# Reactor: An Organic Chemistry Reaction Prediction System

Christopher Michael Maloney
Department of Computer Science
Brown University
Providence, Rhode Island 02912

## Abstract

*There is an infinite set of reactions that could take place between one or more molecules, which stems from the fact that there is an infinite set of possible molecules which can react. While it is impossible to predict the exact outcome and yield of any mixture at a given point on the reaction coordinate [1], a chemist's skill is, in part, determined by how accurately he or she can predict this result. Advancements in areas such as materials science, modern medicine, and even water purification and anti-bacterial household cleaners rely heavily on the ability to predict. At the same time, prediction errors can prove very costly, and even deadly [4]. The complexity of the problem leaves novice chemists completely lost and expert chemists disoriented in an unbounded maze where solutions to synthesis problems are often more likely uncovered by accident than by design. One of the most effective applications of computers to this problem is the creation of large databases which store empirical data, so that the prediction problem can be replaced by a table lookup. This approach is strictly limited to reactions that have already been performed under specific conditions. Attempting to model a reaction in a simulator proves futile for a number of reasons, paramount of which is the fact that our ability to solve complex quantum mechanical systems is insufficient. Reactor attempts to model the mind of the chemist rather than the physical system as a whole. Using logic and heuristics that an expert would apply to a prediction problem, it is possible to solve problems just as well as, if not better than, an actual person. The advantage is that the computer has virtually unlimited memory and time to consider all possibilities, whereas a chemist might forget about a constraint and leave something out. The downside is that the predictions are based on a large and complex set of information and rules, none of which is expendable.*

[1] A measure of progress in a reaction from initial state to equilibrium which is related to time. Since the rate of the reaction is dependant on a number of conditions including external factors like temperature and pressure, a reaction coordinate is used to describe the reaction over time without specifying the length of intervals between steps.

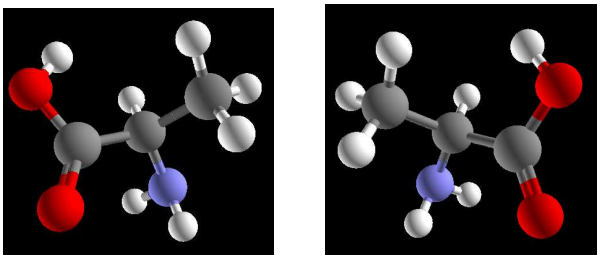

Figure 1: Stereoisomers of the amino acid Alanine. On the left, D-Alanine, and on the right, L-Alanine. They have exactly the same in chemical makeup and connectivity, but the spatial orientation of the four groups connected to the central carbon atom affect their reactivity enough to render the D isomer useless to the human body.

## 1. Introduction

There are a number of approaches to predicting the results of any reaction. The first and most reliable approach is the observation of empirical data, which requires a table lookup into a database. This approach is limited to reactions that have already been performed under specific conditions. The chemist can perform the experiment himself, but this can be extremely expensive and time-consuming. Furthermore, chemists often wish to conjecture about the results of a multitude of reaction pathways before selecting one that he or she likes best, especially when developing methods of synthesis for a target molecule. One might think that a solution to this problem is to search the database of empirical information and find the closest inexact match and assume that the results will be similar. Unfortunately, two very similar molecules may react in completely different ways. In fact, two molecules with the exact same chemical make-up and connectivity can react differently due to minute differences in three-dimensional orientation. There have been somewhat successful attempts to compare activity of biological molecules based on electrostatic three-dimensional contour maps, but their method applies primarily to non-covalent interactions and their success is greatly limited by spatial orientation [1].

A common approach to solving many problems is to build a model. Looking at empirical data, we can see that given some reactants and conditions, we end up with certain products. How does this transformation occur? It is clear that certain principles hold during the reaction process - energy and matter are conserved. We can deduce that electrons transfer from one atom to another, and these electronic changes in the molecule affect the final structure. The best model would provide a time-dependant mapping of electrons, protons, and neutrons to some position in space. Heisenberg's uncertainty principle, however, states that we cannot determine the position and momentum of a particle at any given time [2]. Quantum mechanics provides a way to describe a probability distribution for both of these properties, such that we can say exactly where a particle isn't and define an area where it could be. Nonetheless, it is possible to approximate the atomic orbitals of many-body and many-electron systems (that is to say, their probability density functions are defined). The molecular orbital (MO) theory provides information about the distribution of energy and position of one or two elections, which, in linear combination, can be used to approximate energy levels. Other systems model the electronic wave function without orbitals, but regardless of the approach most of the ab initio (quantum mechanics based) methods are limited to small molecules and require immense amounts of computational power.

The results of these calculations yield solutions to a number of properties about a static molecular electronic structure. A reaction coordinate consists of a constantly shifting electronic structure. Even on a discrete timeline, the number of calculations increases immensely with each timestep, and this number is augmented by the fact that an accurate model would incorporate large quantities of each molecule. Furthermore, reactivity is based on chaos. Molecules can only react with one another when they are spatially oriented in the correct fashion and happen to collide. Therefore, a physically accurate model would incorporate the physics of molecules moving around in space, which includes the attractive forces of each molecule on every other molecule which is an exponential explosion. Modeling the reactivity between molecules accurately falls under the Levinthal paradox - what happens almost instantaneously in nature is seemingly impossible to compute in our lifetime. In the meantime, it seems necessary to seek alternative or approximate methods to solving the prediction problem.

For the purposes of predicting reactions in this system, we will turn to the chemist. How does he or she predict the outcome of reactions? Upon examination of the wealth of data on chemical reactions, a number of trends become apparent. It is typical to categorize reactions by the process that occurs. One of the most well known examples of this type of abstraction is the acid-base reaction. An acidic

Figure 2: An acid-base reaction between hydroflouric acid and water. Water, which acts as a base in this situation, removes a proton from the acid to produce a flourine ion and water's conjugate acid, which is weaker than hydroflouric acid.

molecule will donate a hydrogen atom (a proton) to a basic molecule (proton acceptor). This transformation can be described pictorially with arrows representing the flow of electrons as in figure 2. The basic molecule contains an area of high electron density. These electrons will attract the electron-deficient hydrogen atom, which will relinquish its single electron to the acid in favor of the extra electrons on the basic molecule.

In many reactions, atoms are stripped away or added to double and triple bonds resulting in elimination or addition reactions. Parts of molecules can be replaced, resulting in substitution reactions. In some cases two parts of the molecule will simply dissociate from each other to become ions (much like salt when placed in water, where NaCl becomes $Na^+$ and $Cl^-$). The mechanisms that describe these transformations produce recognizable patterns which recur under similar conditions. The basic operation that is common to any chemical reaction is the flow of electrons from one atom to another, either intermolecular or intramolecular. Complex reactions can be described as strings of basic transformations. In modeling these reactions, it is necessary to define these basic steps and to define when they are invoked.

## 2. Mixture Analysis

In order to predict what will happen in a molecular system, it is necessary to gather as much information as possible regarding the molecules and their electronic structures. The primary focus is to determine the various regions of electron density. Each atom is comprised of a nucleus (protons and neutrons) and orbitals filled with electrons. A charged atom will contain either more electrons than protons (negative charge) or fewer electrons than protons (positive charge). Certain atoms can accommodate a charge much better than others, especially metals like tin and magnesium. Other atoms, like carbon, are extremely reactive when electrons are missing or are in excess. In determining the electronic structure of a molecule, we are primarily concerned with the valence electrons (those contained in the outer shell). Most organic atoms obey the octet rule, which implies that they are only satisfied with a total of eight electrons in this outer shell. To achieve this configuration, they will react (either

sharing, donating, or taking electrons). Every other atom in the periodic table donates an odd number of electrons to the system, but since electrons tend to operate in pairs, any atom with an odd number of electrons while bonded in a molecule (called a free radical) will react violently.

In valence bonding theory, covalent bonds are said to 'share' electrons so that both atoms can remain stable. In reality, these electrons are not always equally distributed between atoms. The shapes of the molecular bonding orbitals vary greatly as a result of the environment of a given pair of atoms. The electronegativity of atoms in specific environments provides important information regarding the distribution of electrons between any two atoms. Based on the application of empirical data to the connectivity information about the molecule [2], it is possible to identify sites of high and low electron density. For example, a simple carbon-carbon bond will not produce a noticeable dipole because each carbon has equal electron affinities. However, an oxygen-hydrogen bond is very polar because the oxygen is extremely electronegative compared to the hydrogen atom.

The general shapes of the bonding orbitals can be generalized based on their connectivity with other atoms. Carbon atoms tend to adopt hybrid orbital shapes when attached to multiple constituents. The S-shaped orbtial surrounds the atom completely (think of a single electron orbiting a hydrogen atom). The P-shaped orbitals are lobe-shaped with a node in the center, consisting of one positive and one negative lobe. When this orbital is occupied with electrons, the positive lobe grows and the negative (antibonding) orbital shrinks. If carbon is single bonded to four different constituents, it is considered to have $sp^(3)$ hybridization. If one of those bonds is replaced by a double bond (thus removing a constituent to maintain the octet rule), the amount of S-character increases, and the amount of P-character decreases, giving the bond $sp^(2)$ hybridization. Other atoms often adopt similar characteristics as their carbon neighbors, and these hybrid orbitals provide information about the reactivity of molecules.

Other sites of high electron density are identified by double bonds, triple bonds, lone pairs of electrons, and charged species (both radicals and ions). Sites of low electron density are identified by charges, electronegativity, and by areas of high acidity. Empirical data about the acid dissociation constant (pKa) of a given hydrogen atom lends information about electron density. To apply this data, a recursive neighborhood search matches templates to the area of the molecule near the hydrogen atom. Regions of high electron density are labeled as nucleophilic and regions of low electron density are labeled electrophilic in order to match standard terminology.

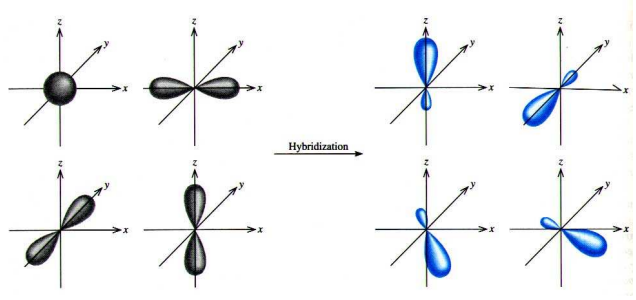


Figure 3: The bonding orbitals of an $sp^(3)$ hybridized atom consist of four orbitals (S, Px, Py, Pz), which, when summed together, produce the shapes on the right (imagine all four shapes co-located at a single origin). Image credit: http://nanotech.sc.mahidol.ac.th

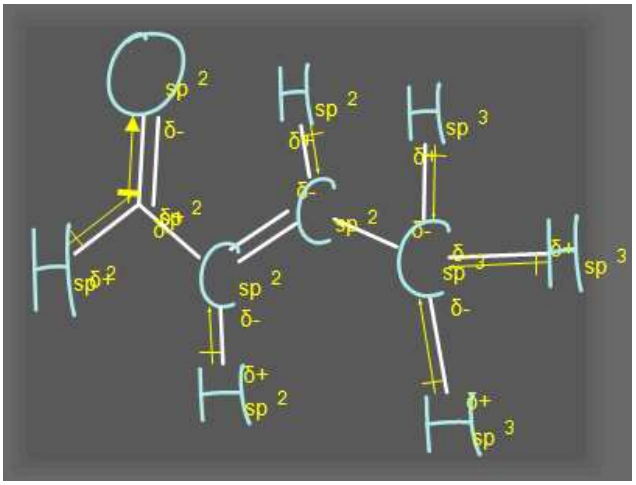


Figure 4: A molecule of 2-butenal, showing the hybridization of orbitals and the polarization of bonds (which produces partially charged dipoles).

[2] Reactor relies on the Pauling scale of electronegativity

# 3. Mechanism Pathways

Given electron density information and orbital shape data, one can produce a list of possible pathways which might occur. Each mechanistic operation comprises its own category based on its overall transformational effect on the molecule. The set of operational steps included in Reactor are proton transfer (acid-base), ionization (dissociation), electrophilic addition, nucleophilic addition, elimination, concerted nucleophilic substitution, concerted elimination, and resonance [3]. Stringing a number of steps together produces a mechanism which represents a possible reaction coordinate. For example, a unimolecular substitution reaction is comprised of a proton transfer followed by an ionization step and a nucleophilic addition step. It is important to understand, however, that each stage of this reaction is separate. At any given point, the reaction could take any number of different paths to produce different final products. It is also possible that any one of these steps is infeasible for any number of reasons. Thus, each step requires careful consideration and analysis.

The overall mechanism could require a single step to reach equilibrium, or, in more complex reactions, dozens. Many reactions produce more than one set of final products (an impure mixture) in various ratios, which requires that the mechanism branch into separate pathways. In certain instances such as the mixed-aldol reaction, this can lead to horrific mixtures of products which are nearly impossible to separate [3]. In such cases, the reaction is not viable as a synthetic method. Reactor is useful for detecting these types of situations since it will consider all possibilities given a set of reactants and conditions. Once equilibrium is achieved, the reaction is complete. In a synthesis problem which requires multiple reactions, a chemist would then introduce additional reactants to invoke another reaction.

## 3.1. Proton Transfer

This step could very well be the most common and straightforward. The epitome of an acid-base reaction, this step involves the transfer of a hydrogen atom from one molecule to another. This reaction will tend towards the production of the weaker acid. Any molecule will behave as an acid as long as there is a more basic molecule present. This situation is reversed in the presence of a stronger acid. Equilibrium lies at a pH of 7.0, which implies that the concentrations of any two molecules are in a ratio that is proportional to their relative acidities. The key observation is that this reaction is reversible. One can add acid or base to the solvent to shift the equilibrium concentrations back and forth.

[3]Rearrangement, Diels-Alder, tautomerism, concerted addition, polymer formation, aromatic, free radical (single electron transfer), and other specialized transformations (such as E1cb and certain oxidations), while important, are left to a future implementation.

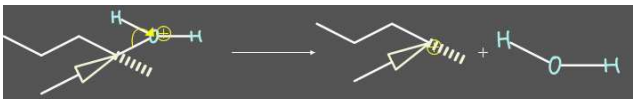

Figure 5: An ionization step. The molecule dissociates into two separate entities.

Under this logic, if the conjugate base (product of the acid-base reaction) is removed by subsequent steps of the reaction mechanism, then the reaction can proceed to completion. Reactor determines relative acidities based on empirical data as described in the mixture analysis section.

## 3.2. Ionization

Ionization requires that the products (often charged intermediates) maintain stability under the conditions of the reaction mixture. Salts ionize readily in water because water is a polar substance. A carbon atom can only withstand a positive charge for a small amount of time and under very specific conditions. Usually this species is stabilized by hyperconjugation (sharing nearby hydrogen atoms), resonance, and connectivity to other carbon atoms. A rule of thumb is that, without stabilization, the carbocation will not form. This means that the carbon atom must be attached to at least two non-hydrogen constituents. The more stable the charge, the more likely it is to form, and the more likely the reaction is to proceed. The group that detaches from the positively charged center must be a weak nucleophile. An alcohol will seldom dissociate from a carbon atom unless treated with an acid or a sulfur or phosphorus compound. The treatment part of the mechanism will take place in previous steps, but it is up to this step to determine whether the ionization is feasible.

## 3.3. Electrophilic Addition

Electrophilic addition involves the reaction of electrons in a pi-bond (double bond) with a polar reagent, comprised of both an electrophilic and nucleophilic counterpart. The electrons in the pi-bond will detach from one of the terminal atoms to seek attachment with an electrophilie such as a proton. This produces a positively charged intermediate and a weak nucleophile. The resultant nucleophile reacts in subsequent steps to produce a final product. Sometimes additional acid-base steps are required to produce a viable nucleophile.

An important consideration is the atom from which the pi-bond will detach. Most additions occur in a Markovnikov fashion, meaning that the charged species will form on the more highly substituted atom. This will result in a more stable intermediate which can react further by nucleophilic addition.

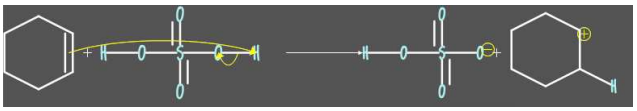

Figure 6: In this electrophilic addition transformation, electrons from a pi-bond attach to an electrophile of a polar solvent (in this case the hydrogen of sulfuric acid).

Figure 7: Nucleophilic addition involes a charged electrophile and a nucleophlie. The reaction produces a racemic mixture of stereoisomers.

## 3.4. Nucleophilic Addition

This is the final step of both addition and unimolecular substitution reactions. It requires the presence of a positively charged atom and a nucleophilic species. This reaction competes with the elimination step, because a weak nucleophile will react with neighboring hydrogen atoms in an acid-base step before attaching to the carbon atom. Even a strong nucleophile is subject to these conditions, although in smaller yields. The temperature of the solution is very important when determining the most likely pathway, because warmer temperatures favor elimination. The neighborhood of the positive center also affects the yield, because atoms bonded to fewer substrates are less hindered and less stable.

An important concern associated with this type of reaction is the stereochemistry of the final product. Stereochemistry refers to the three-dimensional orientation of substrates attached to an atom (see introduction). When four or more different groups are attached to a single atom, there are two non-superimposable versions of the molecule which are mirror images of each other. While the difference may appear subtle, this phenomenon often results in a significant difference in reactivity. A nucleophilic addition reaction will produce a racemic mixture of product, which implies equal ratios of each isomer. The logic behind this result is the fact that a positively charged carbon atom, bonded only to three substituents, will adopt a planar structure with two unbonded lobes, one on either side of the plane. The nucleophile has an equal probability of colliding with either side. The concerted mechanisms described below will produce similar transformations to those ending in a nucleophilic addition, but the stereochemistry of the products will differ (the reactions are stereospecific).

## 3.5. Elimination

The elimination step requires a base and a positively charged species which is beta to a hydrogen atom (meaning the hydrogen atom is separated by two bonds). The base will remove the beta hydrogen (which is often slightly positively charged due to hyperconjugation) leaving the electrons behind to bond with the positively charged center, thus stabilizing the molecule. As discussed above, the presence of a nucleophile and a positively charged center can result in nucleophilic addition. When either nucleophile is strongly basic, the carbocation hindered, or the temperatures very hot, however, elimination is favored.

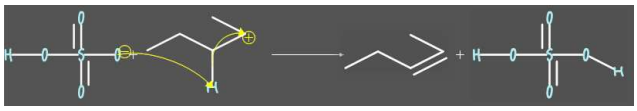

Figure 8: In this elimination, the base removes a hydrogen located beta (two bonds away from) the charged carbocation. In this case, the most highly substituted (Saytzeff) product is formed since the reaction is under thermodynamic control.

Elimination in this fashion (through a positively charged intermediate) will result in a mixture of products. Substitution products aside, there is usually more than one beta hydrogen atom which is a candidate for removal. This particular mechanism is regulated by thermodynamic control, meaning the more highly substituted (and thus more stable) product is more likely to form in a higher yield.

## 3.6. Concerted Nucleophilic Substitution

The pathway for this reaction is more complex than the previous pathways. It essentially performs two transformations in a single step. Instead of splitting the ionization and nucleophilic addition steps into two separate entities, both occur at the same time. The environment of the carbon atom which undergoes substitution is important, because a nucleophile will not have access to a hindered (highly substituted) atom center. This reaction competes with elimination at high temperatures, especially if the nucleophile is a strong base. If the nucleophile is weak, then the unimolecular pathways may proceed after ionization. A lot of this depends on the nature of the carbocation as well, namely its stability in the solvent.

Concerted substitution reactions are sometimes preferred over the unimolecular counterpart because the stereochemical outcome of the reaction is predictable. Regardless of other factors, this pathway will produce the inversion of the reactant's original stereochemical orientation. The substituents attached to the substituted atom will adopt a planar structure as the nucleophile approaches, and then invert once the attachment occurs and return to a tetrahedral configuration.

Figure 9: In this concerted nucleophilic substitution process, the leaving group dissociates from the carbon atom while the bromine ion attaches to the backside, resulting in inverted stereochemistry.

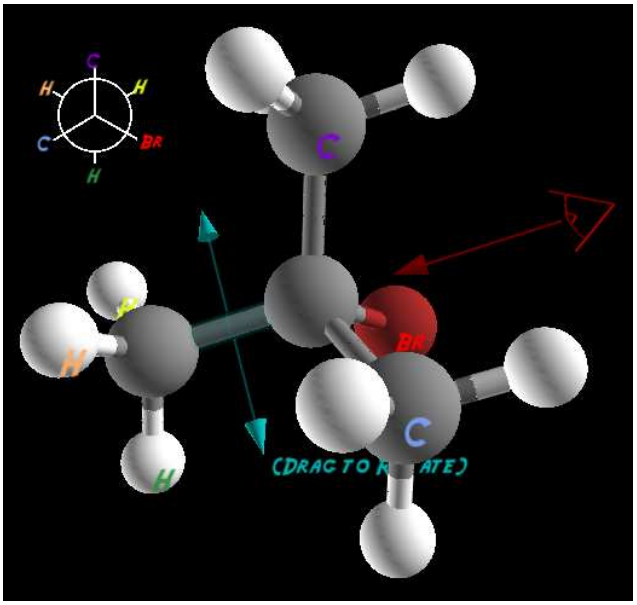

Figure 10: Shown here is the anti-periplanar configuration of a hydrogen atom (peach label) with a bromine leaving group (red label), a requirement for concerted elimination reactions.

## 3.7. Concerted Elimination

Much like concerted substitution, concerted elimination combines the ionization of a leaving group and the elimination steps into a single transformation. This reaction will only occur under specific conditions, however. A base is required to remove a proton which is beta to the carbon atom, only the hydrogen must have proper orientation in this case, because the electron transfer and dissociation of the leaving group occurs simultaneously with the removal. Specifically, the hydrogen must adopt an antiperiplanar [4] orientation to the leaving group. In a cyclic compound, these two substituents must be trans and diaxial. This orientation information is retrieved from a three-dimensional molecular mechanics simulation which provides conformational cues about chemical structure.

Other conditions are important to consider as well, such as the solvent and temperature in order to prevent side-reactions. At the moment, these considerations are left to the chemist since he or she determines the initial conditions of the reaction. Reactor will consider all side reactions as possible pathways.

[4] looking down the axis of the bond between the carbon and the alpha carbon, the leaving group and hydrogen must be collinear in the projection plane.

## 3.8. Resonance

Resonance structures provide a large amount of stability for charged species. When a charge is present on an atom, electrons can shift around the molecule to delocalize the electron density. This behavior is only possible in certain conditions where the octet rule is maintained (disregarding charge that is already present) and the electron movements are feasible. The procedure is to shift lone pairs and electrons in pi-bonds to neighboring atoms and to count formal charge on each atom. Important contributors to the set of resonance structures are those with similar charges on atoms of the same type. If one can generate two or more resonance structures, then the overall structure is considered a hybrid of the entire set. Thus, any of the structures can react in subsequent steps. The ones that are most likely to react are those which fit the criteria of the proposed transformation.

## 3.9. Reaction Prediction

Listed above is a set of mechanistic operations that can occur under certain conditions. Understanding the characteristics of each transformation is important to understanding reactivity. The difficulty lies in determining which of these steps is likely to proceed at any given point in the reaction coordinate, since they are by no means mutually exclusive. Resonance and branching must be considered at every stage. Sometimes further input from the user is required to proceed with any level of confidence (such as temperature), and at other times the solution is relatively straightforward (choosing between a set of acid-base reactions). Sometimes it will be necessary to draw on past knowledge of reaction rates. The important point is that given all the initial conditions and empirical data resources, the computer can heuristically make the same choices as the chemist, and ideally the computer would assist the chemist to ensure that errors are avoided in complex situations.

## 3.10. Future Work

While the mechanistic transformations described in this paper can describe a large number of organic reactions, there remain a number of specialized transformations that have not yet been implemented due to time constraints. A perfect system would require the consideration of all possible pathways.

The current system provides an aid for determining the mechanism by which a set of reactants will interact to provide final products. Once the prediction accuracy of this

system reaches the level of a real chemist, the next step would be to provide suggestions for the reagents that should be added to a solution to achieve a desired result. In other words, given some reactants and some final products, what reagents would invoke the correct mechanism to render the overall transformation possible?

Given a directed graph of starting reactants, reagents, and outcomes, it would then be possible to provide as input a final product and a database of starting materials with their associated costs, and attempt to solve the shortest path of reactions (i.e. cheapest cost) to producing that product. There are usually many ways to produce a single molecule, and the best method is unclear until all possibilities are considered. It is likely that unexplored pathways exist which could improve upon existing methods of synthesis.

## 3.11. Conclusion

The problem of prediction dates back to the earliest days of alchemy, even before the true nature of atoms and bonds was discovered. Thousands of years of accumulated knowledge make it possible for a chemist to propose transformations which will synthesize new or pre-existing molecules. Only recently has humanity developed the computational capacity to assist in this daunting process, but without a basic understanding of chemical principles, these tools are useless. As we develop better hardware and learn more about the quantum mechanical nature of molecule-molecule interaction, computers prove invaluable in the advancement of chemistry, especially in the fields of medicine and nanotechnology.

By implementing problem solving skills on a computer, the goal is to allow the chemist to focus more on creativity and less on specific details. Hopefully computational chemistry will advance to a state where it can help to avoid errors and help to provide new ways of fighting disease and preventing environmental disaster.

## 3.12. Acknowledgements

I would like to thank Andy van Dam for all of his support and encouragement over the last few years while I conducted this research. His faith in me was often the driving force that kept my wheels rolling late into the night and into the next morning. I would also like to thank Matthew Zimmt for sharing his immense amount of chemical knowledge, and for helping me to understand the limitations in the domain of computational chemistry thus allowing me to create something practical and beneficial within my lifetime. Both of these professors challenged my thought process to its extreme limits, and I would not have made it this far without them.

# References

[1] Hugo Kubinyi. Comparative molecular field analysis (comfa). *BASF AG*, D-67056, 1998.

[2] John P. Lowe. *Quantum Chemistry*. Academic Press, New York, NY, USA, second edition edition, 1993.

[3] Thomas N. Sorrell. *Organic Chemistry*. University Science Books, Sausalito, CA, 1999.

[4] Wikipedia. Qin shi huang — wikipedia, the free encyclopedia, 2008. [Online; accessed 16-May-2008].