On the Energy Minimization of Large Molecules

by

James A. Davey, B.Sc.

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Chemistry

Carleton University

Ottawa, Ontario, Canada

July 2011

Your file Votre référence
ISBN: 978-0-494-83173-1
Our file Notre référence
ISBN: 978-0-494-83173-1

ABSTRACT

Recent work has demonstrated that energy minimization, when applied to large molecular systems, exhibits initial condition sensitivity, i.e. small input perturbations gives large variations between geometry optimized structures. We hypothesize that there are several different causes for this behaviour: (1) energy gradient on input, (2) choice of minimization algorithm, and (3) tethered minimization protocols. In addition, (4) we hypothesize that these variations will alter the outcome of molecular docking simulations. As a proof of principle, by compressing bond lengths, we found that gradient is a contributing factor to spread in output. Hydrogen addition methodology, however, showed no affect. Minimizations done using second order algorithms (truncated Newton) were found to be more sensitive to input perturbation than first order algorithms (conjugate gradient). Tethering protocols were capable of completely eliminating output spread. Application of perturbed initial conditions to molecular docking showed novel effects leading to the new concept of receptor model bias.

ACKNOWLEDGEMENTS

Room 512 in the Steacie building is unlike any other on campus. It lacks a fire exit, its sink can't provide potable water and the majority of lighting fixtures don't work. More importantly, SC512 holds a collection of experience contained in theses written under the guidance of my supervisor Prof. Wright. The content of each thesis varies from early work involving the parameterization of balanced basis sets to recent work on molecular docking studies. And, for me, it's difficult to understand where my thesis and I fit into all that. I came out of my undergraduate degree unprepared for the study of theoretical chemistry. And despite that, Prof. Wright still managed to teach me the ropes, a testament to his skills as an educator and scientist. His lessons and patience are greatly appreciated and I hope that my hard work serves as a suitable expression of my gratitude. There are many other people who deserve acknowledgement for their encouragement and continued support of my studies in the sciences. First, my colleague and friend Matt Anderson for sharing his insight and knowledge. Second, to Dr. Hooman Shadnia for sharing his knowledge and friendship. To my friends Kurt DesRoches, André Vaillant, Steve Legault, Pascal Desjardins, Josselin Desjardins, Rémy Poulin and Matt Alteen for their support and their interest in my research. To Jim Logan for providing the four single core computers used to conduct this research. To my sister, Sarah Davey, my parents and grandparents for their continuing support and interest in my research. And, of course, to my loving girlfriend Julie Levionnois for understanding and sharing my passion for science, and for all the nights that I missed with her to complete this thesis.

TABLE OF CONTENTS

Title page i

Abstract ii

Acknowledgements iii

Table of contents iv

List of abbreviations vii

List of Tables ix

List of Figures x

Chapter 1 – Molecular Mechanics 1

1. 1. Introducing Molecular Mechanics 2

1. 2. The Intramolecular Terms 4

1. 3. The Intermolecular Terms 15

1. 4. Force Fields 22

1. 5. Introducing Energy Minimization 24

1. 6. The Steepest Descent Algorithm 28

1. 7. The Conjugate Gradient Algorithm 30

1. 8. The Truncated Newton Algorithm 31

1. 9. An Example Energy Minimization Problem 33

Chapter 2 – Characterizing Perturbed Energy Minimizations 36

2. 1. Introducing Input Perturbation Sensitivity 37

2. 2. Perturbation Size and Spread in Minimization 41

2. 3. The Effect of Gradient on Input and Input Perturbation Sensitivity 43

2. 4. The Addition of Hydrogens and Input Perturbation Sensitivity 46

2. 5. The Effect of Force Field and Input Perturbation Sensitivity 51

2. 6. The Minimization Algorithm and Input Perturbation Sensitivity 54

Chapter 3 – Constrained Energy Minimizations 58

3. 1. Previous Methods Involving Constrained Energy Minimization 59

3. 2. The Function of Tether Schemes to Reduce Spreads in Minimizations 63

Chapter 4 – Molecular Docking and Receptor Model Bias 72

4. 1. Molecular Simulations Involving Energy Minimization 73

4. 2. Refinement Procedures in Docking Simulation 75

4. 3. Introducing a New Approach to Docking Simulation 83

4. 4. Characterizing a Set of Receptor Conformations 85

4. 5. Docking a Set of Receptor Conformers 94

4. 6. Invalidating $RMSD_{Xtal}$ and E_{tot} as Qualifiers of a Valid Docking Receptor..... 103

Chapter 5 – Conclusions 105

Chapter 6 – Methods 110

6. 1. Hardware and Software Specifications 111

6. 2. An Example Energy Minimization Problem 112

6. 3. Perturbation of Molecular Coordinates 117

6. 4. The Preparation of Proteins for Experimentation 121

6. 5. Bond Compression Experiment 124

6. 6. Hydrogen Addition Experiment 125

6. 7. Tethered Energy Minimization Experiments 125

6. 8. Docking Simulation Procedures 129

References 138

LIST OF ABBREVIATIONS

1GWR - human estrogen receptor-alpha ligand binding domain dimer (protein)

1GWR-TM - truncated human estrogen receptor-alpha ligand binding domain (model)

1HDO - human biliverdin IX beta reductase (protein)

1KPI - mycolic acid cyclopropane synthase CmaA2 (protein)

1TNF - tumour necrosis factor-alpha

1UBQ - ubiquitin (protein)

2Y1R - MecA-ClpC molecular machine (protein)

3APU - alpha1-acid glycoprotein variant A (protein)

3M9H - prokaryotic ubiquitin like protein

AMBER-99 - assisted model building with energy refinement, version 99 (force field)

BCI - bond-charge increment (model)

BDE - bond dissociation energy

CEM - constrained energy minimization

CG - conjugate gradient (minimization algorithm)

CHARM-27 - chemistry at Harvard macromolecular mechanics, version 27 (force field)

DFT - density functional theory

E2 - 17β-estradiol

ERα - estrogen receptor-alpha

$\mathbf{E_{int}}$ - interaction energy between ligand and protein (descriptor)

$\mathbf{E_{tot}}$ - total potential energy (descriptor)

FF - force field

gtest - minimization termination gradient

LdG - london dG (descriptor)

MM - molecular mechanics

MMFF-94s - Merk mixed force field, version 94s (force field)

MOE - molecular operating environment (software)

PDB - protein data bank

PES - potential energy surface

QSAR - quantitative structure-activity relationship

RMS - root mean square

$RMSD_{pw}$ - pair-wise root mean square deviation

$RMSD_{Xtal}$ - root mean square deviation to crystal structure

RMSG - root mean square gradient

SD - steepest descent (minimization algorithm)

SF - scoring function

SVL - scientific vector language (programming language)

TEM - tethered energy minimization

TN - truncated Newton (minimization algorithm)

TNF-α - tumour necrosis factor-alpha

UEM - unconstrained energy minimization

UFF - universal force field (force field)

VSEPR - valence-shell electron-pair repulsion (theory)

ZPE - zero-point energy

LIST OF TABLES

Table 1.1 – Summary of parameters for various force fields 23

Table 1.2 – Energy minimizations of 1,2-dibromoethane 35

Table 2.1 – Unperturbed and perturbed energy minimizations of 1KPI 39

Table 2.2 – Various perturbation sizes for minimizations of 1UBQ 41

Table 2.3 – Bond compression and energy minimization with 1UBQ 44

Table 2.4 – Input gradient after default and protonate 3D hydrogen addition 48

Table 2.5 – Minimization of default and protonate 3D hydrogen addition structures 49

Table 2.6 – Input gradient for 1UBQ with various force fields 52

Table 2.7 – Minimizations of 1UBQ with various force fields 53

Table 2.8 – Energy minimizations with SD, CG and TN algorithms 56

Table 3.1 – Energy minimization with previous published tether protocols 62

Table 3.2 – Tethered energy minimization schemes 65

Table 3.3 – Tethered energy minimization of perturbed structures 66

Table 3.4 – The behaviour of TEM-4 with 1GWR-TM 70

Table 4.1 – Examples of molecular simulations not involving minimization 73

Table 4.2 – Estradiol and A-CD ligand series for docking studies 78

Table 4.3 – QSARs for three docking refinement methods 80

Table 4.4 – Docking refinement procedure and spread in descriptors 81

Table 4.5 – Notation used for prepared receptor models 86

Table 4.6 – Receptor properties before and after minimization 87

Table 4.7 – Property variations for the set of receptors 91

Table 4.8 – Single variable QSAR for docking studies in 25 receptors 95

Table 4.9 – Multiple variable QSAR for docking studies in 25 receptors 98

Table 4.10– Ligand orientation and H-bond strength for dockings in the receptor set. 100

Table 6.1 – Preparation of protein structures for experiments 123

LIST OF FIGURES

Figure 1.1 – Illustrating the five major force field terms..... 3

Figure 1.2 – Modelling the bond stretch term 6

Figure 1.3 – Newman projections for ethane conformations 9

Figure 1.4 – Modelling the bond torsion term 11

Figure 1.5 – Illustrating the cross potential terms 14

Figure 1.6 – Modelling the van der Waals potential term 21

Figure 1.7 – The potential energy surface of 1,2-dibromoethane 25

Figure 1.8 – Newman projections for minimization of 1,2-dibromoethane 33

Figure 2.1 – Perturbed and unperturbed energy minimizations of 1KPI 38

Figure 2.2 – Hydrogen addition using default and protonate 3D methods 47

Figure 3.1 – Schematic drawing of the TEM method 63

Figure 3.2 – TEM-4 results for 1GWR-TM..... 67

Figure 3.3 – UEM and TEM-4 results for three test proteins 69

Figure 4.1 – The relationship between minimization input and output structures 74

Figure 4.2 – An overview of molecular docking simulation 76

Figure 4.3 – Heavy atom overlay of the prepared seed and 24 perturbed receptors..... 87

Figure 4.4 – Hydrogen bonding network for the set of minimized receptors..... 89

Figure 4.5 – Examining the residues responsible for variations in interaction energy 93

Chapter 1

MOLECULAR MECHANICS

1. 1. Introducing Molecular Mechanics

The experiments presented in this thesis involve the energy minimization of large molecules in a molecular mechanics environment. As a result, it is important to introduce the methodology behind molecular mechanics along with its parameterization and the process of energy minimization. It is only with a clear understanding of the fundamental concepts presented in this first chapter that the problems presented by molecular modeling in later chapters can be successfully explored and understood.

There exists an arsenal of methodologies available to the computational chemist for the modeling of molecular structures. Quantum mechanical methods, such as Hartree-Fock,[1] allow for the computation of the electronic energy of a molecule using the full set of electrons while semi-empirical methods, such as Austin model 1,[2] approximate electronic energy as a function of the valence electrons. When modeling large molecules, such as proteins, molecular mechanics (MM) is often the chosen level of theory because the *ab initio* and semi-empirical methods are too time consuming to implement efficiently.[3a] MM models differ from the quantum mechanical methods because electrons are not explicitly included in the calculation. Instead, the MM approach represents atoms as point charges having mass and represents the bonds as springs having force constants. This approach requires the parameterization of force constants, partial charges, etc. to accurately model both the intramolecular (bonded) terms and the intermolecular (non-bonded) terms. As a result, MM attempts to implicitly account for the electronic affects that would otherwise be directly calculated

in semi-empirical or *ab initio* methods. The key potential terms that constitute a force field (FF) correspond to the molecular motions and interactions shown in Figure 1.1.

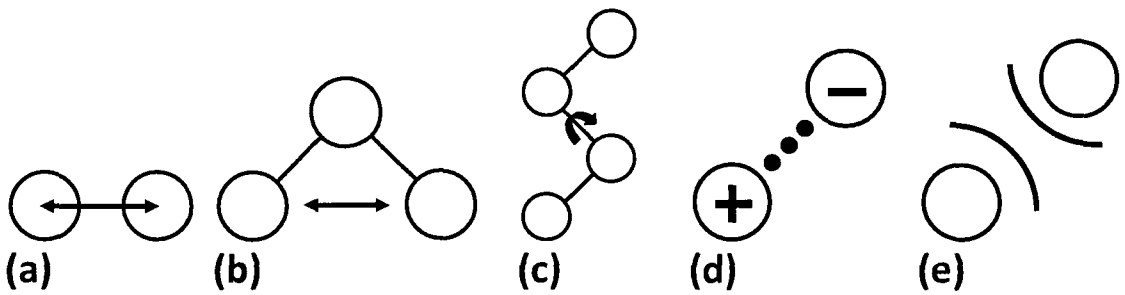


Figure 1.1 – Illustrating the five major force field terms
An illustration of the five key force field terms which include the (a) bond stretch, (b) bond bend, (c) bond torsion, (d) electrostatic, and (e) van der Waals terms.

The bond stretch term is involved in the computation of the potential energy penalty for deviation from ideal bond lengths for two atoms. The bond bend term evaluates the potential energy of the angle between three atoms bonded in series. The torsion potential term computes the potential energy for bond rotations. And the two intermolecular terms, the van der Waals and electrostatic potential terms, account for the non-bonded interactions between pairs of atoms in the system.

MM models involve calculation of the potential energy from the sum of the intramolecular and intermolecular terms whereby deviations from ideal bond lengths, angles, torsions, etc. result in energetic penalties to the potential energy of the system. The first chapter introduces the MM terms which comprise the potential energy surface (PES), the various sets of parameters used in MM referred to as force fields (FF), and finally the process of molecular energy minimization.

1. 2. The Intramolecular Potential Terms

The intramolecular (bonded) terms include three major terms to describe molecular motions for (i) the stretching of a bond between two atoms, (ii) the opening and closing of angles comprised of three atoms and two bonds, and (iii) the rotation about a bond.[3a] These three terms are referred to as the stretch, bend and torsion terms respectively.

1.2.1. The Stretch Potential Term

The stretch term can be modelled using a variety of functions, however the function most often used is based off of the Morse potential curve.[4] The Morse curve is an empirical formulation and is provided in equation 1.1, having the form:

$$V(r - r_e) = D_e \left[1 - e^{-\beta(r-r_e)}\right]^2 \tag{1.1}$$

where V is the potential energy, D_e is the depth of the potential energy well, r is the bond length, and r_e is the equilibrium bond length. β represents the relation, $\beta = \sqrt{k/2D_e}$, where k is the bond's force constant. The frequency of bond vibration, v, is related to the bond's force constant[4] given by $v = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}}$. This requires that the reduced mass between two atoms, μ, be calculated $\mu = \frac{m_i\, m_j}{m_i + m_j}$, having mass m_i and m_j for each atom respectively. The depth of the potential energy well, $D_e = BDE + \frac{1}{2}hv$ is calculated from the sum of the zero-point energy (ZPE), via the frequency of the vibration of the bond ($\frac{1}{2}hv$) for the molecule at ground state, and the bond dissociation

energy (BDE). While the Morse potential curve can adequately model the stretch and compression of a bond, the fact that the equation contains three parameters that need to be calculated (i.e. the depth of the potential well, and the bond's deviation from ideal length) before the potential energy can be evaluated makes the approach unsuitable for the efficient computation of stretch terms for large molecular structures.[3b] Instead, one of two approaches are taken to simplify the calculation. In simple MM models, Hooke's law is used to approximate the bond stretching term, as shown in equation 1.2 below.

$$V(r - r_e) = \frac{1}{2}k(r - r_e)^2 \tag{1.2}$$

This approach only requires the computation of a single term. A more rigorous approach involves fitting multiple polynomials with a Taylor series approximation,[5a, 5b] shown in equation 1.3, having the form:

$$V(r - r_e) = V(r) + V^{(1)}(r)(r - r_e) + \frac{V^{(2)}(r)}{2!}(r - r_e)^2 + \frac{V^{(3)}(r)}{3!}(r - r_e)^3 + \frac{V^{(4)}(r)}{4!}(r - r_e)^4 \tag{1.3}$$

which is typically formulated to the fourth order polynomial. Taylor series expansion to the fourth order polynomial for the Morse potential curve about $r = r_e$ gives the general form presented in equation 1.4.

$$V^{(0)}(r - r_e) = D_e\left[1 - e^{-\beta(r-r_e)}\right]^2 = 0$$

$$V^{(1)}(r - r_e) = 2D_e\beta e^{-\beta(r-r_e)} + 2D_e\beta e^{-2\beta(r-r_e)} = 0$$

$$V^{(2)}(r - r_e) = -2D_e\beta^2 e^{-\beta(r-r_e)} + 4D_e\beta^2 e^{-2\beta(r-r_e)} = 2D_e\beta^2$$

$$V^{(3)}(r - r_e) = 2D_e\beta^3 e^{-\beta(r-r_e)} - 8D_e\beta^3 e^{-2\beta(r-r_e)} = -6D_e\beta^3$$

$$V^{(4)}(r - r_e) = -2D_e\beta^4 e^{-\beta(r-r_e)} + 16D_e\beta^4 e^{-2\beta(r-r_e)} = 14D_e\beta^4$$

$$V(r - r_e) = \frac{k}{2}(r - r_e)^2 - \frac{k}{2}\sqrt{\frac{k}{2D_e}}(r - r_e)^3 + \frac{7k^2}{48D_e}(r - r_e)^4 \qquad (1.4)$$

Figure 1.2 shows the Morse potential curve for diatomic hydrogen fitted with the harmonic and Taylor series functions. Construction of the Morse potential curve for diatomic hydrogen required the experimental values of the bond force constant, k = 575 N/m,[6a] and solution of the reduced mass, μ = 8.37E-28 kg using equation 1.3, to calculate the frequency of vibration, v = 1.32E14 Hz, using equation 1.2. The depth of the potential energy was calculated from the experimental BDE = 435990 J/mol at 298 K,[6b] to be D_e = 435990 J/mol (or 104.204 kcal/mol), using equation 1.4. Finally, the experimental value for the ideal bond length of diatomic hydrogen was retrieved, r_e = 0.7414 Å (or 7.414E-11 m).[6c]

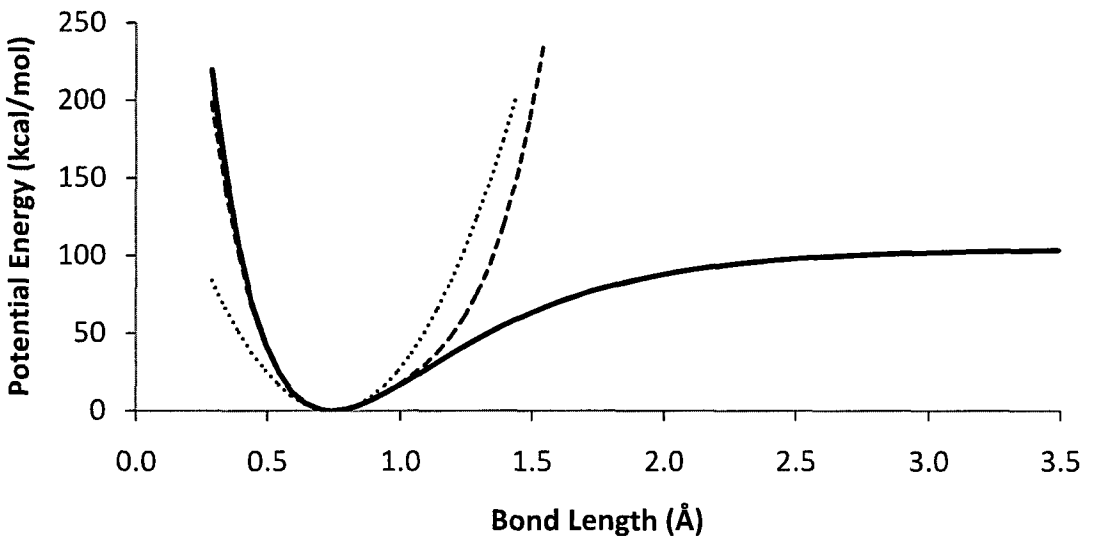


Figure 1.2 – Modelling the bond stretch term
Modelling the bond stretch term for diatomic hydrogen using the Morse potential function (solid line), the harmonic potential curve (dotted line), and Taylor series expansion to the fourth order polynomial (dashed line).

As shown in Figure 1.2, the Morse potential curve models the change in bond distance for the diatomic hydrogen molecule from repulsion ($r < r_e$), passing through the potential energy minimum involving maximum bonding orbital overlap ($r = r_e$), to less than ideal orbital overlap and attraction ($r > r_e$), and eventually bond dissociation with no orbital interaction or attraction ($r > 3.5$ Å).

The harmonic approximation of the bond stretch term can sufficiently approximate the Morse potential curve when close to the minimum for diatomic hydrogen. However, when bond lengths deviate far from the ideal value it can be seen that the harmonic approximation fails to model the Morse potential function.[3b] Use of the Taylor series approximation to the fourth order polynomial shows better fitting of the stretch term over greater ranges in bond length. The Taylor series approximation is particularly accurate at modelling the compression of the bond compared to the stretching of the bond. If better accuracy for the approximation of the bond stretch term is required, the Taylor series can be expanded to include polynomials of greater order.

1. 2. 2. The Bend Potential Term

The second major intramolecular component of MM, the bend term, models the energy penalty as applied to the angle between a set of three atoms bonded in series [A_1–A_2–A_3].[3c] Ideal bond angles in MM loosely conform to the empirical parameters established by valence-shell electron-pair repulsion (VSEPR) theory. VSEPR theory describes molecular geometry on the basis that electron pairs, whether involved in

bonding (shared) or not involved in bonding (lone), are arranged in the geometry that provides maximum distance between any two pairs.[7-8] Additional factors can cause deviation from the bond angles described by VSEPR theory. For example, two neighbouring and strongly electronegative atoms (i.e. fluorine) will likely have greater bond angles due to columbic repulsion. Additionally, atoms having a large atomic radius (i.e. iodine) will have greater bond angles due to steric clashes. The contribution of all these factors are accounted for by the ideal bond angle and the parameterization of the MM model. Generally, the MM bending term is modelled in a similar fashion to the stretch term by using either a harmonic potential curve,[3c] as shown in equation 1.5:

$$V(\theta - \theta_e) = \frac{1}{2}k(\theta - \theta_e)^2 \qquad (1.5)$$

or, in the case of highly strained or more complex potential curves, by using multiple higher-order polynomials.

1. 2. 3. The Torsion Potential Term

The third major intramolecular term is the MM torsion term. The torsion term models the potential energy penalty as a result of deviations from multiple ideal dihedral angles across a set of four atoms [A_1–A_2–A_3–A_4]. The torsion term is a critical component to MM because the majority of molecular conformations, especially in the case of proteins, are due largely in part to the rotations about chemical bonds.[3d] For example, the ethane molecule has three potential energy minima and maxima when rotating 360° about the CH_3–CH_3 bond. Figure 1.3 shows the Newman projections corresponding to the three minima and maxima of ethane.

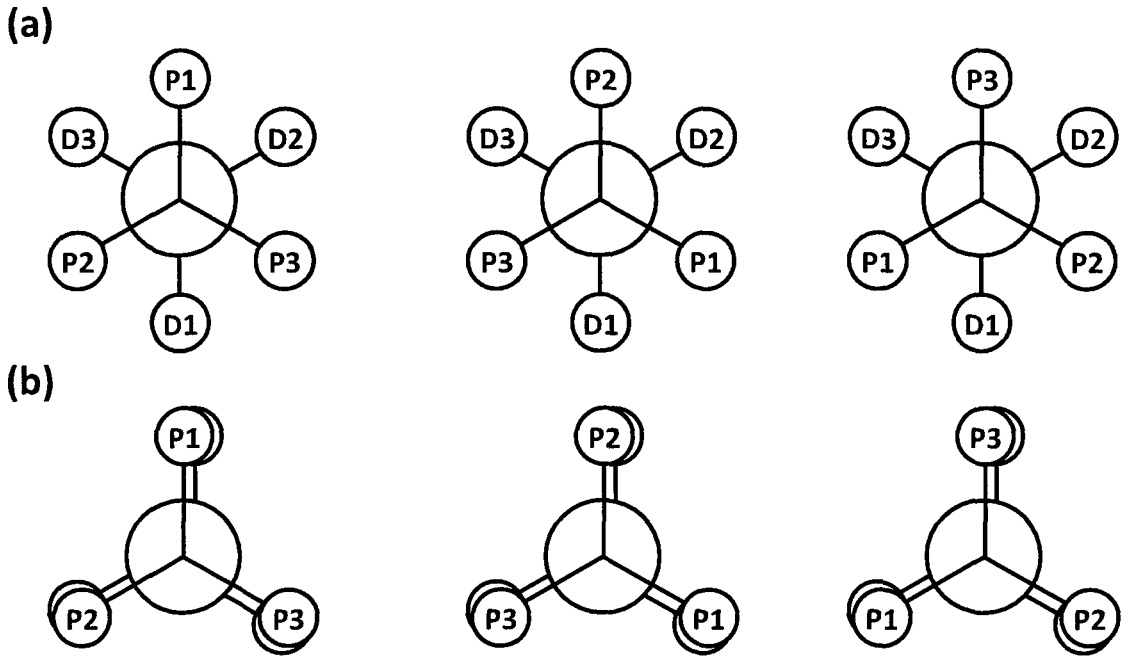


Figure 1.3 – Newman projections for ethane conformations
The six Newman projections involving 60° clockwise rotations of the proximal carbon for the ethane molecule which illustrate (a) the three potential energy minima, and (b) the three potential energy maximal. Hydrogens belonging to the proximal carbon are labelled as P# and hydrogens belonging to the distal carbon are labelled as D#.

As shown in Figure 1.3, the three potential energy minima correspond to the staggered arrangements of each methyl group while the three potential energy maxima correspond to the eclipsed arrangement of each methyl group. The barriers to rotation arise as a result of hyper-conjugation between the bonding orbitals and anti-bonding orbitals of the C–H chemical bonds of opposite carbons. The hyper-conjugation between bonding and anti-bonding orbitals has the greatest orbital overlap when ethane is in its staggered conformation and no overlap when ethane is in its eclipsed conformation.[9]

The modelling of bond rotation is made more complex due to the increased number of minima about the torsion potential term. A single ideal value for the

dihedral angle [A_1–A_2–A_3–A_4] about the bond [A_2–A_3], like those used previously for stretch and bend potential curves, is insufficient to model the potential curve for torsions. Instead, the energy profile for dihedral angle rotations is represented by a cosine function[3d] as shown in equation 1.6, having the form:

$$V(\omega) = \frac{k}{2}[1 + \cos(m\omega - \gamma)] \tag{1.6}$$

where, the energy penalty to all the dihedral angles (ω) about the rotatable bond sum to give the potential energy for the torsion term of that rotatable bond. In the case of ethane a total of nine such dihedral angles that must be evaluated (i.e. [$H_{1,1}$–C_1–C_2–$H_{2,1}$], [$H_{1,2}$–C_1–C_2–$H_{2,1}$], [$H_{1,3}$–C_1–C_2–$H_{2,1}$], [$H_{1,2}$–C_1–C_2–$H_{2,1}$], [$H_{1,2}$–C_1–C_2–$H_{2,2}$], [$H_{1,2}$–C_1–C_2–$H_{2,3}$], [$H_{1,3}$–C_1–C_2–$H_{2,1}$], [$H_{1,3}$–C_1–C_2–$H_{2,2}$], [$H_{1,3}$–C_1–C_2–$H_{2,3}$]). The number of minima across the torsion potential curve is set as an integer, m, and the particular dihedral angle's contribution to the energy penalty is given by k.[10] The parameter k specifies the potential energy barrier (saddle points) between two local minima for a given collection of four atoms in a torsion. It is important to note that intramolecular and intermolecular interactions can alter the height and location of the barrier. The phase for the cosine function, which determines the location of the minima and maxima, is adjusted using the parameter γ. Figure 1.4 shows the potential curves for rotations about the carbon-carbon bonds in ethane and ethene. Bond torsion parameters were retrieved from the universal force field (UFF).[11]

As shown in Figure 1.4, the bond torsion potential curves for both ethane and ethene are different in both height, phase, and periodicity. The height of the potential

barrier is 2 kcal/mol for ethane and 45 kcal/mol for ethene, as parameterized in UFF.[11] Calculation of the ethane torsion potential curve required the evaluation of nine dihedral terms while the ethene torsion potential curve required that only four dihedral terms be evaluated. This means that the force constant, k, for ethane and ethene dihedral terms were $^{2}/_{9}$ kcal/mol and $^{45}/_{4}$ kcal/mol respectively. The phase and number of minima for each potential curve is also different. For ethane, rotation about the single bond requires that the three potential energy minima (m = 3) exist in which hydrogens lie in the staggered conformation with dihedral angles at -60°, +60° and ±180°, as was shown in Figure 1.3, with a potential energy maxima passing through γ = 0°. For ethene, rotation about the double bond requires that two potential energy minima (m = 2) exist in which hydrogens lie in the plane of the molecule at 0° and ±180°. A potential energy maximum for ethene passes through γ = +90°.

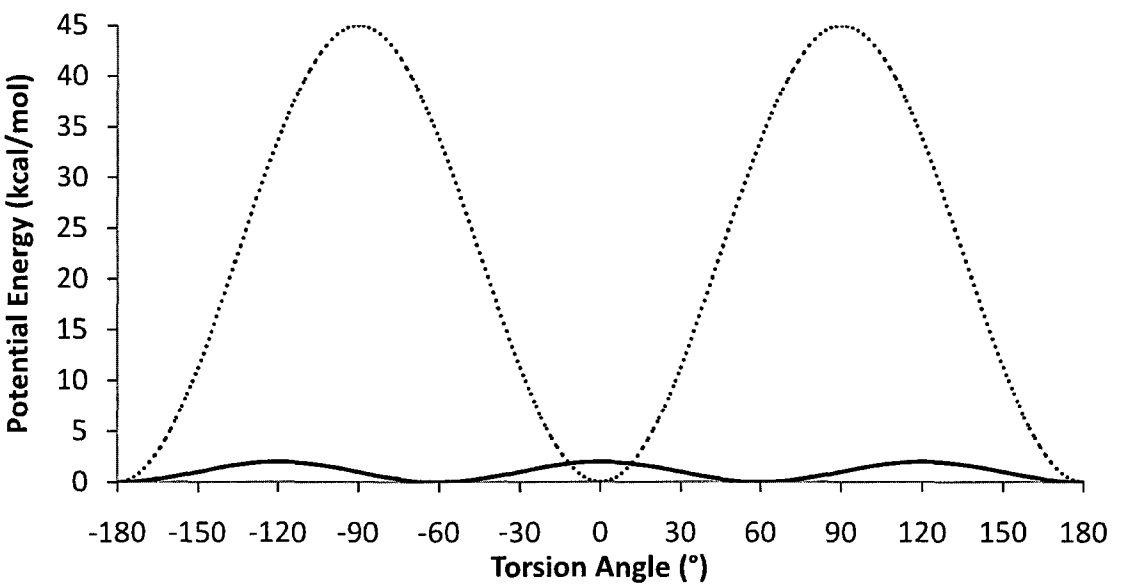


Figure 1.4 – Modelling the bond torsion term
Modelling the bond torsion term for rotations about the carbon-carbon single and double bonds of ethane (solid line) and ethene (dotted line) respectively.

As with both the stretch and bend potential terms, the cosine series can be expanded to improve the fit for the experimental torsion potential curve. The expansion of the cosine function is accomplished via Fourier series expansion[5c, 12] as shown in equation 1.7 with three terms:

$$V(\omega) = \sum_{n=1}^{N} k_n X_n$$

$$X_n = \frac{1}{2}[1 - \cos(n\omega)]$$

$$k_n = -\frac{4}{j}\sum_{i=1}^{j-1} \cos(n\omega_i) \cdot V(\omega_i) - \frac{2}{j}\cos(n\omega_n) \cdot V(\omega_n)$$

$$V(\omega) = \frac{k_1}{2}(1 + \cos\omega) + \frac{k_2}{2}(1 - \cos 2\omega) + \frac{k_3}{2}(1 + \cos 3\omega) \qquad (1.7)$$

where, X_n represents the function and k_n the coefficient evaluated for the n^{th} term. Series expansion requires that the potential energy $V(\omega_i)$ be evaluated for a set of data points of total size (j). The order of the term (n) determines the number of data points considered. Specifically, the evaluation of the first term ($n = 1$) requires two data points ($j = n + 1$). Thus, fitting the expanded series is referred to as a least squares fitting procedure.[15]

The fact that the torsion potential energy term is applied across sets of four atoms makes it a particularly costly MM term to evaluate. For example, a simple benzene ring involves 24 such dihedrals which all require computation. As a result, the various FF's available are highly variable depending on the requirement for accurate measurement of the various torsion terms. Some FF methods eliminate the torsion potential term entirely and account for torsion terms via the van der Waals term, while

other FF methods are parameterized to ignore the identity of the terminal atoms involved in the torsion.[3d]

A special case of the torsion potential term is referred to as the out-of-plane bending term. This term arises to account for the presence of π-bonds which have an optimum configuration when the torsion is in an coplanar arrangement. The term is often referred to as an improper torsion because its computation involves a set of four atoms that are not bonded in series, for example, the ketone belonging to cyclobutanone.[3e] Because π-bonds experience maximum bonding overlap in a coplanar configuration, the torsion potential term is simplified because there exists only one ideal value (0°) for configuration of the improper torsion. There are several ways in which the out-of-plane bonded term can be modeled. Equations 1.8, 1.9 and 1.10 provide the harmonic approximations for the improper torsion potential terms evaluated in one of three manners respectively.

$$V(\omega) = k[1 - \cos 2\omega] \tag{1.8}$$

$$V(\theta) = \frac{k}{2}\theta^2 \tag{1.9}$$

$$V(h) = \frac{k}{2}h^2 \tag{1.10}$$

In the case of equation 1.8, the improper torsion potential term is evaluated in the same fashion as the torsion term. Equations 1.9 and 1.10 take a different approach where the term is evaluated by either its angular deviation or its height deviation from the plane respectively.[3e]

1. 2. 4. The Cross Potential Terms

In addition to the three key intramolecular potential terms (i.e. stretch, bend and torsion), MM can include additional potential terms that involve the coordination of two or more of the basic potential terms. These potential terms which include multiple molecular motions are referred to as cross potential terms. While the incorporation of cross potential terms is not required to achieve accurate energy minimized molecular geometries, the use of cross potential terms does provide a higher degree of accuracy for modelling the potential surface. As a result, cross potential terms often provide increased accuracy to experimental values for properties that are sensitive to changes in the potential surface (i.e. molecular vibration).[3f] Figure 1.5 includes an illustration of various cross potential terms.

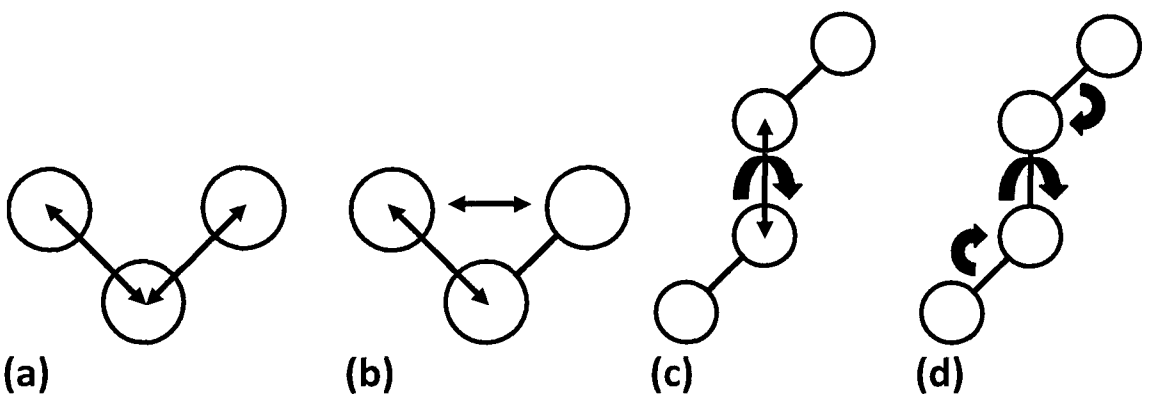


Figure 1.5 – Illustrating the cross potential terms
An illustration of cross potential terms which incorporate two or more molecular motions into a single term. Shown are the (a) stretch-stretch, (b) stretch-bend, (c) stretch-torsion, and (d) bend-torsion cross potential terms.

The cross potential terms, as shown in Figure 1.5, incorporate two or more molecular motions into one concerted potential term. For example, a stretch-bend

cross potential term can be derived for water. As the bond angle [H_1–O–H_2] is decreased from its equilibrium value, the [H_1–O] and [O–H_2] bond lengths increase to compensate for the repulsion between each hydrogen.[13] The stretch-bend potential term is provided in equation 1.11:

$$V(r_{\mathrm{H1}}, r_{\mathrm{H2}}, \theta) = \frac{k_{r_{\mathrm{H1}}, r_{\mathrm{H2}}, \theta}}{2}\left[\left(r_{\mathrm{H1}} - r_{e_{\mathrm{H1}}}\right) + \left(r_{\mathrm{H2}} - r_{e_{\mathrm{H2}}}\right)\right](\theta - \theta_e) \qquad (1.11)$$

which has the form like a harmonic potential, having only a single force constant ($k_{r_{\mathrm{H1}}, r_{\mathrm{H2}}, \theta}$) to describe the cross potential term.[3f] The non-bonded potential terms (i.e. van der Waals and electrostatic terms) can be used in place of cross potential terms to account for coupled molecular motions.[14]

1. 3. The Intermolecular Potential Terms

The intermolecular potential terms represent the non-bonded interactions present in the system. In molecular mechanics, the intermolecular potential terms are separated into two categories, (i) the electrostatic interactions and (ii) the van der Waals interactions.[15] MM methods evaluate these interactions between pairs of atoms which means that for a non-bonded system of n atoms there are a total of $n(n-1)/2$ number of interactions that must be evaluated for both of intermolecular potential terms. Some FFs use the non-bonded terms to account for the cross potential terms. In the case where of the stretch-stretch or stretch-bend cross potential terms are modelled using the intermolecular potential terms, the interaction is referred to as a

1,3-interaction because the term is computed for the first and third atoms from the set of atoms that comprise the angle [A_1–A_2–A_3].[14] In the case where the stretch-torsion or bend-torsion cross potential terms are modelled using the intermolecular potential terms, the interaction is referred to as a 1,4-interaction because the term is computed for the first and fourth atoms from the set of atoms that comprise the torsion [A_1–A_2–A_3–A_4].[14] Thus, the number of intermolecular potential terms evaluated for a given system is dependent on the FF methodology employed.

1. 3. 1. The Electrostatic Potential Term

The more electronegative an atom is, the more it attracts electrons. Thus, the distribution of charge across a given molecule is dependent on its molecular composition and geometry. A method known as central multipole expansion which uses the electric moments, or multipoles, can be used to describe the distribution of charge across a molecule by representing the distribution as an infinite series of point charges.[3g] However, in many cases, it is sufficient to represent the distribution of charge across a molecule via the lowest order non-zero moment. For the case of ions and small singly charged molecules, which bear a formal charge (i.e. Na^+, I^-, CO_3^-), the charge (q) is sufficient to describe distribution of charge. This is because the distribution of charge about a formally charged center typically adheres to a spherical geometry and is the only charge, having the greatest magnitude, in the molecule.

The majority of non-charged molecules (i.e. H_2O, CH_3OH) have charge distribution that can be sufficiently modelled using the dipole (μ). A dipole represents

the distribution of charge as a vector.[3g] The formulation of the dipole in Cartesian coordinates, $\vec{\mu} = \sum_{n=i}^{N} q_i r_i$, is a vector (rank order 1 tensor) where, each point charge (i) having magnitude (q) and location ($r = [x - x_0, y - y_0, z - z_0]$) from a defined point of reference (i.e. center of mass). Solution of the moments along the x, y, and z axes would yield the components of the dipole moment $\mu_x = \sum q_i x_i$, $\mu_y = \sum q_i y_i$, and $\mu_z = \sum q_i z_i$ respectively.

In the case of symmetric linear non-charged molecules or a planar and symmetric non-charged molecule (i.e. H_2, N_2, CO_2, C_6H_6) the dipole moment is equal to zero. Thus, the next electric moment, the quadrupole (Θ), is required to describe the charge distribution about the molecule.[3g] The quadrupole is solved as a rank order 2 tensor as shown in equation 1.12:

$$\Theta = \begin{bmatrix} \sum q_i (y_i + z_i)^2 & -\sum q_i x_i y_i & -\sum q_i x_i z_i \\ -\sum q_i y_i x_i & \sum q_i (x_i + z_i)^2 & -\sum q_i y_i z_i \\ -\sum q_i z_i x_i & -\sum q_i z_i y_i & \sum q_i (x_i + y_i)^2 \end{bmatrix} \tag{1.12}$$

once again, the coordinates are all defined with respect to a point of reference like the center of mass. The diagonal terms in the tensor are referred to as the principal components of the quadrupole while the off-diagonal terms in the tensor are referred to as the products of the quadrupole. Diagonalization[5d] of the quadrupole tensor yields the three Eigenvalues Θ_x, Θ_y, and Θ_z which correspond to the principal axes of the quadrupole.[5e] In this fashion, the tensor describes the charge distribution across the molecule as a three dimensional spherical shape. If all three principal axes of the quadrupole are equal the quadrupole field has a shape of a perfect sphere. Varying the

principal axes of the quadrupole alters the shape of the sphere in all three axes about the point of reference.

The charge distribution of a molecule can be evaluated to an infinite number of multipoles. However, in many cases it is sufficient to evaluate the first non-zero multipole.[16] Calculation of an electrostatic interaction between two point charges q_i and q_j is computed using Coulomb's law shown in equation 1.13:

$$V(q_i, q_j) = \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{1.13}$$

where, r_{ij} is the distance between the two point charges and ε_0 is the permittivity of space in a vacuum.[6e] The interaction of charge distributions can be expanded as an infinite series with charge-charge ($V \propto 1/r$), charge-dipole ($V \propto 1/r^2$), dipole-dipole ($V \propto 1/r^3$), dipole-quadrupole ($V \propto 1/r^4$), etc. interaction terms.[16]

In MM, the charge distribution across a molecule is represented by a series of point charges which reside at or near the center of each atom in the system.[15] This approach, referred to as the partial atomic charge model, is used in an attempt to recreate the charge distribution about molecules because electrons are not explicitly modelled in MM. Each atom in a system has a parameterized partial charge assigned based on its identity and bonded neighbours. Parameterization of partial atomic charges can be accomplished through a variety of methods, all of which are parameterized to fit charge distributions determined by either higher level *ab initio* computation or from experiment. These methods generally follow one of four

approaches: (i) an approach published by Cox and Williams whereby a set of partial atomic charges are fitted using the least-squares procedure,[17] (ii) a procedure employed by Chirlian and Francl which uses Lagrange multipliers in place of least-squares to the fit a sphere of point charges about an atom,[18] (iii) the model proposed by Gasteiger and Marsili which relates partial atomic charge to the formal charge of an atom attenuated by the electronegativity of neighbouring bonded atoms,[19] and (iv) the bond-charge increment (BCI)[20] method which acts in a similar manner to the Gasteiger and Marsili model but also accounts for some non-bonded charge interactions and aromatic ring attenuation of the partial charges.

The electrostatic potential term is then evaluated between sets of partial atomic charges using Coulomb's law. There exist additional modified electrostatic potential curves which attempt to implicitly account for the effects of interactions between multiple point charges in a system. A modified version of Coulomb's law, the distance dependent dielectric model, scales the electrostatic potential term by the square of the distance, as shown having the general form in equation 1.14.

$$V(q_i, q_j) = \frac{q_i q_j}{4\pi\varepsilon_0 {r_{ij}}^2} \tag{1.14}$$

The distance dependent dielectric model dampens the permittivity of space in the system in an attempt to implicitly account for solvent interactions.[3h] While the distance dependent dielectric relation in equation 1.15 has no physical basis for the calculation of potential energy (in fact, equation 1.15 would return the force of the electrostatic term), careful parameterization of the term using a sigmoidal distance

dependency has been demonstrated to effectively model electrostatic interactions for biological systems.[21]

1. 3. 2. The van der Waals Potential Term

The van der Walls potential term arises from the attractive and repulsion interactions between two neutral atoms. The attractive interaction is due to dispersive forces between induced dipoles of the two interacting atoms. The repulsion interaction is due to exchange forces which arise from the unfavourable interaction of electrons in close proximity.[31] Modelling the van der Waals potential term is accomplished using a Lennard-Jones potential curve, shown in equation 1.15, having the general form:

$$V(r_{ij}) = ke_{ij}\left[\left(\frac{\sigma}{r_{ij}}\right)^{n} - \left(\frac{\sigma}{r_{ij}}\right)^{m}\right] \tag{1.15}$$

where, e_{ij} is the depth of the potential energy well, r_{ij} is the distance between the two atoms, and σ is the distance on the repulsion portion of the function where the potential energy is equal to 0. The Lennard-Jones potential curve is a composite of two functions where the repulsion portion of the curve is represented by $\left(\frac{\sigma}{r_{ij}}\right)^{n}$ and the attractive portion of the curve is represented by $-\left(\frac{\sigma}{r_{ij}}\right)^{m}$.[31] Typically, the powers n and m have values of 12-10, 12-9 or 12-6 respectively. The constant, k, is given by the relation $k = \frac{n}{n-m}\left(\frac{n}{m}\right)^{m/(n-m)}$.[31] In MM, these values are parameterized and assigned based on the two atoms involved in the interactions. Figure 1.6 shows the individual attractive and repulsion functions which form the 12-6 Lennard Jones potential curve

for two interacting argon atoms. Construction of the curve required experimental data for the depth of the potential energy well, e_{ij} = 0.233977 kcal/mol, and the location inter-atomic distance where the potential energy is 0, σ = 3.419 Å.[22]

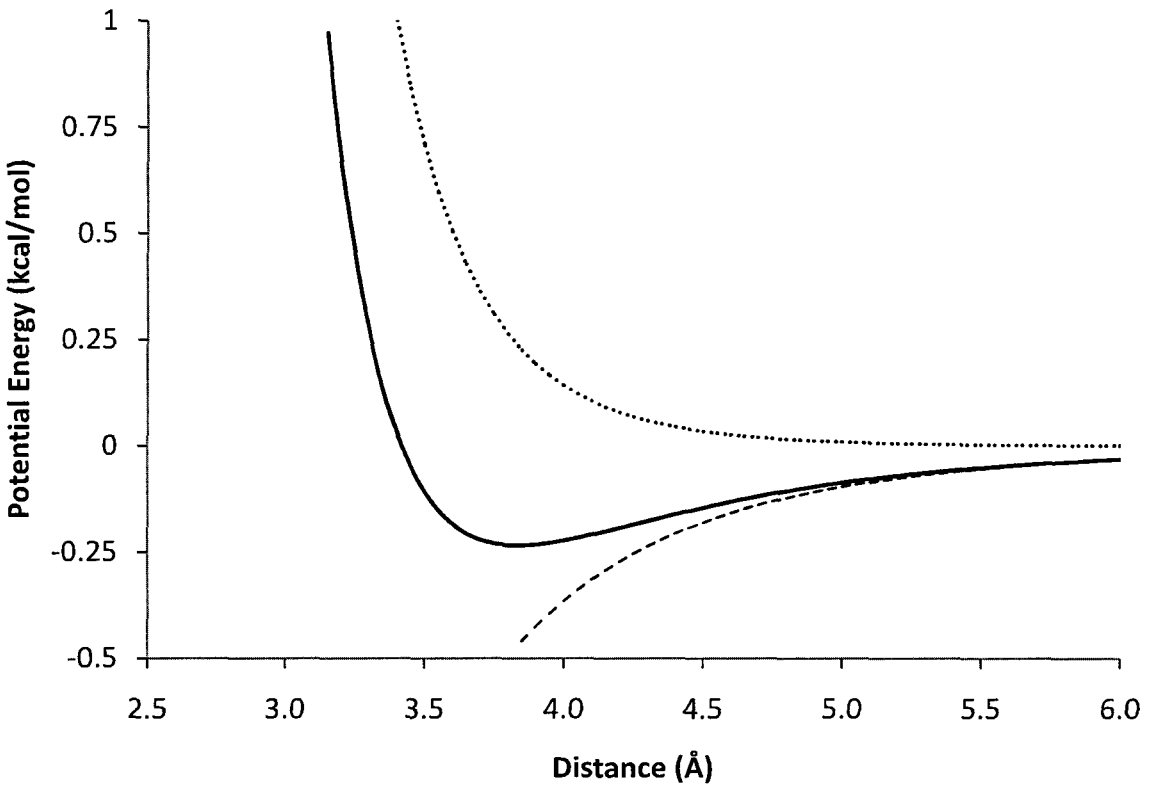


Figure 1.6 – Modeling the van der Waals potential term
The van der Waals potential energy curve for two interacting atoms in the gas phase modelled using a Lennard-Jones 12-6 curve (solid line) with attractive (dashed line) and repulsion (dotted line) portions of the curve.

As shown in Figure 1.6, the 12-6 Lennard-Jones curve models the van der Waals interaction between two atoms using a composite of two functions. Selection of repulsion and attraction parameters that are larger (i.e. n > 12 and m > 6) will make each portion of the curve steeper.

1. 4. Force Fields

A collection of parameters which define the intermolecular and intramolecular terms in a MM model is called a force field (FF). There exist a variety of FFs parameterized for various types of molecules. Table 1.1 summarizes the implementation and parameterization of the intramolecular and intermolecular potential terms for three typically used FFs: MMFF-94s, AMBER-99 and CHARMM-27.

The Merk mixed forced field, version 94s (MMFF94s) has been parameterized for small organic molecules and use in ligand-receptor docking studies.[23-24] The assisted modeling building with energy refinement, version 99 (AMBER-99) FF, belongs to a family of FF created by Kollman which has been parameterized for the modelling of proteins and nucleic acids.[25-26] The FF, chemistry at Harvard macromolecular mechanics, versions 22 and 27 (CHARMM-22 & CHARMM-27) are another set of FFs which have been parameterized for proteins and nucleic acids respectively.[27-28] In addition to having different parameters, each FF may also have an entirely different method of calculating particular terms or may include or omit additional terms depending on its purpose. The AMBER-99 and CHARMM-27 FFs are referred to as class I force fields because all potential terms are not expanded beyond their basic first term nor do either of the FFs include explicit cross potential terms.[29] The MMFF-94s FF is referred to as a class II force field because its potential terms have been expanded to higher order polynomials and because evaluation of some of the cross potential terms are explicitly included.

Table 1.1 – Summary of parameters for various force fields

Potential Terms	MMFF-94s	AMBER-99	CHARMM-27
Parameterization	• for gas phase small organic molecules • parameterized via computational data	• for proteins and nucleic acid • parameterized via computational data	• for nucleic acid • parameterized via computational data
Stretch	• 4th order Taylor polynomial	• 2nd order Taylor polynomial	• 2nd order Taylor polynomial
Bend	• 3rd order polynomial	• 2nd order polynomial	• 2nd order polynomial
Torsion	• Fourier expansion of cosine series to 3rd term • improper torsions are a separate term	• Fourier expansion of cosine series to 1st term • improper torsions are included	• Fourier expansion of cosine series to 1st term • improper torsions are included
Cross Terms	• stretch-bend term explicitly included • 1,4 interactions via electrostatic and van der Waals terms	• no cross terms are explicitly modelled • 1,4 interactions via electrostatic and van der Waals terms	• no cross terms are explicitly modelled • 1,4 interactions via electrostatic and van der Waals terms
Electrostatic	• partial charges from bond-charge increment model	• partial charges by regression from peptide models	• partial charges by regression from water models
van der Waals	• 14-7 Lennard-Jones potential curve	• 12-6 Lennard-Jones potential curve • polar hydrogens do not have radii	• 12-6 Lennard-Jones potential curve

Parameters for MMFF-94s, AMBER-99 and CHARM-27 are published and available from references 23, 25, and 27

Each FF has been parameterized for a separate purpose and each FF contains considerably different parameters. In addition to the differences in the organization between class I and class II FFs, there is variation in the method of parameterization of the intermolecular potential terms. The MMFF-94s FF evaluates the Lennard-Jones

potential term as a 14-7 potential curve[30] while the AMBER-99 and CHARMM-27 FFs each use the standard 12-6 Lennard-Jones potential curve. As well, the parameterization of partial charges is considerably different. The MMFF-94s FF derives partial charges from the bond-charge increment (BCI) model. The AMBER-99 FF uses a two step regression method, referred to as a restrained electrostatic potential fitting method, where charges are assigned on shells around the atom corresponding to their van der Waals radii.[31] As well, the AMBER-99 FF treats polar hydrogens as though they have no van der Waals radius. The partial charges for CHARMM-27 were fitted to *ab initio* calculations of nucleic acid interactions with explicit water. Each of the three force fields share the similar application of intermolecular potential terms where only 1,4 and non-bonded interactions are evaluated while 1,2 or 1,3 interactions are not evaluated.

The behaviour and computation results will vary for a modelled system depending on the force field used due to the various differences in parameterization and implementation of each force field.

1. 5. Introducing Energy Minimization

The potential energy minimum is important because it forms the basis for the thermodynamics of the ground state.[32] This means that molecular modelling in computational chemistry inevitably involves an energy minimization procedure to find the potential energy minimum. Energy minimization, also referred to as geometry optimization, involves the solution of a potential energy minimum from an input

geometry. Examples of common molecular modeling practices that involve energy minimization include (i) homology modeling,[33] (ii) molecular docking studies,[34] and (iii) conformational analysis.[35] It is important to consider that energy minimization algorithms can only return a potential energy minimum within the vicinity of the input geometry.[36a] Figure 1.7 illustrates how, for a given function such as a potential energy surface (PES), only the closest potential energy minimum from input is found and that it may not be lowest energy minimum on the PES.

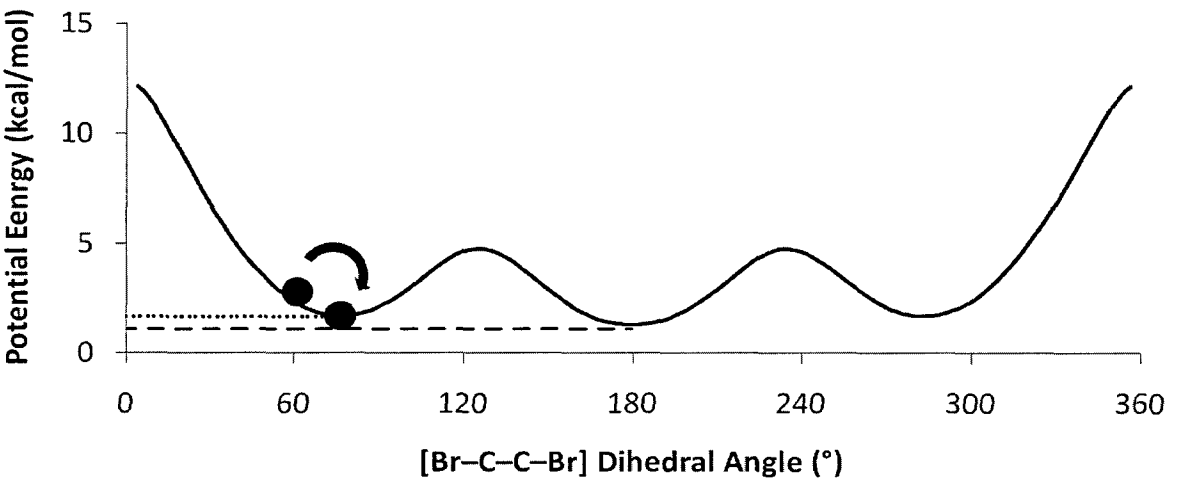


Figure 1.7 – The potential energy surface of 1,2-dibromoethane
The potential energy surface of 1,2-dibromoethane for clockwise rotations about the Br–C–C–Br dihedral angle having eclipsed conformations at 0°, 120° and 240° rotations and staggered conformations at 75°, 180° and 285° rotations.

As shown in Figure 1.7, the PES of 1,2-dibromoethane has three potential energy minima as modeled using MMFF94s FF with the distance dielectric solvent model implemented in the molecular operating environment (MOE).[37] Two of these minima are found at 75° and 285° having a potential energy of 1.68 kcal/mol while the third minima resides at 180° and has a potential energy of 1.27 kcal/mol. A distinction is made between the two sets of minima whereby the lowest potential energy minimum is

referred to as the global energy minimum while all other local minima are referred to as local energy minima. The difference in potential energy is due to the 1,4 van der Waals and electrostatic interactions between the large bromine atoms. This PES differs from the bond torsion potential term example presented in Figure 1.5 as the potential energy wells of ethane are all degenerate since all terminal atoms are all the same.

If an input geometry for 1,2-dibromoethane having a dihedral angle of 60° between the [Br–C–C–Br] dihedral angle is subject to energy minimization, the algorithm will only find the local minimum at 75°.[36a] This is due to the function of minimization algorithms. Minimization algorithms use first derivative, and sometimes second derivative, information to navigate the PES. As a result, minimization algorithms can find the closest potential energy minimum to the input configuration of molecule. To solve the location of the global energy minimum, multiple input configurations must be subjected to energy minimization. This process is referred to as a conformational search and can be accomplished in a systematic fashion whereby bond torsions are rotated to produce input structures. Additional methods, such as Monte Carlo or molecular dynamics simulation, can also be used to search the PES.[36b] To solve the location of the global energy minimum for the example using 1,2-bromoethane, an input geometry having a [Br–C–C–Br] dihedral angle greater than 120° but less than 240° would have to be provided to the minimization algorithm.

The potential energy of a molecular system is computed as a function of its coordinates by summing all the intramolecular and intermolecular potential energy terms in the system as shown in equation 1.16:

$$V(x,y,z) = \sum E_{str} + \sum E_{bend} + \sum E_{tor} + \sum E_{cross} + \sum E_{vdW} + \sum E_{ele} \qquad (1.16)$$

where, the sum of all the stretch (str), bend (bend), torsion (tor), cross (cross), van der Waals (vdW) and electrostatic (ele) potential terms gives the potential energy of the system. The first derivate of the PES, referred to as the gradient, provides the direction of the minimum from the input location as well as the slope on the surface. The goal of any minimization algorithm is to find a point on the potential surface that satisfies $\nabla = 0$ and $\nabla^2 \geq 0$ where, the gradient (∇) of the function is equal to zero and the second derivative (∇^2) of the function is greater than or equal to zero.[3j] In the Cartesian coordinate system, the gradient for a molecule is computed for each atom (i) along all three axes [x, y, z] by introducing displacements [$\pm x, \pm y, \pm z$] and computing the average change in the potential energy as demonstrated in equation 1.17:

$$\nabla = \begin{bmatrix} \frac{V_{+x} - V_{-x}}{2dx_i} & \frac{V_{+y} - V_{-y}}{2dy_i} & \frac{V_{+z} - V_{-z}}{2dz_i} \\ \vdots & \vdots & \vdots \\ \frac{V_{+x} - V_{-x}}{2dx_n} & \frac{V_{+y} - V_{-y}}{2dy_n} & \frac{V_{+z} - V_{-z}}{2dz_n} \end{bmatrix} \qquad (1.17)$$

which, gives rise to a Jacobian matrix, identified as ∇, of width 3 for the partial derivatives about the three axes and of length n, the total number of atoms in the system.[5f] Thus, the magnitude of the gradient for an atom on the potential surface is computed by the relation shown in equation 1.18:

$$|\nabla_i| = (\nabla_i \cdot \nabla_i)^{1/2} \tag{1.18}$$

given by the dot product of the gradient for an atom with itself under the square root. Because the potential surface is made up of n atoms, it is more informative to compute the root mean square gradient (RMSG). The RMSG is calculated for a system as shown in equation 1.19.

$$\mathrm{RMSG} = \sqrt{\frac{\sum_{k=1}^{n} |\nabla_k|^2}{n}} \tag{1.19}$$

Because computational methods are involved which have accuracy limited to the precision of the floating point, the exact location of the minimum cannot be computed. Instead, the minimization algorithm must find a configuration for the molecule whose RMSG falls below a threshold referred to as the convergence criteria for the algorithm (gtest). A minimization algorithm finds the location of the minimum by repeating two tasks, (i) determining the direction of the step, and (ii) determining the size of the step. Each step taken is referred to as an iteration and the algorithm proceeds until the convergence criteria is satisfied. The choice of minimization algorithm determines how step direction and size are computed.

1. 6. The Steepest Descent Algorithm

The steepest descent (SD) minimization algorithm is classified as a first derivative minimization algorithm because it only uses first derivative information to locate the position of minima on the PES. The direction of step at each iteration (k) is calculated as

a normalized vector in the direction of the gradient. Equation 1.20 provides the SD solution of direction step (s):

$$s_k = -\frac{\nabla_k}{\|\nabla_k\|} \tag{1.20}$$

where, the array of gradient vectors is divided by the norm of the vector to return the 3N dimensional unit vector corresponding to direction of step. In this manner, the SD algorithm takes the step in the steepest direction at iteration k towards the minimum.[3k] It is important to note that the sign of s_k is negative, if the direction of step was assigned a positive value then the method would solve for the nearest local maximum.[39] Next, the SD minimization algorithm must determine the size of step to take on the PES. Step size is typically determined by one of two methods. The first, the line search method, involves the iterative solution of a minimum along the path specified by s_k. The second method, the arbitrary step method, involves a similar search along s_k involving the scaling of arbitrary step sizes.

The line search method plots solves points along s_k where the first point corresponds to the potential energy at the current position (k). The line search method attempts to find a second and third point whereby the potential energy of the second point is less than both of the first and third points. In this manner, the location of the minimum along s_k is bracketed somewhere between the first and third points. The distance between the first and third points is then gradually reduced until the second point corresponds to the minimum value along s_k.[3k] Because this approach can be computationally expensive, it is common to fit polynomials to the set of points along s_k.

A quadratic function often makes for a more effective fit of s_k as the RMSG of the system approaches the minimum.[3l]

The second method, the arbitrary step approach, a default step size is applied across the path s_k. If the potential energy is reduced, then the step size is increased by a default factor, and if the potential energy is increased, then the step size is decreased by a default factor. The process is repeated until a point found along s_k no longer reduces the potential energy of the system.[3l] The process is continued until the convergence criteria for RMSG of the system is met. While the arbitrary step approach requires fewer evaluations of points along the path s_k, the method can exhibit poor convergence when close to the minimum because multiplication by gross arbitrary factors can lead to continuous over stepping of the potential energy minimum.[36c]

1. 7. The Conjugate Gradient Algorithm

The conjugate gradient (CG) algorithm is also a first derivative minimization algorithm. CG is different from SD in that it takes orthogonal steps from the previous iterations gradient. Because first step performed with CG has no previously evaluated gradient, the first direction of search is typically solved using the SD minimization algorithm. Orthogonality of step s_k is enforced using a scalar constant which is applied to the previous iteration's gradient as shown in equation 1.21:

$$s_k = -\nabla_k + \gamma_k \nabla_{k-1} \tag{1.21}$$

where, g_k is the current iteration's gradient, g_{k-1} is the previous iteration's gradient and $\gamma_k$is the scalar constant. The calculation of the scalar constant, $\gamma_k = \frac{\nabla_k \cdot \nabla_k}{\nabla_{k-1} \cdot \nabla_{k-1}}$, ensures that each successive step is orthogonal from the last positions gradient.[3m]

Because the CG minimization algorithm formulates step directions which are orthogonal to both the previous iteration's gradient and step direction, the CG method can converge in fewer iterations compared to its SD counterpart. As a result, the CG minimization algorithm is particularly efficient at converging over parabolic surfaces.[3m] Step size is then solved using either the line search or arbitrary step approach as described for the SD algorithm.

1. 8. The Truncated Newton Algorithm

The truncated Newton (TN) family of energy minimization algorithms differ from the previously discussed SD and CG algorithms because TN is a second order energy minimization algorithm. The approach to computing the step direction and size (s_k) is accomplished by solving Newton's equation in one concerted solution:

$$\nabla^2{}_k s_k = -\nabla_k \qquad (1.22)$$

where, s_k is a vector that transforms the second derivative matrix (Hessian) into the gradient (Jacobian) for the system.[39] In this manner, the solution of s_k is calculated under the assumption that the PES between its current configuration (V_k) and the configuration at the next step (V_{k+1}) is a quadratic function. The exact solution of s_k

does not need to be determined, in fact, the majority of TN minimization algorithms employ a CG method to solve equation 1.22.

Numerical formulation of the Hessian matrix would follow the same method as that for the Jacobian matrix (shown in equation 1.17). The difference being that the matrix evaluates differences in the gradient about the current configuration of the system to give the Hessian matrix ($\nabla^2{}_k$) with a size of $3n \times 3n$. The TN minimization algorithm is so named a truncated algorithm because the full Hessian matrix is not computed. Instead, a sparse matrix referred to as a band matrix which has pseudo tri-diagonal form, as shown in equation 1.23.

$$\nabla^2{}_k = \begin{bmatrix} \frac{\partial^2 V}{\partial x_1^2} & \frac{\partial^2 V}{\partial x_1 \partial y_1} & \frac{\partial^2 V}{\partial x_1 \partial z_1} & & & & \\ \frac{\partial^2 V}{\partial y_1 \partial x_1} & \frac{\partial^2 V}{\partial y_1^2} & \frac{\partial^2 V}{\partial y_1 \partial z_1} & \cdots & & 0 & \\ \frac{\partial^2 V}{\partial z_1 \partial x_1} & \frac{\partial^2 V}{\partial z_1 \partial y_1} & \frac{\partial^2 V}{\partial z_1^2} & & & & \\ & \vdots & & \ddots & & \vdots & \\ & & & & \frac{\partial^2 V}{\partial x_n^2} & \frac{\partial^2 V}{\partial x_n \partial y_n} & \frac{\partial^2 V}{\partial x_n \partial z_n} \\ & 0 & & \cdots & \frac{\partial^2 V}{\partial y_n \partial x_n} & \frac{\partial^2 V}{\partial y_n^2} & \frac{\partial^2 V}{\partial y_n \partial z_n} \\ & & & & \frac{\partial^2 V}{\partial z_n \partial x_n} & \frac{\partial^2 V}{\partial z_n \partial y_n} & \frac{\partial^2 V}{\partial z_n^2} \end{bmatrix} \quad (1.23)$$

The solution of s_k is dependent on the nature of the Hessian matrix ($\nabla^2{}_k$). To ensure convergence of the TN algorithm, the line search is typically implemented along s_k. This double checking procedure is referred to as a globalization strategy.[39]

1. 9. An Example Energy Minimization Problem

We will consider an example energy minimization problem involving the 1,2-dibromoethane molecule presented previously in Figure 1.7. This problem was solved from an input geometry where the [Br–C–C–Br] dihedral angle set at 60°. The energy minimization will be conducted using the Molecular Operating Environment (MOE)[37] program, an MMFF94s[25-26] FF with distance dependant dielectric solvation model, and MOE's implementation of the SD, CG and TN algorithms. To reproduce this experiment, explicit instructions are available in the methods chapter (6. 1. Hardware and Software Parameters & 6. 2. An Example Energy Minimization Problem). Figure 1.8 shows the Newman projections for the input and output configurations of the 1,2-dibromoethane molecule. From this input configuration, Figure 1.8(a), none of the energy minimization algorithms should be capable of finding the global energy minimum for 1,2-dibromoethane having a [Br–C–C–Br] dihedral angle of 180°. Instead, the local minimum with a [Br–C–C–Br] dihedral angle of 75°, in Figure 1(b) is found.

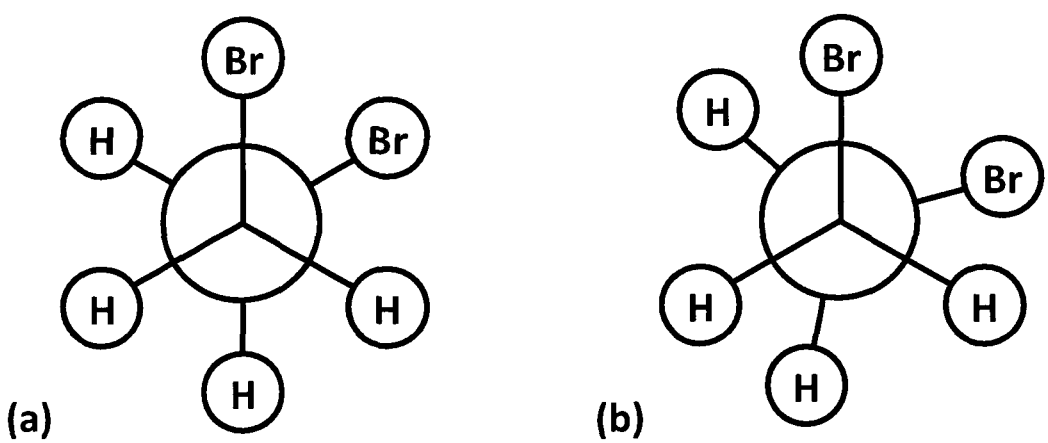


Figure 1.8 – Newman projections for minimization of 1,2-dibromoethane
The Newman projections of 1,2-dibromoethane for the **(a)** input configuration before energy minimization, and the **(b)** output conformation returned by energy minimization.

As shown in Figure 1.8, the expected 60° dihedral angle between proximal and distal substituents corresponding to the staggered conformation of ethane, is disturbed to approximately a 75° dihedral angle because of 1,4-interactions between the two bromine atoms to give the expected energy minimized geometry shown in Figure 1.8(b). Table 1.2 follows the energy minimization from input configuration to output conformation for the three energy minimization algorithms. Table 1.2 tracks the total potential energy (E_{tot}) and root mean square gradient (RMSG) at each iteration for the three enegry minimization algorithms, steepest descent (SD), conjugate gradient (CG) and truncated Newton (TN). The energy minimizations were run to convergence with a termination criteria of RMSG = 0.001 kcal/mol·Å.

As shown in Table 1.2, each of the three energy minimization algorithms find the same local minimum having a potential energy of 1.68740 kcal/mol. As can be seen, each algorithm is initiated from an identical input structure having an input RMSG of 3.64371 kcal/mol·Å. However, the efficiency for each algorithm is variable. The TN algorithm is the most efficient solving the energy minimization in 7 iterations. The CG algorithm converges at iteration 40 while the SD algorithm returns the local minimum after 179 iterations. The performance of each algorithm for the minimization of molecular systems is already well documented, particularly the poor rate of convergence observed for the SD algorithm.[36c, 40]

Table 1.2 – Energy minimizations of 1,2-dibromoethane

Iteration No.	Steepest Descent		Conjugate Gradient		Truncated Newton	
	E_{tot}	RMSG	E_{tot}	RMSG	E_{tot}	RMSG
0	3.64371	7.14864	3.64371	7.14864	3.64371	7.14864
1	3.17572	8.89632	3.17572	8.89632	2.43546	2.30353
2	2.92713	4.90674	2.64103	6.72776	1.76249	2.35198
3	2.79797	5.15073	2.43241	2.69283	1.69059	0.36883
4	2.70157	4.25285	2.36242	3.53330	1.68752	0.11620
5	2.62712	3.59898	2.25711	2.18749	1.68740	0.01470
6	2.56626	3.33611	2.17726	3.26741	1.68740	0.00318
7	2.51579	3.11769	2.05047	4.12019	1.68740	0.00080
8	2.47359	2.70435	1.95939	3.10119		
9	2.43593	2.85021	1.81005	2.67710		
10	2.40385	2.47346	1.77368	2.29340		
20	2.20321	1.82001	1.68772	0.12942		
30	2.07953	1.69746	1.68741	0.01364		
31	2.06896	1.49173	1.68741	0.00814		
32	2.05906	1.45310	1.68741	0.01514		
33	2.04942	1.41581	1.68741	0.01601		
34	2.04007	1.38023	1.68741	0.01150		
35	2.03096	1.34587	1.68740	0.01388		
36	2.02200	1.41542	1.68740	0.01232		
37	2.01339	1.26820	1.68740	0.00445		
38	2.00470	1.48262	1.68740	0.00272		
39	1.99631	1.30926	1.68740	0.00162		
40	1.98838	1.27645	1.68740	0.00092		
50	1.91803	1.10073				
60	1.86489	1.14367				
70	1.71980	3.32694				
80	1.69337	0.18677				
90	1.69199	0.18336				
100	1.69096	0.12739				
179	1.68740	0.00096				

*total energy (E_{tot}) in kcal/mol and root mean square gradient (RMSG) in kcal/mol·Å

With this information in mind, we can now begin to tackle the energy minimization of large molecules. The proceeding chapter will present and characterize large molecule minimization and the problems which arise due to its sensitivity to input perturbations and conditions.

Chapter 2

CHARACTERIZING PERTURBED ENERGY MINIMIZATIONS

2. 1. Introducing Input Perturbation Sensitivity

In 2008, Williams and Feher were the first to publish results which demonstrated that energy minimizations of large molecules gave alternate local minima when subjected to tiny coordinate perturbations before minimization.[41] They explained that a molecular system can only be sensitive to perturbation if the PES is complex enough to yield alternative local minima from near identical input structures. They showed that the spread in output structures scaled with the complexity of the molecular system which was approximated by heavy atom count. Williams and Feher also demonstrated that hardware and software differences can affect the outcome of an energy minimization. For this reason, a description of the hardware and software used to run experiments presented in this thesis can be found in the methods chapter (6. 1. Hardware and Software Specifications). The experiments conducted by Williams and Feher drew attention to the fact that initial condition sensitivity introduces substantial ambiguity into molecular simulation and computation.

The previous chapter introduced molecular mechanics (MM) and three energy minimization algorithms along with a sample problem where 1,2-dibromoethane was energy minimized to its local minimum (E_{tot} = 1.68 kcal/mol, [Br–C–C–Br] ≈ 75°). To introduce the concept of initial condition sensitivity and its presence in energy minimization we will consider repeated energy minimizations of a more complex PES. To illustrate that energy minimizations of large molecular structures exhibit initial condition sensitivity, the protein mycolic acid cyclopropane synthase CmaA2 (1KPI),[42]

was subjected to ten repeated energy minimizations, to a termination gradient (gtest) of 0.001 kcal/mol·Å, from both unperturbed and perturbed (0.001 Å displacements) input structures. A detailed copy of the procedure to introduce atomic perturbations to molecular structures can be found in the methods chapter (6. 3. Perturbation of molecular coordinates). The preparation methods for 1KPI can be found in the methods chapter (6. 4. The Preparation of Proteins for Experimentation). Figure 2.1 shows heavy atom overlays for the ten unperturbed and perturbed 1KPI proteins before and after minimization.

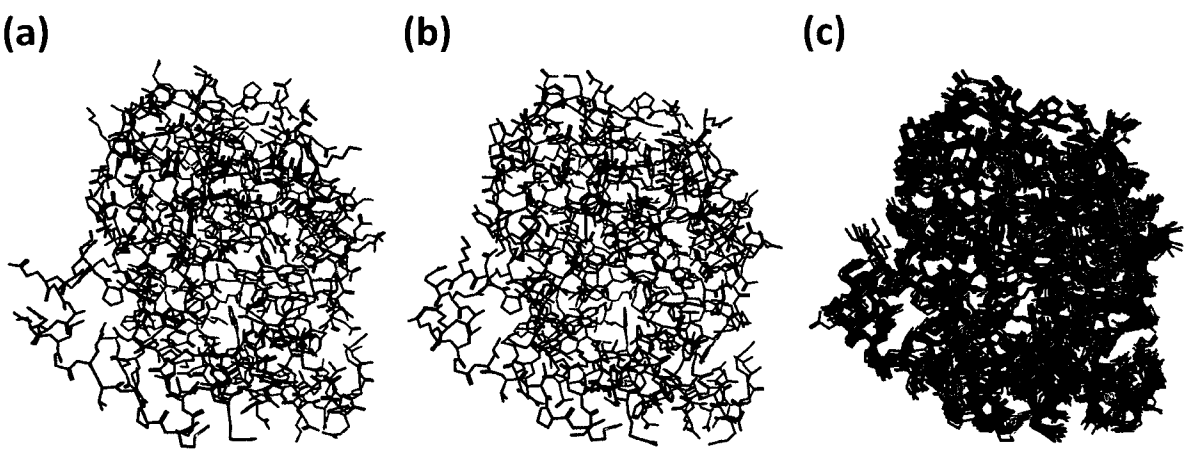


Figure 2.1 – Unperturbed and perturbed energy minimizations of 1KPI
Heavy atom overlays for **(a)** ten unperturbed and ten perturbed structures before energy minimization, **(b)** ten unperturbed structures after energy minimization, and **(c)** ten perturbed structures after energy minimization.

As shown in Figure 2.1(a), the seed and perturbed structures before energy minimization are visually identical. Minimizations of the seed structure set, shown in Figure 2.1(b), return ten identical structures while minimizations of the perturbed structure set, Figure 2.1(c), return ten unique local minima for 1KPI. Although the perturbed minimized structures in Figure 2.1(c) all appear to have the same

characteristic fold and shape, the property variations listed in Table 2.1 show that there is considerable variation between the ten structures.

Table 2.1 – Unperturbed and perturbed energy minimizations of 1KPI

Structure	Unperturbed		Perturbed	
	E_{tot} (kcal/mol)	RMSD (Å)	E_{tot} (kcal/mol)	RMSD (Å)
Before Minimization	3684.32 ± 0.00	0.000 ± 0.000	3688.91 ± 1.29	0.002 ± 0.000
After Minimization	-790.05 ± 0.00	0.000 ± 0.000	-817.87 ± 35.85	0.807 ± 0.116

To quantify the results, the total potential energy (E_{tot}) and pair-wise RMSD ($RMSD_{pw}$) properties of the input and output structure sets will be examined. In the case of E_{tot}, the spreads (standard deviation) are compared. $RMSD_{pw}$ is measured by comparing the RMSD between pairs of structures in the set, such that there are n(n-1)/2 total comparisons. In the case of $RMSD_{pw}$, the mean values between input and output are compared. As shown in Table 2.1, the unperturbed set of ten identical input structures returns no spread in E_{tot} (-790.05 ± 0.00 kcal/mol) and identical coordinates ($RMSD_{pw}$ of 0.000 ± 0.000 Å) after energy minimization. However, for the perturbed structure set the spread in output E_{tot} (-871.87 ± 35.85 kcal/mol) is one order of magnitude greater than that of the perturbed input set (3688.91 ± 1.29 kcal/mol). The mean $RMSD_{pw}$ is two orders of magnitude greater for the output (0.807 ± 0.116 Å) compared to the input (0.002 ± 0.000 Å). Because the minimization algorithm is deterministic and the spread in output properties are much greater than input properties, the minimizations of the 1KPI PES can be described to exhibit initial condition sensitivity. The definition of "much greater" is not explicitly provided by Williams and Feher in their 2008 paper. However, careful study of the minimizations

described by Williams and Feher to be sensitive to initial conditions, shows that all test structures had a spread in output properties that were at least one order of magnitude greater compared to spreads on input.

Thus, energy minimizations of a molecular system that are subjected to perturbations on input must meet the following qualifications in order to be considered to exhibit initial condition sensitivity.

(i) the energy minimization process must be deterministic, that is repeated energy minimization from the same (identical) input structure return identical geometry optimized structures, and

(ii) the spread in output properties, i.e. total potential energy or pair-wise RMSD, must be at least one order of magnitude greater than the spread in the same properties for the set of perturbed input structures

The atomic coordinates of proteins are typically determined via X-ray crystallography. Ligand and some solvent molecules are often included in crystallographic structures, however, the modeling of solvent is typically done implicitly. The resolution of conventional crystallographic structures varies depending on the protein. Generally the resolution of protein structures is greater than 1 Å. In the case of 1KPI, resolution of the crystallographic structure was accomplished to 2.65 Å. This means that the resolution of atomic coordinates involves an uncertainty in bond lengths which falls between 0.01 to 0.02 Å.[43] For the experiment presented in Table 2.1, the

perturbations introduced to each atomic coordinate (0.0017 Å) were far below the resolution of the atomic coordinates.

2. 2. Perturbation Size and Input Perturbation Sensitivity

The initial experiments conducted by Williams and Feher in 2008 did not examine the effect of perturbation size on the resulting spread on perturbed minimized structures. To investigate if such a dependence does exist, an experiment whereby energy minimizations of input perturbed structures generated with various perturbation magnitudes must be conducted. For this experiment, the protein ubiquitin (1UBQ)[44] refined to 1.8 Å resolution was subjected to atomic coordinate displacements of 0.0005 Å, 0.001 Å, 0.005 Å, and 0.01 Å to produce thirty structures for each perturbed set. For perturbation of atomic coordinates and preparation of 1UBQ see the methods chapter (6. 3. & 6. 4. respectively). Each structure was subjected to minimization (gtest = 0.001 kcal/mol·Å) in MOE[37] using the TN algorithm with the MMFF94s FF[23-24] and a distance dependent dielectric model.[40] Table 2.2 summarizes the results for the of spread in energy minimized structures and perturbation size.

Table 2.2 – Various perturbation sizes for minimizations of 1UBQ

Perturbation Size	Before Minimization		After Minimization	
	E_{tot} (kcal/mol)	$RMSD_{pw}$ (Å)	E_{tot} (kcal/mol)	$RMSD_{pw}$ (Å)
Unperturbed	1429.31 ± 0.00	0.000 ± 0.000	-1315.87 ± 0.00	0.000 ± 0.000
0.0005 Å	1429.79 ± 0.58	0.001 ± 0.000	-1292.71 ± 33.92	0.681 ± 0.125
0.001 Å	1430.73 ± 1.30	0.002 ± 0.000	-1297.67 ± 28.91	0.676 ± 0.134
0.005 Å	1465.73 ± 6.80	0.012 ± 0.000	-1291.67 ± 34.10	0.669 ± 0.136
0.01 Å	1571.85 ± 14.05	0.024 ± 0.000	-1305.90 ± 31.45	0.617 ± 0.147

Experiment conducted using Intel Celeron 2.6 GHz processor

Before energy minimization the unperturbed, 0.0005 Å, 0.001 Å, 0.005 Å and 0.01 Å perturbation input sets have deviation from crystal structure ($RMSD_{xtal}$) values of 0.0000 Å, 0.00087 Å, 0.0017 Å, 0.0087 Å and 0.017 Å respectively. As shown in Table 2.2, the input E_{tot} mean and spread increases with perturbation size, i.e. 1429.31 ± 0.00, 1429.79 ± 0.00, 1430.73 ± 1.30, 1465.73 ± 6.80 and 1571.85 ± 14.05 kcal/mol for 0.000, 0.0005, 0.001, 0.005, 0.01 Å perturbation sizes respectively. The $RMSD_{pw}$ results for input structure sets reflect the perturbation size and difference between structures as expected. Examination of the spreads in E_{tot} for minimized structures shows that the range of perturbation sizes do not affect the property variations of structures after energy minimization. For example, the spread in E_{tot} remains unchanged (33.92, 28.91, 34.10, 31.45 kcal/mol) for perturbed structures regardless of their respective perturbation size (0.0005, 0.001, 0.005, 0.01 Å). The mean E_{tot} values are also very similar as they all fall within one standard deviation of each other's E_{tot} distribution. The mean $RMSD_{pw}$ also shows that the spread in minimized structures is not affected by the range of perturbation sizes tested. For example, the mean $RMSD_{pw}$ (0.681 ± 0.125, 0.676 ± 0.134, 0.669 ± 0.136 and 0.617 ± 0.147 Å) for each distribution (perturbation sizes 0.0005, 0.001, 0.005, 0.01 Å respectively) lie within one standard deviation of the other distributions.

This experiment demonstrates that the spread in output properties is not dependant on the magnitude of perturbation for the range in displacements tested and applied to the set of structures prior to energy minimization. However, it is important to note that the two largest perturbed structure set (0.005 and 0.01 Å) have E_{tot} spreads

of the same order of magnitude for input (6.80 and 14.05 kcal/mol) and output (34.10 and 31.45 kcal/mol) structures. As well, the atomic coordinates in most crystal structures are resolved to an approximate certainty between 0.01 and 0.02 Å, which means that input structures arising from 0.01 Å displacements cannot be considered as identical starting points. As a result, perturbation magnitudes greater than 0.01 Å are not suitable for the study of initial condition sensitive minimization.

2. 3. The Effect of Gradient on Input and Input Perturbation Sensitivity

In Williams and Feher's 2008 publication the spread in output properties, specifically E_{tot}, was compared before and after energy minimization of their test structures.[41] While complexity of the PES is a requirement for the initial condition sensitivity of minimization, the input gradient can also affect spreads in output results. Small perturbations to input structures in a region of steep gradient will cause larger property variations. These variations have the potential to drastically alter the direction of descent or initial step size of an energy minimization which would result in the solution of alternate local minima. Larger input variations are expected to lead to more divergent output properties providing that the PES is already complex enough to warrant input condition sensitivity. It is also probable that the number of iterations required to arrive at a potential energy minimum will be increased when minimization is initiated from a steeper slope, since in general this means that the system is further from a minimum, and increasing the number of iterations is expected to contribute to

divergent output. To first establish that input gradient is a factor governing the spread in output results, the protein structure 1UBQ was subjected to bond compressions prior to perturbation and energy minimization. Table 2.3 shows the results of three bond compression tests, where bond lengths were compressed by 0%, 10% and 25% of their original length. Each compressed structure was subjected to perturbations involving atomic coordinate displacements of 0.0005 Å to generate a total of thirty structures for each of the three test sets. The energy minimizations were conducted using the same parameters as those used for the experiment present in Table 2.2 except that calculation was performed using Intel Pentium 4, 3.20 GHz single-core processors. The methods chapter (6. 5. Bond Compression Experiment) lists the procedural details to compress bonds in a molecular system.

Table 2.3 – Bond compression and energy minimization with 1UBQ

Property	0% Compression	10% Compression	25% Compression
Input RMSG[a]	31.86 ± 0.03	138.24 ± 0.06	539.95 ± 0.12
Output E_{tot}[b]	-1333.05 ± 20.42	-280.59 ± 21.10	-300.80 ± 23.15
Output $RMSD_{pw}$[c]	0.637 ± 0.134	0.794 ± 0.160	1.131 ± 0.210

[a]root mean square gradient (RMSG) in units kcal/mol·Å
[b]total potential energy (E_{tot}) in units kcal/mol
[c]pair-wise root mean square deviation ($RMSD_{pw}$) in units Å

Table 2.3 shows that even for 0% compression, the output $RMSD_{pw}$ (0.637 Å) far exceeds the $RMSD_{pw}$ for structures before energy minimization (0.0015 Å), demonstrating that 1UBQ is sensitive to input perturbations. Table 2.3 also shows that as the bonds are compressed by 0, 10 and 25% both the mean value and the spread in values (measured by the standard deviation) for the initial RMSG increases with the amount of compression. To test if the input gradient can influence the spread of

structures resulting from minimization, both the spread in E_{tot} and mean values for $RMSD_{pw}$ following energy minimization were examined. The spread in the E_{tot} does not increase very much (20.42, 21.10, 23.15 kcal/mol), while the mean values for $RMSD_{pw}$ do increase substantially with increasing bond compression (0.637, 0.794, 1.131 Å).

The number of iterations required for convergence to the minimum for the seed structure having 0, 10 and 25% compression was monitored to see whether this was correlated with the spread in output $RMSD_{pw}$. Optimization required a total of 208, 167 and 190 iterations respectively. Thus, the number of iterations required is not correlated with the initial RMSG. However, the output values of RMSD do roughly correlate with input RMSG.

The bond compression experiment conclusively links RMSG on input as a contributing factor to the input perturbation sensitivity of energy minimization. However, the bond compression experiment involved gross manipulation of input structures that are not physically realistic. Because it is desirable to characterize and reduce output spreads in minimized structures, or account for the source of these spreads, procedures currently available in the literature that might allow for users to reduce RMSG prior to minimization were studied.

It is important to note that the data for perturbed energy minimizations of 1UBQ shown in Tables 2.2 and 2.3 was collected using different computer hardare. The E_{tot} values for each experiment (-1292.82 ± 33.92 and -1333.05 ± 20.42 kcal/mol for Tables 2.2 and 2.3 respectively) are different, albeit they fall within at least two sigma of

the mean of each other. Williams and Feher have demonstrated that hardware specifications can dramatically alter the result of an energy minimization.[41] Given that these calculations can be altered by hardware and software specifications, all additional experiments presented in this thesis were conducted using identical Pentium 4, 3.20 GHz processors.

2. 4. The Addition of Hydrogens and Input Perturbation Sensitivity

In Williams and Feher's first paper detailing perturbed minimization they suggested that the discrepancy between energy minimization spreads conducted with MOE and Discovery Studio were in part due to hydrogen addition to the input structures.[41] Because hydrogen atoms are left unresolved in crystal structures, their addition into models can increase the RMSG on input due to steric clashes. To test this hypothesis, five protein structures were prepared with two hydrogen addition methods: default hydrogen addition and hydrogen addition with the function Protonate 3D.[45]

Introduction of hydrogen atoms in default orientations results in the hydrogen atoms being placed according to the geometry of the parent atom. For example, if the hydrogen is being added to a methyl carbon then hydrogen is added in an sp3 arrangement. However, for atoms having no specific orientation such as water, the addition of hydrogen is ambiguous and has the potential to result in steric clashes with neighbouring atoms in the molecular system. Protonate 3D is a built in MOE function which introduces hydrogen atoms based on internal and interaction energy scores for a

number of possible hydrogen configurations.[45] A unary-quadratic optimization process is then performed on the total number of possible hydrogen rearrangements in the system. Ultimately, the objective of the Protonate 3D function is to introduce hydrogens in the lowest potential energy arrangement. An example of hydrogen addition by both methods to a network of five water molecules is shown in Figure 2.2.

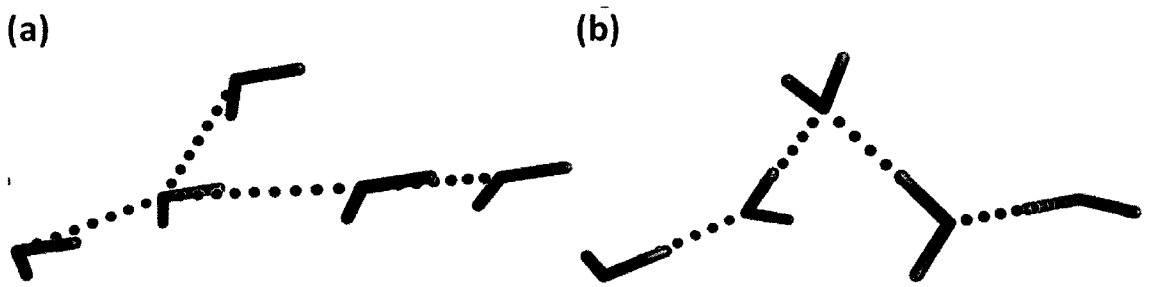


Figure 2.2 – Hydrogen addition using default and protonate 3D methods
Hydrogen addition to the same set of five water molecules using **(a)** default hydrogen addition, and **(b)** protonate 3D hydrogen addition. Hydrogen bonds (black dots) are drawn for clarity.

The default addition procedure is shown in Figure 2.2(a). Each set of hydrogens is added in the same orientation and the sequential orientations of four molecules in Figure 2.2(a) results in the formation of three hydrogen bonds, but the parallel orientation of the upper molecule will be repulsive with the second molecule in the lower sequence. In contrast, Figure 2.2(b) (which has the same oxygen positions) shows the result from using Protonate 3D, where all five water molecules have been reoriented in order to achieve a stable network. Similarly, Protonate 3D attempts to optimize hydrogen placement for amino acids containing rotatable O-H, S-H, S-C or C-N bonds on the side chains.

The five protein structures used in this experiment were: ubiquitin (1UBQ),[44] MecA-ClpC molecular machine (2Y1R),[46] alpha1-acid glycoprotein variant A (3APU),[47] the human estrogen-alpha ligand binding domain (1GWR),[48] and mycolic acid cyclopropane synthase CmaA2 (1KPI).[42] Two sets of structures, prepared using either default hydrogen addition or protonate 3D hydrogen addition, consisting of thirty perturbed structures were created for each protein and subjected to energy minimization using the TN algorithm available in MOE.[37] Once again, the MMFF94s FF and distance dielectric dependant solvation model were employed.[23,40] Energy minimizations were conducted to a final gtest of 0.001 kcal/mol·Å. The details of hydrogen addition are available in the methods chapter (6. 5. Hydrogen Addition Experiment). Table 2.4 shows the input RMSG for all, heavy and hydrogen atoms for the five proteins tested using default and protonate 3D hydrogen addition methodologies.

Table 2.4 – Input gradient after default and protonate 3D hydrogen addition

Default	**$RMSG_{tot}$**	**$RMSG_{hydrogens}$**	**$RMSG_{heavy\ atoms}$**
1UBQ	33.36 ± 0.03	21.26 ± 0.03	43.12 ± 0.04
2Y1R	89.32 ± 0.51	121.75 ± 0.74	27.88 ± 0.03
3APU	24.77 ± 0.03	22.77 ± 0.06	26.57 ± 0.02
1GWR	53.02 ± 0.12	67.84 ± 0.18	30.85 ± 0.03
1KPI	41.59 ± 0.10	52.62 ± 0.16	26.46 ± 0.02
Protonate 3D	**$RMSG_{tot}$**	**$RMSG_{hydrogens}$**	**$RMSG_{heavy\ atoms}$**
1UBQ	31.86 ± 0.03	16.84 ± 0.02	42.90 ± 0.05
2Y1R	22.71 ± 0.02	16.38 ± 0.02	27.85 ± 0.03
3APU	20.69 ± 0.02	11.56 ± 0.02	26.75 ± 0.03
1GWR	30.75 ± 0.06	30.59 ± 0.13	30.91 ± 0.03
1KPI	20.74 ± 0.01	12.34 ± 0.01	26.56 ± 0.02

*root mean square gradient (RMSG) values in units kcal/mol·Å

Comparing the initial RMS gradient for structures prepared with default hydrogen addition and Protonate 3D, it is clear that the gradient for the system can be heavily influenced. The changes in input RMSG arises solely from the addition of hydrogen atoms to the PES exclusively. For the five structures (1UBQ, 2Y1R, 3APU, 1GWR and 1KPI) the mean all atom initial RMS gradient ($RMSG_{tot}$) resulting from default hydrogen addition (33.36, 89.32, 24.77, 53.01 and 41.59 kcal/mol·Å) are larger than the mean $RMSG_{tot}$ obtained using Protonate 3D (31.86 , 22.71, 20.69, 30.75 and 20.74 kcal/mol·Å). The spreads in these distributions are also reduced by using Protonate 3D. For example, the most extreme case involves 2Y1R, where the standard deviation is reduced from 0.51 to 0.02 kcal/mol·Å. It is only for the smallest system, 1UBQ, where this reduction in $RMSG_{tot}$ spread is not observed. In general, the initial RMSG and its spread are smaller and have a sharper distribution when Protonate 3D is used for hydrogen atom addition to a crystal structure. The effect of hydrogen addition methodology on energy minimization is shown in Table 2.5 with E_{tot} and $RMSD_{pw}$ values.

Table 2.5 – Minimization of default and protonate 3D hydrogen addition structures

Default	**It. N°**	**E_{tot} (kcal/mol)**	**$RMSD_{pw}$ (Å)**
1UBQ	233	-1296.28 ± 33.09	0.672 ± 0.168
2Y1R	199	-1035.60 ± 29.80	0.651 ± 0.100
3APU	212	-962.21 ± 39.55	0.570 ± 0.125
1GWR	310	-338.33 ± 62.07	0.808 ± 0.110
1KPI	419	-2464.82 ± 35.08	0.661 ± 0.107
Protonate 3D	**It. N°**	**E_{tot} (kcal/mol)**	**$RMSD_{pw}$ (Å)**
1UBQ	134	-1333.05 ± 20.42	0.631 ± 0.135
2Y1R	130	-1031.76 ± 18.70	0.483 ± 0.086
3APU	169	-1020.17 ± 40.02	0.625 ± 0.114
1GWR	198	-356.10 ± 45.95	0.671 ± 0.138
1KPI	242	-2545.29 ± 55.03	0.700 ± 0.100

*iteration number (It. N°) provided for seed structure minimization only

The E_{tot} is generally improved on going from Default addition to Protonate 3D. For example, the E_{tot} for default hydrogen addition and energy minimized structures of 1UBQ (-1296.28 ± 33.09 kcal/mol), 3APU (-962.21 ± 39.55 kcal/mol) and 1KPI (-2464.82 ± 35.08 kcal/mol) is improved with Protonate 3D hydrogen addition resulting in E_{tot} values for 1UBQ (-1333.05 ± 20.42 kcal/mol), 3APU (-1020.17 ± 40.02 kcal/mol) and 1KPI (-2545.29 ± 55.03 kcal/mol). However, it should be noted this is not always the case as observed for 2Y1R having an E_{tot} (-1035.60 ± 29.80 kcal/mol) for minimizations with default hydrogen addition which has no improvement in E_{tot} (-1031.76 ± 18.70 kcal/mol) when prepared with Protonate 3D. The standard deviation in E_{tot}, which is the measure of input perturbation sensitivity, is only improved for three out of five structures. $RMSD_{pw}$ values show clear evidence that the structures all have divergent output and are sensitive to perturbations, since the $RMSD_{pw}$ increases from 0.0015 Å to approximately 0.6 Å for the five structures using either method. The change in $RMSD_{pw}$ on going from default to Protonate 3D hydrogen addition methods shows no clear trend, since only three out of the five structures show a decrease in divergence. To try to gain insight into why hydrogen atom addition gave inconclusive results, the number of iterations required to reach the final minimum are shown in Table 2.6. All structures prepared with Protonate 3D required fewer iterations to reach minimum, by a factor of approximately 1.5 compared to the default hydrogen addition structure sets. For example, the largest seed structure 1KPI required 242 iterations using Protonate 3D while minimizations of 1KPI structures prepared with default hydrogen orientations required 419 iterations. This result was unexpected, since it is expected that an increase

in input perturbation sensitivity (larger spreads) should result from an increased number of iterations. There are two possible reasons for these observations, it may be the case that variations in initial RMSG are not large enough to outweigh the complexity of the PES. As a result, any improvements to the input RMSG of a system will be masked. In support of this hypothesis, consider the bond compression experiments in Table 2.2. The initial gradient for 10% compression was larger than any of the entries for default hydrogen in Table 2.3 (138 compared to 122 iterations), yet the effects of gradient on bond compression were barely detectable on output spreads (no effect on E_{tot}, small effect on $RMSD_{pw}$). The second possible reason is that the improvements in input $RMSG_{tot}$ afforded by the Protonate 3D function are not substantial enough to significantly affect spreads in minimized structures. Examining Table 2.5 it can be seen that hydrogen addition methodology alters the RMSG for hydrogen atoms but has little to no affect on the RMSG of heavy atoms. For example, 2Y1R has input RMSG for hydrogen atoms which is improved from default placement (121.75 ± 0.74 kcal/mol·Å) by using Protonate 3D (16.38 ± 0.02 kcal/mol·Å), while the RMSG for heavy atoms is unchanged from default placement (27.88 ± 0.03 kcal/mol·Å) when using Protonate 3D (27.85 ± 0.03 kcal/mol·Å).

2. 5. The Effect of Force Field and Input Perturbation Sensitivity

The differences in minimized structure spread noted by Williams and Feher between the MOE and Discovery Studio software packages could have been due to

either the choice of force field (FF) used to model each MM system or the hardware and software parameters involved in running each experiment. FF choice could have a major impact on energy minimization because the parameterization of the FF will affect the RMSG on input for the system. If a particular FF is poorly parameterized for proteins compared to another FF, than the RMSG on input will be substantially greater. To investigate this possibility, energy minimizations were conducted on 1UBQ using the MMFF94s,[23-24] AMBER99,[25-26], and CHARM22/27[27-28] force field implementations available in MOE. An overview of each force field, their parameterization and function is provided in section 1.5. The experiment was conducted on ten perturbed structures prepared using the Protonate 3D hydrogen addition methodology.[45] Table 2.6 shows the input RMSG for the 1UBQ test protein set modelled with each of the four force fields.

Table 2.6 – Input gradient for 1UBQ with various force fields

	MMFF94s	AMBER99	CHARMM22/27
RMSG[a]	31.88 ± 0.04	160.35 ± 1.62	105.09 ± 0.95
E_{tot}[b]	723.77 ± 0.63	310.77 ± 4.94	1020.22 ± 3.03

[a]root mean square gradient (RMSG) values in units kcal/mol·Å
[b]total potential energy (E_{tot}) value in units kcal/mol

As shown in Table 2.6, the spreads in E_{tot} increase with increasing mean values in RMSG. For example, the standard deviation for E_{tot} (0.63 kcal/mol) modelled with MMFF94s (RMSG = 31.88 ± 0.04 kcal/mol·Å) is smaller than the standard deviation for E_{tot} (4.94 kcal/mol) when modelling with AMBER99 (RMSG = 160.35 ± 1.62 kcal/mol·Å). Of the three FFs tested, MMFF94s models 1UBQ with the smallest RMSG on input while

AMBER 99 models 1UBQ with the largest. The results (E_{tot} and $RMSD_{pw}$) for energy minimization of 1UBQ under each of the three force fields are presented in Table 2.7.

Table 2.7 – Minimizations of 1UBQ with various force fields

	MMFF94s	AMBER99	CHARMM22/27
$RMSD_{pw}$[a]	0.675 ± 0.133	0.672 ± 0.144	0.676 ± 0.133
E_{tot}[b]	-1344.03 ± 17.16	-1973.71 ± 29.59	-1582.91 ± 16.90

[a]pair-wise root mean square deviation ($RMSD_{pw}$) values in units Å
[b]total potential energy (E_{tot}) value in units kcal/mol

As shown in Table 2.7, the minimizations of 1UBQ in all force fields is sensitive to input perturbation. 1UBQ perturbed structures had a mean input $RMSD_{pw}$ (0.002 Å) that was two orders of magnitude less than the minimized sets of structures (0.675 ± 0.133, 0.672 ± 0.144, 0.676 ± 0.133 Å for MMFF94s, AMBER99 and CHARMM22/27 respectively). Despite the difference in input gradient, energy minimizations under each of the three FFs return nearly the same $RMSD_{pw}$. However, spreads in the total energy between minimized structure sets are different. For example, the standard deviation for the MMFF94s minimized set is 17.16 kcal/mol while the standard deviation for the AMBER99 minimized set is 29.59 kcal/mol. These results are consistent with the 2008 paper published by Williams and Feher which used the spread in total potential energy to characterized spreads in minimized structures and not the pair-wise RMSD of structures. These results show that the spread in potential energy between two systems can provide results which do not reflect the actual pair-wise RMSD values.

2. 6. The Minimization Algorithm and Input Perturbation Sensitivity

Another possible application involving gradients where gradient effects may be related to input perturbation sensitivity involves energy minimization algorithms, because they involve the repeated recalculation of gradients throughout an energy minimization. It has already been reported by Kini and Evans in 1991 that different minimization algorithms can substantially alter the resulting minima from an identical starting point.[49] If the gradient is a factor governing the spread in output of minimizations for perturbed structures it is important to consider that the algorithm employed for minimization may also influence the resulting spread in minima obtained. In this experiment the resulting property variations before and after minimizations was investigated with three minimization algorithms commonly used in computational chemistry: steepest descent (SD), conjugate gradient (CG) and truncated Newton (TN). Sections 1.5 through 1.8 in chapter 1 introduce energy minimization procedures in more detail. Briefly, SD is a first order minimization algorithm that locates a minimum by taking steps in the same direction as the first derivative.[3l,3m] SD exhibits extremely poor convergence when close to the minimum so it is rarely used exclusively.[36c] CG is also a first order minimization algorithm, however, it locates minima by taking steps which are orthogonal to the previous iteration's gradient to solve for the location of the nearest local minimum.[40] TN is a second order minimization algorithm that solves the location of local minima by approximation of the Newton equation.[39] Of the three algorithms discussed, TN usually takes the fewest iterations to reach a minimum.[40]

Table 2.8 shows results for minimizations using different algorithms with various termination criteria for the five sets of structures. The total potential energy (E_{tot}) and pair-wise ($RMSD_{pw}$) and crystal structure RMSDs ($RMSD_{Xtal}$) are reported. Of the three algorithms, CG and TN were run on their own and in combination with each other.

It is important to note that minimizations using SD were so inefficient that they were unable to reach a final convergence threshold of 0.001 kcal/mol·Å on their own in under 20,000 iterations, and so SD was only tested in combination with CG and TN algorithms. Each of the 5 structure sets (1UBQ, 2Y1R, 3APU, 1GWR and 1KPI) was prepared using Protonate 3D and perturbations were introduced using random 0.0005 Å displacements applied to each coordinate of each atom. Once again, the number of iterations required for minimization of each seed structure was tracked.

All minimizations, regardless of optimization algorithm employed or structure set studied, exhibit input perturbation sensitive behavior as their spread in output properties is at least one order of magnitude greater than the spread in their input properties. For example, the 30 perturbed structures of 1UBQ have a spread in input E_{tot} (0.56 kcal/mol) which is much smaller than the spread for output E_{tot} (10-20 kcal/mol) regardless of the specific minimization algorithm employed. What is clearly observed is that the spread in output structures can be drastically different depending on which minimization algorithm is employed. Examining 1UBQ, minimization completed with just the TN algorithm returns the largest spread (20.42 kcal/mol) whereas the CG, TN combination returns the smallest (10.27 kcal/mol).

Table 2.8 – Energy minimizations with SD, CG and TN algorithms

	Algorithm*		E_{tot} (kcal/mol)	$RMSD_{xtal}$ (Å)	$RMSD_{pw}$ (Å)
1UBQ	Input		723.00 ± 0.56	0.0009 ± 0.0000	0.0014 ± 0.0000
	TN 0.001	(It. 134)	-1333.05 ± 20.42	0.8055 ± 0.1116	0.6373 ± 0.1389
	CG 0.001	(It. 2639)	-1356.89 ± 10.54	0.6806 ± 0.0317	0.2446 ± 0.0705
	CG 0.1, TN 0.001	(It. 1064)	-1354.20 ± 10.27	0.6828 ± 0.0288	0.2378 ± 0.0647
	SD 1, CG 0.1, TN 0.001	(It. 1303)	-1353.14 ± 12.36	0.6977 ± 0.0151	0.2179 ± 0.0811
2Y1R	Input		1597.39 ± 0.51	0.0009 ± 0.0000	0.0014 ± 0.0000
	TN 0.001	(It. 130)	-1031.76 ± 18.70	0.9912 ± 0.0535	0.4882 ± 0.0847
	CG 0.001	(It. 3758)	-1004.66 ± 11.53	0.9679 ± 0.0327	0.3410 ± 0.1049
	CG 0.1, TN 0.001	(It. 1044)	-999.89 ± 10.21	0.9650 ± 0.0286	0.3239 ± 0.1064
	SD 1, CG 0.1, TN 0.001	(It. 1217)	-995.83 ± 7.36	1.0024 ± 0.0372	0.2315 ± 0.1072
3APU	Input		1834.65 ± 0.57	0.0009 ± 0.0000	0.0015 ± 0.0000
	TN 0.001	(It. 169)	-1020.17 ± 40.02	0.9177 ± 0.0788	0.6274 ± 0.1123
	CG 0.001	(It. 3910)	-967.98 ± 29.39	0.8497 ± 0.0301	0.2250 ± 0.0968
	CG 0.1, TN 0.001	(It. 890)	-961.65 ± 17.77	0.8415 ± 0.0139	0.2008 ± 0.0867
	SD 1, CG 0.1, TN 0.001	(It. 1306)	-974.38 ± 26.57	0.8763 ± 0.0457	0.3026 ± 0.1084
1GWR	Input		4089.56 ± 0.86	0.0009 ± 0.0000	0.0016 ± 0.0000
	TN 0.001	(It. 198)	-356.10 ± 45.95	1.0979 ± 0.0814	0.6652 ± 0.1361
	CG 0.001	(It. 4397)	-380.20 ± 20.82	1.1969 ± 0.0375	0.3499 ± 0.1034
	CG 0.1, TN 0.001	(It. 1319)	-377.35 ± 22.02	1.1874 ± 0.0373	0.3374 ± 0.0975
	SD 1, CG 0.1, TN 0.001	(It. 1804)	-384.84 ± 29.76	1.1673 ± 0.0277	0.4861 ± 0.1585
1KPI	Input		2433.75 ± 0.77	0.0009 ± 0.0000	0.0014 ± 0.0000
	TN 0.001	(It. 242)	-2545.29 ± 55.03	0.8896 ± 0.0703	0.6918 ± 0.1014
	CG 0.001	(It. 2639)	-2609.32 ± 30.31	0.8162 ± 0.0387	0.4604 ± 0.1158
	CG 0.1, TN 0.001	(It. 1110)	-2607.75 ± 30.55	0.8160 ± 0.0410	0.4656 ± 0.1225
	SD 1, CG 0.1, TN 0.001	(It. 1450)	-2594.50 ± 22.62	0.8551 ± 0.0530	0.4621 ± 0.1220

*algorithm heading includes structure set and termination criteria for each test in kcal/mol·Å as well as iterations required for seed structures to reach minimum in parentheses as (It.)

Finally, considering the $RMSD_{pw}$, CG provided output structures having reduced $RMSD_{pw}$ values over structures minimized with TN. This reduction in output $RMSD_{pw}$ (TN vs. CG) could be as little as 30% in the case of 2Y1R or as much as 64% in the case of 3APU. Looking at the entire set of five structures, it is the [CG, TN] method which is

most effective at reducing $RMSD_{pw}$ for three of five structures (2Y1R, 3APU, 1GWR) although the alternate method involving [SD, CG, TN] is also competitive (1UBQ, 2Y1R).

To try to understand the source of these variations, the number of iterations required for the seed structures to reach their minimum was monitored. The number of iterations with CG is at least one order of magnitude greater than required for TN, e.g. for 1UBQ TN required 134 iterations and CG required 2639. For the same structure the [CG, TN] combination required an intermediate value of 1064. These observations suggest that the second order derivative method (TN) is more sensitive to input perturbations than the first-order CG method, despite the reduced number iterations required to reach minimum. This seems reasonable when one considers that if there is indeed a divergent effect due to gradient, then a method involving the gradient of the gradient (TN) is likely to amplify the sensitivity of the minimization to perturbation.

Now that several factors which influence initial condition sensitive energy minimizations have been characterized, our experiments can be directed towards common molecular modeling techniques. Because energy minimization is ubiquitous in computational chemistry and a prerequisite to almost every study, a variety of minimization strategies have been developed to steer energy minimizations using constraints. These processes are examined in detail in the next chapter.

Chapter 3

CONSTRAINED ENERGY MINIMIZATIONS

3. 1. Previous Methods Involving Constrained Energy Minimization

Constrained energy minimization (CEM) is a common approach applied to the energy minimization of proteins. The intended goal of a CEM procedure is to generate geometry-optimized models having the smallest possible deviation from their crystal structure ($RMSD_{Xtal}$) or to provide structures having lower total potential energy values (E_{tot}).[36] These CEM procedures are typically employed prior to molecular modeling experiments such as docking or mutation analysis in hopes of enhancing the success of the study. Two generalized approaches to constrained minimization have been previously published in the literature, the first method (referred to as CEM-1) involves the use of parabolic tethers which are assigned according to groupings of atoms. For example, a publication in 1995[50] describes a procedure in which tethers were applied in series, where the structure is first minimized with all heavy atoms strongly tethered but with hydrogens free to relax; the structure is then minimized with backbone atoms tethered while allowing sidechain atoms to relax, and finally the structure is minimized without the use of tethers. Multiple variations on this procedure have been published, however the general approach is the same.

A second approach, involves constrained energy minimization based on use of shells (CEM-2). This method involves the use of linear tethers which are assigned to atoms according to their distance from a molecule of importance, for example a ligand.[51] As reported in the paper by Shadnia *et al.*, CEM-2 grouped atoms of the ligand and residues having atoms within 6, 12 and 18 Å into shells (L, S1, S2 and J

respectively).[51] The CEM-2 protocol was performed on a truncated version of the estrogen receptor alpha ligand binding domain referred to as 1GWR-TM.[20] Upon truncation of 1GWR-TM, the hydrogen caps from severed bonds were considered part of the 18 Å shell (J). Hydrogen atoms belonging to each shell, excluding the final shell (J) consisting of capped hydrogens, were also provided with separate tether constants in the protocol and are grouped into sets (L_h, S1_h and S2_h). The constrained optimization method then performs 9 iterations in which tethers are gradually released to relax the model, where additional tether is applied to L and S1 and less tether is applied to S2 and J. In the paper by Shadnia *et al.* the specific tether protocol employed was referred to as 'extended thaw'.

A tether, or constraint, is an addition potential energy term that is applied to a system such that deviation from a specified geometry (i.e. position) results in an energy penalty to the potential energy. A tether potential term often has one of two forms corresponding to a linear curve and a parabolic curve. A linear tether potential term, shown in equation 3.1, has the form:

$$V_{tether} = \sum_{i=1}^{n} k\sqrt{(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2} \qquad (3.1)$$

where, the energy penalty is summed for all n atoms in the system, k is the linear tether weight (units kcal/mol·Å), $[x_i, y_i, z_i]$ is the position of the atom (i), and $[x_0, y_0, z_0]$ is the local of the tether for the atom (i). A parabolic tether potential term, shown in equation 3.2, has the form:

$$V_{tether} = \sum_{i=1}^{n} k[(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2] \quad (3.2)$$

where, k is the parabolic tether weight (units kcal/mol·Å^2).[36a]

Because each type of CEM protocol is frequently used in the computational study of proteins, it is of interest to examine how these procedures behave in response to input perturbations. Table 3.1 shows the results of two CEM energy minimizations, where CEM-1 involved the use of parabolic tethers (k = 50 kcal/mol·Å^2) assigned by atom type and CEM-2 involve the use of linear tethers by shell definition. The minimizations were performed on a truncated structure of the estrogen receptor-alpha ligand binding domain (1GWR-TM).[20] 1GWR-TM is studied because it is the structure used in estrogen ligand docking studies performed by the Wright lab.[52] Protocol for the preparation of 1GWR-TM and its perturbation are provided in the methods chapter (6. 3. & 6. 4.) while the procedure for tethered minimization can be found in (6. 5. Tethered Energy Minimization Experiments). Both the seed (unperturbed original) structure and thirty perturbed structures (displacement = 0.0005 Å) were minimized using a conventional unconstrained energy minimization (UEM) and both CEM-1 and CEM-2 to a final termination gradient (gtest) of 0.001 kcal/mol·Å. The experiments were performed in the molecular operating environment (MOE)[37] using the built in truncated Newton (TN) algorithm with MMFF94s force field (FF)[23-24] and a distance dependant dielectric solvation model.[20]

Table 3.1 – Energy minimization with previous published tether protocols

Structure Set		Input	UEM	CEM-1	CEM-2
Seed	E_{tot}[a]	2770.65	-100.05	-144.19	-166.07
	E_{int}[a]	-56.85	-71.65	-74.40	-72.54
	$RMSD_{Xtal}$[b]	0.0000	0.8483	0.8574	1.0416
	$RMSD_{pw}$[b]	0.0000	0.0000	0.0000	0.0000
Perturb	E_{tot}[a]	2771.46 ± 1.31	-112.25 ± 15.40	-131 ± 30.70	-120.59 ± 21.44
	E_{int}[a]	-56.85 ± 0.04	-72.31 ± 1.71	-74.57 ± 0.78	-74.60 ± 1.75
	$RMSD_{Xtal}$[b]	0.0009 ± 0.0000	0.8946 ± 0.0434	0.8908 ± 0.0641	0.9753 ± 0.0571
	$RMSD_{pw}$[b]	0.0016 ± 0.0000	0.6173 ± 0.1007	0.4785 ± 0.1365	0.6857 ± 0.1679

[a]total energy (E_{tot}) and interaction energy (E_{int}) in units kcal/mol
[b]crystal structure RMSD ($RMSD_{Xtal}$) and pair-wise RMSD ($RMSD_{pw}$) in units Å

As shown in Table 3.1, energy minimizations with either CEM method is still susceptible to initial condition sensitivity. For example, the mean pair-wise RMSD ($RMSD_{pw}$) for each structure set (0.6173 ± 0.1007, 0.4785 ± 0.1365, 0.6857 ± 0.1679 Å for UEM, CEM-1 and CEM-2 respectively) was two orders of magnitude greater than the input $RMSD_{pw}$ (0.0016 ± 0.0000). At least one order of magnitude in spread is observed for the other three properties, total energy (E_{tot}), interaction energy between ligand and receptor (E_{int}) and the RMSD to the crystal structure ($RMSD_{Xtal}$). Examining the $RMSD_{Xtal}$ properties obtained for minimizations of the seed structure, it can be seen that both CEM methods fail to achieve their goal to reduce deviation from the crystal structure (0.8483 Å for UEM in comparison to 0.8574 and 1.0416 Å for CEM-1 and CEM-2 respectively). The same lack of improvement is observed for the perturbed structure sets.

3. 2. The Function of Tether Schemes to Reduce Variation in Minimizations

The purpose of the previous CEM approaches, however, was not to address the divergent output behaviour of minimization. The exact manner in which the tether constants are reduced can drastically reduce the spread in the four output properties. An overview of a new approach is illustrated schematically in Figure 3.1.

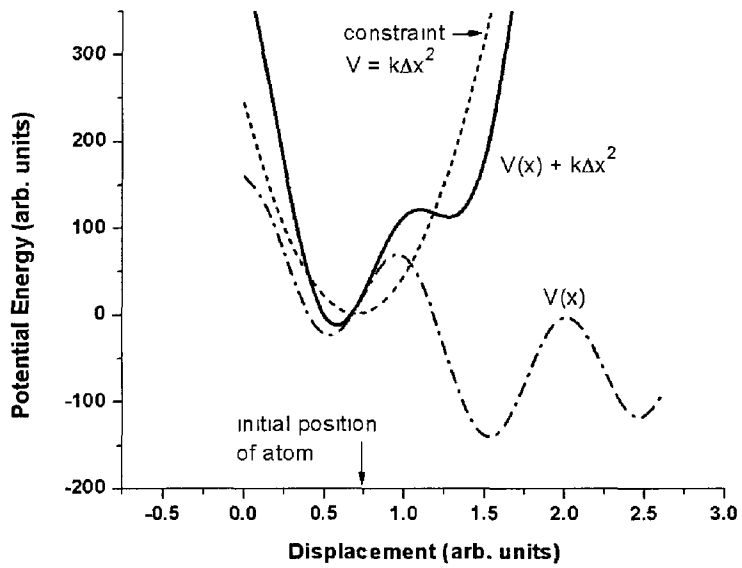


Figure 3.1 – Schematic drawing of the TEM method
A complex PES containing densely-packed minima (dashed-dotted line) has an imposed parabolic constraint (dashed line) at the current location of an atom on the surface. The sum of the two surfaces (solid line) restricts accessibility of the surrounding local minima. Imagine created in Origin 8.0 courtesy of Prof. Wright.

Here V(x) represents the true potential energy function, showing many minima and maxima. The atom is not initially at a potential minimum but is positioned on a sloping region of V(x) (see arrow on the Figure). A parabolic tether potential of the form shown in equation 3.2 is applied, centered at the initial position of the atom. An energy minimization is carried out on the composite surface (solid line) until a convergence

threshold is met, after which the tether constant is reduced and the process is repeated with the position of the tether reset to match the new coordinates of the atom.

This approach is referred to as a tethered energy minimization (TEM). Although the process is no different from a CEM, it is the concept and manner in which tethers are applied that makes them different. Specifically, the protocol calls for the application of tether constants to all atoms in an impartial fashion during the TEM procedure. As a result the system gradually heads, in an unbiased manner, toward the local minimum near where the parabolic potential is centered. The fact that tethers can be used to restrict accessible local minima is central to the method. As a result, the composite function (tether plus PES) is made less complex. It is also important to note that the method is in no way capable of searching for a global minimum using this procedure. However, finding a singular local minimum is desirable for several purposes, i.e. in preparing a protein receptor for ligand binding, or in the docking of ligands into protein cavities. For example, Feher and Williams conducted two studies on receptor score variations in which they characterized and suggested parameters to reduce variation. Their motivation and purpose of publication was to make readers aware of the fact that near identical placements of ligands during docking studies can give drastically different results.[53-54] Importantly, the TEM process serves to removes the ambiguity in discussion of output properties which necessarily accompanies a spread in output values.

Parabolic tethers employed during TEMs were set using MOE's MM function option (tetherWeight). Table 3.2 presents four TEM schemes which iteratively cycle

through decreasing values of tether constants to reproducibly minimize perturbed structures. The tether schemes, TEM-1 and TEM-2, involve linear decrements in tether constants while the tether schemes, TEM-3 and TEM-4 involve logarithmic decrements in tether constants. At each step the location of the tether is reset to match the current coordinates of the system. For example, in the case of TEM-4, k was set to an initial value of $10^{2.00}$ kcal/mol·Å^2 and the energy of the system was minimized with tethers present on all atoms until an RMSG of 0.001 kcal/mol·Å was achieved. This process was then repeated for logarithmically decreasing tether constants ($10^{2.00}$, $10^{1.99}$, $10^{1.98}$...) until the system was minimized using a final tether constant of $10^{-2.00}$ kcal/mol·Å. Following that, all tethers were removed and the structure was minimized to a root mean square gradient (RMSG) of 0.001 kcal/mol·Å. A detailed procedure for TEM procedures is provided in the methods chapter (6. 7. Tethered Energy Minimization Experiments).

Table 3.2 – Tethered energy minimization schemes

Scheme	No. Steps	Tether Constants[a]
TEM-1	119	100, 99, 98... 1, 0.9, 0.8... 0.1, 0.09, 0.08... 0.01, 0.00
TEM-2	200	100, 99, 98... 1, 0.99, 0.98... 0.01, 0.00
TEM-3	82	$10^{2.0}$, $10^{1.95}$, $10^{1.90}$... $10^{-2.00}$, 0.00
TEM-4	402	$10^{2.00}$, $10^{1.99}$, $10^{1.98}$... $10^{-2.00}$, 0.00
TEM-5	4002	$10^{2.000}$, $10^{1.999}$, $10^{1.998}$... $10^{-2.00}$, 0.00

[a]tether constants in units kcal/mol·Å^2

Energy minimizations using the TEM schemes listed in Table 3.2 were performed on four protein test structures. The test structure sets included: ubiquitin (1UBQ),[44] human biliverdin IX beta reductase (1HDO),[13] 1GWR-TM,[48] and a prokaryotic ubiquitin like protein (3M9H).[14] Each structure has a heavy atom count of 602, 1544, 2043 and

2088 respectively. Using a range of system sizes, the function of these TEM schemes can be tested to account for the effect of complexity. Each structure was subjected to thirty perturbations (displacements of 0.0005 Å) and energy minimizations with the TEM-4 scheme. Only TEM schemes one through three were tested on 1GWR-TM. This was done to determine which TEM scheme to use for the remaining three structures. TEM-5 was only tested on the largest structure 3M9H. Force field and minimization parameters were set in the same fashion as those used to test the CEM methods in section 3.1. Table 3.3 lists the results of energy minimization with the four TEM schemes.

Table 3.3 – Tethered energy minimization of perturbed structures

Structure Set		E_{tot} (kcal/mol)	$RMSD_{Xtal}$ (Å)	$RMSD_{pw}$ (Å)
Input	1UBQ	1215.44 ± 0.68	0.0009 ± 0.0000	0.0016 ± 0.0000
	1HDO	12460.78 ± 3.30	0.0009 ± 0.0000	0.0016 ± 0.0000
	1GWR-TM	2771.46 ± 1.31	0.0009 ± 0.0000	0.0016 ± 0.0000
	3M9H	1928.34 ± 0.88	0.0009 ± 0.0000	0.0016 ± 0.0000
UEM	1UBQ	-240.29 ± 10.60	1.1161 ± 0.1119	0.6679 ± 0.2284
	1HDO	-158.83 ± 28.95	1.0835 ± 0.0932	0.7901 ± 0.1533
	1GWR-TM	-112.25 ± 15.40	0.8946 ± 0.0434	0.6173 ± 0.1007
	3M9H	-3374.56 ± 36.81	1.7809 ± 0.1534	1.4362 ± 0.3016
TEM-1 1GWR-TM		-97.90 ± 2.38	0.9176 ± 0.0073	Not Tested
TEM-2 1GWR-TM		-99.02 ± 0.35	0.9289 ± 0.0021	Not Tested
TEM-3 1GWR-TM		-100.81 ± 0.35	0.9387 ± 0.0021	Not Tested
TEM-4	1UBQ	-241.05 ± 0.00	1.1191 ± 0.0000	0.0007 ± 0.0001
	1HDO	-128.02 ± 0.00	0.9745 ± 0.0001	0.0008 ± 0.0001
	1GWR-TM	-97.92 ± 0.00	0.9323 ± 0.0001	0.0009 ± 0.0001
	3M9H	-3352.48 ± 3.66	2.2087 ± 0.0562	0.3164 ± 0.2790
TEM-5 3M9H		-3353.41 ± 0.00	2.1765 ± 0.0012	0.0009 ± 0.0001

Consider results for 1GWR-TM, first, the spread in E_{tot} for input, UEM, and TEM-1 through 4 gives standard deviations of 1.31, 15.40, 2.38, 0.35, 0.06 and 0.00 kcal/mol,

respectively. This important result shows that while UEM is sensitive to input perturbation, the linear schemes TEM-1 and TEM-2 have returned the output to a non-divergent region (less than a factor of 10 increase in dispersion compared to input), the coarse log scheme TEM-3 has almost eliminated dispersion, and the fine log scheme, TEM-4, has eliminated the spread in results giving a singular result of -97.92 ± 0.00 kcal/mol. Note that the total energy for TEM-4 is not as low as that from UEM (-112.25 kcal/mol) but it does lie within the range of values obtained from UEM (-112.25 ± 15.40 kcal/mol). Similar results are observed for $RMSD_{Xtal}$. Examining the $RMSD_{pw}$ confirms that there is almost no deviation between the 30 values taken pair-wise.

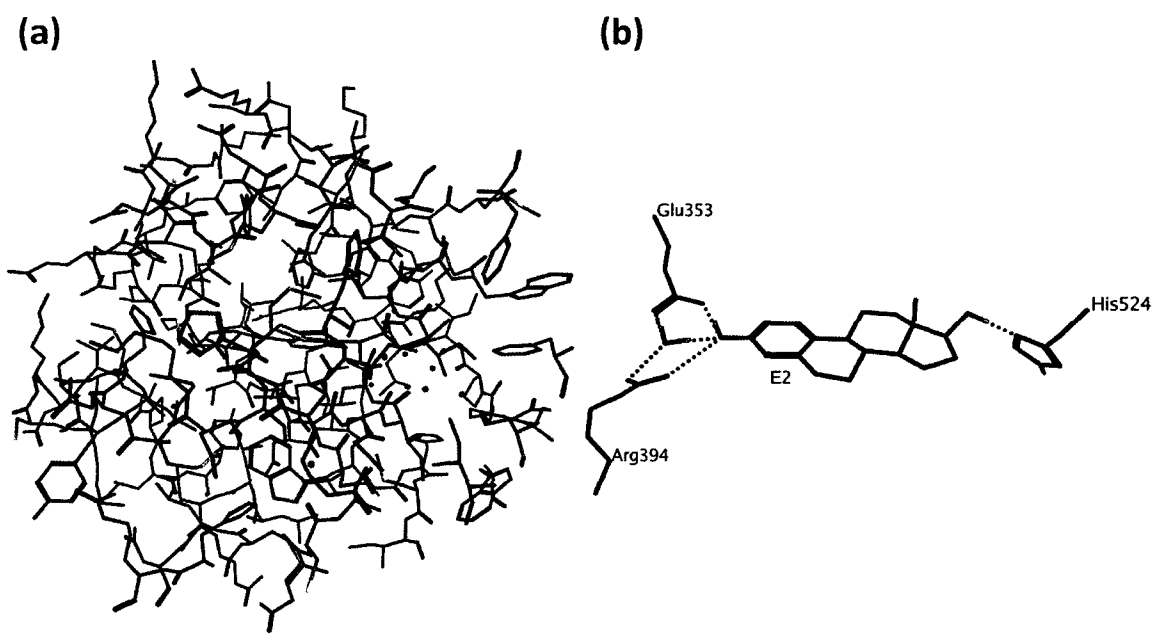


Figure 3.2 – TEM-4 results for 1GWR-TM
Heavy atom overlays for 30 perturbed structures of 1GWR-TM subjected to TEM-4 **(a)** the entire truncated model, and **(b)** a close-up of the active site. Hydrogen bonds (black-dots) drawn for clarity.

It is important that the energy minimum obtained using the TEM procedure correspond to a physical representation of the structure. Figure 3.2 depicts heavy atom

overlays for the thirty TEM-4 perturbed 1GWR-TM structures as well as the hydrogen bond network identified to be key to ligand binding.[55] Overlays of structures in Figure 3.2(a) appear to be identical, in agreement with results in Table 3.3. The overlays of all the active sites, showing only the key hydrogen-bonding residues (Glu353, Arg394 and His524), water (HOH2009) and ligand (E2), demonstrate that the local minimum has the correct configuration for ligand binding. A hydrogen bond interaction is considered only if the two partners are within a 3.5 Å radius. Because TEM-4 had the most success at narrowing output property spread for 1GWR-TM, TEM-4 was then tested on the additional three proteins also listed in Table 3.3. The smallest system 1UBQ has a standard deviation for E_{tot} after UEM which is about two orders of magnitude greater than that for the input structure (10.60 compared to 0.68 kcal/mol) so even the smallest chosen protein with 1231 atoms shows signs of minimizations that are sensitive to perturbation. The tethered minimization TEM-4, on the other hand, results in a single value for E_{tot}, with a standard deviation of 0.00 kcal/mol. $RMSD_{Xtal}$ results, comparing UEM and TEM output spreads, are reduced from 0.1119 Å (UEM) to 0.0000 Å (TEM-4). Perturbed structures, once again, returned similar output mean values for Etot (-240.29 ± 10.60, -241.05 ± 0.00 kcal/mol) and $RMSD_{Xtal}$ (1.1161 ± 0.1119, 1.1191 ± 0.0000 Å) for UEM and TEM-4 respectively. Again TEM-4 returned a singular geometry optimized structure within the range of the distribution of UEM structures. The patterns are similar for the larger system 1HDO, which shows signs of input perturbation sensitivity after using UEM, whereas there is almost no spread in E_{tot} using TEM-4. It is only for the largest system 3M9H (4284 atoms) that the tethering scheme begins to show signs of

instability. Comparing input and UEM standard deviation for E_{tot} (0.88 vs. 36.81 kcal/mol) demonstrates that structures after minimization begin to diverge (greater than an order of magnitude in output/input standard deviation). As might be expected, increasing the number of steps in the tether scheme (i.e. TEM-5) to further simplify the PES can further decrease the spread of results.

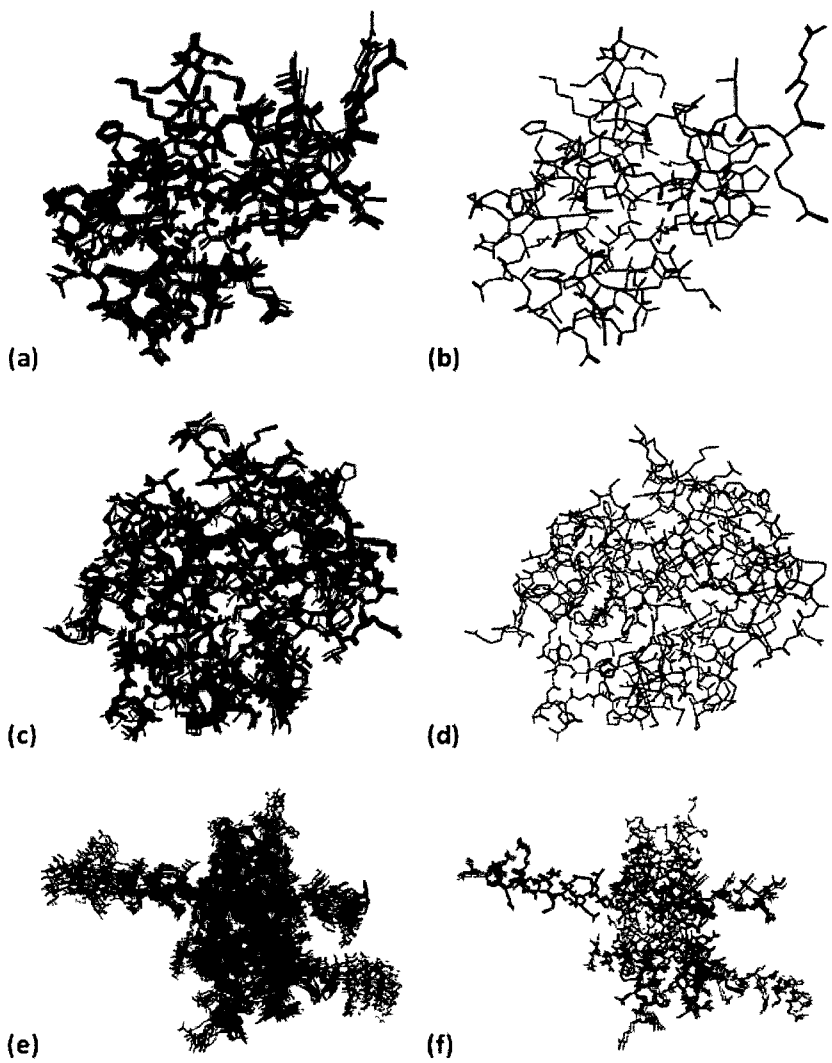


Figure 3.3 – UEM and TEM-4 results for three test proteins
Heavy atom overlays for 30 perturbed structures of **(a)** 1UBQ subjected to UEM, **(b)** 1UBQ subjected to TEM-4, **(c)** 1HDO subjected to UEM, **(d)** 1HDO subjected to TEM, **(e)** 3M9H subjected to UEM, and **(f)** 3M9H subjected to TEM-4.

Figure 3.3 shows the results of UEM and TEM-4 protocols applied to 1UBQ, 1HDO and 3M9H. As can be seen, UEM heavy atom overlays produce spreads in structures while the TEM-4 method, except for 3M9H, shows no spread in output.

It is also of interest to keep track of output variables during the course of a tethered minimization run. Table 3.4 shows the result for thirty tethered minimizations using perturbed structures of 1GWR-TM.

Table 3.4 – The behaviour of TEM-4 with 1GWR-TM

Step No.	K	E_{tot}	RMSG
Input		2773.3425 ± 2.9592	65.3941 ± 0.3390
1	100.000	1553.2301 ± 0.1695	7.0873 ± 0.0021
51	31.623	301.2245 ± 0.1894	0.4320 ± 0.0009
101	10.000	137.6220 ± 0.0235	0.1498 ± 0.0001
151	3.162	-12.0088 ± 0.0402	0.0764 ± 0.0005
201	1.000	-81.8965 ± 0.0115	0.0221 ± 0.0000
251	0.316	-94.4080 ± 0.0031	0.0032 ± 0.0000
301	0.100	-97.9213 ± 0.0001	0.0008 ± 0.0002
351	0.032	-97.9214 ± 0.0001	0.0005 ± 0.0001
401	0.010	-97.9214 ± 0.0001	0.0005 ± 0.0001
Output	0.000	-97.9214 ± 0.0001	0.0005 ± 0.0001

[a]tether constant: k (kcal/mol·Å^2)
[b]total energy: E_{tot} (mean ± stdev, kcal/mol)
[c]RMS gradient: RMSG (mean ± stdev, kcal/mol·Å)

As the tether constants are decreased the spread in total energy values also decreases, e.g. for tether constants [100, 10, 1, 0.1, 0.01 kcal/mol·Å^2] the spread in values of E_{tot} is [0.1695, 0.0235, 0.0115, 0.0001, 0.0001 kcal/mol], respectively. Convergence is reached at a tether constant of 0.1, suggesting that the last decade (from 0.1 to 0.01) is unnecessary. The convergence to a final output value of E_{tot} = -97.9214 ± 0.0001 kcal/mol is very stable. This means that all the perturbed structures

are being "focused" toward the final (single) output value for both E_{tot} and RMSG. This focusing behavior is the opposite of the divergent behaviour encountered during UEM. It is important to note that while the spread in perturbed minimized structures is decreased, no two structures take an identical path to the minimum.

The experiments in chapter 3 have demonstrated that the energy minimization of proteins with commonly applied tethering procedures remains sensitive to input conditions. More importantly, minimizations with the two CEM methods show that the published protocols actually fail to give energy minimized structures with decreased coordinate deviation from their crystal structure (compared to unconstrained energy minimizations). The final experiment presented in this chapter demonstrates that the modified tethering protocol, where tethers are applied in an unbiased manner and gradually released, can be used to simplify the PES. Minimizations of the simplified PES have been shown to return near identical results from perturbed input structures.

The next chapter in this thesis examines the relationship between structure sets before and after energy minimization. The next chapter also investigates the effect of three docking energy minimization methods.

Chapter 4

MOLECULAR DOCKING AND RECEPTOR MODEL BIAS

4. 1. Molecular Simulations Involving Energy Minimization

The molecular simulation of proteins spans a variety of fields which include, homology modeling,[36e] rotamer (mutation) searching,[36f] and ligand docking simulation.[36g] These three types of simulation, are all performed on large proteins involving combinatorial problems. For example, homology modeling involves the use of a template protein to create a model of a protein whose structure has not been solved. Rotamer and mutation searching involves the substitution of amino acid side chains, and ligand docking studies involve combinations of ligand conformers and orientations in an active site. Given the magnitude of these problems, the typical strategy is often to conduct partial energy minimizations or to conduct the experiment without any energy minimization. Another approach involves eliminating the number of inputs to be minimized by using a pruning function that selectively chooses which structures should or should not be energy minimized. Table 4.1 lists a collection of papers which were conducted without energy minimization.

Table 4.1 – Examples of molecular simulations not involving minimization

Ref.	Simulation	Description
55	Docking	Rigid docking of aryl-hydrocarbons into a homology model
56	Rotamer	Rigid rotamer search on red fluorescent proteins
57	Docking	Rigid docking of glutamate racemase inhibitors
58	Homology	Rigid homology modeling of hen egg white lysozyme

While the experiments in Table 4.1 were not conducted using energy minimization, this does not mean that they were ineffective. The common element to these studies is that the best result, (i.e. the top ranked docking pose, receptor model,

mutation) does not always match the experimental data. Instead, a more poorly ranked result gave the more favourable mutation when considering experimental data.[56] This approach is sufficient when the objective of the experiment is to eliminate a large number of potential solutions and narrow down the correct solution for experimentation. The design of lead compounds, for instance in docking simulation, requires that the correct orientation, pose and score for a given ligand be predicted to match experimental data.

An energy minimization process applied to a large number of inputs can yield results with one of two outcomes, (i) the order of output is related to the order of input structures, or (ii) the order of output is not related to the order of input structures. Figure 4.1 illustrates these two potential outcomes.

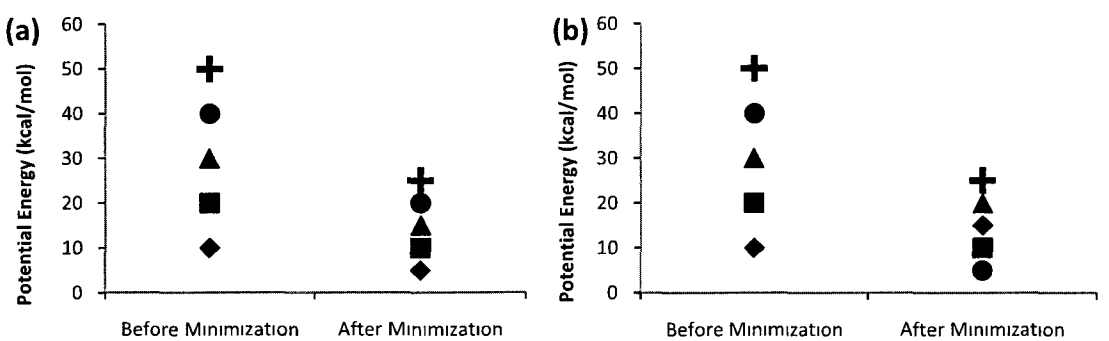


Figure 4.1 – The relationship between minimization input and output structures
The two potential outcomes of an energy minimization, where **(a)** the ranking of the input structures determines the ranking of the output structures, and **(b)** the ranking of the input structures does not determine the ranking of the output structure.

The majority of cases are similar to the schematic shown in Figure 4.1(b) where the ranking of the output structure is not consistent with the ranking of the input structure. This has had a substantial impact on the molecular modeling field, and as a

result several of the common pruning or searching functions (i.e. dead end elimination, mean field elimination) that would typically be implemented only for input structures[60] have been modified to involve energy minimization.[61-62] The fact that the input structure is not a physical representation of the potential energy minimum means that a successful docking simulation that can reproduce the orientation and predict the binding affinity of a ligand with a receptor must be conducted by scoring poses at energy minima.

There are two experiments in this chapter, the first investigates how flexible docking methodology influences the variation and spread in results and, the second investigates the impact of receptor geometry on the outcome of a molecular docking simulation.

4. 2. Refinement Procedures in Docking Simulation

The goal of a docking simulation is to correctly identify the binding mode between two molecular structures. A classic example of docking simulation involves small molecular weight ligands and their binding to large protein structures, i.e. 17β-estradiol (**E2**) and its agonist action upon forming a bound complex with estrogen receptor-alpha (**ERα**).[52,63] The general approach to docking simulation is illustrated in Figure 4.2. First, the ligand set must be prepared which involves a conformer search of the ligands for input. The receptor target must also be prepared, typically this involves the addition of hydrogens and an energy minimization. Once the active site is defined, a

database of ligand input poses is generated. These poses can be refined by molecular dynamics or energy minimization. The means of refinement can be highly variable depending on the docking simulation. Methods that do not involve energy minimization are referred to as rigid dockings. After refinement, the poses are scored and the best pose for each ligand is correlated to experimental data to predict binding affinity, the result is a quantitative structure-activity relationship (QSAR).[3n]

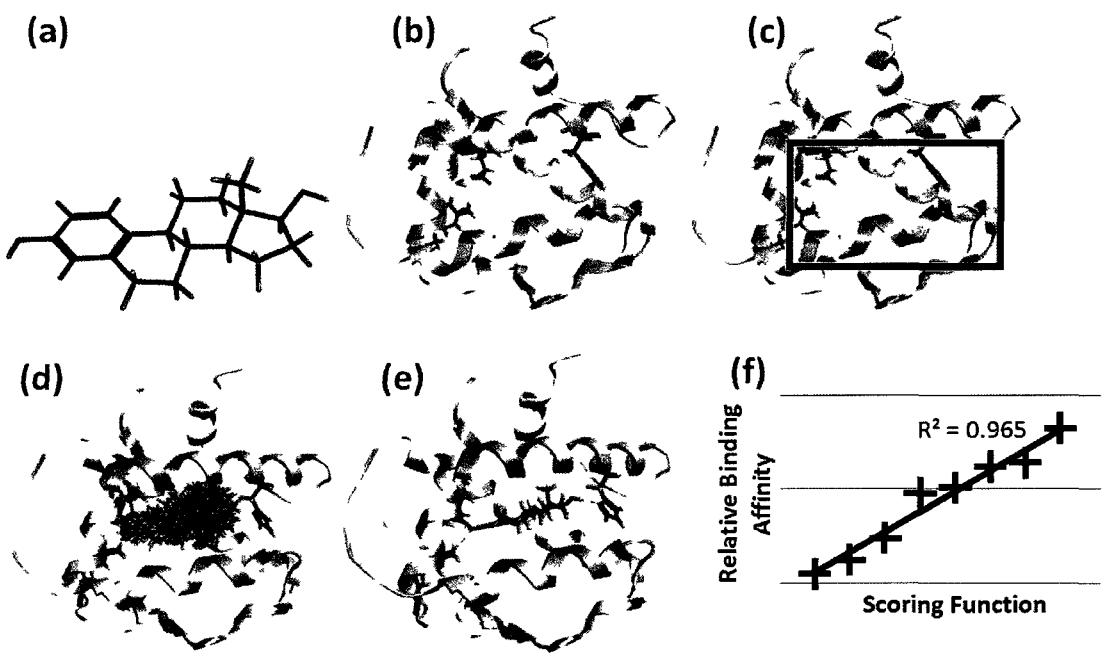


Figure 4.2 – An overview of molecular docking simulation
A schematic of the ligand-receptor docking process where, **(a)** ligands are prepared by performing a conformer search, **(b)** the receptor target is prepared via energy minimization, **(c)** the docking site is identified, **(d)** the docking site is populated with ligands, **(e)** each ligand pose is refined, **(f)** the simulation results are scored and compared to experimental data.

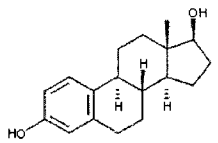
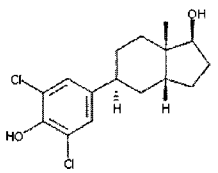
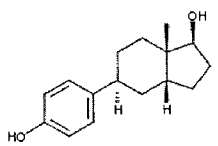
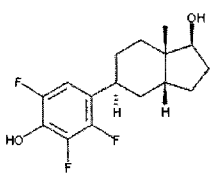
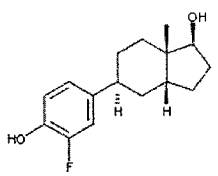
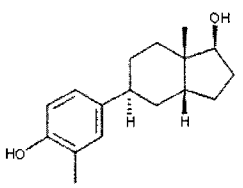
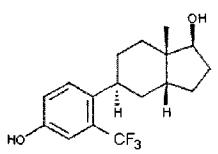
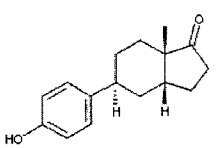
Feher and Williams published two additional papers in 2009 and 2010 on rigid docking score variability and provided guidelines to reduce the spread in results.[53-54] However, no study has been conducted on an analysis of pose variability in molecular

docking simulations involving energy minimization. As mentioned previously, the method of pose refinement can be highly variable depending on the docking methodology. For example, studies of **ERα** by Wright *et al.* involve the uDock application which performs a constrained energy minimization to a partial minima.[52] Variability in a docking study could originate for any number of reasons. For example, in chapter 2 we saw that gradient of the system and choice of minimization algorithm can alter the final minimized structure. In chapter 3 we saw that different constraint treatments during minimization altered the final minimization result. The process is further complicated by initial condition sensitivity because the systems involved are often large and the ligand input poses are not near a potential energy minimum. Therefore, it is important to study and understand how the refinement protocol employed during docking can alter the outcome and spread in results of the simulation.

The results and spreads for three refinement procedures were examined in this experiment, (i) an unconstrained energy minimization (UEM) of input poses, (ii) a constrained energy minimization (CEM) of input poses followed by an UEM, and (iii) a fixed receptor minimization (FIX), followed by CEM and UEM of input poses. The docking simulations were run on a truncated model of **ERα** referred to in chapter 3 as 1GWR-TM (see 6. 4. for preparation methods of 1GWR-TM). The 1GWR-TM model was subjected to the TEM-4 procedure described in chapter 3, and the ligand set used in docking simulation is shown in Table 4.2. Input positioning of ligands was completed using the triangle matcher function in MOE.[37] A copy of the program used to conduct the docking experiments and a description of the methodology is available in the

methods chapter (6. 8. Docking Simulation Procedures). The docking input poses were minimized using a cascade of conjugate gradient (CG) and truncated Newton (TN) algorithms to termination gradients of 0.1 and 0.001 kcal/mol·Å respectively. The simulations were conducted in MOE[37] using the MMFF94s[23-24] FF with a distance dependent dielectric solvent model.[20] Eight ligands were used in this docking experiment.

Table 4.2 – Estradiol and A-CD ligand series for docking studies

Ligand		Log RBA[a]	Conf.	Ligand		Log RBA[a]	Conf.
E2		2.0000	6	**1h**		-2.3979	6
1a		0.1673	6	**1k**		-0.7305	12
1b		0.0170	12	**1l**		0.2430	12
1f		1.9528	12	**2**		-2.0969	2

[a]Log RBA values taken from ref. 10

As shown in Table 4.2, the majority of modifications to each ligand are found on the A-ring of the steroid like structures. Binding affinity is provided as the log of the relative binding affinity (RBA). For example, RBA data is relative to 100% binding for **E2**

(therefore $\log RBA_{E2}$ = 2.00), thus ligand **1a** binds with 1.5% affinity (relative to **E2**). The naming of compounds docked in this study matches the nomenclature provided by Wright *et al.*[52]

Each of the three docking refinement procedures tested involved identical input poses. To test for the spread in results for the docking simulations, each input pose was perturbed a total of eight times with a displacement magnitude of 0.001 Å about the center of mass. This means that the input pose of each ligand was translated in the active site [± 0.001, ± 0.001, ± 0.001] without modification to the internal coordinates of the ligand. Three potential energy based descriptors (and their spreads) were computed for each output pose: the interaction energy between the ligand and the receptor (E_{int}), the deformation energy of the ligand from its global minimum (ΔE_L), and the deformation energy of the receptor from its local minimum without ligand present (ΔE_R). Choice of the final docked pose was selected in the same manner as published by Wright *et al.* where poses with a ΔE_L greater than 10 kcal/mol were discarded and each pose was rank ordered according to its E_{int} (most negative value is retrieved). **E2** was included as a control because a successful docking study should be capable of identically reproducing the orientation of the native ligand bound to the target receptor. Thus, the RMSD of the docked **E2** relative to the native structure was computed ($RMSD_L$).

Two QSARs were calculated using the logRBA data presented in Table 4.2. The first QSAR was a single variable linear regression of logRBA with E_{int} (QSAR-1), and the second QSAR was a multiple linear regression of logRBA with E_{int}, ΔE_L and ΔE_R (QSAR-2).

The results of the docking simulation are shown in Table 4.3. Construction of a QSAR using MOE is demonstrated in the methods chapter 6. 8.

Table 4.3 – QSARs for Three Docking Refinement Methods

Refinement Procedure	$RMSD_L$ (Å)	R^2 QSAR-1	R^2 QSAR-2
(i) UEM	0.086	0.53	0.93
(ii) CEM UEM	0.090	0.53	0.83
(iii) FIX CEM UEM	0.000	0.74	0.90

(QSAR-1) single variable linear regression of logRBA with E_{int}
(QSAR-2) multiple linear regression of logRBA with E_{int}, ΔE_L and ΔE_R

As shown in Table 4.3, only the third docking refinement procedure (FIX CEM UEM) was capable of returning a docked **E2** pose identical to the orientation of the ligand in the prepared and minimized receptor. However, it should be noted that the first (UEM) and second (CEM UEM) procedures could also predict the orientation of the **E2** to 0.086 and 0.090 Å respectively. This result is central to the method and an important indicator of a successful docking. **E2** serves as the control ligand because its coordinates when bound to the receptor in the active site are explicitly known. If the goal of a docking procedure is to return these coordinates, then it is essential that the control ligand orientation is reproduced.[64-65] The single variable linear regressions were poor for the first two refinement procedures (R^2 = 0.53 for both) while the third, which was successful at reproducing the native **E2** pose, was predictive (R^2 = 0.74). The multiple linear regressions were all predictive regardless of refinement procedure.

Considering these results, the third refinement procedure (FIX CEM UEM) is best suited for docking A-CD ligands in 1GWR-TM.

Just as Feher and Williams observed variations in score from input perturbation, refinement procedures applied to perturbed input poses also have variations. The variations in descriptors for top ranked ligand poses docked with each of the three refinement protocols are presented in Table 4.4. The properties of each minima found by the three refinement procedures can be drastically different.

Table 4.4 – Docking refinement procedure and spread in descriptors

	Ligand[a]	E_{int} (kcal/mol)	ΔE_L (kcal/mol)	ΔE_R (kcal/mol)
(i) UEM	**E2**	-74.67 ± 0.00	2.89 ± 0.00	-2.48 ± 0.00
	1a	-73.70 ± 0.29	8.94 ± 0.15	-5.75 ± 12.25
	1b	-72.16 ± 0.00	7.55 ± 0.00	-0.20 ± 0.01
	1f	-81.83 ± 0.83	8.33 ± 0.35	9.72 ± 4.14
	1h	-72.78 ± 0.00	9.20 ± 0.01	13.46 ± 0.06
	1k	-70.90 ± 0.15	9.22 ± 0.06	2.04 ± 0.20
	1l	-73.99 ± 0.05	9.06 ± 0.01	4.22 ± 1.46
	2	-60.68 ± 0.15	5.36 ± 0.12	-9.03 ± 0.02
(ii) CEM UEM	**E2**	-74.90 ± 0.08	2.89 ± 0.00	1.76 ± 1.50
	1a	-73.52 ± 0.38	8.95 ± 0.21	-6.96 ± 8.31
	1b	-72.16 ± 0.00	7.55 ± 0.00	-0.21 ± 0.01
	1f	-81.09 ± 0.13	8.37 ± 0.02	2.93 ± 0.25
	1h	-73.01 ± 0.28	9.08 ± 0.09	10.49 ± 2.50
	1k	-70.58 ± 0.15	9.59 ± 0.10	-6.82 ± 2.17
	1l	-73.29 ± 0.04	8.09 ± 0.01	15.09 ± 0.84
	2	-60.74 ± 0.00	5.42 ± 0.00	-9.02 ± 0.00
(iii) FIX CEM UEM	**E2**	-71.12 ± 0.00	2.59 ± 0.00	0.00 ± 0.00
	1a	-72.87 ± 0.02	8.57 ± 0.00	5.67 ± 0.02
	1b	-67.51 ± 0.11	8.09 ± 0.01	-5.52 ± 0.05
	1f	-78.27 ± 0.00	4.94 ± 0.00	-2.22 ± 0.00
	1h	-64.52 ± 0.18	5.45 ± 0.02	18.42 ± 0.13
	1k	-68.62 ± 0.01	8.78 ± 0.01	4.91 ± 0.52
	1l	-70.80 ± 0.00	6.11 ± 0.00	5.25 ± 0.00
	2	-60.69 ± 0.15	5.38 ± 0.12	-9.03 ± 0.02

[a]Docking refinement procedure indicated under the ligand heading

Consider **E2**, the third refinement procedure was only the only docking simulation capable of returning the orientation of the native ligand. The E_{int}, for the 1GWR-TM receptor minimized using the TEM-4 tether scheme, is -71.12 kcal/mol which is the same for refinement method three (-71.12 ± 0.00 kcal/mol). Other variations in the obtained minima can be seen, for example ligand **1h** has E_{int} values of -72.78 ± 0.28, -73.01 ± 0.28, and -64.52 ± 0.18 kcal/mol for the three refinement procedures (i) UEM, (ii) CEM UEM, and (iii) FIX CEM UEM respectively. It is clear that the pose obtained is dependent on the refinement procedure used. Of the three descriptors, spreads in output results are greatest for the ΔE_R descriptor while spreads are the smallest for the ΔE_L descriptor. The spread in results also appear to be ligand dependent. For example, ligand **1f** docked using the first refinement procedure (UEM) has spreads (0.83, 0.35, 4.14 kcal/mol for Eint, ΔE_L, and ΔE_R respectively) that are much greater than those for **E2** (all 0.00 kcal/mol). Thus, magnitude of spread is also dependent on the refinement procedure used. Careful examination of Table 4.4 shows that refinement procedure three (FIX CEM UEM) generally has the smallest spreads. This is similar to the reduction of results observed in chapter 3 when using TEM methods because the complexity of the potential energy surface has been substantially reduced when first performing the energy minimization with the fixed receptor.

These results demonstrate that refinement methodology and ligand choice clearly affect spread and variability in pose results. Of the three methods tested, refinement procedure three (FIX CEM UEM) was the only docking minimization that returned the correct orientation of **E2** while still maintaining predictive capability for

both single variable and multiple variable QSARs. Another important aspect of this study is that the 1GWR-TM model energy minimized using the tether scheme methodology (TEM-4) used for docking in this experiment gives a receptor model that is predictive for ligand-receptor binding.

4. 3. Introducing a New Approach to Docking Simulation

Preparation of the receptor model in the previous experiment required energy minimization prior to the docking simulation. However, as was demonstrated in chapter 2 and by Williams and Feher, the energy minimization of large molecules is sensitive to input perturbation.[41] When studying large systems, i.e. a ligand-protein docking study, it is generally impractical, or even irrelevant, to seek a global potential minimum. Instead, the target becomes a local potential minimum of relevance to the biochemical mechanism of interest. In general, the presence of input perturbation sensitivity in energy minimization poses a difficult problem because it is not clear how to decide if a particular receptor model is correct, when there is a spread of other possible receptor models. When considering receptors models for docking, the common solution to this problem is to choose the receptor model which has either the lowest total energy or the least coordinate deviation from the crystal structure.[36c] However, the act of choosing one particular receptor from the set may then result in "receptor model bias" because the results obtained from choosing between equally valid starting protein geometries may result in significantly different ligand binding results.

Receptor model bias has been discussed in other contexts in the literature. For example, in a paper published by Jain in 2008, a docking study with PPAR-gamma returned different binding modes for the native ligand which were dependent on the tautomeric state of a histidine and change of rotamer of a tyrosine hydroxyl group in the active site.[66] Two additional methods that allow for the exploration of different receptor conformations include the use of multiple crystallographic structures, as in a paper by Ferrari *et al.* involving docking studies of T4 lysozyme.[67] Another method involves the use of molecular dynamics.[68] Despite the case made by these three papers, namely that the consideration of multiple conformers is important in lead optimization, a literature search failed to find a docking study that investigated the effect of alternate local minima generated from near identical input structures.

The output from a docking calculation is a correlation between predicted and observed ligand relative binding affinities (RBA). As stated above, input perturbation sensitivity can affect the calculation both for the preparation of the receptor and, as was shown in the previous experiment, in the determination of the docked poses. In this case, input perturbation sensitivity can be considered to be a nuisance. In fact, Feher and Williams have stated that its presence requires a lengthy error analysis plus additional calculations to determine the effect of input perturbation sensitivity on error bounds for the property of interest.[53-54]

This next experiment demonstrates how docking results can differ based on receptor model choice by using the same set of ligands but varying the receptor model

geometry. Receptor model variations are introduced by using input perturbations before energy minimization, as a result all receptor models originate from near identical and valid input structures. This chapter demonstrates that the results from closely related receptors can be very different indeed. This may be one (largely unexplored) reason why there is so much variability in the success of molecular docking calculations.[69]

There are three objectives of this experiment. First, to establish that input perturbation sensitivity can be used to generate multiple predictive models by generating a set of 25 different but equally valid receptors and proving that their energy minimization leads to divergent results. The second, to examine the importance of receptor model bias on the selection of the best docked poses and the correlations which result from the different choices. Finally, it is important to determine whether the traditional rationale for receptor model choice (i.e. total energy or deviation from crystal structure are traditionally used to choose a receptor model) has any validity .

4. 4. Characterizing a Set of Receptor Conformations

As discussed in chapter 2, assessing an energy minimization as to whether it is sensitive to input perturbation requires that the spread in output properties must be much greater than the spread in input properties before minimization. Once again, 1GWR-TM and the ligand set in Table 4.2 was used in the docking simulation. Coordinates of 24 variations on the seed structure were generated by using random

perturbations to each atomic coordinate in the system of ± 0.001 Å. Just as before, these coordinate displacements all fall far below the resolution of the crystallographic structure determined to 2.4 Å.[41] Energy minimization was then performed on the seed structure, S*, and each of its 24 variants, S'_1* through to S'_{24}*. Confusion is possible as to what is being calculated so notation is provided in Table 4.5 showing the definitions adopted for this experiment.

Table 4.5 – Notation used for prepared receptor models

Notation	Structure type
S	Seed structure (with heavy-atom crystal coordinates)
S*	Seed structure minimized in presence of estradiol
S'_4	Perturbed seed structure, receptor ID number 4
S'_4*	Perturbed seed structure, receptor ID number 4, minimized in presence of estradiol
S*(**L**)	Seed minimized in presence of estradiol, followed by removal of the native estradiol and docked with ligand **L**
S'_4*(**L**)	Perturbed seed receptor number 4, minimized in presence of estradiol, followed by removal of the native estradiol and docked with ligand **L**

Here the symbol ′ implies perturbation to the seed coordinates, and * implies energy minimization in the presence of estradiol, so the notation S'_4*(**1a**) implies that the seed structure has been perturbed and energy-minimized in the presence of E2 to give perturbed receptor number 4 which was then used to dock the ligand **1a**, etc. Thus there will be a set of 25 receptors (seed +24 perturbed) prepared for docking studies with each of the 8 ligands presented in Table 4.2.

Figure 4.3(a) shows an overlay of the heavy atoms for the seed structure and the 24 receptor models after perturbation but before minimization. This figure shows that

the atomic displacements are too small to be distinguished visually. Figure 4.3(b) shows an overlay of the seed structure and the 24 receptor models after minimization. The effects of input perturbation sensitivity are clearly visible as receptor models in the overlay are clearly distinct.

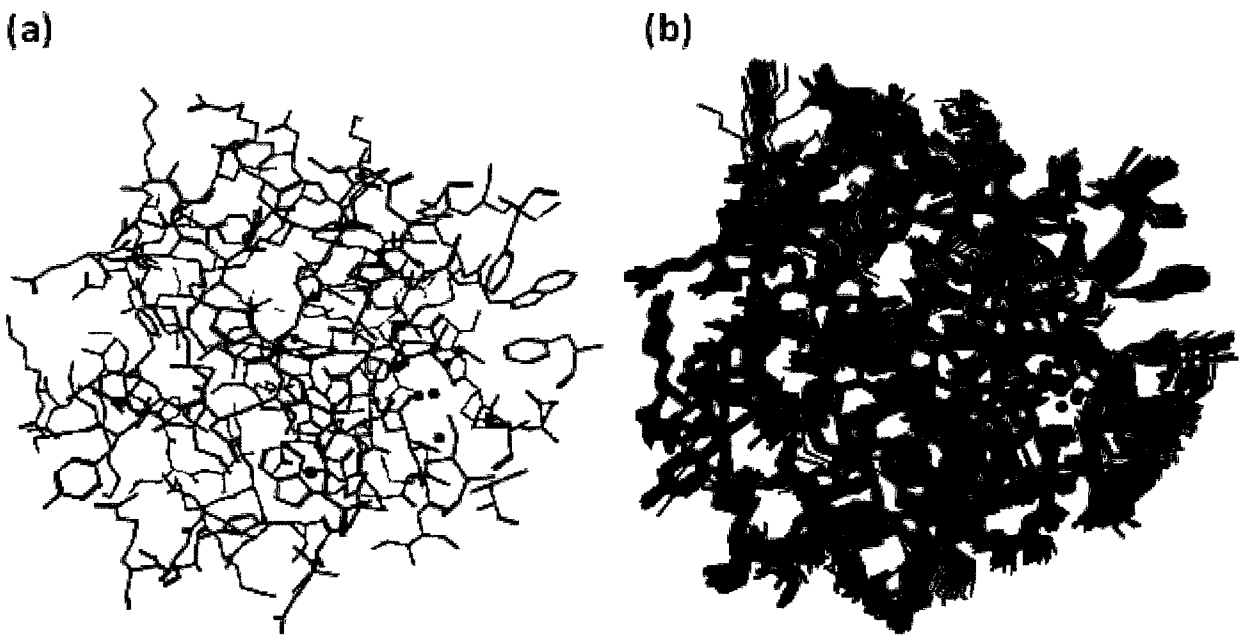


Figure 4.3 – Heavy atom overlay of the prepared seed and 24 perturbed receptors
Seed and perturbed structure heavy atom overlays **(a)** before energy minimization, and **(b)** after energy minimization

It is also important to monitor the difference between several variables before and after minimization, including the total energy (E_{tot}), the interaction energy (E_{int}), the root mean square deviation from the crystal geometry ($RMSD_{Xtal}$) and the pair-wise root mean square deviation between the 25 different receptor models ($RMSD_{pw}$). These property values are characterized in Table 4.6 by their mean value and the standard deviation.

Table 4.6 – Receptor properties before and after minimization

Property	Before Minimization	After Minimization
E_{tot} (kcal/mol)	2060.39 ± 1.11	-136.12 ± 16.25
E_{int} (kcal/mol)	-56.12 ± 0.09	-72.94 ± 2.63
$RMSD_{Xtal}$ (Å)	0.0017 ± 0.0003	0.9424 ± 0.0662
$RMSD_{pw}$ (Å)	0.0024 ± 0.0002	0.6914 ± 0.1329

*variations are for a seed structure and 24 perturbed structures

Table 4.6 shows that the standard deviation for E_{tot}, E_{int} and $RMSD_{Xtal}$, all of which are measures of spread, increase by more than an order of magnitude after minimization. For example, the standard deviation in E_{tot} increases from 1.11 to 16.25 kcal/mol and the standard deviation in E_{int} increases from 0.09 to 2.63 kcal/mol. The RMSD between any two receptor structures ($RMSD_{pw}$) after minimization is 0.6914 Å. This result is consistent with input perturbation sensitivity since energy minimizations starting from nearly identical receptors have created a broad range of receptor geometries.

It is unclear at this point as to which receptors will be suitable or unsuitable for docking study and predicting ligand binding affinity. However, the arbitrary choice of one member of the set to choose exclusively for docking may result in "receptor model bias". Figure 4.4(a) shows the arrangement of the ligand **E2** in the crystal structure 1GWR, along with the important residues Glu353. Arg394, His524 and HOH2009, where hydrogen bonds are indicated as dotted lines. The hydrogen bonding network has been identified by Prathipati *et al.* as an important factor contributing about half of the enthalpy of ligand binding, the other half coming from hydrophobic interactions (not shown).[52] For the optimized receptors, the contacts between the ligand and the receptor were visually examined to see if they resembled those shown in Figure 4.4(a).

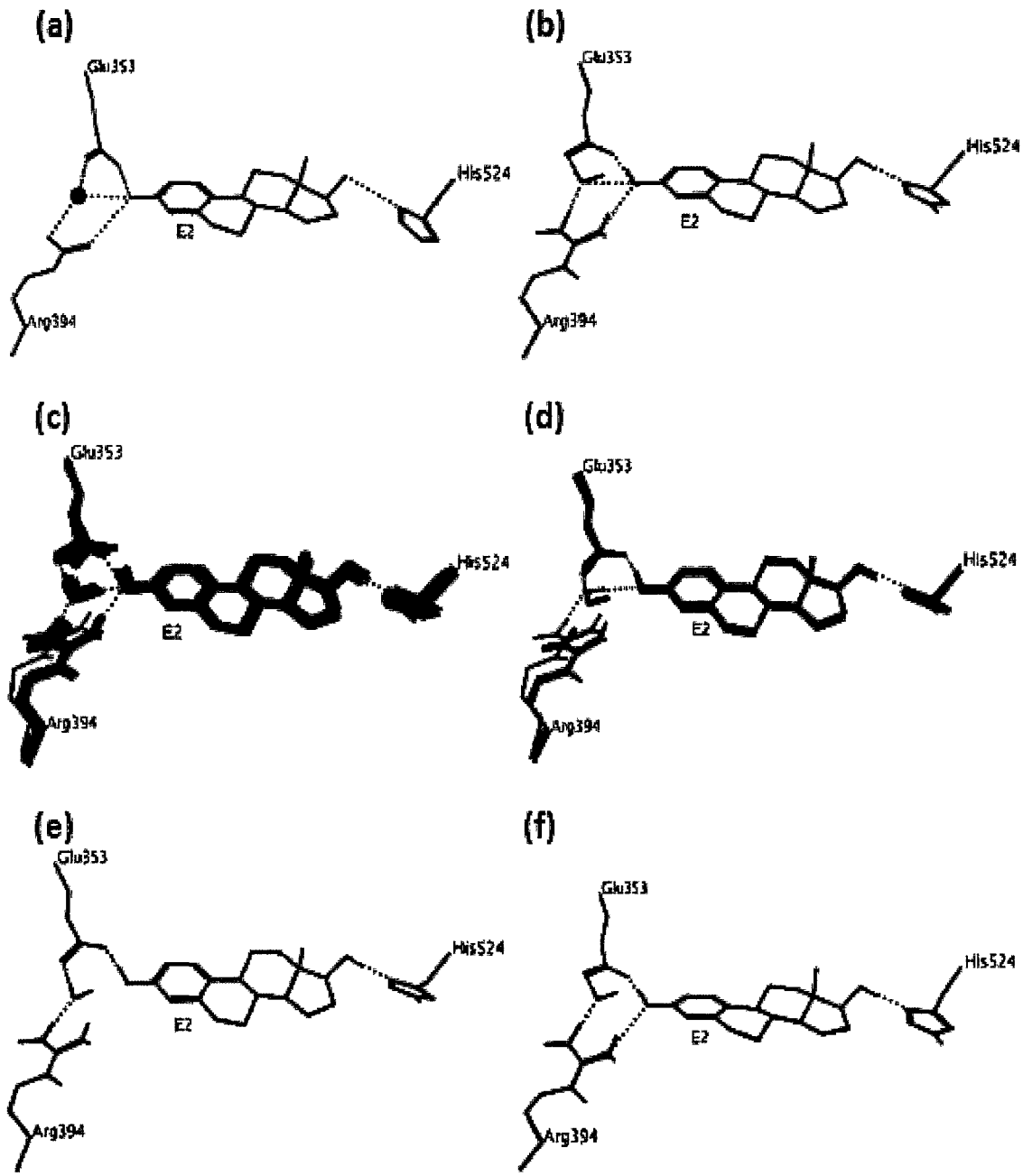


Figure 4.4 – Hydrogen bonding network in the set of minimized receptors
Close-up of the estrogen ligand **E2** bound to the receptor (1GWR) with H-bonds (black dots) involving glutamic acid (Glu353), arginine (Arg394), histidine (His524) and a bridging water molecule (HOH2009). Non-polar hydrogens and remaining residues are omitted for clarity. **(a)** crystal structure from PDB file, **(b)** S* showing correct H-bond network, **(c)** $S'_{1\text{-}19}$* showing correct H-bond network, **(d)** overlay of $S'_{20\text{-}22}$* missing H-bond between Arg394 and **E2**, **(e)** S'_{23}* missing both H-bonds between Arg394, HOH2009 and **E2**, and **(f)** S'_{24}* missing the H-bond between HOH2009 and **E2**.

Figure 4.4(b) shows the optimized seed structure where the polar hydrogens are included explicitly. It can be seen that the orientation of the **E2** ligand and the pattern of hydrogen bonds is virtually identical to that in the crystal structure. From the set of 24 perturbed (optimized) receptors, 19 of them have the same bonding pattern as the seed structure. These are shown as an overlay in Figure 4.4(c), where minor differences can be seen but the hydrogen bonding pattern is identical. However, Figure 4.4(d) shows an overlay of three receptors having a different H-bonding pattern where a hydrogen bond between Arg394 and E2 is missing [donor and acceptor atoms must be within a 3.5 Å radius to be considered as an H-bond interaction]. Figures 4.4(e) and (f) show two other receptor structures where different hydrogen bonds are missing. Thus, from the set of 24 perturbed structures 19 showed the correct H-bonding pattern, identified as $S'_{1\text{-}19}$*, and 5 did not, identified as $S'_{20\text{-}24}$*.

As shown in Table 4.6 and Figures 4.3 and 4.4 that the distribution of minimized seed and perturbed receptors is divergent, and that the ligand orientation and contacts in the receptor may be visually different. However, it is desirable to more explicitly quantify the magnitude of these variations. The data are given in Table 4.7, which lists minimized receptor values for E_{tot} , E_{int} , the electrostatic component of the interaction energy (E^{el}_{int}), the van der Waals component of the interaction energy (E^{vdW}_{int}) and $RMSD_{Xtal}$, where the ligand is the native **E2** from crystal structure.

Table 4.7 – Property variations for the set of receptors

	E_{tot}	E_{int}	E^{el}_{int}	E^{vdW}_{int}	$RMSD_{Xtal}$
S*	-134.38	-74.81	-69.71	-5.10	1.005
S'_1*	-112.79	-76.36	-71.65	-4.71	0.861
S'_2*	-168.83	-76.31	-73.56	-2.75	0.941
S'_3*	-149.76	-75.63	-74.22	-1.41	0.986
S'_4*	-146.25	-75.38	-70.05	-5.33	0.881
S'_5*	-122.08	-75.18	-71.28	-3.90	0.874
S'_6*	-123.39	-74.98	-70.68	-4.30	0.925
S'_7*	-138.00	-74.71	-73.66	-1.05	1.110
S'_8*	-125.33	-74.58	-70.87	-3.71	0.941
S'_9*	-134.30	-74.23	-72.89	-1.34	1.055
S'_{10}*	-125.15	-73.97	-69.39	-4.58	0.930
S'_{11}*	-138.17	-73.97	-72.16	-1.81	1.020
S'_{12}*	-122.88	-73.82	-70.43	-3.39	0.932
S'_{13}*	-133.03	-73.61	-70.88	-2.73	0.932
S'_{14}*	-149.44	-72.97	-69.83	-3.14	0.944
S'_{15}*	-114.40	-72.51	-64.41	-8.10	0.897
S'_{16}*	-139.02	-71.57	-65.88	-5.69	0.930
S'_{17}*	-120.64	-70.37	-64.27	-6.10	0.864
S'_{18}*	-173.62	-68.89	-66.00	-2.89	1.044
S'_{19}*	-135.21	-68.85	-61.91	-6.94	0.958
S'_{20}*	-128.04	-67.04	-59.54	-7.50	0.857
S'_{21}*	-152.10	-71.44	-63.61	-7.83	0.882
S'_{22}*	-162.63	-69.89	-64.75	-5.14	0.996
S'_{23}*	-135.05	-68.77	-60.71	-8.06	0.904
S'_{24}*	-118.53	-73.62	-67.37	-6.25	0.890

[a]Total energy, interaction energy, electrostatic and van der Waals interaction energy components in kcal/mol, RMSD to crystal structure in Å.

Table 4.7 shows that there is a considerable variation in receptor properties for the 25 different receptors. The total energy varies from its most negative value for S'_{18}* (-173.62 kcal/mol) to its most positive value for S'_1* (-112.79 kcal/mol), thus spanning a range of over 60 kcal/mol. The seed structure, S*, has a total energy (-134.38 kcal/mol) that lies near the mean value of the distribution (-136.12 kcal/mol).

Figure 4.4 showed that receptors 1-19 all had the correct hydrogen bonding pattern, and it can be seen that Table 4.7 groups these receptors in order of increasing E_{int} , from -76.36 kcal/mol for S'_1* to -67.04 kcal/mol for S'_{19}*. The remaining receptors 20-24, which have an imperfect hydrogen bonding network, nevertheless have E_{int} values which fall within this same range. Breaking down the interaction energy into its two components (E^{el}_{int} and E^{vdW}_{int}) we can see that even though two receptors have very similar interaction energies, for example, receptors S'_1* and S'_2* (-76.36 and -76.31 kcal/mol, respectively) the breakdown between electrostatic and van der Waals components can be different by several kcal/mol. Thus even for the receptors $S'_{1\text{-}19}$* with the optimal hydrogen bonding network the electrostatic component varies by over 14 kcal/mol and the van der Waals component can vary over 7 kcal/mol. Thus the receptor structures are interacting differently with the ligand.

The RMSD between the heavy-atom crystal coordinates and the optimized receptor coordinates is near 1 Å for the set of receptors. The best agreement with the crystal coordinates occurs for S'_{20}* at 0.857 Å. Traditionally, receptor models would be energy minimized using one of the two constrained energy minimization procedures studied in chapter 3 to improve $RMSD_{Xtal}$ values. However, neither protocol was employed in this experiment because the results in chapter 3 demonstrate that these protocols do not actually achieve better $RMSD_{Xtal}$ values. It is interesting to note that receptor S'_{20}* has the best $RMSD_{Xtal}$ value of the set (0.857 Å) while its interaction energy (-67.04 kcal/mol) is the worst (least negative).

While it is clear that there is considerable variation between the receptors in the set, it is not clear where the variations in ligand-receptor interaction energy are coming from. Figure 4.5 shows the range of interaction energy values between the ligand **E2** and several key residues in the active-site cavity. The per-residue interaction energy was computed with the method described by Shadnia *et al.* (methods chapter 6. 8.).[51]

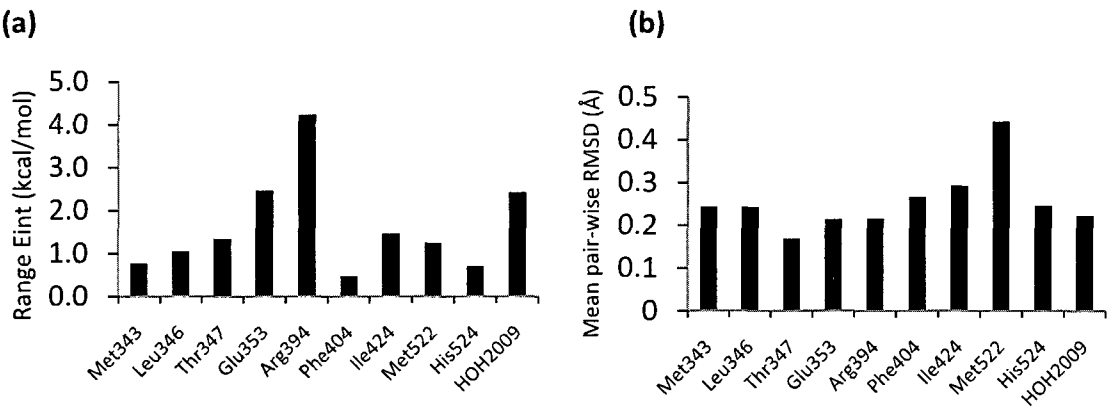


Figure 4.5 – Examining the residues responsible for variations in interaction energy
Variation in the interaction energy across 25 receptor models **(a)** the range in residue interaction energy, and **(b)** the mean pair-wise RMSD for the residues specified in (a).

Figure 4.5(a) shows that the greatest variations in interaction energy are from residues Glu353 and Arg394 as well as the water molecule HOH2009, which form the hydrogen bond network with the ligand at the A-ring. Non-polar amino acids have ranges in interaction energy that are all below 1 kcal/mol. Examining the mean pair-wise RMSD, Figure 4.5(b), for each residue shows that the residues generally deviate from each other between 0.2 and 0.3 Å. One residue, Met522, has the largest pair-wise RMSD (0.45 Å) however its range in interaction energy is small (1.2 kcal/mol) in comparison to the charged residues in the active site. The variation and range in interaction energy appears to be particularly sensitive if the interaction is electrostatic.

Now that the set of receptors for docking have been characterized the docking study can be performed. The experiment presented in the next section will demonstrate that these variations have a substantial impact on the outcome of the docking study.

4. 5. Docking a Set of Receptor Conformers

The evidence provided has shown that the seed and set of perturbed energy-minimized receptors have variations in the total energy, interaction energy, and RMSD to the crystal structure. However, it remains to be shown whether these variations are substantial enough to return a significant variation in docking results. To study this important feature of the receptors, docking was carried out using the same eight ligands presented in Table 4.2.[52,63] Each ligand was docked into all 25 of the receptor models. Briefly, 200 poses were generated for each ligand and each pose was subjected to two refinement steps: a fixed receptor minimization followed by a constrained receptor minimization. This refinement procedure is identical to the third refinement procedure tested in chapter 4.2 except that the receptor-ligand complex was not subject to a final unconstrained energy minimization. This adjustment to the refinement procedure was made to be consistent and comparable to similar docking studies by Wright *et al.* which are minimized to the same partial minima.[52] Ligand poses were then rank-ordered by the interaction energy, with the most negative interaction energy corresponding to the top pose. Ligand poses that had to deform 10 kcal/mol or greater from their global

minimum were discarded from the set. Table 4.8 shows a quantitative structure activity relationship (QSAR) between the interaction energy and the experimental log RBA for all of the docked ligands in each of the 25 receptors. The computer program implementing the docking procedures employed in this experiment can be found in the methods chapter (6. 8. Docking Simulation Procedures).

Table 4.8 – Single variable QSAR for docking studies in 25 receptors

Receptor	QSAR		Interaction Energy (kcal/mol)							
	R^2	RMSE	E2	1a	1b	1f	1h	1k	1l	2
S*	0.83	0.630	-74.82	-71.31	-71.60	-74.77	-64.30	-71.23	-70.48	-60.26
S'_1*	0.82	0.642	-76.36	-73.77	-73.35	-75.23	-57.63	-71.89	-70.88	-61.14
S'_2*	0.77	0.729	-76.31	-74.60	-75.64	-76.69	-65.86	-72.98	-71.52	-62.30
S'_3*	0.73	0.790	-75.63	-71.55	-71.91	-73.41	-65.51	-71.47	-70.26	-59.10
S'_4*	0.63	0.919	-75.38	-73.04	-72.90	-71.44	-62.42	-73.71	-70.34	-60.99
S'_5*	0.76	0.749	-75.19	-72.50	-73.10	-74.45	-64.82	-72.83	-70.17	-61.74
S'_6*	0.86	0.579	-74.98	-70.46	-71.75	-73.65	-60.41	-70.76	-69.09	-60.58
S'_7*	0.77	0.732	-74.71	-71.53	-72.00	-74.74	-62.38	-73.34	-70.16	-60.48
S'_8*	0.82	0.645	-74.58	-71.63	-69.98	-71.22	-57.35	-68.93	-67.66	-58.83
S'_9*	0.82	0.643	-74.23	-71.07	-71.65	-75.37	-61.39	-72.07	-69.07	-60.42
S'_{10}*	0.78	0.706	-73.97	-71.35	-70.54	-72.56	-63.15	-71.01	-68.98	-60.22
S'_{11}*	0.81	0.657	-73.97	-71.03	-70.60	-74.59	-63.93	-71.19	-70.37	-59.79
S'_{12}*	0.72	0.803	-73.82	-71.94	-70.83	-71.35	-60.61	-70.49	-69.15	-59.37
S'_{13}*	0.77	0.722	-73.62	-70.77	-71.44	-73.31	-59.46	-71.86	-68.32	-60.31
S'_{14}*	0.86	0.571	-72.97	-71.16	-69.99	-74.19	-63.46	-69.17	-68.56	-60.29
S'_{15}*	0.67	0.879	-72.51	-69.36	-71.24	-69.34	-58.96	-70.72	-68.94	-57.75
S'_{16}*	0.86	0.574	-71.57	-67.74	-69.08	-72.80	-60.54	-68.20	-68.52	-57.66
S'_{17}*	0.59	0.977	-70.37	-69.80	-71.11	-69.43	-61.62	-70.23	-68.73	-57.97
S'_{18}*	0.68	0.865	-68.89	-66.28	-66.70	-68.75	-59.77	-67.94	-66.83	-54.11
S'_{19}*	0.56	1.010	-68.85	-66.94	-69.17	-69.32	-61.99	-68.82	-65.62	-54.66
S'_{20}*	0.44	1.137	-67.04	-66.86	-68.80	-66.27	-61.42	-68.08	-65.52	-55.85
S'_{21}*	0.65	0.898	-71.44	-69.09	-71.16	-69.89	-62.67	-69.31	-68.71	-57.82
S'_{22}*	0.54	1.036	-69.89	-67.17	-68.97	-69.34	-61.51	-71.58	-67.15	-56.62
S'_{23}*	0.71	0.815	-68.77	-65.93	-68.28	-68.33	-58.95	-67.42	-67.07	-56.22
S'_{24}*	0.81	0.665	-73.63	-70.70	-71.09	-71.96	-60.94	-70.12	-69.59	-59.68

As shown in table 4.8, the various receptor models return a range of correlation coefficients between E_{int} and log RBA ($0.44 \leq R^2 \leq 0.86$) for the set of eight ligands. Three perturbed receptors (S'_6*, S'_{14}*, S'_{16}*) return the best QSAR correlation coefficient ($R^2 = 0.86$). For these correlations the rms error in prediction of RBA is almost down to the level of 0.5 log unit. However, the variation in E_{int} between these three receptors for the set of ligands is considerable, even though the correlation coefficient is the same. Thus, the native ligand **E2** has an E_{int} of -74.98, -72.97 and -71.57 kcal/mol in each of the three receptors (S'_6*, S'_{14}*, S'_{16}*) respectively. A comparison of the E_{int} values for the docking of the remaining set of ligands in these three receptors also shows significant variations. For example, the E_{int} values for ligand **1h** are similar in receptors S'_6* (-60.41 kcal/mol) and S'_{16}* (-60.54 kcal/mol) but different for receptor S'_{14}* (-63.46). For another example, ligand **1a** has E_{int} values of -70.46, -71.16, and -67.74 kcal/mol in the three receptors (S'_6*, S'_{14}*, S'_{16}*) respectively.

It is important to note that the variation in E_{int} values obtained for the set of ligands does not simply scale according to their native **E2** E_{int} values. That is, if ligand **E2** has E_{int} values that are different by 2 kcal/mol for two receptors, this does not mean that a newly docked ligand **L**, will have the same 2 kcal/mol difference. In fact, the ranking by E_{int} of the docked poses is different for each of the three receptors. Receptor S'_6* ranks the set of ligands as {**E2**, **1f**, **1b**, **1k**, **1a**, **1l**, **2**, **1h**} from best to worst E_{int} while receptors S'_{14}* and S'_{16}* rank the set of ligands as {**1f**, **E2**, **1a**, **1b**, **1k**, **1l**, **1h**, **2**}, and {**1f**, **E2**, **1b**, **1l**, **1k**, **1a**, **1h**, **2**} respectively. In fact, the correct ranking of ligands by

experimental RBA values {**E2**, **1f**, **1l**, **1a**, **1b**, **1k**, **2**, **1h**}, was not obtained by any of the three receptors having the best QSAR correlation coefficients.

The seed structure, S*, had the fourth best QSAR correlation coefficient (R^2 = 0.83) and lies one sigma above the mean QSAR correlation coefficient ($R^2 = 0.73 \pm 0.11$) for the set of receptors. The worst QSAR correlation coefficient belonged to S'_{20}* at R^2 = 0.44. S'_{20}* is pictured in figure 2d having the native **E2** ligand lacking the H-bond with Arg394. This variation in docking results demonstrate that receptor model bias, as described by Jain *et. al.*,[66] is not just limited to different receptor crystal structures. In fact, receptor model bias is present even when docking into the same receptor model prepared having near identical receptor coordinates.

A second and more sophisticated QSAR model was constructed using the same docking output poses obtained by using E_{int}. This second QSAR model involved a multiple linear regression of three energy terms, the interaction energy between the ligand and receptor (E_{int}), the deformation energy of the ligand (ΔE_L), and the deformation energy of the receptor (ΔE_R) to predict log RBA.[52] Results for this second QSAR model are given in Table 4.9.

Use of a multiple linear regression model significantly improved the correlation results. Now the five receptor models having the best QSAR correlation coefficient were 6, 1, 11, 15, and 21, with very high correlation coefficients of (0.98, 0.96, 0.96, 0.96 0.96), respectively. The rms error for these correlations was now reduced to ca. 0.3 log units. The multiple regression coefficients were very similar for the three energy

descriptors for these five best receptor models, giving -0.304, -0.249, -0.301, -0.284, -0.311, respectively, for the same five receptor models. The same is true for the relative importance for the set of descriptor coefficients. The seed structure S*, with R^2 = 0.95, had the third best correlation coefficient from the set of receptors. The worst QSAR correlation coefficient in the set was given by S'_{19}* (R^2 = 0.69).

Table 4.9 – Multiple variable QSAR for docking studies in 25 receptors

Receptor	QSAR		Descriptor Coefficient			Descriptor Relative Importance		
	R^2	RMSE	E_{int}	ΔE_R	ΔE_L	E_{int}	ΔE_R	ΔE_L
S*	0.95	0.345	-0.283	0.083	-0.259	1.000	0.333	0.428
S'_1*	0.96	0.309	-0.249	0.156	-0.345	1.000	0.482	0.417
S'_2*	0.83	0.619	-0.311	0.138	-0.224	1.000	0.237	0.330
S'_3*	0.92	0.420	-0.268	0.026	-0.295	1.000	0.051	0.528
S'_4*	0.76	0.748	-0.301	0.023	-0.309	1.000	0.117	0.404
S'_5*	0.91	0.466	-0.318	0.142	-0.372	1.000	0.250	0.495
S'_6*	0.98	0.193	-0.304	0.146	-0.225	1.000	0.318	0.355
S'_7*	0.89	0.504	-0.293	-0.016	-0.252	1.000	0.042	0.363
S'_8*	0.95	0.356	-0.271	0.209	-0.342	1.000	0.366	0.444
S'_9*	0.94	0.370	-0.298	0.028	-0.281	1.000	0.073	0.365
S'_{10}*	0.94	0.376	-0.340	-0.005	-0.343	1.000	0.020	0.409
S'_{11}*	0.96	0.300	-0.301	0.100	-0.258	1.000	0.306	0.462
S'_{12}*	0.88	0.537	-0.264	0.179	-0.451	1.000	0.368	0.590
S'_{13}*	0.81	0.660	-0.327	0.093	-0.100	1.000	0.238	0.132
S'_{14}*	0.90	0.475	-0.350	0.008	-0.149	1.000	0.034	0.236
S'_{15}*	0.96	0.301	-0.284	0.195	-0.388	1.000	0.640	0.519
S'_{16}*	0.87	0.551	-0.266	0.029	-0.109	1.000	0.047	0.161
S'_{17}*	0.94	0.358	-0.330	0.234	-0.431	1.000	0.408	0.618
S'_{18}*	0.82	0.648	-0.259	-0.029	-0.246	1.000	0.084	0.410
S'_{19}*	0.69	0.841	-0.283	0.105	-0.283	1.000	0.332	0.521
S'_{20}*	0.92	0.433	-0.422	0.301	-0.618	1.000	0.585	0.720
S'_{21}*	0.96	0.286	-0.311	0.164	-0.457	1.000	0.352	0.619
S'_{22}*	0.91	0.468	-0.359	0.348	-0.463	1.000	0.458	0.667
S'_{23}*	0.90	0.473	-0.230	0.086	-0.350	1.000	0.237	0.650
S'_{24}*	0.86	0.575	-0.320	0.174	-0.116	1.000	0.251	0.137

Many of the 3-term correlations for the receptor models now have excellent correlation coefficients greater than 0.9. The ranking of predicted log RBA in most of these cases is identical or almost identical to the correct experimental order as measured by binding affinity. The coefficients given by multiple linear regression have relatively small variations between the most highly predictive models. However, it is also important to consider the orientation and interaction that each docked pose has with its receptor.

Table 4.10 describes the binding orientation and H-bond network for the set of eight ligands in all twenty-five receptors. The color scale indicates the electrostatic interaction energy (E^{el}_{int}) between the docked pose in its particular receptor (specifically residues Glu353, Arg394 and His524, along with HOH2009). This is directly relevant to the H-bond strength of the network. This property was computed using the method described by Shadnia *et al.* (methods chapter 6. 8.).[51] As an example, ligand **E2** docked into perturbed receptor 1, i.e. S'^{1}(**E2**), has an E^{el}_{int} value of light green corresponding to the range $-55 < E^{el}_{int} < -51$ kcal/mol. (2) The binding orientation is indicated by the number in each box, which corresponds to the ligand RMSD relative to the native **E2** found in the prepared and minimized receptor. For example, ligand **1a** docked into the seed structure, S*(**1a**), has its positional coordinates compared to the **E2** ligand found in the row containing S* (0.24 Å). In this manner, the orientation of a ligand **L** with respect to the crystallographic ligand **E2** can be compared. The RMSD value describing ligand orientation was calculated from three "landmark" atoms common to all docked ligands. The three atoms included: the oxygen on carbon-3 of the A-ring, carbon-9 of the B-ring

which is links to the A-ring and carbon-18 corresponding to the methyl group on the D-ring. The RMSD calculation is described in detail in the methods chapter 6. 8.

Table 4.10 – Ligand orientation and H-bond strength for dockings in the receptor set

Receptor	Ligand (L) E2	1a	1b	1f	1h	1k	1l	2
S*	0.00	0.24	0.23	4.30	1.24	1.27	0.23	0.36
S'1*	0.00	0.23	0.18	1.17	4.31	1.16	0.39	0.33
S'2*	0.00	0.30	3.27	0.53	1.29	0.23	0.30	0.45
S'3*	0.00	3.66	3.53	1.13	1.28	1.32	1.17	0.49
S'4*	0.00	0.15	0.21	1.18	4.53	0.16	0.18	0.84
S'5*	0.00	1.17	0.19	0.88	1.28	1.17	0.23	0.87
S'6*	0.00	1.18	1.17	1.18	0.50	1.33	1.17	0.98
S'7*	0.00	0.24	1.15	1.14	4.38	1.30	1.30	0.87
S'8*	0.00	0.10	0.24	1.16	0.48	1.17	1.18	1.21
S'9*	0.00	1.15	1.15	1.12	4.40	1.36	1.12	0.55
S'10*	0.00	0.14	0.71	0.89	4.44	1.25	0.75	1.08
S'11*	1.33	1.09	1.09	4.48	1.20	1.24	1.09	0.94
S'12*	0.00	0.14	0.24	1.13	0.54	1.14	1.17	0.26
S'13*	3.33	0.21	0.19	0.53	4.45	0.21	0.28	1.24
S'14*	0.00	0.19	1.19	4.33	4.28	1.19	1.34	0.97
S'15*	3.33	3.23	0.23	4.25	1.36	3.22	3.25	3.35
S'16*	0.00	1.12	1.13	0.68	0.54	1.14	1.11	0.00
S'17*	0.00	0.19	0.16	0.84	0.41	0.18	0.17	1.07
S'18*	0.00	1.12	1.12	1.31	4.28	1.14	1.11	0.00
S'19*	0.00	0.31	0.33	4.31	1.21	0.35	1.27	0.35
S'20*	3.33	3.25	0.22	4.21	1.20	3.22	3.32	1.10
S'21*	0.00	0.20	0.16	0.77	0.95	3.17	3.24	0.40
S'22*	0.00	0.24	0.22	0.43	0.85	1.25	1.13	1.09
S'23*	3.33	3.25	0.20	3.80	3.94	3.27	3.45	1.38
S'24*	0.00	0.19	0.16	0.40	0.95	3.17	3.30	1.59

[a]Electrostatic Interaction Energy Legend

E^{el}_{int} > -51 | -51 to -55 | -55 to -59 | -59 to -63 | -63 to -67 | E^{el}_{int} < -67

Table 4.10 lists RMSD values of 0.00 Å for the column labelled E2, showing that dockings of **E2** starting from the perturbed receptors are all able to identically reproduce the bound pose **E2** found in their respective structure. For example, an RMSD of 0.00 Å for **E2** (newly) docked into the seed structure, S*(**E2**), has the same orientation and placement as the native **E2** in the prepared and minimized receptor seed model, S*. The same is true for **E2** (newly) docked into the perturbed structure, S'_1*(**E2**), having the same orientation as the native **E2** in the prepared and minimized receptor perturbed model, S'_1*. The exact orientation and placement of the native **E2** is different depending on the minimized receptor model. The range in colors, representing the electrostatic interaction energy between docked ligand and H-bond residues, vary for each docking of **E2**. These variations, from orange to dark green, reflect the variation in receptor models (as characterized in Table 5.3 and shown graphically in Figure 5.2). Therefore, it is possible that a docking in two receptor models, i.e. S'_1* and S'_1*, can reproduce placement of **E2** for each receptor (RMSD = 0.00 Å), but still have different E^{el}_{int} values.

Next, consider the binding orientations which belong to receptor models that return good QSAR correlations ($R^2 > 0.90$). If we consider a larger ligand, e.g. **1f**, the variation in binding poses is much greater compared to the results for a smaller ligand like **1a**. S*(**1f**) suggests that the binding mode achieves poor electrostatic interactions with the hydrogen bond network residues (light red: -51 to -55 kcal/mol) and that the ligand binds in a completely different orientation (flipped) than the native **E2** ligand (RMSD = 4.30 Å). Another receptor model, S'_1*(**1f**), has a binding orientation in which

the 18-methyl group is rotated into the back of the cavity and a slightly improved electrostatic interaction energy (light yellow: -59 to -63 kcal/mol). Two additional receptor models, S'_{10}*(**1f**) and S'_{22}*(**1f**), have binding orientations with better electrostatic interaction energy (light green: -63 to -67 kcal/mol) and ligand orientations which better resemble the binding mode of **E2** at 0.83 Å and 0.48 Å respectively.

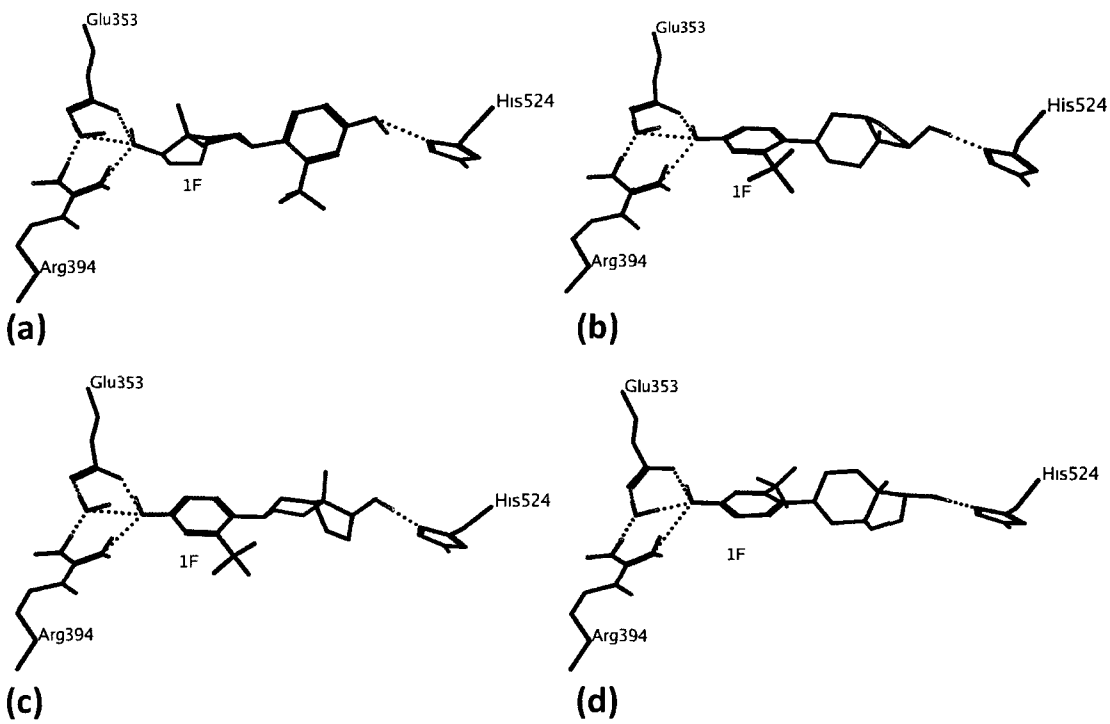


Figure 4.5 – Binding orientations of ligand 1f docked in four predictive receptors
Binding pose of ligand **1f** in the **(a)** seed receptor (S*), and the perturbed receptors **(b)** S'_{10}*,**(c)** S'_{10}*, and **(d)** S'_{22}*. Non-polar hydrogens omitted and H-bonds (black dots) drawn for clarity.

As shown in Figure 4.5(a), the binding orientation for S*(**1f**) corresponds to a flipped pose in the active site (RMSD: 4.30 Å). The binding orientation of S'_1(**1f**), shown in Figure 4.5(b), is rotated in the active site (RMSD: 1.17 Å). The poses for 10 (**1f**) and 22 (**1f**) have similar E^{el}_{int} values (both are light green) however their binding orientations, as

shown in Figure 4.5(c) and (d) respectively, have the trifluoro group from the A-ring on opposite faces of the active site. Careful study of Table 4.10 along with the consideration of the examples shown in Figure 4.5 demonstrate that the small variations in receptor models can have a substantial impact on the orientation of the docked poses in the active site.

While many of the receptor models gave excellent correlations ($R^2 > 0.90$), the variation in binding modes can involve flipping or rotation of the ligand depending on which receptor model is chosen. These results highlight the deficiencies of traditional docking methodologies that only consider a single geometry for the target receptor. Thus, the different binding orientations should all be considered in the design of new ligands, because they all originate from receptor models having very similar predictive capabilities.

4. 6. Invalidating $RMSD_{Xtal}$ and E_{tot} as Qualifiers of a Valid Docking Receptor

The total energy (E_{tot}) and the RMSD to crystal structure ($RMSD_{Xtal}$) are two receptor properties often thought to be representative of model quality.[3c,49] Had this docking study selected the receptor model with the lowest deviation from crystal structure (receptor properties shown in Table 4.7) then the perturbed receptor S'_{20}* would have been used. This receptor had the worst single variable QSAR with a correlation coefficient of 0.44. Had the lowest energy receptor model been choosen instead, S'_{18}*, then the docking study would also have had a poor single variable QSAR

with a correlation coefficient of 0.68. This observation is contrary to the common practice where preference is given to the choice of receptor model which has either the lowest total energy or the least coordinate deviation from the crystal structure.[3c] Given that we have demonstrated receptor model bias exists even for near identical input receptors, it is perhaps not surprising that a single property that globally describes each receptor (e.g. E_{tot} or $RMSD_{Xtal}$) is not an effective indicator of a predictive docking model.

Chapter 5

CONCLUSIONS

The experiments presented in this thesis explore large molecule minimization and its sensitivity to input perturbations. The phenomena, as it relates to the complexity of the potential surface, was originally investigated by Williams and Feher in 2008.[41] The second chapter of this thesis characterizes how input gradient and related factors influence the spread in perturbed energy minimizations.

The magnitude of perturbation was varied and it was found that different magnitudes of displacement did not influence the divergence of minimized structures. The bond compression experiment conducted with 1UBQ established a "proof of principle" that input gradient is a factor that can influence the spread in output properties. For this simple but hypothetical example it was found that larger values of input RMS gradient return greater ranges of output values. A more practical experiment, beginning from crystal structure positioning of the heavy atoms, involved the choice of hydrogen addition methodology which can also affect the initial gradient. However, using the method with least gradient (Protonate 3D) did not lead to any significant improvement in output spreads relative to hydrogen atom addition using default placements. Monitoring the number of iterations did not clarify this result, supposedly the lack of effect can be attributed to the fact that input perturbation sensitivity arising from molecular complexity was a stronger determining factor than that arising from initial gradient. Finally, input perturbed minimizations were performed using different optimization algorithms. Of the three algorithms tested, which included steepest descent, conjugate gradient and truncated Newton, it was the [CG, TN]

combination algorithm which returned the smallest spreads in output structures with the fewest number of iterations.

The third chapter investigates the use of constraints applied to energy minimization and the resulting distribution of local minima upon perturbation of input structures. The first experiment explored previously published approaches to constrained energy minimization typically employed in the preparation of the structures for molecular simulation. The objective of these algorithms is to give an energy minimized structure which has either a lower potential energy or smaller deviation from the crystal structure compared to the same input structure subjected to an unconstrained energy minimization. The experiments demonstrated that neither of the two approaches studied satisfied either objective and that the spread of minimized structures was either as divergent or more divergent compared to unconstrained energy minimization methods. The second experiment in this chapter involved the application of new tether schemes applied in an unbiased and systematic fashion to reduce the sensitivity of large structure minimization to input perturbation. Each tether scheme was applied with the intention to reduce accessibility of additional local minima on the potential energy surface. It was found that tether schemes having very gradual reduction of tether weights were successful at reducing output spreads of perturbed and minimized structures to near zero.

The fourth chapter examined the effect of minimization strategies on the refinement process in docking simulation as well as the effect of receptor model

geometry on the outcome of a docking experiment. The first experiment involved the testing of three refinement strategies which were found to significantly influence the results and variation of poses in docking studies. It was found that refinement method three (fixed receptor minimization, flexible receptor constrained energy minimization, and unconstrained energy minimization) returned the smallest spreads in descriptor properties. It was also found that refinement method three was the only method that could reproduce the binding orientation of the native ligand, **E2**, in estrogen receptor-alpha (**ERα**) docking studies. The last experiment presented in this thesis is found in chapter 4. The objective of the experiment was to investigate the effect of receptor model bias on docking studies. The experiment made use of minimization and input perturbation sensitivity to create the multiple receptor conformations used in docking studies. Results of the experiment demonstrated that a range of correlations could be obtained and that there was no consensus in binding orientation among the most predictive receptors. These results highlight deficiencies in traditional docking methodologies which rely on potential energy based descriptors and single receptor docking targets. The experiment also demonstrated, contrary to current scientific opinion, that the most predictive receptors were not the receptor models which had the lowest potential energy or least deviation from crystal structure.

The work in this thesis characterizes how input perturbation sensitivity influences the results of energy minimization. It shows that the complex potential energy landscape of proteins is an important factor often left unconsidered in molecular modeling studies. More importantly, the experiments presented in this thesis show that

the phenomena is more than just a nuisance that introduces substantial uncertainty in results. Instead, initial condition sensitivity is a useful and constructive tool that allows one to produce and explore multiple valid protein geometries.

Chapter 6

METHODS

6. 1. Hardware and Software Specifications

All of the calculations conducted and presented in this thesis were performed on four Intel Pentium 4, 3.20 GHz single-core CPU computers with the hyperthreading option disabled (with the exception of the experiment in 2.2 which was completed using an Intel Celeron 2.6 GHz processor). This precaution was taken because some energy minimization software packages are known to introduce irreproducibility in multi-core/hyperthreading applications. The specification of single-core CPU with hyper-threading disabled is important. There appears to be a race condition present in MOE2009's parallel implementation of minimization where identical inputs produce variable output. The same test run on a single core non-threaded processor is completely deterministic returning no variation in output. This feature has been characterized as a race condition based on the description provided by Netzer and Miller.[70] Each computer ran Microsoft Windows XP professional version 2002, service pack 3. The experiments were all conducted using the Molecular Operating Environment (MOE) software, version 2009.[37]

Setup and modification of the force field and the potential energy model in MOE can be accomplished by accessing the potential setup window: MOE | Window | Potential Setup > Prompts Potential Setup Window (PSW), where the choice of force field for experimentation can be found in the PSW window: PSW | Load > List of Force Fields (i.e. MMFF94s, AMBER99, CHARM27, etc.), and the choice of solvent model: PSW | Solvation: > List of Solvent Models (i.e. Distance, Gas Phase, Born, etc.), can all be

accessed. It is important that changes to the solvation model can only take affect if Apply is clicked in the PSW. An alternate approach to modifying the potential energy model involves the use of scientific vector language (SVL). SVL commands can be input to MOE using the command line interface accessed by: MOE | SVL > Prompts the Command Line Interface (SVL>),

and the following commands:

SVL> pot_Load '$MOE/lib/mmff94s.ff' // example for loading MMFF94s force field

SVL > pot_Setup [] // example of function which will return list of potential parameters

allow for the modification of the force field and potential energy model respectively.

6. 2. An Example Energy Minimization Problem

The example energy minimization problem considered involves 1,2-dibromoethane. To construct 1,2-dibromoethane the molecule builder must first be accessed using the following command: MOE | Edit | Build | Molecule > Prompts Molecule Build Window (MBW).

To construct 1,2-dibromoethane, begin by creating a carbon atom by clicking on the carbon button (C) in the MBW. Notice that the carbon atom is created with all valences occupied with hydrogen atoms (methane molecule) and that one hydrogen is selected (pink). Next, in the MBW, with a hydrogen selected, click on the carbon again

to create the ethane molecule. Next, select two hydrogens (each belonging to opposite carbons) and click on the bromine button in the MBW creating the 1,2-dibromoethane molecule.

Once the molecule is constructed we need to alter the dihedral angle to conform to the input provided for the example having a [Br-C-C-Br] dihedral angle of 60°. To alter the dihedral angle, select the bromine atoms and the carbon atoms. In the MBW, the dihedral angle can now be set by typing 60 into the field marked 'New Dihedral Angle (Deg):', and clicking apply.

Now that the molecule is constructed with the correct input dihedral angle we can procedure with an energy minimization. First, partial charges must be computed for the system: MOE | Compute | Partial Charges > Prompts Partial Charge Window (PCW) | Apply | OK.

Make sure that the correct force field and solvent model are loaded. Next, the energy minimization can be conducted from: MOE | Compute | Energy Minimize > Prompts the Energy Minimize Window (EMW).

Set the gradient to 0.001 and click okay. The molecule should now be minimized with a print out of the iteration, potential energy and gradient in the SVL command window. To disable or enable particular minimization algorithms during minimization, one must use the SVL command window when performing energy minimization. An example command is provided:

MM [sd_maxit:100, cg_maxit:100, maxit:100, sd_gtest:1, cg_gtest:0.1, gtest:0.001]

where, the max number of iterations for steepest descent, conjugate gradient and truncated Newton algorithms is given to sd_maxit, cg_maxit and maxit respectively, and the termination gradient for each of the tree algorithms is indicated in sd_gtest, cg_gtest and gtest respectively.

Because a substantial number of energy minimizations must be conducted on several molecules, the following SVL script acts on a database of structures to energy minimize and report on parameters of importance to this thesis.

```
#svl
#set main 'Minimizinator'

// ============================================================================================
// Database Energy Minimization Function                                                  ====
// ============================================================================================

// Created by James Davey
// Wright Lab (SC512) Carleton University
// An SVL Script for MOE created July 28/2011

// Function Description  ======================================================================
// The minimizinator is a SVL script designed to energy minimize entries in a database belonging to a specified  //
// molecule field  The function allows the user to perform an energy minimization using three algorithms  steepest //
// descent (SD), conjugate gradient (CG), and truncated Newton (TN)  The function computes system properties for //
// the final energy minimized structures (Etot, Eint, RMSG, RMSDxtal and RMSDpw)                                  //
// ============================================================================================

// User Defined Parameters  ===================================================================

  const Path            = 'c /users/smp-user/desktop/SVL SCRIPTS THESIS/',                    //(a)
  const Database        = 'Perturbed 1GWR-TM mdb',                                            //(b)
  const LigandChain     = 2,                                                                  //(c)
  const ForceField      = '$MOE/lib/mmff94s ff',                                              //(d)
  const PotentialOpt    = [],                                                                 //(e)
  const InputField      = 'Pert Structure',                                                   //(f)
  const MM_Opt          = [  sd_maxit 10000,
                             sd_gtest 1 0,
                             cg_maxit 10000,
                             cg_gtest 0 1,
                             maxit 10000,
                             gtest 0 001],                                                   //(g)

// The minimizinator requires that (a) the directory and path of the project be specified in Path as a token  The (b)  //
```

```
// database base containing sturctures to be minimized be provided as a  mbd file in Database as a token, and (c)    //
// the ligand chain be specified in LigandChain as an integer  The (d) force field file is specified in ForceField as a    //
// token indicating the location of the force field file  Modification of (e) the potential setup parameters (i e          //
// distance, gas phase, etc ) can be specified in PotenialOpt as a vector  The (f) input field in the database            //
// containing the structures to be minimized is entered into InputField as a token and (g) the energy minimization       //
// options for MOE's function MM is provided in the vector MM_Opt                                                        //
// ======================================================================================================

// Function Declaration    ==========================================================================

  function PartialCharge,
  function Potential,
  function pro_Superpose,
  function ComputeEint,
  function ComputeRMSG,
  function MM,


// Minimizinator Function    ==========================================================================

function Minimizinator [],

pot_Load ForceField,
pot_Setup PotentialOpt,

local i, j, LigandAtoms, CurrentEntry, dbkey, EntryKey, EntryNo,
local mol, Etot, Eint, RMSG, RMSDxtal, RMSDpw,

dbkey = db_Open [tok_cat [Path, Database], 'read-write'],
EntryKey = db_Entries dbkey,
EntryNo = db_nEntries dbkey,

db_EnsureField [dbkey, tok_cat ['MM ', InputField], 'molecule'],
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' Etot'], 'float'],
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' Eint'], 'float'],
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' RMSG'], 'float'],
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' RMSDxtal'], 'float'],

 for i = 2, EntryNo loop,
 db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' RMSDpw ', totok i], 'float'],
 endloop,

 for i = 1, EntryNo loop,
 mol_Create cat db_ReadFields [dbkey, EntryKey(i), InputField],
 LigandAtoms = cat cAtoms ((Chains [])(LigandChain)),
 aSetCharge [Atoms [], first PartialCharge [Atoms [], 'FF']],
 MM MM_Opt,

 Etot = first Potential [dX 0],
 Eint = ComputeEint LigandAtoms,
 RMSG = ComputeRMSG [],
 mol = mol_Extract Atoms [],
 mol_Create cat db_ReadFields [dbkey, EntryKey(i), 'Seed Structure'],
 RMSDxtal = pro_Superpose [Chains [], ['atom_set' 'All']],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField]],[mol]]],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' Etot']],[Etot]]],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' Eint']],[Eint]]],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' RMSG']],[RMSG]]],
```

```
db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' RMSDxtal']],[RMSDxtal]]],
Close [force 1],
endloop,


 for i = 1, (EntryNo - 1) loop,
 for j = (i + 1), EntryNo loop,
 mol_Create cat db_ReadFields [dbkey, EntryKey(i),
                                    tok_cat ['MM ', InputField]],
 mol_Create cat db_ReadFields [dbkey, EntryKey(j),
                                    tok_cat ['MM ', InputField]],
 RMSDpw = pro_Superpose [Chains [], ['atom_set' 'All']],
 Close [force 1],
 db_Write [dbkey, EntryKey(i),
        tag [[tok_cat ['MM ', InputField, ' RMSDpw ', totok j]], [RMSDpw]]],
 endloop,
 endloop,


endfunction,

// ComputeEint  ================================================================================

function ComputeEint x,

local Eint = third pot_eleEnergy x + third pot_vdwEnergy x + third pot_solEnergy x,
return Eint,

endfunction,

// ComputeRMSG  ================================================================================

function ComputeRMSG [],

local Jtot = tr second Potential [],
local RMSGtot = sqrt ((add rot3d_vDot [Jtot, Jtot]) / (length Jtot)),
return RMSGtot,

endfunction,
```

Evaluation of the potential energy of a molecule can be done in the MOE program using: MOE | Compute | Potential Energy, or in the SVL command line with: SVL> first Potential [dX:0]. Superposition and computation of the RMSD between two protein structures can be accomplished using MOE's pro_Superpose function found in the sequence editor: MOE | SEQ > Prompts the Sequence Editor (SE). SE | Homology |

Superpose > Prompts the Superpose Function Window. In the case of all experiments presented in this thesis, all heavy atoms were included in protein RMSD calculations.

6. 3. Perturbation of Molecular Coordinates

Perturbed structures were generated by applying a displacement of ± 0.001 Å (or otherwise specified magnitude) to each coordinate of each atom according to a random number (R) where {0 < R< 1}; for R < 0.5 the displacement is negative and otherwise positive. If the displacement magnitude was ± 0.001 Å then all perturbed structures have an atomic RMSD with respect to the seed structure of 0.0017 Å, as expected from the definition RMSD = SQRT[$\sum[\Delta x_i^2 + \Delta y_i^2 + \Delta z_i^2]/N$], where N = number of atoms. The introduced perturbations were well below the level of the crystal atomic resolution.

Because a large number of structures must be perturbed prior to energy minimization, an SVL script is provided below to perform the perturbation and compute system properties (i.e. total energy, RMS gradient, etc.). The output for a file is saved to a database.

```
#svl
#set main 'Perturbinator'

//=============================================================================================
// Atomic Coordinate Perturbation Function                                              ======
//=============================================================================================

// Created by James Davey
// Wright Lab (SC512) Carleton University
// An SVL Script for MOE created July 28/2011

// Function Description ========================================================================
// The perturbinator is an SVL script designed to perturb the atomic coordinates of an input structure by a      //
// specified displacement value. The molecular properties (Etot, Eint, RMSDxtal, RMSDpw, RMSGtot,                 //
// RMSGhydrogens, RMSGheavyatoms, and atomic coordinates) are all calculated for the seed (input) structure and //
// perturbed structure sets. The input (seed) structure is renamed in the function as 'S0 w/Output' and each     //
// subsequent perturbed structure is named S1, S2, S3,   with the Output name proceeding The atomic coordinates//
// of each perturbed structure are randomly perturbed in all coordinate directions by a specified displacement (+/-) //
```

```
// according to a random number (R) where {0 < R < 1}, for R > 0 5 the displacement is positive and otherwise      //
// negative  User defined parameters are set below                                                                 //
//=================================================================================================================

// User Defined Parameters ========================================================================

  const Path              = 'c /users/smp-user/desktop/SVL SCRIPTS THESIS/',                          //(a)
  const Input             = '1GWR-TM moe',                                                           //(b)
  const Output            = 'Perturbed 1GWR-TM mdb',                                                 //(c)
  const CreateFiles       = 0,                                                                       //(d)
  const Displacement      = 0 001,                                                                   //(e)
  const StructureNo       = 5,                                                                       //(f)
  const LigandChain       = 2,                                                                       //(g)
  const ForceField        = '$MOE/lib/mmff94s ff',                                                   //(h)
  const PotentialOpt      = [],                                                                      //(i)

// The perturbinator requires that (a) the directory and path of the project be specified in Path as a token  The (b)  //
// prepared input (seed) structure must also be supplied as a  moe file in Input as a token, and (c) the output       //
// structures are a database which must be named in Output as a token  mdb format  If  moe files are also required  //
// then (d) the CreateFiles const variable can be set to 1 (otherwise 0)  The (e) displacement magnitude must be      //
// provided in Displacement as a float, and the (f) number of perturbed structures to generate must be indicated in  //
// StructureNo as an integer  The (g) ligand chain index number must be provided in LingandChain as an integer       //
// The (h) force field file is specified in ForceField as a token indicating the location of the force field file    //
// Modification of (i) the potential setup parameters (i e  distance, gas phase,etc ) can be specified in PotenialOpt  //
// as a vector                                                                                                      //
//==================================================================================================================

// Function Declaration ===========================================================================

  function PartialCharge,
  function Potential,
  function pro_Superpose,
  function ComputeEint,
  function ComputeRMSG,
  function PerturbCoordinates,

// Perturbinator Function ========================================================================

function Perturbinator [],

pot_Load ForceField,
pot_Setup PotentialOpt,

local i, j, LigandAtoms, CurrentEntry, dbkey,
local SeedMol, SeedRMSDpw, SeedEtot, SeedEint, SeedRMSGtot, SeedRMSGheavy, SeedRMSGhydrogen,
local PertMol, PertRMSDxtal, PertRMSDpw, PertEtot, PertEint, PertRMSGtot, PertRMSGheavy,
local PertRMSGhydrogen,

dbkey = db_Open [tok_cat [Path, Output], 'create'],

db_CreateField [dbkey, 'Seed Structure', 'molecule'],
db_CreateField [dbkey, 'Seed Etot', 'float'],
db_CreateField [dbkey, 'Seed Eint', 'float'],
db_CreateField [dbkey, 'Seed RMSGtot', 'float'],
db_CreateField [dbkey, 'Seed RMSGheavy', 'float'],
db_CreateField [dbkey, 'Seed RMSGhydrogen', 'float'],
db_CreateField [dbkey, 'Seed RMSDxtal', 'float'],
```

```
 for i = 2, StructureNo loop,
 db_CreateField [dbkey, tok_cat ['Seed RMSD ', totok i], 'float'],
 endloop,

db_CreateField [dbkey, 'Pert Structure', 'molecule'],
db_CreateField [dbkey, 'Pert Etot', 'float'],
db_CreateField [dbkey, 'Pert Eint', 'float'],
db_CreateField [dbkey, 'Pert RMSGtot', 'float'],
db_CreateField [dbkey, 'Pert RMSGheavy', 'float'],
db_CreateField [dbkey, 'Pert RMSGhydrogen', 'float ],
db_CreateField [dbkey, 'Pert RMSDxtal', 'float'],

 for i = 2, StructureNo loop,
 db_CreateField [dbkey, tok_cat ['Pert RMSD ', totok i], 'float'],
 endloop,

 for i = 1, StructureNo loop,
 ReadAuto tok_cat [Path, Input],
 LigandAtoms = cat cAtoms ((Chains [])(LigandChain)),
 aSetCharge [Atoms [], first PartialCharge [Atoms [], 'FF']],

 cSetName [Chains [], tok_cat ['S0 ', tok_drop [Input, -4]]],
 cSetTag [Chains [], tok_cat ['S0 ', tok_drop [Input, -4]]],
 SeedEtot = first Potential [dX 0],
 SeedEint = ComputeEint LigandAtoms,
 [SeedRMSGtot, SeedRMSGheavy, SeedRMSGhydrogen] = ComputeRMSG [],
 SeedMol = mol_Extract Atoms [],
 CurrentEntry = db_Write [dbkey, 0, [   'Seed Structure'      SeedMol,
                                        'Seed Etot'           SeedEtot,
                                        'Seed Eint'           SeedEint,
                                        'Seed RMSGtot'        SeedRMSGtot,
                                        'Seed RMSGheavy'   SeedRMSGheavy,
                                        'Seed RMSGhydrogen'   SeedRMSGhydrogen,
                                        'Seed RMSDxtal'       0]],

 PerturbCoordinates Displacement,
 cSetName [Chains [], tok_cat ['S', totok i, ' ', tok_drop [Input, -4]]],
 cSetTag [Chains [], tok_cat ['S', totok i, ' ', tok_drop [Input, -4]]],
 PertEtot = first Potential [dX 0],
 PertEint = ComputeEint LigandAtoms,
 [PertRMSGtot, PertRMSGheavy, PertRMSGhydrogen] = ComputeRMSG [],
 PertMol = mol_Extract Atoms [],
 mol_Create cat db_ReadFields [dbkey, CurrentEntry, 'Seed Structure'],
 PertRMSDxtal = pro_Superpose [Chains [], ['atom_set' 'All']],
 db_Write [dbkey, CurrentEntry, [   'Pert Structure'      PertMol,
                                    'Pert Etot'           PertEtot,
                                    'Pert Eint'           PertEint,
                                    'Pert RMSGtot'        PertRMSGtot,
                                    'Pert RMSGheavy'      PertRMSGheavy,
                                    'Pert RMSGhydrogen'   PertRMSGhydrogen,
                                    'Pert RMSDxtal'       PertRMSDxtal]],
 Close [force 1],
 endloop,

local EntryKey = db_Entries dbkey,
local EntryNo = db_nEntries dbkey,

  for i = 1, (EntryNo - 1) loop,
```

```
  for j = (i + 1), EntryNo loop,
  mol_Create cat db_ReadFields [dbkey, EntryKey(i), 'Seed Structure'],
  mol_Create cat db_ReadFields [dbkey, EntryKey(j), 'Seed Structure'],
  SeedRMSDpw = pro_Superpose [Chains [], ['atom_set' 'All']],
  Close [force 1],
  mol_Create cat db_ReadFields [dbkey, EntryKey(i), 'Pert Structure'],
  mol_Create cat db_ReadFields [dbkey, EntryKey(j), 'Pert Structure'],
  PertRMSDpw = pro_Superpose [Chains [], ['atom_set' 'All']],
  Close [force 1],
  db_Write [dbkey, EntryKey(i), tag [[tok_cat ['Seed RMSD ', totok j],
                                          tok_cat ['Pert RMSD ', totok j]],
                                     [SeedRMSDpw,
                                          PertRMSDpw]]],
  endloop,
  endloop,

 if CreateFiles == 1 then
 mol_Create cat db_ReadFields [dbkey, EntryKey(1), 'Seed Structure'],
 SaveAs tok_cat [first cName Chains [], ' moe'],
 Close [force 1],
  for i = 1, EntryNo loop;
  mol_Create cat db_ReadFields [dbkey, EntryKey(i), 'Pert Structure'],
  SaveAs tok_cat [first cName Chains [], ' moe'],
  Close [force 1],
  endloop,
 endif,

endfunction,

// ComputeEint ================================================================================

function ComputeEint x,

local Eint = third pot_eleEnergy x + third pot_vdwEnergy x + third pot_solEnergy x,
return Eint,

endfunction,

// ComputeRMSG ================================================================================

function ComputeRMSG [],

local Jtot = tr second Potential [],
local Jheavy = Jtot | aElement Atoms [] <> 'H',
local Jhydro = Jtot | aElement Atoms [] == 'H',

local RMSGtot = sqrt ((add rot3d_vDot [Jtot, Jtot]) / (length Jtot)),
local RMSGheavy = sqrt ((add rot3d_vDot [Jheavy, Jheavy]) / (length Jheavy)),
local RMSGhydro = sqrt ((add rot3d_vDot [Jhydro, Jhydro]) / (length Jhydro)),

return [RMSGtot, RMSGheavy, RMSGhydro],
endfunction,

// PerturbCoordinates ================================================================================

function PerturbCoordinates Disp,

local R, posvec, k,
```

```
posvec = rep [0, ((length Atoms []) * 3)];

 for k = 1, length posvec loop;
 R = randU 1;
  if R > 0.5 then posvec(k) = posvec(k) + Disp;
  else posvec(k) = posvec(k) - Disp;
  endif;
 endloop;
posvec = split [posvec, length Atoms []];
aSetPos [Atoms [], (aPos Atoms []) + (posvec)];
endfunction;
```

6. 4. The Preparation of Proteins for Experimentation

A variety of proteins were used in the experiments presented in this thesis. The preparation of all proteins, excluding the truncated model of estrogen receptor-alpha (1GWR-TM), were prepared in a similar fashion. First, the crystal structure of the desired protein was retrieved from the protein data bank (PDB). The PDB, found at the URL: http://www.pdb.org/pdb/home/home.do, is a publically accessible internet database which contains the atomic coordinates of proteins resolved by x-ray diffraction and nuclear magnetic resonance spectroscopy techniques. For convenience, the protein structures used in this thesis are all referred to by their PDB ID. The PDB ID is an alphanumeric label used to identify and find structures in the database. Thus, to find the crystal structure coordinates of ubiquitin (identified as 1UBQ), one merely needs to search the PDB with 1UBQ.

Preparation of the truncated estrogen receptor-alpha model (1GWR-TM) is slightly more complicated than the typical procedure required to prepare the majority of protein structures in this thesis above. First, 1GWR was downloaded from the PDB and loaded into MOE. Since only one monomer is required to model the active site, the

first monomer A corresponding to chains 1, 3, 5 and 7 was deleted. Deletion of the chains involves the sequence editor: MOE | SEQ > Prompts the sequence editor, and selecting the chains to be deleted: SEQ | Select Chains: 1, 3, 5, 7. Ensure that the sequence editor is synchronized to the MOE window with: SEQ | Selection | Synchronize, and then selected all the atoms belonging to monomer A by: SEQ | Selection | Atoms | of selected chains. These atoms can be deleted leaving a single monomer with new chain index numbers for monomer B (1), the cofactor peptide (2), the native estrogen ligand (3), and the waters associated with the structure (4). At this point, there are a total of 1989 heavy atoms in the system. The cofactor peptide of monomer B (now chain index number 2) is also deleted leaving 1913 heavy atoms in the system. The crystal structure of monomer B has several residues that are incomplete and are missing atoms. To find the incomplete amino acids the residue selector can be used which is prompted by: SEQ | Selection | Residue Selector > Prompts the residue selector (RS). Incomplete residues can then be found by searching missing atoms under amino acid atom data (from the 'more:' menu). The incomplete amino acids (Leu306, Tyr331, Tyr459, Leu446, Lys467, Leu469, Lys492, Lys531, Leu536) can be fixed by using the protein mutation tool (SEQ | Edit | Protein | Mutate) mutating first to alanine and then back to original amino acid. Repeating this procedure for all of the incomplete amino acids gives a structure with 1958 heavy atoms. Hydrogens are then added (MOE | Edit | Hydrogens | Add Hydrogens) to give a total of 3980 atoms in the system. The hydrogen bonding network between the ligand **E2**, the receptor residues (Glu353, Arg394 and His524) and the water molecule HOH2009 was adjusted to achieve the

network as described by Prathipati *et al.*[55] Specifically, coordinate adjustments were made to: the A-ring hydroxyl hydrogen of **E2** [7.946, 3.014, -18.141], the D-ring hydroxyl of **E2** [-3.514, 3.502, -14.085] and the hydrogens of HOH2009 [10.084, 2.075, -17.063] and [9.456, 2.943, -16.434]. Following coordinate adjustment, the model was truncated omitting residues further than a 12 Å radius from the ligand. This is accomplished using the atom selector accessed from: MOE | Select > Prompts the atom selector (AS). First selecting **E2**, the selection proximity is extended to 12 Å by: AS | Raduis: 12, Proximity | Extend: Residue. Then the selections is inverted in the atom selector and the newly selected atoms (external to a 12 Å distance from the ligand) are deleted. This leaves a total of 2043 for the final 1GWR-TM seed receptor.

Table 6.1 – Preparation of protein structures for experiments

PDB ID	Experiment(s)	Preparation Protocol
1GWR	2.4, 2.6	No modification to the PDB file, hydrogens added in default orientations unless otherwise specified
1GWR-TM	3.1, 3.2, 4.2, 4.3, 5.2, 5.3, 5.4	Truncated model prepared as described above
1HDO	3.2	All solvent deleted and hydrogens were added in default orientations
1KPI	2.4, 2.6	No modification to the PDB file, hydrogens added in default orientations unless otherwise specified
1UBQ	2.2, 2.3, 2.4, 2.5, 2.6, 3.2	No modification to the PDB file, hydrogens added in default orientations unless otherwise specified (solvent deleted in 3.2)
2Y1R	2.4, 2.6	Use of only one monomer, hydrogens added in default orientations unless otherwise specified
3APU	2.4, 2.6	Use of only one monomer, hydrogens added in default orientations unless otherwise specified
3M9H	3.2	All solvent deleted and hydrogens were added in default orientations

Table 6.1 summarizes all of the protein structures used for experiments in this thesis along with their preparation method and section.

6. 5. Bond Compression Experiment

To demonstrate that gradient was a factor influencing the divergence of perturbed minimizations, an experiment involving the compression of bond lengths was performed. Before subjecting bonds to compression, all solvent was removed and hydrogens were introduced to the 1UBQ seed structure using Protonate 3D (function described below). The bond lengths of all non-ring atoms were compressed by 0%, 10% and 25% of their original length using the function aSetBond built into MOE. The function aSetBond performs an internal coordinate adjustment to the desired bond lengths on any two neighbouring atoms provided that they do not belong to a ring. The function aSetBond requires two arguments, the first is an array of vectors each comprised of two elements corresponding to the atom keys that make up a bond. The second argument is the bond length value in angstroms. An example using the aSetBond function can involve two hydrogen atoms having atom keys A1 and A2 such that the function is called by: aSetBond [[A1, A2], [0.7414]] to set the hydrogen-hydrogen bond distance to 0.7414 Å. The coordinates of the seed structure were then perturbed by random displacements of 0.0005 Å to create a test set containing 30 different structures.

6. 6. Hydrogen Addition Experiment

Hydrogen atoms can be added using the more sophisticated command Protonate 3D.[45] When employing the command Protonate 3D the following options were used: Titrate: None (to prevent changes in protonation state), Flip: None (to prevent the flipping of tautomers such as the imidazole of histidine or the carboxylic acid groups belonging to aspartic acid or glutamic acid) and van der Waals: 9-6 (to lessen the repulsion of the van der Waals interactions).

6. 7. Tethered Energy Minimization Experiments

Two types of tethering functions were employed for the experiments described in chapter 3. The tether function belonging to CEM-2 involved linear tethers while the remaining protocols (CEM-1, TEM-1 through -4) involved parabolic tethers. Linear tethers can be set in MOE using the function aSetTether which requires the argument: [[Atom Keys], [[x_1,... x_n], [y_1,... y_n], [z_1,... z_n]], tether weight, 0, 0]. Parabolic tethers are set in the MM function with the option, tetherWeight, sent within the MM argument. The option tetherWeight must be sent with a vector in which each element represents the tether weight for the atom with the same index number in the system.

CEM-1 involved a parabolic tethering protocol consisting of three steps, where (Step-1) all heavy atoms were tethered k = 50 kcal/moll·Å^2, (Step-2) backbone atoms were tethered k = 50 kcal/moll·Å^2, and (Step-3) all atoms are minimized without tether. CEM-2 involved a linear tethering protocol consisting of nine steps. The protein is

divided into shells assigned based on proximity to the ligand (**L**). Shell 1 (**S1**) is comprised of residues within a 6 Å radius of **L** while shell 2 (**S2**) is comprised of residues between a 6 to 12 Å radius of **L**. The set identified as **J** consists of hydrogen atoms that are introduced upon severing bonds when truncating the 1GWR-TM model. Each set is further divided into heavy atom and hydrogen atom groups. The heavy atoms of the ligand are identified as **L** while the hydrogens are identified as **L_h**. The nomenclature is the same for the shells (**J** has no heavy atoms). The tether protocol is listed as follows with tether weights in kcal/mol·Å and the termination gradient (gtest): (Step-1) L = 1000, L_h = 100, S1 = 1000, S1_h = 100, S2 = 100, S2_h = 10, J = 10, gtest = 0.05; (Step-2) L = 100, L_h = 10, S1 = 100, S1_h = 10, S2 = 10, S2_h = 0, J = 5, gtest = 0.05; (Step-3) L = 100, L_h = 5, S1 = 100, S1_h = 5, S2 = 5, S2_h = 0, J = 0, gtest = 0.05; (Step-4) L = 30, L_h = 5, S1 = 30, S1_h = 5, S2 = 5, S2_h = 0, J = 0, gtest = 0.01; (Step-5) L = 10, L_h = 5, S1 = 10, S1_h = 5, S2 = 5, S2_h = 0, J = 0, gtest = 0.01; (Step-6) L = 5, L_h = 3, S1 = 5, S1_h = 3, S2 = 0, S2_h = 0, J = 0, gtest = 0.001; (Step-7) L = 3, L_h = 2, S1 = 3, S1_h = 2, S2 = 0, S2_h = 0, J = 0, gtest = 0.001; (Step-8) L = 2, L_h = 0, S1 = 2, S1_h = 0, S2 = 0, S2_h = 0, J = 0), gtest = 0.0001; and, (Step-9) no tether, gtest = 0.0001.

The TEM protocols involve the use of parabolic tethers which are applied to each atom in an unbiased manner which are gradually released. Because a large number of structures were subjected to tethered energy minimization, an SVL script is provided below to perform the protocol and calculate system properties (i.e. total energy, RMS gradient, etc.). The SVL script expects an input database which also receives program outputs.

```
#svl
#set main 'TetheredMinimizinator'

// ============================================================================================
// Database TEM Function                                                                  ====
// ============================================================================================

// Created by James Davey
// Wright Lab (SC512) Carleton University
// An SVL Script for MOE created July 28/2011

// Function Description  ======================================================================
// The minimizinator is a SVL script designed to energy minimize entries in a database belonging to a specified  //
// molecule field  The function allows the user to perform an energy minimization using three algorithms  steepest //
// descent (SD), conjugate gradient (CG), and truncated Newton (TN)  The function computes system properties for //
// the final energy minimized structures (Etot, Eint, RMSG, RMSDxtal and RMSDpw)                                  //
// ============================================================================================

// User Defined Parameters  ===================================================================

  const Path              = 'c /users/smp-user/desktop/SVL SCRIPTS THESIS/',                    //(a)
  const Database          = 'Perturbed 1GWR-TM mdb',                                           //(b)
  const LigandChain       = 2,                                                                 //(c)
  const ForceField        = '$MOE/lib/mmff94s ff',                                             //(d)
  const PotentialOpt      = [],                                                                //(e)
  const InputField        = 'Pert Structure',                                                  //(f)
  const TEM               = [100, 10, 1, 0 1, 0 01, 0]                                         //(g)

// The minimizinator requires that (a) the directory and path of the project be specified in Path as a token  The (b)  //
// database base containing sturctures to be minimized be provided as a  mbd file in Database as a token, and (c)    //
// the ligand chain be specified in LigandChain as an integer  The (d) force field file is specified in ForceField as a  //
// token indicating the location of the force field file  Modification of (e) the potential setup parameters (i e       //
// distance, gas phase, etc ) can be specified in PotenialOpt as a vector  The (f) input field in the database         //
// containing the structures to be minimized is entered into InputField as a token and (g) the tether scheme must    //
// provide in the vector (g) which lists the tether weights to be applied at each step during the tether scheme  TEM  //
// must end with a tether weight of 0 for the system to arrive at a full minimum                                    //
// ============================================================================================

// Function Declaration  ======================================================================

  function PartialCharge,
  function Potential,
  function pro_Superpose,
  function ComputeEint,
  function ComputeRMSG,
  function MM,


// TetheredMinimizinator Function  ============================================================

function TetheredMinimizinator [],

pot_Load ForceField,
pot_Setup PotentialOpt,

local i, j, k, LigandAtoms, CurrentEntry, dbkey, EntryKey, EntryNo,
local mol, Etot, Eint, RMSG, RMSDxtal, RMSDpw,
```

```
dbkey = db_Open [tok_cat [Path, Database], 'read-write'];
EntryKey = db_Entries dbkey,
EntryNo = db_nEntries dbkey;

db_EnsureField [dbkey, tok_cat ['MM ', InputField], 'molecule'];
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' Etot'], 'float'];
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' Eint'], 'float'],
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' RMSG'], 'float'];
db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' RMSDxtal'], 'float'],

 for i = 2, EntryNo loop;
 db_EnsureField [dbkey, tok_cat ['MM ', InputField, ' RMSDpw ', totok i], 'float'];
 endloop;

 for i = 1, EntryNo loop,
 mol_Create cat db_ReadFields [dbkey, EntryKey(i), InputField],
 LigandAtoms = cat cAtoms ((Chains [])(LigandChain));
 aSetCharge [Atoms [], first PartialCharge [Atoms [], 'FF']],

 for k = 1, length TEM loop;
 MM [gtest·0 001, maxit 1000, tetherWeight (rep [TEM(k), length Atoms []])],,
 endloop;

 Etot = first Potential [dX 0],
 Eint = ComputeEint LigandAtoms;
 RMSG = ComputeRMSG [];
 mol = mol_Extract Atoms [];
 mol_Create cat db_ReadFields [dbkey, EntryKey(i), 'Seed Structure'];
 RMSDxtal = pro_Superpose [Chains [], ['atom_set' 'All']],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField]],[mol]]];
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' Etot']],[Etot]]];
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' Eint']],[Eint]]],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' RMSG']],[RMSG]]],
 db_Write [dbkey, EntryKey(i),tag [[tok_cat ['MM ', InputField, ' RMSDxtal']],[RMSDxtal]]];
 Close [force 1],
 endloop,


  for i = 1, (EntryNo - 1) loop;
  for j = (i + 1), EntryNo loop;
  mol_Create cat db_ReadFields [dbkey, EntryKey(i),
                                tok_cat ['MM ', InputField]],
  mol_Create cat db_ReadFields [dbkey, EntryKey(j),
                                tok_cat ['MM ', InputField]];
  RMSDpw = pro_Superpose [Chains [], ['atom_set' 'All']],
  Close [force 1];
  db_Write [dbkey, EntryKey(i),
        tag [[tok_cat ['MM ', InputField, ' RMSDpw ', totok j]], [RMSDpw]]];
  endloop;
  endloop;


endfunction;

// ComputeEint  ================================================================================

function ComputeEint x;
```

```
local Eint = third pot_eleEnergy x + third pot_vdwEnergy x + third pot_solEnergy x;
return Eint;

endfunction;

// ComputeRMSG  ================================================================================

function ComputeRMSG [];

local Jtot = tr second Potential [];
local RMSGtot = sqrt ((add rot3d_vDot [Jtot, Jtot]) / (length Jtot));
return RMSGtot;

endfunction;
```

6. 8. Docking Simulation Procedures

The docking simulation procedures tested and used in this thesis were conducted in MOE using the triangle matcher placement method. The triangle matcher algorithm is the default ligand placement methodology in MOE and it involves the superposition of ligand atoms over receptor site points (atom sets of three). Three pose refinement strategies were tested, (i) an unconstrained energy minimization, (ii) a constrained and unconstrained energy minimization protocol and (iii) a fixed, constrained and unconstrained energy minimization protocol. A molecular docking simulation can be performed in MOE by loading the prepared and minimized receptor target (with ligand) into the MOE window. Next, the ligand atoms are selected and the docking function is prompted with the command: MOE | Compute | Simulations | Dock > prompts the docking window (DW). From the docking window the pose generation, refinement strategy and scoring methodology can be specified. To generate a set of input poses to be docked the following commands are required: (1) Receptor: Unselected Atoms, (2) Site: Selected Atoms, (3) Ligand Atoms: Selected Atoms, (4)

Placement: Triangle Matcher, (5) Retain : (specify number of poses), and (6) Rescoring 1: none. These commands generate a database of ligand poses, each of which must be placed into the receptor target (without the native ligand present) and subjected to one of the three refinement procedures. The SVL script to conduct the docking is provided below. The program requires the prepared and energy minimized receptor in a database file with the native ligand still in place. The program accepts multiple receptor targets in the database, each requires an index number in the field rseq. The docking program also requires a ligand database with ligands identified in fields with an index number 'mseq' and a conformer number 'conf'. The ligand database includes all ligands to be docked and their conformers. This means that a conformer search must be completed for each ligand. A conformer search can be performed in MOE with: MOE | Compute | Conformations | Conformational Search > prompts the conformational search window (CSEARCH). To conduct a systematic search of the conformer set (ring variations not explored) the following commands must be issued: CSEARCH | Method: Systematic, RMS Gradient: 0.001, RMSD Limit: 0.05 and Strain CutOff: 10. The ligand to be searched must be loaded in the MOE window. Special care must also be taken when examining the output of the conformer search to ensure that no duplicates are present. The program also requires the receptor energy and global ligand energy from which ΔE_{rec} and ΔE_{lig} are to be calculated from.

```
#svl
#set main 'Dockinator'

// ================================================================================================
// Molecular Docking Simulation Program                                                      =====
// ================================================================================================
```

```
// Created by James Davey
// Wright Lab (SC512) Carleton University
// An SVL Script for MOE created July 28/2011


// Function Description ==========================================================================
// The dockinator is an SVL program that performs a molecular docking simulation across a set of perturbed          //
// structures provided in a  mdb database  Potenial enegry descriptors (or free energy descriptors if specified) are  //
// used to score docking poses  The receptor database requires that the original xtal structure be provided along    //
// with the minimized receptor with ligand present and minimized receptor without ligand present  The baseline       //
// energy for each receptor required to compute deltaEreceptor                                                      //
// must be provided in the receptor database  A receptor ID field must be included in the database (integer         //
// numbers unique for each receptor)  Ligands provided in a ligand database must include the conformers to be       //
// searched during docking simulation  Each ligand is given a unique integer ID and each conformer within each     //
// ligand must be numbered as well  The binding affinity for each ligand must also be provided in the database      //
// along with the global minimum energy for each ligand                                                             //
// Docking input poses are generated using MOE's docking function DockFile which has a variety of pose generation //
// options  The poses are refined in three stages  (1) a fixed receptor energy minimization, (2) a constrained      //
// receptor energy minimization, and (3) a full unconstrained minimization  After the first refinement stage,       //
// redundant ligand poses can be discarded  The second constrained energy minimization stage involves the          //
// application of parabolic tethers to atoms in the receptor  Atoms in the receptor are defined in shells based on //
// their distance from the ligand  Both the radius and tether weight applied to two shells (S1 and S2) can be       //
// specified by the user  User defined parameters are set below                                                     //
// ================================================================================================================


// User Defined Parameters  ======================================================================


  // Project Location  ===========================================================================
  const Path= 'c /users/smp-user/desktop/SVL SCRIPTS THESIS/',                                    //(a)
  const ReceptorDatabase          = 'Perturbed 1GWR-TM mdb',                                      //(b)
  const LigandDatabase            = 'csearch mdb',                                                //(c)
  // The project path is specified in (a) Path as a token  The receptor database (b) is provided in ReceptorDatabase //
  // as a token and the ligand database (c) is provided in LigandDatabase as a token as well                        //

  // Receptor Database  ==========================================================================
  const ReceptorIndexField        = 'rseq',                                                       //(d)
  const ErecField                 =  MM Pert Structure Etot',                                     //(e)
  const GrecField                 =  Grec',                                                       //(f)
  const XtalReceptorField         = 'Seed Structure',                                             //(g)
  const TargetReceptorField       = 'MM Pert Structure',                                          //(h)
  const LigandChainIndex          = 2,                                                            //(i)
  // Each receptor in the database must have a unique receptor ID number (d) specified in a field listed by        //
  // ReceptorIndexField  The baseline (e and f) potential and free energy used to calculate deltaEreceptor is      //
  // provided in ErecField and GrecField  The unminimized crystal structure (g) is provided in XtalReceptorField and //
  // the target receptor field is provided in (h) TargetReceptorField  The ligand index chain in the target receptor //
  // must be specified in LigandChainIndex as a token                                                             //

  // Ligand Database  ============================================================================
  const LigandIndexField          = 'mseq',                                                       //(j)
  const BindingAffinityField      = 'logRBA',                                                     //(l)
  const LigandField               = mol',                                                        //(m)
  const globalEligField           = 'globalElig',                                                 //(n)
  const globalGligField           = 'globalGlig',                                                 //(o)
  // The ligand database have list each ligand with a unique integer for identification purpose (j) under         //
  // LigandIndexField as a token  Each conformer of each ligand must also be given an index number (k)            //
  // LigandConformerField  The binding affinity is to be listed under the field identified (l) as BindingAffinityField //
  // The coordinates of each ligand for docking should be placed in (m) LigandField and the global energy and free //
```

```
  // energy should be identified in (n) and (o)                                                    //

  // Docking Parameters  ============================================================================
  const DockingDatabase          = 'Docking Output mdb',                                          //(p)
  const PoseKeep                 = ['Number' 3, 'EintDifference' 1, 'EintCutOff' 20],             //(q)
  const ForceField               = '$MOE/lib/mmff94s ff',                                         //(s)
  const PotentialOpt             = [],                                                            //(t)
  const DockingOpt               = [csearch      0,
                                    sel_ent_only  0,
                                    maxpose       200,
                                    wall          ['',0,[0,0,0],[1000000,1000000,1000000],0],
                                    placement     'Triangle Matcher',
                                    placement_opt [ timeout 300, nretpose 1000 ],
                                    scoring       'None',
                                    scoring_opt   [],
                                    refine        'None',
                                    refine_opt    [],
                                    rescoring     'None',
                                    rescoring_opt [],
                                    remaxpose     30,
                                    dup_placement 1,
                                    dup_refine    1],                                            //(u)
  const S1                       = [TetherWeight 0, Radius 6],                                    //(v)
  const S2                       = [TetherWeight 10, Radius 10],                                  //(w)

   // The docking database is specified under (p) DockingDatabase as a token  The number of redundant docked     //
   // poses to be retained after the first refinement step (q) is listed as an integer in PoseKeep              //
   // Force field and potential setup options should be                                                         //
   // provided to (s) ForceField and (t) PotentialOpt  Options for docking pose inputs using MOE's DockFile program  //
   // are specified in (t) DockingOpt   The shells required for the 2nd refinement stage are listed in (v) S1 and (w) and //
   // require the TetherWeight and Radius of each shell  Any atoms external to S2 are fixed in the 2nd refinement  //
   // stage                                                                                                     //

// Function Declaration  ========================================================================

  function PartialCharge,
  function Potential,
  function ComputeEint,
  function ComputeEsel,
  function MM,
  function DockFile,
  function SelectionExtendProximity,
  function SelectionExtendResidue,

// Dockinator Function  ========================================================================

function Dockinator [],

pot_Load ForceField,
pot_Setup PotentialOpt,

local ligdbkey = db_Open [tok_cat [Path, LigandDatabase], 'read-write'],
local ligEntryKey = db_Entries ligdbkey,
local ligEntryNo = db_nEntries ligdbkey,

local recdbkey = db_Open [tok_cat [Path, ReceptorDatabase],  read-write'],
local recEntryKey = db_Entries recdbkey,
local recEntryNo = db_nEntries recdbkey,
```

```
local dbkey = db_Open [tok_cat [Path, DockingDatabase],  create ],
db_CreateField [dbkey, ReceptorIndexField, 'int'],
db_CreateField [dbkey, LigandIndexField, 'int'],
db_CreateField [dbkey, BindingAffinityField, 'float'],
db_CreateField [dbkey, 'BoundPose', 'molecule'],
db_CreateField [dbkey, 'deltaEbinding', 'float'],
db_CreateField [dbkey, 'Eint', 'float'],
db_CreateField [dbkey, 'Etot', 'float'],
db_CreateField [dbkey, 'dElig', 'float'],
db_CreateField [dbkey, 'dErec', 'float'],
db_CreateField [dbkey, 'freeElig', 'float'],
db_CreateField [dbkey, 'freeErec', 'float'],
db_CreateField [dbkey, 'boundElig', 'float'],
db_CreateField [dbkey, 'boundErec', 'float'],

local LigandAtoms, ReceptorAtoms, tempdbkey, tempEntryKey, tempEntryNo, LigandChain,
local S1_Index, S2_Index, S3_Index, j, Eint, mol,

mol_Create cat db_ReadFields [recdbkey, recEntryKey(1), XtalReceptorField],
LigandAtoms = cat cAtoms cat ((Chains [])(LigandChainIndex)),
aSetSelected [LigandAtoms, 1],

SelectionExtendProximity [(S1 Radius)],
SelectionExtendResidue [],
aSetSelected [LigandAtoms, 0],
S1_Index = SelectedAtoms [],
aSetSelected [Atoms [], 0],  aSetSelected [LigandAtoms, 1],
SelectionExtendProximity [(S2 Radius)],
SelectionExtendResidue [],
aSetSelected [LigandAtoms, 0],  aSetSelected [S1_Index, 0],
S2_Index = SelectedAtoms [],
aSetSelected [Atoms [], 1],  aSetSelected [LigandAtoms, 0],
aSetSelected [S1_Index, 0],  aSetSelected [S2_Index, 0],
S3_Index = SelectedAtoms [],  aSetSelected [Atoms [], 1],
aSetSelected [LigandAtoms, 0],
ReceptorAtoms = SelectedAtoms [],
S1_Index = indexof [S1_Index, ReceptorAtoms],
S2_Index = indexof [S2_Index, ReceptorAtoms],
S3_Index = indexof [S3_Index, ReceptorAtoms],
Close [force 1],

local i,
for i = 1, recEntryNo loop,

mol_Create cat db_ReadFields [recdbkey, recEntryKey(i), TargetReceptorField],
aSetCharge [Atoms [], first PartialCharge [Atoms [], 'FF']],
MM [pot_charge 0, gtest 0 001],
LigandAtoms = cat cAtoms cat ((Chains [])(LigandChainIndex)),
aSetSelected [Atoms [], 1], aSetSelected [LigandAtoms, 0],
ReceptorAtoms = SelectedAtoms [],  aSetSelected [Atoms [], 0],

DockFile [ReceptorAtoms,
                LigandAtoms,
                tok_cat [Path, LigandDatabase],
                tok_cat [Path, 'temp mdb'],
                DockingOpt],
```

```
tempdbkey = db_Open [tok_cat [Path,'temp mdb'], 'read-write'],
tempEntryKey = db_Entries tempdbkey,
tempEntryNo = db_nEntries tempdbkey,
db_CreateField [tempdbkey, 'Eint', 'float'],

oDestroy ((Chains [])(LigandChainIndex)),
ReceptorAtoms = Atoms [],
aSetFixed [ReceptorAtoms, 1],

 for j = 1, tempEntryNo loop,
 LigandChain = mol_Create cat db_ReadFields [tempdbkey, tempEntryKey(j), 'mol'],
 LigandAtoms = cat cAtoms LigandChain,
 aSetCharge [LigandAtoms, first PartialCharge [LigandAtoms, 'FF']],
 MM [pot_charge 0, gtest 0 001],
 Eint = ComputeEint LigandAtoms,
 mol = mol_Extract LigandAtoms,
 db_Write [tempdbkey, tempEntryKey(j), ['Eint' Eint, 'mol' mol]],
 oDestroy LigandChain,
 endloop,

db_Sort [tempdbkey, ['mseq', 'Eint'], [0, 0]],
tempEntryKey = db_Entries tempdbkey,
tempEntryNo = db_nEntries tempdbkey,
Eint = db_ReadColumn [tempdbkey, 'Eint'],
mol = db_ReadColumn [tempdbkey, 'mseq'],

local KeepIndex = 1,
local TargetValue = first Eint,
local TargetIndex = 1,
local TargetMseq = first mol,
local TopEntry = TargetValue,

 for j = 2, tempEntryNo loop,
  if TargetMseq == mol(j) then
   if Eint(j) < (TopEntry + PoseKeep EintCutOff) then
    if (abs (TargetValue - Eint(j))) < PoseKeep EintDifference then
     if TargetIndex < (PoseKeep Number) then
     TargetIndex = TargetIndex + 1,
     KeepIndex = cat [cat KeepIndex, cat j],
     endif,
    else
    TargetIndex = 1,
    KeepIndex = cat [cat KeepIndex, cat j],
    TargetValue = Eint(j),
    endif,
   endif
  else
  TargetMseq = mol(j),
  TargetValue = Eint(j),
  TopEntry = Eint(j),
  TargetIndex = 1,
  KeepIndex = cat [cat KeepIndex, cat j],
  endif,
 endloop,

 for j = 1, length KeepIndex loop,
 LigandChain = mol_Create cat db_ReadFields [tempdbkey, tempEntryKey(KeepIndex(j)), 'mol'],
 aSetFixed [Atoms [], 0],
```

```
 db_Write [dbkey, 0,
          tag [[totok ReceptorIndexField, totok LigandIndexField, 'BoundPose', 'Eint'],
          [i, (db_ReadFields [tempdbkey, tempEntryKey(KeepIndex(j)), 'mseq']),
          (mol_Extract Atoms []), (db_ReadFields [tempdbkey, tempEntryKey(KeepIndex(j)), 'Eint'])]]],
 oDestroy LigandChain,
 endloop,

Close [force 1],
db_Close tempdbkey,

local EntryKey = db_Entries dbkey,
local EntryNo = db_nEntries dbkey,

local Erec = db_ReadColumn [recdbkey, ErecField],
local rseq = db_ReadColumn [recdbkey, ReceptorIndexField],
local Elig = db_ReadColumn [ligdbkey, globalEligField],
local RBA = db_ReadColumn [ligdbkey, BindingAffinityField],
local mseq = db_ReadColumn [ligdbkey, LigandIndexField],

Erec = tag [totok rseq, Erec],
Elig = tag [totok mseq, Elig],
RBA  = tag [totok mseq, RBA],

 for j = 1, EntryNo loop,
 db_Write [dbkey, EntryKey(j), tag [
          ['freeElig', 'freeErec', BindingAffinityField],
          [Elig (totok db_ReadFields [dbkey, EntryKey(j), LigandIndexField]),
           Erec (totok db_ReadFields [dbkey, EntryKey(j), ReceptorIndexField]),
           RBA (totok db_ReadFields [dbkey, EntryKey(j), LigandIndexField])]]],
 endloop,

endloop,

db_Close ligdbkey,
db_Close recdbkey,
db_Sort [dbkey, ['rseq', 'mseq', 'Eint'], [0, 0, 0]],
EntryKey = db_Entries dbkey,
EntryNo = db_nEntries dbkey,

local dElig, dErec, deltaE, Gcomplex, deltaG, Etot,

 for i = 1, EntryNo loop,
 mol_Create cat db_ReadFields [dbkey, EntryKey(i), 'BoundPose'],
 aSetCharge [Atoms[], first PartialCharge [Atoms [], 'FF']],
 LigandChain = last Chains [],
 LigandAtoms = cat cAtoms LigandChain,
 aSetSelected [Atoms [], 1],  aSetSelected [LigandAtoms, 0],
 ReceptorAtoms = SelectedAtoms [],  aSetSelected [Atoms [], 0],
 aSetFixed [ReceptorAtoms, 1],
 MM [pot_charge 0, gtest 0 001],
 aSetFixed [Atoms [], 0],
 local S1Atoms = get [ReceptorAtoms, S1_Index],
 local S2Atoms = get [ReceptorAtoms, S2_Index],
 local S3Atoms = get [ReceptorAtoms, S3_Index],
 aSetFixed [S3Atoms, 1],
 S1Atoms = indexof [Atoms [], S1Atoms],
 S2Atoms = indexof [Atoms [], S2Atoms],
 S1Atoms = S1Atoms <> 0,  S1Atoms = S1Atoms * (S1 TetherWeight),
```

```
S2Atoms = S2Atoms <> 0,  S2Atoms = S2Atoms * (S2 TetherWeight),
local TV = S1Atoms + S2Atoms,
MM [pot_charge 0, cg_maxit 1000, cg_gtest 0 1, maxit 1000, gtest 0 001, tetherWeight TV],
aSetFixed [Atoms [], 0],
MM [pot_charge 0, cg_maxit 1000, cg_gtest 0 1, maxit 1000, gtest 0 001],
Eint = ComputeEint LigandAtoms,
Elig = ComputeEsel LigandAtoms,
Erec = ComputeEsel ReceptorAtoms,
dElig = Elig - db_ReadFields [dbkey, EntryKey(i), 'freeElig'],
dErec = Erec - db_ReadFields [dbkey, EntryKey(i), 'freeErec'],
Etot = first Potential [dX 0],
deltaE = Etot - (db_ReadFields [dbkey, EntryKey(i), 'freeElig'] +
                 db_ReadFields [dbkey, EntryKey(i), 'freeErec']),

db_Write [dbkey, EntryKey(i),['BoundPose' (mol_Extract Atoms []),
                              'deltaEbinding' deltaE,
                              'Eint' Eint,
                              'Etot' Etot,
                              'dElig' dElig,
                              'dErec' dErec,
                              'boundElig' Elig,
                              'boundErec' Erec]],
Close [force 1],
endloop,

db_Sort [dbkey, ['rseq', 'mseq', 'Eint'], [0, 0, 0]],
endfunction,

function ComputeEint x,
local Eint = third pot_eleEnergy x + third pot_vdwEnergy x + third pot_solEnergy x,
return Eint,
endfunction,

function ComputeEsel x,
local Esel =    second pot_strEnergy x +
                second pot_angEnergy x +
                second pot_torEnergy x +
                second pot_oopEnergy x +
                second pot_stbEnergy x +
                second pot_eleEnergy x +
                second pot_vdwEnergy x +
                second pot_solEnergy x,
return Esel,
endfunction,
```

To describe ligand interactions with individual amino acids in the active site the interaction energy can be decomposed by a method described by Shadnia *et al.*[51] The potential energy model in MOE allows for non-bonded potential terms to be disabled

based on "atom state". Atom state is indicated by an integer, i.e. 0 or 1, where atoms of the same state (1 and 1) will have their interactions computed in the system. Atoms that are not of the same state (1 and 2) will not have their interactions computed. By default, all atoms have their state set to 0 which is defined as "wildcard". Atom states of 0 can interact with all states (for example, 0 and 1 or 0 and 2). Consider the computation of the interaction between the native estrogen ligand (E2) and glutamic acid 353 (Glu353) in the active site. Selecting Glu353 and E2 and setting their state to 1 while all other atoms in the system have a state set to 2 means that interactions between E2 and Glu353 are the only computed. To do this, the function aSetState is used which accepts a vector of atom keys and a vector of their respective state numbers. For example, if Glu353 and E2 were to only selected atoms in the system, then the command aSetState [SelectedAtoms, 2] would set their state to 2. Selecting E2 and computing the interaction energy (MOE | Compute | Potential Energy) would print the interaction energy and the break down components (electrostatic and van der Waals) in the SVL window.

Ligand orientations were also described by comparing atoms on docked ligands to atoms belonging to the native ligand in the receptor model from crystal structure. The RMSD is computed between atoms i and j by: $RMSD = \left(\frac{(x_i,y_i,z_i)\cdot(x_j,y_j,z_j)}{2}\right)^{1/2}$. In the case of the docking experiments conducted in this thesis, three land mark atoms were used in the RMSD calculation. These three atoms (oxygen of carbon 3, carbon 9 and carbon 18) are common to all atoms in the ligand set.

REFERENCES

1. Slater, J. C. A simplification of the Hartree-Fock method. *Phys. Rev.*, **1951**, 81(3):385-390.

2. 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.

3. Leach, A. R. *Molecular Modeling Principles and Applications 2nd edition*; Pearson Education, Limited: Essex, 2001; (a) p. 165, (b) pp. 171-172, (c) p. 173, (d) pp. 174-176, (e) p. 177, (f) pp. 178-180, (g) pp. 182-185, (h) pp. 202-203, (i) pp. 205-207, (j) p. 255, (k) p. 262, (l) pp. 263-264, (m) pp. 265-266, (n) pp.661-667.

4. Moore, W. J. *Physical chemistry 4th edition*; Prentice-Hall, Incorporated: New Jersey, 1972; p. 769.

5. Perrin, C. L. *Mathematics for Chemists*; John Wiley and Sons, Incorporated: New York, 1970; (a) pp. 30-35, (b) pp. 80-81, (c) pp. 98-100,(d) p. 264, (e) pp. 231-232, (f) pp. 195-197.

6. Lide, D. R.; Frederikse, H. P. R. *CRC Handbook of Chemistry and Physics*; The chemical rubber company: Boca Raton, 1994; (a) p. **9**-74, (b) p. **9**-73, (c) p. **9**-18, (d) p. **9**-75, (e) p. **1**-1.

7. Gillespie, R. J. The valence-shell electron-pair repulsion (VSEPR) theory of directed valency. *J. Chem. Educ.*, **1963**, 40(6):295-301.

8. Gillespie, R. J. The VSEPR model revisited. *J. Chem. Soc. Rev.*, **1992**, 21(1):59-69.

9. Mo, Y.; We, W.; Song, L.; Lin, M.; Zhang, Q.; Gao, J. The magnitude of hyperconjugation in ethane: a perspective from *ab initio* valence bond theory. *Angew. Chem.*, **2004**, 116(15):2020-2024.

10. Mayo, S. L.; Olafson, B. D.; Goddard III, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.*, **1990**, 94(26):8897-8909

11. 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. Am. Chem. Soc.*, **1992**, 114(25):10024-10035.

12. Chung-Phillips, A. Methods for the Fourier-series expansion of torsional energies. *J. Comput. Chem.*, **1989**, 10(5):733-747.

13. Dang, L. X.; Pettitt, B. M. Simple intramolecular model potentials for water. *J. Am. Chem. Soc.*, **1987**, 91(12):3349-3354.

14. Palmo, K.; Mannfors, B.; Krimm, S. Balanced charge treatment of intramolecular electrostatic interactions in molecular mechanics energy functions. *Chem. Phys. Lett.*, **2003**, 369(3-4):367-373.

15. Jensen, J. H. *Molecular Modelling Basics*; Taylor and Francis Group, Limited Liability Company: Boca Raton, 2010, p. 25-26.

16. Buckingham, A. D. Molecular quadrupole moments. *Q. Rev. Chem. Soc.* **1959**, 13(3):183-214.

17. Cox, S. R.; Williams, D. E. Representation of the molecular electrostatic potential by a net atomic charge model. *J. Comput. Chem.*, **1981**, 2(3):304-323.

18. Chirlian, L. E.; Francl, M. M. Atomic charges derived from electrostatic potentials: a detailed study. *J. Comput. Chem.*, **1987**, 8(6):894-905.

19. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron*, **1980**, 36(22):3219-3228.

20. Bush, B. L.; Bayly, C. I.; Halgren, T. A. Consensus bond-charge increments fitted to electrostatic potential or field of many compounds: application to MMFF94 training set. *J. Comput. Chem.*, **1999**, 20(14):1495-1516.

21. Smith, P. E.; Pettitt, B. M. Modeling solvent in biomolecular systems. *J. Phys. Chem.*, **1994**, 98(39):9700-9711.

22. White, J. A. Lennard-Jones as a model for argon and test of extended renormalization group calculations. *J. Chem. Phys.*, **1999**, 111(20):9352-9356.

23. Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.*, **1996**, 17(5-6)490-519.

24. Halgren, T. A. MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.*, **1999**, 20(7)720-729.

25. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, **1986**, 7(2):230-252.

26. Cieplak, P.; Caldwell, J.; Kollman, P. Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and n-methyl acemaide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *J. Comput. Chem.*, **2001**, 22(10):1048-1057.

27. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher III, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Wtanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.*, **1998**, 102(18):3586-3616.

28. MacKerell, A. D.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolym.*, **2000**, 56(4):257-265.

29. Martin, M. G. Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor-liquid coexistence curves and liquid densities. *Fluid Phase Equilibr.*, **2006**, 248(1):50-55.

30. Halgren, T. A. The representation of van der Waals (vdW) interactions in molecular mechanics force fields: potential form, combination rules, and vdW parameters. *J. Am. Chem. Soc.*, **1992**, 114(20):7827-7843.

31. Cornell, W. D.; Cieplak, P; Bayly, C. I.; Gould, I. R.; Merz Jr., K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **1995**, 117(19):5179-5197.

32. Graboske Jr., H. C.; Harwood, D. J.; Rogers, F. J. Thermodynamics properties of nonideal gases. I. Free-energy minimization method. *Phys. Rev.*, **1969**, 186(1):210-225.

33. Dudek, M. J.; Ramnarayan, K.; Ponder, J. W. Protein structure prediction using a combination of sequence homology and global energy minimization: II. Energy functions. *J. Comput. Chem.*, **1998**, 19(5):548-573.

34. Abdul Hameed, M. D. M.; Hamza, A.; Liu, J.; Zhan, C. -G. Combined 3D-QSAR modeling and molecular docking study on indolinon derivatives as inhibitors of 3-phosphoinositide-dependent protein kinase-1. *J. Chem. Inf. Model.*, **2008**, 48(9):1760-1772.

35. Renugopalankrishnan, V.; Renugopalakrishnan, M.; Sarkar, B. Conformational analysis of β-glycine, L-alanine, and bisglycinato-Cu(II) complex. *J. Quantum. Chem.*, **1975**, 9(S2):109-116.

36. Höltje, H. -D.; Sippl, W.; Rognan, D.; Folkers, G. *Molecular Modeling Basic Principles and Applications 2nd edition*; Wiley-VCH GmbH and Company: Weinheim, 2003; (a) pp. 17-18, (b) pp. 32-37, (c) p. 19, (d) pp. 120-121, (e) p. 108, (f) p.148, (g) p.149.

37. *Molecular Operating Environment*, version 2009.10; Chemical Computing Group Inc.: Montreal, Québec, Canada, 2009. (b) User manual (http://)

38. Basilevsky, M. V.; Shamov, A. G. The local definition of the optimum ascent path on a multi-dimensional potential energy surface and its practical application for the location of saddle points. *Chem. Phys.*, **1981**, 60(3):347-358.

39. Nash, S. G. A survey of truncated-Newton methods. *J. Comput. Appl. Math.*, **2000**, 124(1-2):45-59.

40. Baysal, C.; Meirovitch, H.; Navon, I. M. Performance of efficient minimization algorithms as applied to models of peptides and proteins. *J. Comput. Chem.* **1999**, 20(4):354-364.

41. Williams, C. I.; Feher, M. The effect of numerical error on the reproducibility of molecular geometry optimizations. *J. Comput. Aided Mol. Des.*, **2008**, 22(1):39-51.

42. Huang, C. -C.; Smith, C. V.; Glickman, M. S.; Jacobs Jr., W. R.; Sacchettini, J. C. Crystal structures of mycolic acid cyclopropane synthase from *Mycobacterium tuberculosis*. *J. Biol. Chem.*, **2002**, 277(13):11559-11569.

43. Jaskolski, M.; Gilski, M.; Dauter, Z.; Wlodawer, A. Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta. Cryst. D.*, **2007**, 65(5):611-620.

44. Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.*, **1987**, 194(3):531-544.

45. Labute, P. Protonate3D: assignment of ionization states and hydrogen coordinates to macromolecular structures. *Protein: Struct. Funct. Bioinf.*, **2009**, 75(1):187-205.

46. Wang, F.; Mei, Z.; Qi, Y.; Yan, C.; Hu, Q.; Wang, J.; Shi, Y. Structure and mechanism of the hexameric MecA-ClpC molecular machine. *Nature*, **2011**, 471(7338):331-335.

47. Nishi, K.; Ono, T.; Nakamura, T.; Fukunaga, N.; Izumi, M.; Watanabe, H.; Suenaga, A.; Maruyama, T.; Yamagata, Y.; Curry, S.; Otagiri, M. Crystal structure of the A variant of human alpha1-acid glycoprotein. *Paper to be published.*

48. Warnmark, A.; Treuter, E.; Gustafsson, J.-A.; Hubbard, R.E.; Brzozowski, A.M.; Pike, A.C.W. Human oestrogen receptor alpha ligand-binding domain in complex with 17beta-oestradiol and TIF2 NRBOX3 peptide. *J. Biol. Chem.*, **2002,** 277(24):21862-21868.

49. Kini, R. M.; Evans, H. J. Molecular modeling of proteins: a strategy for energy minimization by molecular mechanics in the AMBER force field. *J. Biomol. Struct. Dyn.* **1991**, 9(3):475-488.

50. Church, W. B.l Palmer, A.; Wathey, J. C.; Kitson, D. H. Homology modeling of histidine-containing phosphocarrier protein and eosinophil-derived neurotoxin: construction of models and comparison with experiment. *Proteins Struct. Funct. Bioinf.*, **1995**, 23(3):422-430.

51. Shadnia, H.; Wright, J. S.; Anderson, J. M. Interaction force diagrams: new insight into ligand-receptor binding. *J. Comput. Aided Mol. Des.*, **2009**, 23(3):185-194.

52. Wright, J. S.; Shadnia, H.; Anderson, J. M.; Durst, T.; Asim, M.; El-Salfiti, M.; Choueiri, C.; Pratt, M. A. C.; Ruddy, S. C.; Lau, R.; Carlson, K. E.; Katzenellenbogen, J. A.; O'Brien, P. J.; Wan, L. A-CD estrogens. I. Substituent effects, hormone potency, and receptor subtype selectivity in a new family of flexible estrogenic compounds. *J. Med. Chem.*, **2011**, 54(2):433-448.

53. Feher, M.; Williams, C. I. Effect of input differences on the results of docking calculations. *J. Chem. Inf. Model.*, **2009**, 49(7):1704-1714.

54. Feher, M.; Williams, C. I. Reducing docking score variations arising from input differences. *J. Chem. Inf. Model.*, **2010**, 50(9):1549-1560.

55. Prathipati, P.; Saxena, A. K. Evaluation of binary QSAR models derived from LUDI and MOE scoring functions for structure based virtual screening. *J. Chem. Inf. Model.*, **2006**, 46(1):39-51.

56. Pandini, A.; Soshilov, A. A.; Song, Y.; Zhao, J.; Bonati, L.; Dension, M. S. Detection of the TCDD binding-fingerprint within the Ah receptor ligand binding domain by

structurally driven mutagenesis and functional analysis. *Biochem.*, **2009**, 48(25):5972-5983.

57. Chica, R. A.; Moore, M. M.; Allen, B. D.; Mayo, S. L. Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. *Proc. Natl. Acad. Sci.*, **2010**, 107(47):20257-20262.

58. Whalen, K. L.; Pankow, K. L.; Blanke, S. R.; Spies, M. A. Exploiting enzyme plasticity in virtual screening: high efficiency inhibitors of glutamate racemase. *Med. Chem. Lett.*, **2010**, 1(1):9-13.

59. Wilson, C.; Gregoret, L. M.; Agard, D. A. Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J. Mol. Biol.*, **1993**, 229(4):996-1006.

60. Halperin, I.; Buyong, M.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins Struc. Funct. Genet.*, **2002**, 47:409-443.

61. Georgiev, I.; Lilien, R. H.; Donald, B. R. Improved pruning algorithms and divide-and-conquer strategies for dead-end elimination, with application to protein design. *Bioinformatics*, **2006**, 22(14):e174-e183.

62. Yanover, C.; Schueler-Furman, O.; Weiss, Y. Minimizing and learning energy functions for side-chain prediction. *J. Comput. Biol.*, **2008**, 15(7):899-911.

63. Asim, M.; El-Salfiti, M.; Qian, Y.; Choueiri, C.; Salari, S.; Cheng, J.; Shadnia, H.; Bal, M.; Pratt, M. A. C.; Carlson, K. E.; Katzenellenbogen, J. A.; Wright J. S.; Durst, T. Deconstructing estradiol: removal of B-ring generates compounds which are potent and subtype-selective estrogen receptor agonists. *Bioorg. Med. Chem. Lett.*, **2009**, 19(4):1250-1253.

64. Marialke, J.; Tietze, S.; Apostolakis, J. Similarity based docking. *J. Chem. Inf. Model.*, **2008**, 48(1):186-196.

65. Sotriffer, C. A.; Gohlke, H.; Klebe, G. Docking into knowledge-based potentialfields: a comparative evaluation of DrugScore. *J. Med. Chem.*, **2002**, 45(10):1967-1970.

66. Jain, A. N. Bias, reporting, and sharing: computational evaluation of docking methods. *J. Comput. Aided Mol. Des.*, **2008**, 22(3-4):201-212.

67. Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.*, **2004**, 47(21):5076-5084.

68. Stjernschantz, E.; Marelius, J.; Medina, C.; Jacobsson, M.; Vermeulen, M. P. E.; Oostenbrink, C. Are automated molecular dynamics simulations and binding free energy calculations realistic tools in lead optimization? An evaluation of the linear interaction energy (LIE) method. *J. Chem. Inf. Model.*, **2006**, 46(5):1972-1983.

69. 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. Pharmcol.*, **2008**, 153(S1):S7-S26.

70. Netzer, R. H. B.; Miller, B. P. What are race conditions? some issues and formalizations. *LOPAS*, **1992**, 1:74-88.