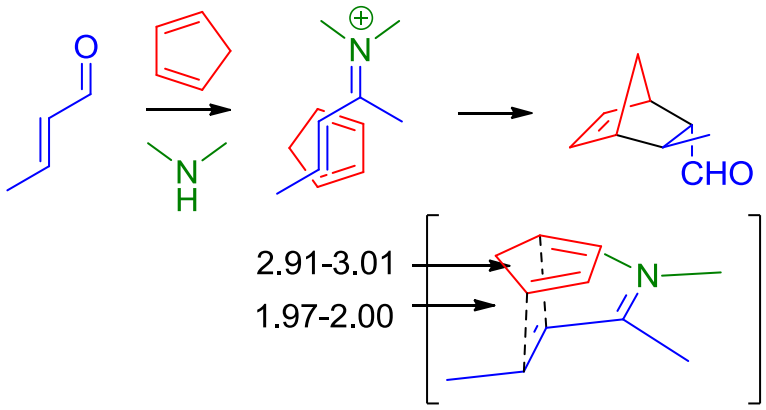


Figure 2.17. Reported TS structures for the asynchronous Diels-Alder reaction. The DFT-derived TSs can be accurately reproduced using $\lambda_{C\text{-}C1} = 0.3$ and $\lambda_{C\text{-}C2} = 0.002$ (for the shortest and longest forming bonds respectively).[8]

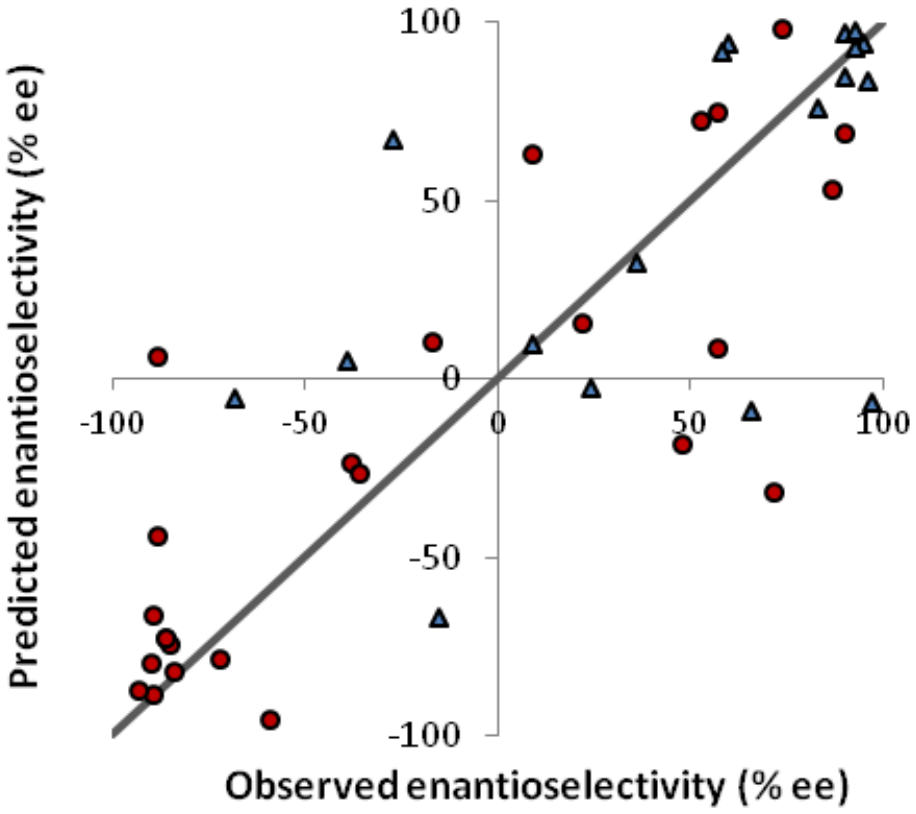


Figure 2.18. Predicted vs. observed enantioselectivities Diels-Alder reactions: *endo* (blue) and *exo* (red) products.

Table 2.3. Identification of stereoselective catalysts for the Diels Alder reaction.

	Av. ee	Top[a]	Middle[a]	Bottom[a]
Diels-Alder - *endo*	63.3%	81.5%	58.8%	50.5%
Diels-Alder - *exo*	66.8%	77.2%	71.0%	45.0%

[a] the ranked list is split into 3 sections (top 33.3%, 33.4-66.6%, bottom 33.3%) and the experimental stereoselectivities averaged

2.5.3 VIRTUAL CHEMIST

With all of these tools in hand, in order to render these tools useable by non-experts, the VIRTUAL CHEMIST interface was developed (Figure 2.19). The sketcher is on top of a workflow made up of the four programs reported herein. Through simple clicks, the

reaction scheme can be drawn and the settings can be chosen. The graphical user interface is intuitive and easy-to-use.

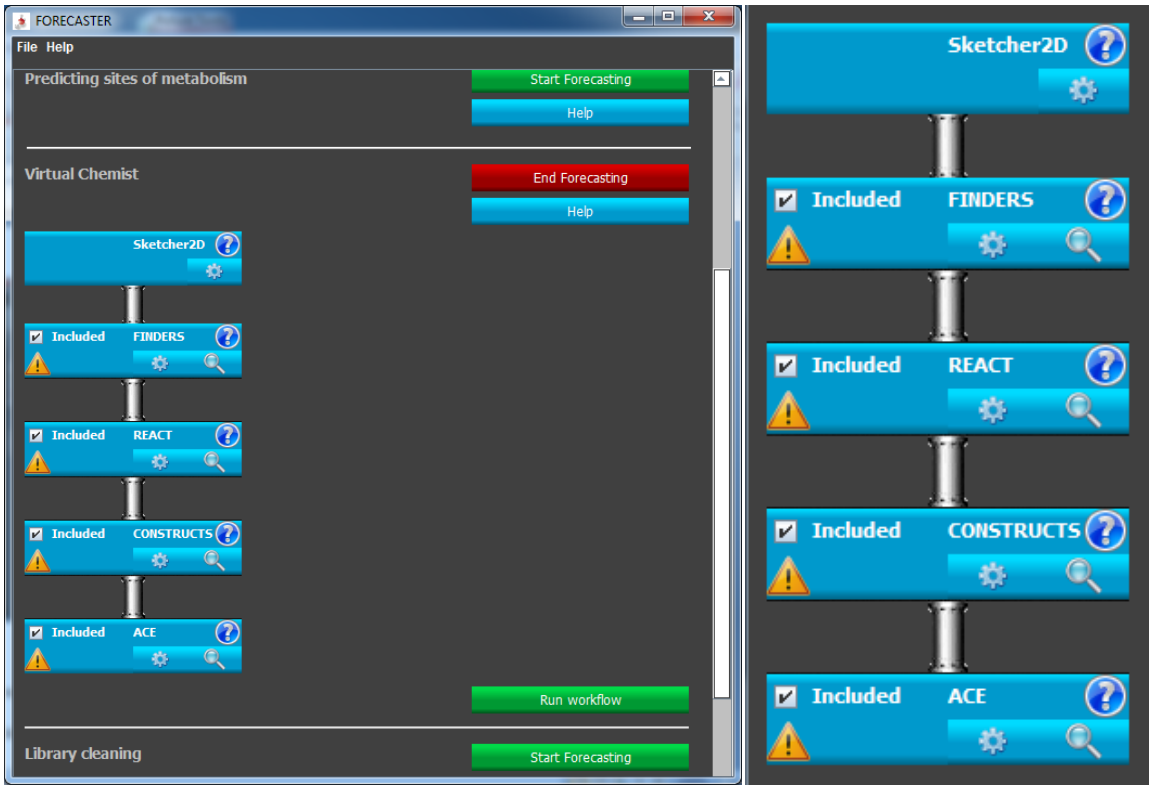


Figure 2.19. The VIRTUAL CHEMIST platform.

2.6 Conclusion

Throughout this work, new implementations that utilize a genotypic representation of molecular structures were applied to multiple facets of asymmetric synthesis research plans. Three new, automated, usable and user-friendly software packages, FINDERS, REACT2D and CONSTRUCTS, were developed to then be followed by an upgraded ACE program. There are few limitations to these algorithms and approaches and the majority of these have been identified and characterized. When competing TSs appear as in the organocatalyzed aldol reaction, ACE is able to correctly identify major isomers albeit with poor enantioselectivity prediction. In the case of the organocatalyzed Diels-Alder reaction, where no competing pathway exists, the accuracy is high enough to enrich a focused library with more stereoselective catalysts. This, along with the integration of

these 4 programs into the VIRTUAL CHEMIST platform, demonstrates that such a computational process, which takes an average of ca. 10h per system (per core), can be an efficient tool to now guide a research plan prior to time-consuming synthesis and testing of new catalysts and not only be a utility in explaining experimental observation.

2.7 Experimental

2.7.1 ACE calculations

Default parameters implemented in ACE have been used for lambda (describing the linear combination of reactants and products) and other genetic algorithm optimization inputs.

2.8 References

1. Alemán, J.; Cabrera, S., Applications of asymmetric organocatalysis in medicinal chemistry. *Chem*. *Soc*. *Rev*. **2013,** *42* (2), 774-793.
2. Marqués-López, E.; Herrera, R. P.; Christmann, M., Asymmetric organocatalysis in total synthesis - A trial by fire. *Nat*. *Prod*. *Rep*. **2010,** *27* (8), 1138-1167.
3. Borhani, D. W.; Shaw, D. E., The future of molecular dynamics simulations in drug discovery. *J*. *Comput*. *Mol*. *Des*. **2012,** *26* (1), 15-26.
4. Durrant, J. D.; McCammon, J. A., Molecular dynamics simulations and drug discovery. *BMC Biol*. **2011,** *9*.
5. Ford, D. D.; Nielsen, L. P. C.; Zuend, S. J.; Musgrave, C. B.; Jacobsen, E. N., Mechanistic Basis for High Stereoselectivity and Broad Substrate Scope in the (salen)Co(III)-Catalyzed Hydrolytic Kinetic Resolution. *J*. *Am*. *Chem*. *Soc*. **2013,** *135* (41), 15595-15608.
6. Wolf, L. M.; Denmark, S. E., A Theoretical Investigation on the Mechanism and Stereochemical Course of the Addition of (E)-2-Butenyltrimethylsilane to

Acetaldehyde by Electrophilic and Nucleophilic Activation. *J*. *Am*. *Chem*. *Soc*. **2013,** *135* (12), 4743-4756.

7. Bahmanyar, S.; Houk, K. N., The origin of stereoselectivity in proline-catalyzed intramolecular aldol reactions. *J*. *Am*. *Chem*. *Soc*. **2001,** *123* (51), 12911-12912.

8. Gordillo, R.; Houk, K. N., Origins of stereoselectivity in Diels-Alder cycloadditions catalyzed by chiral imidazolidinones. *J*. *Am*. *Chem*. *Soc*. **2006,** *128* (11), 3543-3553

9. Lam, Y.-h.; Houk, K. N., Origins of Stereoselectivity in Intramolecular Aldol Reactions Catalyzed by Cinchona Amines. *J*. *Am*. *Chem*. *Soc*. **2015,** *137* (5), 2116-2127.

10. Lam, Y.-h.; Houk, K. N., How Cinchona Alkaloid-Derived Primary Amines Control Asymmetric Electrophilic Fluorination of Cyclic Ketones. *J*. *Am*. *Chem*. *Soc*. **2014,** *136* (27), 9556-9559.

11. Lin, H.; Pei, W.; Wang, H.; Houk, K. N.; Krauss, I. J., Enantioselective Homocrotylboration of Aliphatic Aldehydes. *J*. *Am*. *Chem*. *Soc*. **2013,** *135* (1), 82-85.

12. Schneebeli, S. T.; Hall, M. L.; Breslow, R.; Friesner, R. A., Quantitative DFT Modeling of the Enantiomeric Excess for Dioxirane-Catalyzed Epoxidations. *J*. *Am*. *Chem*. *Soc*. **2009,** *131* (11), 3965-3973

13. Gerosa, G. G.; Spanevello, R. A.; Suárez, A. G.; Sarotti, A. M., Joint Experimental, in Silico, and NMR Studies toward the Rational Design of Iminium-Based Organocatalyst Derived from Renewable Sources. *J*. *Org*. *Chem*. **2015,** *80*, 7626−7634.

14. Pottel, J.; Moitessier, N., Efficient Transition-State Modeling using Molecular Mechanics Force Fields for the Everyday Chemist. *Rev*. *Comp*. *Chem*. **2015,** *29*, in press.

15. Norrby, P. O., Selectivity in asymmetric synthesis from QM-guided molecular mechanics. *J*. *Mol*. *Struct*. *THEOCHEM* **2000,** *506*, 9-16

16. Fristrup, P.; Jensen, G. H.; Andersen, M. L. N.; Tanner, D.; Norrby, P. O., Combining Q2MM modeling and kinetic studies for refinement of the osmium-

catalyzed asymmetric dihydroxylation (AD) mnemonic. *J*. *Organomet*. *Chem*. **2006,** *691* (10), 2182-2198

17. Donoghue, P. J.; Helquist, P.; Norrby, P.-O.; Wiest, O., Prediction of Enantioselectivity in Rhodium Catalyzed Hydrogenations. *J*. *Am*. *Chem*. *Soc*. **2009,** *131* (2), 410-411

18. Lill, S. O. N.; Forbes, A.; Donoghue, P.; Verdolino, V.; Wiest, O.; Rydberg, P.; Norrby, P.-O., Application of Q2MM to Stereoselective Reactions. *Curr*. *Org*. *Chem*. **2010,** *14* (15), 1629-1645.

19. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development and application of a rapid and accurate program based on organic principles. *Angew*. *Chem*. *Int*. *Ed*. **2008,** *47* (14), 2635-2638.

20. Weill, N.; Corbeil, C. R.; De Schutter, J. W.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development of the molecular mechanics-based program ACE and application to asymmetric epoxidation reactions. *J*. *Comput*. *Chem*. **2011,** *32* (13), 2878-2889.

21. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J*. *Chem*. *Inf*. *Comp*. *Sci*. **1988,** *28*, 31-36.

22. Barnard, J. M., Substructure searching methods: Old and new. *J*. *Chem*. *Inf*. *Comp*. *Sci*. **1993,** *33* (4), 532-538.

23. Golovin, A.; Henrick, K., Chemical substructure search in SQL. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (1), 22-27.

24. Crowe, J. E.; Lynch, M. F.; Town, W. G., Analysis of structural characteristics of chemical compounds in a large computer-based file. Part I. Non-cyclic fragments. *Journal of the Chemical Society C: Organic* **1970,** (7), 990-996.

25. Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K. H.; Schneider, G.; Jacoby, E.; Renner, S., A collection of robust organic synthesis

reactions for in silico molecule design. *J*. *Chem*. *Inf*. *Model*. **2011,** *51* (12), 3093-3098.

26. List, B.; Lerner, R. A.; Barbas III, C. F., Proline-catalyzed direct asymmetric aldol reactions [13]. *J*. *Am*. *Chem*. *Soc*. **2000,** *122* (10), 2395-2396.

27. Sakthivel, K.; Notz, W.; Bui, T.; Barbas III, C. F., Amino acid catalyzed direct asymmetric aldol reactions: A bioorganic approach to catalytic asymmetric carbon-carbon bond-forming reactions. *J*. *Am*. *Chem*. *Soc*. **2001,** *123* (22), 5260-5267.

28. Ahrendt, K. A.; Borths, C. J.; MacMillan, D. W. C., New strategies for organic catalysis: The first highly enantioselective organocatalytic diels - Alder reaction [16]. *J*. *Am*. *Chem*. *Soc*. **2000,** *122* (17), 4243-4244.

29. Kano, T.; Tanaka, Y.; Maruoka, K., exo-selective asymmetric Diels-Alder reaction catalyzed by diamine salts as organocatalysts. *Organic Letters* **2006,** *8* (13), 2687-2689.

30. Merino, P.; Marqués-López, E.; Tejero, T.; Herrera, R. P., Enantioselective organocatalytic diels-alder reactions. *Synthesis* **2010,** (1), 1-26.

31. Bhanushali, M.; Zhao, C. G., Developing novel organocatalyzed aldol reactions for the enantioselective synthesis of biologically active molecules. *Synthesis* **2011,** (12), 1815-1830.

32. Bahmanyar, S.; Houk, K. N., Transition states of amine-catalyzed aldol reactions involving enamine intermediates- theoretical studies of mechanism, reactivity, and stereoselectivity. *J*. *Am*. *Chem*. *Soc*. **2001,** *123* (45), 11273-11283.

33. Bahmanyar, S.; Houk, K. N., The origin of stereoselectivity in proline-catalyzed intramolecular aldol reactions. *J*. *Am*. *Chem*. *Soc*. **2001,** *123* (51), 12911-12912.

34. Ji, C.; Peng, Y.; Huang, C.; Wang, N.; Luo, Z.; Jiang, Y., An efficient method for direct aldol reactions catalyzed by pyrrolidine/catechol: The influence of cooperation of Brønsted acidity and hydrogen-bond on the reaction. *J*. *Mol*. *Cat*. *A* **2006,** *246* (1–2), 136-139.

35. Tang, Z.; Jiang, F.; Cui, X.; Gong, L. Z.; Mi, A. Q.; Jiang, Y. Z.; Wu, Y. D., Enantioselective direct aldol reactions catalyzed by L-prolinamide derivatives. *Proc*. *Natl Acad*. *Sci*. *USA* **2004,** *101* (16), 5755-5760.

36. Cobb, A. J. A.; Shaw, D. M.; Longbottom, D. A.; Gold, J. B.; Ley, S. V., Organocatalysis with proline derivatives: Improved catalysts for the asymmetric Mannich, nitro-Michael and aldol reactions. *Org*. *Biomol*. *Chem*. **2005,** *3* (1), 84-96.

37. Gryko, D.; Lipiński, R., L-prolinethioamides - Efficient organocatalysts for the direct asymmetric aldol reaction. *Adv*. *Synth*. *Cat*. **2005,** *347* (15), 1948-1952.

38. Gryko, D.; Chromiński, M.; Pielacińska, D. J., Prolinethioamides versus Prolinamides in Organocatalyzed Aldol Reactions—A Comparative Study. *Symmetry* **2011,** *3*, 265-282.

39. Bahmanyar, S.; Houk, K. N.; Martin, H. J.; List, B., Quantum Mechanical Predictions of the Stereoselectivities of Proline-Catalyzed Asymmetric Intermolecular Aldol Reactions. *J*. *Am*. *Chem*. *Soc*. **2003,** *125* (9), 2475-2479

Introduction to Chapter 3

The application of transition state modeling to proteins while creating accurate computational tools requires a considerable understanding of structural biology. Structure-based design represents a significant scale-up from the small molecule-based approaches seen in the previous chapter. The docking program developed in the Moitessier group, FITTED, and the accessory programs of the FITTED suite will be used throughout the remaining chapters for various applications and the improvements will be described in detail. This chapter reviews the developments of FITTED since its inception in the year 2007 and includes a brief overview of some of the achievements described in later chapters of this thesis. A firm understanding of the milestones in the progression of a software package is required for identifying where certain applications and methodologies may have been overlooked, omitted for technical reasons, or simply unknown.

(This page was left blank intentionally)

Chapter 3:

Docking Ligands into Flexible and Solvated Macromolecules. 8. An Account on the Development of FITTED and other Tools

This chapter has been submitted for publication and is reproduced from the invited review: "Docking Ligands into Flexible and Solvated Macromolecules. 8. An Account on the Development of FITTED and other Tools", Moitessier, N.; Pottel, J.; Therrien, E.; Englebienne, P.; Liu, Z.; Tomberg, A.; Corbeil, C.R.; *Accounts of Chemical Research*, **2015**, submitted. American Chemical Society (2015).

Author Contributions: **Eric Therrien** contributed to the knowledge in sections 3.3 and 3.4.7. **Pablo Englebienne** contributed to the knowledge found in sections 3.4.1 and 3.4.2. **Zhaomin Liu** contributed to G-Quadruplex knowledge in section 3.4.2. **Anna Tomberg** contributed to the knowledge in section 3.4.6 and **Christopher R. Corbeil** contributed to the knowledge of sections 3.4.1, 3.4.2 and 3.4.7. All other sections, figures and overall editing were contributions of the author of this thesis.

3.1 Conspectus

Over the past 20 years, computational methods docking small molecules to proteins have become instrumental in the discovery of novel drugs. Fifteen years ago, we began our first docking-guided drug design project which provided nanomolar metalloproteinase inhibitors and revealed the potential of structure-based drug design. At that time, docking programs primarily considered flexible ligands and rigid proteins. Our subsequent applications of docking programs demonstrated that accounting for protein flexibility and displaceable water molecules, while taking advantage of ligand-based pharmacophores, improved the docking accuracy of existing methods. This prompted us to develop our own program, FITTED, implementing all of these discoveries into a single, automated, and user-friendly software package. Although these simulations are expected to mimic the protein-ligand binding more accurately, several other phenomena, such as binding to metals and covalent inhibitors, must be contemplated to increase its transferability when applied to new therapeutic areas. FITTED has been modified and improved over nearly 10

years by a number of contributors and extensively validated over this period. Since our primary motivation is application-based, several of the concepts behind the evolution of FITTED are rooted in medicinal chemistry projects and collaborations. For example, broad interest in metalloenzymes and, specifically, zinc-metalloenzymes, led us to develop methods considering drug-zinc coordination as well as its effect on the pK_a of surrounding residues. Also, to target covalent enzyme inhibitors, a contentious issue in drug design, as opposed to the usual non-bonded interaction docking, FITTED was updated to identify reactive groups and react them with a given residue (e.g., catalytic residue) when the geometry enables it. This first fully automated covalent docking program was successfully applied to the discovery of covalent prolyl oligopeptidase inhibitors. In order to study drug metabolism by cytochrome P450 enzymes (CYPs), the program IMPACTS was developed as part of FITTED. IMPACTS docks compounds to CYPs and models transition states of oxidation reactions within the catalytic site, and thus predicts the structure of the metabolites. This program opened the door to the use of docking-based methods in biocatalysis.

Our efforts, combined with those of other docking software developers, enabled a better understanding of the complex drug-protein binding process while providing the medicinal chemistry community with useful tools that led to drug discoveries. In this account, we describe, within the historical context, our contributions over the past fifteen years to develop FITTED and additional software programs that have all been integrated into the FORECASTER Platform.

3.2 Introduction

Over the past two decades, traditional medicinal chemistry approaches have been supplemented with a plethora of computational tools, such as docking and quantitative structure-activity relationship (QSAR) methods. In the early 1990s, docking programs primarily investigated the binding between small molecules and rigid proteins. In the late 90s, to better simulate a drug's binding process, Totrov and Abagyan reported a method to dock to flexible proteins,[1] while Lengauer and co-workers implemented discrete water molecules.[2] By 2008, over 60 docking programs were reported.[3]

Our research program relies on the use of predictive computational techniques, specifically docking programs and transition state (TS) predictive tools,[4] to accelerate the drug discovery process. While using existing software for our early medicinal chemistry projects, we rapidly realized that different targets such as nucleic acids, serine proteases, metalloproteinases and various drug classes were treated unequally, and often incorrectly, by all programs. Therefore, we initially added missing features such as protein flexibility and displaceable waters to these programs for a better binding simulation. Although these modifications significantly improved docking accuracy, they also affected the performance (i.e., runtime), user-friendliness, and transferability of the enriched code. In 2007, Moitessier, Corbeil and Englebienne developed FITTED,[5] a docking software integrating protein flexibility and displaceable waters. Since then, the major docking programs, including Surflex,[6] GOLD,[7] AutoDock,[8] and DOCK[9] have also incorporated these two features.[10]

In this account, we wish to summarize the achievements and the current status of our research along with the context for each stage (Figure 3.1). In fact, most of the improvements to the software resulted from the questions arising from our own, or our collaborators', ongoing projects. The foundation of our research programs is to develop software for medicinal chemists based on fundamental knowledge and principles to solve challenges arising in experimental research.

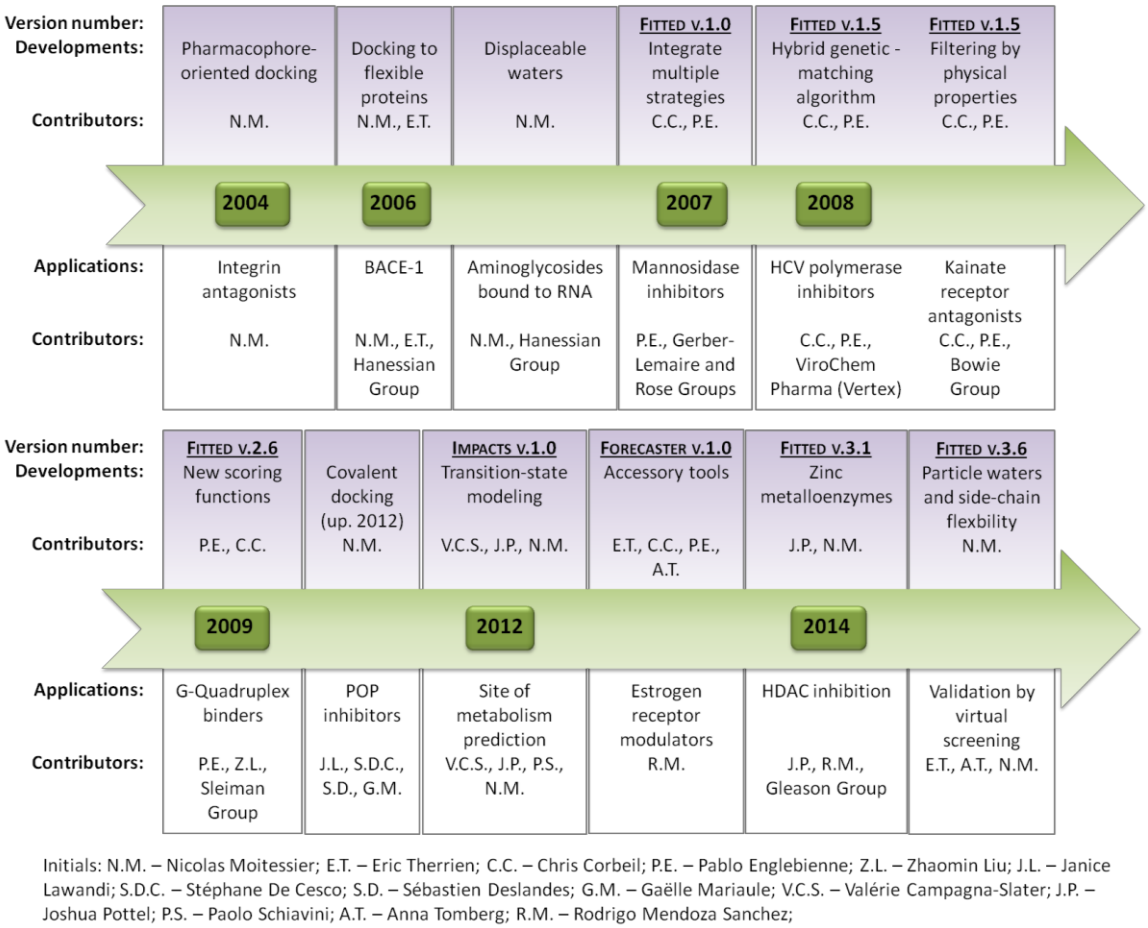


Figure 3.1. Timeline of the development and applications of FITTED.

3.3 The pre-FITTED era

3.3.1 MMP inhibitors 1999-2001

Over fifteen years ago, Moitessier and Therrien, working with Hanessian, were interested in using docking programs to design metalloproteinase (MMP) inhibitors.[11] AutoDock3.0[12] and DOCK4.0,[13] two publicly available docking programs, were tested for their ability to predict the experimentally observed binding modes. Until then, MMP inhibitors were designed in a traditional way but the extensive synthetic efforts to prepare the first series led to no significant inhibitory potency.[14] In parallel, by modifying an existing inhibitor, compound **1** (Figure 3.2), nanomolar sulfonamide derivatives exemplified by **2** and **3** were designed and prepared.[15] When docking was used to

prioritize a subset for synthesis, a number of nanomolar and sub-nanomolar MMP inhibitors were discovered.[16-17]

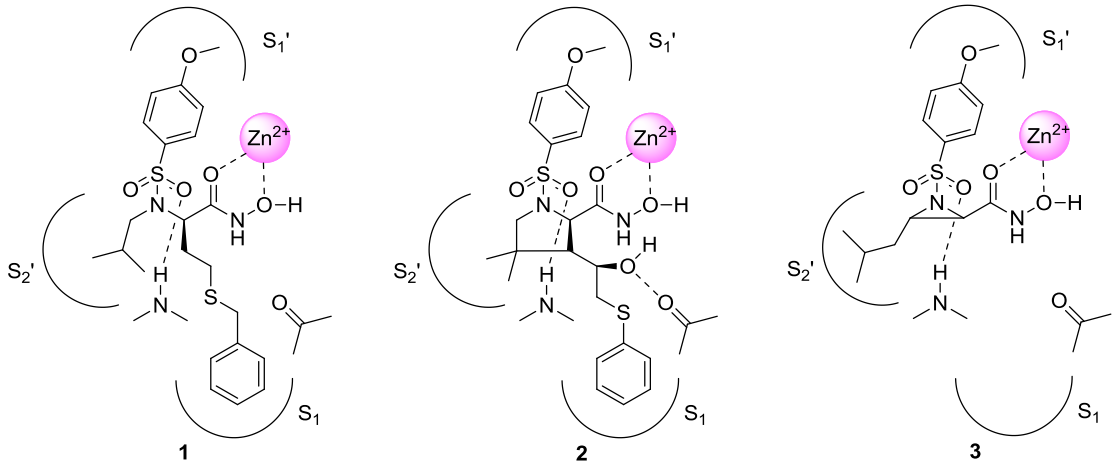


Figure 3.2. MMP inhibitors.

The benefit of using docking programs to guide the design of enzyme inhibitors was clear and so we decided to explore such programs in other medicinal chemistry projects. Unfortunately, this was the first and only time that our use of off-the-shelf software was highly successful. Subsequently, each new project required the implementation of novel ideas and techniques into existing programs to ensure their accuracy.

3.3.2 Integrin antagonists 2002-2003 - pharmacophore oriented docking

In an effort to develop Arg-Gly-Asp (RGD) carbohydrate-based mimetics as $\alpha_V\beta_3$ and $\alpha_{IIb}\beta_3$ antagonists, we planned a combination of computational predictions, advanced parallel and combinatorial syntheses, and cell-based assays. When a crystal structure of the unbound $\alpha_V\beta_3$ receptor came out, AutoDock, FlexX and DOCK were tested for their ability to dock known active antagonists, each featuring a positively charged (mimicking Arg) and a negatively charged (mimicking Asp) functional group. Unfortunately, the latter was predicted to bind to Asp150 while experiments were pointing at Asp218. To correct this, a three-point pharmacophore was derived from reported antagonists, docked

into the *apo* structure and used to orient the docking (Figure 3.3).[18] While this work was ongoing, a crystal structure of this receptor bound to a cyclic peptide was released and validated our proposed binding mode.

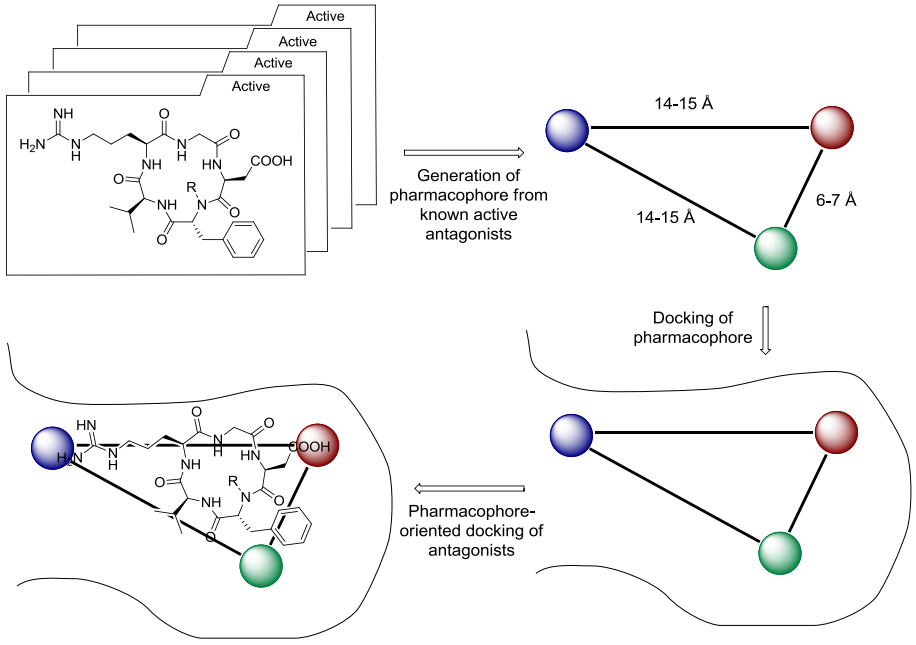


Figure 3.3. Pharmacophore-oriented docking applied to integrin antagonists. Positively charged group (blue bead), negatively charged group (red), aromatic group (green).

3.3.3 BACE-1 inhibitors 2002-2006 - docking to flexible proteins

With several available co-crystal structures, BACE-1, a validated target for Alzheimer's disease therapeutics, has been an excellent candidate for the application of structure-based inhibitor design methods. However, the catalytic di-aspartate protonation state (Figure 3.4) and protein flexibility were critical for optimal BACE-1 inhibitor design, yet difficult to model.[19-20] Moitessier, Therrien and Hanessian teamed up to develop a docking method addressing these challenges.[21]

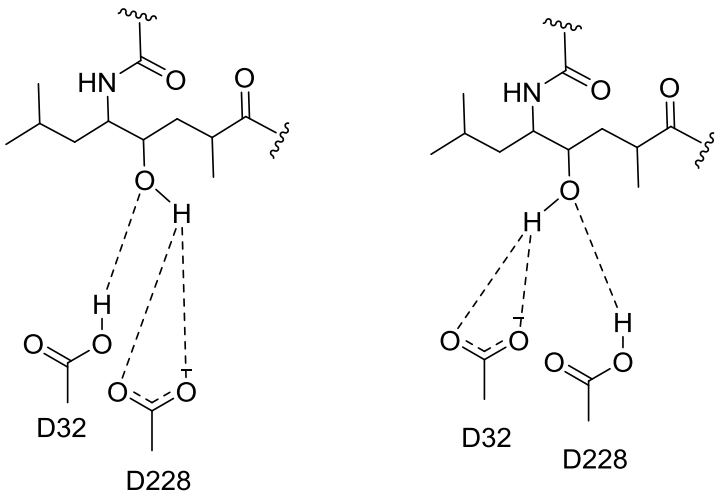


Figure 3.4. Two possible di-aspartate protonation states.

Using the four available BACE-1 crystal structures complexed with different small peptides, we evaluated the accuracy of existing automated docking programs including AutoDock, FlexX and DOCK. Due to the flexibility of both the ligands and the protein, initial docking experiments failed. Our previous successful study of flexible catalysts[22] using a genetic algorithm (GA) led us to optimize flexible protein-ligand complexes with this technique: both the ligand and protein conformations (side chains and backbone) were encoded as chromosomes providing an accurate induced-fit docking method implemented within Accelrys' Insight II. This method was first validated by docking OM99-2 and OM00-3 co-crystallized with BACE-1 with very good accuracy, despite their large number of degrees of freedom (Figure 3.5).

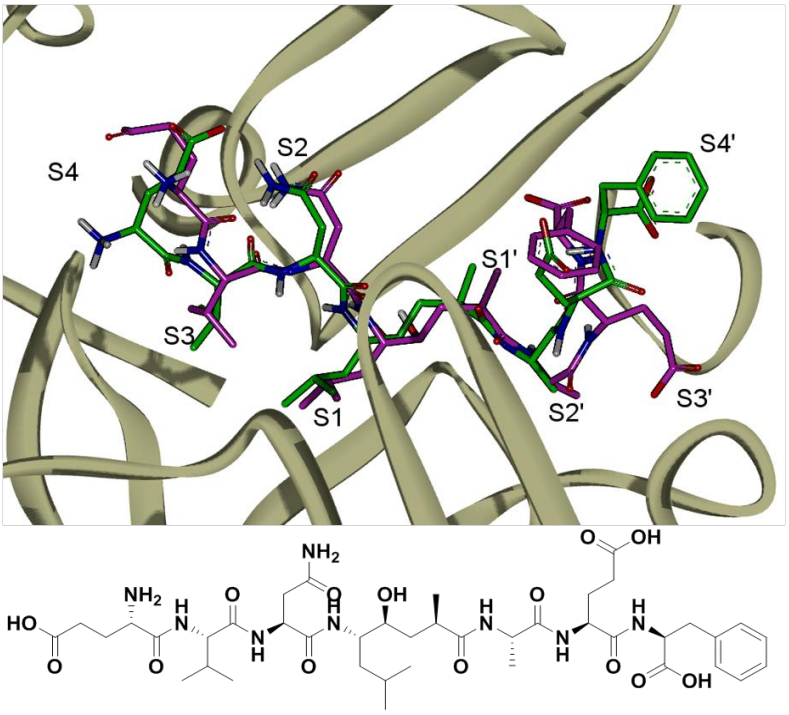


Figure 3.5. BACE-1/OM99-2 crystal structure (green), docked pose (purple), and 2D drawing.

The identification of active compounds requires not only a proper binding mode but also a correct prediction of the free energy of binding, computed with a scoring function. We hypothesized that a better description of the contribution of water binding and entropy loss upon binding would be necessary to discriminate between active and inactive inhibitors. A new force field based scoring function for BACE-1 was developed. RankScore1 included additional terms for hydrogen bonding and a number of rotatable bonds as a surrogate for entropy change upon binding.

$$RankScore1 = c_1 E_{VdW} + c_2 E_{ele} + c_3 E_{HB} + c_4 N_{rot} + c_5 \qquad \textbf{(3.1)}$$

The accuracy of RankScore1 was compared to 15 scoring functions using a validation set of 80 in-house compounds. Combined with our flexible protein docking protocol, RankScore1 outperformed the others for the discrimination of the active compounds. This significant enrichment in active compounds demonstrated the ability of the developed flexible protein docking protocol with the scoring function to accurately predict BACE-1 inhibitors' binding modes. However, this protocol involved multiple computational steps linked together with scripts, which made distribution to medicinal chemists difficult.

3.3.4 Aminoglycoside antibiotics as bacterial RNA binders 2004-2006 - displaceable waters

Another collaborative project with the Hanessian group aimed to use docking methods in guiding the design of aminoglycoside derivatives as potential antibiotics.[23-24] Comparative studies showed that docking programs available at that time were fairly accurately docking molecules to RNA.[25] A close look at crystal structures of aminoglycosides bound to bacterial RNA revealed that the charged ammonium groups of the antibiotics were often not directly interacting with the negatively charged phosphate backbone (Figure 3.6). In fact, it was highly polarized bridging waters connecting the RNA to the aminoglycoside charged groups.

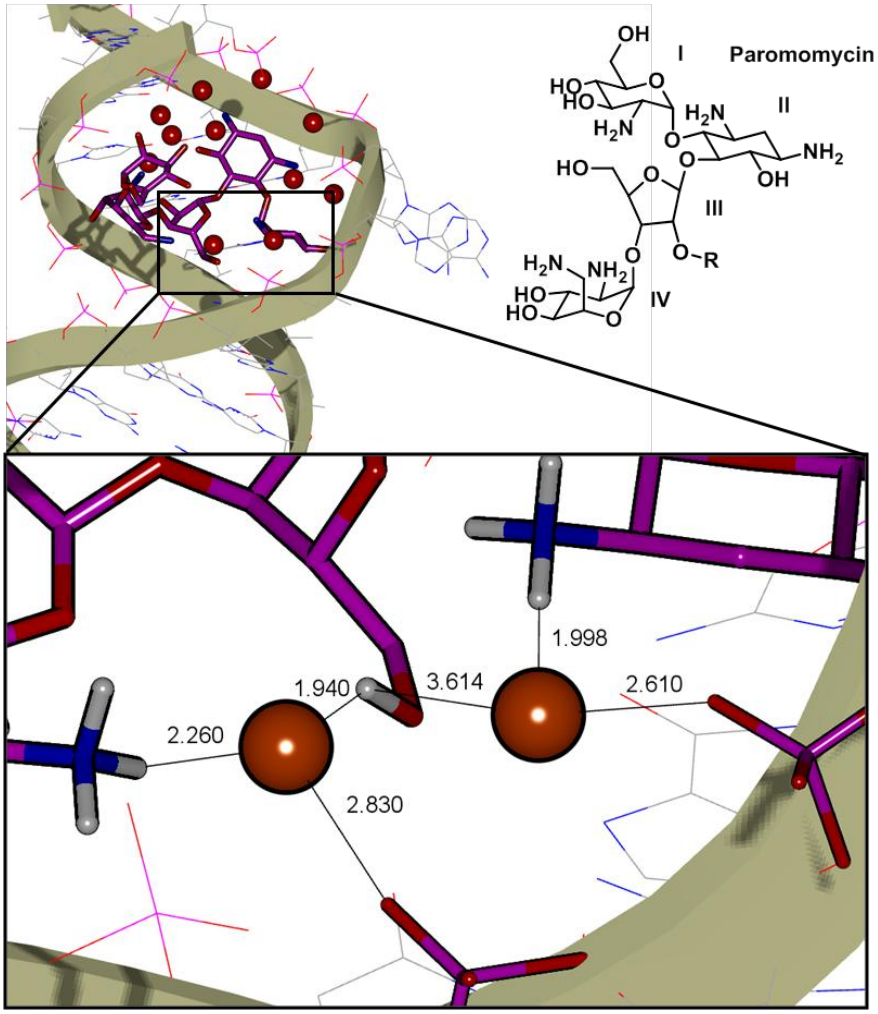


Figure 3.6. Aminoglycoside water-mediated binding to bacterial RNA.

When only one water molecule is involved in the binding, simply selecting the best scoring pose between docking with or without the water is sufficient to evaluate its necessity. In the case of aminoglycosides bound to bacterial RNA however, several waters were needed and the corresponding number of alternative selections of waters increased exponentially, rendering the docking process highly dependent on the set of molecules used. The ability to displace waters during docking would solve this issue. At this time, a study using AutoDock demonstrated that a combination of grids could model side chain motions and water displacement; we thought to expand this approach by displacing several waters simultaneously.

We developed a strategy to compute the interaction between the aminoglycosides and every non-clashing water molecule. This approach simulated the displacement of waters by aminoglycoside functional groups and led to a significantly improved accuracy (lower RMSD – root mean square deviation) in the binding mode prediction with AutoDock (Figure 3.7).

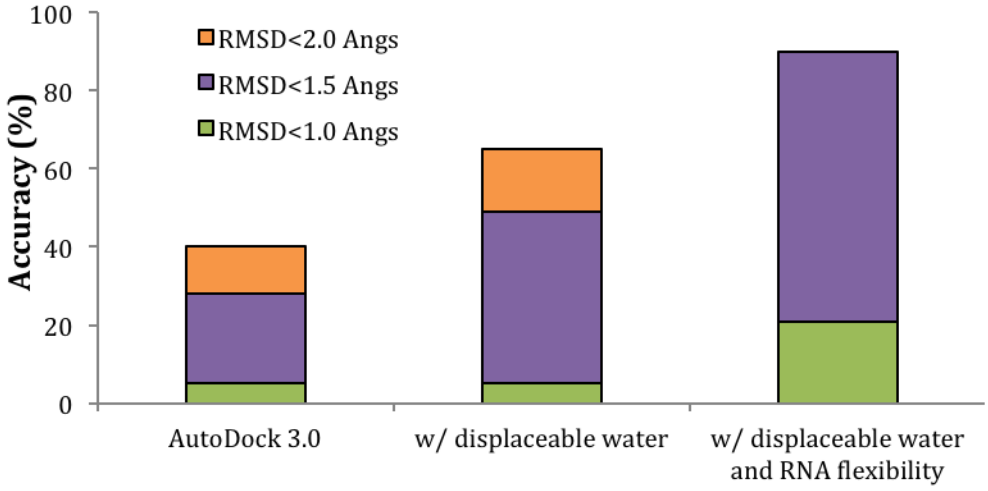


Figure 3.7. AutoDock accuracy in pose predictions with two additional implementations.

3.4 The FITTED era

3.4.1 FITTED 1.0 and 1.5 (Flexibility Induced Through Targeted Evolutionary Description) 2006-2008

At this stage, we had developed, implemented, and applied three strategies/algorithms: pharmacophore-oriented docking, docking to flexible proteins, and docking in the presence of displaceable waters; however, all three of these methods were implemented in various programs. Corbeil, Englebienne and Moitessier took up the challenge of creating a single integrated software and developed FITTED 1.0 (Figure 3.8).[5] The core of FITTED involved a GA to perform the conformational search of the ligand/protein complex, as used in the other implementations. Prior to running FITTED, the input is prepared using SMART, a tool for small molecule atom typing and determining molecular properties, and PROCESS, a tool for creating the binding site cavity and protein interaction sites.

Once the input files are prepared, FITTED creates an initial population by first generating a random ligand conformation, placing it within the binding site using a matching algorithm (and the interaction sites) while checking for clashes using binding site cavity created by PROCESS. This pose undergoes an energy minimization within the binding site and is scored using AMBER/GAFF energy, which continues until the desired population size is reached (Figure 3.8).

The initial population then undergoes evolution. Two ligand poses are selected as parents and their chromosomes undergo crossover, mutation or other operators. The best scoring individuals of the parents and children are kept for the next generation. Once the population converges, the best poses are scored using RankScore and output.

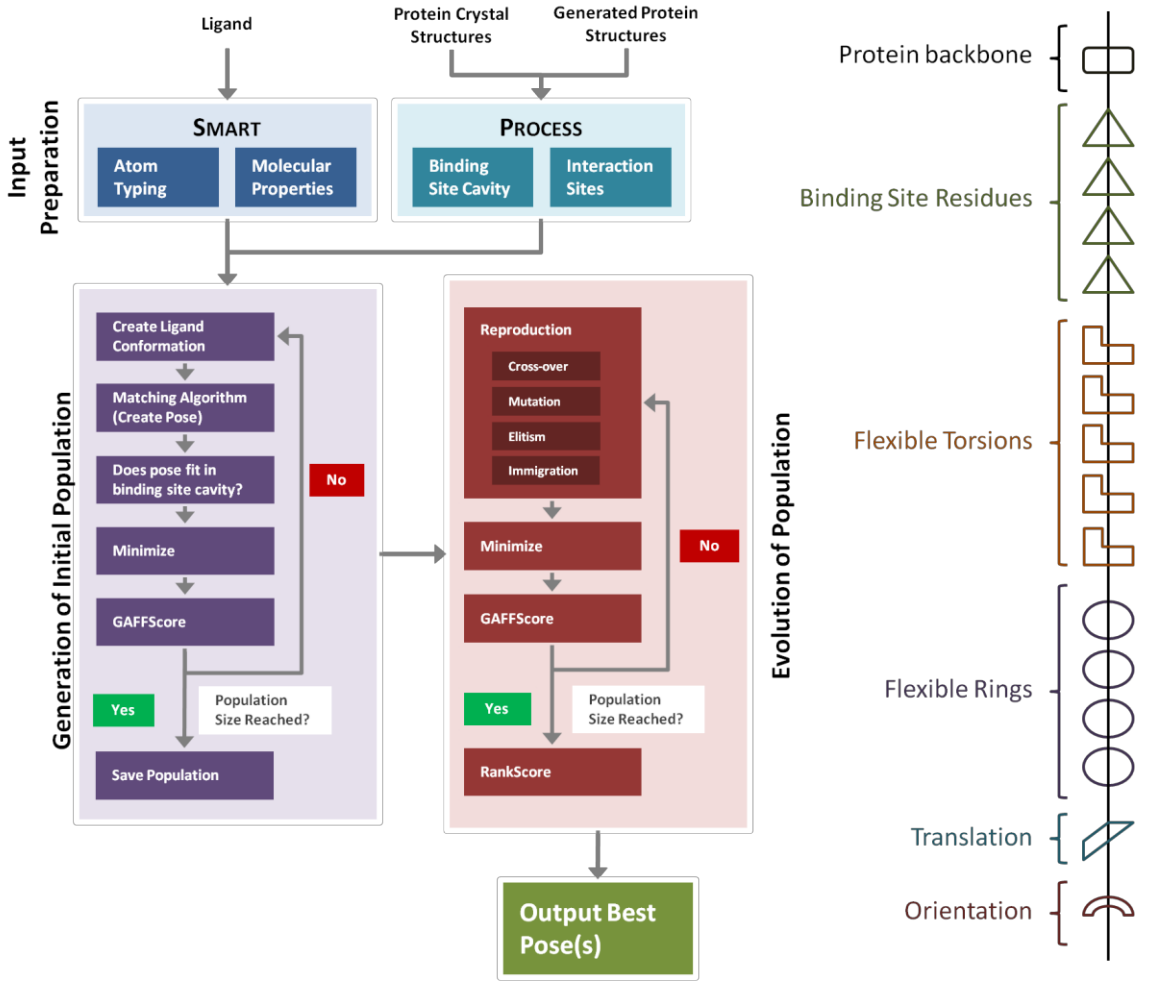


Figure 3.8. Architecture of FITTED 1.0 docking process.

A small set of protein/ligand complexes addressing both protein flexibility and displaceable waters was used for validation.[5] Within this set, FITTED achieved an increased accuracy over the rigid protein model. Figure 3.9 illustrates the success of each feature of FITTED; the correct prediction of the glutamine conformation (up vs. down), the occurrence of waters, and the correct ligand pose. Application to α-mannosidase inhibitors, through a collaboration, was a first test on the transferability of FITTED.[26]

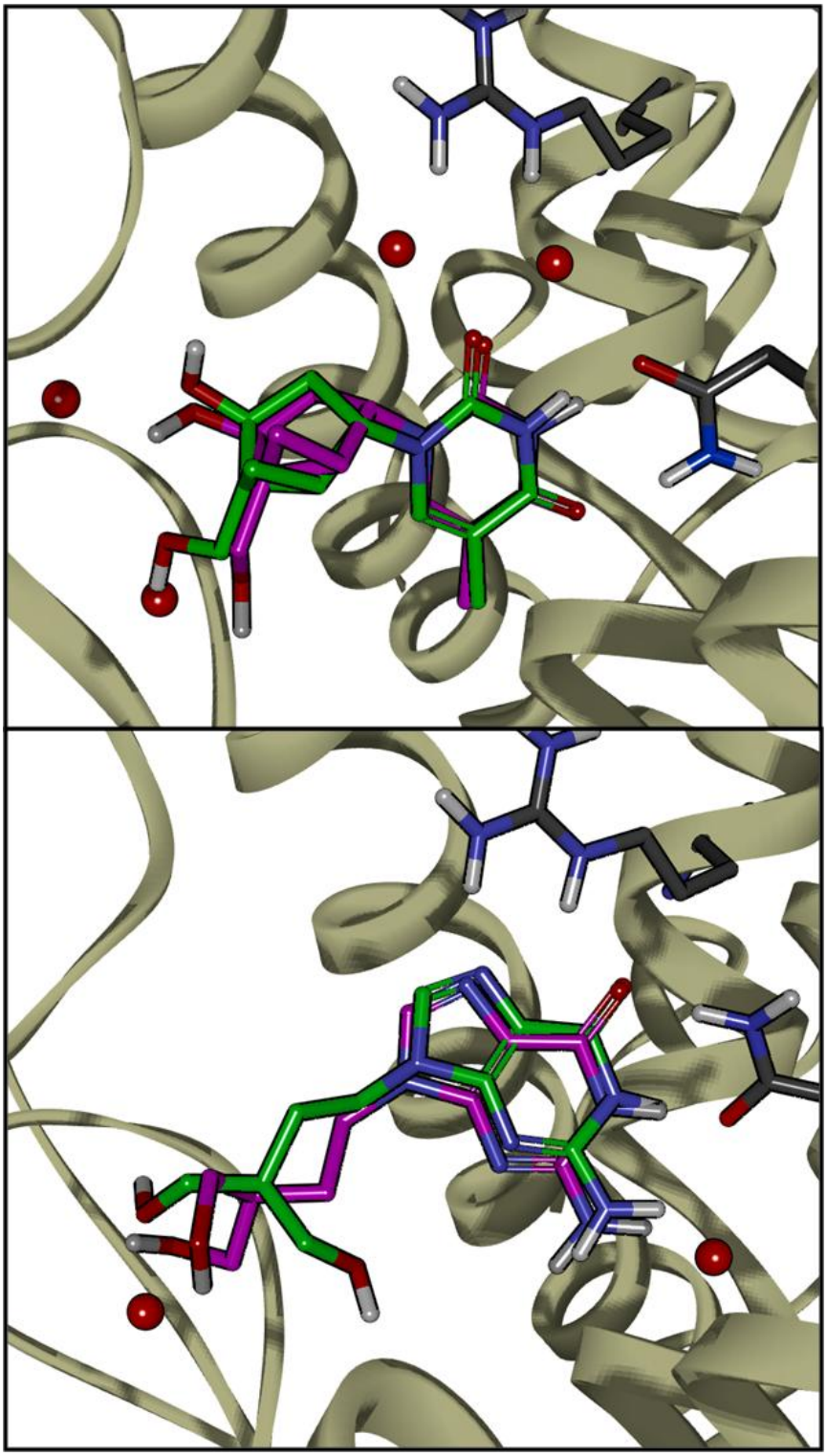


Figure 3.9. Flexible docking applied to two thymidine kinase structures (top: 1e2k and bottom: 1ki3).

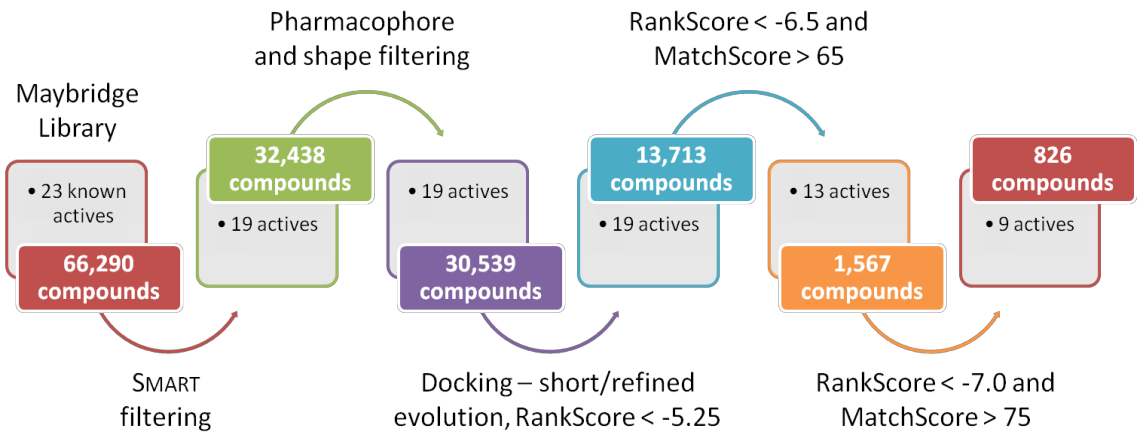


Figure 3.10. Filtering approach implemented in FITTED explained with HCV polymerase as an example.

To enable FITTED to perform a virtual screen (VS) on HCV polymerase, new tools were developed to reduce the size of the library and increase FITTED's speed (Figure 3.10). This involved filtering compounds based on known reactive/toxic groups and Lipinski's rule of five using an early version of REDUCE. The VS of the Maybridge Library identified 826 potential HCV binders and led to the identification of two micro-molar inhibitors.

We next studied the conformational changes of a kainate receptor upon binding. Depending on the type of ligand (full vs. partial agonist), the kainate receptor can adopt an open, closed or intermediate state. Docking of known agonists against an ensemble of crystallographic structures yielded accurate pose predictions and correctly identified the conformational state. Subsequent prospective docking of known partial agonists showed a preference for the closed structure, when most experimentalists thought a more open structure would be preferred. Studies using electrophysiology confirmed their preference for the closed conformation. These investigations also identified a key tyrosine that determined the open/closed state, which in turn was used to design novel binders.[27]

3.4.2 FITTED 2.6 2009

Additional features such as ring flexibility and an improved matching algorithm required FITTED to be re-evaluated on a larger set of complexes. Both self-docking

(ligands/proteins from the same crystal structure) and cross-docking (ligands from a different protein crystal) experiments were performed. This exhaustive validation revealed that protein flexibility was critical while displaceable waters were not.[28] FITTED was comparable to the best docking programs available at this time when using rigid proteins (Figure 3.11). Cross-docking accuracy increased when a flexible protein was used, with FITTED again outperforming the others.

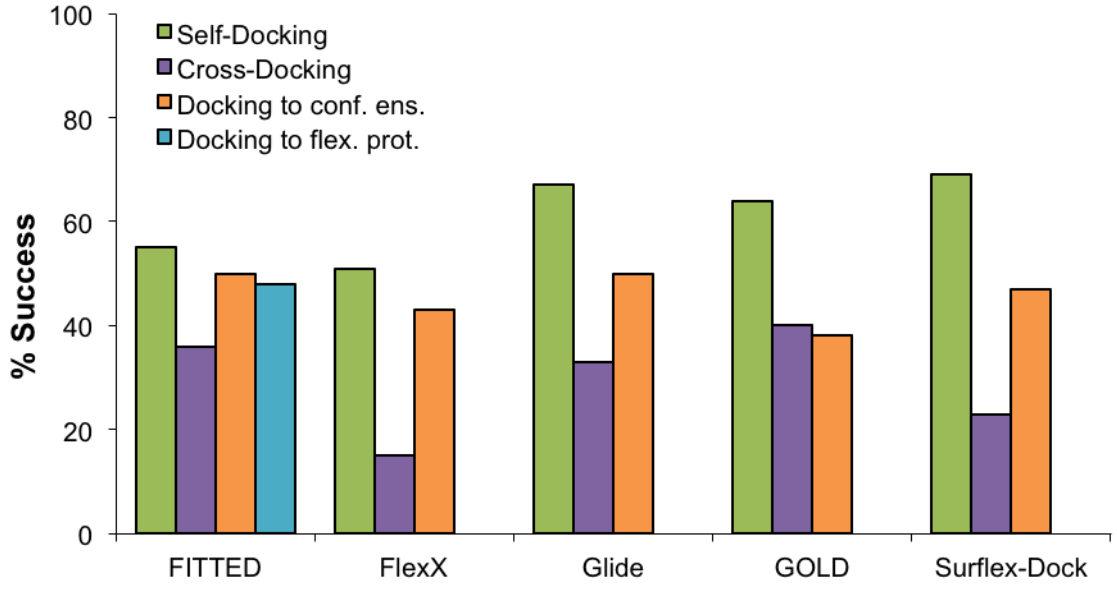


Figure 3.11. Accuracy of FITTED 2.6.

A second study was carried out to determine what role these implementations played on predicting binding affinity.[29] A panel of 18 commercially available scoring functions were applied to our own well-curated, challenging dataset of 209 protein-ligand complexes and, unsurprisingly, most of them did not perform as well as reported. As in the previous study, displaceable water did not have a large impact and should only be used when necessary. On the other hand, protein flexibility had a large impact on a few of the scoring functions, but many others performed similarly (Figure 3.12).

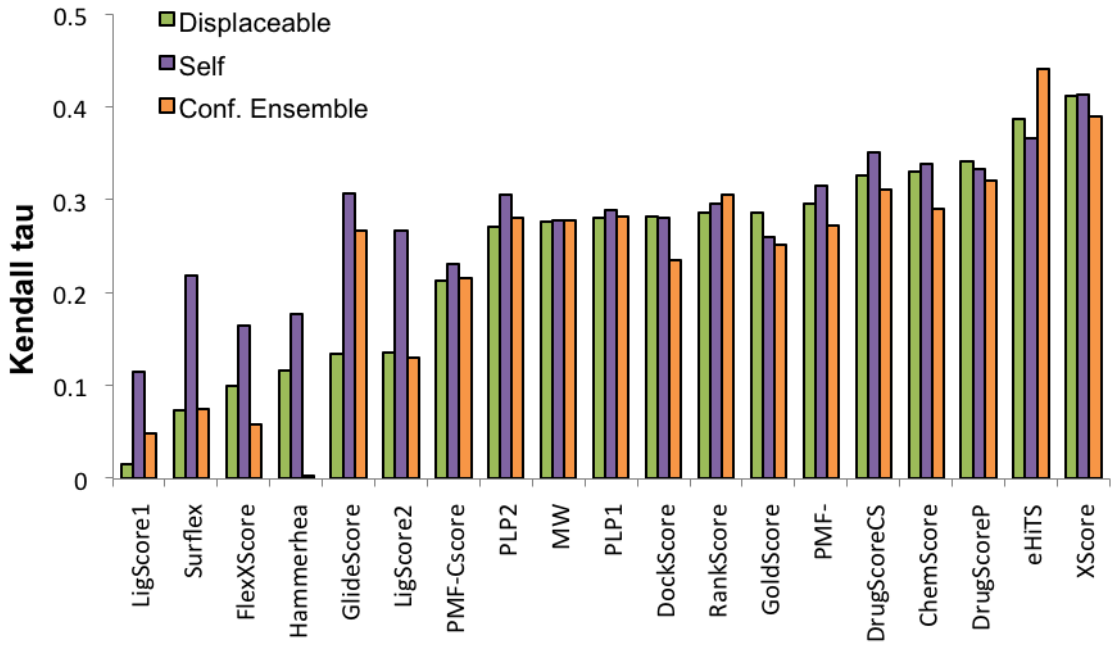


Figure 3.12. Impact of water and protein flexibility on scoring function accuracy.

This last study led to the development of two different scoring function flavors implemented into FITTED 2.6,[30] one for optimizing a ligand for affinity and another for identifying potential VS hits. While the physics remained the same in both cases, the datasets required for training are different and yielded different coefficients for the same formalism.

$$RankScore2(4) = c_1 E_{vdW} + c_2 E_{ele} + C_3 E_{HB} + c_4 \Delta G_{Solvation}$$

$$+c_5 \Delta G_{SASA} + c_6 N_{Water} + c_7 E_{rot} \qquad (\mathbf{3.2})$$

where E_{vdW}, E_{ele} and E_{HB} are the AMBER van der Waals, electrostatic, and hydrogen bond protein-ligand interaction energy, respectively; $\Delta G_{solvation}$ is the GB/SA solvation free energy, ΔG_{SASA} is the solvent accessible surface area, N_{water} is the number of bridging waters and E_{rot} is a weighted score of rotatable bond types with c_n representing fitting weights.

RankScore2 was developed for optimizing ligand affinity and tested on a set of 100 protein-ligand complexes. It was found to be amongst the most accurate scoring functions applied on this dataset.

To enable accurate identification of binders, or "hits", versus non-binders in a VS, RankScore2 was recalibrated using the DUD set,[31] a series of protein targets with known binders and decoy ligands. FITTED was used to dock the ligands and RankScore4 was trained using these poses to maximize the area under a receiver operating characteristic (AU-ROC) curve. Since RankScore4 was developed to identify actives and decoys, it clearly outperformed RankScore2.

Meanwhile, we had started a project aiming at developing G-quadruplex stabilizers as cancer therapeutics in collaboration with the Sleiman, Autexier and Mittermaier groups.[32-35] At first, three crystal structures of G-quadruplex with bound ligands were available. To better address the flexibility of the G-quadruplex and the energy upon binding, we opted for a combined docking-molecular dynamics (MD) simulation strategy using FITTED 2.6. We found that FITTED could be used to generate reasonable starting structures for refinement by MD simulations. Platinum(II) supramolecular squares[32] and platinum(II) phenanthroimidazoles[34-35] were designed. Our group is currently testing FITTED and improving the accuracy with nucleic acid binders.

3.4.3 FITTED and covalent docking 2008-2012

FITTED was modified to handle covalent docking for a separate project aimed at developing covalent prolyl oligopeptidase (POP) inhibitors.[36-38] A large number of reports on covalent drugs prompted us to enable FITTED to auto-identify reactive functional groups, create the covalent bond, and compute its energy in a single run. Thus, upon docking, if the functional group (aldehyde and boronic ester in Figure 3.13) is close enough to the binding site residue, a covalent bond is formed; otherwise, it is assumed to be non-covalent. Other programs/workflows followed our original report including covDock (Schrodinger), covalentDock (AutoDock)[39] and Dockovalent (DOCK).[40] Using this method, we have designed the POP inhibitors shown in Figure 3.14.

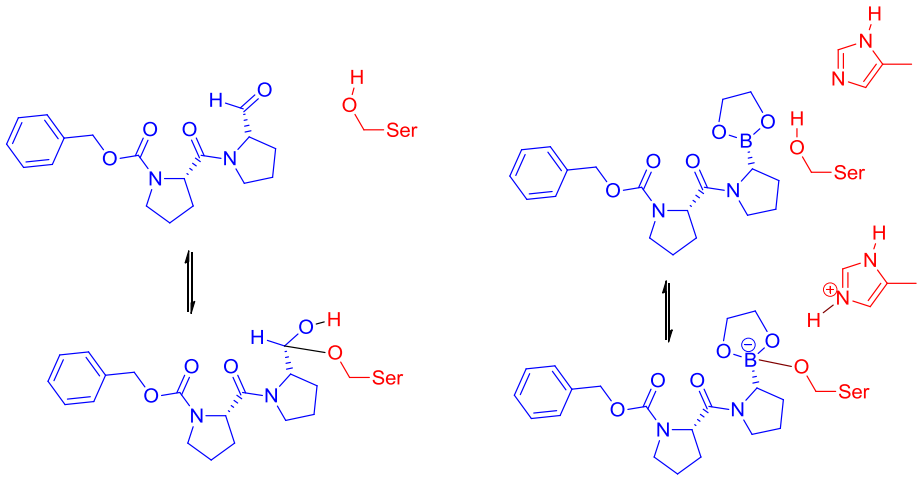


Figure 3.13. Reversible covalent inhibitors.

$t_{1/2}$ = 134 min (HLM)
Clint = 10 (μL/min/mg)

POP: IC_{50} = 200 nM
(human brain-derived endothelial cells)
IC_{50} = 1300 nM (living cells)

FAP: IC_{50} > 10 mM
DPP-IV: IC_{50} > 10 mM
DPP8: IC_{50} > 10 mM
DPP9: IC_{50} > 10 mM

$t_{1/2}$ = 31 min (HLM)
Clint = 13 (μL/min/mg)

POP: IC_{50} = 45 nM
(human brain-derived endothelial cells)
IC_{50} = 500 nM (living cells)

FAP: IC_{50} > 10 mM
DPP-IV: IC_{50} > 10 mM
DPP8: IC_{50} > 10 mM
DPP9: IC_{50} > 10 mM

Figure 3.14. Reversible covalent POP inhibitors.

3.4.4 Metabolism prediction 2011-2012

Enzyme inhibitors were the main focus of our developments until we looked at drug metabolism, specifically CYP-mediated oxidation,[41] and exploited FITTED to visualize the reactive state and provide insight into the site of metabolism (SoM) of drug candidates. Many groups have attempted SoM prediction using docking; however, our approach is

novel, as it models the TS by combining FITTED with ACE[4] (a tool for prediction of stereoselectivity). Furthermore, we applied a reactivity-based score derived from QM data resulting in a trivalent approach (reactivity, docking, TS modeling) that demonstrated greater accuracy.

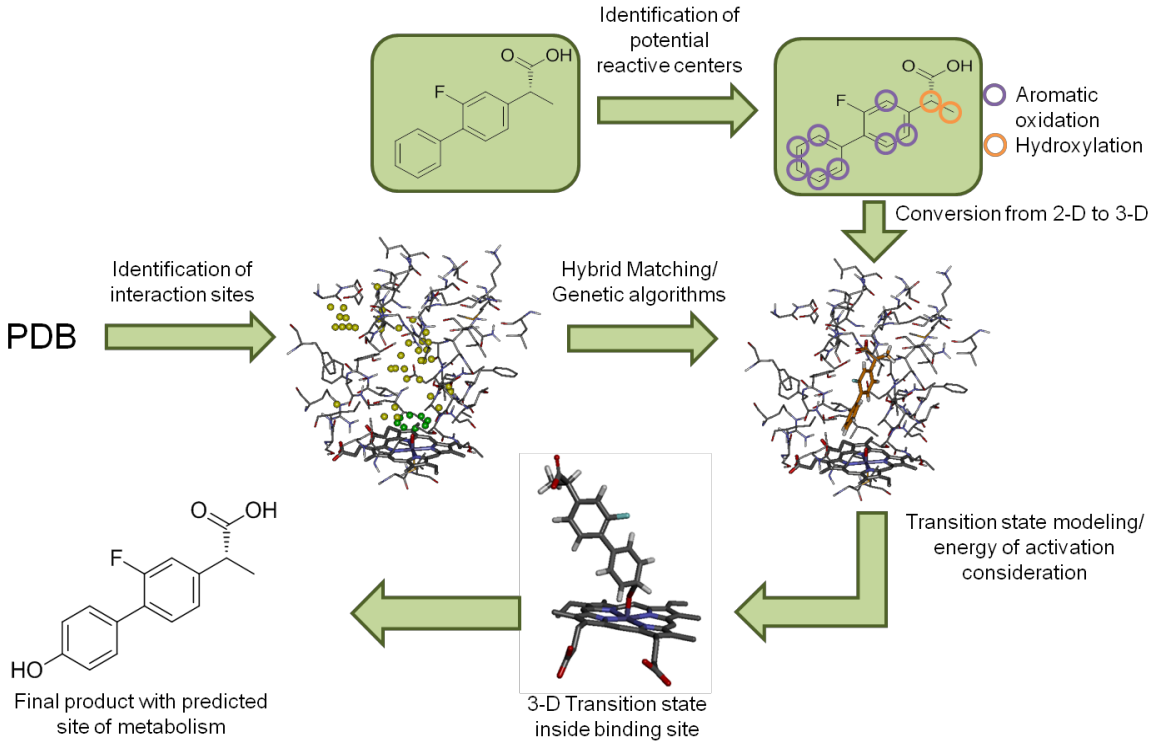


Figure 3.15. Overall approach implemented into IMPACTS.

The initial version of our SoM prediction software, IMPACTS (*In-silico* Metabolism Prediction by Activated Cytochromes and Transition States, Figure 3.15), considered multiple oxidation reaction types and focused on aromatic oxidation and hydrogen abstraction. At first, a dataset of fragments was generated and activation energies were derived by DFT to then be used as penalties within IMPACTS. Next, the P450-substrate complex is built (forcing proximity to iron-oxo), which calculates the TS structure energy. Each potential SoM is given a score (RankScore2 + activation energy), and ranked accordingly. IMPACTS was validated on a diverse dataset of substrates involving several oxidation reactions (e.g., aromatic oxidation, epoxidation, hydroxylation, sulfoxidation) with four of the five major CYP isoforms (Figure 3.16).

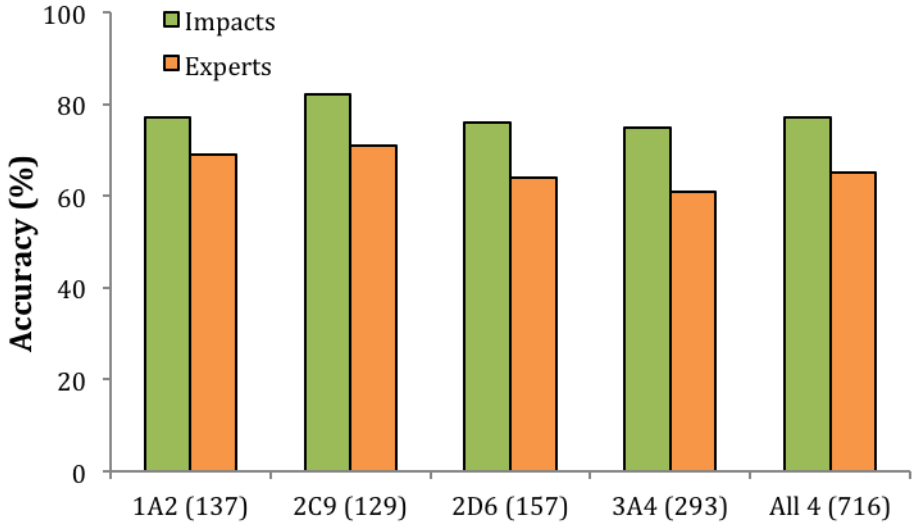


Figure 3.16. % of molecules with an observed SoM in the predicted two SoMs. Number of substrates in the set written in brackets.

As is the standard in this field, a prediction is deemed correct if one of the top-2 SoMs is experimentally observed. We decided to challenge our program by testing against experts in the medicinal chemistry field. The experts were asked to pick 2 SoMs for 716 molecules, and IMPACTS consistently had them outperformed by 6-7%. Although we recognize that this is not a significant representation of the field, it does offer some insight into the place for such software in both academia and industry.

Currently, the success and the abilities of IMPACTS have led to projects investigating CYP inhibition, computer-engineered biocatalysis, and metabolite prediction, the follow-up step to SoM identification.

3.4.5 FITTED 3.1 and metal coordination 2013-2014

In collaboration with the Gleason group, we looked at histone deacetylase (HDAC) inhibitors.[42] HDACs classes I/II are zinc metalloenzymes with inhibitors coordinating the zinc. Often, this interaction is modeled as a purely electrostatic or van der Waals interaction excluding the covalent nature of metal coordination. GOLD and FlexX, like FITTED, use coordination sites to guide the geometry, however metal interaction scoring required improvement. Additionally, in the close proximity to the zinc, a histidine, or

more generally, a basic residue (glutamate in MMPs), is typically present and may bind an acidic proton (Figure 3.17). Accurately modeling these features to predict the binding mode required additional implementations such as specific zinc-coordination and hydrogen bond equations. A bonding consideration is necessary if the ligand is in proximity to the zinc ion, while non-bonding would be more accurate at longer distances. Similarly, the proton shift is modeled with two static representations: if the ligand is in proximity to the zinc, the proton exists on the basic residue; otherwise, it is on the ligand. We thus developed new energy functions to represent the special hydrogen bond, as well as the metal coordination, and implemented them in FITTED.

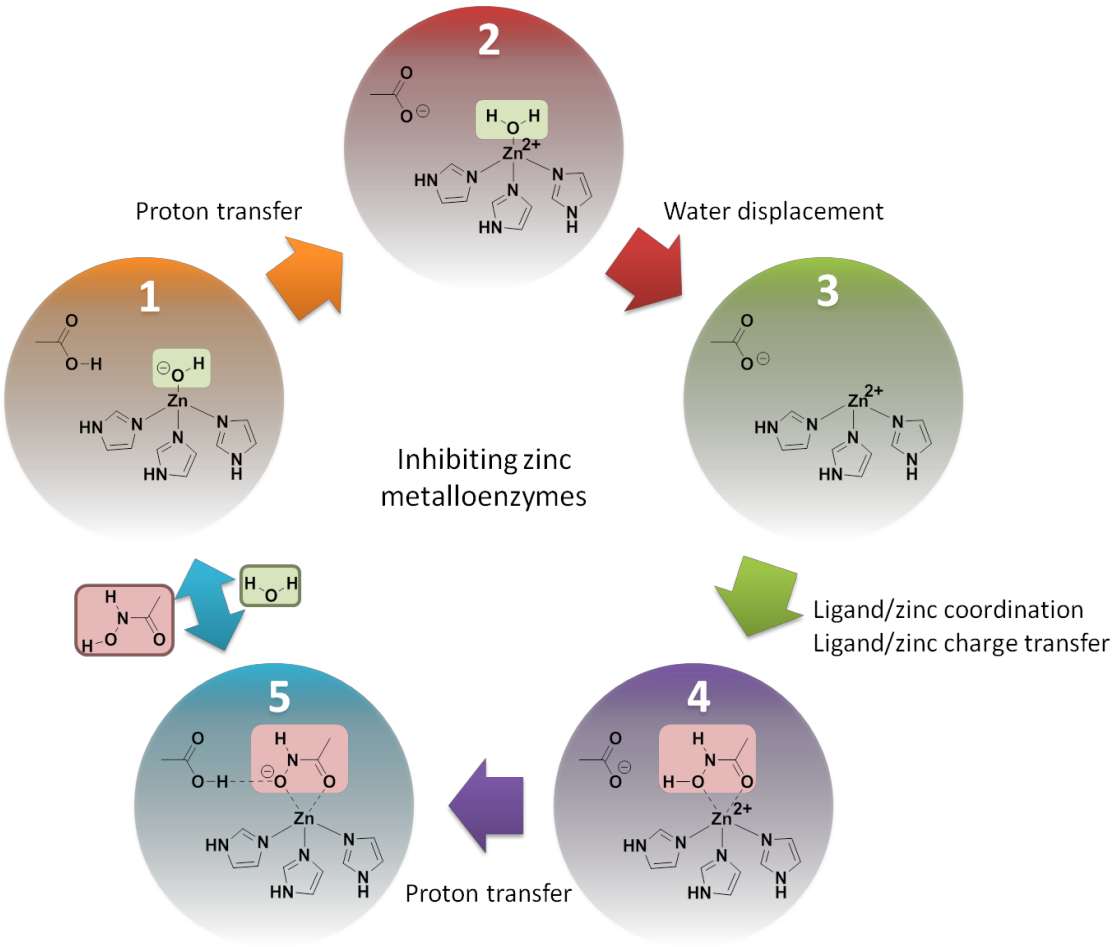


Figure 3.17. Binding process to metalloenzymes.

As validation, 121 complexes were selected for self-docking experiments. The new results were compared with the two previous FITTED implementations (Figure 3.18).

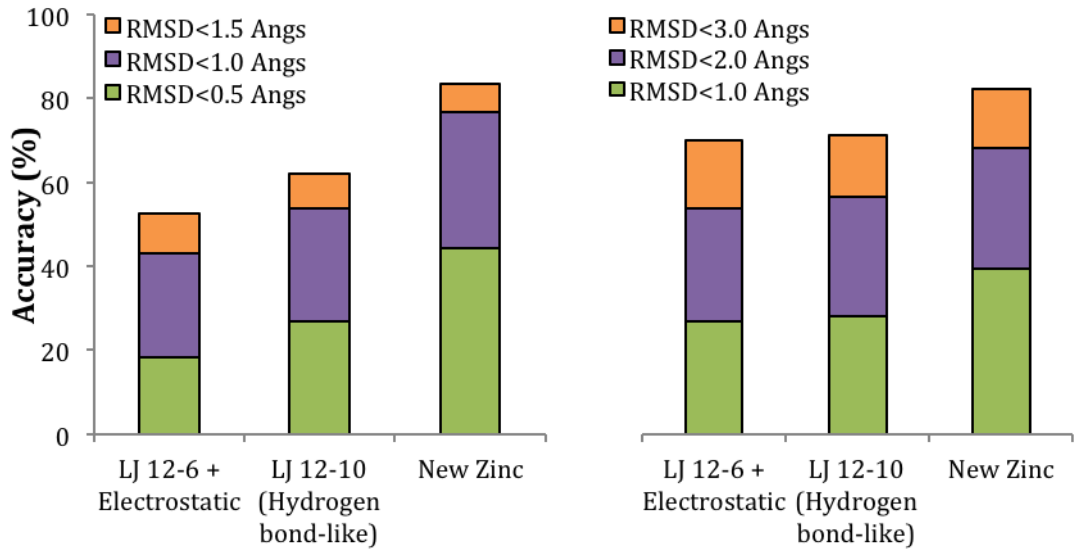


Figure 3.18. Pose prediction accuracy. Left: accuracy of the zinc coordination geometry; right: accuracy of the pose prediction.

These implementations improved the overall pose prediction and zinc binding accuracy. Additionally, a virtual screening of the DUD-e set of metalloenzyme inhibitors yielded an average AU-ROC of 0.869. For comparison, the original FITTED metal considerations led to average AU-ROCs of 0.639 and 0.510. Others followed up our efforts docking to metalloenzymes, such as AutoDock4$_{Zn}$.[43] Application to the design and preparation of HDACi's followed (unreported data).

3.4.6 FITTED and drug discovery 2013-2014

Different models for waters can be used in docking experiments. For example, waters may be represented as spherical particles that are both hydrogen bond donors and acceptors, removing their orientation dependence. We thought that such symmetrical waters, as opposed to those with a directional dipole moment, would result in more reliable docked poses. The resulting FITTED 3.6 was tested exhaustively using the DUD set and the impacts of protein flexibility, zinc-binding, the conserved waters, and the inclusion of water particles were fully evaluated.[44]

The DUD set contained a single crystal structure per target protein. In order to mimic flexibility, conformational ensembles were built using three approaches: experimental

structures, simulated conformations generated, and by allowing side chain flexibility. From the analysis of docking to rigid proteins, it became clear that some targets are more dependent on the protein structure selected for docking than others. Taking the examples of COX-2 and FGFr1 (Figure 3.19), the measured AU-ROC ranged from very poor to high depending on the crystal structure of the same protein. A closer look at these proteins revealed that some of the conformations used may hinder access to the binding site or favor binding of only a certain chemical series, leading to poor scores for active compounds. Including protein flexibility addressed this issue in 1/3 of cases. We therefore suggested that when the protein flexibility is unknown and/or when a diverse set of ligands is studied, including protein flexibility should increase accuracy; however, when the ligands are similar, it is best to use rigid docking as these ligands are expected to bind to a very similar protein conformation. As in the previous studies, using displaceable waters had little effect, or in some cases, hindered the results.

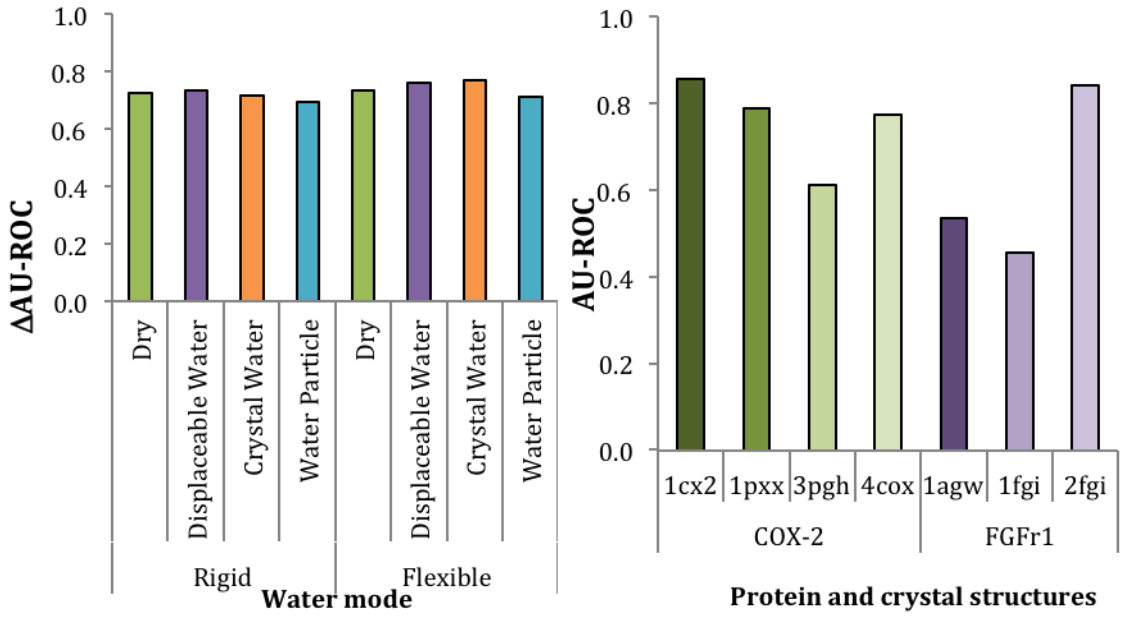


Figure 3.19. VS accuracy dependent on water mode (left) and protein structure (right).

3.4.7 Integrating computational and medicinal chemistry - the FORECASTER platform

To improve the user-experience, FORECASTER, a web-based interface, was developed to allow the user to build drug design workflows (Figure 3.20),[45] eliminating the need to run multiple command-line applications to perform common drug design tasks. As a validation, FORECASTER was used to build virtual combinatorial libraries, filter, and extract a highly diverse library from the NCI database. These focused libraries were then docked to the estrogen receptor (ER); accurately identifying existing ER modulators demonstrated its usability and accuracy.

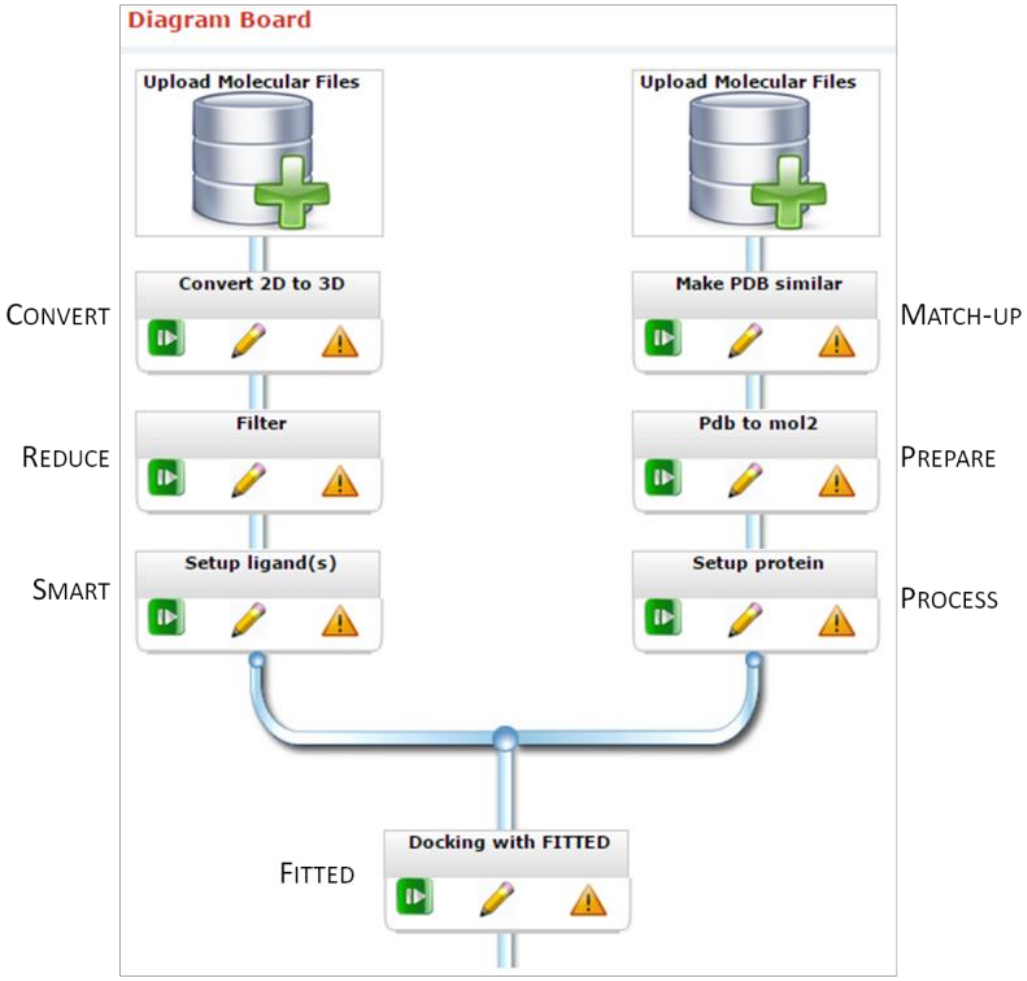


Figure 3.20. Sample workflow on the FORECASTER platform using some of the available tools.

3.5 Conclusion and Perspective

Over the years, we have reported a docking program foundation, FITTED, initially developed for enzyme-inhibitor binding mode prediction. Subsequently, we have expanded its scope for application to VS, metalloenzymes, nucleic acids, covalent inhibitors, and SoM prediction. All of these implementations contributed to successful drug discovery research programs (G-quadruplex binders, POP, HDAC, HCV polymerase and MMP inhibitors, GluK2 antagonists, and PET imaging agents[46]). Over the years, our studies[44] have shown that protein flexibility improves pose prediction and active compound discovery; however, more work is still needed to properly consider the binding free energy of waters. Recently, the first independent comparative study using 6 proteins revealed that FITTED (specifically RankScore) outperformed 15 other scoring functions.[47] This independent success story, along with over 150 licensed academic and industrial groups, demonstrates that FITTED, along with the FORECASTER platform, is a tool medicinal chemists can use to aid in a wide variety of drug discovery challenges with meaningful results.

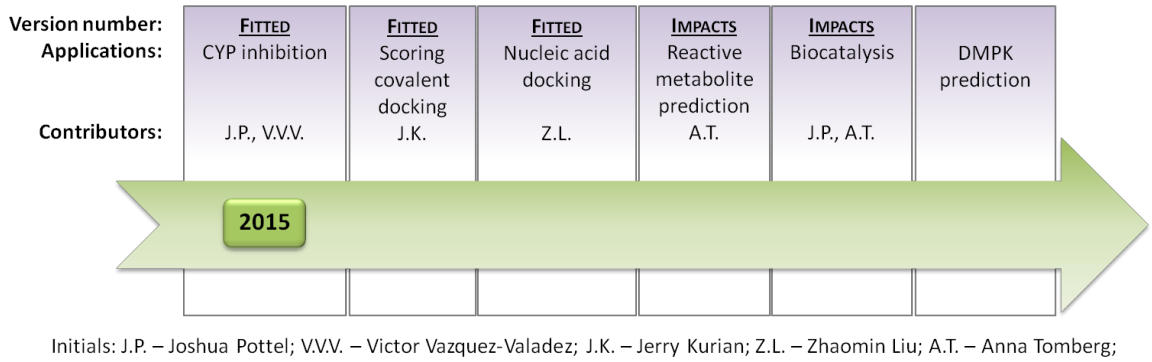


Figure 3.21. Roadmap for the development of FITTED.

While we have been developing and expanding the scope of FITTED (Figure 3.21), the application of docking methods has evolved in many labs from structure-based drug design to metabolism prediction, off-target identification, and more. In principle, any biological process encountering a binding process could be investigated using docking methods. In this context, CYP inhibition, p-gp efflux, and plasma protein binding prediction could be attainable milestones. In the near future, docking methods may be employed to predict several drug properties prior to any experiments.

3.6 Acknowledgment

We acknowledge NSERC (Discovery Grants to NM), CIHR (Operating Grants to NM, Drug Discovery Training program scholarships to JP, ZL and AT), and FRQ-NT (scholarships to JP), as well as industrial partners ViroChem Pharma Montréal, and AstraZeneca R&D Montréal, for financial support and fruitful discussions. We also thank the researchers who contributed to the developments and applications of FITTED over the years.

3.7 References

1. Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Prot*. *Struct*. *Funct*. *Genet*. **1997,** *29* (SUPPL. 1), 215-220

2. Rarey, M.; Kramer, B.; Lengauer, T., The particle concept: Placing discrete water molecules during protein- ligand docking predictions. *Prot*. *Struct*. *Funct*. *Genet*. **1999,** *34* (1), 17-28.

3. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go. *Br*. *J*. *Pharmacol*. **2008,** *153* (SUPPL. 1), S7-S26.

4. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development and application of a rapid and accurate program based on organic principles. *Angew*. *Chem*. *Int*. *Ed*. **2008,** *47* (14), 2635-2638.

5. Corbeil, C. R.; Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J*. *Chem*. *Inf*. *Model*. **2007,** *47* (2), 435-449

6. Jain, A. N., Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J*. *Comp*.*-Aided Mol*. *Des*. **2007,** *21* (5), 281-306.

7. Verdonk, M. L.; Chessari, G.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R., Modeling water molecules in protein-ligand docking using GOLD. *J*. *Med*. *Chem*. **2005,** *48* (20), 6504-6515

8. Morris, G. M.; Ruth, H.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J*. *Comp*. *Chem*. **2009,** *30* (16), 2785-2791.

9. *Dock, http://dock.compbio.ucsf.edu/DOCK_6/index.htm*, 6.7; UCSF: San Francisco, CA, 2015.

10. Corbeil, C. R.; Therrien, E.; Moitessier, N., Modeling reality for optimal docking of small molecules to biological targets. *Curr*. *Comput*.*-Aided Drug Des*. **2009,** *5* (4), 241-263.

11. Hanessian, S.; Moitessier, N.; Therrien, E., A comparative docking study and the design of potentially selective MMP inhibitors. *J*. *Comp*.*-Aided Mol*. *Des*. **2001,** *15* (10), 873-881.

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

13. Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D., DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J*. *Comp*.*-Aided Mol*. *Des*. **2001,** *15* (5), 411-428.

14. Hanessian, S.; Moitessier, N.; Wilmouth, S., Tetrahydrofuran as a scaffold for peptidomimetics. Application to the design and synthesis of conformationally constrained metalloproteinase inhibitors. *Tetrahedron* **2000,** *56* (39), 7643-7660.

15. Hanessian, S.; Moitessier, N.; Gauchet, C.; Viau, M., N-aryl sulfonyl homocysteine hydroxamate inhibitors of matrix metalloproteinases: Further probing of the S1, S1', and S2' pockets. *J*. *Med*. *Chem*. **2001,** *44* (19), 3066

16. Hanessian, S.; Moitessier, N.; Cantin, L. D., Design and synthesis of MMP inhibitors using N-arylsulfonylaziridine hydroxamic acids as constrained scaffolds. *Tetrahedron* **2001,** *57* (32), 6885.

17. Hanessian, S.; MacKay, D. B.; Moitessier, N., Design and synthesis of matrix metalloproteinase inhibitors guided by molecular modeling. Picking the S1 pocket using conformationally constrained inhibitors. *J*. *Med*. *Chem*. **2001,** *44* (19), 3074

18. Moitessier, N.; Henry, C.; Chapleur, Y.; Maigret, B., Combining pharmacophore search, automated docking, and molecular dynamics simulations as a novel strategy for flexible docking. Proof of concept: Docking of arginine-glycine-aspartic acid-like compounds into the alpha(v)beta(3) binding site. *J*. *Med*. *Chem*. **2004,** *47* (17), 4178

19. Polgár, T.; Keseru, G. M., Structure-based β-secretase (BACE1) inhibitors. *Curr*. *Pharm*. *Des*. **2014,** *20* (20), 3373-3379.

20. Cosconati, S.; Marinelli, L.; Di Leva, F. S.; La Pietra, V.; De Simone, A.; Mancini, F.; Andrisano, V.; Novellino, E.; Goodsell, D. S.; Olson, A. J., Protein flexibility in virtual screening: The BACE-1 case study. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (10), 2697-2704.

21. Moitessier, N.; Therrien, E.; Hanessian, S., A method for induced-fit docking, scoring, and ranking of flexible ligands. Application to peptidic and pseudopeptidic β-secretase (BACE 1) inhibitors. *J*. *Med*. *Chem*. **2006,** *49* (20), 5885-5894.

22. Moitessier, N.; Henry, C.; Len, C.; Chapleur, Y., Toward a Computational Tool Predicting the Stereochemical Outcome of Asymmetric Reactions. 1. Application to Sharpless Asymmetric Dihydroxylation. *J*. *Org*. *Chem*. **2002,** *67* (21), 7275-7282.

23. Hanessian, S.; Tremblay, M.; Kornienko, A.; Moitessier, N., Design, modeling and synthesis of functionalized paromamine analogs. *Tetrahedron* **2001,** *57* (16), 3255-3265.

24. Moitessier, N.; Westhof, E.; Hanessian, S., Docking of aminoglycosides to hydrated and flexible RNA. *J*. *Med*. *Chem*. **2006,** *49* (3), 1023-1033

25. Detering, C.; Varani, G., Validation of automated docking programs for docking and database screening against RNA drug targets. *J*. *Med*. *Chem*. **2004,** *47* (17), 4188-4201.

26. Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N., Evaluation of Docking Programs for Predicting Binding of Golgi alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Prot*. *Struct*. *Funct*. *Bioinf*. **2007,** *69* (1), 160-176

27. Schiavini, P.; Dawe, G. B.; Bowie, D.; Moitessier, N., Discovery of novel small-molecule antagonists for GluK2. *Bioorg*. *Med*. *Chem*. *Lett*. **2015**.

28. Corbeil, C. R.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (4), 997-1009

29. Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (6), 1568-1580

30. Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 5. Force-field-based prediction of binding affinities of ligands to proteins. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (11), 2564-2571

31. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J*. *Med*. *Chem*. **2006,** *49* (23), 6789-6801

32. Kieltyka, R.; Englebienne, P.; Fakhoury, J.; Autexier, C.; Moitessier, N.; Sleiman, H. F., A platinum supramolecular square as an effective G-quadruplex binder and telomerase inhibitor. *J*. *Am*. *Chem*. *Soc*. **2008,** *130* (31), 10040-10041

33. Kieltyka, R.; Fakhoury, J.; Moitessier, N.; Sleiman, H. F., Platinum phenanthroimidazole complexes as G-quadruplex DNA selective binders. *Chem*. *Eur*. *J*. **2008,** *14* (4), 1145-1154.

34. Castor, K. J.; Mancini, J.; Fakhoury, J.; Weill, N.; Kieltyka, R.; Englebienne, P.; Avakyan, N.; Mittermaier, A.; Autexier, C.; Moitessier, N.; Sleiman, H. F.,

Platinum(II) phenanthroimidazoles for targeting telomeric G-quadruplexes. *ChemMedChem* **2012,** *7* (1), 85-94.

35. Castor, K. J.; Liu, Z.; Fakhoury, J.; Hancock, M. A.; Mittermaier, A.; Moitessier, N.; Sleiman, H. F., A platinum(ii) phenylphenanthroimidazole with an extended side-chain exhibits slow dissociation from a c-kit G-quadruplex motif. *Chem*. *Eur*. *J*. **2013,** *19* (52), 17836-17845.

36. Lawandi, J.; Toumieux, S.; Seyer, V.; Campbell, P.; Thielges, S.; Juillerat-Jeanneret, L.; Moitessier, N., Constrained Peptidomimetics Reveal Detailed Geometric Requirements of Covalent Prolyl Oligopeptidase Inhibitors. *J*. *Med*. *Chem*. **2009,** *52*, 6672-6684

37. Lawandi, J.; Gerber-Lemaire, S.; Juillerat-Jeanneret, L.; Moitessier, N., Inhibitors of prolyl oligopeptidases for the therapy of human diseases: Defining diseases and inhibitors. *J*. *Med*. *Chem*. **2010,** *53* (9), 3423-3438

38. De Cesco, S.; Deslandes, S.; Therrien, E.; Levan, D.; Cueto, M.; Schmidt, R.; Cantin, L. D.; Mittermaier, A.; Juillerat-Jeanneret, L.; Moitessier, N., Virtual screening and computational optimization for the discovery of covalent prolyl oligopeptidase inhibitors with activity in human cells. *J*. *Med*. *Chem*. **2012,** *55* (14), 6306-6315.

39. Ouyang, X.; Zhou, S.; Su, C. T. T.; Ge, Z.; Li, R.; Kwoh, C. K., CovalentDock: Automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry constraints. *J*. *Comp*. *Chem*. **2013,** *34* (4), 326-336.

40. London, N.; Miller, R. M.; Krishnan, S.; Uchida, K.; Irwin, J. J.; Eidam, O.; Gibold, L.; Cimermančič, P.; Bonnet, R.; Shoichet, B. K.; Taunton, J., Covalent docking of large libraries for the discovery of chemical probes. *Nat*. *Chem*. *Biol*. **2014,** *10* (12), 1066-1072.

41. Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.-D.; Moitessier, N., Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (9), 2471-2483.

42. Pottel, J.; Therrien, E.; Gleason, J. L.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 6. Development and application to the docking of

HDACs and other zinc metalloenzymes inhibitors. *J*. *Chem*. *Inf*. *Model*. **2014,** *54* (1), 254-265.

43. Santos-Martins, D.; Forli, S.; Ramos, M. J.; Olson, A. J., AutoDock4Zn: An Improved AutoDock Force Field for Small-Molecule Docking to Zinc Metalloproteins. *J*. *Chem*. *Inf*. *Model*. **2014,** *54* (8), 2371-2379.

44. Therrien, E.; Weill, N.; Tomberg, A.; Corbeil, C. R.; Lee, D.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 7. Impact of protein flexibility and water molecules on docking-based virtual screening accuracy. *J*. *Chem*. *Inf*. *Model*. **2014,** *54* (11), 3198-3210.

45. Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N., Integrating medicinal chemistry, organic/combinatorial chemistry, and computational chemistry for the discovery of selective estrogen receptor modulatorswith FORECASTER, a novel platform for drug discovery. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (1), 210-224.

46. Bernard-Gauthier, V.; Aliaga, A.; Aliaga, A.; Boudjemeline, M.; Hopewell, R.; Kostikov, A.; Rosa-Neto, P.; Thiel, A.; Schirrmacher, R., *ACS Chem*. *Neurosci*. **2015,** *6*, 260−276.

47. Xu, W.; Lucke, A. J.; Fairlie, D. P., Comparing sixteen scoring functions for predicting biological activities of ligands for protein targets. *J*. *Mol*. *Graph*. *Model*. **2015,** *57*, 76-88.

Introduction to Chapter 4

Docking to metalloenzymes, zinc metalloenzymes specifically, was not a priority during the original development of FITTED. The accuracy of the results was often similar to those of other docking programs however never stellar. The poor accuracy was often attributed to the drawbacks of molecular mechanics tools in modeling metals and d-orbital containing elements. For instance, point-charges were seen as limiting and insufficient when considering the diffuse electron clouds ascribed to transition-metal elements and their geometries. In the context of collaboration with the Gleason group (McGill University), FITTED was used to investigate proper zinc-binding with the goal of providing support for drug discovery and histone deacetylase inhibitor docking. This chapter presents the development of a method modeling a crucial proton-shuttle mechanism found within all zinc metalloenzymes and the measurements supporting the experimental reports of its existence. Additionally, the integration of a new energy function within FITTED is described and demonstrates the improved accuracy of docking to these enzymes.

(This page was left blank intentionally)

Chapter 4:

Docking Ligands into Flexible and Solvated Macromolecules. 6. Development and Application to the Docking of HDACs and other Zinc Metalloenzymes Inhibitors

This chapter is reprinted with permission from: "Docking Ligands into Flexible and Solvated Macromolecules. 6. Development and Application to the Docking of HDACs and other Zinc Metalloenzymes Inhibitors", Pottel, J. ; Therrien, E.; Gleason, J. L.; Moitessier, N.; *Journal of Chemical Information and Modelling*, **2014**, 54(1), 254-265.

Author Contributions: **Eric Therrien** initiated the idea of having docking to metalloenzymes within FITTED with **James L. Gleason** as part of a team grant and they contributed to the knowledge of using the FORECASTER/FITTED suites of programs and the knowledge of HDACs and their inhibitors respectively. All coding and application experiments were contributions of the author of this thesis.

4.1 Abstract

Metalloenzymes are ubiquitous proteins which feature one or more metal ions either directly involved in the enzymatic activity and/or structural properties (i.e., zinc fingers). Several members of this class take advantage of the Lewis acidic properties of zinc ions to carry out their various catalytic transformations including isomerization or amide cleavage. These enzymes have been validated as drug targets for a number of diseases including cancer however, despite their pharmaceutical relevance and the availability of crystal structures, structure-based drug design methods have been poorly and indirectly parameterized for these classes of enzymes. More specifically, the metal coordination component of the process of drugs binding to metalloenzymes has been inadequately predicted by current docking programs, if at all. In addition, several known issues, such as coordination geometry, atomic charge variability and a potential proton transfer from small molecules to a neighboring basic residue, have often been ignored. We report

herein the development of specific functions and parameters to account for zinc-drug coordination focusing on the above-listed phenomena and their impact on docking to zinc metalloenzymes. These atom-type dependent but atomic charge-independent functions enable the simulation of drug binding to metalloenzymes, considering an acid-base reaction with a neighboring residue when necessary with good accuracy.

4.2 Introduction

Over the years, several zinc metalloenzymes have been validated as drug targets. Among these metalloproteins are the matrix metalloproteinases (MMPs) which represent a family of zinc endopeptidases including stromelysin-1 (MMP-3), gelatinases (e.g., MMP-2) and collagenases (e.g. MMP-1 and MMP-9), the carbonic anhydrases, the mannosidases (e.g., α-mannosidase), β-lactamase, phospholipase C, alcohol dihydrogenase and the histone deacetylases (e.g., HDAC-8). As with several other enzymes, docking methods have been tested as tools for inhibitor design and discovery. Our interests in MMP[1-3] and α-mannosidase[4] inhibitors led us to further evaluate docking programs for the design of these classes of inhibitors.

Over the past few decades, docking methods have evolved from simple rigid body assembling tools (rigid body docking) to methods modeling flexible ligand/flexible protein complexes.[5] Our efforts in the field led to the development of a docking program, FITTED (Flexibility Induced Through Targeted Evolutionary Description), which accounts for ligand and protein flexibility as well as for the presence of displaceable water molecules.[6-7] In order to cover a large range of drug classes, further implementations provided FITTED with the ability to model the binding of covalent drugs.[8] In 2007, the first version of FITTED was tested for its ability to dock mannosidase inhibitors which revealed the intricacies of modeling zinc coordination.[4] Later on, we also demonstrated that scoring metalloenzyme inhibitors was poor.[9] Throughout the years, docking programs have been assessed for their ability to dock inhibitors coordinating metal ions (often zinc) of metalloenzymes: AutoDock, DOCK, GOLD and FlexX being widely used.[10-15] Within most of these studies, the metal coordination has been modeled through simple electrostatic and van der Waals interactions and no treatment of the covalent

nature of metal coordination was considered. In addition, the zinc ion atomic charge has often been assumed to be +2 while more advanced molecular dynamics simulations have often required finely tuned charging schemes.[10] Although these approximations provided reasonable overall binding modes and even good enrichments in some studies,[16] scoring the metal coordination is expected to be poorly predictive and more advanced scoring is required, in particular scoring the displacement of water molecules coordinating the zinc ion in the unbound state might be critical in order to identify poor zinc binding groups.[12] In an attempt to account for the covalent nature of zinc coordination, hydrogen bond-like terms have also been evaluated (e.g., GOLD) however this is merely circumventing the problem; a metal coordination is neither a covalent bond nor a hydrogen bond. Despite these limitations, docking-based virtual screening has been reasonably successful but can certainly be improved.[17]

Our recent interest in developing HDAC inhibitor hybrids[18-20] revived our curiosity in improving the ability to dock to metalloenzymes. Here we report our efforts to further develop our docking program in order to accurately model zinc coordination and more specifically to discover novel HDAC inhibitors.

4.3 Theory and Current State

4.3.1 Metalloenzymes and classical molecular mechanics

When simulating the metal-ligand coordination using classical molecular mechanics, two models have been proposed differing in the nature of the coordination bond: the bound model and the non-bound model. First, the bound model, in which the zinc-ligand bond is considered covalent, enables the coordination geometry to be considered with the *normal* bonded terms (bond, angles, and torsions). In contrast, the non-bound model relies on non-directional interactions (electrostatic, van der Waals) and is thus expected to provide less reliable geometries but eliminate the potentially overly-strict constraints placed on a typical covalent interaction such as, for example, a carbon-carbon bond. Docking methods cannot easily rely on the bound model as, following the positioning and orientation of the ligand, the zinc binding group may or may not coordinate the zinc ion. If it is not coordinated, the bound energy must be ignored. In this context, covalent

docking methods exist, in which both covalent and non-covalent poses can be considered and could be extended to metal coordination.[8] When either model is used, the charge transfer between the zinc ion and the coordinating groups varies as a function of the coordination geometry and of the zinc binding group nature; consequently, it remains difficult to simulate the binding process with the traditional point charges and thus, polarizable force fields have been envisioned.[21] The change in pK_a of the zinc binding group, a potential proton transfer to a neighboring basic residue and the presence of water are also to be considered for accurate docking of zinc binding molecules.

4.3.2 Docking to metalloenzymes, coordination geometry, proton and charge transfers and displacement of water molecules

Accurately docking small molecules to metalloenzymes requires consideration of the coordination geometry, charge transfer, change in pK_a and displacement of coordinating water molecules. First, as zinc(II) has a saturated electronic configuration (d^{10}), electrostatic interaction is the major component and the geometry is not as well defined as with other transition metals. In fact, although zinc adopts an octahedral geometry in water through hexacoordination with water molecules, our survey of available crystal structures revealed that the coordination geometry in proteins ranges from tetrahedral to octahedral and may be highly distorted from ideality. As a result, since the coordination geometry is not stringent, simple terms such as electrostatic and van der Waals have been used in several simulation studies.[22]

Second, changes in pK_a and proton transferability significantly modify the zinc coordination energy. As an example, hydroxamic acid-based HDAC8 inhibitors bind to the zinc ion and make an additional two hydrogen bonds with neighboring residues Tyr307 and His142 (Figure 4.1). A close look at another twelve metalloenzymes revealed that a basic residue not coordinated to the zinc ion (e.g., His 142 in HDAC8) and/or a hydrogen bond donating residue (e.g., Tyr307 in HDAC8) can be found in all of these cases. The prevalence of such residues in close proximity to the catalytic zinc may be explained by their role in the enzyme catalytic function. In fact, these residues participate in the catalytic activity by, for instance, converting $L_3Zn(OH_2)$ with L being histidine or

glutamic acid to $L_3Zn(OH)$, hence converting the water molecule into a more reactive hydroxide ion (Figure 4.2). For example, the pK_a of a water molecule bound to the catalytic zinc ion of carbonic anhydrase is as low as 7 while it is 15.7 in bulk water.[23] These changes in pK_a sometimes result in a proton exchange (acid-base reaction). For the zinc coordination energy to be computed correctly, this property must be investigated and considered.

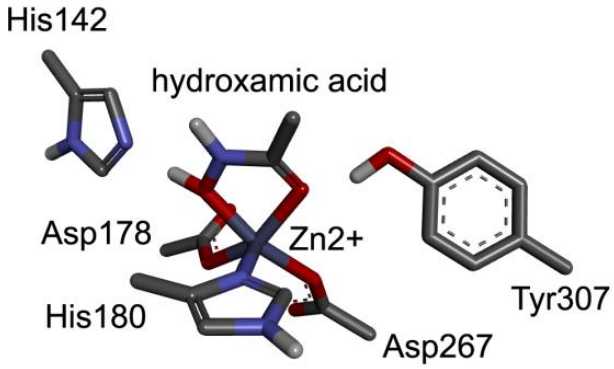


Figure 4.1. HDAC zinc binding site with an hydroxamic acid-containing ligand. Only the zinc binding group of the ligand is shown for clarity (PDB code: 1t67).

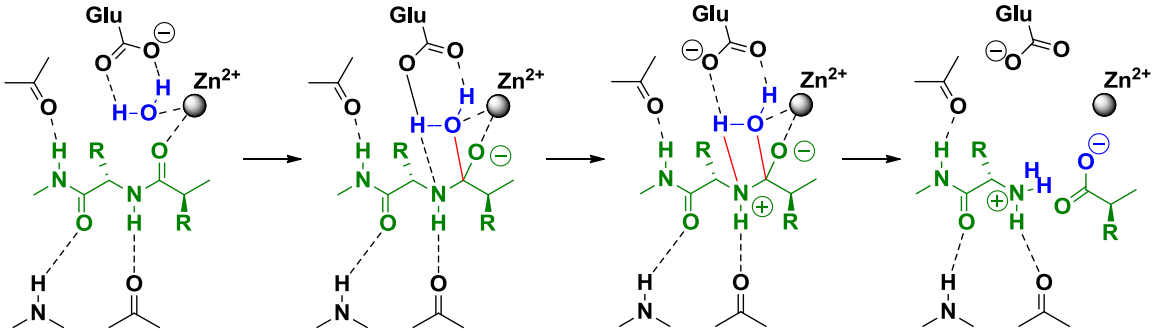


Figure 4.2. Catalytic process of matrix metalloproteinases (Substrate in green, water in blue and enzyme in black)

Some of the above-mentioned basic residues also participate in drug binding through interactions other than usual non-bonded interactions. Previous computational studies have shown that a proton transfer from the drug to the basic residue or change in protonation state of this residue by buffer proton uptake can occur (Figure 4.3). Approximately 10 years ago, Cross and co-workers reported a change in the hydroxamic

acid pK_a of over 3 pK_a units when bound to TACE with a concomitant increase in the pK_a of the neighboring Glu406 of nearly 2 pK_a units.[24] These changes in acidity led the hydroxamic ligand to be more acidic than Glu406 and hence induced a proton shift. The protonation of Glu406 was also proposed when an acetate group was bound to zinc. Earlier this year, this protonation of neighboring Glu when an acetate is bound to zinc was demonstrated experimentally in MMP12.[25] A similar proton transfer was initially proposed for HDAC8,[26] but was recently found to be disfavored by nearly 4.0 kcal/mol (Figure 4.3c).[27] When docking ligands to metalloenzymes, docking programs should therefore include routines to transfer protons when required. To our knowledge, this has never been done.

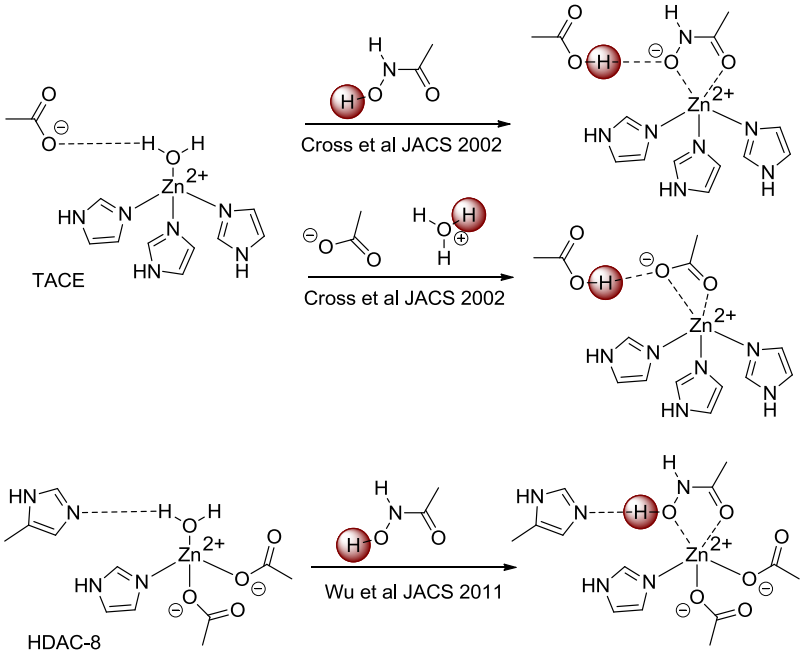


Figure 4.3. Zinc coordination and proton transfer.

As shown in Figures 4.3 and 4.4, one water molecule (or hydroxide ion) is coordinating the catalytic zinc ions (distance lower than 2.5 Å) as observed in several crystal structures such as MMPs (e.g., 1xuc) and thermolysin (e.g., 3tln) while a second one may appear at greater distances (greater than 3 Å) somewhat solvating the first one (e.g., ACE, 1j38). Even enzymes in which zinc coordinates two hydroxyl groups of carbohydrates, such as α-mannosidase, have only one water molecule coordinating zinc when unbound (α-mannosidase, 3bub). Displacement of this water molecule is necessary

for ligand binding. In order to properly score the ligand-metal coordination energy, the water-metal energy should be subtracted. Thus the water coordination energy should be also known.

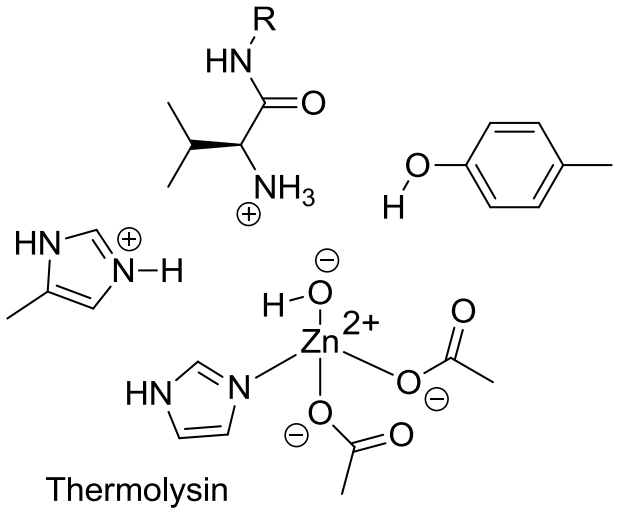


Figure 4.4. Ligand interacting with water molecule which coordinates to zinc (thermolysin, 8tln).

Among the other major factors that docking programs should account for (implicitly or explicitly) is charge transfer. When zinc is coordinated to, for example, 3 histidine residues, its actual charge is not +2 but rather closer to +1 or lower. In addition, the atomic charge also significantly changes when an additional ligand (e.g., a drug) is bound to a free coordination site. For example, Merz and co-workers have developed force field parameters for zinc-containing enzymes in which the zinc atomic charge varies from 0.43 to 0.92 depending on both the ligand (such as water or hydroxide) and coordinated protein residues. In addition, large variations of charge transfer were observed between a neutral (water, charge transfer of 0.17) and negatively charged ligands (hydroxide, 0.41).[28] Consequently, when docking a library of small molecules, a distinction should be made when neutral or charged molecules are considered. Classical molecular mechanics cannot account for these effects, as atomic charges are fixed unless specific charges are developed for each system. In order to account for this charge transfer, polarizable force fields can be used.[21] However their implementation in docking programs can be challenging due to the computational power and time needed to dynamically modify atomic properties.

4.4 Implementation

4.4.1 DFT studies and testing set

We first planned to develop a function that will evaluate the zinc coordination potential energy. This function will be independent of the currently used hydrogen bond, electrostatic or van der Waals terms. To do so, we turned our attention to quantum mechanical (QM) methods. In this area, density functional theory (DFT) studies have been reported although most include data computed solely with optimized structures and not with structures away from their ideal coordination in order to evaluate the coordination energy surface. In addition, in some of these studies, the role played by the basic neighboring residues was not considered.[29] An exhaustive survey of the PDB led us to collect 121 structures of metalloenzymes bound to various ligands and with most having a good resolution, lower than 2.5 Å. The ligands cover a number of zinc binding groups (e.g., *o*-amidoaniline, terminal sulfonamides, hydroxamic acids, thiols), and even include ligands bearing two potential zinc binding group such as captopril co-crystallized with the angiotensin converting enzyme (PDB code: 2x8z, Figure 4.5) or a sulfanyl butanoic acid derivative (3i1u). These latter systems will enable us to test whether the proper zinc binding group can be identified by our docking method if a small molecule features more than one. Some of these ligands do not coordinate to zinc such as the hydrolysis product Val-Trp bound to thermolysin (PDB code: 3tmn, Figure 4.5). These structures were assembled by family (e.g., MMP1, MMP3, MMP8, α-mannosidase, thermolysin, HDAC8, see Appendix 2 for complete set).

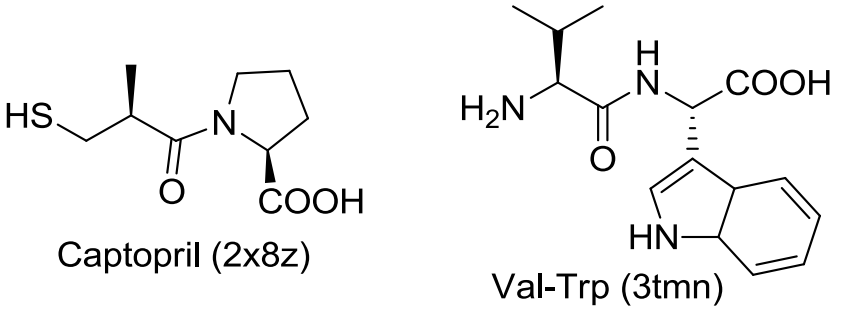


Figure 4.5. Selected ligands co-crystallized with metalloenzymes.

In order to investigate the energetics of the coordination, 19 representative structures out of the 121 were selected for DFT studies. These 22 systems were selected to cover as much of a diverse set of zinc coordination spheres (i.e., coordinating ligand and protein residues) as possible. The key residues and ligands were truncated (the main chain atoms were removed) leading to systems such as the one shown in Figure 4.1. Then hydrogen atoms were added and optimized using GAMESS-US (B3LYP/6-31G*). For the following studies, we kept all the protein atoms frozen as was previously reported.[30] We are aware that DFT functionals are not optimal for these metal coordination energies. Despite the large use of B3LYP in the area, recently reported work by Friedman and co-workers found that on significantly smaller and highly polar systems (e.g., H_3C-S- :::: Zn^{2+}), B3LYP overestimates the interaction energy by about 5% at the equilibrium distance and by even more at longer distance due to an overestimation of the polarization energy.[31] However as mentioned by Friedman and co-workers, the presence of four or five coordinating ligands in our truncated systems should reduce the polarization error hence the overall error of B3LYP. In addition, the size of our systems (four or five coordinating ligands, 40 to 90 atoms) precludes the use of MP2 leaving DFT as an acceptable alternative.

4.4.2 Computing zinc coordination energy

Scripts were developed to move the ligand away from the zinc atom in 0.20 Å increments and single point energy was calculated for each configuration. In order to uncouple the effect of the zinc coordination from the effect of neighboring residues, the same trajectory was alternatively computed with and without these residues; the bound state/unbound state energy difference must be attributed to the zinc coordination energy only if the residue is not present. As mentioned previously and shown in Figure 4.3, the neighboring residues may have significant effects on the binding affinities of zinc binding groups (ZBG). In the case of HDAC8-hydroxamic acid interactions, the overall potential energy drops by as much as 45 kcal/mol when His142 is kept in the truncated system (Figure 4.6). When this additional residue is ignored, the gain in energy of the hydroxamic acid coordination is reduced to about 20 kcal/mol for a difference of

approximately 25 kcal/mol. In practice this histidine residue participates in a strong hydrogen bond with the hydroxamic acid, increasing the polarity of the O-H bond, hence the coordination energy of the oxygen to zinc. In turn, this coordination increases the acidity of this hydrogen, hence the strength of the hydrogen bond. This demonstrates that the two effects (hydrogen bond and zinc coordination) are acting in concert.

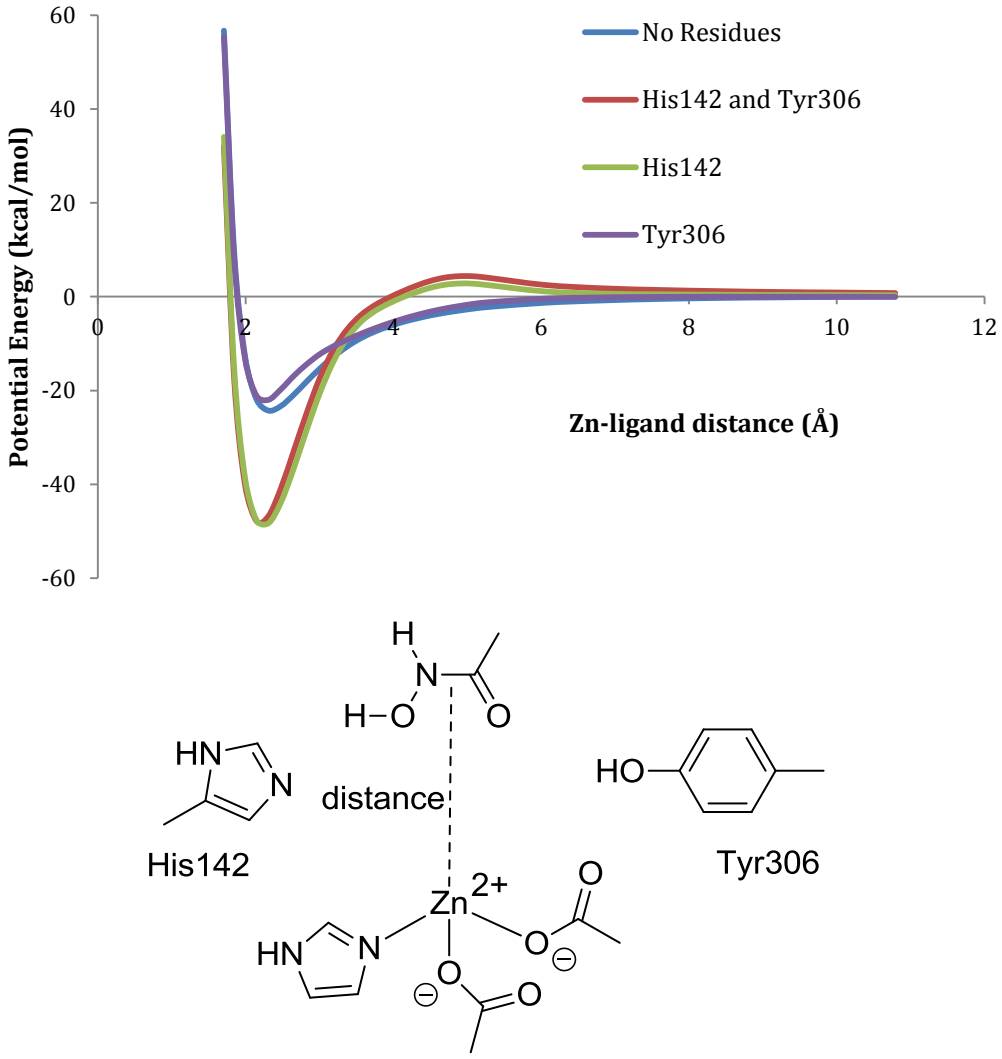


Figure 4.6. Zinc coordination energy for HDAC8 (PDB: 1t67). The shoulder between 4 and 6 Å observed when His142 was kept is due to light steric clashes upon ligand removal.

4.4.3 Computing proton transfer to a neighboring residue

In addition to this zinc coordination, cases where proton transfer is expected were evaluated. Hydroxamic acids are neutral at physiological pH. However, as mentioned above, when approaching the coordination sphere of the zinc ions, the pK_a decreases and the proton may transfer to a neighboring glutamate or histidine residue. To probe this effect, the potential energy of hydroxamate and hydroxamic acid approaching the zinc coordination sphere next to a glutamic acid or a glutamate respectively was first computed (Figure 4.7). Our calculations on the representative systems revealed that at short ligand-zinc distances, the proton of hydroxamic acids, sulfonamides and thiols is likely transferred to the glutamate (via a threonine in the case of carbonic anhydrase). The same calculations were performed when a histidine was the basic residue. In contrast to the glutamate containing systems, we found that the hydroxamic acid proton does not transfer in HDAC8 in which the zinc is neutralized (i.e., coordinated with two aspartates and one histidine). These observations, which are in agreement with the reports from Cross *et al*.[24] and Wu et al.,[27] validated our approach in which the crystallographic Cartesian coordinates were used without further optimization.

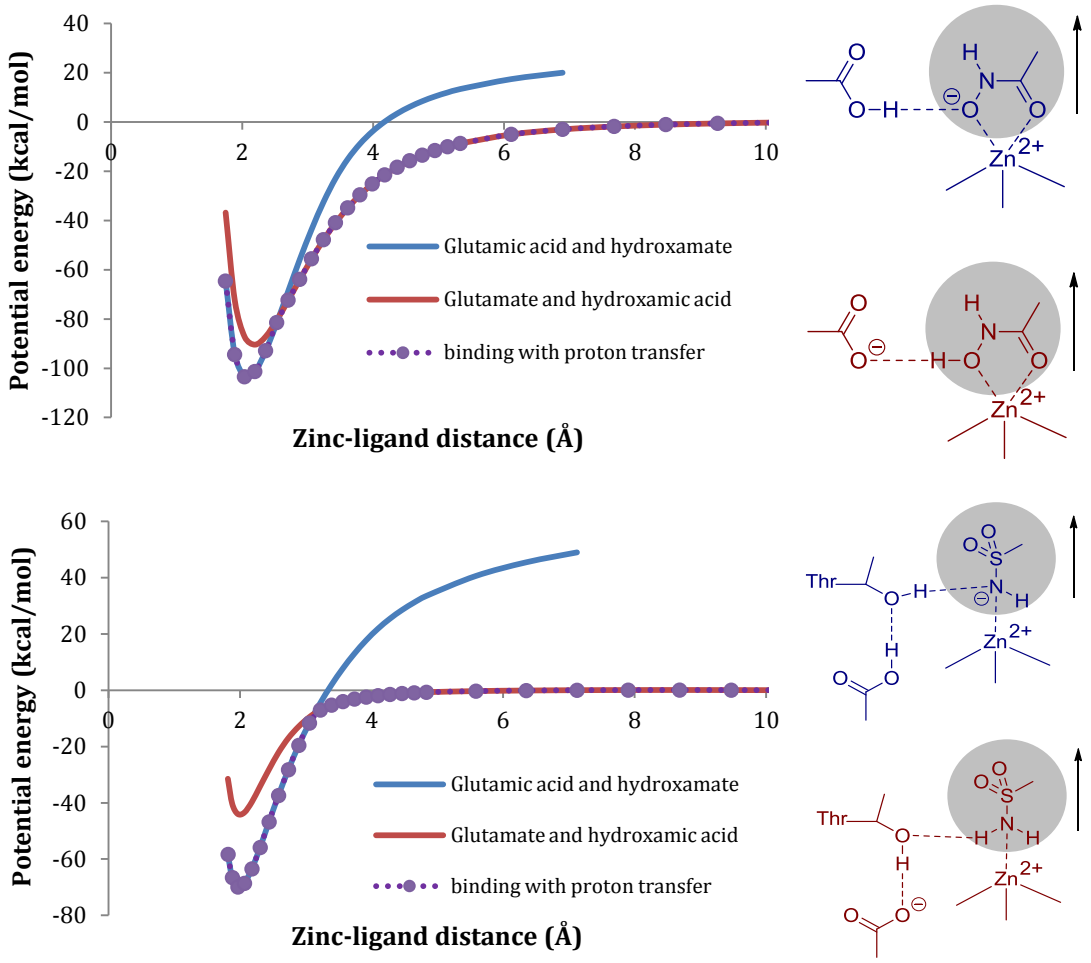


Figure 4.7. The effect of transferring a proton from ligand (hydroxamic acid or sulfonamide) to residue (glutamate with or without a threonine relay). a) 1kbc, human neutrophil collagenase (MMP-8); b) 3s71, carbonic anhydrase. The blue curve shows the favourable situation when the ligand is coordinated and the red shows the favourable situation at longer distances. The purple curve is the combination of the two that is required for docking.

In order to uncouple zinc coordination and proton transfer for implementation into FITTED, investigations were carried out on diversely truncated systems. In contrast to the ligand-zinc coordination energy which was computed with or without neighbouring residues by moving the ligand away, the hydrogen bond strength was evaluated (QM methods mentioned above) with the ligand bound to zinc in its equilibrium position while moving the neighbouring residue away from the ligand along the hydrogen bond coordinate (Figure 4.8).

If we combine the two potential energy curves once more for the 22 representative systems, we see that the distances at which hydrogen bonds occur are solely controlled by

the ionized ligands and glutamic acid residues and the neutral ligand/glutamate energies can be ignored (Figure 4.8). The point at which the two curves cross is assumed to be the energy barrier for the proton transfer to occur. However this is difficult to model with molecular mechanics and thus a flatter-minimum curve was implemented.

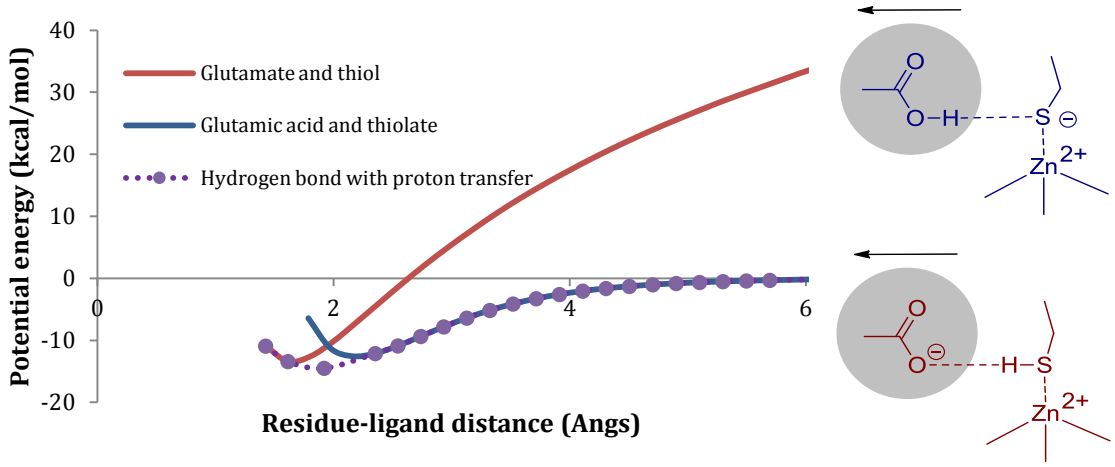


Figure 4.8. The computed interaction energy for selected truncated systems (1uzf, ACE).

A close look at crystal structures also showed that the residue for which no transfer was observed can still form hydrogen bonds that are shorter and stronger than usual. Among the examples is Tyr307 (Figure 4.1) in HDAC8. An approach similar to that used to develop zinc coordination parameters has been applied to develop the corresponding parameters. In a nutshell, the neighbouring basic residues were moved away by 0.20 Å and the corresponding energy was calculated.

4.4.4 Computing water coordination energy

As discussed above, optimal binding energy can only be computed if the energy associated with the displacement of a water molecule is considered. For this purpose, the above systems were used in which the ligand was replaced by a water molecule and the position of the latter was optimized through DFT energy optimization. In some cases, the acidifying effect of the zinc ion led to a spontaneous proton transfer from the water molecule to a glutamate and in contrast, when many aspartic acids are coordinated to zinc, the proton spontaneously transferred from the glutamic acid to the hydroxide ion (Figure 4.9). In order to compute the water coordination energy, the difference in energy

between the lowest in energy bound states (either water-glutamate or hydroxide-glutamic acid, Figure 4.9) and the unbound state was computed. For optimal transferability of these calculations, the set included systems with either Glu (alone or via Thr) or His as a basic residue and either His/His/His, His/His/GA, His/GA/GA, His/His/GA/GA or His/GA/GA/GA as residues in the zinc coordination sphere, where GA represents Glu or Asp.

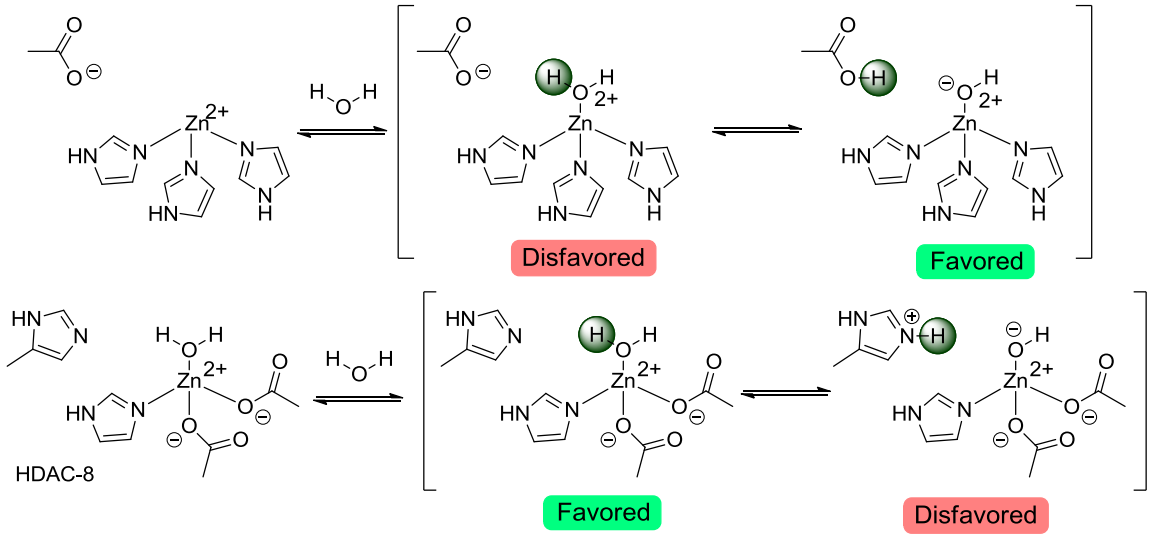


Figure 4.9. Zinc coordination and proton transfer with water molecule in the cases of proximal glutamic acid (top) or histidine (bottom).

4.4.5 Energy function parameterization

As discussed above and shown in Figure 4.10, the binding of small molecules to zinc-containing metalloenzymes is often not a simple bimolecular binding.

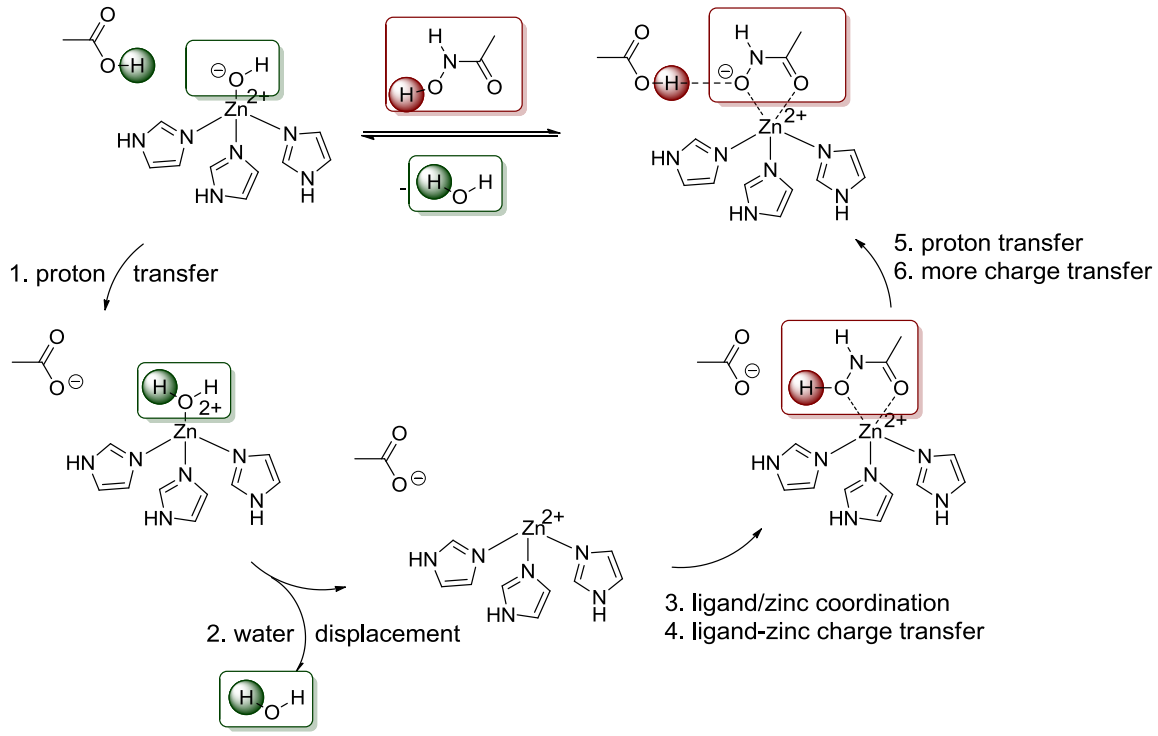


Figure 4.10. Binding process.

In order to properly model the zinc coordination (#3 in Figure 4.10), we proposed to develop a novel energy function term for computing the zinc-ligand interaction. This function should be independent of the charge eliminating the charge transfer effect (#4 and 6 in Figure 4.10). From the collected DFT data, the energy well depth (ε) and zero energy point (σ) were determined for each system and used to derive a Lennard-Jones-like potential equation. The derived equations (Eqs. 4.1-4.3) predict the energy differences with good accuracy (see Figure 4.11).

$$\frac{A}{r^x} - \frac{B}{r^y} - \frac{C}{r^2} \qquad \textbf{(4.1)}$$

$$\frac{A}{(r-0.25)^x} - \frac{B}{(r-0.25)^y} \qquad \textbf{(4.2)}$$

$$A = 4\varepsilon * \sigma^6 \ \ and \ B = 4\varepsilon * \sigma^3 \qquad \textbf{(4.3)}$$

This approach was applied to both zinc coordination and the strong hydrogen bonds described above. (x,y)=(6,3), (8,4) and (10,6) were assessed with A and B being products of the energy well depth and zero points energies as defined in Lennard-Jones' theory,[32] C being a constant term to be trained and r being the distance between the coordinated atom

and zinc. The third term was initially introduced to account for electrostatic interactions that may vary from one system to another. However, we found that upon fitting the curves for each system using the 6-3 relationship, removing this third term was possible without significantly affecting the fit of the DFT and MM curves. Unexpectedly, a shift of 0.25 Å of the position of the minimum of the energy well was observed in the majority of cases. All these observations led us to implement function (2) into FITTED with the parameters given in Section 4.7. Examples of the predicted energy curves are given in Figure 4.11.

In addition to the new zinc-coordination energy equation, upon investigating the special hydrogen bond between the bound ligand and the neighboring basic residue (#5 in Figure 4.10), we determined that a new equation should be implemented for this type of interactions as well. We were able to model this as a 6-3 relationship as well, but with different parameters. The depth of the energy-well upon distancing the basic residue, as shown in Figure 4.8, is far beyond the expected stabilization obtained from currently modeled hydrogen bonds.

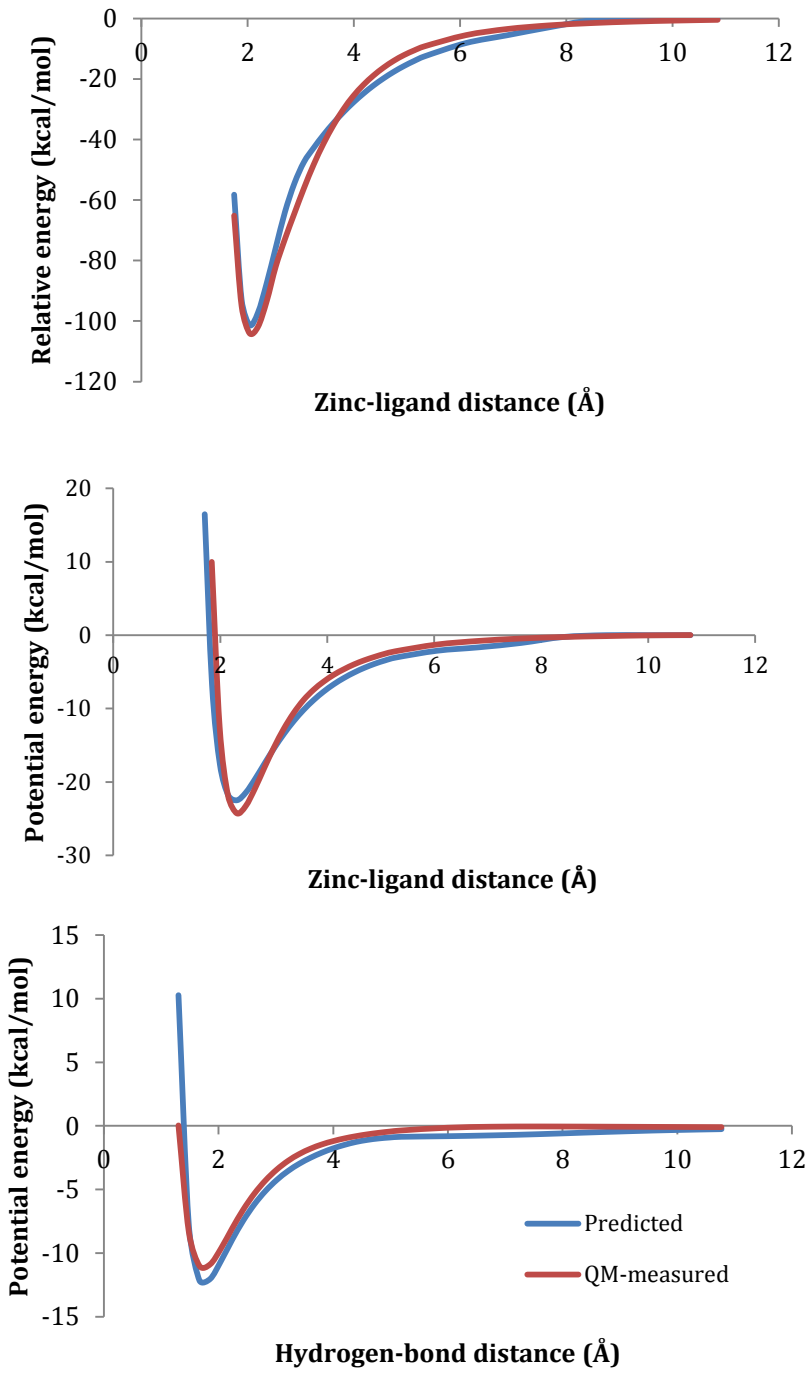


Figure 4.11. FITTED-derived energy curves using a LJ 6-3 equation vs. DFT-derived energy curves. Top: zinc-ligand interaction energy 1kbc; middle: zinc-ligand interaction energy HDAC8 (1t67); bottom: hydrogen bond energy (1kbc).

4.4.6 Implementation

At this stage, the newly implemented FITTED energy function can now properly evaluate the energy of such zinc-ligand systems. Then a routine identifying whether the proton transfer should be carried out based on distances and chemical nature of the coordinating functional group has been implemented. First, a routine identifying acidic zinc binding groups (e.g., hydroxamic acid, terminal sulfonamide and thiols) was implemented into SMART, a program of the FITTED suite used to prepare the small molecules prior to the actual docking. This information is then output in the ligand file. Within FITTED itself, a routine was implemented that can read this information from the ligand file and a switching function was introduced which recognizes whether the zinc binding group (e.g., hydroxamic acid) is close enough to zinc to be ionized or not. If it is ionized, the energy function is applied to the system on the left in Figure 4.12, or if neutral, it is applied to the system on the right. In practice, hydrogens on both the functional group and on the basic residue are present and a list of interactions for each is prepared. The newly implemented routine identifies which of the two lists should be selected.

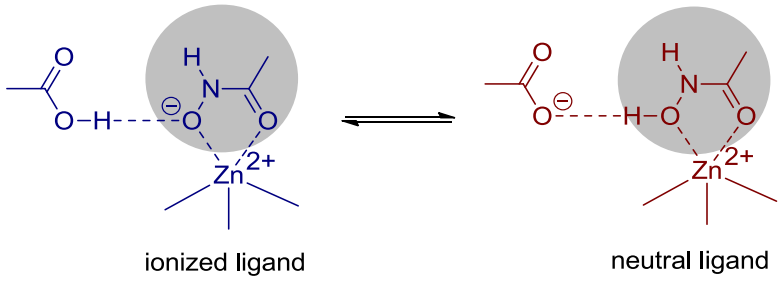


Figure 4.12. Modeling proton transfer.

Since one of the major features of FITTED is the option to displace water molecule, no additional modifications were necessary for displacing the coordinating water in the unbound state. However, scoring the displacement of water molecule was modified to incorporate the water/zinc coordination energy which is highly dependent on the zinc coordination sphere. In parallel to this explicit displaceable water molecule, we also assessed the use of a more implicit approach. For this purpose, scaling factors were

applied to the zinc coordination energies to implicitly account for the displacement of the water molecules coordinating the zinc ion.

4.5 Results and Discussion

4.5.1 Validation - pose prediction

With these implementations in hand, a first set of validation experiments was carried out. For this purpose, self-docking experiments using the 121 complexes selected as a testing set were performed. All these docking experiments were carried out 10 times to ensure that the result was statistically significant. In order to validate the novel energy function and implementations, the accuracy of FITTED, using either the traditional electrostatic/van der Waals energy, the previously implemented hydrogen bond-like term or the current version with an explicit or implicit water molecule coordinating zinc, was assessed. At this stage, we expected a significant improvement of the zinc coordination geometry that will be therefore more accurately scored. In practice, we observed that the novel implementation significantly improved the positioning of the zinc binding group (Figure 4.13 top). As a result, the overall pose prediction is significantly enhanced (Figure 4.13 bottom) with improvement as large as 10% (RMSD lower than 2Å) and even 13% if the best scoring pose of the 10 runs was used for each of the 121 systems. More unexpectedly, the implicit water displacement energy treatment turned out to lead to more accurate predictions.

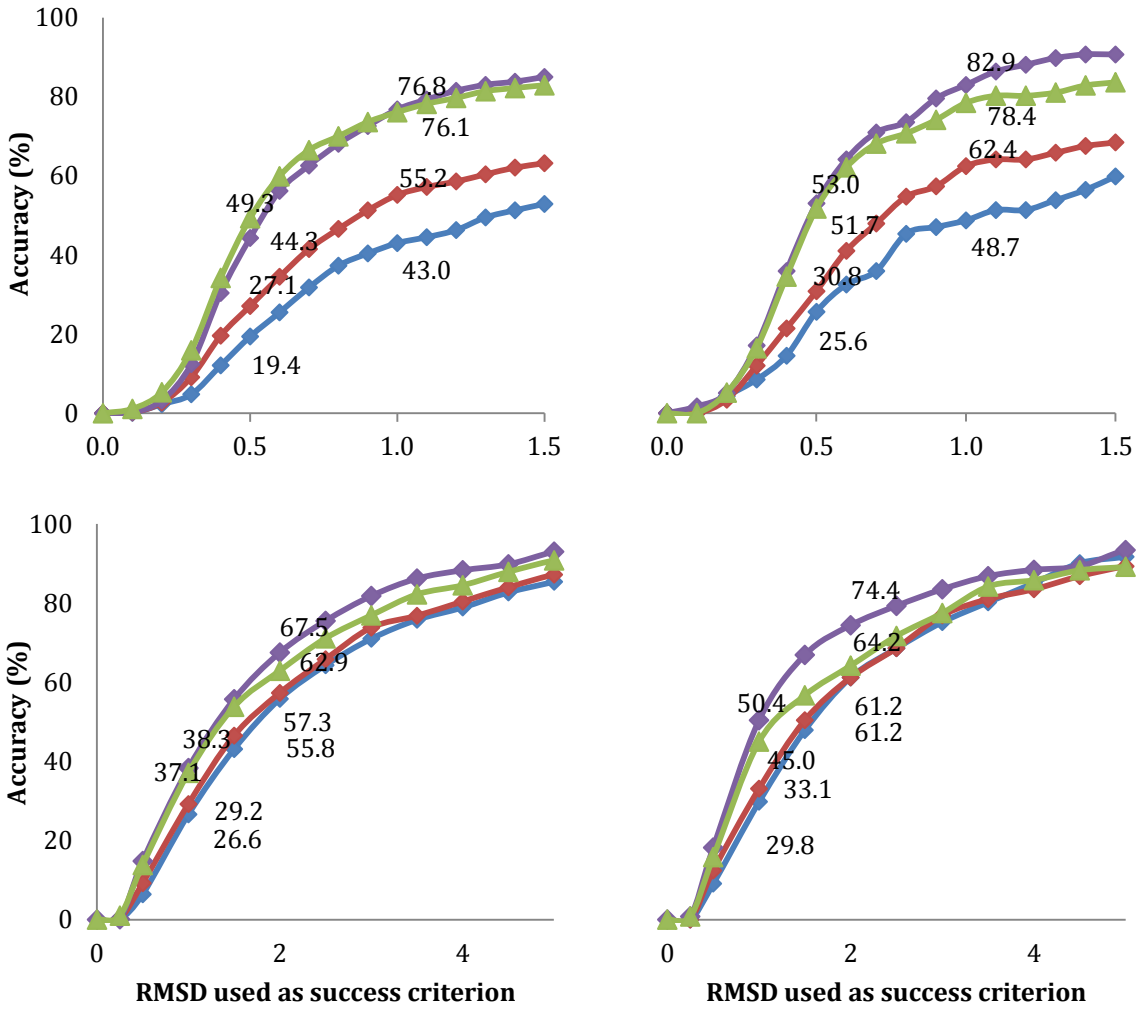


Figure 4.13. Pose prediction accuracy. Blue: 12-6 Lennard-Jones + electrostatic, red: 12-10 Hydrogen bond-like, purple: new implementation with implicit water, green: new implementation with explicit water. Top panel: accuracy of the **zinc coordination** geometry (the RMSD of only the zinc binding group is computed) average over 10 runs (left), best-scoring of the 10 runs (right); bottom panel: accuracy of the **pose prediction** average over 10 runs (left), best-scoring of the 10 runs (right).

We next looked more specifically at HDAC / inhibitor complexes (Table 4.1). While the previous implementations have overall success of 39% (Electrostatic, LJ12-6 treatment of metal coordination) and 48% (Electrostatic, LJ12-10), the novel implementation predicted the pose correctly 71% of the time. 3c0z (HDAC7) and 1t67 (HDAC8) are two examples of significant pose prediction improvement (Figure 4.14). While the previous two implementations predicted poses with RMSDs of about 4 Å in all

10 runs with 3c0z, the novel implementation enabled the prediction of poses within 1.5Å of the observed poses in all 10 runs.

Table 4.1. Docking accuracy on HDACs. The three implementations are compared for their accuracy in predicting the ligand binding modes (implicit water displacement energy mode).

	Implementation	Average RMSD[a]	Lowest RMSD[a]	Success rate[b]
HDAC2 – 3max	Elec./vdW	0.38	0.35	100%
	H-Bond-like	0.38	0.36	100%
	New implem.	0.32	0.28	100%
HDAC4 – 2vqj	Elec./vdW	8.85	2.60	0%
	H-Bond-like	3.62	1.90	10%
	New implem.	3.56	2.11	0%
HDAC7 – 3c0z	Elec./vdW	4.07	3.85	0%
	H-Bond-like	4.38	4.20	0%
	New implem.	1.25	1.09	100%
HDAC7 – 3c10	Elec./vdW	0.92	0.85	90%
	H-Bond-like	1.04	0.82	100%
	New implem.	0.89	0.82	100%
HDAC8 – 1t67	Elec./vdW	2.11	1.51	30%
	H-Bond-like	2.79	1.56	20%
	New implem.	1.46	0.82	90%
HDAC8 – 1t69	Elec./vdW	2.99	1.82	20%
	H-Bond-like	2.13	1.58	50%
	New implem.	3.82	1.27	40%
HDAC8 – 1w22	Elec./vdW	1.55	0.91	60%
	H-Bond-like	1.36	0.92	90%
	New implem.	1.47	0.72	70%
HDAC8 – 3f07	Elec./vdW	2.98	1.90	10%
	H-Bond-like	2.88	1.60	10%
	New implem.	1.59	1.14	70%

[a] over 10 runs. [b] percentage of runs with RMSD < 2Å (out of 10 runs)

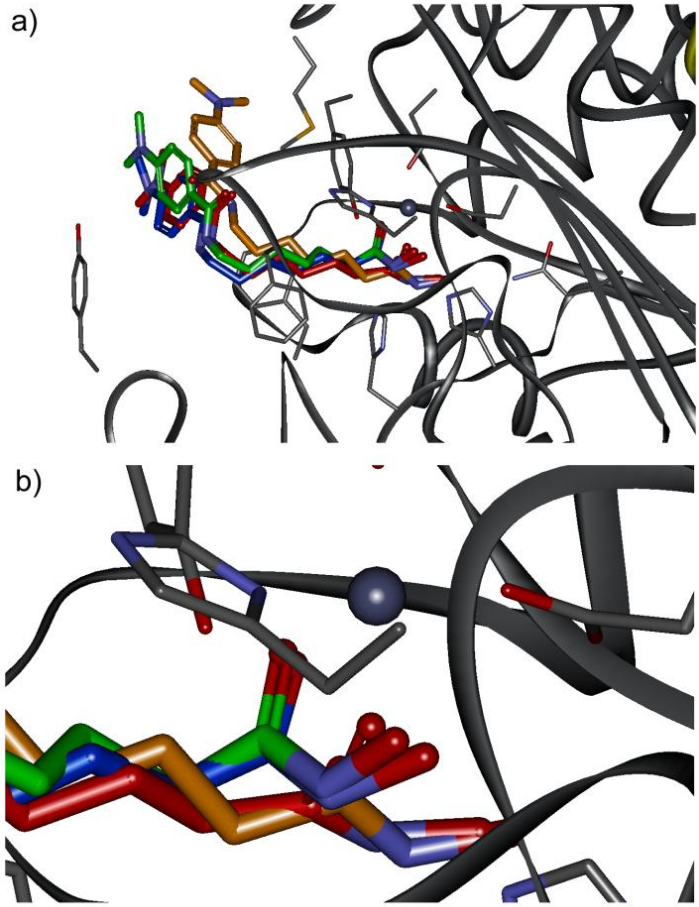


Figure 4.14. Predicted poses: Elec/vdW mode: red, HB-mode: orange, new implementations: green vs. crystal structure (green). a) ligand binding mode; b) zinc coordination.

4.5.2 Validation - virtual screening

At this stage, these novel implementations were applied to the screening of potential metalloenzyme inhibitors. For this purpose, known actives and decoys should be collected. The DUD-e set includes inhibitors and decoys for five metalloenzymes which were used herein. In order to test our entire set of programs, the tautomers from the sets of ligands and decoys from the DUD-e sets were identified and removed and 100 unique ligands and 5000 unique decoys were selected and hydrogens added. All these steps were done using routines of our FORECASTER platform as described in the experimental section. Table 4.2 summarizes the accuracy of these screens. While DOCK provided AU-ROC values ranging from 0.71 to 0.80 with an average of 0.744, FITTED provided AU-ROC

above 0.79 for 4 of the 5 targets used in this validation study with an improved average of 0.829 with an explicit water molecule bound to zinc. Surprisingly HDAC2 provided a lower AU-ROC of 0.67.

Further investigation of this somewhat disappointing result revealed that the set includes several macrocyclic molecules, as exemplified by ChEMBL424189 (trapoxin B, Figure 4.15), as well as several inhibitors with bulky and/or branched cap groups that did not fit into the binding site of HDAC2 co-crystallized with a much smaller ligand (3max, Figure 4.15). The N-terminal L1 loop in HDAC2, which is on the surface lining the opening to the active site, is several amino acids longer than the corresponding loop in HDAC8 (Figure 4.16). The result is that HDAC2 possesses a longer access tube and a more restricted surface that may limit the docking of larger or conformationally restricted cap groups. In addition, it is established in some HDACs, including HDAC8, that the surface is malleable and thus may accommodate larger ligands.[33] The current version of FITTED considers the protein flexibility using sets of experimental protein conformations (i.e., crystal structures). However, all three crystal structures of HDAC2 have small ligands bound and thus it was not possible to model this potential flexibility. In addition, we have found that among the actives included in the DUD-e validation set, some are most likely not simple competitive inhibitors. For example, trapoxin B is a known irreversible covalent inhibitor[34] and that ChEMBL402341, a dithiolethione-containing inhibitor releases H_2S over time (Figure 4.15).[35] Thus the actual active inhibitors might be species other than those in the database.

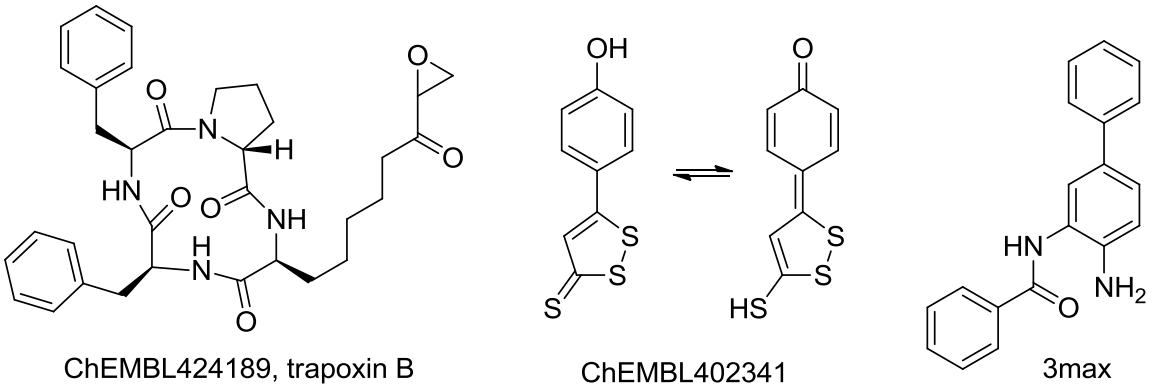


Figure 4.15. HDAC2 inhibitors.

Table 4.2. Area under receiver operating curve.

Enzyme	DOCK[a]	FITTED[b]
ACE	0.72	0.79
CA	0.73	0.89
HDAC2	0.77	0.67
HDAC8	0.80	0.89
MMP13	0.71	0.90

[a] data provided on http://dude.docking.org/targets. [b] HOH explicit

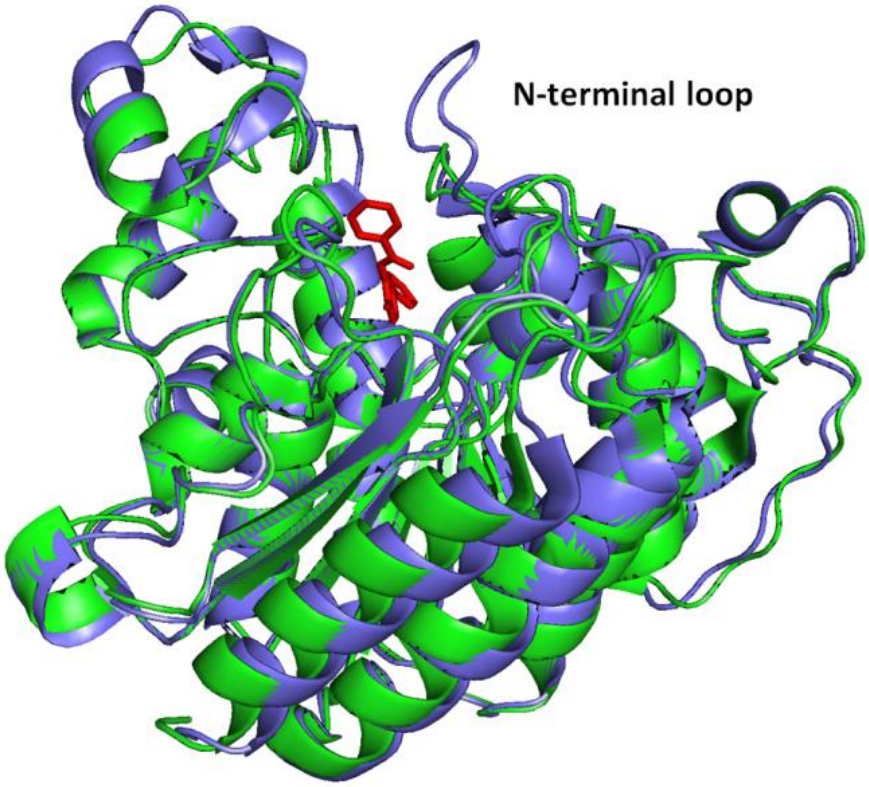


Figure 4.16. Comparison of HDAC2 (3max, blue) and HDAC8 (2v5x, green). The additional length of the *N*-terminal loop increases the length of the access tube. The ligand shown in red is the *N*-(2-aminophenyl)-2-benzamide from the HDAC2 structure (3max).

4.6 Conclusion

In conclusion, our docking program FITTED has been modified to account for zinc coordination of ligands and other related processes such as unusually strong hydrogen bonds and proton exchange with neighboring residues. These implementations significantly improved the pose prediction accuracy over the more traditional electrostatic/Lennard Jones treatment or even the hydrogen bond-like treatment previously implemented. Based on the success of this technique, it is likely that this approach could be applied to other transition metals if desired. A close look at the results showed that docking to HDACs was also more accurate and that identification of actives in screening is possible with this current version of the program.

4.7 Experimental

4.7.1 DFT calculations

All the quantum mechanical calculations were performed using DFT, more specifically the B3LYP functional (Restricted Hartree−Fock) and a custom basis set combining 6-311G* for the zinc atom and 6-31+G* for all other atoms (based on ref. 24). All calculations were performed in vacuum. The B3LYP calculations were performed using GAMESS-US v.Aug2011-64bit. The crystal structures from the PDB were truncated to include the zinc atom, the residues coordinated to zinc (also truncated) and the residues interacting with the ligand. The ligands were also truncated to include only the ZBG. Hydrogen atoms were added to the ligand only where they could interact with the neighboring residue in the active site (i.e. where a proton shift could occur).

4.7.2 Force field parameters

The first step was to freeze all atoms in the system other than the manually-added hydrogen atoms which were optimized using the DFT methods mentioned previously. This was deemed to be the equilibrium state (crystal structure + optimized hydrogen atoms). Next, the system was broken down and the neighboring residues were omitted.

The ligand was moved systematically away from the zinc along a designated vector by steps of 0.2 Å in order to obtain the energy profile (single point energies). Similarly, the neighboring residue was moved systematically along the hydrogen bond vector in order to obtain the hydrogen bond energy profile. These two sets of calculations were performed with the hydrogen atom either on the ligand or on the neighboring residue in order to establish which configuration would be optimal depending on the conditions (i.e. distance of the ligand from the zinc atom); either one configuration was always preferential (no proton transfer) or the two profiles had to be combined (proton transfer). Special considerations were made for the complex systems (carbonic anhydrase) where the hydrogen transfer occurred via a threonine to a glutamic acid.

The generated structures along the vectors were used to compute the FITTED energy and to derive force field parameters for optimal curve matching.

4.7.3 Construction of the testing sets

The sets of decoys and ligands were downloaded from the DUD-e web site (http://dude.docking.org/targets) and further processed as follows. First the hydrogens were removed from all the molecules in order to make the tautomers as similar as possible using a routine of FORECASTER. Then the molecules were clustered by similarity using SELECT and 100 ligands and 5000 decoys with the largest diversity were selected. Hydrogens were added using CONVERT and all these ligands prepared for docking using SMART.

4.7.4 Preparation of the protein files

PREPARE and PROCESS were applied using the specific keyword identifying metalloenzymes (Macromolecule metalloprotein) and other parameters to the default.

4.7.5 Docking with FITTED

Each docking is a set of 3 runs starting from a different seed. The different implementations were identified by a specific keyword (Macromolecule protein/metalloprotein_HB/metalloprotein). For the pose prediction tests each set of 3 runs was carried out 10 times. For the screening, 3 runs for each small molecule were also carried out.

4.7.6 Application of FITTED

Default parameters implemented in FITTED have been used.

4.8 Acknowledgements

We thank FQRNT (Équipe program) for financial support. Calcul Québec and Compute Canada are acknowledged for generous CPU allocations.

4.9 References

1. Hanessian, S.; Moitessier, N.; Therrien, E., A comparative docking study and the design of potentially selective MMP inhibitors. *J*. *Comput*. *Mol*. *Des*. **2001,** *15* (10), 873-881.

2. Hanessian, S.; MacKay, D. B.; Moitessier, N., Design and synthesis of matrix metalloproteinase inhibitors guided by molecular modeling. Picking the S1 pocket using conformationally constrained inhibitors. *J*. *Med*. *Chem*. **2001,** *44* (19), 3074

3. Hanessian, S.; Moitessier, N.; Cantin, L. D., Design and synthesis of MMP inhibitors using N-arylsulfonylaziridine hydroxamic acids as constrained scaffolds. *Tetrahedron* **2001,** *57* (32), 6885.

4. Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N., Evaluation of Docking Programs for Predicting Binding of

Golgi alpha-Mannosidase II Inhibitors: A Comparison with Crystallography. *Proteins: Struct., Funct., Bioinf*. **2007,** *69* (1), 160-176

5. Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R., Towards the development of universal, fast and highly accurate docking//scoring methods: a long way to go. *Br. J. Pharmacol*. **2008,** *153* (S1), S7-S26

6. Corbeil, C. R.; Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model*. **2007,** *47* (2), 435-449

7. Corbeil, C. R.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J. Chem. Inf. Model*. **2009,** *49* (4), 997-1009

8. De Cesco, S.; Deslandes, S.; Therrien, E.; Levan, D.; Cueto, M.; Schmidt, R.; Cantin, L. D.; Mittermaier, A.; Juillerat-Jeanneret, L.; Moitessier, N., Virtual screening and computational optimization for the discovery of covalent prolyl oligopeptidase inhibitors with activity in human cells. *J. Med. Chem*. **2012,** *55* (14), 6306-6315.

9. Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 4. Are popular scoring functions accurate for this class of proteins? *J. Chem. Inf. Model*. **2009,** *49* (6), 1568-1580

10. Marcial, B. L.; Sousa, S. F.; Barbosa, I. L.; Dos Santos, H. F.; Ramos, M. J., Chemically modified tetracyclines as inhibitors of MMP-2 matrix metalloproteinase: A molecular and structural study. *J. Phys. Chem. B* **2012,** *116* (46), 13644-13654.

11. Norris, R.; Casey, F.; FitzGerald, R. J.; Shields, D.; Mooney, C., Predictive modelling of angiotensin converting enzyme inhibitory dipeptides. *Food Chem*. **2012,** *133* (4), 1349-1354.

12. Marques, S. M.; Tuccinardi, T.; Nuti, E.; Santamaria, S.; André, V.; Rossello, A.; Martinelli, A.; Santos, M. A., Novel 1-hydroxypiperazine-2,6-diones as new leads in the inhibition of metalloproteinases. *J. Med. Chem*. **2011,** *54* (24), 8289-8298.

13. Omanakuttan, A.; Nambiar, J.; Harris, R. M.; Bose, C.; Pandurangan, N.; Varghese, R. K.; Kumar, G. B.; Tainer, J. A.; Banerji, A.; Perry, J. J. P.; Nair, B. G., Anacardic acid inhibits the catalytic activity of matrix metalloproteinase-2 and matrix metalloproteinase-9. *Mol*. *Pharmacol*. **2012,** *82* (4), 614-622.

14. Feng, J.; Jin, K.; Zhu, H.; Zhang, X.; Zhang, L.; Liu, J.; Xu, W., A novel aminopeptidase N inhibitor developed by virtual screening approach. *Bioorg*. *Med*. *Chem*. *Lett*. **2012,** *22* (18), 5863-5869.

15. Tuccinardi, T.; Bertini, S.; Granchi, C.; Ortore, G.; Macchia, M.; Minutolo, F.; Martinelli, A.; Supuran, C. T., Salicylaldoxime derivatives as new leads for the development of carbonic anhydrase inhibitors. *Bioorg*. *Med*. *Chem*. **2013,** *21* (6), 1511-1515.

16. Irwin, J. J.; Raushel, F. M.; Shoichet, B. K., Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry* **2005,** *44* (37), 12316-12328.

17. Norris, R.; Casey, F.; FitzGerald, R. J.; Shields, D.; Mooney, C., Predictive modelling of angiotensin converting enzyme inhibitory dipeptides. *Food Chemistry* **2012,** *133* (4), 1349-1354.

18. Tavera-Mendoza, L. E.; Quach, T. D.; Dabbas, B.; Hudon, J.; Liao, X.; Palijan, A.; Gleason, J. L.; White, J. H., Incorporation of histone deacetylase inhibition into the structure of a nuclear receptor agonist. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. **2008,** *105* (24), 8250-8255.

19. Lamblin, M.; Dabbas, B.; Spingarn, R.; Mendoza-Sanchez, R.; Wang, T. T.; An, B. S.; Huang, D. C.; Kremer, R.; White, J. H.; Gleason, J. L., Vitamin D receptor agonist/histone deacetylase inhibitor molecular hybrids. *Bioorg*. *Med*. *Chem*. **2010,** *18* (11), 4119-4137.

20. Fischer, J.; Wang, T. T.; Kaldre, D.; Rochel, N.; Moras, D.; White, J. H.; Gleason, J. L., Synthetically accessible non-secosteroidal hybrid molecules combining vitamin D receptor agonism and histone deacetylase inhibition. *Chem*. *Biol*. **2012,** *19* (8), 963-971.

21. Zhang, J.; Yang, W.; Piquemal, J. P.; Ren, P., Modeling structural coordination and ligand binding in zinc proteins with a polarizable potential. *J*. *Chem*. *Theory Comput*. **2012,** *8* (4), 1314-1324.

22. Kalyaanamoorthy, S.; Chen, Y. P. P., Exploring inhibitor release pathways in histone deacetylases using random acceleration molecular dynamics simulations. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (2), 589-603.

23. Berg, J. M.; L., T. J.; Stryer, L., *Biochemistry 5th Edition, WH Freeman NY* **2002**.

24. Cross, J. B.; Duca, J. S.; Kaminski, J. J.; Madison, V. S., The active site of a zinc-dependent metalloproteinase influences the computed pK_a of ligands coordinated to the catalytic zinc ion. *J*. *Am*. *Chem*. *Soc*. **2002,** *124* (37), 11004-11007.

25. Czarny, B.; Stura, E. A.; Devel, L.; Vera, L.; Cassar-Lajeunesse, E.; Beau, F.; Calderone, V.; Fragai, M.; Luchinat, C.; Dive, V., Molecular Determinants of a Selective Matrix Metalloprotease-12 Inhibitor: Insights from Crystallography and Thermodynamic Studies. *Journal of Medicinal Chemistry* **2013**.

26. Vanommeslaeghe, K.; Van Alsenoy, C.; De Proft, F.; Martins, J. C.; Tourwé, D.; Geerlings, P., Ab initio study of the binding of Trichostatin A (TSA) in the active site of histone deacetylase like protein (HDLP). *Organic and Biomolecular Chemistry* **2003,** *1* (16), 2951-2957.

27. Wu, R.; Lu, Z.; Cao, Z.; Zhang, Y., Zinc Chelation with Hydroxamate in Histone Deacetylases Modulated by Water Access to the Linker Binding Channel. *J*. *Am*. *Chem*. *Soc*. **2011,** *133* (16), 6110-6113.

28. Peters, M. B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M. N.; Merz, K. M., Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *J Chem Theory Comput* **2010,** *6* (9), 2935-2947.

29. Wang, D.; Helquist, P.; Wiest, O., Zinc binding in HDAC inhibitors: A DFT study. *Journal of Organic Chemistry* **2007,** *72* (14), 5446-5449.

30. Vanommeslaeghe, K.; De Proft, F.; Loverix, S.; Tourwé, D.; Geerlings, P., Theoretical study revealing the functioning of a novel combination of catalytic motifs

in histone deacetylase. *Bioorganic and Medicinal Chemistry* **2005,** *13* (12), 3987-3992.

31. Ahlstrand, E.; Spångberg, D.; Hermansson, K.; Friedman, R., Interaction energies between metal ions (Zn2+ and Cd2+) and biologically relevant ligands. *Int*. *J*. *Quantum Chem*. **2013**.

32. Lennard-Jones, J. E., Cohesion. *Proc*. *Phys*. *Soc*. **1931,** *43*, 461-482.

33. Somoza, J. R.; Skene, R. J.; Katz, B. A.; Mol, C.; Ho, J. D.; Jennings, A. J.; Luong, C.; Arvai, A.; Buggy, J. J.; Chi, E.; Tang, J.; Sang, B. C.; Verner, E.; Wynands, R.; Leahy, E. M.; Dougan, D. R.; Snell, G.; Navre, M.; Knuth, M. W.; Swanson, R. V.; McRee, D. E.; Tari, L. W., Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure* **2004,** *12* (7), 1325-1334.

34. Kijima, M.; Yoshida, M.; Sugita, K.; Horinouchi, S.; Beppu, T., Trapoxin, an antitumor cyclic tetrapeptide, is an irreversible inhibitor of mammalian histone deacetylase. *J*. *Biol*. *Chem*. **1993,** *268* (30), 22429-22435.

35. Rossoni, G.; Sparatore, A.; Tazzari, V.; Manfredi, B.; Soldato, P. D.; Berti, F., The hydrogen sulphide-releasing derivative of diclofenac protects against ischaemia-reperfusion injury in the isolated rabbit heart. *Br*. *J*. *Pharmacol*. **2008,** *153* (1), 100-109.

Introduction to Chapter 5

Another group of biologically relevant metalloenzymes are cytochrome P450s. These enzymes are responsible for the majority of metabolism of xenobiotics in the liver. The oxidation of drug molecules by P450s facilitates their excretion from the body. However, it is possible for toxic compounds to be generated and, consequently, predicting the outcomes of the chemical reactions between drugs and P450s is essential to the field of medicinal chemistry. The first step to accurately model this process would be to locate the drug molecule in the binding site of the enzyme. As was shown in the previous chapter, docking to metalloenzymes is not trivial and special considerations were necessary. This chapter presents the development of modeling techniques to capture the biochemical process of P450 oxidation: the inherent reactivity of a drug molecule, the binding of a drug molecule to the reactive state of the enzyme and the transition state of the chemical reaction resulting in a final metabolite. This chapter bridges the first two chapters about small molecule catalysis with the second two chapters about enzymatic inhibition and describes the first time that transition state modeling was integrated into a docking program. The chapter further describes the accuracy of the technique, puts it in context with other available methodologies and compares the accuracy of its predictions to the predictions of pharmacokinetic experts in the field.

(This page was left blank intentionally)

Chapter 5:

Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s

This chapter is reprinted with permission from: "Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s", Campagna-Slater, V.‖; Pottel, J.‖; Therrien, E.; Cantin, L. D.; Moitessier, N.; *Journal of Chemical Information and Modelling*, **2012**, *52*(9), 2471-2483.

Author Contributions: **Valérie Campagna-Slater** encoded the transition-state modeling into FITTED to create the foundation for IMPACTS. **Eric Therrien** contributed to the knowledge of using the FORECASTER/FITTED suites of programs and **Louis-David Cantin** was a co-P.I. on a grant and contributed to the knowledge in medicinal chemistry and metabolism. Other coding and validation experiments were contributions of the author of this thesis.

5.1 Abstract

The metabolism of xenobiotics - and more specifically drugs - in the liver is a critical process controlling their half-life. Developing higher throughput predictive methods of the metabolic stability of xenobiotics and identifying their metabolites is an avenue of research. It is expected that predicting the chemical nature of the metabolites would be an asset for designing safer drugs and/or drugs with modulated half-lives. We have developed IMPACTS (*In-silico* Metabolism Prediction by Activated Cytochromes and Transition States), a computational tool combining docking to metabolic enzymes, transition state modeling and rule-based substrate reactivity prediction to predict the site of metabolism (SoM) of xenobiotics. Its application to sets of CYP1A2, CYP2C9, CYP2D6 and CYP3A4 substrates and comparison to experts' predictions demonstrates its accuracy and significance. IMPACTS identified an experimentally observed SoM in the top 2 predicted sites for 77% of the substrates, while the accuracy of biotransformation experts' prediction was 65%. Application of IMPACTS to external sets and comparison of

its accuracy to those of eleven other methods further validated the method implemented in IMPACTS.

5.2 Introduction

The cytochrome P450s (CYPs) are a group of heme-containing enzymes involved in the metabolism of various xenobiotics and endogenous compounds. In particular, they are involved in the phase-I metabolism of most drugs currently on the market. A majority of these biotransformations are carried out by only 5 isoforms out of the 57 P450s in the human genome, namely CYP1A2, 2C9, 2C19, 2D6, and 3A4.[1,2,3] CYP-mediated chemical modifications of drugs affect their pharmacokinetic properties as microsomal stability often correlates with hepatic clearance and hence with the half-life of drugs in the patients' body. In addition, the produced metabolites can themselves have a pharmacologic effect and intrinsic toxicity.[1] During the development stage, a number of chemical modifications of the lead compounds are often required to reach an acceptable pharmacokinetic profile and to produce a drug candidate. Thus, predicting the metabolic stability of drugs and the binding mode of small molecules in metabolic enzymes in a high throughput manner became a promising avenue of research. In fact, accurately predicting sites of metabolism (SoMs) and the binding mode of small molecules in metabolic enzymes could be useful to flag potential *in vitro* or *in silico* hits, help prioritize experiments, provide key insights enabling the design of more stable compounds with modulated half-life, predict metabolites that may induce toxicity (e.g., CYP1A2-mediated oxidation of aniline leads to carcinogenic metabolites[2]) or even investigate the polymorphism of CYP enzymes and their marked interindividual variability.[3] These multiple applications were the impetus for the development of computational approaches or protocols to predict P450 metabolism of small molecules,[4-6] with a great body of work from the Rydberg group.[7-9] These can be classified as ligand-based (e.g., quantitative structure-activity relationships[10], pharmacophore, quantum mechanical-derived rules,[8,11] descriptors[12]), reactivity-based (e.g., calculation of activation energies of each potential reactive centre by DFT or semi-empirical calculations such as in CypScore or fragment recognition such as in SMARTCyp[8]) and

structure-based (e.g., docking) methods.[5,13-18] A number of methods predicting SoMs have been devised but as stated in a recent perspective article,[6] *most consider a single aspect of the reaction*, as illustrated by ligand-based methods[11] which do not account for the recognition of the substrates by the CYP enzymes. In parallel, structure-based methods are often tested on a single CYP and their transferability to other CYP remains unknown.[19-20] Ultimately, it is expected that predictions would be more accurate if the method considered both CYP protein structures as well as ligand chemical reactivity. Approaches combining ligand reactivity and protein structures have been devised as illustrated by the pioneering work from Cruciani *et al*. (MetaSite)[21] and Oh and co-workers (MLite),[22] which uses a non-atomistic representation of the enzymes. However, despite these efforts, little has been reported on the significance of the predictions. We report herein our efforts towards the development of a fully automated program that combines molecular docking, ligand reactivity estimation and transition state structure modeling to predict the SoM of drugs, with a focus not only on accuracy but also on significance of these predictions. Furthermore, our predictions were compared to those made by biotransformation experts, which revealed the usefulness of such a program.

5.3 Theory and Implementation

5.3.1 Docking and P450-mediated metabolism

Accurately docking small molecules to enzymes requires high resolution protein structures. As of today, crystal structures have been solved for about one third of the human P450 isoforms (including four of the most important five listed above, with 2C19 not yet crystallized), making docking possible.[23-26] However, compared to traditional non covalent drug docking, predicting P450 substrates and SoMs adds a level of difficulty. First, enzymatic catalysis does not only depend on protein-ligand non covalent binding. After the substrate binding event, a reactive orientation of the substrate SoM with respect to the heme is required, allowing the oxidation reaction to take place. The efficiency of the biotransformation also depends on the intrinsic reactivity of the ligand reactive site. In several reported studies, putative substrates have been docked to a P450 structure of interest (either an X-ray structure or homology model[27]) using standard docking

programs, and docking poses obtained have been used to predict most likely metabolites based on distances to the heme group.[19] However, this approach does not take into account intrinsic reactivity of putative SoMs, does not model the chemical transformation, nor does it discriminate between inhibitors and substrates.

5.3.2 Docking and drug reactivity

In order to account for ligand reactivity while docking, rule-based approaches for predicting activation energies based on density functional theory (DFT) calculations, combined with docking to CYP1A2 were proposed to predict SoMs of substrates from both binding energy (based on docking scores) and intrinsic energy (obtained from a rule-based method).[28] Very recently, approaches considering docking to flexible P450[29] and ligand reactivity to predict SoMs of drugs were disclosed.[7,30] Two other main limitations in accurately modeling CYPs, is their promiscuity due to receptor flexibility,[31-32] and the presence of water molecules, which may also be important in drug/CYP binding.[29] In fact, in a recent review, Tarcsay and Keseru listed three major issues to address for accurate docking-based approaches: poor scoring, protein flexibility and presence of water molecules.[5]

5.3.3 Docking, drug reactivity and transition state

To reproduce the process of CYP-mediated metabolism and consider both the thermodynamics of the ligand/enzyme binding and the thermodynamics of the chemical transformation, we thought to combine a docking program and a transition state (TS) modeling program. Docking of TS structures has been previously reported, although the method required development of a TS for each drug prior to the actual docking.[9] In contrast to other docking studies,[19,33] our method not only investigates the protein-ligand non-covalent binding ("binding" in Figure 5.1), but also imposes a proper orientation of the substrate with respect to the heme ("orientation" in Figure 5.1) and considers the intrinsic reactivity of the ligand reactive site ("reaction" in Figure 5.1).[34] This approach should provide a more accurate prediction of the activation energy including the

distortion from optimal TS geometries. In addition, the absence of CYP-specific training throughout the development of this method and its validation on four very different CYPs assessed its transferability to other CYPs.

Although the complete CYP-mediated oxidation cycle includes several steps, the SoM is selected in the sub process shown in Figure 5.1.

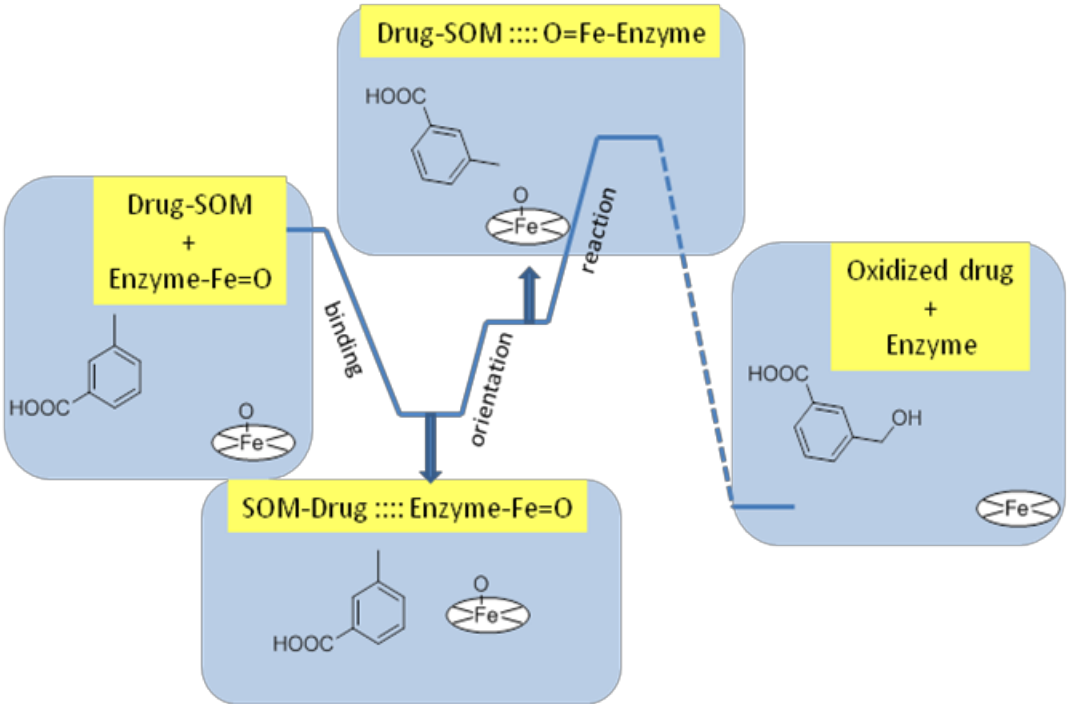


Figure 5.1. Investigated steps in the P450-mediated drug oxidation.

5.3.4 Development of IMPACTS

We have developed and implemented the framework into a SoM prediction program (IMPACTS, *In-silico* Metabolism Prediction by Activated Cytochromes and Transition States), which uses some modified routines of our FITTED docking program[35-36] to predict the CYP-mediated metabolism of small molecules (Figure 5.2). This fully automated program predicts the most likely site(s) of reaction and TS structures of small molecules when reacting with the CYP heme as the activated iron-oxygen species.

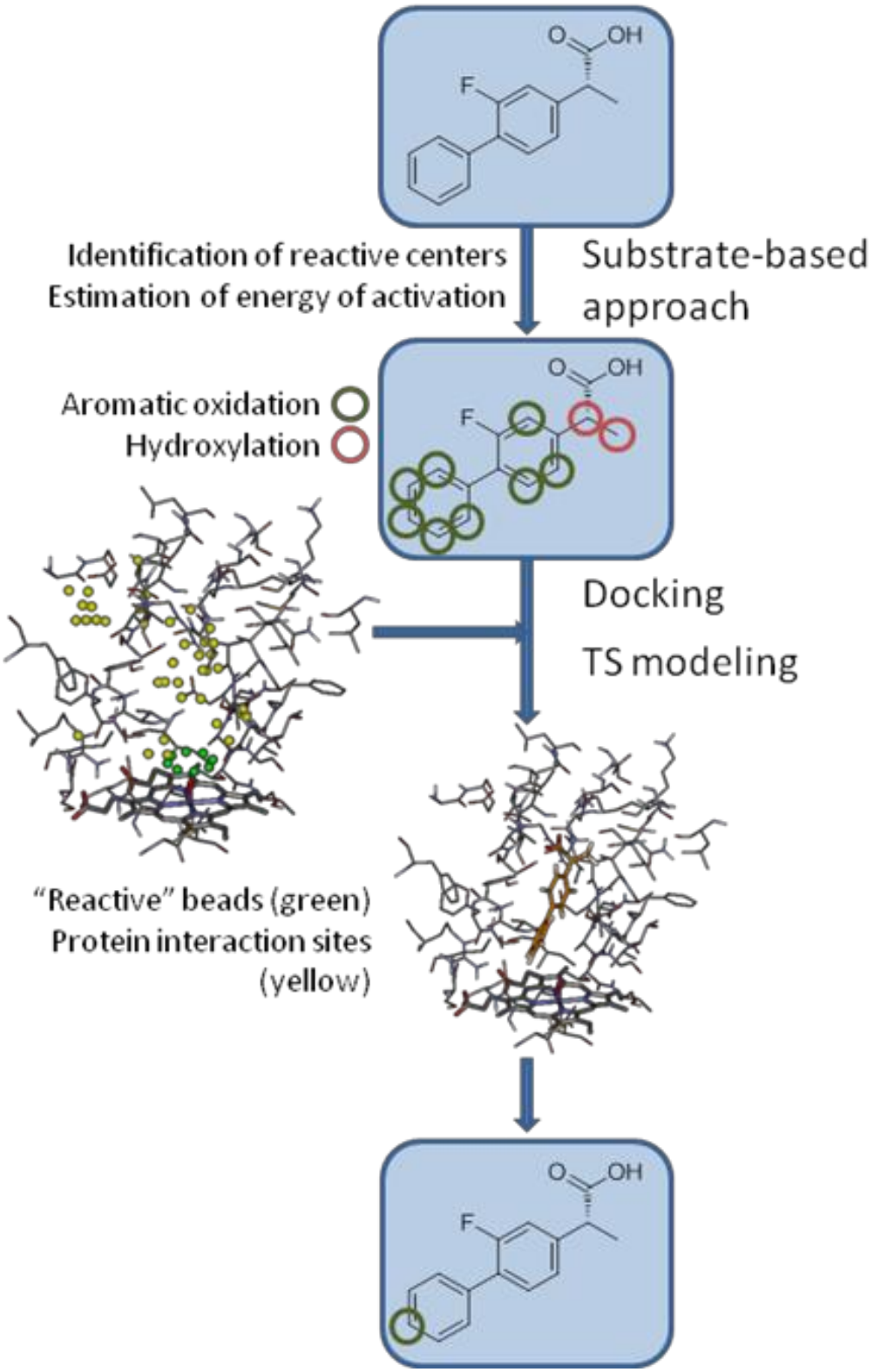


Figure 5.2. Fully automated protocol implemented in IMPACTS. User-input is 2D molecular structure and a selection of one CYP or all 4 major CYPs from a menu. Site of oxidation identified on the phenyl ring.

5.3.5 Identifying SoMs and their reactivity

First, a database of small molecular fragments and their isoform-independent corresponding reactivity (in the form of activation energies) was built. Previously computed data[11,37] was supplemented with additional DFT calculations. For instance, exhaustive DFT calculations were carried out to include the impact of various electron-withdrawing and electron donating groups on the reactivity of phenyl rings. The heme system was modeled using a simpler methoxy radical and energies relative to benzene were computed. In order to test this model, correlation with activation energies obtained using the full heme model was computed (Figure 5.3). Substituents at all positions of the benzene (except that with methoxy) were considered.

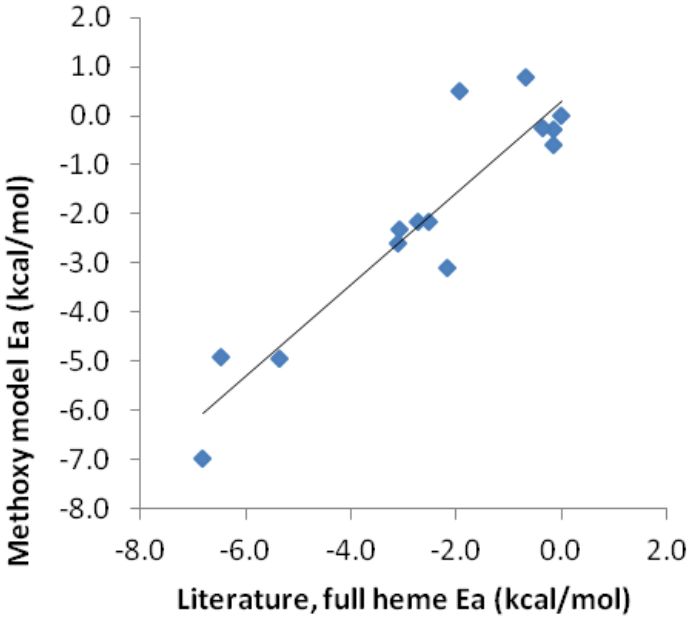


Figure 5.3. Correlation between activation energies (E_a) relative to benzene derived using the methoxy model and the full heme model.[38] r^2=0.86.

We then considered pairs of substituents. All combinations of fifteen groups were carried out and only the minimum was considered for the same positions on mirror sides. The additive effects of these groups were clearly demonstrated. For instance, calculations indicated that the presence of both *p*-OMe and *o*-Me stabilized the radical transition state by 4.5 kcal/mol relative to unsubstituted benzene while *p*-OMe and *o*-Me alone induced a stabilization of 2.4 kcal/mol and 2.3 kcal/mol respectively. Few exceptions arose with

cases of π-stacking or hydrogen bonding. These were evaluated on an individual basis and a static correction factor could be applied to these cases in order to match the additive effect observed. Interestingly, we found that the stabilizing/destabilizing effect of pairs of groups (relative to unfunctionalized benzene) equals the sum of the effect of individual groups (Figure 5.4). As a result, a simple additive rule was implemented in IMPACTS to compute the effects of pairs. The computed relative activation energies are given in Table 5.1.

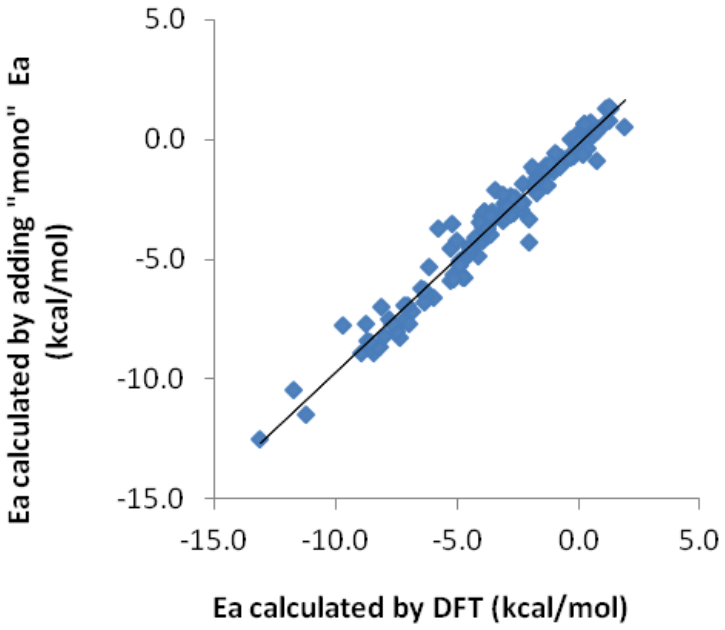
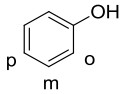
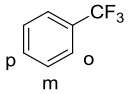
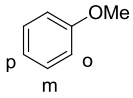
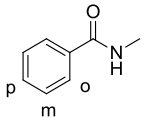
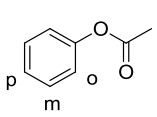
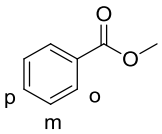
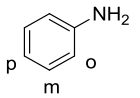
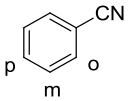
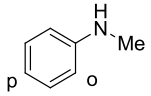
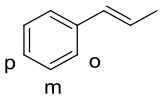
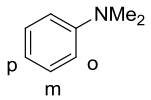
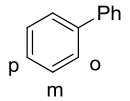
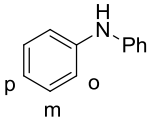
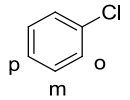
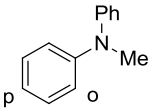
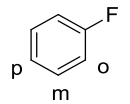
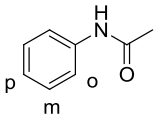
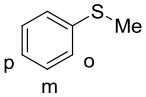
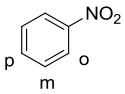
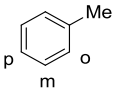
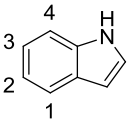


Figure 5.4. Correlation between activation energies (E_a) of bi-functionalized benzene derivatives (e.g., *m*-nitro,*p*-methoxy-benzene) relative to benzene and the sum of the relative energies of individual mono functionalized benzene derivatives (e.g., sum of *m*-nitro-benzene and *p*-methoxy benzene). r^2=0.96.

Table 5.1. Computed activation energies for functionalized benzene (aromatic oxidation). Relative energies (kcal/mol).

Fragment	Ortho	Meta	Para	Fragment	Ortho	Meta	Para
	20.9[a]	20.9[a]	20.9[a]				
	-2.25	-0.35	-2.30		0.04	0.62	0.65
	-3.35	-0.44	-2.45		0.81	0.52	-0.21
	-1.74	-0.45	-0.67		-1.27	0.40	-0.46
	-5.72	-1.72	-5.15		-0.71	0.97	-0.67
	-6.24	-0.36	-4.90		-2.49	-0.91	-3.62
	-7.33	-0.18	-5.28		-4.64	-2.02	-3.87
	-4.67	-0.87	-5.55		-0.65	0.65	-0.67
	-6.85	-1.04	-7.76		-0.96	0.34	-0.45
	-1.98	-0.13	-2.46		-2.59	0.00	-3.08
	-0.76	0.96	0.29		0.79	-0.43	1.27
	-2.29	-0.39	-0.91		1: -4.84 2: -1.69	3: -3.04 4: -4.03	

[a] See ref. [11]

We then turned our attention to hydrogen abstraction. As for the aromatic oxidation, we have found that the methoxy model correlated well with the full heme model (Figure 5.5) and enabled the computation of a large number of derivatives. However, in contrast to aromatic oxidation, the saddle point was deemed a necessary calculation for all hydrogen abstractions, and was thus included; in this case, no additive effects were observed and each necessary pair had to be computed (Table 5.2). The activation energies were computed following Eq. 5.1.

$$E_a = E_{a\ (\text{derivative, methoxy model})} - E_{a\ (\text{ethyl, methoxy model})} + E_{a\ (\text{ethyl, full heme model})} \qquad \textbf{(5.1)}$$

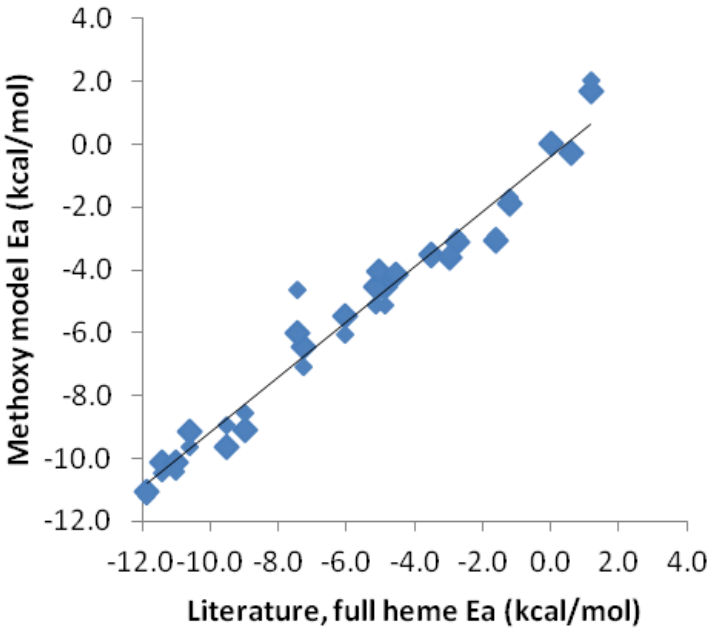
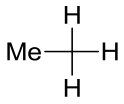
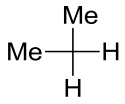
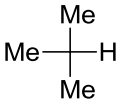
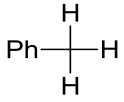
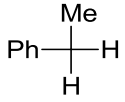
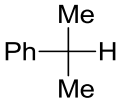
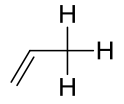
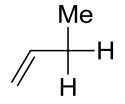
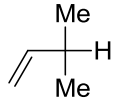
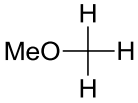
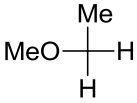
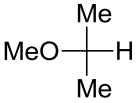
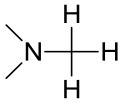
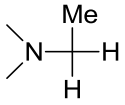
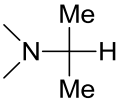
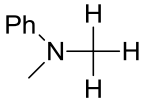
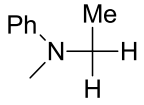
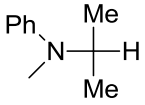
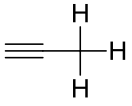
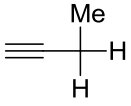
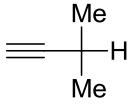
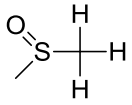
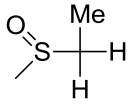
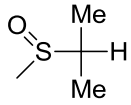
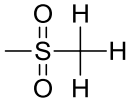
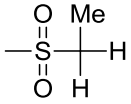
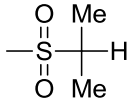
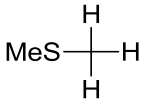
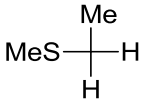
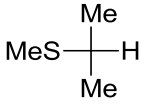
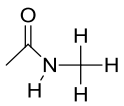
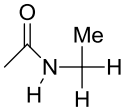
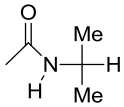
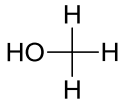
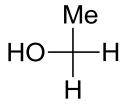
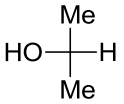
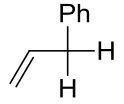
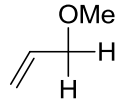
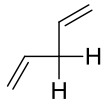
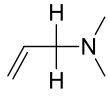


Figure 5.5. Correlation between activation energies relative to ethyl, isopropyl and tertbutyl derived using the methoxy model and the full heme model.[11,39] r^2=0.97

Table 5.2. Computed activation energies for hydrogen abstraction (kcal/mol).

Fragment	ΔΔG	Fragment	ΔΔG	Fragment	ΔΔG
	21.12[a]		18.39[a]		17.85[a]
	16.42		14.39		16.18
	16.00		14.35		15.51
	15.04		13.91		14.31
	10.32		9.63		11.64
	11.50		12.02		14.91
	16.51		14.75		14.82
	18.14		15.80		16.07
	25.53		19.37		20.44
	14.05		13.21		13.42
	15.23		16.57		16.49
	15.84		14.43		13.77
	13.37		11.81		12.56
	10.41		11.43		9.05
	11.77				

[a] See ref. [11]

Next, our program initially developed to prepare ligands for docking,[35] SMART, was modified such that it can identify all putative reactive sites and stabilizing/destabilizing neighboring groups (e.g., *p*-OMe) in a small molecule. This is achieved by comparing the substrate to a database of fragments and groups including those in Table 5.1 and Table 5.2, and assigning the corresponding activation energy. A routine identifying equivalent SoMs was also implemented into SMART (e.g., both *ortho* positions of a mono substituted phenyl, or atoms equivalent through planar symmetry).

5.3.6 IMPACTS

The activation energy values which are pre-computed by SMART are read into IMPACTS (third term in Equation 1). The drug/CYP complex is then built by docking the molecule into the CYP active site using a modified version of the hybrid matching algorithm (MA) / genetic algorithm (GA) from our docking program, FITTED.[35,40] When the MA is applied within IMPACTS, a reactive group is positioned near the P450 heme while another two pharmacophoric groups are randomly selected and placed on complementary interaction site beads located in the CYP enzyme (Figure 5.2). The energy of the system is then computed (Eq. 5.2).

In addition to docking substrates, IMPACTS models the TS as a linear combination of the ground state non-covalently bound substrate and the covalently bound substrate (first two terms in Eq. 5.2). This approach originates from the hypothesis that the TS structure, according to the Hammond-Leffler principle, is located between the reactants and products and can indeed be described as a linear combination of the products and reactants. This approach has been successfully applied in the prediction of TSs with our ACE program, which predicts the stereochemical outcome of asymmetric reactions.[41-42] As in ACE, a weighing factor (λ), allows the user to shift the TS closer to the non-covalently or covalently bound structure following the chemical transformation (first two terms in Eq. 5.2). However, in contrast to ACE, IMPACTS can handle multiple potential reactive sites simultaneously, each leading to a specific TS with its corresponding list of interactions. Herein, an interaction list is defined as a series of bonds, angles, torsions and out-of-plane terms used to describe the structure and energy of the ligand, together with

the van der Waals and electrostatic interactions between the ligand and the protein. The mechanisms of commonly observed reactions[43] have been implemented in this first version of the program, including hydrogen abstraction[38] and aromatic oxidation.[38] The code has been written such that adding a new reaction is straightforward and can be done within a few minutes by the developers as long as the reaction mechanism is known.

$$E_{TS} = (1 - \lambda)E_{substrate} + \lambda E_{product} + E_a + E_{corr} + E_{dock} \quad \textbf{(5.2)}$$

$$Score_{TS} = E_a + S_{dock} \quad \textbf{(5.3)}$$

With this method, the TS is located at the minimum of the linear combination of energy functions (Figure 5.6). However, as an artifact of the method, this minimum is associated with a non zero energy (e.g., nearly 80 kcal/mol in Figure 5.6) which is very dependent on the reaction mechanism. As this value is much lower for hydrogen abstraction than for aromatic oxidation (Figure 5.7), the relative energies of these two reactions cannot be accurately compared unless correcting terms are included (fourth term in Eq. 5.2). To compute these correcting terms, a library of small fragment molecules was built and the energy terms associated with the TSs were computed separately using a modified version of IMPACTS and kept for reference.

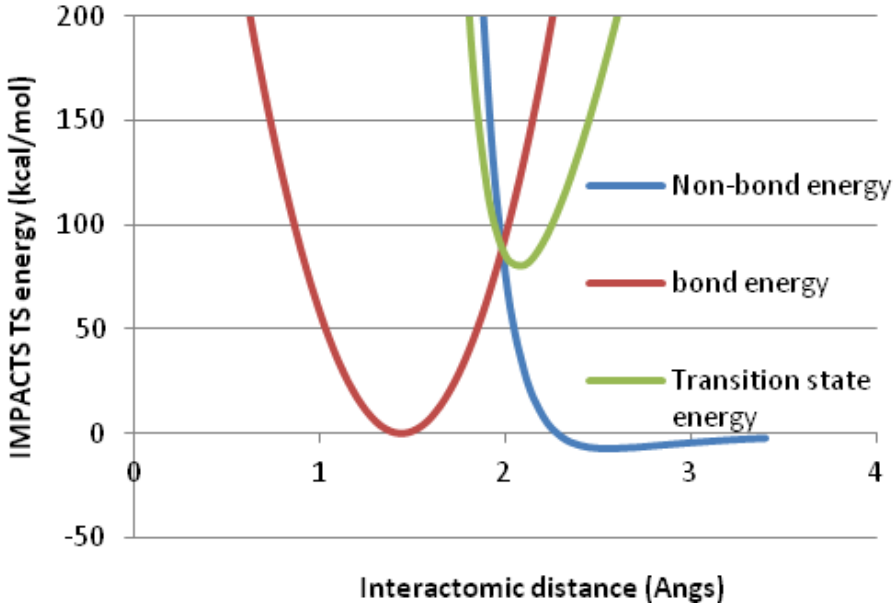


Figure 5.6. TS energy as a linear combination of bonded and non-bonded energies

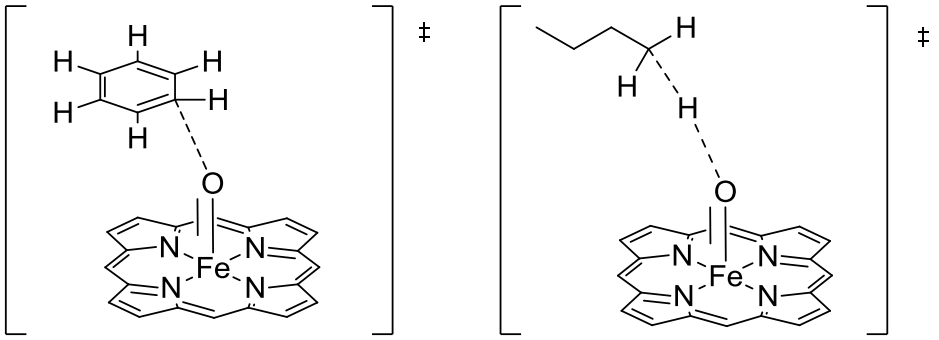


Figure 5.7. TS for aromatic oxidation and hydrogen abstraction. Side chains of the heme on the porphyrin ring are omitted for clarity.

5.3.7 Datasets

In order to measure prediction accuracy, sets of substrates of CYP1A2, CYP2C9, CYP2D6 and CYP3A4 were assembled starting from previously reported sets.[7,19,30,44] In order to evaluate not only the accuracy but also the significance of the predictions, we went back to the primary literature which revealed problems to be considered. First, data available for some compounds was conflicting. For example, the metabolism of Voriconazole by CYP2C9 and Selegiline by CYP2D6 has been observed by Hyland *et al*.[45] and Rittenbach *et al*.[46] respectively while recombinant CYP2C9 and CYP2D6 had no detectable Voriconazole oxidation activities and Selegiline oxidation activities respectively in studies from Murayama *et al*.[47] and Hidestrand *et al*.[48] In two separate reports, CYP2C9 was or not involved in the metabolism of Losartan.[49] Similarly, while CYP2C9 was found to be a major CYP for the metabolism of Sertraline in one report, it was also found to be reacting weakly in another.[50-51] Second, in several cases, drugs experimentally tested were racemic mixtures (e.g., Rosiglitazone[52]) and no data was given on the individual enantiomers. In addition, it is well known that CYP-mediated metabolism may have some level of stereoselectivity (e.g., CYP2D6 preferentially hydroxylates (L)-Trimipramine and preferentially demethylates its enantiomer[53] and (*R*)-Bufuralol is preferentially hydroxylated by CYP2D6[54]). In some cases, the wrong enantiomer was given in the published set and although this would not impact QSAR models, this would expectedly impact the apparent accuracy of any protein structure-based methods such as MetaSite or IMPACTS. Third, depending on the cells used to

express the recombinant enzymes CYP1A2, CYP2D6 and CYP3A4 (human B-lymphoblast cells or baculovirus-infected insect cells), whether they were co-expressed with NADPH-CYP oxidoreductase or not, the azelastine *N*-demethylase activity of these three CYPs varied significantly (by more than two orders of magnitude). Thus, CYP2D6 has been found to be the most reactive CYP by Nakajima *et al*.[55] while CYP1A2 was identified as the most effective by Imai *et al*.[56] Fourth, the substrate concentration is also a factor as shown with the metabolism of FLU-1[57], a metabolite of Fluvastatin. At a concentration of 1 μM, CYP3A4 was the major metabolizing enzyme and CYP2C9 did not react while CYP2C9 was the major metabolizer at a concentration of 100 μM. This same concentration dependence was observed for Sertraline.[51] In some reports, the reported data was unclear. For example, the controversy about the role of CYP3A4 in the metabolism of Ochratoxin A was mentioned[58] but not discussed and CMV423 was described as being oxidized by CYP1A2>CYP3A4>CYP2C9>CYP2D6 although CYP1A2 was less reactive. In this report, the experimental data was presented as an odd result.[59] As an explanation of this variability of results, the role of DMSO as a CYP inhibitor was mentioned by Pearce and co-workers as the formation of some metabolites of Carbazepine were not observed by others who used DMSO-containing incubation mixtures.[60]

In some cases, several metabolites had been found but only one was investigated. For example, three major metabolites for DA-8159 have been found in rats.[61] However, a single one (product of *N*-dealkylation) has been investigated with human CYP, identifying CYP3A4 as the major enzyme.[62] Although the 7-hydroxylation of Chlorpromazine was investigated in 2000[63] it is not until 2010 that the other three major metabolites were investigated.[64] Similarly, Nitrendipine and Nifedipine which are dihydropyridines were aromatized by CYP3A4. However it was mentioned that "*it must be concluded that that P4503A4 is able to oxidize other portion of some of these molecules*".[65] Zolpidem was metabolized into 3 major metabolites (M3, ca 65%, M4, ca 25%, M11, ca 10%) but the contributions of CYP1A2 (ca 8%), 2C9 (ca 31%), 2D6 (ca 17%), 2C19 (ca 2%) and 3A4 (ca 40%) in the metabolism of Zolpidem has only been reported for M3. As a result, CYP1A2 may be producing more M4 than M3. Thus the role of CYP1A2 remains unclear. Similarly, CYP-mediated metabolism of selective

estrogen receptor modulators has been investigated revealing a large number of metabolites.[66] In the case where more than 3 metabolites are reported, we considered only the major three (e.g., metabolism of Ramelteon by CYP1A2[67]). Another issue is the use of animal models (rats and mice) as it has been found that these models were not always accurate to predict the metabolism in humans (e.g., Midazolam in mice vs human[68]). Two additional points that readers should be aware of when developing such a set are given below. First, as the molecular weight of the different CYPs varies, the unit given for activity (whether pmol/min/**pmol** CYP or pmol/min/mg/**mg** CYP) is important as the perception of the role of each of the CYP may differ (see for example, Ketamine metabolism by CYP2B6, CYP3A4 and CYP2C9 using both units[69]). Second, the activity of the CYP is very sensitive to the isoform. *N*,*N*-diethyl-*m*-toluamide is metabolized by CYP2D6*1 (Val374) while CYP2D6 (Met374) does not produce any detectable activity.[70] Data from *in vivo* studies should also be taken with care. For example, the metabolism *in vitro* of Gefitinib was found to produce a number of metabolites including the desmethyl derivative. The latter is a major metabolite found in human plasma but a minor metabolite *in vitro*.[71-72]

Considering all the collected information and the various factors described above, we had to curate the retrieved datasets. As additional criteria, molecules that are too small and/or feature a single or only two potential reactive sites (e.g., butadiene monoxide) were excluded in order to avoid over-simplifying the testing set. Substrates such as Sulfinpyrazone sulphide, Capsaicin, Domperidone and 2-*n*-propyl quinoline were also removed as many metabolites are not clearly identified (several possible regioisomers on aromatic rings). Finally, duplicates were identified and removed. For example, the metabolism of Selegiline[48], Deprenyl[73] and *N*-methyl,*N*-propargylphenylethylamine[46] was investigated independently although these three names refer to the same molecule. This led to sets of 137 CYP1A2 substrates, 128 CYP2C9 substrates, 157 CYP2D6 substrates and 293 CYP3A4 substrates. These sets are supplied online and the PDB codes can be found in Appendix 3.

5.4 Results and Discussion

5.4.1 IMPACTS

In practice, IMPACTS identifies multiple potential reactive sites in a single ligand and creates interaction lists for each of these possible sites (Figure 5.8). The docking procedure then docks the substrate and, for each pose, selects the reactive site closest to the ferryl oxygen as the reactive site. The interaction list corresponding to a bond formation at this particular reactive site is then used to compute forces and potential energy values of this particular TS. As with any docking program, a score is assigned to the proposed binding mode (referred to as a pose) and can be used to rank different poses. Within IMPACTS, the scoring function is composed of the non covalent scoring function RankScore implemented in our docking program FITTED, and the activation energy (Eq. 5.3).

Among the implemented transformations[43] are hydrogen abstraction[38] leading to either hydroxylation of alkyl chains (which also represents the first step in *N*-dealkylation and *O*-dealkylation)[74] or oxidation of aldehydes into carboxylic acids,[75] oxidative deboronation as observed with Bortezomib,[76] aromatic oxidation,[38] thioketone and thiophosphate oxidation,[77] double bond epoxidation, aromatic nitrogen (e.g., pyridine) oxidation, thioether oxidation into sulfoxide, sulfoxide oxidation into sulfone and aniline oxidation. Other more unusual reactions such as P450-mediated conversion of nitriles to amides,[78] and oxidative defluorination[72] have not yet been implemented.

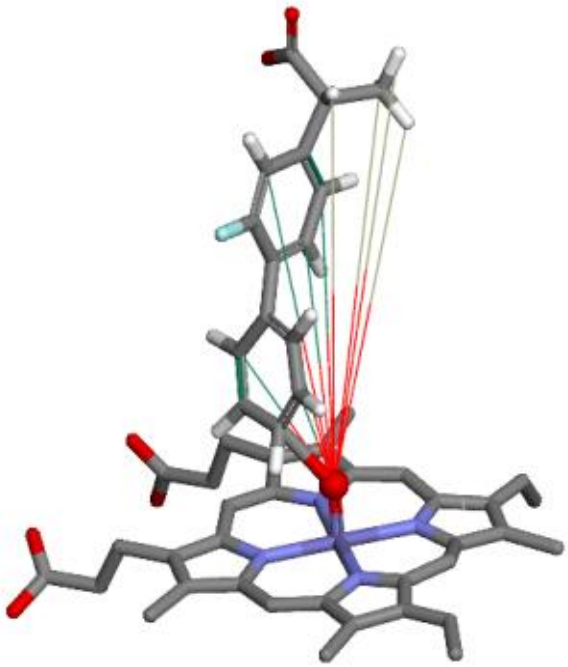


Figure 5.8. TS computed for oxidation of Flurbiprofen with CYP2C9, the various lines represent possible forming bonds (the protein is omitted for clarity).

5.4.2 Measuring accuracy

One important factor to consider in these comparisons is that experimental data will report the most predominant metabolites from incubation samples. However, metabolites that can undergo sequential reactions or are too unstable to be isolated and may be missed. As we are looking at TSs critical for the regioselective oxidation, further rearrangements of an unstable reaction intermediate are not predicted by the program in its current version. For instance, the sulphur atom of the thiophene ring of drugs[79] can first be oxidized (Figure 5.9).[80] This is rapidly followed by a rearrangement and the formally observed metabolite is oxidized on the carbon adjacent to the sulphur. An alternative thiophene oxidation of Tienilic acid and Suprofen has been proposed which goes through an epoxide formation.[81] Oxidative deboronation has also a well-defined mechanism (Figure 5.9). The empty orbital of the boronic acid (e.g., Bortezomib) plays a key role in the oxidation process. For our predictions to be deemed correct, oxidation of the sulphur atom or epoxidation of Tienilic acid and oxidation of the boron atom of

Bortezomib should be proposed as the first transition state and not the oxidation of the adjacent carbons as observed.

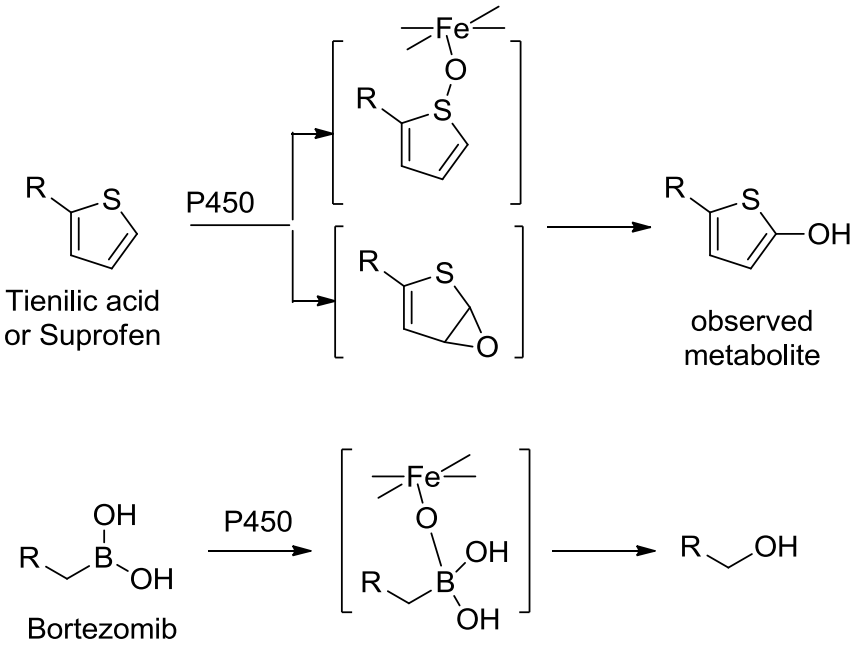


Figure 5.9. Multistep oxidation of Tienilic acid or Suprofen (top) and Bortezomib (bottom)

5.4.3 Applications to CYP1A2, CYP2D6, CYP2C9 and CYP3A4 substrates

In order to assess the accuracy of IMPACTS, testing sets of substrates of CYP1A2, 2C9, 2D6 and 3A4 were assembled. The sets Sheridan *et al*., Danielson *et al*., Vasanthanathan *et al*.[19] and from Rydberg *et al*. [7] were either downloaded or rebuilt and further curated as described in the theory and implementation section. Although CYP2C19 is another major isoform, the lack of crystal structures precluded its use in this work. The success rate of IMPACTS is shown below using the metrics previously described[30] (Table 5.3).

Predictions were made using IMPACTS on the sets of CYP1A2, CYP2C9, CYP2D6 and CYP3A4 substrates and these predictions were defined as correct when one of the top two predicted SoMs has been experimentally observed (top-2 metrics, Table 5.3).[30] A single crystal structure of ligand-bound CYP1A2 (pdb code: 2hi4) and CYP2D6 (3qm4) are available to date, while two and seven crystal structures of ligand-bound CYP2C9

(1r9o, 2og5) and CYP3A4 (1tqn, 1w0e, 1w0f, 1w0g, 2jog, 2vom, 3nxu) respectively have been reported. In order to demonstrate the significance of the predictions, the potential SoMs were identified by IMPACTS and two were randomly selected (Table 5.3). We were pleased to see that the implementation of ligand reactivity significantly increased the overall accuracy from 28% (obtained by random selection) to 60%, and docking to rigid CYPs further increased this accuracy by another 17%. As expected, the accuracy with the most promiscuous CYP3A4 was the highest with the ligand potential SoM reactivity and increased only slightly when docking was considered. This contrasted with the data obtained with the other three "more specific" enzymes which revealed that overall using only the ligand-based rules implemented in IMPACTS was not sufficient. Pharmacophores used to identify potential SoMs for substrates of these three enzymes have been reported indicating that the substrate binding orientation has a significant role in the selection of the SoMs. With these four CYPs, the accuracy is well over 70%.

Table 5.3. Accuracy[a] of IMPACTS in predicting the correct SoMs for respective datasets.

CYP	N[b]	Rand.[c]	E_a[d]	IMPACTS[e]	flexible[f]	Experts[g]	Best expert[h]
1A2	137	31	59	77	-	69 (5)	74
2C9	129	29	59	79-82	50	71 (7)	74
2D6	157	27	49	76	-	64 (4)	65
3A4	293	28	66	72-75	48	61 (6)	71
All 4	716	28	60	77	49	65 (5)	71

[a]% of molecules with an observed SoM in the predicted two SoMs. [b]Number of substrates in the set. [c]Random selection from the potential SoMs identified by IMPACTS. [d]Only the predicted reactivity of the potential SoMs is considered. [e]IMPACTS with a single crystal structure; a range is given if multiple structures were alternatively assessed; [f]IMPACTS when considering protein flexibility. [g]Average predictions by experts standard deviation in brackets. [h]Best predictions from experts.

More unexpectedly, the accuracy of the predictions was lower when protein flexibility was considered. Other reports have shown that this factor improves the accuracy although only slightly.[20,30] A close look at the predictions reveals that the failures are more ligand-dependent (these ligands provide low accuracy regardless of the protein structure used) rather than protein structure-dependent (the failures are similar regardless of the CYP structure used). Thus, we believe that considering protein

flexibility added noise to the calculations rather than improving the protein/substrate modeling. This data also demonstrated that either the promiscuity of the investigated CYPs is a great challenge or that the method as implemented has reached the limit of accuracy of the current potential energy function. We noted that most of the observed SoMs were found in the top 4 suggested (> 90%). The other poorly predicted 10% included large and/or highly flexible molecules which are known to be problematic with docking programs or may require significant conformational changes in the protein.

5.4.4 Experts' predictions

Despite the many reports on methods for SoM prediction,[6] none has specifically questioned the usefulness (accuracy and user-friendliness) of the current methods in the context of drug design, medicinal chemistry and metabolism studies, with, to the best of our knowledge, a single study reporting the prediction of a single biotransformation expert on two medium-sized (N=39, 82) sets.[82] Are these methods accurate enough to be useful? To address this critical issue, we challenged four medicinal chemists and two biotransformation experts each with over ten years of experience. Although these six experimentalists may not be representative of the medicinal chemistry and biotransformation communities, the collected data (Table 5.3) is indicative of what could be considered an accurate and useful method. A web site has been set up to enable these experts to record their predictions on the four sets of substrates (in 2D), i.e., the same input given to IMPACTS. Their predictions were consistently lower than those by IMPACTS by at at least 7%. The accuracy of the predictions of one of the experts were overall closer to that of IMPACTS, although still overall lower by more than 5%.

Interestingly, random selection provides an accurate prediction for as many as one fourth of the substrates with the criteria used in Table 5.3. This accuracy rose to nearly 40% when the top-3 metrics was used.

Overall, this data revealed that our program will be useful for predicting SoMs of small molecules or even libraries of small molecules and for producing three dimensional structures of the TSs of the small molecule/CYP complexes (Figure 5.10).

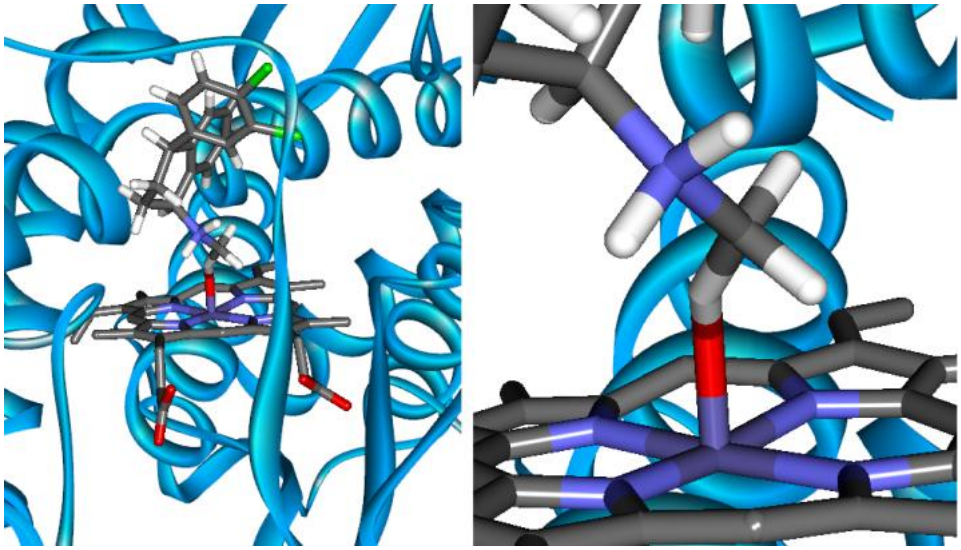


Figure 5.10. *N*-demethylation of Sertraline by CYP2C9.

5.4.5 IMPACTS's performance

In order to assess the accuracy of this first version of IMPACTS, we have retrieved sets of CYP substrates developed by Zaretzki *et al*.[83] and by Afzelius *et al*. [82] These sets were prepared and submitted to IMPACTS, and the predictions were compared to those reported using eleven other methods. These methods include academic (e.g., SMARTCyp) and commercial programs (e.g., StarDrop), ligand-based, structure-based and hybrid (i. e., both ligand and protein structure-based, MetaSite) methods. RS-Predictor makes use of ligand descriptors. A more thorough description of these methods and their use can be found in Zaretzki *et al*. and Afzelius *et al*.'s original publications.[82-83] As was done by Afzelius *et al*., large and highly flexible substrates were not considered by our docking-based method IMPACTS. This pre-selection did not affect much the size of the CYP1A2, CYP2D6 and CYP2C9 substrates sets (one to four molecules per set, less than 2% of these substrates), but did reduce the number of CYP3A4 substrates by 8% as shown in Table 5.4. As the sets are slightly different from one program to the next and as updated and/or improved versions of these methods may have been released since their use by Afzelius *et al*. and Zaretzki *et al*., this comparison should be considered with care and is only used to illustrate the overall performance of IMPACTS, and not to provide a ranking of methods. As can be seen in Table 5.4, this first version of IMPACTS stands well this

comparison with accuracy equivalent to RS-Predictor and superior to all the other methods with CYP2C9 substrates. The accuracy with CYP1A2 substrates is also comparable to that obtained with RS-Predictor. However, IMPACTS was slightly less accurate than other methods for CYP2D6 and CYP3A4 substrates when looking at average accuracy.

Table 5.4. Accuracy of IMPACTS and eleven other methods in predicting the correct SoMs for external datasets.

Isozyme	1A2	2C9	2C9	2C9	2D6	2D6	3A4	3A4	3A4
N[a]	271[c]	98[c]	128[c]	49[d]	134[c]	136[c]	321[c]	154[c]	65[d]
N[b]	269[c]	96[c]	127[c]	45[d]	133[c]	132[c]	295[c]	144[c]	58[d]
IMPACTS	80.5	84.4	76.4	81.8	70.7	71.2	73.2	70.1	82.5
RS-Predictor[83]	83.0	81.6	79.7	-	85.8	79.4	81.9	79.2	-
SMARTCyp[83]	-	67.7	66.9	-	48.5	68.1	73.1	77.2	-
StarDrop[83]	-	77.4	78.4	-	81.5	69.2	77.5	66.9	-
Schrödinger[83]	-	69.6	74.0	-	66.2	70.1	80.2	68.2	-
Sheridan *et al*.[44,83]	-	72.4	-	-	71.9	-	77.4	-	-
MetaSite[44,83]	-	68.8	-	91	65.4	-	61.8	-	87
MetaDock[82]	-	-	-	66	-	-	-	-	67
QMBO[82]	-	-	-	84	-	-	-	-	84
QMSpin[82]	-	-	-	78	-	-	-	-	78
MetaGlide[82]	-	-	-	67	-	-	-	-	65
SporCalc[82]	-	-	-	81	-	-	-	-	81

[a] Number of substrates used by Zaretski *et al*. and Afzelius *et al*. [b] Number for substrates when large and flexible substrates removed following Afzelius *et al*. [c] Sets from Zaretzki *et al*., top-2 metrics was used.[83] [d] Sets from Afzelius *et al*., top-3 metrics was used.[82]

In contrast to ligand-based methods, docking-based methods provide information on the binding mode of the substrates in the active site of CYPs. However, their major drawback[84] is the CPU time required to make predictions. In general, ligand-based predictions can be made within a second or less per compound. Analysis of the data generated to produce Table 5.4 revealed that 77% of the runs took 2 minutes or less per

run and 97% took ten minutes or less. 0.8% of the compounds, mostly 3A4 substrates, were not docked after an hour.

5.5 Conclusion

In conclusion, we have developed a preliminary version of a fully automated program, IMPACTS, for the prediction of the SoMs of drugs. Moreover, IMPACTS provides a 3D picture of the TS of the drug at the active site of cytochromes (Figure 5.10). In addition, this method does not require any training and should be applicable to other CYPs.[6] Knowledge about both the binding mode and the SoM may enable medicinal chemists to design modifications at locations other than the SoMs to disturb or enhance the substrate-CYP recognition. In contrast to other methods, ligand reactivity, binding affinity and proper geometry of the TS are all considered in a single and fully automated run. Other limitations in modeling CYPs include the presence of water molecules, which may also be important in drug/CYP binding.[5,29] Strategies to overcome these challenges and improvements of the current activation energies are currently being investigated in the hope of further increasing the accuracy of IMPACTS. Finally, a comparison between the predictions generated by IMPACTS, those by random selection, those made by experts and those made using another eleven methods demonstrated that IMPACTS can be extremely useful.

5.6 Experimental

5.6.1 Construction of the testing sets

The testing sets were built with care to reduce the noise in the prediction assessment. Here is a list of selection criteria that were used; a detailed discussion is provided in the theory and implementation section. 1. Small molecules such as butadiene monoxide with only one possible reactive site (i.e., a single double bond) have been removed from existing sets as they increase the apparent accuracy but reduce the significance of the predictions; 2. Eicosapentaenoic acid, Sulfinpyrazone sulphide, Capsaicin, Domperidone and 2-*n*-propyl quinoline were also removed as many metabolites are not clearly

identified (several possible regioisomers on aromatic rings). 3. Duplicates were identified and removed. An asterisk was then added to the atoms being oxidized in order to eventually identify whether the predicted SoM is one of the observed SoMs. Sets are given in separate files in mol2 format.

5.6.2 Computation of the activation energies

All the quantum mechanical calculations were performed using DFT, more specifically the B3LYP functional (Unrestricted Hartree-Fock) and the 6-31G* basis-set. All calculations were performed in vacuum. The B3LYP calculations were performed using GAMESS-US v.Aug2011-64bit.

As the reacting methoxy group is not aligned with the benzene ring, the transition state is not perfectly symmetrical. Thus, for the monosubstituted benzene derivatives, the two *meta* and *ortho* positions are not equivalent. The lowest-in-energy of the two *meta* positions and highest-in-energy of the two optimized *ortho* positions were kept for each one. The highest-in-energy of the two minima was regarded as correct for *ortho* position as the difference in energy was often due to hydrogen-bonding effects which are artefacts of the methoxy models. In fact, hydrogen bonds would be limited within the heme system (steric block). In parallel, sterics are the driving force at the *meta* position and thus the lowest-in-energy was regarded as correct

5.6.3 Application of IMPACTS

Default parameters implemented in IMPACTS have been used. IMPACTS has been integrated into our platform FORECASTER[85] for user-friendliness. The user can draw the substrate into a 2D sketcher and select the CYP enzyme with which to predict the SoM. FORECASTER will take care of adding hydrogens, generating a 3D structure and selecting the correct CYP files. IMPACTS and FORECASTER are accessible free of charge to academic users (www.fitted.ca).

5.7 Acknowledgements

We thank AstraZeneca R&D Montréal and NSERC for financial support, Dr. Warner (AstraZeneca) for fruitful discussions and CIHR for a fellowship to VCS (DD training program). We are also grateful to Dr. Raeppel (ChemRF Laboratories), Drs. Projean and Griffin (AstraZeneca), Dr. Ramirez-Molina (GlaxoSmithKline), and Dr. Giroux (Vertex Pharmaceuticals) for their contribution to experts' predictions. Calcul Québec and Compute Canada are acknowledged for generous CPU allocations.

5.8 References

1. Wong, Y. C.; Qian, S.; Zuo, Z., Regioselective biotransformation of CNS drugs and its clinical impact on adverse drug reactions. *Exp*. *Opin*. *Drug Metab*. *Toxicol*. **2012,** *8* (7), 833-854.

2. Shamovsky, I.; Ripa, L.; Börjesson, L.; Mee, C.; Nordén, B.; Hansen, P.; Hasselgren, C.; O'Donovan, M.; Sjö, P., Explanation for Main Features of Structure–Genotoxicity Relationships of Aromatic Amines by Theoretical Studies of Their Activation Pathways in CYP1A2. *J*. *Am*. *Chem*. *Soc*. **2011,** *133* (40), 16168-16185.

3. He, S. M.; Zhou, Z. W.; Li, X. T.; Zhou, S. F., Clinical drugs undergoing polymorphic metabolism by human cytochrome P450 2C9 and the implication in drug development. *Curr*. *Med*. *Chem*. **2011,** *18* (5), 667-713.

4. Zhang, T.; Chen, Q.; Li, L.; Liu, L. A.; Wei, D. Q., In silico prediction of cytochrome P450-mediated drug metabolism. *Comb*. *Chem*. *High Throughput Screening* **2011,** *14* (5), 388-395.

5. Tarcsay, A.; Keserü, G. M., In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Exp*. *Opin*. *Drug Metab*. *Toxicol*. **2011,** *7* (3), 299-312.

6. Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C., Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J*. *Chem*. *Inf*. *Model*. **2012,** *52*, 617-648.

7. Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L., Fast prediction of cytochrome p450 mediated drug metabolism. *ChemMedChem* **2009,** *4* (12), 2070-2079.

8. Rydberg, P.; Gloriam, D. E.; Olsen, L., The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* **2010,** *26* (23), 2988-2989.

9. Rydberg, P.; Hansen, S. M.; Kongsted, J.; Norrby, P.-O.; Olsen, L.; Ryde, U., Transition-State Docking of Flunitrazepam and Progesterone in Cytochrome P450. *J*. *Chem*. *Theory Comput*. **2008,** *4* (4), 673-681.

10. Saraceno, M.; Massarelli, I.; Imbriani, M.; James, T. L.; Bianucci, A. M., Optimizing QSAR Models for Predicting Ligand Binding to the Drug-Metabolizing Cytochrome P450 Isoenzyme CYP2D6. *Chem*. *Biol*. *Drug Des*. **2011,** *78*, 236-251.

11. Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L., SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med*. *Chem*. *Lett*. **2010,** *1* (3), 96-100.

12. Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-w.; Bennett, K. P.; Breneman, C. M., RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J*. *Chem*. *Inf*. *Model*. **2011,** *51* (7), 1667-1689.

13. Pelkonen, O.; Turpeinen, M.; Raunio, H., In vivo-in vitro-in silico pharmacokinetic modelling in drug development: Current status and future directions. *Clin*. *Pharmacokin*. **2011,** *50* (8), 483-491.

14. Czodrowski, P.; Kriegl, J. M.; Scheuerer, S.; Fox, T., Computational approaches to predict drug metabolism. *Exp*. *Opin*. *Drug Metab*. *Toxicol*. **2009,** *5* (1), 15-27.

15. De Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E., Binding mode prediction of cytochrome P450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J*. *Med*. *Chem*. **2005,** *48* (7), 2308-2318.

16. Stjernschantz, E.; Vermeulen, N. P. E.; Oostenbrink, C., Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Exp*. *Opin*. *Drug Metab*. *Toxicol*. **2008,** *4* (5), 513-527.

17. Vaz, R. J.; Zamora, I.; Li, Y.; Reiling, S.; Shen, J.; Cruciani, G., The challenges of in silico contributions to drug metabolism in lead optimization. *Exp*. *Opin*. *Drug Metab*. *Toxicol*. **2010,** *6* (7), 851-861.

18. Sun, H.; Scott, D. O., Structure-based drug metabolism predictions for drug design. *Chem*. *Biol*. *Drug Des*. **2010,** *75* (1), 3-17.

19. Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Jorgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C., Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (1), 43-52.

20. Moors, S. L. C.; Vos, A. M.; Cummings, M. D.; Van Vlijmen, H.; Ceulemans, A., Structure-Based Site of Metabolism Prediction for Cytochrome P450 2D6. *J*. *Med*. *Chem*. **2011,** *54*, 6098-6105.

21. Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R., MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J*. *Med*. *Chem*. **2005,** *48* (22), 6970-6979.

22. Oh, W. S.; Kim, D. N.; Jung, J.; Cho, K. H.; No, K. T., New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. *J*. *Chem*. *Inf*. *Model*. **2008,** *48* (3), 591-601.

23. Williams, P. A.; Cosme, J.; Ward, A.; Angove, H. C.; Vinkovi, D. M.; Jhoti, H., Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature* **2003,** *424* (6947), 464-468.

24. Williams, P. A.; Cosme, J.; Matak Vinkovi, D.; Ward, A.; Angove, H. C.; Day, P. J.; Vonrhein, C.; Tickle, I. J.; Jhoti, H., Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **2004,** *305* (5684), 683-686.

25. Wester, M. R.; Yano, J. K.; Schoch, G. A.; Yang, C.; Griffin, K. J.; Stout, C. D.; Johnson, E. F., The structure of human cytochrome P450 2C9 complexed with flurbiprofen at 2.0-Å resolution. *J*. *Biol*. *Chem*. **2004,** *279* (34), 35630-35637.

26. Coleman, S.; Linderman, R.; Hodgson, E.; Rose, R. L., Comparative metabolism of chloroacetamide herbicides and selected metabolites in human and rat liver microsomes. *Environ*. *Health Persp*. **2000,** *108* (12), 1151-1157.

27. Kjellander, B.; Masimirembwa, C. M.; Zamora, I., Exploration of enzyme-ligand interactions in CYP2D6 & 3A4 homology models and crystal structures using a novel computational approach. *J*. *Chem*. *Inf*. *Model*. **2007,** *47* (3), 1234-1247.

28. Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L., Fast Prediction of Cytochrome P450 Mediated Drug Metabolism. *ChemMedChem* **2009,** *4*, 2070-2079.

29. Hritz, J.; de Ruiter, A.; Oostenbrink, C., Impact of Plasticity and Flexibility on Docking Results for Cytochrome P450 2D6: A Combined Approach of Molecular Dynamics and Ligand Docking. *J*. *Med*. *Chem*. **2008,** *51* (23), 7469-7477.

30. Danielson, M. L.; Desai, P. V.; Mohutsky, M. A.; Wrighton, S. A.; Lill, M. A., Potentially increasing the metabolic stability of drug candidates via computational site of metabolism prediction by CYP2C9: The utility of incorporating protein flexibility via an ensemble of structures. *Eur*. *J*. *Med*. *Chem*. **2011,** *46*, 3953-3963.

31. Ekroos, M.; Sjögren, T., Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. **2006,** *103* (37), 13682-13687.

32. Guengerich, F. P., A malleable catalyst dominates the metabolism of drugs. *Proc*. *Natl Acad*. *Sci*. *USA* **2006,** *103* (37), 13565-13566.

33. Ito, Y.; Kondo, H.; Goldfarb, P. S.; Lewis, D. F. V., Analysis of CYP2D6 substrate interactions by computational methods. *Journal of Molecular Graphics and Modelling* **2008,** *26* (6), 947-956.

34. Zhu, Y.; Silverman, R. B., Revisiting Heme Mechanisms. A Perspective on the Mechanisms of Nitric Oxide Synthase (NOS), Heme Oxygenase (HO), and Cytochrome P450s (CYP450s). *Biochemistry* **2008,** *47* (8), 2231-2243.

35. Corbeil, C. R.; Englebienne, P.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J*. *Chem*. *Inf*. *Model*. **2007,** *47* (2), 435-449.

36. Corbeil, C. R.; Moitessier, N., Docking Ligands into Flexible and Solvated Macromolecules. 3. Impact of Input Ligand Conformation, Protein Flexibility, and Water Molecules on the Accuracy of Docking Programs. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (4), 997-1009

37. Bathelt, C. M.; Ridder, L.; Mulholland, A. J.; Harvey, J. N., Mechanism and structure-reactivity relationships for aromatic hydroxylation by cytochrome P450. *Org*. *Biomol*. *Chem*. **2004,** *2* (20), 2998-3005.

38. Rydberg, P.; Ryde, U.; Olsen, L., Prediction of Activation Energies for Aromatic Oxidation by Cytochrome P450. *J*. *Phys*. *Chem*. *A* **2008,** *112* (50), 13058-13065.

39. Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U., Prediction of Activation Energies for Hydrogen Abstraction by Cytochrome P450. *Journal of Medicinal Chemistry* **2006,** *49* (22), 6489-6499.

40. Corbeil, C. R.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (4), 997-1009.

41. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development and application of a rapid and accurate program based on organic principles. *Angew*. *Chem*. *Int*. *Ed*. **2008,** *47* (14), 2635-2638.

42. Weill, N.; Corbeil, C. R.; De Schutter, J. W.; N., M., Toward a computational tool predicting the stereochemical outcome of asymmetric reactions: Development of the molecular mechanics-based program ACE and application to asymmetric epoxidation reactions. *J*. *Comput*. *Chem*. **2011,** *32* (13), 2878-2889.

43. Guengerich, F. P.; Isin, E. M., Mechanisms of cytochrome P450 reactions. *Acta Chim*. *Slov*. **2008,** *55* (1), 7-19.

44. Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J., Empirical Regioselectivity Models for Human Cytochromes P450 3A4, 2D6, and 2C9. *J*. *Med*. *Chem*. **2007,** *50* (14), 3173-3184.

45. Hyland, R.; Jones, B. C.; Smith, D. A., Identification of the Cytochrome P450 Enzymes Involved in the N-Oxidation of Voriconazole. *Drug Metab*. *Disp*. **2003,** *31* (5), 540-547.

46. Rittenbach, K. A.; Holt, A.; Ling, L.; Shan, J.; Baker, G. B., Metabolism of N-methyl, N-propargylphenylethylamine: Studies with human liver microsomes and cDNA expressed cytochrome P450 (CYP) enzymes. *Cell*. *Mol*. *Neurobiol*. **2007,** *27* (2), 179-190.

47. Murayama, N.; Imai, N.; Nakane, T.; Shimizu, M.; Yamazaki, H., Roles of CYP3A4 and CYP2C19 in methyl hydroxylated and N-oxidized metabolite formation from voriconazole, a new anti-fungal agent, in human liver microsomes. *Biochem*. *Pharm*. **2007,** *73* (12), 2020-2026.

48. Hidestrand, M.; Oscarson, M.; Salonen, J. S.; Nyman, L.; Pelkonen, O.; Turpeinen, M.; Ingelman-Sundberg, M., CYP2B6 and CYP2C19 as the Major Enzymes Responsible for the Metabolism of Selegiline, a Drug Used in the Treatment of Parkinson's Disease, as Revealed from Experiments with Recombinant Enzymes. *Drug Metab*. *Disp*. **2001,** *29* (11), 1480-1484.

49. Yun, C. H.; Lee, H. S.; Lee, H.; Rho, J. K.; Jeong, H. G.; Guengerich, F. P., Oxidation of the angiotensin II receptor antagonist losartan (DuP 753) in human liver microsomes. Role of cytochrome P4503A(4) in formation of the active metabolite EXP3174. *Drug Metab*. *Disp*. **1995,** *23* (2), 285-289.

50. Kobayashi, K.; Ishizuka, T.; Shimada, N.; Yoshimura, Y.; Kamijima, K.; Chiba, K., Sertraline N-Demethylation Is Catalyzed by Multiple Isoforms of Human Cytochrome P-450 In Vitro. *Drug Metab*. *Disp*. **1999,** *27* (7), 763-766.

51. Obach, R. S.; Cox, L. M.; Tremaine, L. M., Sertraline is metabolized by multiple cytochrome P450 enzymes, monoamine oxidases, and glucuronyl transferases in human: An in vitro study. *Drug*. *Metab*. *Disp*. **2005,** *33* (2), 262-270.

52. Baldwin, S. J.; Clarke, S. E.; Chenery, R. J., Characterization of the cytochrome P450 enzymes involved in the in vitro metabolism of rosiglitazone. *Br*. *J*. *Clin*. *Pharmacol*. **1999,** *48* (3), 424-432.

53. Eap, C. B.; Bender, S.; Gastpar, M.; Fischer, W.; Haarmann, C.; Powell, K.; Jonzier-Perey, M.; Cochard, N.; Baumann, P., Steady state plasma levels of the enantiomers

of trimipramine and of its metabolites in CYP2D6-, CYP2C19- and CYP3A4/5-phenotyped patients. *Ther*. *Drug Monit*. **2000,** *22* (2), 209-214.

54. Narimatsu, S.; Takemi, C.; Tsuzuki, D.; Kataoka, H.; Yamamoto, S.; Shimada, N.; Suzuki, S.; Satoh, T.; Meyer, U. A.; Gonzalez, F. J., Stereoselective Metabolism of Bufuralol Racemate and Enantiomers in Human Liver Microsomes. *J*. *Pharmacol*. *Exp*. *Ther*. **2002,** *303* (1), 172-178.

55. Nakajima, M.; Nakamura, S.; Tokudome, S.; Shimada, N.; Yamazaki, H.; Yokoi, T., Azelastine N-Demethylation by Cytochrome P-450 (CYP)3A4, CYP2D6, and CYP1A2 in Human Liver Microsomes: Evaluation of Approach to Predict the Contribution of Multiple CYPs. *Drug Metab*. *Disp*. **1999,** *27* (12), 1381-1391.

56. Imai, T.; Taketani, M.; Suzu, T.; Kusube, K.; Otagiri, M., In Vitro Identification of the Human Cytochrome P-450 Enzymes Involved in the N-Demethylation of Azelastine. *Drug Metab*. *Disp*. **1999,** *27* (8), 942-946.

57. Goda, R.; Nagai, D.; Akiyama, Y.; Nishikawa, K.; Ikemoto, I.; Aizawa, Y.; Nagata, K.; Yamazoe, Y., Detection of a new N-oxidized metabolite of flutamide, N-[4-nitro-3- (trifluoromethyl)phenyl]hydroxylamine, in human liver microsomes and urine of prostate cancer patients. *Drug*. *Metab*. *Disp*. **2006,** *34* (5), 828-835.

58. Simarro Doorten, A. Y.; Bull, S.; Van Der Doelen, M. A. M.; Fink-Gremmels, J., Metabolism-mediated cytotoxicity of ochratoxin A. *Toxicol Vitro* **2004,** *18* (3), 271-277.

59. Bournique, B.; Lambert, N.; Boukaiba, R.; Martinet, M., In vitro metabolism and drug interaction potential of a new highly potent anti-cytomegalovirus molecule, CMV423 (2-chloro 3-pyridine 3-yl 5,6,7,8-tetrahydroindolizine 1-carboxamide). *Br*. *J*. *Clin*. *Pharmacol*. **2001,** *52* (1), 53-63.

60. Pearce, R. E.; Vakkalagadda, G. R.; Steven Leeder, J., Pathways of carbamazepine bioactivation in vitro I. Characterization of human cytochromes P450 responsible for the formation of 2- and 3-hydroxylated metabolites. *Drug Metab*. *Disp*. **2002,** *30* (11), 1170-1179.

61. Choi, S. J.; Ji, H. Y.; Lee, H. Y.; Kim, D. S.; Kim, W. B.; Lee, H. S., In vitro metabolism of a novel phosphodiesterase-5 inhibitor DA-8159 in rat liver preparations using liquid chromatography/electrospray mass spectrometry. *Biomed*. *Chrom*. **2002,** *16* (6), 395-399.

62. Ji, H. Y.; Lee, H. W.; Kim, H. H.; Kim, D. S.; Yoo, M.; Kim, W. B.; Lee, H. S., Role of human cytochrome P450 3A4 in the metabolism of DA–8159, a new erectogenic#. *Xenobiotica* **2004,** *34* (11-12), 973-982.

63. Yoshii, K.; Kobayashi, K.; Tsumuji, M.; Tani, M.; Shimada, N.; Chiba, K., Identification of human cytochrome P450 isoforms involved in the 7- hydroxylation of chlorpromazine by human liver microsomes. *Life Sci*. **2000,** *67* (2), 175-184.

64. Wójcikowski, J.; Boksa, J.; Daniel, W. A., Main contribution of the cytochrome P450 isoenzyme 1A2 (CYP1A2) to N-demethylation and 5-sulfoxidation of the phenothiazine neuroleptic chlorpromazine in human liver—A comparison with other phenothiazines. *Biochem*. *Pharmacol*. **2010** *80*, 1252-1259.

65. Guengerich, F. P.; Brian, W. R.; Iwasaki, M.; Sari, M. A.; Baeaernhielm, C.; Berntsson, P., Oxidation of dihydropyridine calcium channel blockers and analogs by human liver cytochrome P-450 IIIA4. *J*. *Med*. *Chem*. **1991,** *34* (6), 1838-1844.

66. Zhang, Z.; Chen, Q.; Li, Y.; Doss, G. A.; Dean, B. J.; Ngui, J. S.; Silva Elipe, M.; Kim, S.; Wu, J. Y.; DiNinno, F.; Hammond, M. L.; Stearns, R. A.; Evans, D. C.; Baillie, T. A.; Tang, W., In Vitro Bioactivation of Dihydrobenzoxathiin Selective Estrogen Receptor Modulators by Cytochrome P450 3A4 in Human Liver Microsomes: Formation of Reactive Iminium and Quinone Type Metabolites. *Chem*. *Res*. *Toxicol*. **2005,** *18* (4), 675-685.

67. Obach, R. S.; Ryder, T. F., Metabolism of Ramelteon in human liver microsomes and correlation with the effect of fluvoxamine on ramelteon pharmacokinetics. *Drug Metab*. *Disp*. **2010,** *38* (8), 1381-1391.

68. Perloff, M. D.; von Moltke, L. L.; Court, M. H.; Kotegawa, T.; Shader, R. I.; Greenblatt, D. J., Midazolam and Triazolam Biotransformation in Mouse and Human

Liver Microsomes: Relative Contribution of CYP3A and CYP2C Isoforms. *J*. *Pharmacol*. *Exp*. *Ther*. **2000,** *292* (2), 618-628.

69. Hijazi, Y.; Boulieu, R., Contribution of CYP3A4, CYP2B6, and CYP2C9 Isoforms toN-Demethylation of Ketamine in Human Liver Microsomes. *Drug Metab*. *Disp*. **2002,** *30* (7), 853-858.

70. Usmani, K. A.; Rose, R. L.; Goldstein, J. A.; Taylor, W. G.; Brimfield, A. A.; Hodgson, E., In Vitro Human Metabolism and Interactions of Repellent N,N-Diethyl-m-Toluamide. *Drug Metab*. *Disp*. **2002,** *30* (3), 289-294.

71. McKillop, D.; McCormick, A. D.; Millar, A.; Miles, G. S.; Phillips, P. J.; Hutchison, M., Cytochrome P450-dependent metabolism of gefitinib. *Xenobiotica* **2005,** *35* (1), 39-50.

72. Mckillop, D.; Mccormick, A. D.; Miles, G. S.; Phillips, P. J.; Pickup, K. J.; Bushby, N.; Hutchison, M., In vitro metabolism of gefitinib in human liver microsomes. *Xenobiotica* **2004,** *34* (11-12), 983-1000.

73. Grace, J. M.; Kinter, M. T.; Macdonald, T. L., Atypical Metabolism of Deprenyl and Its Enantiomer, (S)-(+)-N,.alpha.-Dimethyl-N-Propynylphenethylamine, by Cytochrome P450 2D6. *Chem*. *Res*. *Toxicol*. **1994,** *7*, 286-290.

74. Rydberg, P.; Ryde, U.; Olsen, L., Sulfoxide, Sulfur, and Nitrogen Oxidation and Dealkylation by Cytochrome P450. *J*. *Chem*. *Theory Comput*. **2008,** *4* (8), 1369-1377.

75. Liu, X.; Wang, Y.; Han, K., Systematic study on the mechanism of aldehyde oxidation to carboxylic acid by cytochrome P450. *J*. *Biol*. *Inorg*. *Chem*. **2007,** *12* (7), 1073-1081.

76. Larkin, J. D.; Markham, G. D.; Milkevitch, M.; Brooks, B. R.; Bock, C. W., Computational investigation of the oxidative deboronation of boroglycine, H2N-CH2-B(OH)2, using H2O and H 2O2. *J*. *Phys*. *Chem*. *A* **2009,** *113* (41), 11028-11034.

77. Jayathirtha Rao, V.; Muthuramu, K.; Ramamurthy, V., Oxidations of thioketones by singlet and triplet oxygen. *J*. *Org*. *Chem*. **1982,** *47* (1), 127-131.

78. Zhang, Z.; Li, Y.; Stearns, R. A.; Ortiz de Montellano, P. R.; Baillie, T. A.; Tang, W., Cytochrome P450 3A4-Mediated Oxidative Conversion of a Cyano to an Amide Group in the Metabolism of Pinacidil. *Biochemistry* **2002,** *41* (8), 2712-2718.

79. Taguchi, K.; Konishi, T.; Nishikawa, H.; Kitamura, S., Identification of human cytochrome P450 isoforms involved in the metabolism of S-2-[4-(3-methyl-2-thienyl)phenyl]propionic acid. *Xenobiotica* **1999,** *29* (9), 899-907.

80. Jean, P.; Lopez-Garcia, P.; Dansette, P.; Mansuy, D.; Goldstein, J. L., Oxidation of Tienilic Acid by Human Yeast-Expressed Cytochromes P-450 2C8, 2C9, 2C18 and 2C19. *Eur*. *J*. *Biochem*. **1996,** *241* (3), 797-804.

81. O'Donnell, J. P.; Dalvie, D. K.; Kalgutkar, A. S.; Obach, R. S., Mechanism-based inactivation of human recombinant P450 2C9 by the nonsteroidal anti-inflammatory drug suprofen. *Drug Metab*. *Disp*. **2003,** *31* (11), 1369-1377.

82. Afzelius, L.; Arnby, C. H.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Raubacher, F.; Weidolf, L., State-of-the-art tools for computational site of metabolism predictions: Comparative analysis, mechanistical insights, and future applications. *Drug Metabol*. *Rev*. **2007,** *39* (1), 61-86.

83. Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K. P.; Olsen, L.; Breneman, C. M., RS-predictor models augmented with SMARTCyp reactivities: Robust metabolic regioselectivity predictions for nine CYP isozymes. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (6), 1637-1659.

84. Rydberg, P., Theoretical Study of the Cytochrome P450 Mediated Metabolism of Phosphorodithioate Pesticides. *Journal of Chemical Theory and Computation* **2012**.

85. Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N., Integrating Medicinal Chemistry, Organic/Combinatorial Chemistry, and Computational Chemistry for the Discovery of Selective Estrogen Receptor Modulators with Forecaster, a Novel Platform for Drug Discovery. *J*. *Chem*. *Inf*. *Model*. **2012,** *52*, 210-224.

Introduction to Chapter 6

The previous chapter describes the accurate modeling of chemical reactions ongoing within a P450 enzyme. It is well documented that enzymes can be used in synthetic chemistry contexts as an efficient, environmentally-friendly, cheap alternative to metal-based catalysts. In this context, P450 enzymes, bacterial or human isoforms, are good candidates to be employed in a laboratory setting. The enhanced specificity and selectivity of biocatalysts is, in this case, a restriction we wish to dissolve by engineering the protein with the goal of producing a different chemical product. With the docking and reaction modeling in hand, virtually engineering the enzyme is left to be accomplished. This chapter presents the first step to a virtual protein engineering software package. The development of statistical libraries and molecular mechanics energy functions to select three-dimensional conformations and accurately mutate side chain residues is described. To limit the number of variables, side chain residues are mutated to themselves – reconstructed – since a mutation to a different amino acid may entail a backbone motion, not covered in this work. Extensive validation experiments are also described within this chapter.

(This page was left blank intentionally)

Chapter 6:

Single-Point Mutation with a Rotamer Library Toolkit: Toward Protein Engineering

This chapter has been submitted for publication and is reproduced from: "Single-Point Mutation with a Rotamer Library Toolkit: Toward Protein Engineering", Pottel, J.; Moitessier, N.; *Journal of Chemical Information and Modelling*, **2015**, submitted. American Chemical Society (2015).

Author Contributions: this chapter consists of software design and statistical evaluation of its accuracy. The reported developments and applications were contributions of the author of this thesis

6.1 Abstract

Biocatalysis, defined as the use of natural catalysts (e.g., enzymes) to perform a chemical transformation, is an ever-growing field of research characterized by efficient, environmentally-friendly and cost-saving chemical reactions. Enzymes are highly praised for their specificity, however this is not without its flaws. Protein engineers have long been hard at work to harness this natural source of regio-, stereo- and chemo-selectivity in order to carry out chemistry (reactions and/or substrates) not previously achieved with these enzymes. This is most commonly achieved by creating a rapid evolution in the flask resulting in mutated enzymes.

The extreme labor demands and exponential number of mutation combinations have induced computational advances in this domain. Due to the multiple factors at play including thermal stability, activity and selectivity, efforts directed at designing enzymes *de novo* have not yet produced the expected breakthrough. As an alternative to *de novo* protein design, we propose to modulate the enzyme chemistry through single point mutations of an existing enzyme that already demonstrates stability and activity. The first step towards this approach is to predict the correct conformations upon mutating residues (i.e., rebuilding side-chains). For this purpose, we opted for a combination of molecular

mechanics and statistical data. Herein, we have developed automated computational tools to extract protein structural information and created conformational libraries dependent on a variable number of criteria for each amino acid side-chain. We have also developed the necessary tool to apply the mutation and optimize the conformation accordingly. We obtained excellent accuracy with an overall average RMSD of 0.91 Å and 1.01 Å for the 18 flexible natural amino acids within two distinct sets of over 3000 and 1500 side-chain residues respectively. Ultimately, since our *in silico* protein engineering outlook involves using our docking software, FITTED/IMPACTS, we applied our mutation protocol to a benchmarked dataset for self- and cross-docking. Our side-chain reconstruction does not hinder our docking software, demonstrating differences in accuracy of 2.1% and 1.6% respectively for sets including over 200 protein structures.

6.2 Introduction

Protein engineering has become a viable and well-sought after approach to complement synthetic chemistry by using biocatalysts for development of new chemical reactions.[1] The concept of biocatalysis can date all the way back to 1858 with Louis Pasteur's use of *Pencillium glaucium* as an enzyme catalyst[2] (fermentation dating back much further). It has become a very active area of research today in the 21st century[3] which aims to provide complementary tools to small molecule-catalyzed asymmetric synthesis. Biocatalyzed transformations can often be carried out under mild reaction conditions, with high stereo- regio-, and chemoselectivity, as well as environmental friendliness. Naturally occurring enzymes carry out a large variety of transformations including kinetic resolution of racemates, regioselective functionalization of molecules and asymmetric chemical transformations.[4] The main drawback of most biocatalysts is their restricted substrate specificity, which has significantly limited their applicability in catalyzing new, industrially-relevant reactions.

Biocatalysis has been classified into three major periods of laboratory development (LD) and the results are well documented:[5] (1) immobilization of native enzymes to carry out specific chemical transformations, (2) engineering of enzymes to enable the binding of new substrates and thus new synthetic routes and (3) directed evolution involving

random selection and screening in order to simulate Darwinian evolution in a rapid fashion.

Currently, we are likely in a fourth period of LD where advances in computational power, tools and understanding have led to a new gold standard: complete enzyme design.[6-7] The combined efforts of experimental and computational research will likely lead to this achievement. While there are excellent reports of success in this domain,[8] it is believed that this technique is far from the potential it could reach. If a parallel set of periods of computational development (CD) in protein engineering is drawn: (1) growth of visual tools for crystal structures and molecular dynamics simulations, (2) *in silico* prediction of mutations in existing enzymes to accept new substrates, (3) iterative process to create a significantly new variety of enzymes. The next stage is where LD meets CD to create new biocatalysts for new chemical reactions altogether.

Many limitations were overcome to achieve (2) and (3) of LD and many remain to be conquered on the computational parallel; we propose that (2) and (3) of CD have not been fully explored as the Holy Grail (complete enzyme design) is being chased too eagerly. Challenges, from the LD perspective, include enhancing protein stability and enzyme activity (regio-, stereo- and chemo-selectivity). Many research groups have attempted to predict protein thermal stability[9-17] using a variety of techniques and software. Similarly, from the CD side, there has been effort to alter regio- and stereoselectivity by way of protein mutation[18-22] or ligand variation.[23] Thus, the community has achieved parts of (2) of CD and now a more efficient, exhaustive predictive power is required to move to, and beyond, (3) of CD where computers can independently suggest one, or more, side-chain mutations to enable a new specificity.

Herein we report our efforts to begin the development of such a software package where it is anticipated that, given an enzyme with a known mechanism of action (e.g., heme-containing monooxygenase) as well as a desired reaction product (e.g., hydroxylated substrate), side-chain mutations will be automatically proposed after considering regio- and stereoselectivities as well as enzyme thermal stability. Specifically, cytochrome $P450_{BM3}$ is known to oxidize fatty acids, however mutations have enabled the oxidation of indole to create indigo; our long-term goal is to be able to

identify these mutations. Furthermore, once validated, we will be able to propose novel mutations in other cytochrome P450s to carry out never-before-seen oxidations.

Many considerations are vital to such a prediction such as side-chain flexibility, backbone motions and thermostability; however, it is believed that making small, subtle changes buried in the binding site (e.g., single point mutation) of a sizeable enzyme will not completely alter protein conformations or stability. Thus, this goal is within reach and may offer an alternative to *de novo* enzyme design where the creation of a completely new protein sequence is based on templates and a desired function.

The very first step to achieve such a milestone is to successfully predict side-chain conformations, often referred to as rotamers, upon modifying one residue in the protein sequence. We report the development of protocols to efficiently generate rotamer libraries and to select the correct rotamer based on a combination of statistics and molecular mechanics (MM). We further demonstrate its capabilities by applying it to "mutate" residues to themselves by removing and reconstructing a given side-chain, herein termed "self-mutation", and then further testing it in reproducing self- and cross-docking results previously reported from our lab.[24-25] This last validation assessed the quality of the structures at the atomic level.

6.3 Theory and Current State

6.3.1 Virtual protein engineering

Experimental protein engineering is often attempted using a technique denoted as iterative saturation mutagenesis (ISM).[26] The first step of this methodology is to identify one or more residues that are deemed relevant and randomly mutate them to create focused libraries to screen.[27] This involves many labor-intensive steps that could be simulated, anticipated and planned by a computer. Furthermore, the number of possibilities when mutating residues grows exponentially with the number of side-chains selected and thus a virtual protocol evaluating these prospective changes would be desired.[28] In order to have a complete design package, it has been proposed that four items are required:[29-30] (1) an accurate protein backbone motion predictor, (2) an

exhaustive list of discrete rotamer conformations for side-chains, (3) an optimization protocol to efficiently search the library of conformations and (4) a method to score, rank and correctly select the "best" side-chain conformations.

In some cases, an entire protein fold can be proposed[31] however a protein structure can often be obtained from crystal structure databanks,[32] homology models,[33] NMR[34] or molecular dynamics simulations.[35] Currently, the most widely known computational tools available to tackle some or all of the above steps are SCHEMA, ProSAR and ROSETTA although there are many other supporting algorithms that can be used to accomplish individual tasks. SCHEMA[36] investigates the effect of DNA recombination on the structural integrity of proteins and has been applied to determine if functional integrity can be retained as well.[37] This software has been used successfully and is well represented in the literature.[38-40] ProSAR[41-42] is a protein structure-activity relationship approach that applies the concepts of QSAR to protein engineering. This is a simulation of directed evolution and can have the desired results.[43] ROSETTA,[44-45] likely the most impacting software, has been shown to be able to design effective enzymes from scratch. This protocol uses quantum mechanics to build a transition state and then uses template protein structures to create a binding cavity to surround this reactive state. It has been successful with several reaction types and has been improved over the years.[7,46-50] We direct the reader to an excellent review on available techniques that tackle the different elements of protein engineering (although not *de novo* design).[51]

6.3.2 Complexity in predicting mutation

The number of properties that are targeted in protein engineering speaks to the complexity of the problem, especially from a computational standpoint. Thus far, we have mentioned many concerns, some of which are subtler than others. The thermostability and entropy change of the protein, including the folding, rigidity and solvent considerations, are major hotspots.[52-56] The general structure prediction, if attempting *de novo* design, is of paramount importance and many methods have been reviewed.[57] What is often not mentioned is the need to accurately simulate the transition-state of the chemical reaction that is under investigation. Within *de novo* design, this is considered by building a

stabilizing binding cavity surrounding a proposed transition-state.[49] However, the effect that the engineered protein will have on this reactive state is unclear and the quantum mechanics process can cause CPU time to be a major concern.

Our approach here is to incorporate transition-state modeling with a mutated protein design. In order to accomplish this task, we will attempt to minimize the number of unknowns and difficult variables to predict. This follows the suggestion from Peisajovich and Tawfik that stated, although "you get what you screen for" in directed evolution, you should also "select for what is already there".[58] Their statement was made on the premise of beginning from a small amount of catalytic activity with the goal of improving it, however we believe this can apply to the entire rational design paradigm. Additionally, Baker outlined many of the remaining challenges ahead and emphasized the risk of low catalytic efficiency regardless of ideal binding site structure in novel designs.[59] At this point, since we plan to mutate binding site residues only, we assume the mechanism of entry will remain unchanged as will the long distance electrostatics and catalytic machinery and consequently need to ensure the accuracy of our side-chain mutations.

6.3.3 Side-chain importance and rotamer libraries

In this vein, if we begin from known protein structures, the first step to successfully predict the effects of a side-chain mutation is to accurately reposition this new residue with respect to its neighbors and in space. In fact, side-chain flexibility is greater than that of the backbone, although more localized, and predicting their conformations has long been a sought after endeavor.[60-61] The conformational space, if considering all conformations for all side-chains, in even a small protein is computationally inaccessible.[62-63] Thus, to achieve reasonable accuracy, many groups have created rotamer libraries[64] – a set of discrete conformations that each amino acid can have within a protein. To apply these libraries, an evaluation of each rotamer is required for selection of the "correct" (i.e., experimentally observed) one, regardless of whether self-mutation or a true mutation is the goal. There are two central approaches to rotamer libraries: statistics[65-71] and force field[72-75] based. There are also those that combine the two.[76-80] For statistical methods, the data used and the clustering properties are critical. If poor

resolution protein structures are used then the output will be no better. Similarly is the case for libraries that are either too specific or non-specific. The clustering by similarity of observed conformations must not be so specific that there is little statistical relevance (for example, 15° in Figure 6.1) yet it must not be so unspecific that the distribution is too wide and not useful (120° in Figure 6.1 loses a lot of the nuances from the actual distribution). In this latter case, a single conformation will be used to represent a large variety of conformations.

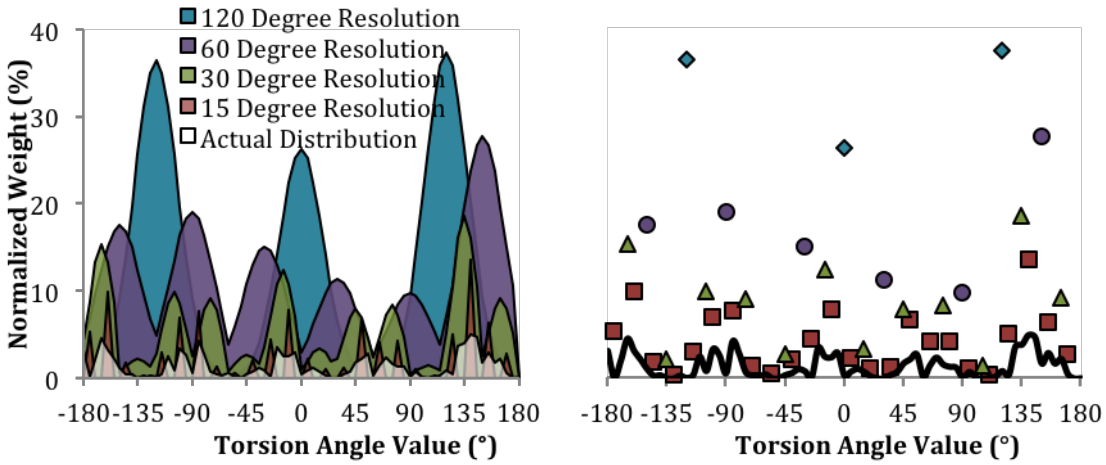


Figure 6.1. The balance for resolution (left) and, consequently, clustering (right) conformational data.

Likewise to backbone conformations, discrete and/or preferred conformations for side-chains can be found in nature that can be exploited. As the structural information contained within the protein data bank (PDB) grows, more specific distributions can be established that can each contain a significant amount of data;[81] libraries have become backbone-dependent for better accuracy. For physics based methods, the primary factor is often steric clashes however other terms have been emphasized such as hydrogen bonds,[66] disulfide bridges[75] and solvation.[82]

6.3.4 Development of a rotamer library

Recently, side-chain conformation was shown to be important for improving protein-protein docking as slight steric clashes can be removed.[83] Current force fields can be very sensitive to attractive or repulsive interactions and thus minor adjustments can have a

large impact on docking results[84] and even stability predictions.[85] Herein we report our development of a new rotamer library/scoring method based on both statistics and physics in order to modify the reconstruction protocol already implemented in PREPARE, a program of the FITTED suite, our docking software.[24] The motivation behind the creation of another rotamer library is three-fold. First, the number of PDB structures has grown significantly in recent years (approx. 10000 new structures per year[32]) and new data results in a more accurate statistical representation. Second, with our own parameters and automated protocol, a more focused/biased library could easily be generated, if necessary; we are not only generating a new rotamer library, but also tools to repeatedly collect new information and create new libraries. For example, if it was identified in a given set of proteins that neighboring residues had a significant effect on side-chain conformations, then we could analyze the database and encode this as a clustering criterion. Or, if we determine that this broad rotamer library is unsatisfactory for modeling side-chains conformations in a given class of proteins, like in the case of DFG-in and –out in the kinase family,[24] it would be quick and easy to only search these enzymes to produce a focused rotamer library. Finally, and perhaps most importantly, for an effective protein engineering software to be used by experimentalist biochemists, the protocol should be integrated and automated in a user-friendly manner. Should the rotamer library/selection criteria exist in a third-party software, this detracts from the likelihood it is used by the general chemistry community.

6.4 Implementation

6.4.1 Statistical library and clustering

To gather the proper data, the entire available PDB was queried and filtered for structures that were deemed appropriate for our study. Since we hope to create a general-purpose software package, we did not discriminate between different protein families or functions. In order to avoid overtraining for a given protein (e.g. α-thrombin has 355 structures deposited as of Aug. 2015), a set of non-redundant protein chains was uploaded from the PDB. Residues with B-factors above 50 (poorly resolved) were removed due to poor resolution resulting in a dataset of 18752 structures which contained a total of

approximately 3.3 million residues, distributed amongst the different amino acids as shown in Table 6.1. Our program PREPARE was modified to read PDB files and collect structural information on each residue. Among the information gathered for a given residue was: torsion angles and the secondary structure – alpha helix, beta sheet or loop – that this residue pertained to; the process was automated and could be applied to any data set of PDB structures.

Table 6.1. Amino acid distribution amongst 18752 PDB structures

Amino Acid	Counts	Amino Acid	Counts
ARG	191239	LYS	203085
ASN	163870	MET	65779
ASP	224480	PHE	165908
CYS	51036	PRO	184096
GLN	136905	SER	232645
GLU	244631	THR	215771
HIS	93218	TRP	55213
ILE	239549	TYR	141297
LEU	387365	VAL	293446

The next challenge was to cluster the large amount of data to obtain a rotamer library that could be evaluated for its ability. We defined the conformation of a residue as being the set of torsion angles that describe the geometry of the side-chain. Thus, we selected different resolution values to create bins, and for each torsion angle we assigned it to a bin according to Eq. 6.1.

$$Code = \frac{\left(Angle + \left(180 + \frac{Resolution}{2}\right)\right)}{Resolution},$$

$$if\ Code = \frac{360}{Resolution}, then\ Code\ is\ set\ to\ 0 \qquad \textbf{(6.1)}$$

The side-chain was then assigned a value corresponding to its bin number, while accounting for the circular redundancy (0° = 360°). The bin value was established according to Eq. 6.2; each individual angle was represented by a 2-digit value, resulting in an 8-digit code. A constant of 10 was added to avoid leading 0s (1 would have to be 01). There are 4 codes since the largest side-chains, arginine and lysine, each have 4

torsion angles that need to be evaluated (the final torsion in arginine is always 180° relative)

$$Bin\ Number = (Code1 + 10) + \big(100 * (Code2 + 10)\big) + \big(10000 * (Code3 + 10)\big) + \big(1000000 * (Code4 + 10)\big) \quad (\mathbf{6.2})$$

Once complete, the bin is reviewed and if it already exists in duplicate then they are grouped together; this method was effective for memory concerns due to minimal information storage. The torsion angle values were continuously summed and never stored, enabling us to obtain an average value for each angle in each cluster. A 2nd iteration and slightly modified algorithm allowed us to also obtain a standard deviation to establish the nature of the distribution of each angle. Both procedures were automated and repeated for all side-chains from all proteins with different resolutions (Table 6.2). The libraries can be found online from Molecular Forecaster Inc.

Table 6.2. Clustering results using different resolution values. Number of clusters with each set of criteria is given.

Res.:	15°			30°			60°			120°		
Sec. Struct.:	Loop	Beta-sheet	Alpha-helix	Loop	Beta-sheet	Alpha-helix	Loop	Beta-sheet	Alpha-helix	Loop	Beta-sheet	Alpha-helix
ARG	9559	22154	19787	3778	5864	5233	751	854	815	80	80	79
ASN	1019	1087	898	302	318	273	83	86	79	18	18	18
ASP	995	1108	997	292	317	305	85	91	79	18	20	19
CYS	24	24	24	12	12	12	6	6	6	3	3	3
GLN	4209	7536	6630	1405	1881	1729	282	355	335	54	54	57
GLU	5326	9056	9187	1569	2097	2097	313	387	365	54	56	55
HIS	885	1038	840	274	301	261	82	80	70	18	18	18
ILE	674	1035	850	239	304	273	68	85	79	19	23	21
LEU	836	1133	880	265	330	271	79	94	85	21	24	19
LYS	8744	19682	18663	3539	5602	5148	735	843	805	80	80	80
MET	2203	3418	2615	810	965	765	171	183	158	27	27	27
PHE	805	946	756	251	277	228	74	80	70	18	18	18
PRO	126	N/A	N/A	46	N/A	N/A	18	N/A	N/A	9	N/A	N/A
SER	24	24	24	12	12	12	6	6	6	3	3	3
THR	83	107	89	42	49	43	20	23	22	9	9	9
TRP	574	781	649	196	240	200	62	69	66	18	18	16
TYR	738	920	747	246	264	234	76	76	68	18	18	19
VAL	76	108	81	35	45	37	17	20	18	7	8	7

In order to select one of these libraries, we evaluated the distribution at 5° resolution – a fine grain – with the goal of establishing how our coarse-grained libraries (60°, 120°) can reproduce the true distribution. For visualization purposes, this was illustrated for only one angle (2-dimensional result) for a small set of amino acids in a given conformation; the chosen angles represent sp^3-sp^3 and sp^3-sp^2 bonds. The coarse-grained plots were obtained assuming a Gaussian distribution using the calculated average angle and standard deviation (Figure 6.2) within each bin.

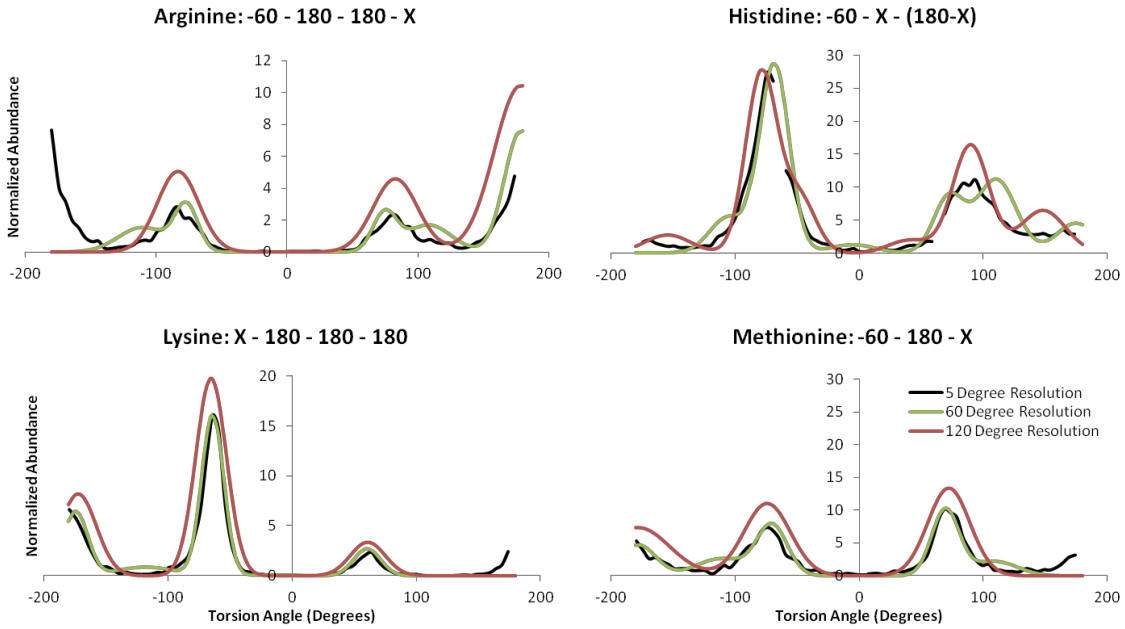


Figure 6.2. Comparing fine and coarse-grained rotamer distributions for different amino acids. The abundance has been normalized to 60° in each case; the relative peak heights are still relevant at 120° however the absolute peak heights are not.

Based on inspection of these curves as well as chemical intuition – the need to explore all gauche/staggered/eclipsed combinations – both 60° and 120° libraries offer potential for success. From this point forward, the finer resolution libraries, 15° and 30°, were omitted for size concerns and a lack of easy representation of the majority of the torsion space.

6.4.2 Testing different side-chain libraries

Since the proposed usage of such a rotamer library is to mutate amino acids within the binding site to affect a protein's function, we "trained" our criteria selection (resolution, library size, conformation scoring) on a set of protein structures that were preferred because of the existence of mutation data for these enzymes, i.e. they could be used in further protein engineering studies. The word "trained" may be inappropriate since the only training is with regard to some of the criteria and does not in any way influence the data output and could essentially not be biased. The set consisted of 98 PDB structures (see Appendix 4 for full list) and using PROCESS part of the FORECASTER platform,[86] we established a list of binding site residues; any side-chain within 7 Å (a default parameter in PROCESS) of the crystallized ligand was identified. This resulted in a data set of 3179 residues distributed amongst the amino acids as shown in Table 6.3.

Table 6.3. Amino acid distribution within the binding site of 98 and 68 PDB structures making up the training and testing sets respectively. The docking set is also presented.

Amino acid	"Training" Set Counts	%	"Testing" Set Counts	%	Difference	Docking Set Count	%
ARG	226	7.11	92	5.36	-1.75	7	3.04
ASN	170	5.35	102	5.94	0.59	10	4.35
ASP	176	5.54	92	5.36	-0.18	11	4.78
CYS	63	1.98	25	1.46	-0.53	3	1.30
GLN	82	2.58	34	1.98	-0.60	7	3.04
GLU	112	3.52	66	3.84	0.32	13	5.65
HIS	113	3.55	52	3.03	-0.53	19	8.26
ILE	253	7.96	203	11.82	3.86	13	5.65
LEU	322	10.13	218	12.69	2.56	27	11.74
LYS	138	4.34	94	5.47	1.13	9	3.91
MET	79	2.49	43	2.50	0.02	12	5.22
PHE	192	6.04	141	8.21	2.17	14	6.09
PRO	164	5.16	91	5.30	0.14	5	2.17
SER	212	6.67	97	5.65	-1.02	24	10.43
THR	246	7.74	131	7.63	-0.11	5	2.17
TRP	153	4.81	34	1.98	-2.83	13	5.65
TYR	185	5.82	50	2.91	-2.91	17	7.39
VAL	293	9.22	153	8.91	-0.31	21	9.13
TOTAL	3179	100	1718	100	0.00	230	100

Next, to evaluate each rotamer library and set of criteria, these ~3200 residues were individually, one at a time, deconstructed from the protein, keeping only the 4 backbone atoms (N, C, O, C_α), and then reconstructed using data from the rotamer library. To rank the different settings, a root mean squared deviation (RMSD) was calculated between the reconstructed and the crystal structure side-chains (excluding the 4 backbone atoms and C_β). In the final instance, an evaluation of the torsions is also given. The progress towards a sensible, well thought out library selection will now be described.

We began with the coarser of the two libraries, a resolution of 120°. Furthermore, we did not consider backbone secondary structure at the outset and used only those conformation clusters within the loop grouping (although these conformations were almost always represented in the beta-sheet and alpha-helix groupings as well). The method of selection of a conformation was varied: a random choice from the library, the most abundantly observed conformation, evaluating every single conformation with a MM force field, as well as allowing some leniency (flexibility by rotating torsion angles clockwise or counter-clockwise within a restricted range) to each average angle were all attempted. The results are summarized in Table 6.4. The "worst" row indicates the maximum RMSD of all the methods listed.

The 60° resolution library was evaluated ("60small", Table 6.4) with the same set of criteria. We used a small set, using only the top 10 conformations per residue and then we expanded the number of clusters considered by the software to 34 (in the rotamer library by Lovell et al.,[68] the library previously implemented in PREPARE, a maximum of 34 conformations were found – "60large", Table 6.4). We thought to probe deeper and expand the library size to top 70, ("60XL", Table 6.4). At this point, we included a measure of probability for the selected conformation ("Scoring", Table 6.4). Each conformation, for each amino acid, was assigned a Boltzmann-weighted energy based on its abundance within the library (Eq. 6.3) in order to add a statistical consideration to the selection process. Consequently, the most observed conformation has an energy penalty of 0 kcal/mol and wide vs. narrow distributions are discriminated in an intelligent manner.

$$E_{conformation} = -k_B T \times \ln\left(\frac{\%\ current\ conformation}{\%\ most\ abundant\ conformation}\right) \quad (\mathbf{6.3})$$

where k_B is the Boltzmann constant and T is the temperature (298K).

Subsequently, we further expanded the number of conformations for arginine and lysine (top 100) without too large a computational cost, and we examined a weighting scheme for the statistical energy, $E_{conformation}$ ("60XXL/w = 0.25-5", Table 6.4). Finally, a consideration for secondary structure was added (the percentages and order of the conformation clusters differed) ("sec", Table 6.4). For comparison, we also evaluated the 120° resolution library including the scoring and the secondary structure and observed similar results.

Table 6.4. Training set results: average RMSD upon side-chain reconstruction using different rotamer libraries and selection criteria. (W.AVG: weighted average; Sc: Scoring, weight in brackets; gradient red to green for improved RMSD)

	Counts:	ARG 226	ASN 170	ASP 176	CYS 63	GLN 82	GLU 112	HIS 113	ILE 253	LEU 322	LYS 138
120	Worst	7.42	3.83	3.88	3.16	5.02	4.99	5.25	3.75	3.68	5.81
	Random	3.80	2.43	2.40	2.08	2.86	2.85	3.52	2.03	2.50	3.29
	Abundant	3.12	1.52	2.02	1.09	2.18	2.12	3.18	1.21	1.54	2.71
	All	2.29	0.81	1.03	0.51	1.36	1.46	0.94	1.22	0.89	1.73
	Flexibility	1.93	0.81	1.02	0.45	1.29	1.46	0.94	1.17	0.94	1.80
	Flex Wide (20-5)	1.98	0.80	1.00	0.42	1.25	1.45	0.91	1.15	1.05	1.98
60small	Worst	7.68	3.65	3.63	3.40	5.11	4.92	5.26	3.74	3.72	5.77
	Random	3.80	2.50	2.51	2.45	2.43	2.29	3.51	2.01	2.32	3.00
	Abundant	3.13	1.50	1.98	1.10	2.17	2.13	3.19	1.21	1.55	2.74
	All	2.47	0.68	0.78	0.54	1.47	1.81	0.99	1.12	1.06	2.20
	Flexibility	2.52	0.66	0.76	0.46	1.44	1.72	0.97	1.09	1.11	2.09
60large	Worst	7.68	3.79	3.76	3.40	5.29	4.92	5.24	3.81	4.06	6.04
	Random	3.82	2.17	2.30	2.45	2.39	2.62	3.18	1.73	2.52	3.12
	Abundant	3.13	1.50	1.98	1.10	2.18	2.13	3.19	1.21	1.55	2.74
	All	2.13	0.89	0.98	0.53	1.12	1.65	0.84	1.21	1.14	1.83
	Flexibility	1.87	0.86	0.95	0.46	1.09	1.58	0.95	1.14	1.18	1.82
	Flex Wide (10-2.5)	1.78	0.85	0.96	0.52	1.17	1.54	0.96	1.15	1.19	1.82
60XL	Flexibility	1.79	0.87	1.03	0.46	1.17	1.61	0.98	1.16	1.19	1.78
	Flex Wide (10-2.5)	2.07	0.84	0.99	0.52	1.20	1.58	0.99	1.17	1.20	1.74
	Flex + Sc (W=1.00)	1.79	0.81	0.91	0.45	1.18	1.47	0.86	1.13	0.59	1.75
60XXL	Flex + Sc (W=0.25)	1.82	0.86	0.97	0.47	1.19	1.63	0.88	1.16	0.94	1.72
	Flex + Sc (W=0.50)	1.83	0.86	0.92	0.46	1.17	1.54	0.91	1.14	0.76	1.69
	Flex + Sc (W=0.75)	1.84	0.84	0.91	0.45	1.19	1.50	0.87	1.14	0.68	1.71
	Flex + Sc (W=1.00)	1.83	0.81	0.91	0.45	1.18	1.47	0.86	1.13	0.59	1.70
	Flex + Sc (W=1.25)	1.82	0.79	0.90	0.45	1.17	1.42	0.82	1.11	0.57	1.69
	Flex + Sc (W=1.50)	1.81	0.78	0.89	0.45	1.16	1.43	0.80	1.11	0.61	1.70
	Flex + Sc (W=1.75)	1.78	0.77	0.89	0.45	1.14	1.41	0.84	1.10	0.63	1.70
	Flex + Sc (W=2.00)	1.78	0.77	0.89	0.44	1.14	1.41	0.82	1.12	0.63	1.69
	Flex + Sc (W=5.00)	1.72	0.66	0.90	0.38	1.23	1.44	0.83	1.10	0.63	1.73
60XXL	Flex + Sc (W=0.00) + sec	1.86	0.89	1.05	0.51	1.14	1.51	1.04	1.16	1.19	1.71
	Flex + Sc (W=1.00) + sec	1.79	0.80	0.89	0.40	1.13	1.44	0.91	1.14	0.56	1.64
	Flex + Sc (W=2.00) + sec	1.75	0.69	0.90	0.33	1.12	1.36	0.91	1.14	0.54	1.61
120 XXL	Flex + Sc (W=1.00) + sec	1.94	0.73	0.93	0.37	1.17	1.41	0.94	1.14	0.56	1.74

		MET	PHE	PRO	SER	THR	TRP	TYR	VAL	AVG	W. AVG
	Counts:	**79**	**192**	**164**	**212**	**246**	**153**	**185**	**293**	**3179**	**3179**
120	Worst	5.66	5.43	3.81	2.66	2.83	7.25	6.42	3.01	4.66	4.49
	Random	3.16	3.82	2.21	1.65	1.75	4.25	4.06	1.81	2.80	2.71
	Abundant	2.09	1.83	0.66	1.59	1.30	3.34	2.62	1.13	1.96	1.87
	All	1.12	0.67	0.68	1.05	0.54	1.17	1.03	0.39	1.05	1.02
	Flexibility	1.06	0.55	0.68	0.97	0.48	0.98	0.84	0.36	0.98	0.96
	Flex Wide (20-5)	1.18	0.54	0.68	0.96	0.52	1.05	0.68	0.37	1.00	0.97
60small	Worst	4.59	5.50	3.25	2.88	2.85	7.12	6.45	2.91	4.58	4.44
	Random	2.67	3.68	0.58	1.76	1.75	4.01	3.44	1.89	2.59	2.52
	Abundant	2.12	1.85	0.64	1.60	1.30	3.45	2.65	1.14	1.97	1.88
	All	0.98	0.63	0.66	1.11	0.56	1.46	1.01	0.38	1.11	1.07
	Flexibility	0.79	0.51	0.67	1.04	0.52	1.23	0.82	0.36	1.04	1.02
60large	Worst	5.09	5.79	3.86	2.88	2.85	7.09	6.59	2.89	4.72	4.57
	Random	2.82	2.74	1.84	1.76	1.76	4.43	3.45	1.97	2.61	2.54
	Abundant	2.12	1.85	0.64	1.60	1.30	3.45	2.65	1.14	1.97	1.88
	All	1.12	0.67	0.67	1.12	0.52	1.20	1.08	0.39	1.06	1.05
	Flexibility	1.04	0.59	0.67	1.05	0.53	1.10	0.90	0.36	1.01	0.99
	Flex Wide (10-2.5)	1.00	0.54	0.67	0.99	0.51	0.98	0.76	0.37	0.99	0.97
60XL	Flexibility	1.24	0.60	0.67	1.05	0.53	1.03	0.86	0.36	1.02	1.00
	Flex Wide (10-2.5)	1.33	0.57	0.67	0.99	0.51	0.93	0.77	0.37	1.02	1.00
	Flex + Sc (W=1.00)	0.80	0.55	0.67	0.94	0.42	1.13	0.85	0.40	0.93	0.89
60XXL	Flex + Sc (W=0.25)	1.08	0.59	0.67	0.99	0.47	1.06	0.87	0.37	0.99	0.96
	Flex + Sc (W=0.50)	0.86	0.56	0.67	0.95	0.44	1.09	0.84	0.38	0.95	0.92
	Flex + Sc (W=0.75)	0.79	0.54	0.67	0.95	0.45	1.13	0.85	0.39	0.94	0.91
	Flex + Sc (W=1.00)	0.80	0.55	0.67	0.94	0.42	1.13	0.85	0.40	0.93	0.89
	Flex + Sc (W=1.25)	0.79	0.54	0.67	0.94	0.42	1.10	0.85	0.41	0.91	0.88
	Flex + Sc (W=1.50)	0.76	0.52	0.67	0.93	0.42	1.15	0.85	0.42	0.91	0.88
	Flex + Sc (W=1.75)	0.76	0.53	0.67	0.92	0.40	1.15	0.84	0.41	0.91	0.88
	Flex + Sc (W=2.00)	0.75	0.54	0.67	0.91	0.44	1.15	0.84	0.43	0.91	0.88
	Flex + Sc (W=5.00)	0.81	0.60	0.67	0.95	0.58	1.06	0.87	0.53	0.93	0.90
60XXL	Flex + Sc (W=0.00) + sec	1.22	0.60	0.67	1.01	0.53	1.06	0.82	0.35	1.02	1.00
	Flex + Sc (W=1.00) + sec	0.76	0.48	0.67	0.96	0.44	1.16	0.77	0.39	0.91	0.87
	Flex + Sc (W=2.00) + sec	0.74	0.50	0.67	0.93	0.50	1.14	0.78	0.45	0.89	0.87
120 XXL	Flex + Sc (W=1.00) + sec	0.76	0.49	0.68	0.93	0.44	0.92	0.80	0.39	0.91	0.88

6.5 Results and Discussion

6.5.1 Analysis of performance

In general, the average RMSD improves as we consider more components of the library and can be attributed to the MM consideration to rank the different potential conformations, as well as greater coverage of the conformational space. For example with ARG shown in Figure 6.3, using 60° resolution, the most abundant conformation (7.66%) gives RMSD=4.05 Å, rank 8 (2.36%), selected from the top 10, has RMSD=2.55 Å and rank 18 (1.51%), selected from the top 34, has RMSD=0.64 Å. In fact, the top 10 conformations in our dataset included less than 85% of the conformation space for 11 of the 18 amino acids and less than 55% for 5 (Table 6.5). The results demonstrated that while some of the smaller amino acids showed improvement, the large, flexible ones became worse. Including the top 34 conformations and subsequently the top 70, only two amino acids being below 80% of the conformational space, led to significant reductions in the RMSD of these floppy side chains. The top 100 conformations were included for ARG and LYS to achieve 80% coverage. This significantly reduces the odds that the correct conformation is not considered, as was the case with the smaller libraries.

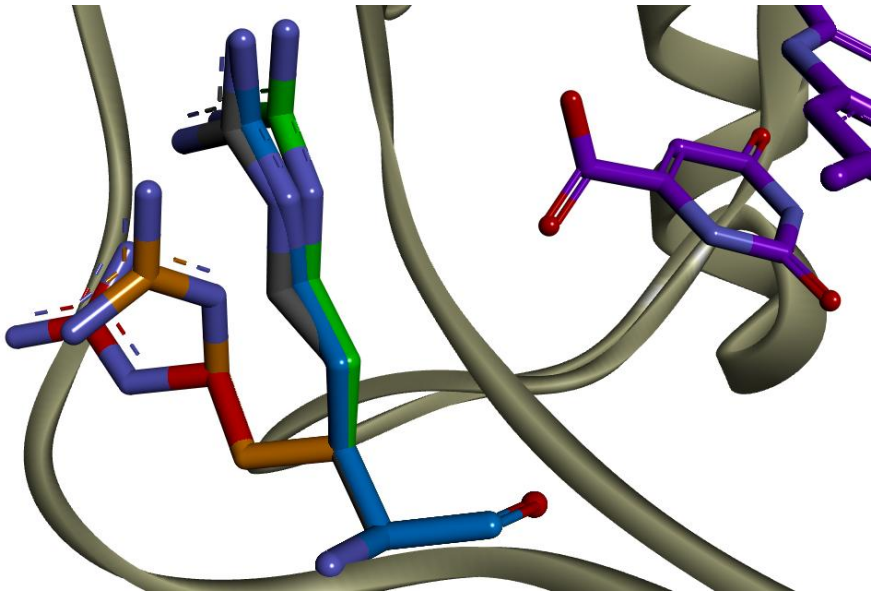


Figure 6.3. Benefit of a larger rotamer library. PDB code: 1OVD; ARG57. Using 60° resolution: most abundant conformation (red), rank 8 (orange), rank 18 (green), with flexibility (blue). Ligand in purple, grey is crystal structure. Hydrogen atoms removed for clarity.

Table 6.5. Coverage of the conformational space (total percentage) for different number of clusters included from a 60° resolution library

	Top 34 (120°)	Top 10	Top 34	Top 70	Top 70 (ARG/LYS 100)
ARG	90.89	32.10	60.40	74.45	79.87
ASN	100.00	78.02	98.54	99.95	99.95
ASP	100.00	82.08	98.88	99.95	99.95
CYS	100.00	100.00	100.00	100.00	100.00
GLN	95.33	50.72	78.78	89.66	89.66
GLU	94.92	53.39	77.50	89.31	89.31
HIS	100.00	80.41	97.92	99.92	99.92
ILE	100.00	93.12	99.12	100.00	100.00
LEU	100.00	90.08	98.93	99.96	99.96
LYS	92.70	49.07	69.12	78.97	83.50
MET	100.00	54.97	83.89	93.10	93.10
PHE	100.00	83.74	98.42	99.98	99.98
PRO	100.00	97.64	100.00	100.00	100.00
SER	100.00	100.00	100.00	100.00	100.00
THR	100.00	99.75	100.00	100.00	100.00
TRP	100.00	78.98	97.85	100.00	100.00
TYR	100.00	84.76	98.66	99.96	99.96
VAL	100.00	99.62	100.00	100.00	100.00

Moreover, as we allow for some flexibility from the average (a range of +/- 5° with 2.5° increments) we see a further refinement where the RMSD drops to 0.32 Å (Figure 6.3). We attempted a much wider flexibility to represent the approximate standard deviations in the clusters (+/- 20° at 5° intervals for 120°, +/- 10° at 2.5° intervals for 60°), however this led to poor success in most cases with a significant increase in computational demand.

With regard to the preference of 60° over the 120° clustering, the results were very similar. In fact, for some residues, such as LEU, the results suggested that 120° resolution was a wiser selection. However, for a flexible residue like ARG, it appeared that 60° was preferred with an average 0.14 Å improvement in RMSD; in some cases, important conformations were not even considered. To demonstrate, Figure 6.4 shows chosen

conformations using 120° resolution and flexibility, rank 33 (0.69%), resulting in an RMSD=4.89 Å and using 60° resolution and flexibility, rank 37 (0.63%), having an RMSD=0.55 Å where this conformation does not exist in the former library. Ultimately, 120° may better model a few of the amino acids based on the bond types (sp^3 – sp^3 vs. sp^2 – sp^3 have different placement/number of minima) however our goal was to opt for one set of criteria to be used as generally as possible without biasing our selection. In the future, one could envision different sized rotamer libraries being used dependent on the amino acid in question or, more specifically, on each bond type.

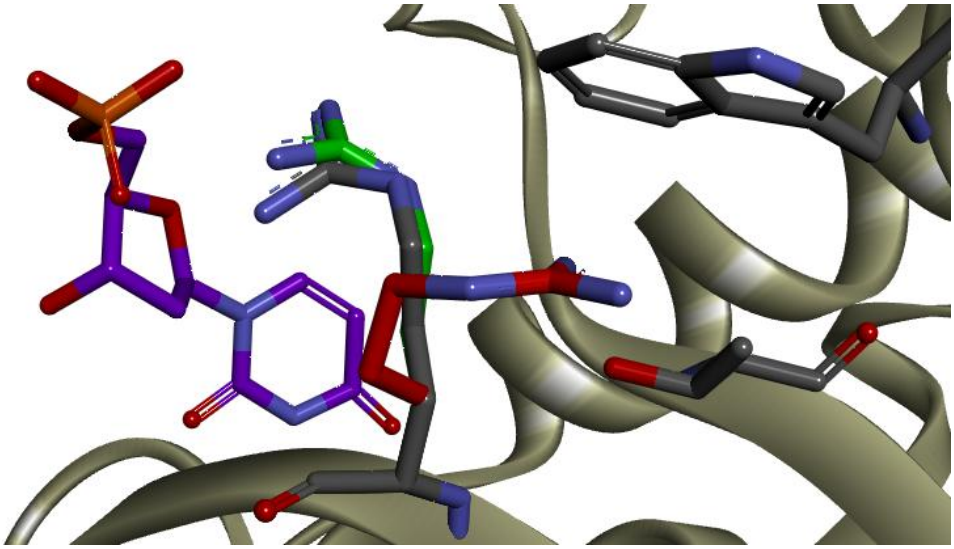


Figure 6.4. Preference for 60° resolution over 120°. PDB code: 1JMF; ARG218. Chosen conformations: using 120° resolution, rank 33 (red) and using 60° resolution, rank 37 (green). Ligand in purple, grey is crystal structure. Hydrogen atoms removed for clarity.

The statistical component within the scoring is where we saw the greatest improvement, especially for LEU and MET. LEU is small, hydrophobic and can often fit in many different conformations. Thus without any statistical consideration, very minor interactions will dominate in MM. For example, using 60° resolution without any statistical weighting, the selected conformation was rank 58 (0.02%) giving an RMSD=1.76 Å, whereas, with scoring W=1, rank 1 (53.7%) was selected yielding an RMSD=0.19 Å (Figure 6.5). MET, on the other hand, is larger and floppier, however the diffuse electron cloud of sulphur atoms can result in discrepancies within force fields[87] and thus a solely MM scoring may be inadequate. Interestingly, the underlying cause of added success with statistical weighting was subtler and no significant electronic effects

were observed; the results were undeniable though. As we can see in Figure 6.5, using 60° resolution without any statistical weighting, rank 42 (0.32%) resulted in an RMSD=3.58 Å, while, with scoring W=1, rank 4 (4.95%) gave a much better RMSD=0.18 Å. The general enhancement in accuracy led us to believe that the combination of statistics (evaluating side-chain intrinsic energy) and MM (measuring the interaction with other residues) considerations would be the direction to advance. To our delight, a weight of 1.00 on the Boltzmann energy was satisfactory and would indicate little over-training for our system.

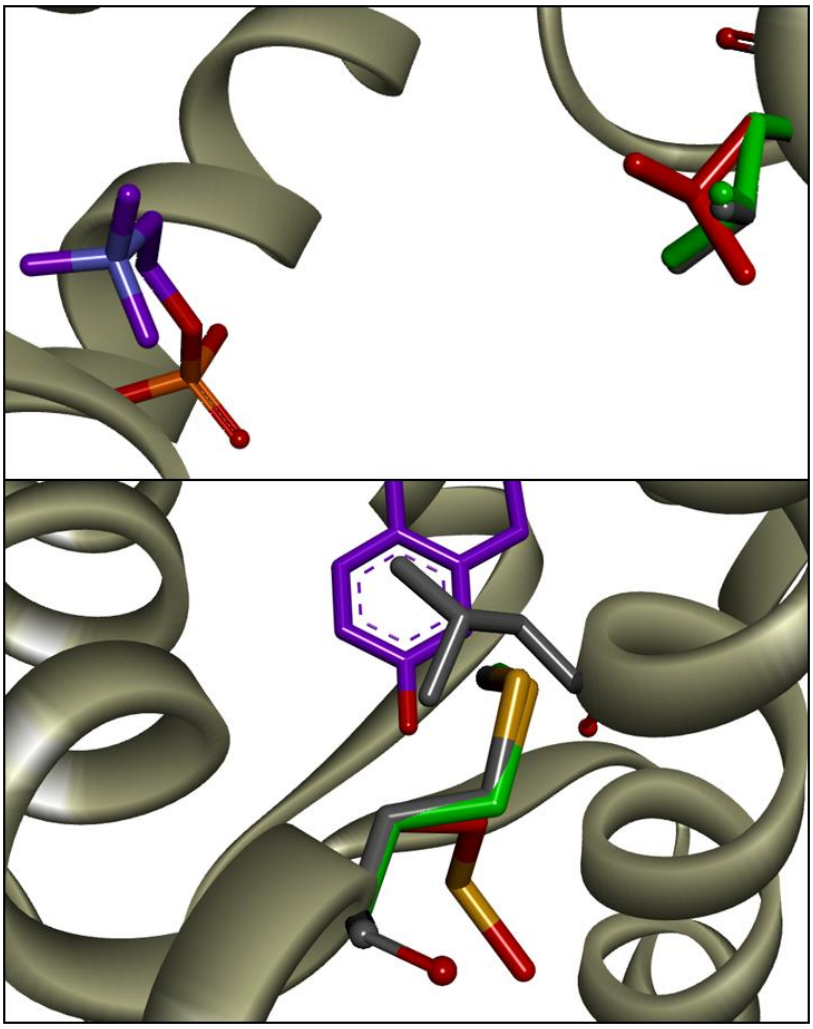


Figure 6.5. Statistical consideration improves LEU and MET self-mutations. Top: PDB code: 3UJD; LEU14. Using 60° resolution: rank 58 (red), scoring W=1, rank 1 (green). Bottom: PDB code: 1QKT; MET357. Using 60° resolution: rank 42 (red), scoring W=1, rank 4 (green). Ligands in purple, grey is crystal structure residues. Hydrogen atoms removed for clarity.

The statistics play an important role when interactions with the ligand were not the driving force behind the observed conformation. Conversely, the MM component could be overpowering. As shown in Figure 6.6, a number of issues can arise. The correct conformation would have been rank 27 (0.92%), an energetic preference of 1 kcal/mol over the selected one at 0.17%, although the final angle is 30° away from the average angle value – outside the 5° flexibility; the flexibility can remain too narrow (i.e., even finer resolution required). Additionally, a water molecule interacts with the crystal structure ARG, however this interaction was not considered during the mutation procedure and thus the favorable interaction was with the ligand instead. This demonstrated the overpowering by MM, disregarding the statistically higher ranked conformation – even with W=5.0, the selection remains the same. It also showed that additional factors such as water molecules could modulate the side-chain conformations, a process that our program does not account for currently.

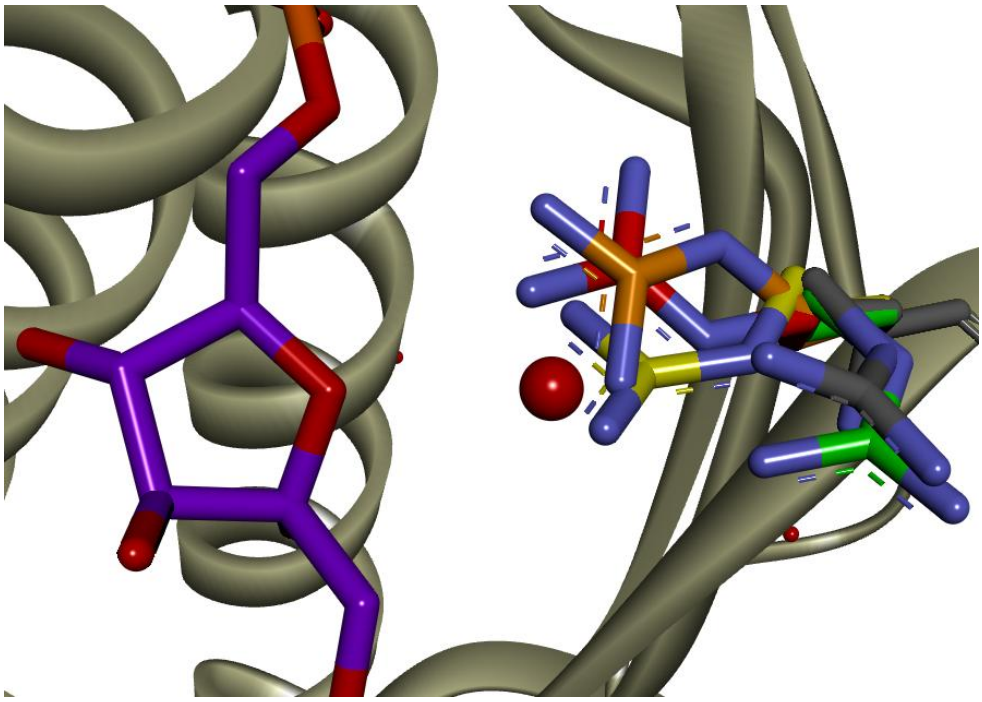


Figure 6.6. Summary of the statistical and MM limitations. PDB code: 1CNQ; ARG140. Using 60° resolution: top 10 and 34, selected rank 4 (2.85%) (Red, RMSD=3.10 Å), top 70, selected rank 58 (0.32%) (Orange, RMSD=2.71 Å), top 100, with scoring W=1, selected rank 87 (0.17%) (Yellow, RMSD=2.46 Å). Correct conformation, rank 27 (0.92) (Green). Ignored water molecule (red sphere), crystal structure (grey) and ligand (purple). Hydrogen atoms removed for clarity.

The secondary structure was the final, optional condition we chose to test; overall the accuracy was slightly improved and this combination demonstrated the best average RMSD for all the criteria. Although nearly identical sets of conformations were collected whether the residues were on loops, α-helices or β-sheets, their occurrence varied, and consequently, so were the weights assigned to each of them. Thus, this slight improvement is likely the result of refined statistical scoring of the residue side-chains incorporating the backbone-dependency. These results were satisfactory and thus we proceeded to evaluate the transferability of the 60° resolution library, using MM and statistical energy components and secondary structure for conformation selection, to a completely different (testing) data set of protein structures.

6.5.2 Validation – testing set for self-mutation

This new rotamer library and selection criteria were now applied to a testing set of 68 PDB structures, also obtained from a set of proteins having crystallographic mutation data for future protein engineering studies. The distribution amongst the 1718 amino acids was very similar to the original data set however the structures are completely different (Table 6.3). To this set, we applied the 60° resolution library using the statistical and MM evaluation with and without the secondary structure consideration. Additionally, the 120° resolution library was evaluated due to the apparent success when combined with the final selection method. However, with this testing set, the 120° library performed worse and reinforced our ultimate selection of the 60° library (Table 6.6).

Table 6.6. Testing set results: average RMSD upon side-chain reconstruction using select rotamer libraries and selection criteria (continued gradient from Table 6.4).

		60XXL		120XXL
	Counts	Flex + Scoring (1.00)	Flex + Scoring (1.00) + sec	Flex + Scoring (1.00) + sec
ARG	92	2.31	1.85	2.17
ASN	102	0.79	0.83	0.86
ASP	92	0.67	0.66	0.67
CYS	25	0.91	0.88	0.88
GLN	34	1.54	1.55	1.49
GLU	66	1.58	1.50	1.51
HIS	52	0.92	0.86	0.91
ILE	203	1.54	1.55	1.60
LEU	218	0.67	0.67	0.68
LYS	94	2.03	1.96	2.02
MET	43	0.90	0.91	0.89
PHE	141	0.64	0.68	0.65
PRO	91	0.61	0.61	0.63
SER	97	1.36	1.40	1.37
THR	131	0.43	0.43	0.49
TRP	34	1.08	1.01	1.06
TYR	50	0.61	0.49	0.47
VAL	153	0.31	0.29	0.29
AVG	1718	1.05	1.01	1.03
W.AVG	1718	1.00	0.97	1.00

Since the standard in the field of rotamer libraries is to determine the success according to torsion matching, we evaluated this metric as well (Table 6.7). In the context of protein engineering and binding site optimization, we believe the RMSD is a more relevant measure, as the overall shapes of the side-chains are crucial. Although the first torsion angle can differ by 90°, for example, the general shape of a floppy residue like arginine or lysine can right itself at the polar head – the vital point of interaction – and have a relatively low RMSD (Figure 6.7). Conversely, by torsion accuracy a prediction may be within 40° at each angle however the overall conformation and position of this

polar head can be distinctly different, having an RMSD over 3.5 Å. Furthermore, a general trend can be seen relating the RMSD and the torsion accuracy demonstrating that at low RMSD, the torsion accuracy must be high; however at high RMSD, where the conformation is positioned incorrectly, the torsion accuracy can be high and misleading (Figure 6.7). Regardless of the metric to evaluate success, the proof that this implementation is valuable can be shown by carrying out self- and cross-docking experiments since this is the foundation of our protein engineering research plan.

Table 6.7. Torsion accuracy for both training and testing in percentage. Predicted torsions (χ) within 40° of the crystal structure are deemed correct. If first χ is incorrect, subsequent χ are also incorrect.

		ARG	**ASN**	**ASP**	**CYS**	**GLN**	**GLU**	**HIS**	**ILE**	**LEU**	**LYS**
	N:	226	170	176	63	82	112	113	253	322	138
Training Set	χ_1	78.0	86.9	82.9	93.8	79.5	75.2	96.5	76.1	95.4	78.4
	χ_2	65.6	71.4	73.1		65.1	54.0	61.1	50.7	87.0	65.5
	χ_3	37.0				54.2	35.4			87.0	53.2
	χ_4	30.0									25.9
	N:	92	102	92	25	34	66	52	203	218	94
Testing Set	χ_1	79.3	82.4	89.2	72.0	88.2	77.3	96.2	63.6	91.3	69.9
	χ_2	67.4	73.5	76.3		55.9	62.1	67.3	29.6	83.0	51.6
	χ_3	37.0				17.6	47.0			83.0	30.1
	χ_4	19.6									17.2
		MET	**PHE**	**PRO**	**SER**	**THR**	**TRP**	**TYR**	**VAL**	**AVG**	**W. AVG**
	N:	79	192	164	212	246	153	185	293	3179	3179
Training Set	χ_1	90.1	100	44.3	63.9	87.3	90.8	95.2	91.3	83.6	83.8
	χ_2	87.7	94.4	38.3			82.4	92.5		70.6	71.5
	χ_3	76.5								57.2	60.7
	χ_4									27.9	28.4
	N:	43	141	91	97	131	34	50	153	1718	1718
Testing Set	χ_1	88.4	95.8	58.2	44.3	90.8	97.1	100	96.1	82.2	81.5
	χ_2	81.4	89.4	51.6			85.3	94.1		69.2	66.8
	χ_3	67.4								47.0	56.5
	χ_4									18.4	18.4

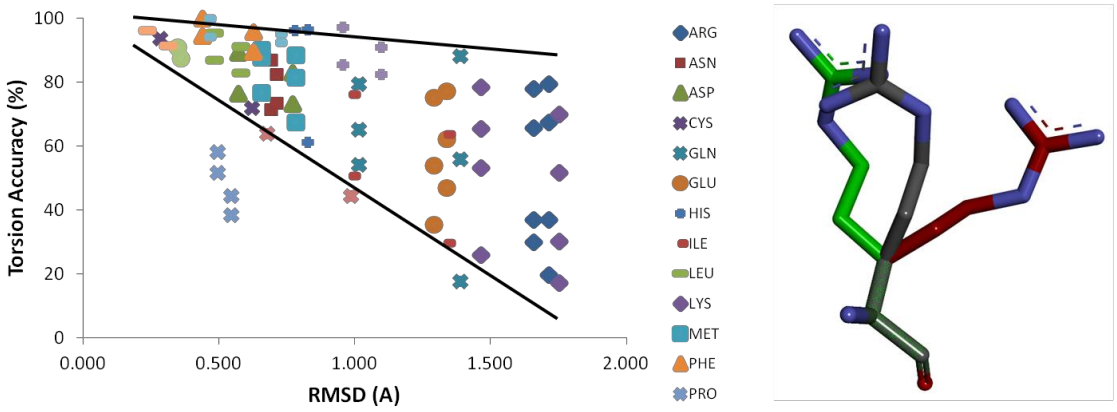


Figure 6.7. Left: Relationship between RMSD and torsion accuracy. Trends defined by black lines. Right: Arginine residue and the demonstration of pitfalls for torsion accuracy. Grey – crystal structure. Green – torsion incorrect, RMSD=1.76 Å. Red – torsion correct, RMSD=3.69 Å.

6.5.3 Validation – pose prediction in self- and cross-docking

To verify the success of our protocol, we ran docking studies since this was essentially the purpose that we will investigate with the mutated side-chains for protein engineering. Our proposed biocatalysis software is a combination of protein engineering and protein reaction modeling.[88] For the latter, we require accurate docking results in order to position the ligand correctly. Consequently, our docking protocol could not be inhibited by an unrealistic binding cavity created from unrealistic side-chain mutations. Therefore, we tested the ability of our self-mutations not only to reproduce crystal structures but also docking results or perhaps improve them. In fact, with no practical reference, it would have been difficult to know whether the low RMSD we measured is low enough for practical applications. The available libraries have often been used for protein modeling in which the side chain conformation is not as important as the overall folding. In our case, moving an atom by as little as 1.0 Å can be detrimental to the substrate binding.

Ultimately, the accuracy of docking a small molecule into a binding site must be conserved upon self-mutation in order to instill confidence in the ability of protein engineering software when molecular dynamics simulation is not an option due to computational power costs. Using a previously established,[24] appropriate screening set of

230 PDB structures that cover a variety of shapes, sizes, functions and ligands (see Appendix 4 for list of PDB codes), we carried out self-docking experiments with the latest version (v. 3.7) of the FITTED software. Subsequently, using PROCESS, a binding site residue within 3 Å of the ligand was identified and then, using our new rotamer library, self-mutated. The self-docking was repeated with the "mutant" structure. To our delight, the average RMSD of the self-mutated residue side-chains, 0.88 Å, was similar to that of the development stages, as was the amino acid distribution (Table 6.3), and furthermore, the success for docking was virtually unchanged, within 2.2%, or 5 fewer successes (using a 2 Å metric, the standard in the field). The causes for failure or success can be attributed to a number of influencing factors. First, let us examine an example where self-mutation improved the docking result (Figure 6.8, top left). In this carbonic anhydrase, a PHE was randomly mutated (RMSD=4.40 Å) and opened additional space for the ligand to be docked correctly. In some cases, the docking was better after self-mutation, however the side chain conformation was almost unchanged (RMSD=0.06 Å) (Figure 6.8, top right). This can once again be attributed to poor convergence or scoring issues due to very minor MM energy differences between the very different poses; the correct conformation was in fact seen but not the lowest in energy and this highly scored, poor conformation was not observed in the self-mutant docking. Unfortunately, the self-mutations can have negative consequences as well (Figure 6.8, bottom left). In this instance, the self-mutation of TYR is incorrect (RMSD=6.14 Å) and resulted in a poor docking orientation. Its placement allowed the ligand more space for a large bond rotation on the opposite side. Coincidentally, a docked pose with RMSD=1.50 Å, better than the wild type (WT) docking, was observed but poorly scored. Once again, a poorly docked pose, this time after self-mutation, can be unrelated to the poorly reconstructed side chain (Figure 6.8, bottom right). As an illustration, the self-mutation of ILE was excellent (RMSD=0.10 Å), yet the docked pose was completely inverted.

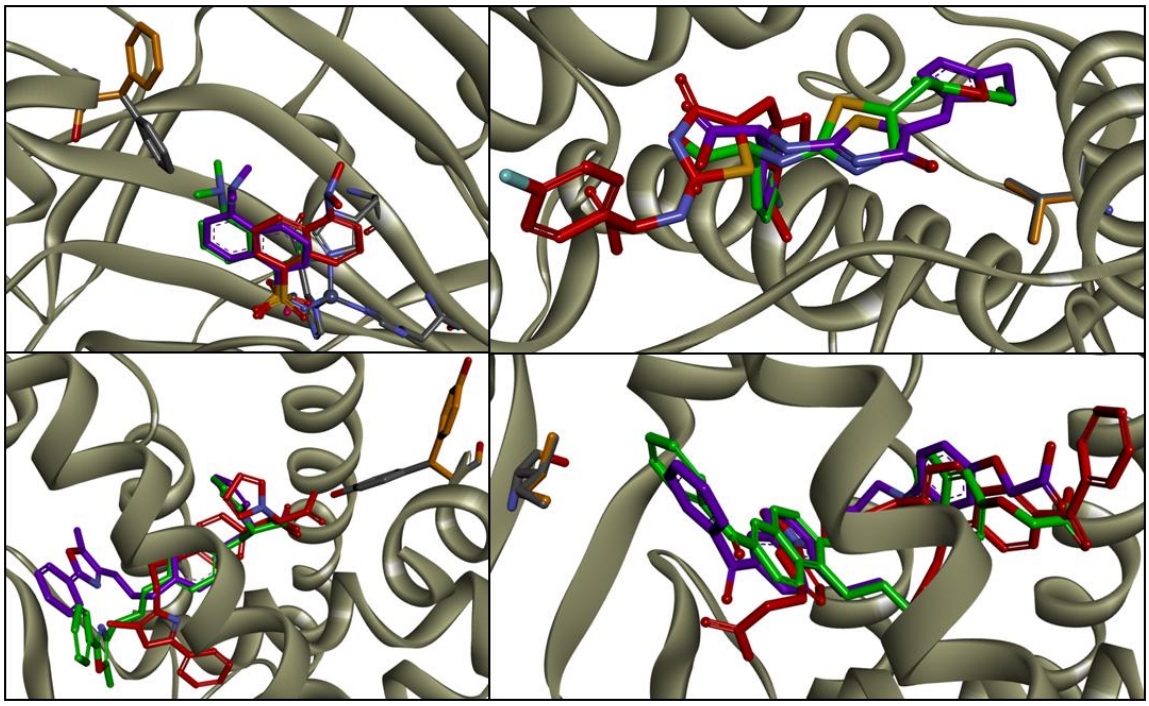


Figure 6.8. Self-docking successes and failures. Crystal structures in grey, crystal ligands in purple, hydrogen atoms removed for clarity. Top left: PDB 1OKL, PHE131; mutated PHE in orange, WT docking in red (RMSD=3.92), mutant docking in green (RMSD=1.50). Top right: PDB 2HWI, VAL485; mutated VAL in orange, WT docking in red (RMSD=7.50), mutant docking in green (RMSD=1.20). Bottom left: PDB 2Q8S, TYR473; mutated TYR in orange, WT docking in green (RMSD=2.91), mutant docking in red (RMSD=6.08). Bottom right: PDB 2HWQ, ILE249; mutated ILE in orange, WT docking in green (RMSD=1.69), mutant docking in red (RMSD=9.25).

In order to investigate if the self-mutation was the most likely culprit when it comes to poor docked poses in the modified protein structures, we broke down the results into the four categories: when both docked poses were accurate, when either and only one was correct and when both failed. The results are summarized in Table 6.8. What we observed was that the most significant average RMSD difference in side chains occurs when the WT docking was wrong, but the post-self-mutation docking was acceptable. The lowest average RMSD was seen when the WT docking was correct and the self-mutation appeared to have a negative impact. The diminished change in the side chain conformation suggested that these were convergence or MM scoring issues primarily.

Table 6.8. Summary of side chain mutation RMSD and consequential docking results

	Self-Docking			Cross-Docking	
	Instances		Average side-chain RMSD (Å)	Instances	
WT good / MUT bad	19	(8.2%)	0.614	64	(7.6%)
WT good / MUT good	117	(50.9%)	0.800	356	(42.3%)
WT bad / MUT good	14	(6.1%)	1.677	50	(5.9%)
WT bad / MUT bad	80	(34.8%)	0.942	372	(44.2%)

To further validate the effectiveness of mutation prediction we applied the same protocol to cross-docking, a true test for any docking software package. Again, both the WT and self-mutated structures were used, the same ones having been generated in the self-docking test. Our self-mutated structures consistently performed well compared to the WT structures with 842 complexes that were formed, demonstrating a difference in accuracy of only 1.6%, or 14 fewer successes (using a 2.6 Å metric[24]). As seen in Table 6.8, there remains the distinction where the self-mutation had an effect on the docked pose, both for good and for bad. A few examples are show in Figure 6.9. First, a reasonable self-mutation of GLU (RMSD=0.77 Å) resulted in a complete displacement of the docked pose; the correct conformation was not observed at all. This was likely due to a disruption of the hydrogen bond that was formed between the GLU and the ligand, albeit after such a small shift. It is worth noting that the self-docking with this protein structure suffered the same consequences (WT: RMSD =0.31 Å, MUT: RMSD=9.58 Å). Self-mutation also had positive impacts in the cross-docking. In the example shown, MET (RMSD=1.18 Å) was shifted slightly over, giving proper room for the docked pose upon cross-docking. Without this minor motion, the ligand was flipped and the pose was incorrect. This followed the same general trend observed in the self-docking experiments.

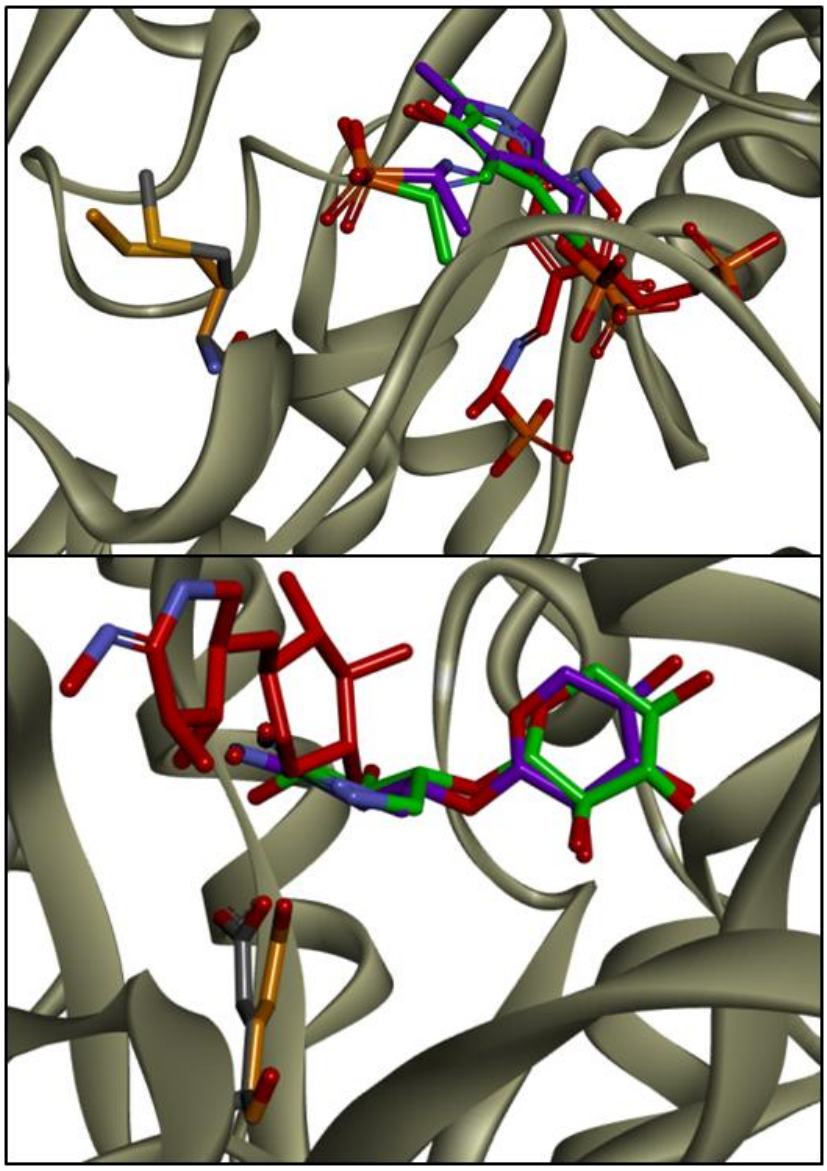


Figure 6.9. Cross-docking successes and failures. Crystal structures in grey, crystal ligands in purple, hydrogen atoms removed for clarity. Top: PDB 1M0Q docked into 1M0O, MET57; mutated MET in orange, WT docking in red (RMSD=6.27), mutant docking in green (RMSD=1.09). Bottom: PDB 1FH9 docked into 1FH8, GLU233; mutated GLU in orange, WT docking in green (RMSD=0.53), mutant docking in red (RMSD=7.70).

The overall docking results are summarized in Table 6.9. The docking accuracy, regardless of RMSD metric, remained quite similar over a long range of RMSD (Figure 6.10) demonstrating that our side-chain mutation protocol was satisfactory in the sense that it did not largely impact the docking ability of FITTED and thus should prove viable for our protein engineering software.

Table 6.9. Docking results for both wild-type and self-mutated protein structures

	Self-Docking		Cross-Docking	
	WT	Self-mutated	WT	Self-mutated
Instances (N)	230	230	842	842
Average RMSD of docked pose (Å)	2.58	2.78	3.74	3.84
Success (N)	136	131	420	406
Success (%)	59.1	57.0	49.9	48.2

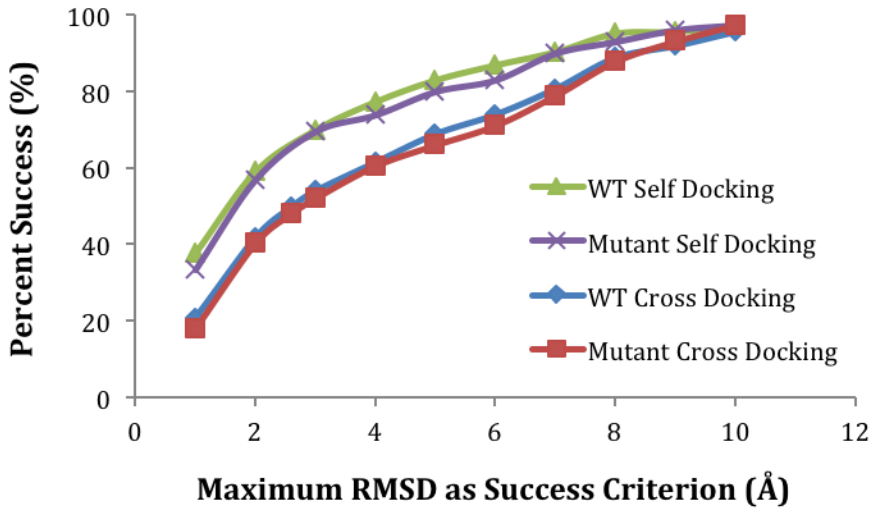


Figure 6.10. Overall accuracy of docking to both WT and mutant structures over a range of success criteria

6.6 Conclusion

In conclusion, our docking program PREPARE has first been modified to collect structural data from available protein/ligand crystal structures. Additional routines were implemented to properly rebuild mutated side-chain amino acids within the binding site of protein structures. This was accomplished using a combination of statistical and MM selection criteria upon building a new rotamer library. The generation of conformation clusters remains a separate protocol and could be applied to a focused set of structures if this is required. The high quality of the produced mutations was confirmed by both low deviation from crystal structures and accurate docking of a set of protein/ligand

complexes. Based on the success of this technique, it is likely that this approach could be applied to actual side-chain mutations in the context of protein engineering. Currently, an algorithm to guide the backbone motion upon changing a residue is under investigation with the goal of combining these two techniques. Re-building random side-chains prior to docking experiments may provide better docking accuracy by essentially cleaning the structures; however a significantly more in-depth study is required.

6.7 Experimental

6.7.1 Construction of the "training" and testing sets

The sets of PDB codes were downloaded from the Platinum (Protein-ligand affinity change upon mutation) database website (http://bleoberis.bioc.cam.ac.uk/platinum/browse) and further processed to select only those with single point mutations within the binding site. The exhaustive list of binding site residues to mutate was identified using PROCESS and a ligand distance of 7 Å (Ligand_Cutoff 7). Sets can be found in Appendix 4.

6.7.2 Preparation of the protein files

FORECASTER routines, PREPARE and PROCESS, were applied using the specific keyword identifying mutation (Mutation 1), the appropriate flexibility mode (Flexibility_Mode Quick), the rotamer library resolution (StatsResolution 120) and other parameters set to the default. The RMSD was calculated by extracting the appropriate side-chain residue coordinates and comparing the mutated and wild-type conformations.

6.7.3 Construction of the docking sets

The sets of PDB were obtained from a previous study[24] with two omissions due to limitations of the current mutation software implementation; molecules binding between two protein domains (i.e. between chain A and chain B), cannot be properly handled in the current version. The PDB codes can be found in Appendix 4.

6.7.4 Preparation of the protein files for docking

MATCH-UP, PREPARE and PROCESS were applied with the default parameters and the specific keyword identifying metalloenzymes when applicable.

6.7.5 Docking with FITTED

Default parameters implemented in FITTED were used. The subversion 4207 of the FORECASTER platform and programs has been used for this study.

6.8 Acknowledgements

We thank FRQ-NT (Équipe program to NM and scholarship to JP) for financial support. Calcul Québec and Compute Canada are acknowledged for generous CPU allocations.

6.9 References

1. Li, Y.; Cirino, P. C., Recent advances in engineering proteins for biocatalysis. *Biotech*. *Bioeng*. **2014,** *111* (7), 1273-1287.
2. Pasteur, L., Mémoire sur la fermentation de l'acide tartrique. *C*. *R*. *Acad*. *Sci*. **1858,** *46*, 615-618.
3. Reetz, M. T., Biocatalysis in organic chemistry and biotechnology: Past, present, and future. *J*. *Amer*. *Chem*. *Soc*. **2013,** *135* (34), 12480-12496.
4. Clouthier, C. M.; Pelletier, J. N., Expanding the organic toolbox: A guide to integrating biocatalysis in synthesis. *Chemical Society Reviews* **2012,** *41* (4), 1585-1605.
5. Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K., Engineering the third wave of biocatalysis. *Nature* **2012,** *484* (7397), 185-194.

6. Hçhne, M.; Bornscheuer, U. T., Protein engineering from scratch is maturing. *Angew*. *Chem*. *Int*. *Ed*. **2014,** *53* (5), 1200-1202.

7. Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D. S.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008,** *453* (7192), 190-195.

8. Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N., Computational enzyme design. *Angew*. *Chem*. *Int*. *Ed*. **2013,** *52* (22), 5700-5725.

9. Chadha, N.; Tiwari, A. K.; Kumar, V.; Milton, M. D.; Mishra, A. K., In silico thermodynamics stability change analysis involved in BH 4 responsive mutations in phenylalanine hydroxylase: QM/MM and MD simulations analysis. *J*. *Biomol*. *Struct*. *Dyn*. **2015,** *33* (3), 573-583.

10. Bhattacharya, S.; Lee, S.; Grisshammer, R.; Tate, C. G.; Vaidehi, N., Rapid Computational Prediction of Thermostabilizing Mutations for G Protein-Coupled Receptors. *J*. *Chem*. *Theo*. *Comput*. **2014,** *10* (11), 5149-5160.

11. Giollo, M.; Martin, A. J.; Walsh, I.; Ferrari, C.; Tosatto, S. C., NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genom*. **2014,** *15*, S7.

12. Chen, J.; Yu, H.; Liu, C.; Liu, J.; Shen, Z., Improving stability of nitrile hydratase by bridging the salt-bridges in specific thermal-sensitive regions. *J*. *Biotech*. **2013,** *164* (2), 354-362.

13. Park, H. J.; Joo, J. C.; Park, K.; Kim, Y. H.; Yoo, Y. J., Prediction of the solvent affecting site and the computational design of stable Candida antarctica lipase B in a hydrophilic organic solvent. *J*. *Biotech*. **2013,** *163* (3), 346-352.

14. Huang, X.; Gao, D.; Zhan, C.-G., Computational design of a thermostable mutant of cocaine esterasevia molecular dynamics simulations. *Org*. *Biomol*. *Chem*. **2011,** *9* (11), 4138-4143.

15. Joo, J. C.; Pohkrel, S.; Pack, S. P.; Yoo, Y. J., Thermostabilization of Bacillus circulans xylanase via computational design of a flexible surface cavity. *J. Biotech.* **2010,** *146* (1–2), 31-39.

16. Tian, J.; Wang, P.; Gao, S.; Chu, X.; Wu, N.; Fan, Y., Enhanced thermostability of methyl parathion hydrolase from Ochrobactrum sp. M231 by rational engineering of a glycine to proline mutation. *FEBS Journal* **2010,** *277* (23), 4901-4908.

17. Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D., A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003,** *332* (2), 449-460.

18. Roiban, G.-D.; Reetz, M. T., Expanding the toolbox of organic chemists: directed evolution of P450 monooxygenases as catalysts in regio- and stereoselective oxidative hydroxylation. *Chem. Comm.* **2015,** *51* (12), 2208-2224.

19. McIntosh, J. A.; Farwell, C. C.; Arnold, F. H., Expanding P450 catalytic reaction space through evolution and engineering. *Curr. Opin. Chem. Biol.* **2014,** *19* (0), 126-134.

20. McIntosh, J. A.; Coelho, P. S.; Farwell, C. C.; Wang, Z. J.; Lewis, J. C.; Brown, T. R.; Arnold, F. H., Enantioselective Intramolecular C□H Amination Catalyzed by Engineered Cytochrome P450 Enzymes In Vitro and In Vivo. *Angew. Chem. Int. Ed.* **2013,** *52* (35), 9309-9312.

21. Seifert, A.; Vomund, S.; Grohmann, K.; Kriening, S.; Urlacher, V. B.; Laschat, S.; Pleiss, J., Rational design of a minimal and highly enriched CYP102A1 mutant library with improved regio-, stereo- and chemoselectivity. *ChemBioChem* **2009,** *10* (5), 853-861.

22. Oliver, C. F.; Modi, S.; Sutcliffe, M. J.; Primrose, W. U.; Lian, L.-Y.; Roberts, G. C. K., A Single Mutation in Cytochrome P450 BM3 Changes Substrate Orientation in a Catalytic Intermediate and the Regiospecificity of Hydroxylation. *Biochemistry* **1997,** *36* (7), 1567-1572.

23. Polic, V.; Auclair, K., Controlling substrate specificity and product regio- and stereo-selectivities of P450 enzymes without mutagenesis. *Bioorg*. *Med*. *Chem*. **2014,** *22* (20), 5547-5554.

24. Therrien, E.; Weill, N.; Tomberg, A.; Corbeil, C. R.; Lee, D.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 7. Impact of protein flexibility and water molecules on docking-based virtual screening accuracy. *J*. *Chem*. *Inf*. *Model*. **2014,** *54* (11), 3198-3210.

25. Corbeil, C. R.; Moitessier, N., Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J*. *Chem*. *Inf*. *Model*. **2009,** *49* (4), 997-1009.

26. Reetz, M. T.; Carballeira, J. D., Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat*. *Prot*. **2007,** *2* (4), 891-903.

27. Acevedo-Rocha, C. G.; Hoebenreich, S.; Reetz, M. T., Iterative saturation mutagenesis: A powerful approach to engineer proteins by systematically simulating darwinian evolution. In *Methods in Molecular Biology*, 2014; Vol. 1179, pp 103-128.

28. Chica, R. A.; Doucet, N.; Pelletier, J. N., Semi-rational approaches to engineering enzyme activity: Combining the benefits of directed evolution and rational design. *Curr*. *Opin*. *Biotech*. **2005,** *16* (4), 378-384.

29. Davey, J. A.; Chica, R. A., Multistate approaches in computational protein design. *Prot*. *Sci*. **2012,** *21* (9), 1241-1252.

30. Bommarius, A. S.; Blum, J. K.; Abrahamson, M. J., Status of protein engineering for biocatalysts: How to design an industrially useful biocatalyst. *Curr*. *Opin*. *Chem*. *Biol*. **2011,** *15* (2), 194-200.

31. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **2003,** *302* (5649), 1364-1368.

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

33. Arnold, K.; Kiefer, F.; Kopp, J.; Battey, J. N. D.; Podvinec, M.; Westbrook, J. D.; Berman, H. M.; Bordoli, L.; Schwede, T., The Protein Model Portal. *J*. *Struct*. *Func*. *Gen*. **2009,** *10* (1), 1-8.

34. Schneider, M.; Fu, X.; Keating, A. E., X-ray vs. NMR structures as templates for computational protein design. *Prot*. *Struct*. *Func*. *Bioinf*. **2009,** *77* (1), 97-110.

35. Allen, B. D.; Nisthal, A.; Mayo, S. L., Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. *Proc*. *Nat*. *Aca*. *Sci*. **2010,** *107* (46), 19838-19843.

36. Voigt, C. A.; Martinez, C.; Wang, Z. G.; Mayo, S. L.; Arnold, F. H., Protein building blocks preserved by recombination. *Nat*. *Struct*. *Biol*. **2002,** *9* (7), 553-558.

37. Meyer, M. M.; Silberg, J. J.; Voigt, C. A.; Endelman, J. B.; Mayo, S. L.; Wang, Z. G.; Arnold, F. H., Library analysis of SCHEMA-guided protein recombination. *Prot*. *Sci*. **2003,** *12* (8), 1686-1693.

38. Currin, A.; Swainston, N.; Day, P. J.; Kell, D. B., Synthetic biology for the directed evolution of protein biocatalysts: Navigating sequence space intelligently. *Chem*. *Soc*. *Rev*. **2015,** *44* (5), 1172-1239.

39. Johnson, L. B.; Huber, T. R.; Snow, C. D., Methods for library-scale computational protein design. In *Methods in Molecular Biology*, 2014; Vol. 1216, pp 129-159.

40. Bommarius, A. S.; Paye, M. F., Stabilizing biocatalysts. *Chem*. *Soc*. *Rev*. **2013,** *42* (15), 6534-6565.

41. Fox, R., Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J*. *Theo*. *Biol*. **2005,** *234* (2), 187-199.

42. Fox, R.; Roy, A.; Govindarajan, S.; Minshull, J.; Gustafsson, C.; Jones, J. T.; Emig, R., Optimizing the search algorithm for protein engineering by directed evolution. *Prot*. *Eng*. **2003,** *16* (8), 589-597.

43. Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S.; Grate, J.; Gruber, J.; Whitman, J. C.; Sheldon, R. A.; Huisman, G. W., Improving catalytic function by ProSAR-driven enzyme evolution. *Nat*. *Biotech*. **2007,** *25* (3), 338-344.

44. Zanghellini, A.; Jiang, L.; Wollacott, A. M.; Cheng, G.; Meiler, J.; Althoff, E. A.; Röthlisberger, D.; Baker, D., New algorithms and an in silico benchmark for computational enzyme design. *Prot*. *Sci*. **2006,** *15* (12), 2785-2794.

45. Kuhlman, B.; Baker, D., Native protein sequences are close to optimal for their structures. *Proc*. *Nat*. *Aca*. *Sci*. **2000,** *97* (19), 10383-10388.

46. Baker, D., Protein folding, structure prediction and design. *Biochem*. *Soc*. *Trans*. **2014,** *42* (2), 225-229.

47. Pearson, A. D.; Mills, J. H.; Song, Y.; Nasertorabi, F.; Han, G. W.; Baker, D.; Stevens, R. C.; Schultz, P. G., Trapping a transition state in a computationally designed protein bottle. *Science* **2015,** *347* (6224), 863-867.

48. Gordon, S. R.; Stanley, E. J.; Wolf, S.; Toland, A.; Wu, S. J.; Hadidi, D.; Mills, J. H.; Baker, D.; Pultz, I. S.; Siegel, J. B., Computational design of an α-gliadin peptidase. *J*. *Amer*. *Chem*. *Soc*. **2012,** *134* (50), 20513-20520.

49. Siegel, J. B.; Zanghellini, A.; Lovick, H. M.; Kiss, G.; Lambert, A. R.; St.Clair, J. L.; Gallaher, J. L.; Hilvert, D.; Gelb, M. H.; Stoddard, B. L.; Houk, K. N.; Michael, F. E.; Baker, D., Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science* **2010,** *329* (5989), 309-313.

50. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas Iii, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008,** *319* (5868), 1387-1391.

51. Verma, R.; Schwaneberg, U.; Roccatano, D., Computer-aided protein directed evolution: A review of web servers, databases and other computational tools for protein engineering. *Comput*. *Struct*. *Biotech*. *J*. **2012,** *2* (3).

52. Frembgen-Kesner, T.; Andrews, C. T.; Li, S.; Ngo, N. A.; Shubert, S. A.; Jain, A.; Olayiwola, O. J.; Weishaar, M. R.; Elcock, A. H., Parametrization of backbone flexibility in a Coarse-grained Force Field for Proteins (COFFDROP) derived from all-atom explicit-solvent molecular dynamics simulations of all possible two-Residue reptides. *J*. *Chem*. *Theo*. *Comput*. **2015,** *11* (5), 2341-2354.

53. Davey, J. A.; Chica, R. A., Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles. *Prot*. *Struct*. *Func*. *Bioinf*. **2014,** *82* (5), 771-784.

54. Compiani, M.; Capriotti, E., Computational and theoretical methods for protein folding. *Biochemistry* **2013,** *52* (48), 8601-8624.

55. Tokuriki, N.; Tawfik, D. S., Stability effects of mutations and protein evolvability. *Curr*. *Opin*. *Struct*. *Biol*. **2009,** *19* (5), 596-604.

56. Suárez, M.; Jaramillo, A., Challenges in the computational design of proteins. *J*. *Roy*. *Soc*. *Interf*. **2009,** *6* (SUPPL. 4), S477-S491.

57. Pantazes, R. J.; Grisewood, M. J.; Maranas, C. D., Recent advances in computational protein design. *Curr*. *Opin*. *Struct*. *Biol*. **2011,** *21* (4), 467-472.

58. Peisajovich, S. G.; Tawfik, D. S., Protein engineers turned evolutionists. *Nat*. *Meth*. **2007,** *4* (12), 991-994.

59. Baker, D., An exciting but challenging road ahead for computational enzyme design. *Prot*. *Sci*. **2010,** *19* (10), 1817-1819.

60. Lee, C.; Subbiah, S., Prediction of protein side-chain conformation by packing optimization. *J*. *Mol*. *Biol*. **1991,** *217* (2), 373-388.

61. Summers, N. L.; Karplus, M., Construction of side-chains in homology modelling. Application to the C-terminal lobe of rhizopuspepsin. *J*. *Mol*. *Biol*. **1989,** *210* (4), 785-811.

62. Chazelle, B.; Kingsford, C.; Singh, M., A semidefinite programming approach to side chain positioning with new rounding strategies. *Inf*. *Journ*. *Comput*. **2004,** *16* (4), 380-392.

63. Pierce, N. A.; Winfree, E., Protein design is NP-hard. *Prot*. *Eng*. **2003,** *15* (10), 779-782.

64. Ponder, J. W.; Richards, F. M., Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J*. *Mol*. *Biol*. **1987,** *193* (4), 775-791.

65. Bhuyan, M. S.; Gao, X., A protein-dependent side-chain rotamer library. *BMC Bioinf*. **2011,** *12 Suppl 14*.

66. Krivov, G. G.; Shapovalov, M. V.; Dunbrack Jr, R. L., Improved prediction of protein side-chain conformations with SCWRL4. *Prot*. *Struct*. *Func*. *Bioinf*. **2009,** *77* (4), 778-795.

67. Shetty, R. P.; De Bakker, P. I. W.; DePristo, M. A.; Blundell, T. L., Advantages of fine-grained side chain conformer libraries. *Prot*. *Eng*. **2003,** *16* (12), 963-969.

68. Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C., The penultimate rotamer library. *Prot*. *Struct*. *Func*. *Gen*. **2000,** *40* (3), 389-408.

69. Bower, M. J.; Cohen, F. E.; Dunbrack Jr, R. L., Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J*. *Mol*. *Biol*. **1997,** *267* (5), 1268-1282.

70. Dunbrack Jr, R. L.; Karplus, M., Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat*. *Struct*. *Biol*. **1994,** *1* (5), 334-340.

71. Dunbrack Jr, R. L.; Karplus, M., Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J*. *Mol*. *Biol*. **1993,** *230* (2), 543-574.

72. Francis-Lyon, P.; Koehl, P., Protein side-chain modeling with a protein-dependent optimized rotamer library. *Prot*. *Struct*. *Func*. *Bioinf*. **2014,** *82* (9), 2000-2017.

73. Harpole, T. J.; Grosman, C., Side-chain conformation at the selectivity filter shapes the permeation free-energy landscape of an ion channel. *Proc*. *Nat*. *Aca*. *Sci*. **2014,** *111* (31), E3196-E3205.

74. Renfrew, P. D.; Craven, T. W.; Butterfoss, G. L.; Kirshenbaum, K.; Bonneau, R., A rotamer library to enable modeling and design of peptoid foldamers. *J*. *Amer*. *Chem*. *Soc*. **2014,** *136* (24), 8772-8782.

75. Koehl, P.; Delarue, M., Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J*. *Mol*. *Biol*. **1994,** *239* (2), 249-275.

76. Alexander, N. S.; Stein, R. A.; Koteiche, H. A.; Kaufmann, K. W.; McHaourab, H. S.; Meiler, J., RosettaEPR: Rotamer Library for Spin Label Structure and Dynamics. *PLoS ONE* **2013,** *8* (9).

77. Miao, Z.; Cao, Y.; Jiang, T., RASP: Rapid modeling of protein side chain conformations. *Bioinformatics* **2011,** *27* (22), 3117-3122.

78. Lu, M.; Dousis, A. D.; Ma, J., OPUS-Rota: A fast and accurate method for side-chain modeling. *Prot*. *Sci*. **2008,** *17* (9), 1576-1585.

79. Peterson, R. W.; Dutton, P. L.; Wand, A. J., Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Prot*. *Sci*. **2004,** *13* (3), 735-751.

80. Liang, S.; Grishin, N. V., Side-chain modeling with an optimized scoring function. *Prot*. *Sci*. **2002,** *11* (2), 322-331.

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

82. Mendes, J.; Baptista, A. M.; Carrondo, M. A.; Soares, C. M., Implicit solvation in the self-consistent mean field theory method: Sidechain modelling and prediction of folding free energies of protein mutants. *J*. *Comput*. *Mol*. *Des*. **2001,** *15* (8), 721-740.

83. Moghadasi, M.; Mirzaei, H.; Mamonov, A.; Vakili, P.; Vajda, S.; Paschalidis, I. C.; Kozakov, D., The impact of side-chain packing on protein docking refinement. *J*. *Chem*. *Inf*. *Model*. **2015,** *55* (4), 872-881.

84. Mashiach, E.; Schneidman-Duhovny, D.; Andrusier, N.; Nussinov, R.; Wolfson, H. J., FireDock: a web server for fast interaction refinement in molecular docking. *Nucl*. *Aci*. *Res*. **2008,** *36* (Web Server issue), W229-232.

85. Davey, J. A.; Chica, R. A., Optimization of rotamers prior to template minimization improves stability predictions made by computational protein design. *Prot*. *Sci*. **2015,** *24* (4), 545-560.

86. Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N., Integrating medicinal chemistry, organic/combinatorial chemistry, and computational chemistry for the discovery of selective estrogen receptor modulatorswith FORECASTER, a novel platform for drug discovery. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (1), 210-224.

87. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr, A. D., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J*. *Comp*. *Chem*. **2010,** *31* (4), 671-690.

88. Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.-D.; Moitessier, N., Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s. *J*. *Chem*. *Inf*. *Model*. **2012,** *52* (9), 2471-2483.

Chapter 7:

Conclusion and Perspective

7.1 Concluding remarks

In the world of chemoinformatics and software development for chemists, the usability and user-friendliness is often overlooked when it comes to simulating routine experiments, as opposed to visualization software that normally interfaces to other software to carry out experiments. Within the research presented in this thesis, this has been a focus; automating and simplifying software was just as crucial as developing the methodology. Moreover, the automation can be even more difficult to encode than the chemical theory, thus requiring a significant effort. Consequently, a successful research contribution in this domain is not only the accuracy of the proposed software but the reflection of its usability by it being employed by research groups worldwide.

At the risk of sounding unhinged, the VIRTUAL CHEMIST platform could modernize synthetic chemistry and every field that it influences. The idea behind this program is to aid organic chemists in planning and focusing their synthesis for asymmetric catalysis. There is no such automated software that guides chemistry from start to finish and its development was, in fact, propelled by this need in previous work not described in this thesis.[1] Chapter 2 presents the evaluation of the steps involved in experimental synthesis, from ordering chemicals to testing a catalyst, the automation of each of these steps virtually, and their integration into one easy-to-use online platform. In short, the software will search any given catalog for desired scaffolds, based on a reaction scheme (the only other input from the experimentalist), which are then used to virtually create organocatalysts by combinatorial chemistry. It then builds a 3-dimensional transition state for a given asymmetric reaction with these catalysts and finally evaluates and ranks the enantioselectivity with good accuracy. The VIRTUAL CHEMIST is still in its infancy and its integration within the Moitessier group to discover new asymmetric catalysts is beginning. Once polished, such software could transform the purpose of computational chemistry from rationalization to instruction and ultimately bring it to the forefront of

chemical synthesis. The VIRTUAL CHEMIST is now a registered copyright and will be distributed by Molecular Forecaster Inc.

The work presented in chapter 4 regarding zinc metalloenzymes sought to develop a protocol for these enzymes in order to improve the docking software, FITTED. The identification of ongoing biochemistry within the binding cavity jumped the accuracy 10-20% for these enzymes and was recently featured in an account on the development of FITTED. However, not only did this research improve the in-house docking software, it preceded the research of other groups,[2-4] AutoDock for example (one of the most widely used), seeking the same advances in their own software.[5] Furthermore, the approach, identifying and modeling dynamic chemical and biochemical phenomena, was commended by the community. Consequently, the software has been used by the Gleason group[6-7] and has been implemented in the teaching curriculum at McGill University in the undergraduate research labs. The software is also distributed by Molecular Forecaster Inc. and free for academic research.

The work presented in chapter 5 is often cited as being a unique software development in the field of site of metabolism prediction. The trivalent approach taken has been praised as insightful.[8-14] Additionally, it demonstrates some of the best accuracy in the community and has been used by experimentalists in our group on medicinal chemistry projects involving prolyl-oligopeptidase,[15] and abroad to evaluate kinase inhibitors.[16] The abilities of the software have further been exploited to identify sites of metabolism and then predict the consequential metabolites formed.[17] Its use exhibits the reliability of the predictions as well as the user-friendliness of a computational tool. Moreover, Molecular Forecaster Inc. markets and sells the software while making it freely available to academic groups worldwide.

Building off of the advances reported in chapter 5, chapter 6 presents efforts towards an accurate, an eventually automated, protein engineering software. The versatility of this computational tool has yet to be fully developed or explored. The foundation has been built towards single point mutations whereas most research groups have been relatively unsuccessful with different virtual engineering approaches.

Overall, this thesis presents a body of work based on a philosophy that computational modeling should be: (1) guided by reason and chemical justification and (2) relevant to the laboratory chemists – produces accurate research and software that is employed by organic and medicinal chemists. The significance of this research led the writing of two invited reviews presented in chapters 1 and 3, demonstrating the considerable interest from the scientific community. This thesis aims to bring this philosophy to future research projects in chemical theory and computational chemistry.

7.2 Future opportunities

Organocatalysis in computational chemistry is a budding field of research with many untapped potential applications. In the context of the developments presented in this thesis, there are many improvements that could be envisioned. For one, the overall speed of this software has not reached its maximum. The novel string, or genotype, representation of a molecule resulted in relatively efficient comparisons; however the number of comparisons, especially when searching for duplicates, can be enormous. A more focused searching algorithm rather than exhaustive assessments could solve this problem. Furthermore, one could imagine a software protocol to identify "hotspots" that are more characteristic as opposed to the implemented "rare atom" technique as a starting point. Moreover, an optional, automated similarity clustering protocol should be included to limit the size of chemical libraries and thus reduce the runtime without sacrificing meaningful and interesting results.

From a new application perspective, this software platform and automated chemical tool could be used to generate new mechanistic insights. ACE could be "inverted" to search for transition states – the output would be the transition state, or a list of plausible ones – given reacting and product configurations input by the user. While the conformation space and 3D space can be expansive, reactive sites on molecules (where a bond is likely to be broken or formed) can already be identified, similarly to IMPACTS, the only difference being the catalytic machinery is not fixed as it is within the enzyme. Optimization algorithms can be written to search potential reactive states and the current ACE technology can be used to evaluate reaction barriers. While these activation energies

may not be accurate in absolute value, they have been shown to be able to rank enantioselectivity in a relative fashion and therefore, with a normalization factor for different reaction types, are promising for ranking the transition states themselves.

Taking this notion of mechanistic discovery one step further, it could be possible to apply the VIRTUAL CHEMIST in reverse to create a retrosynthetic analysis tool. As opposed to being given the reaction scheme, this would be the unknown in the experiment. If given a chemical product and a library of known chemical reactions and their associated transition states, it may be possible to work backwards through ACE to arrive at a library of catalysts and substrates. Using reactive sites within these molecules, REACT2D could be reversed in a similar fashion with FINDERS used to identify catalog molecules for production of the catalysts. These are just some of the ways the VIRTUAL CHEMIST platform could be expanded for new capabilities and used to explore the vast chemical space.

In-silico biocatalysis is far from its end-point and more research is required to obtain an accurate, reliable protein engineering software package. It appears that modeling chemical reactions within binding sites has reached acceptable precision. Unfortunately, protein structures are large biological systems with many long-range effects dominating the functionality in some cases. Currently, research is ongoing in the Moitessier group to supplement the side-chain mutation protocols presented in this thesis with a simulation of backbone motion upon residue mutation. Once accomplished, the three algorithms (side-chain mutation, backbone motion, IMPACTS), must be integrated and validated by modeling existing wild-type and corresponding mutant crystal structures. Subsequently, testing can begin to match reactivity to the data provided by IMPACTS regarding the activation barriers of the oxidations. Only once this milestone is achieved can new biochemical reactions be conceived. Additionally, multi-point mutations may be a possibility however this concept entails more structural motions, i.e. more technical difficulties.

In general, computational chemistry, when focusing on applicability and usability within the organic chemistry community, will continue to be limited by the restrictions imposed by molecular mechanics since quantum mechanics is neither time-efficient nor

easily applied. Molecular mechanics must be improved for metal atoms and reach accuracies better than 1 kcal/mol (sometimes insufficient to distinguish two different states) to be applied to many relevant catalytic cycles; until then, computational chemists will run simulations to rationalize experimental findings and never guide synthetic design. Furthermore, there is a psychological hurdle to overcome for experimentalists and the community at large to accept computational data. There is seemingly no amount of laboratory data for a satisfactory validation of a modeling technique; this is in part due to the required shortcuts for time-sensitivity and in part due to a lack of trust. Through the experiences reported in this thesis, it appears that it will remain difficult for this barrier to be broken for the foreseeable future. It is not all bleak, however, computational insights are often sought after and remain an active component in a majority of high impact synthetic and medicinal chemistry projects and the field will continue to prosper and blossom.

7.3 References

1. Bezanson, M.; Pottel, J.; Bilbeisi, R.; Toumieux, S.; Cueto, M.; Moitessier, N., Stereo- and regioselective synthesis of polysubstituted chiral 1,4-oxazepanes. *J*. *Org*. *Chem*. **2013,** *78* (3), 872-885.

2. Chaskar, P.; Zoete, V.; Röhrig, U. F., Toward on-the-fly quantum mechanical/molecular mechanical (QM/MM) docking: Development and benchmark of a scoring function. *J*. *Chem*. *Inf*. *Model*. **2014,** *54* (11), 3137-3152.

3. Hou, X.; Du, J.; Liu, R.; Zhou, Y.; Li, M.; Xu, W.; Fang, H., Enhancing the sensitivity of pharmacophore-based virtual screening by incorporating customized ZBG features: A case study using histone deacetylase 8. *J*. *Chem*. *Inf*. *Model*. **2015,** *55* (4), 861-871.

4. Martin, D. P.; Blachly, P. G.; McCammon, J. A.; Cohen, S. M., Exploring the influence of the protein environment on metal-binding pharmacophores. *J*. *Med*. *Chem*. **2014,** *57* (16), 7126-7135.

5. Santos-Martins, D.; Forli, S.; Ramos, M. J.; Olson, A. J., AutoDock4Zn: An improved AutoDock force field for small-molecule docking to zinc metalloproteins. *J*. *Chem*. *Inf*. *Model*. **2014,** *54* (8), 2371-2379.

6. Kaldre, D.; Wang, T. T.; Fischer, J.; White, J. H.; Gleason, J. L., Optimization of histone deacetylase inhibitor activity in non-secosteroidal vitamin D-receptor agonist hybrids. *Bioorg*. *Med*. *Chem*. **2015,** *23* (15), 5035-5049.

7. Mendoza-Sanchez, R.; Cotnoir-White, D.; Kulpa, J.; Pottel, J.; Moitessier, N.; Mader, S.; Gleason, J. L., Design, synthesis and evaluation of antiestrogen and histone deacetylase inhibitor molecular hybrids. *Bioorg*. *Med*. *Chem*. **2015**, submitted.

8. Huang, T. W.; Zaretzki, J.; Bergeron, C.; Bennett, K. P.; Breneman, C. M., DR-Predictor: Incorporating flexible docking with specialized electronic reactivity and machine learning techniques to predict CYP-mediated sites of metabolism. *J*. *Chem*. *Inf*. *Model*. **2013,** *53* (12), 3352-3366.

9. Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G., Predicting drug metabolism: Experiment and/or computation? *Nat*. *Rev*. *Drug*. *Disc*. **2015,** *14* (6), 387-404.

10. Nielsen, L. M.; Linnet, K.; Olsen, L.; Rydberg, P., Prediction of cytochrome P450 mediated metabolism of designer drugs. *Curr*. *Top*. *Med*. *Chem*. **2014,** *14* (11), 1365-1373.

11. Olsen, L.; Oostenbrink, C.; Jørgensen, F. S., Prediction of cytochrome P450 mediated metabolism. *Advanced Drug Delivery Reviews* **2015,** *86*, 61-71.

12. Raunio, H., Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes. *Frontiers in Pharmacology* **2015,** *6* (MAY).

13. Rydberg, P., Reactivity-Based Approaches and Machine Learning Methods for Predicting the Sites of Cytochrome P450-Mediated Metabolism. In *Drug Metabolism Prediction*, 2014; pp 265-292.

14. Tyzack, J. D.; Williamson, M. J.; Torella, R.; Glen, R. C., Prediction of cytochrome P450 xenobiotic metabolism: Tethered docking and reactivity derived from ligand molecular orbital analysis. *J*. *Chem*. *Inf*. *Model*. **2013,** *53* (6), 1294-1305.

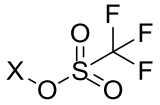
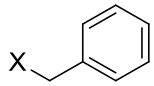
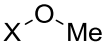
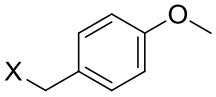
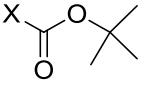
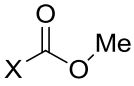
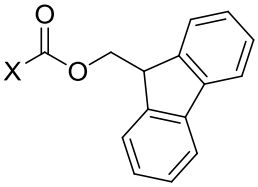
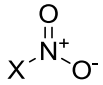
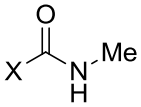
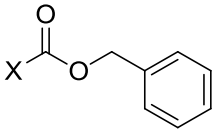
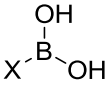
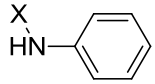
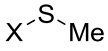
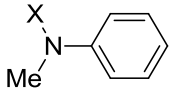
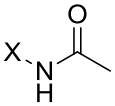
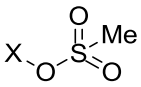
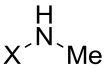
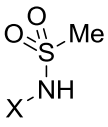
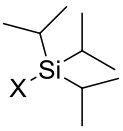
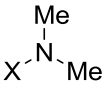
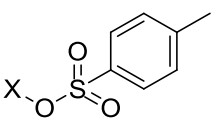
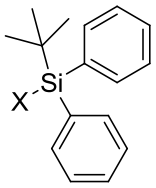
15. Schiavini, P.; Pottel, J.; Moitessier, N.; Auclair, K., Metabolic Instability of Cyanothiazolidine-Based Prolyl Oligopeptidase Inhibitors: A Structural Assignment Challenge and Potential Medicinal Chemistry Implications. *ChemMedChem* **2015,** *10* (7), 1174-1183.

16. Bernard-Gauthier, V.; Aliaga, A.; Boudjemeline, M.; Hopewell, R.; Kostikov, A.; Rosa-Neto, P.; Thiel, A.; Schirrmacher, R., Syntheses and evaluation of carbon-11- and fluorine-18-radiolabeled pan-tropomyosin receptor kinase (Trk) inhibitors: Exploration of the 4-aza-2-oxindole scaffold as Trk PET imaging agents. *ACS Chemical Neuroscience* **2015,** *6* (2), 260-276.

17. Tomberg, A.; Pottel, J.; Liu, Z.; Labute, P.; Moitessier, N., Investigation and Prediction of the P450-Mediated Oxidation of Aromatic Compounds into Highly Toxic Epoxide Derivatives. *Angew*. *Chem*. *Int*. *Ed*. **2015**, in press.

Appendix 1

Table A1.1 List of available protecting and leaving groups. X only shown for clarity.

Label	2D Structure	Label	2D Structure	Label	2D Structure
H		Et		CN	
CF3		tBu		OTf	
F		Ac		Bn	
Cl		OMe		PMB	
Br		OH		Boc	
I		COOMe		Fmoc	
NO2		CONHMe		Cbz	
NH2		BOH2		NHPh	
Ph		SMe		NMePh	
NHAc		OMs		TMS	
NHMe		NMs		TIPS	
NMe2		Ots		TBDPS	
Me					

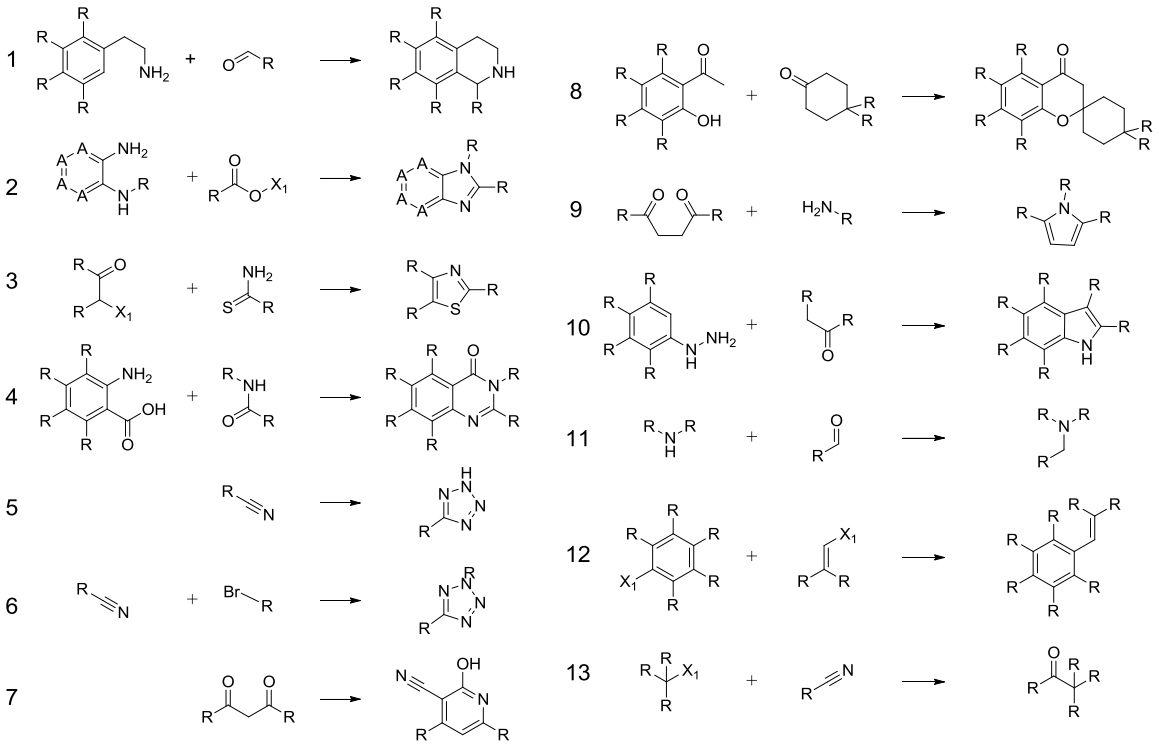


Figure A1.1. List of chemical reactions used to validate FINDERS and REACT2D.

Appendix 2

Table A2.1. Set used for the validation study. The ones used for QM studies are shown in bold

Entry	PDB code	Entry	PDB code	Entry	PDB code	Entry	PDB code
1	1a85	32	1jj9	63	2gc1	94	**3kgq**
2	1a86	33	**1kbc**	64	2gc2	95	3kne
3	**1b3d**	34	1lde	65	2oc2	96	3l3n
4	1b8y	35	1mmb	66	2ow6	97	3m04
5	1biw	36	**1mmp**	67	2qdm	98	**3max**
6	1bn4	37	1mmq	68	2tmn	99	3nxq
7	1bnn	38	1mnc	69	2usn	100	3p25
8	1bqo	39	1o86	70	**2vqj**	101	3p58
9	1bzm	40	1okl	71	**2wd2**	102	3p5a
10	1bzs	41	1ps3	72	2x8z	103	3qyk
11	**1caq**	42	1r33	73	2x94	104	**3s71**
12	1cbx	43	1r34	74	2x96	105	3s8x
13	1ciz	44	1r41	75	**2x97**	106	3tmn
14	1cxv	45	1sln	76	2xhm	107	3v2m
15	1d8m	46	**1t67**	77	**2zxg**	108	456c
16	1de5	47	1t69	78	3b2p	109	4dwv
17	1fbl	48	1thl	79	3b34	110	4tln
18	1g4o	49	1tlp	80	3bto	111	4tmn
19	1hdq	50	1tmn	81	3bup	112	5tln
20	1hee	51	1tqt	82	**3c0z**	113	5tmn
21	**1hfc**	52	1ttm	83	3c10	114	6cpa
22	1hfs	53	1usn	84	3cpa	115	6tmn
23	1hww	54	**1uze**	85	3d4z	116	7cpa
24	1hxk	55	**1uzf**	86	3dx2	117	7tln
25	1hy7	56	1w22	87	3ejp	118	830c
26	**1i76**	57	2c6n	88	3ejs	119	8cpa
27	1iy7	58	2ctc	89	3f07	120	8tln
28	1jan	59	2fl8	90	3fgd	121	966c
29	**1jao**	60	2fla	91	3fvl		
30	**1jap**	61	2f7r	92	3fx6		
31	1jaq	62	2gc0	93	3il u		

Appendix 3

A3.1 Accuracy of IMPACTS using top 1 to 4 as metrics

Table A3.1. Accuracy[a] of IMPACTS in predicting an observed SoM for respective datasets.

		Top 1			Top 2		
		Rand.[c]	E_{act}[d]	IMPACTS[e]	Rand.[c]	E_{act}[d]	IMPACTS[e]
CYP1A2	137	16	39	59	31	59	77
CYP2C9	129	15	38	56	29	59	79-82
CYP2D6	157	14	31	56	27	49	76
CYP3A4	293	15	42	50-53	28	66	72-75
All 4	716	15	37	49	28	60	77
		Top 3			Top 4		
		Rand.[c]	E_{act}[d]	IMPACTS[e]	Rand.[c]	E_{act}[d]	IMPACTS[e]
CYP1A2	137	44	81	88	56	85	91
CYP2C9	129	41	74	88-90	53	80	92-94
CYP2D6	157	38	66	86	49	76	90
CYP3A4	293	40	81	82-84	50	87	87-89
All 4	716	41	74	86	52	82	90

[a] % of molecules with an observed SoM in the predicted one, two, three or four SoMs referred to as top 1, top 2, top 3 and top 4. [b] Number of substrates in the set. [c] Random selection from the SoMs identified by IMPACTS. [d] Only the predicted reactivity of the SoMs is considered. [e] IMPACTS with a single crystal structure; if multiple structures were alternatively assessed, a range is given.

The computations have been done in triplicates using different seeds for the random number generator. The standard deviation never exceeds 3.0.

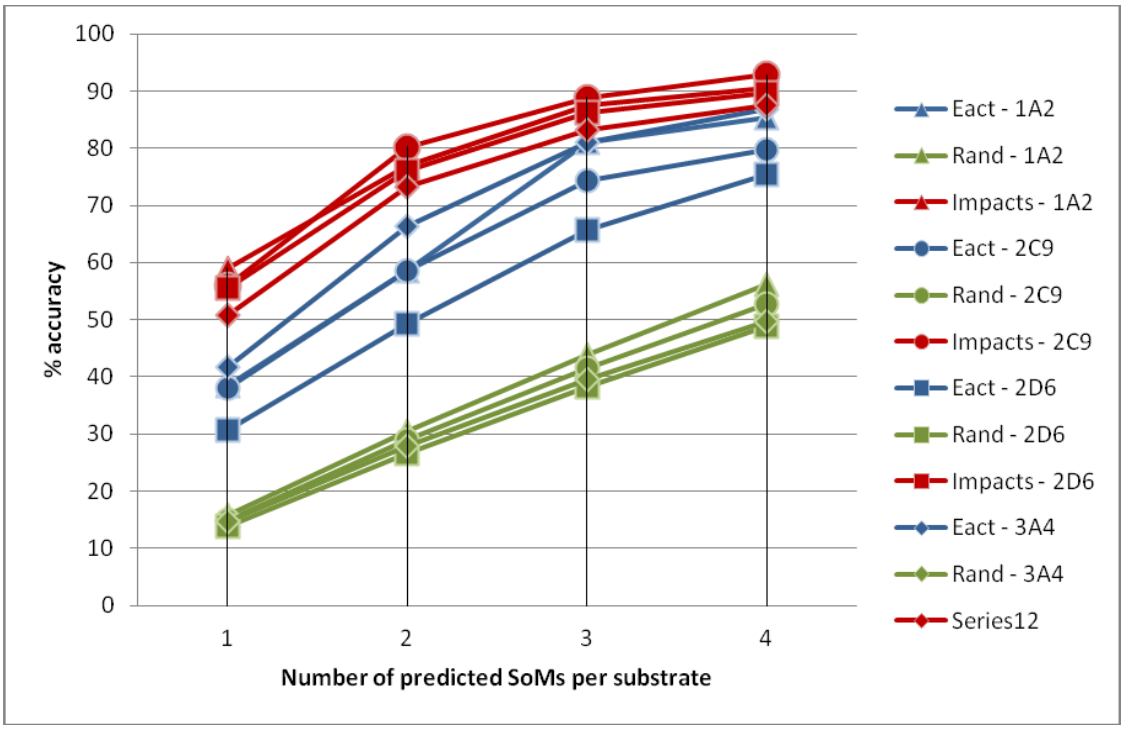


Figure A3.1. Accuracy of IMPACTS when Top 1, Top 2, Top 3 and Top 4 predicted SoMs are considered. Red: IMPACTS; Blue: Energy of activation only (ligand-based method); Green: random selection; Triangles: CYP 1A2; Circles: CYP2C9; Squares: CYP2D6 and Rhombus: CYP3A4.

A3.2 Construction of the testing sets

The testing sets were built with care to reduce the noise in the prediction assessment as discussed in the main text. These sets are given in separate files in mol2 format.

The pdb codes for the crystal structures are given below:

CYP1A2: 2HI4

CYP2C9: 1R9O, 1OG2 and 1OG5

CYP2D6: 3QM4

CYP3A4: 1TQN, 1W0E, 1W0F, 1W0G, 2J0D, 2V0M, 3NXU

A3.3 Experimental procedures

Default parameters implemented in IMPACTS have been used. IMPACTS has been implemented into our platform FORECASTER[1] for user-friendly use. The user can draw the substrate into a 2D sketcher and select the CYP enzyme with which to predict the SoM. FORECASTER will take care of adding hydrogens, generating a 3D structure and selecting the correct CYP files (Figure A3.2). IMPACTS and FORECASTER are accessible free of charge to academic users.

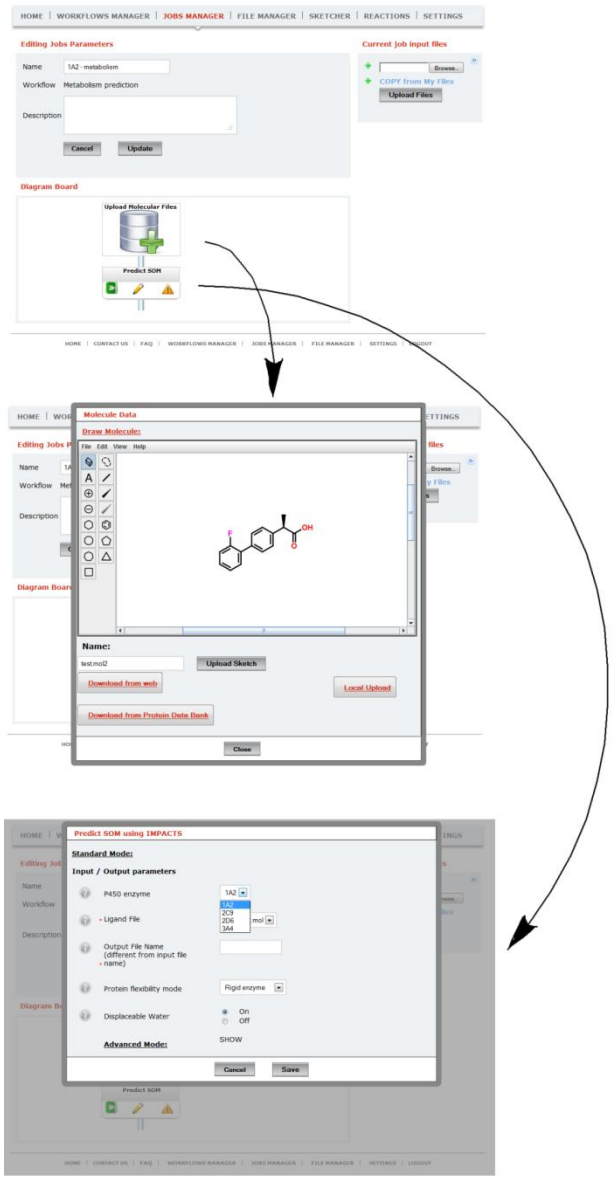


Figure A3.2. Integration in FORECASTER.

A3.4 References

1. Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N., Integrating Medicinal Chemistry, Organic/Combinatorial Chemistry, and Computational Chemistry for the Discovery of Selective Estrogen Receptor Modulators with Forecaster, a Novel Platform for Drug Discovery. *J. Chem. Inf. Model.* **2012,** *52*, 210-224.

Appendix 4

Table A4.1. PDB codes used for "training" conformational library selection

Entry	PDB code	Entry	PDB code	Entry	PDB code
1	1a4h	32	1x7z	63	3f8z
2	1amk	33	1yxi	64	3f91
3	1cnq	34	1zoa	65	3fjx
4	1flm	35	1zzr	66	3fk0
5	1flv	36	1zzs	67	3fs6
6	1jmf	37	1zzu	68	3g0e
7	1jmg	38	2aog	69	3g0f
8	1jmh	39	2nnp	70	3gug
9	1jmi	40	2ony	71	3ijw
10	1jqx	41	2onz	72	3lbo
11	1nja	42	2pym	73	3ld5
12	1njc	43	2pyn	74	3lep
13	1nje	44	2q63	75	3lqg
14	1ovd	45	2qfs	76	3lq1
15	1p6n	46	2qft	77	3lz5
16	1q6e	47	2rde	78	3m4h
17	1q6g	48	2tdm	79	3n0h
18	1qds	49	2yge	80	3n0m
19	1qkt	50	2ygf	81	3n0s
20	1qy1	51	3a20	82	3nu9
21	1thy	52	3a6q	83	3oxc
22	1tpw	53	3a6r	84	3rdo
23	1tsv	54	3aj5	85	3ry2
24	1tsy	55	3am3	86	3s3v
25	1u5b	56	3am5	87	3ug2
26	1us0	57	3cyx	88	3uj9
27	1wli	58	3d1x	89	3ujc
28	1wlk	59	3d1y	90	3ujd
29	1x7w	60	3dgl	91	
30	1x7x	61	3dgo	92	
31	1x7y	62	3f8y	93	

Table A4.2. PDB codes used for "testing" conformational library selection

Entry	PDB code	Entry	PDB code	Entry	PDB code
1	1gz3	32	3dt7	63	3rhq
2	1gz4	33	3dtb	64	3rhr
3	1ykl	34	3fjz	65	3tkw
4	1ykp	35	3fk1	66	3um5
5	2aoc	36	3fra	67	3um6
6	2aod	37	3frb	68	3um8
7	2ca8	38	3fre		
8	2caq	39	3frf		
9	2f8f	40	3fy8		
10	2idw	41	3fy9		
11	2ien	42	3fyv		
12	2ieo	43	3fyw		
13	2ito	44	3kyg		
14	2ity	45	3lzs		
15	2jbz	46	3lzu		
16	2nmz	47	3m3c		
17	2o2q	48	3m3e		
18	2o2r	49	3m3o		
19	2q64	50	3moe		
20	2qd7	51	3mof		
21	2vfe	52	3moh		
22	2vfg	53	3nu3		
23	2vfh	54	3nu4		
24	2vfi	55	3nu5		
25	2wds	56	3nu6		
26	2wdy	57	3nuj		
27	3aj6	58	3nuo		
28	3cyw	59	3pca		
29	3d1z	60	3qgt		
30	3d20	61	3rhj		
31	3dt4	62	3rho		

Table A4.3. PDB codes used for docking studies (part 1)

Entry	PDB code	Entry	PDB code	Entry	PDB code	Entry	PDB code
1	1a4g	32	1f0s	63	1iel	95	1o5r
2	1a4q	33	1f4e	64	1iem	96	1o86
3	1a8i	34	1f4f	65	1igz	97	1okl
4	1agw	35	1f4g	66	1jao	98	1ony
5	1ah0	36	1fcx	67	1jaq	99	1ozl
6	1ah3	37	1fcy	68	1jj9	100	1p44
7	1b8n	38	1fcz	69	1jkx	101	1p4g
8	1b8o	39	1fd0	70	1ki2	102	1pax
9	1b8y	40	1fgi	71	1ki7	103	1ps3
10	1ba8	41	1fh7	72	1kim	104	1pxx
11	1biw	42	1fh8	73	1kv2	105	1q4g
12	1bju	43	1fh9	74	1l2i	106	1r34
13	1bn4	44	1fhd	75	1l2s	107	1rt1
14	1bnn	45	1fm9	76	1lhg	108	1rth
15	1c1b	46	1g4o	78	1llb	109	1s7y
16	1c2t	47	1gar	79	1m0n	110	1s9t
17	1c3e	48	1ggn	80	1m0o	111	1sd3
18	1cde	49	1gi6	81	1m0q	112	1sqa
19	1ciz	50	1gi8	82	1m17	113	1sr7
20	1cx2	51	1gj5	83	1mmb	114	1t9s
21	1d8m	52	1gj7	84	1mnc	115	1tbf
22	1db1	53	1gja	85	1mv9	116	1thl
23	1e2k	54	1gpn	86	1mvc	117	1tlp
24	1e2n	55	1gwq	87	1ndw	118	1tnk
25	1ecv	56	1gwr	88	1nhu	119	1tt1
26	1efy	57	1h1d	89	1nl9	120	1ttm
27	1ejn	58	1hw8	90	1nsd	121	1txi
28	1eko	59	1hwi	91	1nwl	122	1utp
29	1eqg	60	1hww	92	1o0m	123	1uy6
30	1eve	61	1hxk	93	1o0n	124	1uyd
31	1f0r	62	1ie8	94	1o0o	125	1uze

Table A4.4. PDB codes used for docking studies (part 2)

Entry	PDB code	Entry	PDB code	Entry	PDB code	Entry	PDB code
1	1uzf	32	2axa	63	2q61	95	3fuc
2	1v7a	33	2ayl	64	2q8s	96	3fug
3	1vid	34	2b35	65	2qe2	97	3g8n
4	1vrt	35	2bdj	66	2qe5	98	3pax
5	1vru	36	2bz5	67	2qn7	99	3pgh
6	1wbo	37	2ckm	68	2qn8	100	3std
7	1wbv	38	2cl5	69	2rg6	101	3tmn
8	1wxy	39	2fl8	70	2rgp	102	4cox
9	1xgj	40	2fgi	71	2src	103	4pax
10	1xkk	41	2fvc	72	2std	104	4std
11	1xoz	42	2gc8	73	2uwd	105	5std
12	1xp0	43	2gir	74	2uwl	106	5tmn
13	1xp1	44	2gs6	75	2w8y	107	8tln
14	1xpc	45	2h7l	76	2x23		
15	1y3u	46	2h7n	78	2xbw		
16	1y57	47	2hai	79	2xir		
17	1yae	48	2har	80	2z7g		
18	1yol	49	2has	81	2zff		
19	1yvf	50	2hwi	82	2zgb		
20	1ywn	51	2hwq	83	2zno		
21	1zgc	52	2iog	84	2zvj		
22	1zuc	53	2iok	85	3b68		
23	2a3i	54	2o5d	86	3bel		
24	2aa2	55	2oc2	87	3ccw		
25	2aa5	56	2oiq	88	3ccz		
26	2aa7	57	2ouz	89	3cdb		
27	2ack	58	2oye	90	3d90		
28	2ail	59	2p16	91	3dt3		
29	2ai2	60	2p1v	92	3ekr		
30	2ao6	61	2pax	93	3ert		
31	2ax9	62	2pnu	94	3fc6		

Rotamer libraries with 60° resolution are available online from Molecular Forecaster Inc.