Calculation of Phenolic O-H Bond Dissociation Enthalpies and Optimization of Docking Scoring Functions

by

James Anderson

A Thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Science
in

Chemistry
Carleton University
Ottawa, Ontario, Canada
September 2011

Your file Votre référence
ISBN: 978-0-494-83180-9
Our file Notre référence
ISBN: 978-0-494-83180-9

Abstract

The bond dissociation enthalpy (BDE) of phenol is very sensitive to changes in the substituents added This work will attempt to derive a set of substituent constants for *ortho*-, *meta*-, and *para*-halogenated phenols (F,Cl,Br) using the computational B3LYP functional A multiple linear regression (MLR) was used to derive these constants on exclusively halogenated training sets, resulting in fits with $R^2 > 0\ 99$ As a test, BDEs were predicted for mixed polyhalogenated phenols, demonstrating the additivity and generality of this approach and resulting in good correlations Accurately predicting the binding affinities for ligands has been a challenge for computational chemists for decades The goal of this work is to develop an energy-based scoring function that will be iteratively optimized to match experimental binding affinities using an MLR fit It was found that after 4-7 iterations the process converged with excellent correlations This method is not prone to overfitting and may be of general use in improving docking predictions

Acknowledgments

There are many people who have played a significant role in developing me and helping me reach this point. I would like to thank my supervisor, Dr. Jim Wright, for your many wise discussions and your experience that has been a great aid to my own learning. Dr. Hooman Shadnia, who played the role of a friend and mentor when I first joined the lab, proved an invaluable resource for my many questions. I still hum "Here Comes the Weekend" on Friday afternoons. Jamie Davey has spent more time with me over the last two years than any other colleague. Although a fellow student, you are also a role model for the passion you put into every project. When I put on glasses I will always get this strange urge to shout out "It's like I can touch you!".

I would also like to thank my family who has been a great encouragement in my pursuit of knowledge. From the early days before I could walk when my father would discuss chemistry while holding me in his arms to the later years when my mother home schooled my sister and myself all the way through high school, you two have been incredible! Sister, the best sister I have ever grown up with, thank you for your friendship and all the fun we had. I also thank my dear wife and best friend, Shana, for the encouragement you have been while I was working late nights to finish this thesis. You know how to bring out a smile, even when I don't think it's possible

Finally, I owe my greatest thanks to my Heavenly Father who has given me the chance to learn such a fascinating and challenging topic as chemistry and has also given me the love of learning.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronyms	Definition
BDE	Bond Dissociation Enthalpy
CPCM	Conductor-like Polarisable Continuum Model
DFT	Density Functional Theory
DPs	Docked Poses
E2	17-β Estradiol
ERα	Estrogen Receptor alpha
EDG	Electron-Donating Group
EWG	Electron-Withdrawing Group
G09	Gaussian 09
H-Bond	Hydrogen Bond
MAD	Mean Absolute Deviation
MMFF	Merck Molecular Force Field
MOE	Molecular Operating Environment

MLM2	Medium Level Model 2
MLR	Multiple Linear Regression
PDB	Protein Data Bank
RBA	Relative Binding Affinity
ROS	Radical Oxygen Species
SVL	Scientific Vector Language
SF	Scoring Function

Chapter 1

Introduction

1.1 Part I. Substituent Effects

1.1.1 Oxidation

Living organisms are exposed to oxidation every day. Because of this, organisms have developed methods of controlling the presence of these radicals and, thus protecting themselves [1, Wright 1997]. Oxygen containing radicals, known as reactive oxygen species (ROS) are commonly found in the body.

Examples of such molecules are the superoxide radical anion ($^{\bullet}O_2^-$), hydrogen peroxide (H_2O_2), the hydroxyl radical ($^{\bullet}OH$), lipid peroxide and lipid peroxyl radicals (ROOH and ROO$^{\bullet}$) and many others [2, Colton and Gilbert 2002, p. 24]. These are present in the body and have various effects. In some cases these can be beneficial to living organisms. Leukocytes are one example; they generate hydrogen peroxide as a mechanism to kill bacteria [3, Klebanoff 1970]. However, in many cases this causes cellular damage which may even lead to cell death. When an ROS is generated in the body it will abstract the nearest accessible hydrogen. This uses up the ROS, but it generates another radical in the molecule that lost the hydrogen [4, Halliwell and Gutteridge 1992].

One of the molecules that is often attacked by ROS are membrane fatty acids. The allylic hydrogen atoms are easily abstracted, leaving a carbon centered radical. Molecular oxygen reacts readily to form peroxyl radicals which can in turn abstract hydrogen, continuing the cycle (see Figure 1). Thus, ROS can damage the properties of the lipid membrane rather quickly [2, Colton and Gilbert 2002, p. 610].

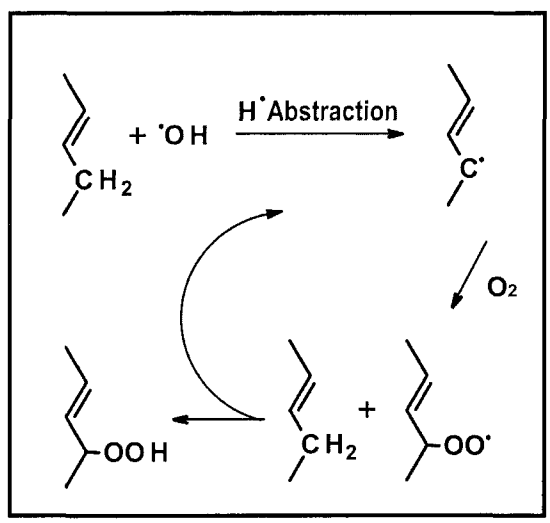

Figure 1: Mechanism of lipid peroxidation.

One method for controlling unwanted oxidation is through the use of low concentrations of antioxidants. These are compounds that quench the radicals without causing damage themselves.

One of the key components found in many natural and commercial antioxidants is the phenol. They make good antioxidants because the phenolic O-H bond is relatively weak [5, Klein 2006b] and thus the hydrogen atom is easily abstracted from the phenol by the radical.

1.1.2 Bond Dissociation Enthalpy

A measure of the strength of a bond is the bond dissociation enthalpy (BDE), or the enthalpy change upon breaking the bond. Stronger bonds have larger BDEs and weaker bonds have smaller BDEs. A BDE is calculated by the reaction,

$$X - Y \rightarrow X^{\bullet} + Y^{\bullet} \tag{1}$$

where species X and Y are bonded together, $X^{\bullet}$ is a radical species X and $Y^{\bullet}$ is the other radical species. The BDE for this prototypical reaction is the sum of the enthalpies of the radicals $X^{\bullet}$ and $Y^{\bullet}$ minus the enthalpy of the parent compound X-Y or,

$$BDE_{X-Y} = H^{o}_{298,X^{\bullet}} + H^{o}_{298,Y^{\bullet}} - H^{o}_{298,X-Y} \tag{2}$$

Specifically, the example that will be examined in detail in this work is the calculation of BDEs for various phenolic compounds. The calculation for phenol is,

$$BDE_{PhO-H} = H^{o}_{298,PhO^{\bullet}} + H^{o}_{298,H^{\bullet}} - H^{o}_{298,PhO-H} \tag{3}$$

Two factors that affect the magnitude of the BDE are the stability of the parent compound and the stabilities of the radical products. Anything that increases the energy gap between the parent and radical increases the BDE. If a substituent stabilizes the parent compound or destabilizes the radical, it increases the BDE. If a substituent is added that stabilizes the radical product or destabilizes the parent compound, then the BDE will be decreased.

Compared with many compounds containing a hydroxyl, the BDE_{OH} of phenol is remarkably low. For example, the BDE_{OH} for a saturated alcohol is around 105 kcal/mol, which is about 20 kcal/mol higher than that of phenol. The reason phenol has such a low BDE is that the radical is resonance stabilized. That is, the unpaired

electron is delocalized across the molecule (see Figure 2) lowering the energy [6, Rappoport 2003].

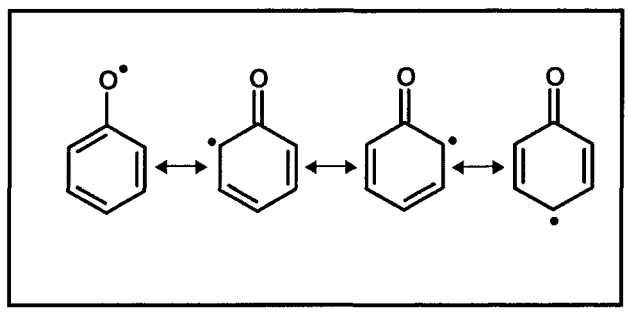

Figure 2: Resonance forms of the phenoxyl radical.

1.1.3 Anti-oxidants

There are many sources of oxidants, as even air is full of O_2, it follows that there are many compounds used for antioxidant purposes. Several common compounds with antioxidant properties are shown in Figure 3, each of which contains one or more phenolic O-H bonds that are the primary place of hydrogen abstraction by ROS.

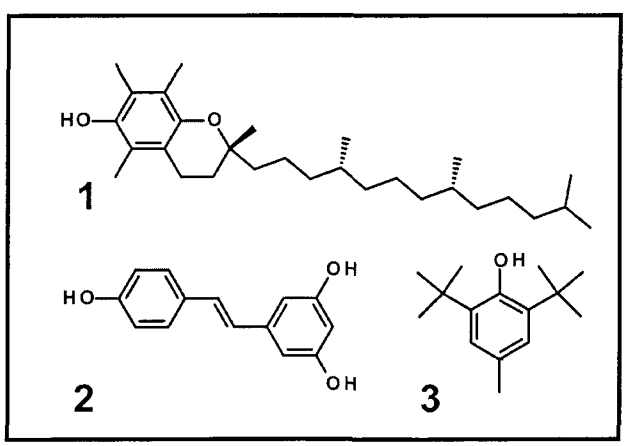

Figure 3: Common antioxidant compounds featuring phenolic hydroxyls in their backbone. **1** α-tocopherol, **2** resveratrol, **3** butylated hydroxytoluene

There are several naturally occurring forms of vitamin E, but the most potent antioxidant form, α-tocopherol, is shown in Figure 3. The phenolic hydrogen atom on vitamin E is donated to the radical species, when acting as an antioxidant [7, Burton and Traber 1990]. The second example, resveratrol, is a compound that is naturally occurring in some plants and has been found to have antioxidant properties [8, Fremont 1999]. Thirdly, Butylated hydroxytoluene (BHT) is an antioxidant that is commonly used as an additive [9, Bjorkhem 1991] [10, Branen 1975]. In each case, the antioxidant capability comes from the phenolic hydroxyls.

Another common use for phenolic compounds is as synthetic organic materials [11, Denisov and Khudyakov 1987] and their radicals also have important roles in biological and industrial applications [12, Halliwell and Gutteridge 1989]. Small quantities of antioxidants are added to synthetic polymer products to protect them from oxidation [13, Zhu, Zhang and Fry 1997]. It is important to be able to understand what factors lead to an increase or decrease in the BDE_{OH}, since they have such important roles [14, Chandra 2002].

Phenols protect from oxidation by donating their H-atom to the radical (see Figure 4). Although this generates a radical from the antioxidant molecule, they are chosen such that their radical form does not react with the protected molecules.

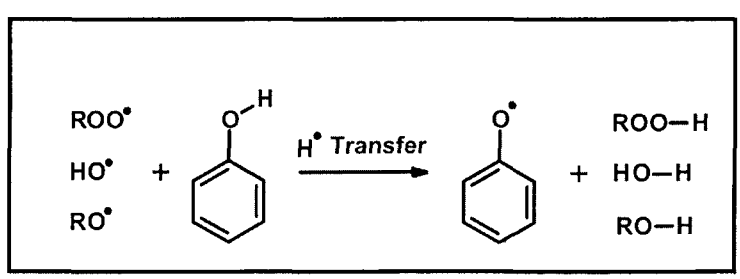

Figure 4: Scheme of phenolic antioxidant donating H-atom to ROS and quenching the radical.

The effectiveness of an antioxidant depends primarily on the strength of the O-H

bond. Although phenols in general have weak O-H bonds, there is significant variation depending on the other substituents attached to the phenol. It is important that the phenol have a weak bond (low BDE) because an antioxidant with a phenolic BDE greater than that of the ROS, will not permit H transfer. For reference, the gas phase BDEs of some ROS products are shown in Table 2, along with the BDE of phenol.

Molecule	BDE_{OH} (kcal/mol)[a]
PhO-H	87.2
ROO-H[b]	89.9
HOO-H	87.9
RO-H[b]	104.9
HO-H	119.0

Table 2: Gas phase BDE values for ROS products and phenol for comparison.
[a] All BDEs obtained from [13, Zhu, Zhang and Fry 1997] except for BDE PhO-H, which was taken from [15, Nix 2006].
[b] R represents a saturated carbon chain.

The BDE of unsubstituted phenol is roughly the same as those of peroxides and hydroperoxides, but is 20 - 30 kcal/mol lower than those of alcohols and water. In this case, phenol would be an effective antioxidant against alkoxyl and hydroxyl radicals, but not work as well for the others. It is of vital importance to know the BDEs of the antioxidant and the target radicals to ensure that the anti-oxidant will be effective.

1.1.4 Substituent Effects on Phenolic BDE_{OH}

The ability to raise or lower the BDE_{OH} of phenol by simply adding or changing substituents has the potential to allow one to design a molecule that has all the desired anti-oxidant properties. Assuming one has found a molecule containing a phenol that has the solubility, metabolic stability or other properties that are desired, the BDE can be 'tuned' to the appropriate range.

The BDE_{OH} can be raised by the addition of electron withdrawing groups (EWG) due to a combination of one or both field/inductive effects and resonance effects [16, Zhang et al 2001] As examples, a p-NO_2 raises the BDE_{OH} by 6 kcal/mol through a combination of these effects, while p-CF_3 raises the BDE_{OH} by 4 kcal/mol using only inductive effects [17, Santes and Simoes 1998] [18, Hansch and Taft 1991]

In contrast, the addition of an electron donating group (EDG) will cause the lowering of the BDE A group such as p-OMe lowers the BDE by 5 kcal/mol by resonance effects, even though it has a slight EWG inductive effect Similarly, p-NH_2 also lowers the BDE_{OH} by 10 kcal/mol by resonance effects despite the inductive EWG effect [17, Santes and Simoes 1998]

Both resonance and inductive effects contributed to the substituent effect in *ortho* and *para* positions, but the *meta* position is only affected by inductive effects [19, Brown, Okamoto and Ham 1957] As a result, the p-OMe group has a -5 kcal/mol effect on the BDE but m-OMe has no effect on the BDE due to the loss of the resonance effects [17, Santes and Simoes 1998] The substituent effects of halogens are of particular interest in this work They are some of the most electronegative elements and thus have a EWG inductive effect However, they are also able to donate electrons from their lone pairs, producing an EDG resonance effect The experimental change in BDE values are contained in Table 3 [17, Santes and Simoes 1998]

The position of the substituent determines which of the inductive or resonance effects dominates Switching a halogen through *ortho*, *meta* and *para* causes the effect to go from decreasing to increasing and back to decreasing the BDE

There are other methods of lowering the BDE_{OH} also A set of compounds commonly used as non-staining anti-oxidants in the synthetic polymer industry contain phenols with two t-butyl groups at the *ortho* positions The t-butyl groups lower the BDE_{OH} primarily because the steric interactions from the bulky groups raise the

	Substituent		
Position	**Fluorine**	**Chlorine**	**Bromine**
ortho	-1.9 (± 2)	-0.7 (± 1)	-1.7 (± 2)
meta	1.4 (± 2)	1.2 (± 1)	–
para	-1.0 (± 1)	-0.2 (± 1)	0.5 (± 1)

Table 3: Effect of substituents on gas phase BDE values relative to unsubstituted phenol.
[a] All BDEs obtained from [17, Santes and Simoes 1998].

energy of the parent molecule. An added benefit of the t-butyl groups is that they shield the phenoxyl radical, to some extent, from further reactions [13, Zhu, Zhang and Fry 1997].

1.1.5 Experimental Value for BDE_{OH}

Although the nature of the phenolic OH bond is of great importance, there is still significant variability in the literature over the experimental value of the BDE_{OH} of unsubstituted phenol. In 2005, Mulder reported that experimental gas phase BDE_{OH} values ranged from 85.8 to 91 kcal/mol [20, Mulder 2005]. One complication to this problem is that many different experiments are used, with some values measured in gas phase while others were measured in liquid phase. Those measured in the liquid phase need to be converted to 'gas-phase' but the assumptions made to correct for hydrogen bonding to the solvent may not be appropriate. The majority of the errors in these 'gas-phase' BDEs were from experiments using highly polar solvents [5, Klein 2006b]. Poor assumptions during conversion or the use of inaccurate auxiliary data can cause the values to be off by several kcal/mol [20, Mulder 2005]. Because of the variations found in the literature, Santos and Simoes published a review in 1998 of approximately 90 literature references to the O-H bond in phenol and substituted

phenols. In this review they predict the most likely gas phase BDE_{OH} for unsubstituted phenol at 88.7 ± 0.5 kcal/mol [14, Chandra 2002]. More recently, however, a report has been published using new data from a combination of gas-phase and liquid-phase experiments where the experimental gas phase BDE_{OH} was calculated to be 86.7 ± 0.7 kcal/mol [20, Mulder 2005]. They mentioned the value selected by Santos and Simoes, but chose not to comment on the discrepancy. It has been mentioned, however, that Santes and Simoes

> "feel that new theoretical and experimental data may invalidate some of their selections because of large discrepancies in the experimental results"

[14, Chandra 2002]. This more recent result by Mulder has been confirmed by several more measurements including one very precise value by Nix et al. where they recorded the gas phase BDE_{OH} as being 85.8 ± 0.1 kcal/mol at 0K [15, Nix 2006] which increases to 87.2 ± 0.1 kcal/mol at 298K [21, Wang and Tang 2010]. This will be used as the reference experimental value in this thesis, as it matches with the average value measured by Mulder and it has higher precision.

1.1.6 Theoretical Calculation of BDE_{OH}

One of the advantages of theoretical calculations is the speed and simplicity with which they can be carried out. Theoretical calculations can also be used to examine compounds that are difficult or impossible, to synthesize in a lab. One such example is phenol with an *ortho* substituent. While it is computationally possible to examine the isomer with the OH facing toward the substituent separate from the OH facing away, in an experiment they are not possible to separate [1, Wright 1997].

It has also been suggested that using a set of substituted phenols it would be possible to determine the effect on the BDE_{OH} for each substituent at each position [1, Wright 1997]. Assuming the initial set of compounds was calculated accurately, this

might enable BDEs to be predicted without any calculations Instead BDEs could be found by adding these terms together

An idea similar to this has already been around since 1937, when Hammett introduced his parameters This constant 'σ' represents the electronic effect of replacing the H in either the *meta* or *para* positions of benzene with a different substituent Hammett initially published 14 σ values that he tabulated and another 17 that were derived from the literature He later extended this set to 44 substituent effects in 1940 [22, Shorter 2000] These values were obtained from a set of *meta-* and *para-*substituted benzoic acids that were synthesized They were put in solution and the equilibrium constants were measured for the equilibrium in Figure 5

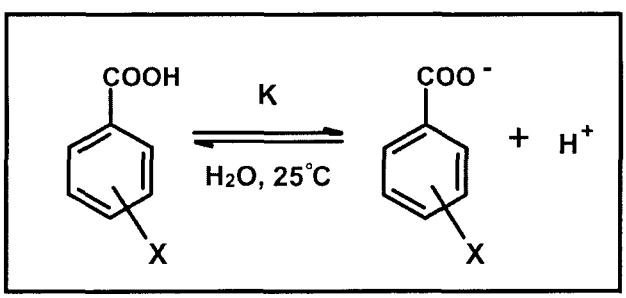

Figure 5: The equilibrium that led to the Hammett parameters

Based on this equilibrium, the σ parameter is defined by the equation,

$$\sigma_{m\,p} = logK - logK_0 = -pKa + (pKa)_0 \tag{4}$$

where K_0 is the equilibrium constant for benzoic acid and K is the equilibrium constant for substituted benzene The pKa values are arbitrarily defined in H_2O at 298K

It is generally true that side chain reactivity for *m-* and *p-* substituted benzenes correlate linearly with σ Thus, we get the two equations that comprise the Hammett

equation,

$$logk/k_0 = \sigma\rho \tag{5}$$

where k is the rate constant of the side chain reaction, and,

$$logK/K_0 = \sigma\rho \tag{6}$$

where K is the side chain equilibrium.

Although these parameters do not measure the strength of the O-H bond, it has been observed numerous times in literature that these parameters correlate extremely well with the substituent effects on the BDE_{OH} of phenol and have been used to calculate predicted ΔBDE values [23, Brinck et al 1997] [24, dos Santos et al 2008].

1.1.7 Quantum Mechanics

When talking about quantum mechanics it is always good to start at the beginning, the Schrodinger equation, which can be written as,

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + V\right\}\Psi(x,y,z) = E\Psi(x,y,z) \tag{7}$$

Where $\nabla^2 = \frac{\delta^2}{\delta x^2} + \frac{\delta^2}{\delta y^2} + \frac{\delta^2}{\delta z^2}$. This is often shortened by defining the Hamiltonian operator, $H = -\frac{\hbar^2}{2m}\nabla^2 + V$. This gives us,

$$H\Psi = E\Psi \tag{8}$$

However, except for special cases this equation can only be solved exactly for systems containing only one electron. A method that produces an approximation to the Schrodinger equation is the Hartree-Fock approximation. This assumes that the Born-Oppenheimer approximation is true, or that the nuclei are fixed relative to the electrons. This allows the wavefunction to be split into the nuclear component and

the electronic component. This allows us to solve the Schrodinger equation for the electrons alone. The other assumption in Hartree-Fock is that the wavefunction can be described as a Slater determinant. [25, Molecular Modeling 2001]

The most popular strategy for the solution to the Hartree-Fock equations is to write each spin orbital as a linear combination of single electron orbitals which are usually called basis functions. The ideal choice for approximating an atomic orbital (basis function) would be a Slater type orbital; however since Slater functions are computationally expensive, Gaussian functions are typically used. Thus, Hartree-Fock attempts to solve the electronic structure of atoms by creating a wavefunction with each electron. [25, Molecular Modeling 2001]

Another method used to determine the electronic structures of atoms is known as density functional theory (DFT). The main assumption in DFT is that the total electronic energy is related to the overall electronic density. DFT tries to calculate the overall electronic density and thus, the total electronic energy. This has the benefit of reducing the problem from considering an electron density that relies on the coordinates of each electron (N electrons with 3N coordinates) down to the coordinates of the overall electronic density (N electrons with 3 coordinates). [25, Molecular Modeling 2001]

The functional used in this work is known as the B3LYP functional. It is a hybrid Hartree-Fock/DFT method which incorporates the gradient correction to the exchange functional by Becke and the Lee-Yang-Parr correlation functional with the local spin density approximation exchange result with the form,

$$E_{\chi C} = (1 + a_0)E_{\chi}^{LSDA} + a_0 E_{\chi}^{exact} + a_{\chi}\Delta E_{\chi}^{B88} + a_c \Delta E_c^{PW91} \tag{9}$$

Where $a_0 = 0.2$, $a\chi, = 0.72$, and $a_c, = 0.81$ as parameterized using the G2 molecule

data set, E_X^{exact} is the exact exchange energy, ΔE_X^{B88} is Becke's 1988 gradient correction, ΔE_c^{PW91} is the 1991 gradient correction for correlation of Perdew and Wang and E_X^{LSDA} is the electron-gas parameterization that Becke took from work published in 1992 by Perdew and Wang. [25, Molecular Modeling 2001] [26, Becke 1993] [27, Lee 1988]

1.2 Part II. Iterative Scoring Functions

1.2.1 Estrogen

The estrogen receptor is very important. There are two subtypes, ERα and ERβ, which have been the focus of much research in an attempt to develop drugs with beneficial effects on these receptors [28, Dahlman-Wright et al 2006]. These may be used to treat symptoms or diseases such as hot flashes and osteoporosis as well as having the potential of being used in the treatment for breast cancer [29, MacGregor and Jordan 1998].

The natural ligand that binds into the estrogen receptors is the steroid estradiol (E2) and is shown in Figure 6. Upon binding, the receptor changes conformation such that it can dimerize with another ligand bound estrogen receptor. A number of other factors bind to the protein dimer and the whole complex moves to the nucleus, where DNA transcription can begin [28, Dahlman-Wright et al. 2006]. The goal of much effort in drug design is to find a synthetic compound that will bind in the same manner, causing the same effects, as the natural ligand so that this process can be jump started when required.

A series of synthetic estrogens were developed previously [30, Wright et al 2011] [31, Asim et al 2009] that have varying activity on the receptors. They have been called 'A-CD' compounds, since they share the same ABCD ring backbone as E2

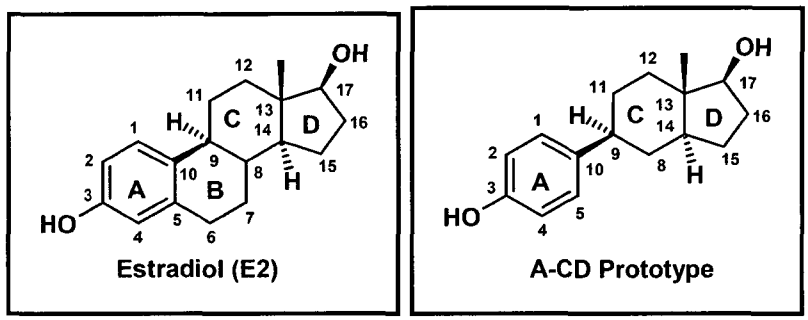

Figure 6: Estradiol the native estrogen ligand (left) and the synthetic A-CD estrogen prototype (right).

except in the fact that the B-ring is missing, shown in Figure 6. For ease of comparison the standard steroid carbon numbering is retained.

The key to binding and activating the estrogen receptors is found in two H-bonding groups at either end of the binding pocket. The C3-OH forms an H-bond network at one end while the C17-OH binds to the other side. The key residues around the C3 OH are Glu353, Arg394 and HOH2009 with His524 at the other side (Figure 7).

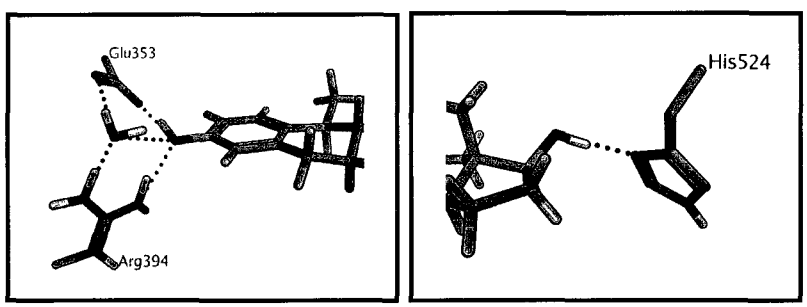

Figure 7: Hydrogen bonding network for estradiol bound to ERα as optimized from the crystal structure 1GWR. The H-bonding network surrounding the 3-OH (left) and the H-bond at the 17-OH end (right).

A compound will generally bind well to the ER receptors if it contains the ability to maintain these H-bonding networks. This part of the work will use computational means, such as potential energy calculations and molecular docking to analyze the interactions of the A-CD estrogens in an attempt to predict their binding affinities

accurately.

1.2.2 Molecular Mechanics

Calculation of electronic configurations using quantum mechanics for small molecules is relatively easy. However, as the number of atoms in the system increases, the number of electrons does so even more quickly. When dealing with large molecules such as proteins, the computational time required to run calculations becomes unfeasible, even when only a few electrons per atom are considered. In order to keep calculations on a reasonable timescale, certain approximations must be made. Molecular mechanics ignores electrons and computes molecular structures and properties by treating atoms as balls held together by springs representing bonds. The atoms are given charges and radii while bonds have an equilibrium length and stretch/compression penalties The potential energy of a particular conformation of the system is given as a sum of the bonded and the non-bonded terms, with each molecular mechanical force-field having slight variations on exactly how these terms are calculated The force-field used in this work is known as MMFF94s, which was developed by Merck [32, Halgren 1996][1]. The non-bonded terms calculated in this force-field are electrostatic, or charged, and van der Waals, or hydrophobic, while the bonded terms are bond stretch, bond angle bend and dihedral (or torsion), with two other terms , stretch-bend and out-of-plane that are used to parameterize specific aspects of this force-field. The five key terms are shown in Figure 8.

[1]The parameters for MMFF94 can be found at ftp //ftp wiley com/public/journals/jcc/suppmat/17/490 and the modifications to out-of-plane and torsion parameters can be found at ftp //ftp wiley com/public/journals/jcc/suppmat/20/720/

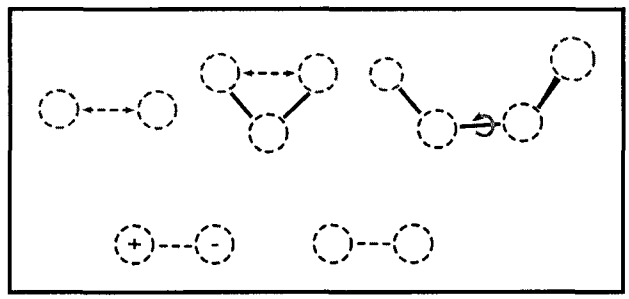

Figure 8: The five key terms that make up a molecular mechanics force-field: (from top left to bottom right) bond stretching, angle bending, torsion, electrostatics and van der Waals.

1.2.3 Bond stretching

A bond between two atoms has an ideal, or equilibrium, length where it is at its lowest energy. Either stretching or compressing this bond from that distance will increase the energy by a certain amount. The specific features of the bond depend on the properties of atoms that form the bond. In general, Hooke's law is the equation that is used to describe the variance of energy over bond length which has the form,

$$E_{stretch} = \frac{k}{2}(r_x - r_e)^2 \tag{10}$$

k is the stretch constant of the bond, r_e is the equilibrium length of the bond and r_x is the bond length after stretching or compressing. While the Hooke's law is reasonable for bond lengths near equilibrium, it behaves poorly as the bond is further from equilibrium. A more sophisticated way to deal with the bond stretching term involves the expansion of the Morse potential using Taylor series. The MMFF94s force-field truncates it to the quartic term, as shown below,

$$E_{stretch} = K_{stretch}(r_x - r_e)^2[1 + cs(r_x - r_e) + 7/12(cs^2(r_x - r_e)^2)] \tag{11}$$

where, cs is the cubic-stretch constant.

1.2.4 Angle bending

When three or more atoms are involved in a molecule the potential for angle bending occurs. The bending of an angle is also commonly described using Hooke's law,

$$E_{angle} = \frac{k}{2}(\theta_x - \theta_e)^2 \tag{12}$$

where, k is the angle bending constant, θ_x is the distorted angle and θ_e is the equilibrium angle. This can again be expanded by including higher order terms (MMFF94s only includes one more),

$$E_{angle} = K_0(\theta_x - \theta_e)^2[1 + cb(\theta_x - \theta_e)] \tag{13}$$

where, K is again the force constant, θ is the angle between three atoms bonded together and cb is the cubic-bend constant. However, when the angle is linear or almost linear, the angle term changes to,

$$E_{angle,linear} = K_{linear}(1 + cos\theta) \tag{14}$$

1.2.5 Torsion

A torsional angle is made between four consecutively bonded atoms i, j, k and l. The angle is defined as the angle between the plane made by atoms i, j and k and the plane made by atoms j, k and l. As the torsional angle increases beyond the optimum value, the energy increases. The energy term is calculated as follows,

$$E_{torsion} = 0.5[V_1(1 + cos\theta) + V_2(1 + cos2\theta) + V_3(1 + cos3\theta)] \tag{15}$$

where V_1, V_2 and V_3 are force constants that relate to the atoms involved in the dihedral.

1.2.6 Electrostatic

The first of the non-bonded terms, the electrostatic term accounts for the interactions between charges over space. It is calculated using a buffered coulombic form,

$$E_{electrostatic} = (q_i q_j)/(D(r_{ij} + \delta)) \tag{16}$$

where q_i and q_j are the partial charges of the atoms i and j, D is the dielectric constant, r_{ij} is the distance between atoms i and j and δ is a buffering constant that is equal to 0.05 Å.

1.2.7 Van der Waals

The van der Waals energy term describes the interactions that cannot be described by electrostatic interactions. Unlike electrostatic interactions, where the term is either attractive or repulsive, van der Waals interactions have both attractive and repulsive terms (see Figure 9).

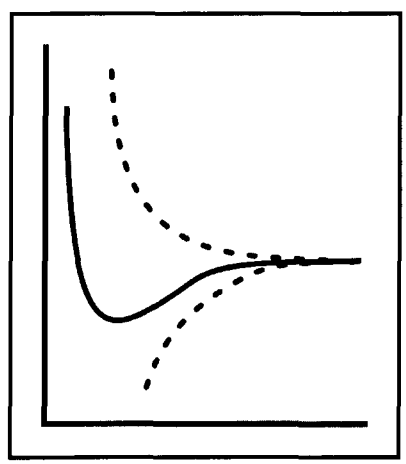

Figure 9: The attractive (lower dashed curve) and repulsive (upper dashed curve) terms sum together to form the picture.

At larger distances the attractive forces dominate and as the distance decreases the

repulsive force become more prominent. Dispersive, or instantaneous induced dipole, forces are the source of the attractive component Repulsions between electrons begin as the atoms are brought closer together Electrons with the same quantum numbers are not permitted in the same system, and so the electron density shifts away from the internuclear area and we see repulsions between the two nuclei One of the most commonly used functions that models these interactions is the Lennard-Jones function,

$$E_{L-J} = k\epsilon[(\frac{\sigma}{r})^n - (\frac{\sigma}{r})^m], \qquad k = [\frac{n}{(n-m)}](\frac{n}{m})^{\frac{m}{(n-m)}} \tag{17}$$

where ϵ is the well depth, σ is the collision diameter, n is the exponent for the repulsive term and m is the exponent for the attractive term. Typical values for n and m that are commonly used are 12 and 6 The MMFF94s force-field uses a variant of this form labeled a 'buffered 14-7' potential

$$E_{vdW} = \epsilon(\frac{1+\delta}{\rho+\delta})^{(n-m)}(\frac{1+\gamma}{\rho+\gamma} - 2) \tag{18}$$

where $\rho = r/r^\bullet$, $r^\bullet$ is the distance where the energy reaches a minimum and both δ and γ are constants applying to all interactions between the atoms For the 'buffered 14-7' potential, the terms are chosen as n = 14, m = 7, $\delta = 0.07$ and $\gamma = 0\ 12$. The equation then becomes,

$$E_{vdW} = \epsilon(\frac{1.07r^\bullet}{r+0.07r^\bullet})^7(\frac{1\ 12r^{\ 7}}{r_7+0\ 12r^{\ 7}} - 2) \tag{19}$$

1.2.8 Out-of-plane

The two other terms in the MMFF94s force-field are not present in every force-field The first is the out-of-plane term and is used to keep atoms in plane that would otherwise be out of the plane. For example, the oxygen on cyclobutanone as in Figure 10,

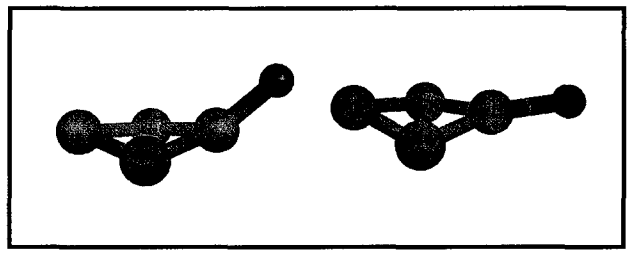

Figure 10: Cyclobutanone with hydrogens omitted for clarity showing C=O non-planar (left) and planar(right).

would be modeled with the C=O out of the plane of the ring using the stretch and angle terms. Since force-fields do not calculate electronic properties, they do not account for the π-bonding energy, which is optimized when planar. Thus a term is added that forces the oxygen back into plane with the ring. The oop term used in MMFF94s is,

$$E_{oop} = K_{oop}(\chi_{ijk,l})^2 \tag{20}$$

where K is the force constant, χ is the Wilson wag angle between bond jl and plane ijk.

1.2.9 Stretch-bend

The last term in the MMFF94s force-field, stretch-bend, is a cross term. Cross terms happen when the effect of one action causes another. There are cross terms between each type of motion, ie. stretch-stretch, stretch-torsion, bend-bend, bend-torsion and stretch-bend. However, stretch-bend is only one that is modeled in MMFF94s. One example of this type of motion would occur when an angle between two atoms was decreased and the atom bonds concurrently lengthened to reduce the interaction between the atoms. Below is the equation used to describe the energetic of this motion,

$$E_{s-b} = [K_{ijk}(r_{ij} - r_{ij}^0) + K_{kji}(r_{kj} - r_{kj}^0)](\theta - \theta^0) \quad (21)$$

where K_{ijk} and K_{kji} are the force constants that couple the stretching of the ij and kj bonds to the ijk angle.

All these terms are summed together to calculate the total energy of the molecular mechanics system. The simplicity of these equations makes it possible to calculate the energy of large molecules, such as proteins.

1.2.10 Molecular Docking

The goal of molecular docking is to predict and analyze the combined interactions of multiple molecules when complexed together. The most common use for docking is currently to predict the binding poses of compounds in protein targets. One of the earliest examples of a docking program was the program published by Kuntz and co-workers [33, Kuntz et al 1982], called DOCK. The average docking can be performed relatively quickly compared with the process of synthesizing and testing a compound experimentally. Thus, it is much more economical to perform tests *in silico* first, and then only test the compounds that are predicted to be promising. Although there has been vast improvement in this field since the start, it still requires significant improvements before achieving highly accurate docking [34, Moitessier et al. 2008].

The general procedure of docking involves a target, usually a protein, and a set of candidate compounds to be tested for binding to that target. The computational method optimizes the placement of each compound in the target, analyzes the best placement for each compound and returns a ranking of the compounds (Figure 11).

Currently, docking is primarily used to identify important binding modes, enrich a compound pool, and analyzing key drug interactions. Although docking can usually find important binding modes, accurately predicting binding affinities has had much

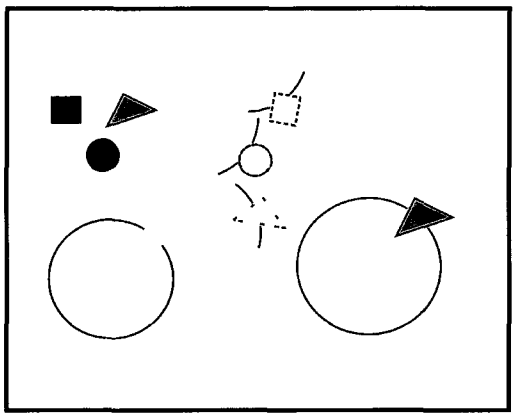

Figure 11: Overview of docking Starting with target and set of compounds (left), determining best fit for each compound (middle) and final best pose for top compound (right)

worse success [35, Warren 2006] Some of the challenges that must be faced when performing a docking calculation include the high computational demands required to calculate the interactions between the many atoms, dealing with many ligand conformations, accounting for protein flexibility and conformational changes as well as interactions with other factors Docking protocols have many different ways of dealing with each of these challenges

In an attempt to find the balance between computational demands and accuracy, different force-fields have been developed that merge hydrogens to their respective carbons [36, Jorgensen and Tirado-Rives 1988] At the other end of the spectrum, some force-fields include more than the standard number of terms, such as MMFF94s, and some programs have integrated molecular mechanics and quantum mechanical calculations as in Gaussian's ONIOM [37, Dapprich et al 1999] [38, Rahu and Merz 2005]

The conformation of a ligand may change depending on the features of the system that surrounds it If prior knowledge of the preferred conformation is not known, then one way to determine the best ligand conformation is to try all possibilities

and see which one is ranked highest. Different docking programs attempt to find the optimum balance in computation versus accuracy by using more or less rigorous ligand conformation searching methods [39, Merz 2010].

Modeling protein flexibility has been a challenge for many docking programs. One possibility is to dock into multiple conformations of the same protein [34, Moitessier et al 2008], while another variation is the ensemble docking algorithm used by Huang and Zou where multiple structures are used and the optimum is selected [40, Huang and Zou 2006] [41, Stjernschantz and Oostenbrink 2010].

One of the largest areas in computational drug design is the development of scoring functions (SFs) to rank-order a set of docked poses. They are generally classified into three categories in the literature: empirical, knowledge based and force-field based.

In empirical SFs, the energy of binding is calculated from many smaller components, such as number of H-bonds, molecular weight, surface area or number of rotatable bonds. Each term is assigned a scaling factor that is optimized by regression to match observed binding affinities (Equation 22) [34, Moitessier et al 2008].

$$\Delta G_{bind} = \Delta G_0 + \Delta G_{H-bond}(\#H-bonds) + \Delta G_{MW}(MW) + \Delta G_{SA}(SA) + \Delta G_{Rot}(\#Rot) \quad (22)$$

One downside to these methods is the requirement for a training set of experimental binding affinities.

Knowledge based SFs do not require binding affinity data as they are built by extracting structural information from sets of crystal structures of ligand-protein complexes. The interactions of the ligand with the protein are grouped into atom-type pairs and statistical analysis is performed on the distribution of these pairs to convert this into potentials. Thus, pair interactions that do not occur frequently are given lower significance than those that appear more frequently [42, Velec 2005] [34, Moitessier et al 2008].

Finally, force-field based SFs are designed to reproduce the energy of ligand binding by the summing various force-field energy terms. These terms often require scaling factors to be applied to achieve predictiveness [34, Moitessier et al 2008] [43, Hecht 2009].

One of the main goals of developing scoring methods is to perform virtual screening of large compound databases. The massive number of compounds that are searched in such a screening requires that the time spent analyzing each compound be kept at a minimum. This work will assume that the compound database has already been refined and that we are working with a few ligands selected for lead optimization. Since we will be working with ligand sets that contain <100 ligands, we have the ability to examine them in greater depth than is possible with virtual screening.

The procedure that is examined here involves the iterative optimization of a scoring function aimed at accurately predicting relative binding affinities [30, Wright 2011]. The SF uses force-field terms and a solvation energy term, combining them to match with observed binding affinities using a multiple linear regression (MLR) analysis. This function is iteratively optimized by reselecting top poses based on the old SF and calculating the SF again based on this new set of poses. Creation of the set of DPs is the most computationally demanding step and only needs to be performed once. Sorting and iterating through the DPs to optimize the SF requires much less time.

A general overview of the steps involved in the process of iterative scoring function optimization is shown in Figure 12.

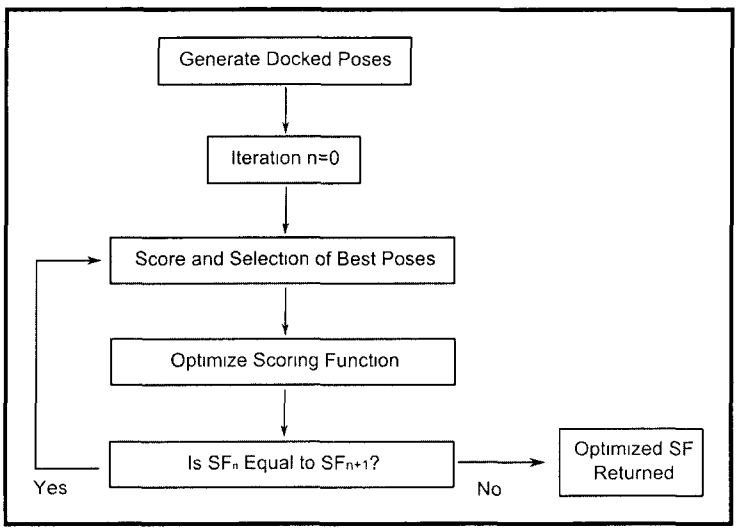

Figure 12: The steps involved in an iterative optimization of a SF, note that the computationally intensive step, generation of DPs only occurs once.

This kind of iterative pose selection has been highlighted previously in the literature, often with reference to the dangers of overfitting. In one example, Huang and Zhou used an iterative procedure to derive interaction potentials for ligand-protein interactions using a knowledge based procedure [44, Huang and Zou 2006a]. Another example is the empirical scoring function that Martin and Sullivan [45, Martin and Sullivan 2008] describe called 'AutoShim' where the SF uses an iterative procedure to select poses. As will be shown in the following pages, the procedure for iterative pose selection employed here is simple and predictive as well as avoiding the danger of overfitting.

Chapter 2

Methods

Calculations run on Gaussian 09/03 were submitted to the High Performance Computing Virtual Laboratory (HPCVL), a cluster of powerful Sun computers located across several universities and colleges in Ontario. Calculations run in the Molecular Operating Environment 2009 were performed on a local cluster of AMD dual core processors running Window XP professional SP3. For an overview of the programs used in this work, see Appendix 1.

2.1 Part I. Substituent Effects

The program Gaussian [46, G09] was used for all DFT calculations involving substituent effects. The Molecular Operating Environment (MOE) [47, MOE2009] was used to generate three dimensional atom coordinates for the compounds and create input files to be run in Gaussian.

The compounds were generated in MOE and then saved into a MOE database. A script written in MOE's Scientific Vector Language (SVL) written by H. Shadnia was used to extract the atom coordinates and insert them into a template Gaussian input file with the desired runtime parameters. These Gaussian input files were then submitted to the HPCVL cluster and processed.

Calculation of gas phase phenolic O-H BDEs were carried out according to the MLM2 method [48, DiLabio et al. 1999], which is defined as follows. The geometry of the structures was first optimized using the OPT command in Gaussian with the B3LYP hybrid functional [49, G09 User Reference] [26, Becke 1993] [27, Lee 1988] and the 6-31G(d) basis set. Using the optimized coordinates, vibrational frequencies were also calculated at the same level using the FREQ=ReadIsotope command. This allows the calculations to be performed at alternate temperatures, pressures, frequency scale factors or isotopes. The default values for temperature (298.15 K), pressure (1.0 atm) and isotope were chosen, however, the frequency scale factor was set to 0.9806. This value was found to be appropriate previously [48, DiLabio et al 1999]. The vibrational analysis provided the thermal correction for enthalpy values for both the phenol (ArOH) and the radical (ArO$^\bullet$). Electronic energies of the compounds were calculated using the UB3LYP functional and the larger 6-311+G(2d,2p) basis set for the phenols and the ROB3LYP/6-311+G(2d,2p) level for radicals. Summing the thermal correction for enthalpy with the electronic energy provided the enthalpy for the compound. The enthalpy of H$^\bullet$ was calculated by setting the electronic energy to its exact value of -0.50000 a.u. This was done to achieve BDEs in good agreement with experiment [50, Johnson et al 2003].

The BDE_{OH} for the formation of the phenoxyl radical was calculated by:

$$BDE_{OH} = \Delta H^o_{298,rxn} = H^o_{298,ArO\bullet} + H^o_{298,H\bullet} - H^o_{298,ArOH} \tag{23}$$

BDE_{OH} values were calculated for each compound in the same manner. Changing method/basis set can result in large variances in BDE_{OH} values, however, the relative difference between BDE_{OH} values remains constant [14, Chandra 2002]. Thus, BDE

differences (ΔBDEs) were calculated as follows

$$\Delta BDE = BDE_{X-phenol} - BDE_{phenol} \quad (24)$$

A positive value of ΔBDE designates a stronger substituted phenolic OH bond and a negative value designates a weaker OH bond

Once calculated, the ΔBDEs were correlated to the number and type of substituents by performing a multiple linear regression (MLR) with ΔBDE as the dependent variable and the independent variables being the type (F, Cl or Br) and position (*ortho*$_{toward}$, *ortho*$_{away}$, *meta* and *para*) A set of Hammett-like coefficients were returned from the MLR that could be used to calculate the predicted ΔBDE for an arbitrarily substituted phenol as follows

$$\Delta BDE_{(pred)} = C_{X\ H-bond}(\#H-bond\,X) + C_{X\ o}(\#o-X) + C_{X\ m}(\#m-X) + C_{X\ p}(\#p-X) \quad (25)$$

where the terms are the number of internal H-bonds between OH and the halogen substituent, the number of halogen substituents at each position (*ortho*, *meta* and *para*), X represents any of F, Cl or Br and the coefficients (C) are the parameters returned by the MLR

2.2 Part II. Iterative Scoring Functions

MOE 2009 was used extensively for structure preparation, potential energy minimizations, docking calculations, energy reading and final pose analysis All docking calculations and energy minimizations used the MMFF94s force field in the gas phase [32, Halgren 1996], as implemented in the MOE software, unless otherwise mentioned This force field was developed for energy minimizations and was parameterized to treat both proteins and small molecules well, making it useful in medicinal chemistry [51, Halgren 1999] A distance-dependent dielectric constant of 1 was used

with a cutoff at 10Å.

2.2.1 Receptor Preparation

Three dimensional coordinates, in the form of X-ray crystallographic data, were obtained for human recombinant ERα from the Protein DataBank [52, RCSB]. The file with PDB ID: 1GWR was chosen as it contained the native E2 ligand bound in a ligand binding domain (LBD) homo dimer. The dimer is in complex with coactivator peptides, has a good resolution of 2.4Å and also contains the associated water molecules [53, Warnwark 2002]. For the purposes of this docking study, only one of the two monomers, chain B, was kept, along with its bound ligand and water molecules.

Heavy atom coordinates are contained in the PDB file, but the hydrogen atoms are missing. The hydrogen atoms must be included before energy minimizations can be performed. The program MOE has the option to add hydrogens by default, which assigned charged residues according to pH 7, giving the standard geometry for each atom type contained in the PDB file. In reality, many hydrogen atoms do not have completely standard geometry. Thus, the resulting file was high in energy, as many hydrogens were placed in orientations that clashed with other atoms. These clashes were dealt with later by energy minimization using the default minimization scheme in MOE. This scheme uses a combination of three methods: steepest descent, conjugate gradient and truncated Newton. Energy minimizations begin with steepest descent until the gradient of the system reaches 1000 kcal/mol Åthen conjugate gradient takes over. The minimization continues with conjugate gradient until the gradient reaches 100 kcal/mol Å, at which point truncated Newton is used until the system is fully minimized.

The crystal structure provides vital information on the orientation of atoms and residues that provides the basis for further analysis and interpretation of the protein. However, some parts of the structure are incomplete and must be corrected. This

often happens towards the outside of the protein, where there is more flexibility in the amino acid residues. One example of such an incomplete residue is Lys531 (Figure 13) which is missing the ϵ-carbon and the amino at the end of the side chain.

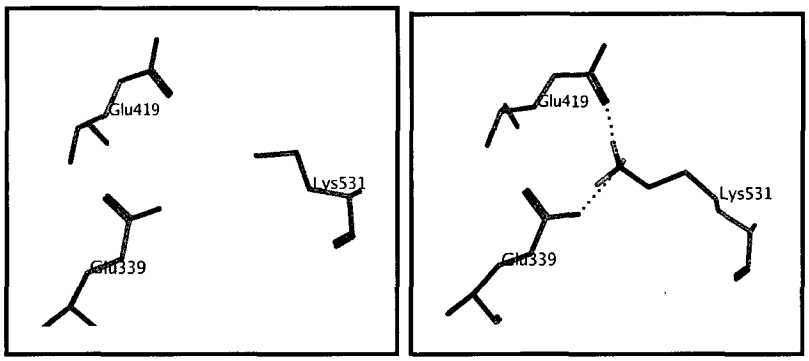

Figure 13: Incomplete Lys531 residue as found in the 1GWR monomer B crystal structure (left). Fixed Lys531 residue with complete side chain and restored hydrogen bond interactions with Glu339 and Glu419 (right).

In this example it is clear the optimal orientation of the side chain is facing the nearby charged Glu residues enabling them to participate in a H-bonding network with Lys. The full residue was added in using a function in MOE called the Rotamer Explorer that generates a list of possible mutations ranked by its own scoring function. Each rotamer was compared with the side chain atoms available in the crystal structure. The highest ranked rotamer that matched with the crystal structure was chosen and the side chain was allowed to relax through a coarse energy minimization. The restored H-bond network can also be seen in Figure 13.

Another area of the protein that requires particular care is the orientation of the hydrogens in the active site. There are several key residues, Arg394, Glu353, His524 and HOH2009 that participate in binding E2 within the active site. The OH on C3 of E2 forms an important H-bonded bridge with Arg394, Glu353 and HOH2009 [54, Prathipati 2006]. At the other end of E2, the OH on C17, also participates in an

important H-bond with His524. Initially, the hydrogens are placed such that the H-bond network is disrupted. The hydrogens on E2 and HOH were adjusted (as shown in Figure 14) to approximate the ideal H-bond network.

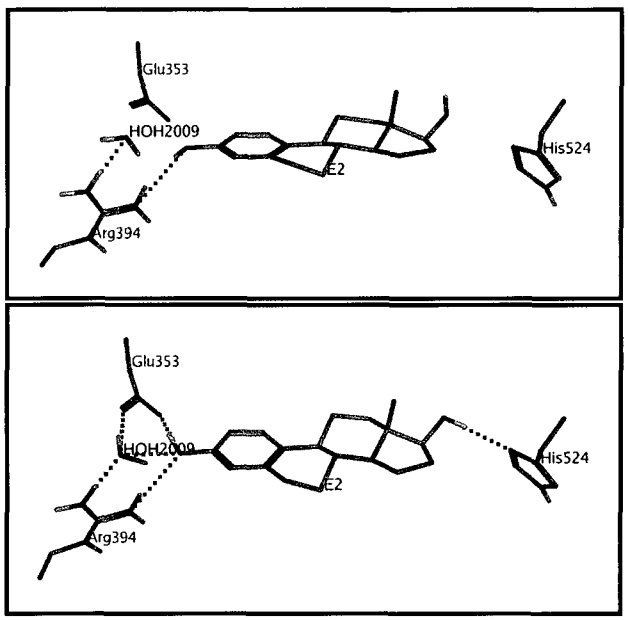

Figure 14: Hydrogen positions as added by default in MOE (top). Manually adjusted hydrogens on E2 and HOH to approximate the ideal H-bond network (lower).

The prepared complex was then energy minimized to remove the hydrogen clashes and optimize bond lengths, angles and protein dihedral angles for the force field. Our in-house program H_pdb_thaw which was written in MOE's Scientific Vector Language (SVL) was used to perform energy minimizations and create shells. Protein shells were defined to reduce computation time as well as to accommodate the docking protocol. A protein shell was defined as being the residues within a certain distance from the ligand. All residues with at least one atom within 6Å of the ligand were defined as the first shell, S1. The second shell, S2, was defined as the residues outside the S1 shell and having at least one atom closer than 12Å from the ligand. The remaining residues were all deleted and the broken bonds were capped with hydrogen atoms.

These 'capping' hydrogen atoms were labeled, S3 (See Figure 15).

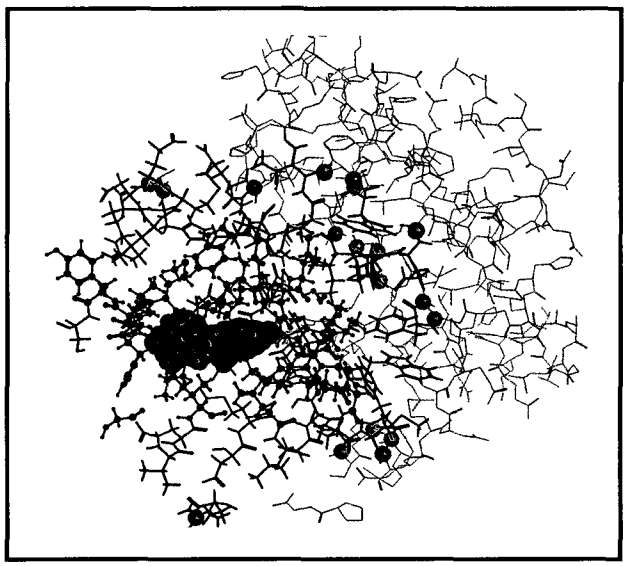

Figure 15: The shells of the receptor coloured with red (ligand), green (S1), blue (S2), orange (S3) and grey (discarded atoms).

S1 included all the active site atoms, containing 45 residues (664 atoms), S2 included 84 residues (1295 atoms) and S3 contained 25 atoms. The full monomer B contains around 4000 atoms, so the protein has been effectively reduced to half its previous size.

The thawing program attempts to reduce the distortion on the ligand and active site of the protein by initially keeping those atoms virtually fixed using strong tethers. Tethers are artificial parabolic potentials of the form kdx_i^2, where dx_i is the displacement from the initial atomic coordinate x_i and k is the tether constant. The minimization starts with outer atoms having low tether and inner atoms strongly tethered. The next step of minimization has lowered tethers. After each round of minimization, the tethers are gradually lessened, starting from the outer atoms. All tethers were removed for the last round of minimization to create a fully thawed crystal structure in the MMFF94s force field.

Although the next step was not necessary for the docking procedure or calculation of the required energy terms, it was useful in the analysis of the docked poses (DPs). A graphical representation of the surface of the active site was created using MOE. The ligand atoms, bound inside the receptor active site, were selected and the MOE option: Compute||Surfaces and Maps||Interaction(VdW). The image below (Figure 16) shows the generated active site.

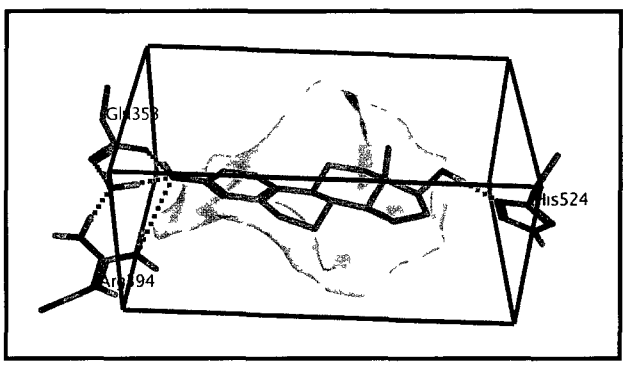

Figure 16: The active site surface as rendered by MOE in grey and the docking box defining the active site area with a few Angstrom buffer drawn in red lines.

In order to speed up docking, a docking box was also defined. This is an area of the protein in which the docking ligands are to be placed. This allows the docking program to ignore the irrelevant parts of the protein and only focus on the defined area of interest. A docking box was created that surrounded the whole active site, with a few angstroms buffer space in each direction (Figure 16).

2.2.2 Ligand Preparation

The compounds examined in this work are variants of the basic A-CD structure. That is, they are the standard steroid backbone with the B-ring removed. For ease of comparison between E2 and these compounds, the carbon atom numbering is the same as that of the standard steroid ring numbering. The stereochemistry of all

stereo centers in A-CD is the same as E2, except at C14, where the hydrogen faces up. These compounds are labeled the 'cis' series, since the C14 hydrogen faces in the same direction as the methyl. The 'trans' A-CD series contains compounds with the C14 hydrogen facing down. The A-ring was modified by adding substituents at positions 2, 4 and 5 on the A-ring. These substituents included F, Cl, CH_3 and CF_3. The compounds labeled 1a-1l and 1o have a hydroxyl group at C17, as in E2, while compound 1p has a carbonyl group at C17, as in estrone. Compound 1q has the C17 hydroxyl and also contains a 17-ethynyl group. There are also three compounds with inverted chirality at C9 (2a, 2e, 2f), four compounds with a double bond in the C-ring between C8 and C9 (3a, 3b, 3e, 3f) and four with a double bond between C9 and C11 (4a, 4b, 4e, 4f). The structures can be seen in the image below (Figure 17).

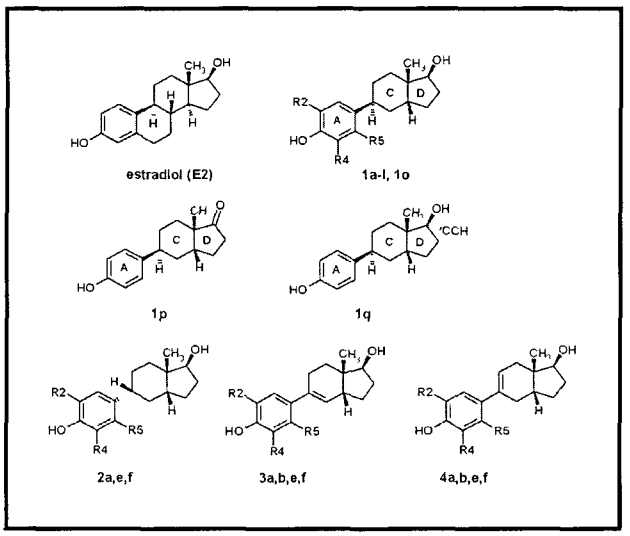

Figure 17: Structures and numbering scheme for A-CD compounds.

All of the experimental binding data are measured relative to the binding affinity of estradiol. The relative binding affinities (RBAs) were all gathered using the same procedure at the same lab, which gives consistency across the data. The RBAs were determined by a competitive radiometric binding assay using estradiol as the

standard [55, De Angelis et al 2005]. For a full listing of the compounds, their substituents and their experimental RBAs, see Table 4. The set of compounds has binding affinities ranging from very poor binders to strong binders, almost 5 orders of magnitude, which will be acceptable for testing a SF.

Compound	R2	R4	R5	RBA α (%)
E2	—	—	—	100 00
1a	H	H	H	1 47 (0 26)
1b	H	F	H	1 04 (0 20)
1c	H	H	F	27 3 (0 69)
1d	H	H	Cl	52 3 (12 0)
1e	H	H	Me	2 82 (0 45)
1f	H	H	CF3	89 7 (14 0)
1g	F	F	H	0 040 (0 002)
1h	Cl	Cl	H	0 004 (0 001)
1i	H	F	F	4 62 (0 93)
1j	F	H	F	0 380 (0 090)
1k	F	F	F	0 186 (0 010)
1l	H	Me	H	1 75 (0 50)
1o	H	Cl	H	0 065 (0 001)
1p	Cl	H	Cl	0 008 (0 001)
1q	H	Me	Me	0 088 (0 006)
2a	H	H	H	0 061 (0 002)
2e	F	F	H	0 006 (0 001)
2f	Cl	Cl	H	0 004 (0 001)
3a	H	H	H	0 246 (0 02)
3b	H	H	Cl	195 (35)
3e	F	H	F	0 038 (0 005)
3f	H	H	CF3	189 (1 0)
4a	H	H	H	0 415 (0 08)
4b	H	H	Cl	59 9 (7 9)
4e	F	H	F	0 436 (0 08)
4f	H	H	CF3	122 (17)

Table 4: Experimental values of relative binding affinities (RBAs) for the full set of *cis*-A-CD ligands bound to ERα. Error bars in parentheses.

There were also several *trans-* compounds used in this work, in addition to the series of *cis-* compounds. There are 7 in total, 5 from the *trans-* 1 series and 2 from the *trans-* 4 series. The complete list, using an equivalent numbering scheme, is as follows: 1b-t, 1c-t, 1e-t, 1f-t, 1i-t, 4b -t and 4c-t. The experimental binding affinities

are shown in Table 5 below.

Compound	R2	R4	R5	RBA α (%)
1b-t	H	F	H	1 7
1c-t	H	H	F	4 2
1e-t	H	H	Me	0.47
1f-t	H	H	CF3	4 8
1i-t	H	F	F	0 92
4b-t	H	H	Cl	9 7
4c-t	H	H	F	21 8

Table 5: Experimental values of relative binding affinities (RBAs) for the *trans-* '-t' A-CD ligands bound to ERα.

Having defined the set of compounds to be docked, the next step was to generate the 3D structures This was done in MOE, building each framework and adding the appropriate substituents A database containing conformers of each compound was generated using the systematic conformer search feature available in MOE All rotatable bonds were systematically rotated by 20 degrees and each conformer was energy-minimized All redundant conformers were removed, leaving a docking input database containing 6 - 12 conformers of each compound

2.2.3 Docking Protocol

The docking algorithm takes each ligand conformer and places it in the center of the docking box that was defined previously Then the conformer is given a randomly generated rotation and translation If the conformer still remained inside the docking box the initial energy of the complex and interaction with the S1 shell was calculated The complex was then partially energy-minimized and the energy terms re-calculated Finally, the complex was minimized fully and the final energy terms were calculated In each minimization, the ligand and S1 shell atoms were unconstrained, S2 shell atoms were tethered and S3 shell atoms fixed This procedure is repeated until 50 poses have been saved for each conformer, resulting in 500-1000 DPs per ligand

Once a pose was selected and fully minimized, there were several energy terms that were calculated The first term measured was the energy of the ligand in that particular pose The energy of the receptor was approximated by calculating the energy of the S1 shell and the energy of interaction between the ligand and the S1 shell was also measured These were combined with several other energy terms to generate the iterative SF that will be discussed later on

The details of how the force-field energy terms were calculated follows As stated above, the energy of the docked ligand pose in complex was measured (E'_L) as well as the lowest energy conformer of the free ligand (E_L) These two terms were combined

to calculate the change in internal ligand energy upon binding to the receptor, $\Delta E_L = E'_L - E_L$. It should be noted that E'_L only takes into account the internal energy of the ligand and disregards the interactions with the receptor. In effect, this is calculating how far above the lowest energy conformer is the particular pose in question. Another way of looking at this is that the term ΔE_L is actually a calculation of the deformation penalty that must be paid to bind the ligand in this pose. For this particular ligand set the ΔE_L values ranged from 3 - 8 kcal/mol for good poses. The energy of the receptor in complex (E'_R) was approximated by taking the energy of the S1 shell of the protein. A true value of ΔE_R should include a measurement of the lowest energy value of the free receptor (E_R) as well, but since it will remain constant for all DPs, we have arbitrarily set it to be 330 kcal/mol. This gave a range of ΔE_R values starting around 30 - 60 kcal/mol for good poses. The interaction energy between the ligand and the receptor (E_{int}) was measured and constituted the third term in the SF. Good poses of E2 docked into the receptor produced values of E_{int} around -73 kcal/mol. The last term used in the scoring function was the energy of de-solvating the ligand from water. This term was calculated in Gaussian 03 with geometry optimized using B3LYP/6-31G(d) level. The conductor-like polarisable continuum model (CPCM) was selected in a second step with the parameters, HF/6-31G(d) SCRF=(CPCM, READ, SOLVENT=WATER) in the first line and PCMDOC, RADII=UAHF and SCFVAC at the end of the input file. Contained in the output file was the fourth energy term, ΔG_{solv}.

Iteration toward finding an optimum SF using the four energy terms described above began with finding the 'best' DP for each ligand. Each pose was analyzed using the London dG scoring function which is the default in MOE. The poses were then rank ordered based on this SF. The top pose for each ligand was chosen to be the one with the most negative value of the London dG score. The four energy terms for each of these 'best' poses were used in generating the first optimized SF. This was done by

performing a multiple linear regression (MLR) over each of the independent variables (the energy terms) with respect to the log of the binding affinity, the dependent variable. The result of the MLR was a SF with coefficients for each energy term optimized to reproduce the experimental binding affinity with the form of:

$$logRBA_{(pred)} = SF = C_1 + C_2(E_{int}) + C_3(\Delta E_L) + C_4(\Delta E_R) + C_5(\Delta G_{solv}) \quad (26)$$

All poses were re-evaluated with the new SF and ordered once more. The new 'best' poses were selected and the regression was performed on this new ligand set. The second iteration SF was again used to evaluate the DPs. This process continued until convergence was reached, which was defined as no further changes in the SF upon subsequent iterations. This process resulted in an iteratively optimized scoring function that could be used to predict binding affinities for docked poses.

To be able to see the predictive ability of the optimized scoring function, the set of ligands will be split into a training set and a separate test set. A 17 ligand training set was selected from the 27 ligands listed in Table 4. The training set contained the reference ligand E2, the parent compound 1a and 15 other ligands chosen to include a representative set of high (>10%), medium (0.1 - 10%) and low (< 0.1%) RBAs. An additional criterion for the training set was that it must include compounds from each set of structure types. The first structure type was A-CD with a saturated C-ring (1a - 1q), those with unsaturated C-rings (the 3- and 4- series) and those with inverted stereochemistry at C9 (the 2- series). The remaining 10 compounds made up the test set. These criteria were used to create four different test sets and their respective training sets.

The scoring function was tested by docking the test set and calculating the predicted log RBA using the optimized final SF for each of the DPs resulting from the docking. The top ranked poses based on this SF were then selected as the best pose

and the predicted binding affinity was correlated to the experimental log RBAs. A good correlation indicates that the SF has predictive value for those compounds. This process was performed for each definition of the training and test sets.

A second test was performed to see if randomized, or nonsense, experimental data could also produce a correlation that appeared reasonable. The data was randomized by scrambling all of the log RBA values for each compound. This was accomplished by randomly swapping log RBA values until each compound had a different log RBA assigned to it. To ensure that the random process did not accidentally roughly correlate to the true values, the log RBA_{expt} were correlated with the new log $\mathrm{RBA}_{scrambled}$ and they had no significant correlation. The iterative procedure was again performed on the new scrambled training set to test convergence.

Chapter 3

Results and Discussion

3.1 Part I. Substituent Effects

The compounds examined here all have one feature in common, that is they are all derived from phenol. They only difference is in the type, position and number of substituents. This difference, however, gives rise to the varying strength of the phenolic OH bond, measured as the ΔBDE. There are five different positions that a substituent might be placed on a mono-substituted phenol. The five different positions are displayed in Figure 18 for a mono-fluorinated phenol. For *ortho*-fluorinated phenol, the 'toward' position is the lowest energy state. The 'away' position is also a potential minimum, but the internal OH — F hydrogen bond is only possible in the 'toward' position. Although reduced in strength from an optimally oriented H-bond, it still plays a significant role in stabilizing the compound. A linear water-HF dimer H-bond has a strength of around 10 kcal/mol [56, Kollman and Allen 1969] in gas phase, thus even a severe reduction in that interaction due to the bent conformation still leaves a significant interaction.

Although there is no internal H-bonding in the *meta* positions, each position is different due to long-range interactions. There is only one arrangement possible for the *para* substituent, giving a total of five positions.

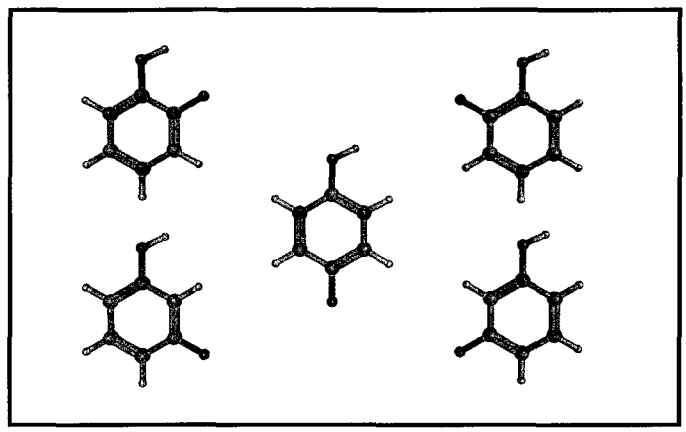

Figure 18: Possible positions for mono-fluorinated phenols Top left, *ortho* 'toward', top right, *ortho* 'away', middle, *para*, bottom left, *meta* 'toward', bottom right, *meta* 'away'

3.1.1 Training Set

Since the BDE value changes with each different substituent arrangement, BDE_{OH} values were calculated for each compound using the MLM2 method, described in Methods The differences in values, or ΔBDEs, were also calculated and these values are recorded in Table 6 for compounds that were exclusively fluorinated, chlorinated or brominated The calculated BDE value for unsubstituted phenol was 87 48 kcal/mol This agrees well with the gas-phase value of 87 2 $\pm$ 0 1 kcal/mol reported by Nix et al [15, Nix 2006]

Now we shall look at the effect of substituents added to each position, starting with the *ortho* position Although halogens are usually classified as electron withdrawing groups (EWG), when in the *ortho* position on a phenol, a halogen typically acts as a classical electron donor by lowering the BDE_{OH} This can be seen by the ΔBDE values for *ortho*-F phenol (away) of -1 76 kcal/mol, *ortho*-Cl ph (away) of -1 46 kcal/mol and *ortho*-Br ph (away) of -1 35 kcal/mol However, something changes when looking at the *ortho*-halogen (toward) conformations The ΔBDE values increase for toward *ortho*-halogenated phenols For example, *o*-F ph (toward) has

Compound	Position	F BDE	Cl BDE	Br BDE	F ΔBDE	Cl ΔBDE	Br ΔBDE
Phenol	NA	87 48	87 48	87 48	**0.00**	**0.00**	**0.00**
Ph 2 X	toward	88 46	89 24	89 38	**0.98**	**1.76**	**1.90**
Ph 2 X	away	85 72	86 02	86 13	**-1.76**	**-1.46**	**-1.35**
Ph 3 X	toward	88 31	88 49	88 44	**0.83**	**1.01**	**0.96**
Ph 3 X	away	88 48	88 48	88 40	**1.00**	**1.00**	**0.92**
Ph 4 X	NA	85 46	86 20	86 44	**-2.02**	**-1.28**	**-1.04**
Ph 2,6 X	toward	86 94	87 84	88 05	**-0.54**	**0.36**	**0.57**
Ph 2,4 X	toward	86 58	87 83	88 11	**-0.90**	**0.35**	**0.63**
Ph 2,4 X	away	83 91	84 68	84 91	**-3.57**	**-2.80**	**-2.57**
Ph 3,4 X	toward	86 06	87 08	87 25	**-1.42**	**-0.40**	**-0.23**
Ph 3,4 X	away	86 23	87 03	87 17	**-1.25**	**-0.45**	**-0.31**
Ph 3,5 X	NA	89 61	89 58	89 44	**2.13**	**2.10**	**1.96**
Ph 2,3 X	toward	- -	89 91	90 00	**- -**	**2.42**	**2.52**
Ph 2,3 X	away	- -	86 54	86 44	**- -**	**-0.94**	**-1.04**
Ph 2,6,4 X	toward	85 19	86 38	86 67	**-2.29**	**-1.10**	**-0.81**
Ph 3,5,4 X	NA	87 15	87 99	- -	**-0.33**	**0.51**	**- -**

Table 6: Calculated BDEs for phenols exclusively halogenated (F, Cl, Br). Values in kcal/mol.

a ΔBDE of 0 98 kcal/mol which is 2.74 kcal/mol larger than the away ΔBDE of -1.76 kcal/mol. This can be explained by an internal H-bond that stabilizes the phenol. This will effectively raise the BDE_{OH} since the H-bond stabilizes the parent molecule. The internal H-bond raises the BDE to such a degree that it actually reverses the trend from decreasing to increasing the BDE_{OH}. It can be seen that the OH — F internal H-bond must make up the difference between toward and away conformers. This allows us to calculate the contribution of the H-bond by difference in the two conformers. Thus, the calculation of the internal *ortho*-F H-bond was found to be 2.74 kcal/mol. Similarly, the internal H-bonds for *ortho*-Cl and *ortho*-Br are 3.21 and 3.25 kcal/mol, respectively At first it would seem odd that the most electronegative atom, fluorine, does not have the strongest H-bond. However, Baker and Kaeding observed this phenomenon in a publication in 1959. They found that the H-bonding of *ortho*-halophenols was strongest for Br, then Cl and finally F [57, Baker and Kaeding 1959].

This toward/away problem at the *ortho* position complicates calculation of substituent constants It is common in literature to see discussions only for *meta*- and *para*-substituents [14, Chandra 2002] [16, Zhang 2001] [58, Bean 2002] Instead, the constant for *ortho* is often approximated from the *para* substituent constant [59, Charton 1960]

The *meta*-halogenated phenols do not have the possibility of forming internal H-bonds Although this does make them simpler to deal with, there is still a small difference between toward and away conformations The ΔBDE value for *meta*-F phenol is +0 83 kcal/mol for toward and +1 00 kcal/mol for the away position, for a difference of 0 17 kcal/mol The phenoxyl radical formed is identical in both cases, so the difference must come from the parent phenol conformations The difference in ΔBDEs can be explained by aligning of O-H and C-X dipoles in the away position, thus stabilizing the parent and raising the BDE_{OH}

Both Cl and Br had much smaller variation between toward and away conformations, with differences of only 0 01 and 0 03 kcal/mol, respectively In each case the differences between the two *meta* conformations are small enough to permit us to assume there is only a single type of *meta*-substituent The average *meta*-substituent effects are +0 91, +1 00 and +0 94 kcal/mol for F, Cl and Br respectively In each case, the BDEs are being increased by adding a halogen to the *meta* position by approximately the same degree This is the effect that is seen from classical electron-withdrawing groups This is what was expected, as inductive effects form the primary contribution to the *meta* substituent effects and halogens are EWGs inductively

As mentioned previously, the *para* position has no H-bonding potential and only one conformation The calculated ΔBDEs for F, Cl and Br respectively are -2 02, -1 28 and -1 04 kcal/mol In each case, the halogens are lowering the BDE_{OH} in the way typical of a classical electron-donating group, which was the effect that was expected for the *para* position Halogens have EDG resonance effects and the para

position has contributions from both inductive and resonance effects

The calculated monohalogenated ΔBDE values are collected below along with the experimental values contained in Table 3 in the Introduction

Calculated	**Fluorine**	**Chlorine**	**Bromine**
ortho	-1 68	-1 42	-1 35
meta	1 02	1 02	0 97
para	-2 02	-1 39	-1 04
Experimental	**Fluorine**	**Chlorine**	**Bromine**
ortho	-1 9 (± 2)	-0 7 (± 1)	-1 7 (± 2)
meta	1 4 (± 2)	1 2 (± 1)	–
para	-1 0 (± 1)	-0 2 (± 1)	0 5 (± 1)

Table 7: Experimental and calculated ΔBDE values for monohalogenated phenols
[a] Experimental ΔBDEs obtained from [17 Santes and Simoes 1998]

A quick look shows that the major trends between the calculated and experimental values are correct. In general, *ortho* halogens reduce the BDE, *meta* halogens increase the BDE and *para* halogens decrease the BDE Bromine is the exception for the *para* substituents as it actually increases the BDE However, the trend across the halogens at this position remains the same, with F < Cl < Br Finally, the resonance donor effect of F is about twice as strong as the other halogens [60, Swain and Lupton 1968], which can be seen in the lower calculated ΔBDEs at the *ortho* and *para* positions A graph of the correlation between the calculated and experimental ΔBDE values for monohalogenated phenols shows that there is a moderate trend between the values (Figure 19)

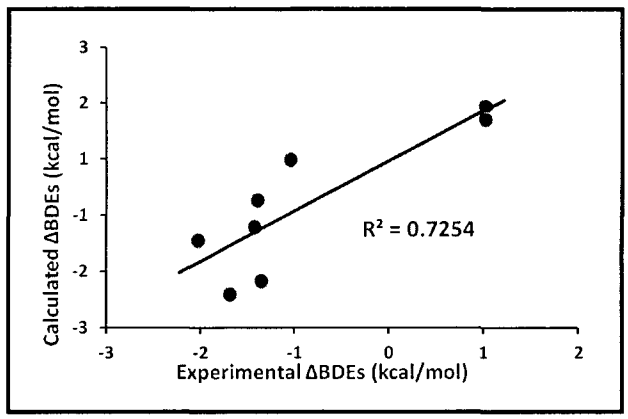

Figure 19: The correlation between experimental and calculated ΔBDE_{OH} values shows the calculated values match.

Despite minor differences in trends, all of the calculated ΔBDEs fall within the uncertainties of the experimental ΔBDEs.

A general set of substituent constants were generated using multiple linear regression, producing a different equation for each halogen type. The independent variables were # of H-bonds, # of *o*-substituents, # of *m*-substituents and # of *p*-substituents and the dependent variable was the calculated ΔBDE_{OH}. Each equation must be calibrated using a training set of compounds, again, one training set for each halogen type. The training set chosen for fluorinated phenols included each of the 5 monofluorinated compounds from Figure 18, *ortho*, *ortho-* and *meta*, *meta-* difluorophenol as well as unsubstituted phenol. This gives a total of 8 compounds in the training set where ΔBDE values were calculated for each compound using the B3LYP functional and basis sets as described in Methods. All equations were forced to have an intercept at zero to create equations that only depended on the substituent constants.

Each of the equations generated by the training sets is shown in Equations 27-29 and is also collected in Table 8 along with errors on parameters, R^2 values and MADs.

$$\Delta BDE = 2.74 * \#_{H-bonds} - 1.68 * \#_{ortho-F} + 1.02 * \#_{meta-F} - 2.02 * \#_{para-F} \quad (27)$$

$$\Delta BDE = 3.19 * \#_{H-bonds} - 1.42 * \#_{ortho-Cl} + 1.02 * \#_{meta-Cl} - 1.39 * \#_{para-Cl} \quad (28)$$

$$\Delta BDE = 3.25 * \#_{H-bonds} - 1.35 * \#_{ortho-Br} + 0.97 * \#_{meta-Br} - 1.04 * \#_{para-Br} \quad (29)$$

Halogen	Type	n	HB	ortho	meta	para	R2	MAD
Fluorine	Training	8	2 74 (0 18)	-1.68 (0.10)	1 02 (0 05)	-2 02 (0 13)	0 9957	0.07
Chlorine	Training	8	3 19 (0 06)	-1 42 (0.04)	1 02 (0 02)	-1 39 (0 03)	0 9988	0.03
Bromine	Training	8	3 25 (0 04)	-1.35 (0.02)	0 97 (0 01)	-1 04 (0 03)	0 9998	0.01
Fluorine	Expanded	14	- -	- -	- -	- -	0 9832	0.12
Chlorine	Expanded	16	- -	- -	- -	- -	0 9852	0.09
Bromine	Expanded	15	- -	- -	- -	- -	0.9687	0 14

Table 8: Parameter values and their associated errors for training and expanded sets using halogens of one type only, where n = # of data values. All values in kcal/mol. MAD = mean absolute deviation.

The equations generated from each of the training sets all fit the data extremely well, with correlation coefficients $R^2 > 0.995$. The MADs are also very small when compared with the range of the ΔBDEs contained in the training sets. The ranges of ΔBDEs for the training sets are shown (Low - High): F (-2.02 - 2.13); Cl (-1.46 - 2.10); Br (-1.35 - 1.90). Thus, the fluorine training set has a range of 4.15 kcal/mol with a MAD of 0.07 kcal/mol, which is 1.7%. The chlorine training set had a range of 3.56 kcal/mol and a MAD of 0.03 kcal/mol, or 0.8%. The bromine training set had a range of 3.25 kcal/mol and a MAD of 0.01 kcal/mol, which was 0.3%.

The good quality of correlation coefficients and the low MAD values from the equations derived from the training sets indicate that they will be able to predict[1] ΔBDE_{OH} of halogenated phenols not contained in the training set.

[1]From now on, calculated will refer to ΔBDEs calculated in Gaussian using the B3LYP protocol and predicted will refer to ΔBDEs found using the equations derived from the training sets.

3.1.2 Expanded Set

In order to test the generality of the equations and the additivity of the substituent constants, expanded sets of data were created using more substituents The parameters derived from the training sets were used to predict ΔBDEs in order to determine the additivity of the substituent constants The quality of the fit was measured by the MAD between the calculated and predicted ΔBDEs The compounds used in these tests were still exclusively halogenated, but contained combinations of dihalogenated, *o,o,p*- and *m,m,p*-substituted compounds, which can also be found in Table 6 These expanded sets contained between 14 - 16 compounds, which included the training set compounds and had expanded ΔBDE spreads of F (-3 57 - 2 13 kcal/mol), Cl (-2 80 - 2 42 kcal/mol), Br (-2 57 - 2 52 kcal/mol) Although the MAD values increased in data set to 0 15, 0 09 and 0 14 kcal/mol for F, Cl and Br respectively, they are still significantly smaller than the data spread The fluorine expanded set had a data spread of 5 70 kcal/mol and a MAD of 0 15 kcal/mol, or 2 6% The chlorine and bromine expanded sets had data spreads of 5 22 and 5 09 kcal/mol with MADS of 0 09 and 0 14 kcal/mol, or 1 7% and 2 8% respectively In addition to the excellent correlations from the training sets, with R^2 values > 0 995, this expanded test demonstrates once more that the parameters obtained from the training sets have good predictive value and additivity

3.1.3 Mixed Polyhalogenated Phenols

Having established that the training set parameters could predict ΔBDEs for the expanded compounds, we now considered a more challenging problem Three sets of mixed compounds were built containing combinations of double and triple substitutions on the phenol The first set contained mixed F and Cl substituents, the second mixed F and Br substituents and the third mixed Cl and Br substituents The

mixed double substituted phenols were comprised of combinations of *ortho/para*, *ortho/ortho*, *meta/para* and *meta/meta* substituents, with both toward and away conformations For example, *ortho/para* in the mixed F and Cl set had four conformers, *o*-F,*p*-Cl phenol (toward and away) as well as *o*-Cl,*p*-F phenol (toward and away). The triple substituted phenols were comprised of combinations of *ortho/ortho/para* and *meta/meta/para* using any two of the three halogens (F, Cl, Br) used in this work. A full listing of the mixed polyhalogenated compounds can be seen in Table 9, along with their calculated and predicted ΔBDE values.

		Calculated ΔBDE			Predicted ΔBDE		
Compound	**H-bond**	**F/Cl**	**F/Br**	**Cl/Br**	**F/Cl**	**F/Br**	**Cl/Br**
Phenol	NA	0 00	0 00	0 00	0 00	0 00	0 00
Ph 2,6 X	toward	-0 24	-0 13	- -	-0 36	-0 29	- -
Ph 2,6 X	away	0 07	0 19	- -	0 08	0 23	- -
Ph 2,4 X	toward	-0 28	-0 12	0 54	-0 33	0 02	0 72
Ph 2,4 X	away	-2 96	-2 79	-2 62	-3 07	-2 72	-2 47
Ph 4,2 X	toward	-0 26	-0 14	- -	-0 26	-0 11	- -
Ph 4,2 X	away	-3 39	-3 28	- -	-3 44	-3 37	- -
Ph 3,4 X	toward	-0 74	-0 55	- -	-0 38	-0 03	- -
Ph 3,4 X	away	-0 58	-0 40	- -	-0 38	-0 03	- -
Ph 4,3 X	toward	-1 14	-1 19	- -	-1 00	-1 05	- -
Ph 4,3 X	away	-1 15	-1 24	- -	-1 00	-1 05	- -
Ph 3,5 X	toward	2 01	1 88	1 98	2 03	1 98	1 98
Ph 3,5 X	away	2 17	2 08	2 00	2 03	1 98	1 98
Ph 2,6 X 4 Y	toward	-1 77	-1 62	-0 96	-2 01	-1 66	-0 71
Ph 2,4 X 6 Y	toward	-2 08	-1 99	-1 04	-2 38	-2 31	-0 98
Ph 2,4 X 6 Y	away	-1 80	-1 72	- -	-1 94	-1 79	- -
Ph 4 X 2,6 Y	toward	-1 61	-1 43	-0 94	-1 68	-1 46	-0 83
Ph 6 X 2,4 Y	toward	-1 28	-1 02	-0 88	-1 31	-0 82	-0 56
Ph 6 X 2,4 Y	away	-1 57	-1 34	-0 89	-1 75	-1 33	-0 63
Ph 3,5 X 4 Y	toward	0 21	0 41	0 73	0 64	0 99	0 99
Ph 3,4 X 5 Y	toward	-0 38	-0 50	0 46	0 01	-0 04	0 59
Ph 3,4 X 5 Y	away	-0 20	-0 28	0 49	0 01	-0 04	0 59
Ph 4 X 3,5 Y	toward	-0 18	-0 31	0 42	0 02	-0 09	0 54
Ph 5 X 3,4 Y	toward	0 40	0 56	- -	0 64	0 94	- -
Ph 5 X 3,4 Y	away	0 22	0 35	- -	0 64	0 94	- -

Table 9: Calculated and predicted ΔBDEs for mixed polyhalogenated (F, Cl, Br) compounds All values in kcal/mol.

The predicted ΔBDE was calculated for each compound in the mixed dataset using the previously derived parameters. The ΔBDE contribution from each type of substituent needed to be calculated and then summed together. For example, (toward) *ortho*-F,*para*-Cl phenol has one internal hydrogen bond with F, one *ortho*-F substituent and one *para*-Cl substituent. Using parameters collected in Table 8, the predicted ΔBDE is,

$$\Delta BDE = \#HBs(F)*2.74 + \#ortho(F)*-1.68 + \#para(Cl)*-1.39 = -0.33$$

This is quite close to the calculated value of -0.28 kcal/mol, with abs. dev. = 0.05 kcal/mol. The predicted ΔBDEs were calculated and stored in Table 9 along with their respective calculated ΔBDEs. The correlation between predicted and calculated ΔBDE is excellent, as seen in Figure 20.

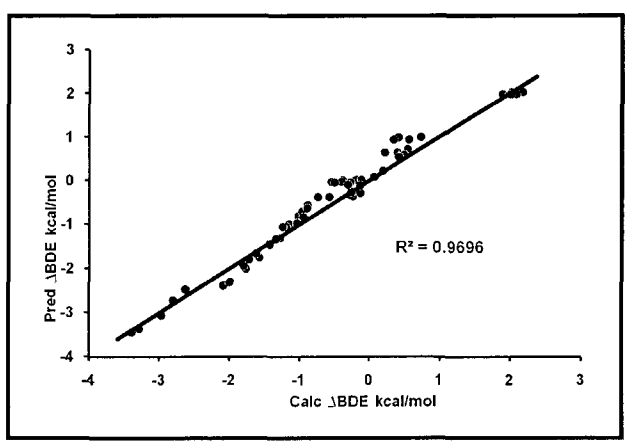

Figure 20: Correlation between predicted ΔBDE and calculated ΔBDE values for the mixed polyhalogenated phenols.

The other test for goodness of fit, the MADs, was also excellent. The MAD for the full F/Cl set, which includes the training set, is MAD = 0.18 kcal/mol which is in the order of the MADs of the training sets. The same process was performed for the F/Br and Cl/Br sets obtaining a MAD = 0.21 kcal/mol for F/Br and 0.15 kcal/mol for Cl/Br. When the MAD was calculated for the combined mixed sets, the total

MAD was 0.18 kcal/mol.

The results have demonstrated that the training set was well fitted using an MLR and the variables # of H-bonds, # of substituents of each position (*ortho*, *meta*, *para*) and type of substituent (F, Cl, Br). The predicted ΔBDE values were very well correlated with calculated values for polyhalogenated phenols when halogens are of the same type and these parameters also work even when different types of halogens are mixed together. This indicates excellent additivity of the substituent effects.

3.2 Part II. Iterative Scoring Functions

The 17 ligand training set contained the reference ligand E2, the parent compound 1a and 15 other ligands, was chosen as described in Methods and labeled Training Set 1 and the remaining 10 compounds made up Test Set 1. The docking procedure performed on each conformer gave 16,172 docked poses for the training set and 9,186 DPs for the test set. Each pose in the training set was evaluated using the London dG scoring function and the 'best pose' for each ligand was chosen as having the most negative value of London dG. These 'best poses' were selected and shown in Table 10 showing the experimental log RBA values, the calculated London dG scores as well as the four energy terms E_{int}, ΔE_L, ΔE_R and ΔG_{solv}. Having selected the top ranked DPs based on the current London dG SF, the analysis can begin. The first issue that can be seen with these poses is that the E_{int} values are significantly higher than the -70 kcal/mol range that was expected, several ΔE_L values are high (ie. 13.6 kcal/mol for 1c) and the range of ΔE_R is also high (1o is as high as 74 kcal/mol).

Ligand	log $RBA_{(expt)}$	E_{int}	ΔE_L	ΔE_R	ΔG_{solv}	London dG	SF_1
E2	2 00	-53 89	6 11	61 01	-12 40	-14 39	0 22
1a	0 17	-54 32	6 29	36 54	-10 69	-13 57	0 93
1b	0 02	-52 85	6 87	37 94	-8 09	-13 93	0 43
1c	1 44	-62 04	13 59	51 44	-10 07	-14 24	-0 54
1e	0 45	-58 09	5 77	49 79	-9 58	-13 59	0 43
1f	1 95	-65 22	4 12	50 40	8 20	-14 94	0 61
1h	-2 40	-56 73	9 97	50 02	-5 97	-14 03	0 65
1k	-0 73	-59 76	7 51	56 24	-7 43	-14 63	-0 30
1l	0 24	-56 30	7 25	49 12	-8 96	-14 42	0 14
1o	-1 19	-47 03	9 70	73 82	-8 59	-14 41	-1 41
1p	-2 10	-52 62	12 59	54 84	-11 15	-12 86	-0 60
2a	-1 21	-62 19	9 77	50 80	-11 75	-14 54	0 23
2e	-2 22	-56 94	10 84	76 86	-8 70	-15 26	-1 44
3e	-1 42	-60 49	6 50	44 74	-12 32	-14 97	0 94
3f	2 28	-57 21	8 73	40 93	-8 98	-14 69	0 28
4b	1 78	-58 88	9 17	33 51	-9 45	-14 06	0 60
4e	-0 36	-45 27	12 82	66 70	-11 67	-14 99	-1 19

Table 10: Experimental log RBA values and calculated energy descriptor set (E_{int} , ΔE_L, ΔE_R and ΔG_{solv}) for the 17 ligands in Training Set 1 Docked poses were taken from the top-ranked (most negative) values based on the London dG Scoring Function All energy values in kcal/mol

The MLR procedure was performed on Training Set 1 to generate the following SF

SF_1 (log RBA_{pred}) = 0 570 - 0 0225*E_{int} - 0 135*ΔE_L - 0 0387*ΔE_R - 0 130*ΔG_{solv}

Giving the correlation coefficient R^2 = 0.245 and std dev = 1 32 log units The correlation is very weak (see Figure 21), which is supported by the poor correlation coefficient and the large standard deviation This could indicate several things, namely that the docking method/force-field model is unable to return physically significant results, that the energy terms used in the MLR only account for part of the story or that the 'best' poses selected by the initial scoring function were not truly that great

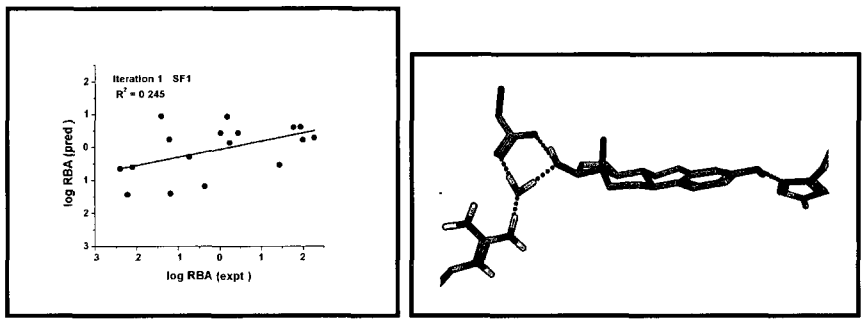

Figure 21: Initial correlation between predicted and experimental log RBA values for Training Set 1, based on an MLR over four energy terms and pose selection based on the London dG SF (left). Binding mode of E2, showing the top-ranked docked pose according to the London dG SF (right).

.The second part of Figure 21 shows the top ranked binding mode of E2 as chosen using the London dG function. This binding mode is flipped 180° from the crystal structure and so it is likely that the chosen pose is not a very good pose at all. A quick scan of the docked poses produces several other poses that appear to be significantly better visually and with much more negative E_{int} values.

The binding affinities for all DPs were calculated using SF_1, the newly generated scoring function. The top pose, or the pose with the highest predicted RBA, was selected for each ligand. The new selection of top poses is contained in Table 11.

Ligand	log $RBA_{(expt)}$	E_{int}	ΔE_L	ΔE_R	ΔG_{solv}	SF_2
E2	2 00	-68 11	2 95	26 36	-12 40	1 39
1a	0 17	-64 64	4 88	19 22	-10 69	0 22
1b	0 02	-60 47	5 30	27 23	-8 09	1 12
1c	1 44	-65 51	5 00	19 59	-10 07	0 37
1e	0 45	-66 28	4 40	23 17	-9 58	0 66
1f	1 95	-71 35	6 16	22 09	-8 20	1 18
1h	-2 40	-55 92	3 69	40 00	-5 97	1 95
1k	0 73	65 04	5 04	28 74	7 43	0 03
1l	0 24	67 41	5 13	19 48	8 96	0 80
1o	-1 19	64 29	6 82	33 61	-8 59	1 19
1p	2 10	51 41	2 92	26 35	-11 15	2 27
2a	1 21	62 64	6 64	25 69	11 75	-1 29
2e	-2 22	-61 17	7 38	35 73	8 70	2 22
3e	1 42	62 19	4 74	20 75	-12 32	0 41
3f	2 28	-75 31	5 61	17 26	-8 98	2 47
4b	1 78	-70 79	4 21	15 54	-9 45	2 09
4e	-0 36	-68 59	7 25	21 53	11 67	-0 01

Table 11: Docked poses for Training Set 1 re-ranked using the SF_1 scoring function The most positive value of SF_1 for each ligand was used to generate this table Energy values in kcal/mol

This data in the table clearly shows that there has been a general improvement in the selected 'best' poses chosen from the set of DPs E_{int} values have been lowered from the previous range of -45 to -62 kcal/mol to a more reasonable range of -51 to -75 kcal/mol The same trend appears in both the ΔE_L and ΔE_R values, where they have much smaller deformation penalties These terms were used for the MLR procedure to produce a second scoring function, SF_2, as follows

SF_2 (log RBA_{pred}) = -10 85 - 0 223*E_{int} - 0 410*ΔE_L - 0 0446*ΔE_R + 0 0462*ΔG_{solv}

The correlation coefficient obtained was R^2 = 0 853 with a std dev = 0 582 log units This is clearly a much improved fit, without major outliers (see Figure 22)

The visual check on the binding mode for the 'best' E2 pose selected based on SF_1

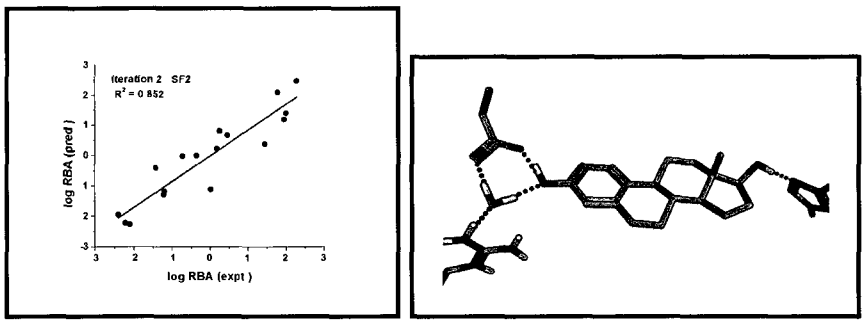

Figure 22: Correlation between the predicted log RBA calculate using SF_2 and the experimental log RBA values for Training Set 1 (left). Top-ranked binding mode of E2, as given by the top-ranked pose from SF_2 (right).

shows that the pose is oriented correctly. There is a twist in the side chain of Arg394 that makes the H-bond with E2 less than ideal, but this is a significantly improved pose compared with the previous one.

The DPs were next re-ranked using SF_2 and again an MLR was performed on the new set of 'best' poses. This was continued until the current scoring function was the same as the previous one, ie. $SF_n = SF_{n-1}$. The change of correlation coefficient R^2 over iteration number is shown in Figure 23. It can be seen that the function approaches convergence around iteration 4 and has achieved it by iteration 8.

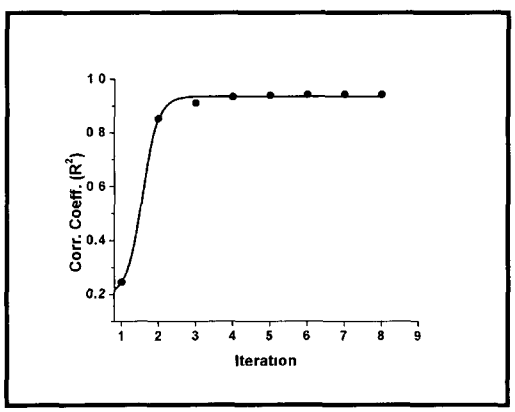

Figure 23: Change in correlation coefficient as a function of iteration number for Test Set 1.

Not only does the process converge, but it has also reached an excellent correlation of $R^2 = 0.942$ and std dev $= 0.365$ log unit. The resulting scoring function is

SF_8 (log RBA_{pred}) = -24.23 - 0.362*E_{int} - 0.413*ΔE_L + 0.0266*ΔE_R + 0.0268*ΔG_{solv}

The final set of poses was selected using SF_8 and the terms are in Table 12. It can be seen that the interaction energies are much improved, with many values more negative than -70 kcal/mol. The values of ΔE_L and ΔE_R are also in an acceptable range with all ΔE_L <10 kcal/mol and all ΔE_R <66 kcal/mol.

Ligand	log $RBA_{(expt)}$	E_{int}	ΔE_L	ΔE_R	ΔG_{solv}	SF_8
E2	2.00	-73.23	3.22	47.88	-12.40	1.92
1a	0.17	-70.99	5.63	40.63	-10.69	-0.04
1b	0.02	-71.47	8.08	50.89	-8.09	-0.53
1c	1.44	-72.26	4.68	38.08	-10.07	0.76
1e	0.45	-72.87	6.15	43.92	-9.58	0.55
1f	1.95	-73.87	4.30	52.38	-8.20	1.93
1h	-2.40	-65.29	7.82	65.32	-5.97	-2.23
1k	-0.73	-69.97	5.45	55.36	-7.43	0.15
1l	0.24	-69.55	4.19	49.39	-8.96	0.31
1o	-1.19	-66.51	7.09	62.51	-8.59	-1.62
1p	-2.10	-62.68	3.15	34.93	-11.15	-2.19
2a	-1.21	-71.28	8.66	44.48	-11.75	-1.11
2e	-2.22	-66.88	8.38	52.66	-8.70	-2.29
3e	-1.42	-69.84	6.27	36.46	-12.32	-0.87
3f	2.28	-75.24	4.30	40.58	-8.98	2.10
4b	1.78	-76.82	5.42	34.71	-9.45	2.04
4e	-0.36	-71.11	6.01	39.85	-11.67	-0.19

Table 12: Docked poses and their SF terms as selected by the SF_8 Scoring Function. Energy values in kcal/mol.

The calculated log RBA values are very close to the experimental values, as can be seen by the excellent correlation in Figure 24. There are no outliers in the fit and the values are well spread across the range of log RBA values.

The second half of Figure 24 shows that the top binding mode for E2 now contains

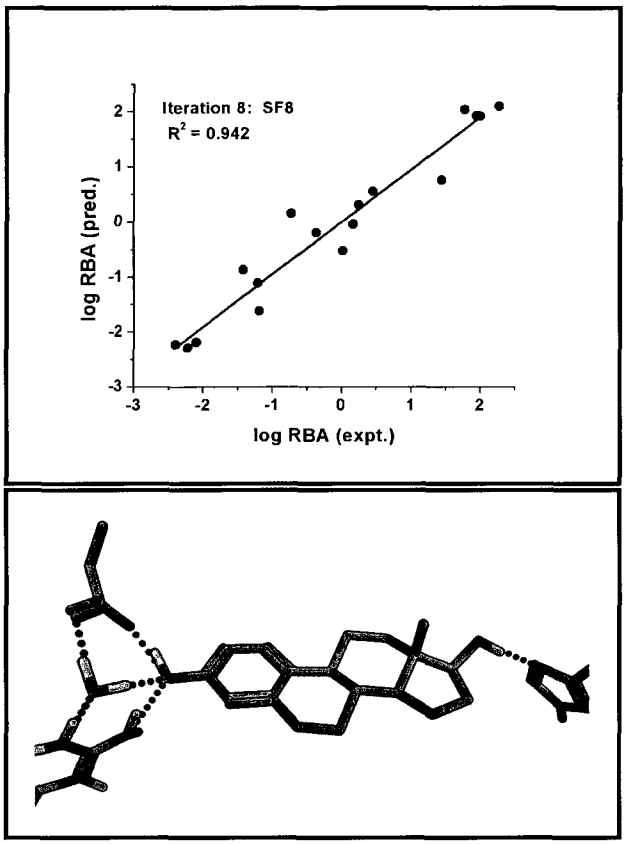

Figure 24: Correlation between the predicted log RBA calculate using SF_8 and the experimental log RBA values for Training Set 1 (left). Top-ranked binding mode of E2, as given by the top-ranked pose from SF_8 (right).

all the features that are present in the crystal structure binding mode.

3.2.1 Validation on Test Set 1

Once the final iteratively optimized scoring function was obtained, the predicted log RBA values were calculated for the docked poses of Test Set 1. The poses in Test Set 1 were re-ranked using SF_8 and the top pose was selected for each ligand. These top poses were selected and the terms can be seen in Table 13, where it can be observed that the values of log RBA(pred.) are reasonably close to those of the experimentally obtained log RBA(expt.).

Ligand	log $RBA_{(expt)}$	E_{int}	ΔE_L	ΔE_R	ΔG_{solv}	log RBA (calc.)
1d	1 72	-74 19	5 03	40 30	-9 18	1 40
1g	1 40	-66 60	4 99	56 57	-8 53	-0 88
1i	0 66	73 81	7 27	50 32	-7 13	0 66
1j	-0 42	-69 50	6 36	43 83	-7 25	0 70
1q	-1 06	-68 43	4 81	50 04	-9 19	-0 33
2f	2 40	66 49	10 25	67 16	-6 86	-2 77
3a	-0 61	-71 01	4 32	30 11	11 77	0 20
3b	2 29	-71 73	3 51	45 69	-10 19	1 25
4a	-0 38	-69 90	4 03	46 48	-11 22	0 37
4f	2 09	-77 54	4 62	41 20	8 32	2 83

Table 13: Energy terms for poses of Test Set 1 as selected by SF_8, the converged scoring function determined by Training Set 1 Energy values in kcal/mol

In addition to the small variation between log RBA(pred) and log RBA(expt), the individual terms are quite good as well For example, the values of E_{int} range from -66 to -77 kcal/mol All the values of ΔE_L are below 10 5 kcal/mol and all the ΔE_R terms are less than 68 kcal/mol

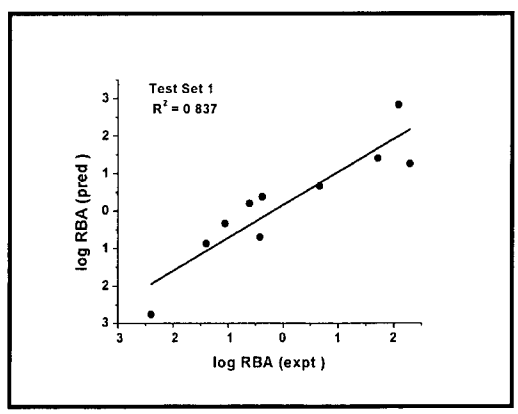

Figure 25: Calculated vs experimental log RBAs for Test Set 1, using coefficients from SF_8

The correlation between experimental and calculated log RBA values for Test Set 1 (Figure 25), although weaker than that of the training set, still has a significant

correlation coefficient $R^2 = 0.837$. This indicates that the SF developed using Training Set 1 has predictive value even among ligands not used to calibrate the scoring function.

3.2.2 Validation on Trans A-CD Compounds

Another test for the scoring function is to see how well it predicts compounds that have different structures. The *trans-* data set contains compounds that have reasonably large differences in the main structure. Although the scoring function was not trained on these compounds, it will be tested on them to see if it can predict the binding affinity of *trans-* accurately. As with Test Set 1, the DPs were ranked using SF8 and the top pose of each was retained. The top poses for each compound are contained in Table 14 and produced a moderate correlation with the experimental log RBAs as shown in Figure 26.

Ligand	log $RBA_{(expt.)}$	log RBA (calc.)
1b-t	0.23	-0.52
1c-t	0.62	0.91
1e-t	-0.33	0.43
1f-t	0.68	0.83
1i-t	-0.04	0.48
4b-t	0.99	1.57
4c-t	1.34	1.2

Table 14: Predicted binding affinites for top poses of the *trans-* A-CD set as selected by SF8.

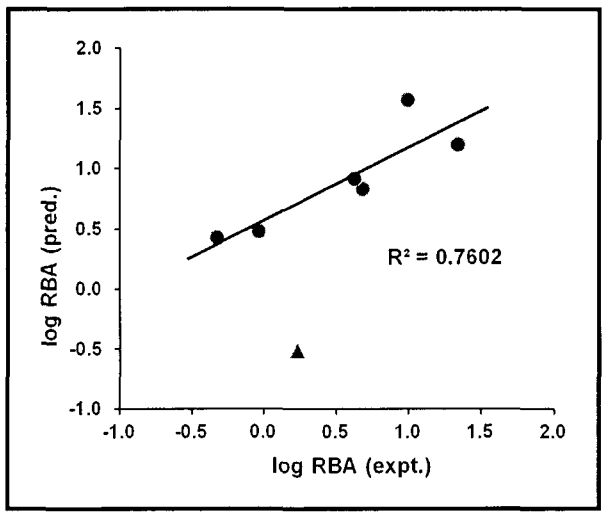

Figure 26: Calculated vs. experimental log RBAs for *trans-* A-CD set, using coefficients from SF_8.

Although the correlation was rather poor, with $R^2 = 0.45$, elimination of one outlier increased the correlation to $R^2 = 0.76$. The data range is only just under 2 kcal/mol which is rather low. Compared to the *cis-* training and test sets with data ranges of 4 - 5 kcal/mol this is a poor data spread. There are possibly two factors that account for the worse prediction of *trans-* vs *cis-* compounds. First, the *trans-* compounds may bind in significantly different ways than the *cis-* compounds. If this were the case, it would be unreasonable to assume that a SF trained on *cis-* A-CD compounds would be predictive of this new binding mode. The second possibility is that the data spread is so small that it is difficult to produce a good correlation. Improving the spread by including more experimental data might increase the correlation to make it comparable to that of Test Set 1. Unfortunately, the only way to test this hypothesis is to synthesize more compounds to fill out the set.

3.2.3 Validation Against Overfitting

There is a danger of overfitting when using iterative methods to fit experimental data. This has been mentioned by several authors, including Martin et al. [45, Martin and Sullivan 2008] and various sources referenced within. Thus, before much confidence can be put in the procedure, there should be several validation tests performed to characterize the potential downfalls of this method. One possibility is that this iterative process is fitting noise and, thus, producing a correlation which appears unrealistically good. Another possibility is that the specific training set chosen randomly happened to give a fit that worked for the test set as well.

The first possibility will be examined by randomly scrambling the experimental RBA data and performing the iterative optimization over this new data set. The scrambling was accomplished by swapping values of each ligand randomly with that of another. This method was chosen rather than assigning completely new RBA values as it gave an identical range and distribution of RBA values in the end. To test the second possibility several different training sets were generated, along with their respective test sets. This will allow us to determine if it was random chance that the chosen training set produced a fit that also gave acceptable predictions for the test set. If the iterative method works, it will be expected that there will be a certain similarity between each fitted equation generated by each test set.

Scrambled RBAs

The first test for this method was the scrambling of experimental RBAs to observe the extent of noise fitting. The resulting training set was completely uncorrelated to the initial RBA values, as shown in Figure 27.

The same iterative procedure was carried out for this scrambled set as for the previous training sets. It can be seen that the iterative procedure did not produce

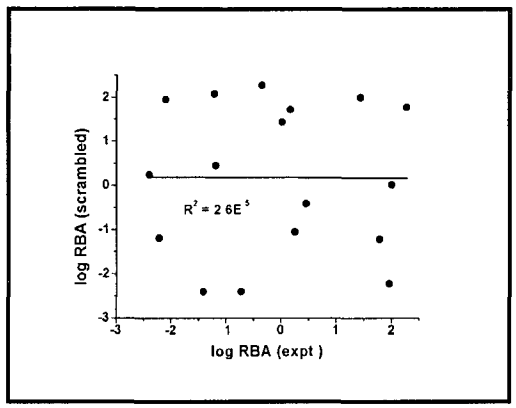

Figure 27: Correlation between scrambled log RBAs and experimental log RBAs, without a significant fit.

the same smooth convergence as in the previous examples. Instead, there were wild oscillations in the correlation coefficient over the first iterations, ending with a final cyclic convergence after about 30 iterations (Figure 28). The final correlation coefficient cycled at values with $R^2 < 0.2$.

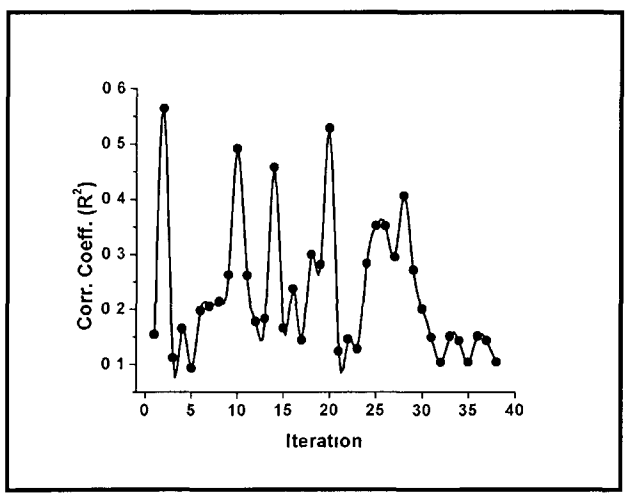

Figure 28: Value of the correlation coefficient (R^2) as a function of iteration number for the scrambled training set.

The poor quality of the correlations and the lack of smooth convergence show that the iterative scoring function treats random noise differently than it does meaningful

data. When given data with no correlation to the real binding affinities, the iterative scoring function fails to produce a significant correlation.

Additional Training and Test Sets

Three additional training sets, 2, 3 and 4, were generated using the same selection criteria as for Training Set 1. Each set contained 17 ligands with compounds from series 1, 2, 3 and 4 and with RBA values spread across low, medium and high RBA values. Due to the fact that it is our reference compound, E2 was also included in each training set. As can be seen in Figure 29, each training set converged in around 5-7 iterations. In each case, the correlation coefficient also converged around $R^2 = 0.9$ or better.

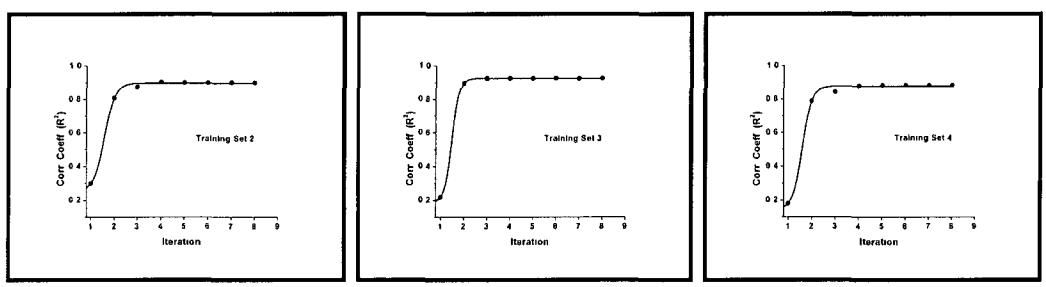

Figure 29: Convergence properties for the additional training sets, showing the change of correlation coefficient R^2 over iteration. (left) Training Set 2, (middle) Training Set 3, (right) Training Set 4.

The coefficient values for the first and last iterations of Training Sets 1, 2, 3 and 4 as well as the number of iterations required to reach convergence are contained in Table 15. The correlation obtained for the test sets as well as their composition are also contained in the table.

Training Set	Initial R^2	Final R^2	Number of iterations
1	0 245	0 942	8
2	0 300	0 900	6
3	0 219	0 925	3
4	0 181	0 883	6
Test Set	**Composition**	R^2	
1	1d,1g,1ı,1ȷ,1q,2f,3a,3b,4a,4f	0 834	none
2	1c,1f,1h,1l,1o,2a,3e,3f,4b,4e	0 786	none
3	1b,1e,1f,1l,1o 2e,3b,3e,4a,4b	0 859	none
4	1b,1c,1f,1h,1p,2a,3a,3f,4e,4f	0 927	none

Table 15: Properties of the Training and Test Sets 2, 3 and 4 are shown Data for Set 1 are included for comparison

The four training sets all converge with good consistency, with final R^2 values in the range of 0 91 $\pm$ 0 03 The convergence indicates that the scoring functions are focusing in on the best pose for each ligand Each of the test sets also produced significant correlations, with the $R^2 = 0\ 85 \pm 0\ 1$ This test has demonstrated that not only is the iterative procedure robust in its ability to obtain similar results regardless of exact training and test sets, but it is also capable of producing fits with significant predictive power (R^2 >0 78 for test sets)

Chapter 4

Conclusion

4.1 Part I. Substituent Effects

The BDE_{OH} for unsubstituted phenol was calculated to be 87.48 kcal/mol, which agrees well with the experimental value of 87.2 ± 0.1 kcal/mol.

Comparison to available literature BDE values shows that the calculated monohalogenated ΔBDEs fit within the uncertainties of the experimental values and correlate with a coefficient $R^2 = 0.73$.

The halogens also show the expected change from EWG effects in the *meta* position to EDG effects in either the *ortho* or *para* positions.

The training produced correlation coefficients $R^2 > 0.99$ and the following equations:

$$\Delta BDE = 2.74 * \#_{H-bonds} - 1.68 * \#_{ortho-F} + 1.02 * \#_{meta-F} - 2.02 * \#_{para-F}$$

$$\Delta BDE = 3.19 * \#_{H-bonds} - 1.42 * \#_{ortho-Cl} + 1.02 * \#_{meta-Cl} - 1.39 * \#_{para-Cl}$$

$$\Delta BDE = 3.25 * \#_{H-bonds} - 1.35 * \#_{ortho-Br} + 0.97 * \#_{meta-Br} - 1.04 * \#_{para-Br}$$

The substituent constants obtained were found to predict BDE_{OH} values accurately for a mixed polyhalogenated test set, producing an overall correlation coefficient of $R^2 = 0.97$ and demonstrating the additivity of the derived constants.

This work has shown that a simple MLR fit on calculated ΔBDE_{OH} values can

produce substituent constants for *ortho-*, *meta-* and *para-* halogenated compounds. These constants can be used like Hammett parameters to accurately predict ΔBDE values for arbitrarily substituted compounds. This has potential use in the development of anti-oxidants as these can be used to quickly and easily predict the BDE_{OH} of a suggested compound before synthesis begins.

This work could be improved by calculating parameters for a variety of substituents that are commonly used. This would allow prediction of ΔBDE_{OH} values for a greater number of compounds and increase the applicability of these constants. Further work in this area would involve designing promising compounds based on these constants, developing and testing them.

4.2 Part II. Iterative Scoring Functions

The iterative procedure for Training Set 1 converged smoothly and achieved and excellent correlation coefficient of $R^2 = 0.942$ with std. dev. = 0.365 log unit.

The final optimized scoring function is:

SF_8 (log $RBA_{pred.}$) = -24.23 - 0.362*E_{int} - 0.413*ΔE_L + 0.0266*ΔE_R + 0.0268*ΔG_{solv}.

This scoring function was tested on Test Set 1 and produced a correlation coefficient of $R^2 = 0.837$ which shows that the scoring function has predictive value among similar compounds not contained in the training set.

A small set of trans A-CD compounds were also predicted using SF8 and produced a correlation coefficient of $R^2 = 0.45$ ($R^2 = 0.76$ with one outlier), which was not ideal. However, the data spread was quite low for this test set and so it is hypothesized that increasing the number of data points would improve the correlation.

A test was performed to check against overfitting of the data. First, the experimental data was scrambled and the iterative procedure was performed. This optimization lacked the smooth convergence of the previous training set and produced a

poor correlation of $R^2 < 0.2$. This showed that the optimization was not overfitting scrambled data in the same way as it was real binding affinities.

Three additional training and test sets were produced to see if similar results could be obtained with different combinations of training and test compounds. In each case smooth convergence was observed resulting in excellent predictive abilities. Correlations coefficients had a range of $R^2 = 0.91 \pm 0.03$ for the training set and $R^2 = 0.85 \pm 0.1$ for the test set.

These tests have demonstrated that the iterative procedure is capable of iteratively optimizing a scoring function by optimizing the selection of docked poses until convergence is achieved. The results were excellent for cis A-CD compounds and it appears that more data may improve the fit for trans A-CD compounds.

A future direction for furthering this work would be to increase the experimental data on the trans A-CD series to better test the power of the scoring function. Using this procedure to optimize scoring functions for other receptors, such as ERβ or the androgen receptor would demonstrate the generality of this method. In future, this method could be used to test compounds not already synthesized and produce insight into promising candidates.

List of References

[1] J S Wright, D. J. Carpenter, D. J. McKay, and K. U. Ingold. *Journal of the American Chemical Society* **119**(18), 4245–4252 (1997).

[2] D. Colton, Carol; Gilbert. *Reactive Oxygen Species in Biological Systems.* Kluwer Academic Publishers (2002).

[3] S. J Klebanoff. *Science* **169**(3950), 1095–1097 (1970).

[4] B. Halliwell, G. J.M., and C. Cross. *Journal of Laboratory and Clinical Medicine* **119**(6), 598–620 (1992).

[5] E. Klein and V. Lukes. *Chemical Physics* **330**(3), 515 – 525 (2006).

[6] Z. Rappoport. *The Chemistry of Phenols* PATAI'S Chemistry of Functional Groups. John Wiley & Sons. ISBN 9780470869451 (2004).

[7] G. W. Burton and M. G. Traber. *Annual Review of Nutrition* **10**(1), 357–382 (1990).

[8] L. Frmont, L. Belguendouz, and S. Delpal. *Life Sciences* **64**(26), 2511 – 2521 (1999).

[9] I. Bjorkhem, A. Henriksson-Freyschuss, O. Breuer, U. Diczfalusy, L. Berglund, and P. Henriksson. *Arteriosclerosis, Thrombosis, and Vascular Biology* **11**(1), 15–22 (1991).

[10] A. Branen. *Journal of the American Oil Chemists' Society* **52**, 59–63. 10.1007/BF02901825 (1975).

[11] E. T. Denisov and I. V. Khudyakov. *Chemical Reviews* **87**(6), 1313–1357 (1987).

[12] B. Halliwell and J. Gutteridge. *Free radicals in biology and medicine.* Clarendon Press. ISBN 9780198552949 (1989).

[13] Q. Zhu, X.-M. Zhang, and A J. Fry. *Polymer Degradation and Stability* **57**(1), 43 – 50 (1997).

[14] A. K. Chandra and T. Uchimaru. *International Journal of Molecular Sciences* **3**(4), 407–422 (2002).

[15] M. G. D. Nix, A. L. Devine, B. Cronin, R. N. Dixon, and M. N. R. Ashfold. *Phys. Chem Chem. Phys.* **125**(13), 133318 (13 pages) (2006).

[16] H.-Y. Zhang, Y.-M. Sun, and D.-Z. Chen. *Quantitative Structure-Activity Relationships* **20**(2), 148–152 (2001).

[17] R. M. B. dos Santos and J. A. M. Simes **27**(3), 707–739 (1998).

[18] C. Hansch, A. Leo, and R W. Taft. *Chemical Reviews* **91**(2), 165–195 (1991).

[19] H. C. Brown, Y. Okamoto, and G Ham. *Journal of the American Chemical Society* **79**(8), 1906–1909 (1957).

[20] P. Mulder, H.-G. Korth, D. A. Pratt, G. A. DiLabio, L. Valgimigli, G. F. Pedulli, and K. U. Ingold. *The Journal of Physical Chemistry A* **109**(11), 2647–2655 (2005).

[21] L. Wang and A. Tang. *International Journal of Chemical Kinetics* **43**(2), 62–69 (2011).

[22] J. Shorter. *ChemInform* **31**(33), no–no (2000).

[23] T. Brinck, M. Haeberlein, and M. Jonsson. *Journal of the American Chemical Society* **119**(18), 4239–4244 (1997).

[24] D. J. V. A. dos Santos, A. S. Newton, R. Bernardino, and R. C. Guedes. *International Journal of Quantum Chemistry* **108**(4), 754–761 (2008).

[25] A. R. Leach *Molecular modelling. principles and applications.* Pearson Education. Prentice Hall. ISBN 9780582382107 (2001).

[26] A. D. Becke. *Journal of Chemical Physics* **98**(7), 5648–5652 (1993).

[27] C. Lee, W. Yang, and R. G. Parr. *Phys. Rev. B* **37**, 785–789 (1988).

[28] K. Dahlman-Wright, V. Cavailles, S. A. Fuqua, V. C. Jordan, J. A. Katzenellenbogen, K. S. Korach, A. Maggi, M. Muramatsu, M. G. Parker, and J.-k Gustafsson. *Pharmacological Reviews* **58**(4), 773–781 (2006).

[29] J. I. Macgregor and V. C. Jordan. *Pharmacological Reviews* **50**(2), 151 196 (1998).

[30] J S. Wright, H. Shadnia, J. M. Anderson, T. Durst, M. Asim, M. El-Salfiti, C. Choueiri, M. A. C. Pratt, S. C. Ruddy, R. Lau, K. E. Carlson, J. A. Katzenellenbogen, P. J. OBrien, and L. Wan. *Journal of Medicinal Chemistry* **54**(2), 433–448 (2011).

[31] M. Asim, M. El-Salfiti, Y. Qian, C. Choueiri, S. Salari, J Cheng, H. Shadnia. M. Bal, M. C. Pratt, K. E. Carlson, J. A. Katzenellenbogen, J. S. Wright, and T. Durst. *Bioorganic & Medicinal Chemistry Letters* **19**(4), 1250 – 1253 (2009).

[32] T. A. Halgren. *Journal of Computational Chemistry* **17**(5-6), 490–519 (1996).

[33] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. *Journal of Molecular Biology* **161**(2), 269 – 288 (1982).

[34] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil. *British Journal of Pharmacology* **153**(S1), S7–S26 (2008).

[35] G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head. *Journal of Medicinal Chemistry* **49**(20), 5912–5931 (2006).

[36] W. L. Jorgensen and J. Tirado-Rives. *Journal of the American Chemical Society* **110**(6), 1657–1666 (1988).

[37] S. Dapprich, I. Komromi, K. S. Byun, K. Morokuma, and M. J. Frisch. *Journal of Molecular Structure: THEOCHEM* **461-462**, 1 – 21 (1999).

[38] K. Raha and K. M. Merz. *Journal of Medicinal Chemistry* **48**(14), 4558–4575 (2005).

[39] K. M. Merz. *Journal of Chemical Theory and Computation* **6**(5), 1769–1776 (2010).

[40] S.-Y. Huang and X. Zou. *Proteins: Structure, Function, and Bioinformatics* **66**(2), 399–421 (2007).

[41] E. Stjernschantz and C. Oostenbrink. *Biophysical Journal* **98**(11), 2682 - 2691 (2010).

[42] H. F. G. Velec, H. Gohlke, and G. Klebe. *Journal of Medicinal Chemistry* **48**(20), 6296–6303 (2005).

[43] D. Hecht and G. B. Fogel. *Current Computer - Aided Drug Design* **5**(1), 56–68 (2009).

[44] S.-Y. Huang and X. Zou. *Journal of Computational Chemistry* **27**(15), 1866–1875 (2006).

[45] E. J. Martin and D. C. Sullivan. *Journal of Chemical Information and Modeling* **48**(4), 861–872 (2008).

[46] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. "Gaussian 09 Revision A.1." Gaussian Inc. Wallingford CT 2009.

[47] "Molecular operating environment (moe), version 2009.10." Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2009.

[48] G. A. Dilabio, D. A. Pratt, A. D. Lofaro, and J. S. Wright. *The Journal of Physical Chemistry A* **103**(11), 1653–1661 (1999).

[49] http://www.gaussian.com/g_tech/g_ur/k_dft.htm. "Gaussian user reference."

[50] E. R. Johnson, O. J. Clarkin, and G. A. DiLabio. *The Journal of Physical Chemistry A* **107**(46), 9953–9963 (2003).

[51] T. A. Halgren. *Journal of Computational Chemistry* **20**(7), 720–729 (1999).

[52] www.rcsb.org. "Rcsb."

[53] A. Warnmark, E. Treuter, J.-k. Gustafsson, R. E. Hubbard, A. M. Brzozowski, and A. C. W. Pike. *Journal of Biological Chemistry* **277**(24), 21862–21868 (2002).

[54] P. Prathipati and A. K. Saxena. *Journal of Chemical Information and Modeling* **46**(1), 39–51 (2006).

[55] M. De Angelis, F. Stossi, K. A. Carlson, B. S. Katzenellenbogen, and J. A. Katzenellenbogen. *Journal of Medicinal Chemistry* **48**(4), 1132–1144 (2005)

[56] P. A. Kollman and L. C. Allen. *Journal of Chemical Physics* **52**(10), 5085–5094 (1970).

[57] A. W. Baker and W. W. Kaeding. *Journal of the American Chemical Society* **81**(22), 5904–5907 (1959).

[58] G P. Bean. *Tetrahedron* **58**(50), 9941 – 9948 (2002).

[59] M. Charton. *Canadian Journal of Chemistry* **38**(12), 2493–2499 (1960).

[60] C. G. Swain and E. C. Lupton. *Journal of the American Chemical Society* **90**(16), 4328–4337 (1968).

Appendix A

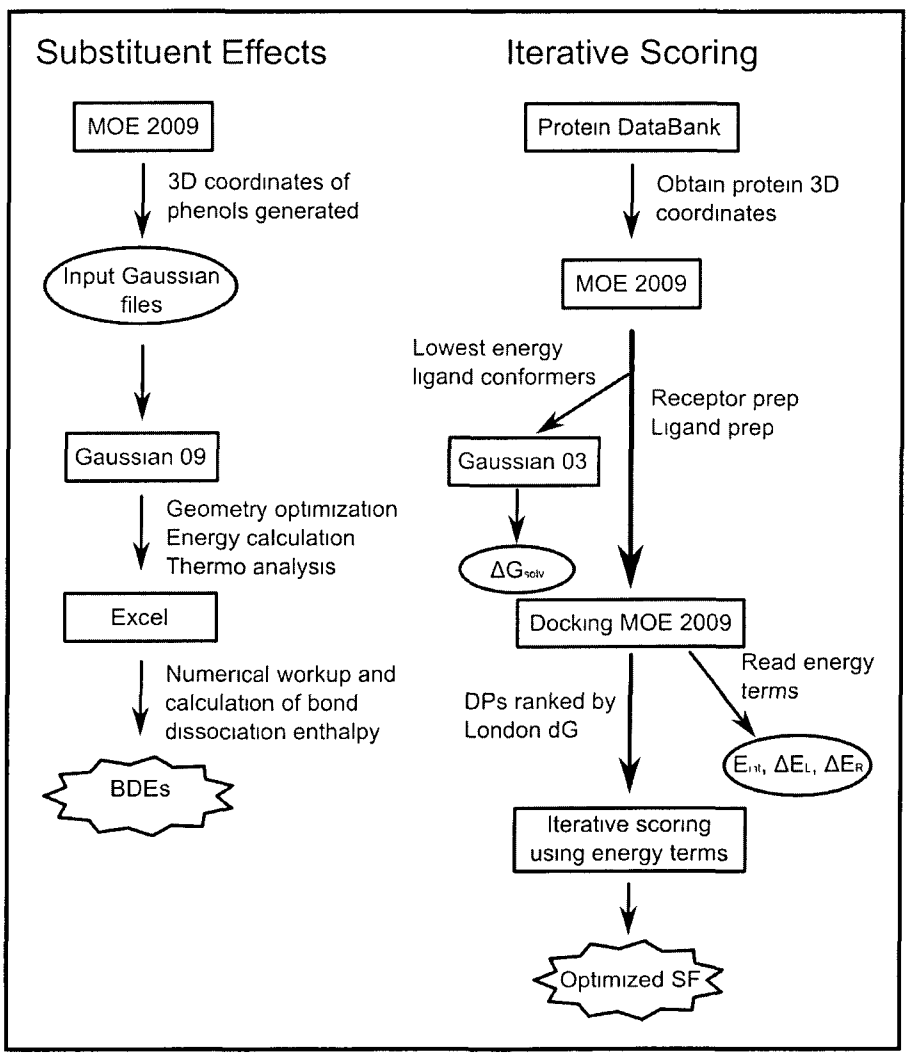

Figure 30: A flow chart showing the various steps involved in the calculations