Generation of Conformations for Sketched Molecule Diagrams

Dana Tenneson

April 2007

Abstract

In chemistry, molecules are drawn on paper and chalkboards as diagrams of lines, letters, and symbols which represent not only the atoms and bonds in the molecule but encode cues to the 3D geometry of the molecule. Recent efforts into pen-based input methods for chemistry software have made progress at allowing chemists to input 2D diagrams of molecules into a computer simply by drawing them on a digitizer tablet. However, the task of interpreting these parsed sketches into proper 3D models is largely unsolved due to the difficulty in making the models satisfy both the natural properties of molecules and the structural cues made explicit in the drawing. This thesis proposal presents a framework for solving this problem via augmenting molecular mechanics equations to include drawing-based constraints. Common drawing-based constraints are identified and algorithms are addressed for building molecule models using these equations and leveraging the structure cues.

1 Overview

1.1 Chemistry sketches as a 2D Language

Molecules are three-dimensional (3D) structures that chemists need to depict on two-dimensional (2D) paper for purposes of communication. As such, a standard notation exists for drawing molecule diagrams which conveys a great deal of 3D information to a chemist with expert knowledge about molecular structure. This 2D notation is commonplace in chemistry classrooms, laboratories, and publications. It provides a standardized means of conveying information about molecule shapes the same way architects use drafting techniques to convey information about buildings.

1.2 The Goal of Generating 3D Models from Inked Structures

This thesis proposal revolves around the task of taking molecule sketches made on a computer using a Tablet PC or similar digital inking technology and interpreting them into 3D models that correspond to the sketch. Traditionally, when chemists need to create 3D models of molecules, they use 3D molecular modeling programs. These programs allow for direct manipulation of atoms and bonds in a 3D model using a standard point-and-click GUI and are time consuming to use. The speed of this process inhibits quick "cocktail napkin" style brainstorming. The 3D modeling programs also require a certain level of specific expertise with the software to use correctly. Being able to draw a molecule using the same notation the chemist is already accustomed to and getting the correct 3D shape would remove this obstacle. The automatic generation of 3D structure from molecule sketches would also allow for digital lab book initiatives such as SmartTea (m. c. schraefel *et al.*, 2004),(Butler, 2005) to automatically store structure data of performed experiments in databases for search purposes. This database structure data could then be searched by techniques such as (Clark *et al.*, 1994). New techniques for chemistry interfaces using interactive drawing to create both 2D and 3D molecule models have potential value to a wide range of educational and industrial chemistry software packages.

1.3 Related Work

1.3.1 Interpreting Sketches in Other Domains

Several efforts have been made into pen-based interfaces for input of information in domains which have standard 2D notations. Circuit diagrams (Edwards & Chandran, 2000),(Gennari *et al.*, 2005), mathematics (LaViola & Zeleznik, 2004), (Labahn *et al.*, 2006), music (Forsberg *et al.*, 1998), to name a few. Most similar to this task are interpreting sketches into 3D shapes for architecture ((Zeleznik *et al.*, 1996), Google's SketchUp) and models ((Igarashi *et al.*, 1999), (Karpenko & Hughes, 2006), (Nealen *et al.*, 2005)). All of these examples leverage some amount of knowledge of the relevant domain to interpret a given sketch. Likewise, this thesis proposal specifically addresses a means of using chemistry domain knowledge for the task of understanding and generating 3D structures of sketched molecules.

1.3.2 ChemOffice and Other Commercial Diagram-based Model Builders

CambridgeSoft's ChemOffice software suite contains the closest equivalent we can find to solving the problem of interpreting 3D structures from 2D sketches. Their ChemDraw program is a industry standard for creating molecule diagrams suitable for print. Using a traditional point-and-click-based interface, chemists can create typeset versions of their diagrams. Chem3D is their popular 3D modeling tool which contains a somewhat obscure interface for 3D modeling using the ChemDraw interface. A plugin linking ChemDraw to Chem3D allows the user to use the ChemDraw interface as a starting point for creating 3D structures in Chem3D. The user thereby need not start the molecule construction process from scratch, but receives a model with the correct atoms and connectivity. However, the user still needs to tweak this model and manually

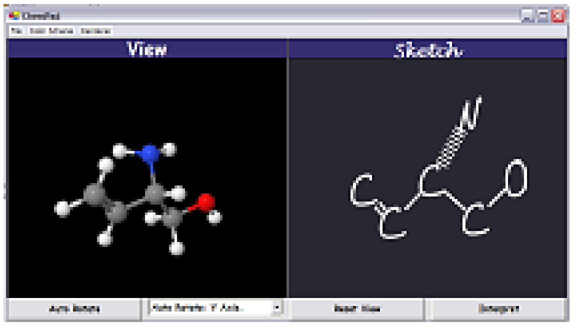


Figure 1: The interface for ChemPad 2004-2006.

perform the tasks this thesis proposes to solve regarding conformational search. Essentially, the user needs to ensure the 3D model matches the expectations put forth by the structure diagram. Similar capabilities can be found in ACD's ChemSketch and Accelrys' Sketcher and Converter programs.

1.3.3 ChemPad

An early version of this work, ChemPad (Tenneson & Becker, 2005) (shown in Figure 1) was started in 2004 . ChemPad used the Fluid Inking (Zeleznik & Miller, 2006), (Zeleznik *et al.*, 2004) ink gesture library to define single-stroke gestures for components of chemistry sketches which were close to the standard notations. For instance, drawing a "Cl" for Chlorine consisted of drawing a cursive "C" and "l" without lifting the pen. Building the 3D model was performed by an algorithm which viewed atoms and bonds as if they are plastic model kit pieces and fit them together based directly on geometrical cues found in the sketch. This scheme for molecule model construction, its benefits, and its shortcomings are described later in this proposal.

1.3.4 New Research In Molecule Sketch Input

Since the initial release of ChemPad, two new research efforts have started to address the task of ink-based input for molecule drawings. (Ouyang & Davis, 2006) interprets inked molecule drawings into cdxml (ChemDraw XML) files. This work batch processes full molecule drawings without requiring the gesture-based input found in ChemPad, but lacks 2D interactivity and 3D model generation. (Bryfczynski, 2006) accepts ink input of molecules with an interface suitable for chemistry students in an introductory course. Additionally, (Tenneson & Maloney, 2007) is a new ink interface for ChemPad which relaxes the single-stroke gesture requirement and makes several more improvements to the inking interface. This provides an interface which requires less explicit training for chemists to use.

1.4 Thesis Focus

While a number of interesting research questions exist regarding user interfaces for molecule sketches, this thesis proposal will primarily be focused on the task of generating 3D models, or conformations, for an already understood molecule sketch. The ongoing work on the 2D input aspect of ChemPad is primarily being conducted by Christopher Maloney as part of his master's thesis.

2 Conformation Generation

Conformation generation is the choosing of a 3D location for every atom in a molecule. In actuality, molecules are not rigid 3D structures, but atoms in motion held in certain shapes by the forces between atom electrons and protons. Small molecules may transition between only a few preferred states, but the number of possibilities grows exponentially as the molecule size increases. It is, however, still useful to think of molecules as having specific structures because some conformations are much lower in energy, and therefore more likely to exist at any particular moment, than others. Chemists prefer to think of molecules as assuming the conformations which minimize the molecule's energy since the molecules are close to those structures most of the time.

2.1 Preprocessing for Conformation Generation

Before starting the conformation generation process for a given molecule sketch, the drawing's 2D structure must be parsed and interpreted into a connectivity graph for the molecule. This graph contains the atoms which are part of the molecule, the bonds that exist within the molecule, the 2D locations of the components of the drawing, and the notational cues (such as wedges, dashes, charges, etc.) drawn onto the drawing, but lacks the 3D coordinates of the atoms and any knowledge of the forces between the atoms. This step can be completed with an inking input system such as (Tenneson, 2005a), (Tenneson & Maloney, 2007),(Ouyang & Davis, 2006), (Bryfczynski, 2006) or even be side-stepped using a point-and-click interface such as that found in ChemDraw and many other commercial chemistry applications. This graph can then be used to assign molecular mechanics atom types (symbols) to the atoms which can be completed using a system such as (Wang *et al.*, 2006). With this information gathered, the conformation generation systems described in this section can be completed.

2.2 Modeling Kits

A student with a plastic ball-and-stick molecule modeling kit can easily generate a conformation for a given molecule by plugging pieces together. Modeling kits can accurately depict many molecules' preferred conformations accurately because atoms tend to prefer to bond in specific 3D shapes. These shapes are dependant on the atom in question, the number of neighboring atoms, and the

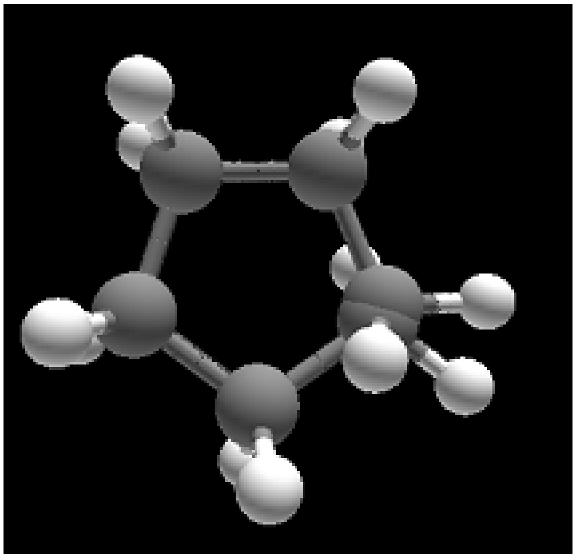


Figure 2: A crude interpretation of cyclohexane taking the "straight" bonds in the diagram as being coplanar.

order of the bonds to those atoms. The student is then responsible for using their knowledge of chemistry to assemble these pieces into reasonable conformations. Admittedly, this can be difficult for a beginning student since the model kit does not provide any feedback representing the energy content of torsional and steric strains in conformations.

2.3 Heuristics for Interpreting Sketches into 3D Shape

By leveraging the power of the modeling kit on a computer, a conformation generation algorithm can be derived. Given a 3D template for each piece in a modeling kit and a table of when to use each piece, conformation generation can be performed by connecting templates based on heuristics dependent of the way the molecule was drawn. Starting with the carbon backbone of an organic molecule, atoms connected by regular "straight" bond notations can be added in such as way as to ensure local planarity (anti, or syn periplanar conformations). Wedge and Dash bonds indicate deviations from the local plane and atoms connected by them are connected to the appropriate outbound connection on the atom template. Details of this template and heuristic based algorithm appear in (Tenneson, 2005a).

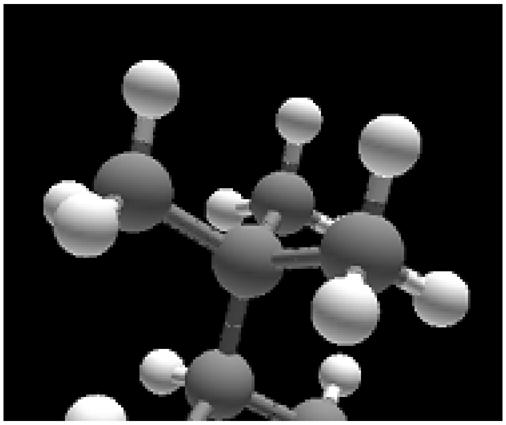


Figure 3: A t-butyl functional group with the carbons properly aligned to maximize parallelity.

2.4 Shortcomings of Heuristic-based Model Building

Unfortunately, building molecule models in this fashion overlooks three important factors in the shape of molecules. First, there is no check to prevent interpenetration of atom nuclei. A simple example of this is cyclohexane. Cyclohexane consists of six carbons bonded to each other in a ring and each carbon is attached to 2 hydrogen atoms. The standard way to draw cyclohexane is as a hexagon - a shape which implies that the carbons are coplanar. However, when six tetrahedral carbons (with angles between bonds of about 109 degrees) are placed in together in a planar ring, two of the carbons interpenetrate as seen in Figure 2. This is highly improbable because of the natural repulsion of positively charged nuclei. Having a simple interpenetration check wouldn't solve the problem either since nuclei can be pushed together a little by other structure constraints.

The second problem with the template and heuristic method is that torsional angles between atoms are ignored. This problem exists in plastic model kits as well as computer modeling and chemistry students are taught to build their models to minimize torsional strain. While a well-drawn molecule diagram does a good job of showing the anti-periplanar nature of a carbon backbone, side branches lack information about this constraint. A simple example here is a t-butyl functional group (a carbon attached to three other carbons). In this example, the hydrogen and carbon atoms in the t-butyl group should align themselves to maximize parallelity amongst themselves and with other atoms in the molecule as shown in Figure 3. However, without drawn cues to show this, all torsonial angles look similarly good.

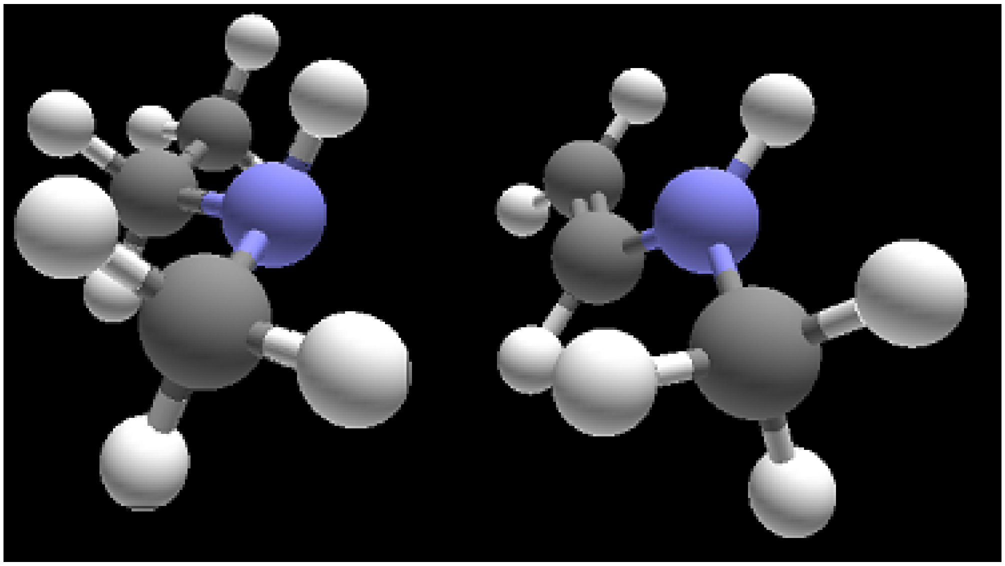


Figure 4: The nitrogen in both molecules are connected to two carbons and a hydrogen. The one on the left is in a pyramidal shape, while the one on the right has a planar shape due to the resonance with the double-bonded carbon neighbor.

The third problem with the template and heuristic method is that the above mentioned rules for choosing an atom template have exceptions. Foremost in the exceptions is dealing with changes due to resonance. For example, a nitrogen atom bonded to three other atoms with single bonds normally takes a pyramidal shape. However, if it is adjacent to a carbon participating in a double bond to a different atom as shown in Figure 4, electrons will be shared between the double bond and the nitrogen changing the nitrogen's shape to trigonal-planar. While accounting for any one of these rule exceptions is not difficult, determining and compensating for all such rules exceptions is a separate research task.

2.5 Using Molecular Mechanics and Gradient Descent to Improve Generated Models

Attempting to make molecule models adhere to physical rules is not a new problem for computational chemists. Molecular mechanics attempts to simplify quantum mechanics principles into molecule energy formulas that can be calculated relatively quickly on a computer. While models exist for computing quantum mechanics calculations, they require vastly more time to complete and are therefore unsuitable for use in a user interface technique. A primer on molecular mechanics can be found in Appendix B

Once the energy of a molecule conformation can be calculated using molecular mechanics, creating a "correct" conformation is a matter of minimizing the conformation energy. The energy function is defined over the domain $\mathbb{R}^{3*N}$ where N is the number of atoms in the molecule. Due to the size of the domain, global minimization is a very difficult problem. Traditionally, conjugate-gradient and annealing algorithms have proven most effective at solving these problems (NIH, n.d.) although (Wang, 2000) makes an interesting use of Branch and Bound after discretizing the solution space.

Because molecular mechanics energy formulas are approximations, there are different equations, or force fields, used for approximating different types of molecules. For instance the AMBER force field (Ponder & Case, 2003) (Weiner *et al.*, 1984) is designed for approximating energies of proteins, while the GAFF force field (Wang *et al.*, 2004) modifies AMBER to approximate energies of simple organic compounds. For the purposes of this thesis proposal, the GAFF force field is used, but the techniques described generalize to other force fields.

2.6 Why This Loses Handwriting Cues

Taking a model built by the template and heuristic method and performing a molecular mechanics optimization upon it does generate a conformation that is chemically feasible and satisfies the problems described earlier. Assuming a global energy minima is found, a cyclohexane molecule is popped out of its planar conformation and into the near-planar "chair" conformation. t-butyl groups align their members properly to minimize torsional constrains. Resonance shape exceptions are accounted for in the force field and corrected.

Unfortunately, a number of new problems can arise during minimization. Because molecular mechanics has no concept of handwritten cues, certain structure requirements detected by the handwriting heuristics can be lost during optimization. The most catastrophic of these losses can actually result in the wrong molecule being generated by the system. The two molecules pictured in Figure 5 are carvone molecules which have the same formula and the same connectivity, but different 3D structures. Molecular mechanics considers these molecules to be equally optimal and either solution is acceptable to a minimization task. The molecules are different however. The one on the right is the molecule responsible for the taste of spearmint while the molecule on the left is the one responsible for the taste of caraway. Because differences in molecules such as these are not detectable in the connectivity of the molecule, chemists use special notations in their sketches which convey the correct 3D shape.

Another class of problems that arises after optimizing conformations is that of handwritten notations which purposely depict molecules in non-optimal conformations. Since molecules are constantly changing, it is often useful to draw a molecule in a higher energy conformation that is chemically relevant, but occur less often than the optimal one. For example, a carbon backbone can be drawn with an eclipsed conformation to show a transition state necessary for a reaction to occur. A molecule drawn as such would be forced into the optimal anti-periplanar conformation despite the chemist's intention.

Similarly, chemists sometimes think of molecules in non-optimal conformations when there are alternate conformations which are close to the optimal in energy and convey and fit more readily into existing mental frameworks. For example, in GAFF a 2-butanol molecule has an optimal energy when the oxygen is anti-periplanar to the ethyl end of the carbon backbone. However, chemists prefer to think of the methyl end assuming the anti-periplanar position. A molecule sketch disambiguates the intention of the chemist, but molecular mechanics optimization will always pick the optimal energy.

Additionally, sometimes a molecule sketch will be made from an alternate view angle or even multiple view angles as shown in the pictures of Cineole and Diosgenin in Figure 6. Therefore, these sketches use perspective cues to indicate 3D structure which could not be interpreted using the most naïve understanding of molecule sketch view angles.

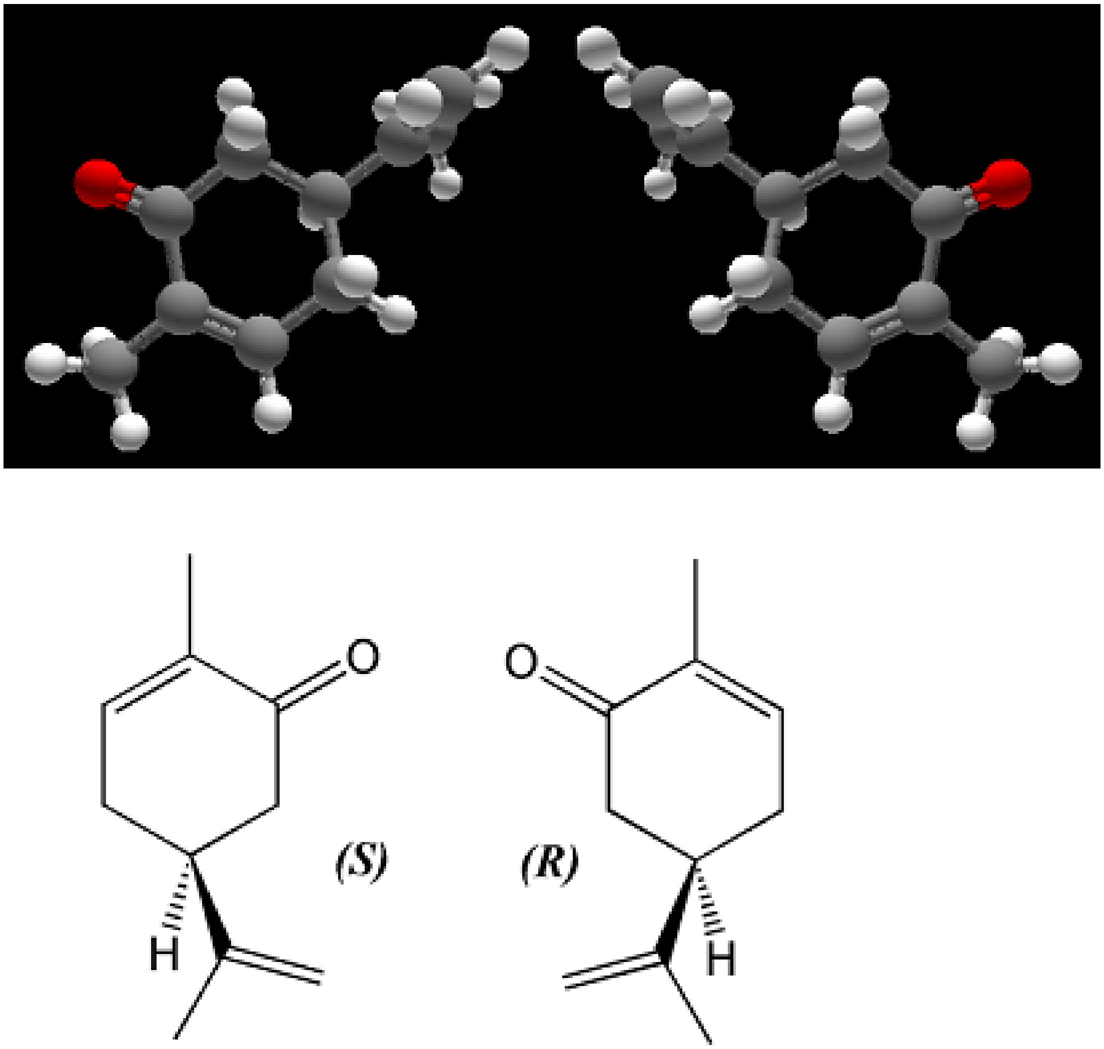
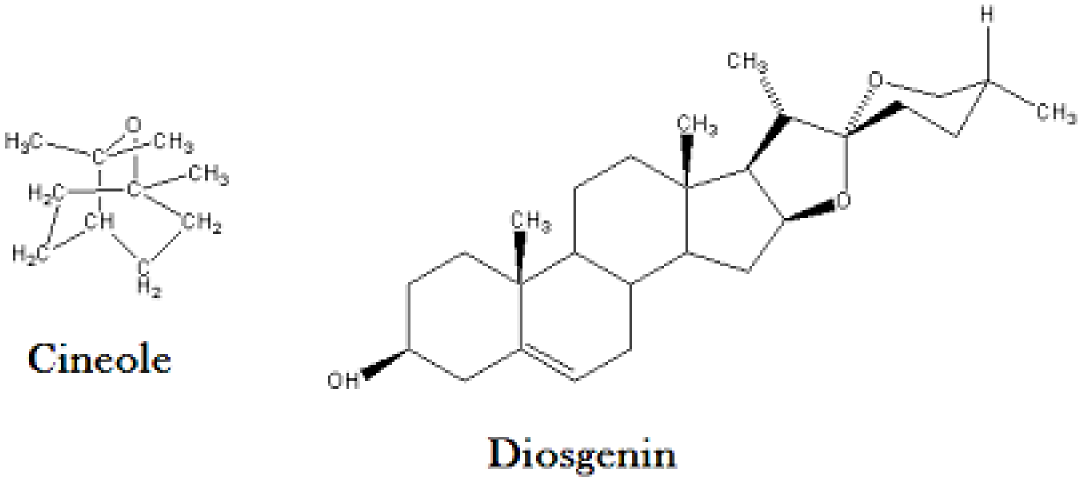


Figure 5: Two versions of the carvone molecule. S-carvone is the smell and flavor of caraway. R-carvone is the smell and flavor of spearmint.



Cineole

Diosgenin

Figure 6: Typeset diagrams of the cineole and diosgenin molecules. The cineole molecule is drawn from the side showing an explicit "boat" structure to the ring. The diosgenin molecule contains several rings drawn from the "top-down" view and a ring at the right drawn from the side indicating its relationship with the neighboring ring and the equatorial position of its methyl group.

2.7 Adding Molecular Mechanics Terms to Compensate for Diagram Structure Cues

The primary proposed contribution of this thesis is the augmenting of a force field to contain additional information about the handwriting-based constraints that should guide minimization. This augmented force field will make it possible for an optimization algorithm to (i) reject conformations which are chemically incorrect and (ii) accept conformations which are chemically suboptimal, but which are the intention of the user and explicitly marked as such in the drawing.

Molecular Mechanics force fields are assembled as large sums of terms each of which penalize specific energetically unlikely relationships. These sums are detailed in Appendix B. The proposed technique for modification of the force field is not to change any of the existing terms which determine chemical feasibility, but to add additional terms which penalize relationships that contradict the drawing. These additional terms must adhere to two constraints to ensure that existing gradient-based algorithms can still be used for optimization. First, the energy function of the entire force field must remain continuous, so that the force field has a gradient. Second, the gradient of the augmentation terms must have a known analytical solution so that optimization can be performed quickly.

Since the augmentation terms are essentially "tacked-on" to the end of an existing molecular mechanics force field, this framework for combining chemical constraints with drawing constraints should generalize to all force fields and to drawing notations beyond the scope of this thesis. For example, this framework could be extended to protein sketching by using a protein-based force field and by defining the notational cues specific to proteins in molecular mechanics terms. The calculations would become much more difficult and therefore would not be feasible for current hardware, but the potential exists.

2.7.1 Preventing "Incorrect" Models

The types of errors an energy optimization can make during conformation generation with a standard force field fall under the general classification of stereochemistry errors. In these cases, some structural constraint in the molecule prevents one conformation from changing to another low energy conformation without breaking bonds thereby creating a chemical difference between the conformations.

Our first stereochemistry error arises due to the naturally tetrahedral shape atoms such as carbon assume when bonded to four different molecule substructures. There are two different ways to attach these four different substructures that are not rotationally equivalent in 3D (right-handed vs. left-handed). The atoms with four different substructures are called stereocenters and their chirality is denoted as being either R (Rectus - clockwise) or S (Sinister - counterclockwise). In the drawing notation, at least one bond at a stereocenter is drawn

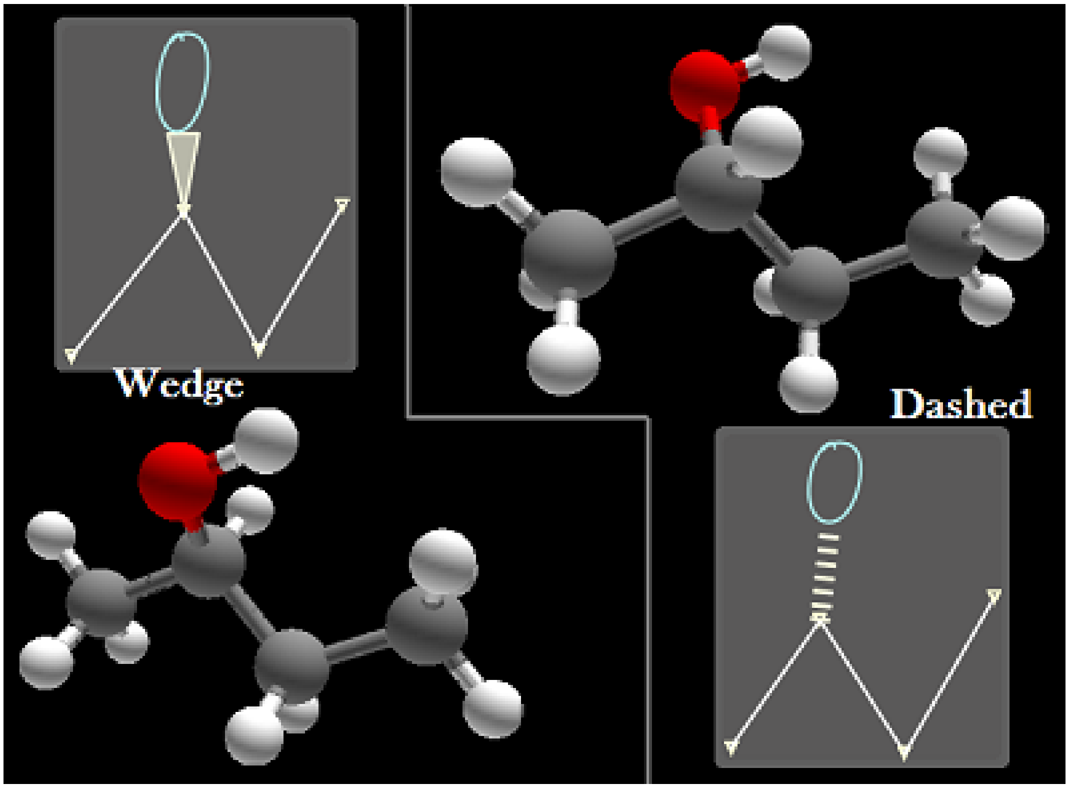


Figure 7: Effects of the wedge and dashed notation at a stereocenter. The wedge bond on the left pushes the oxygen forward while the dashed bond on the left pushes the oxygen back.

with a special notation to denote the proper 3D structure. A wedge bond would show the connected atom as being in front of the stereocenter while a dashed bond would show the connected atom as being behind the stereocenter. Wedge and dashed bond examples drawn in ChemPad are shown in Figure 7.

Another classification of stereochemistry errors comes from the inability of molecules to rotate around double bonds. Normally, molecules can rotate around their single bonds if struck with enough energy without breaking the bond. However, the sharing of electrons in a π-bond found in double (and higher order) bonds prevents this kind of rotation. Therefore, the relative arrangement of atoms connected to the atoms participating in the π-bond is fixed. In this case, no special notation is needed in the molecule drawing, the arrangement of the atoms is directly depicted by the notation. A correct conformation must preserve the direction of the turns made by these atoms with the atoms participating in the π-bond.

Our final classification of stereochemistry errors occur in cyclic molecules where atoms are bonded in a ring which prevents completely free and independent rotation about a one σ-bond (single bond) of the ring. While cyclic molecules are not necessarily held rigid in their shape and many, such as cyclohexane, can twist slightly to change between energetically feasible conformations, there are still relationships between atoms connected to the ring which must remain in all such conformations. In particular, atoms connected to the ring can be defined as being on the same side (cis) or opposite sides (trans) of the 'plane' of the ring and will remain so through conformational change. As opposed to the previous stereochemistry terms, this relationship is not defined over a single four atom region of the molecule, but involves a region bounded only by the size of the ring.

While this set does not encompass the entire range of stereochemistry errors that can occur during conformation generation, it does cover an important subset. These three stereochemistry differences are taught early in organic chemistry courses because of their relevance to basic organic molecules.

2.7.2 Understanding Notational Intentions

Less critical than the stereochemistry errors above where a molecule chemically different from the intended one is created, conformational errors can occur where the molecule is chemically correct, but not in the conformation that was drawn. These discrepancies arise primarily from differences between the shape the molecule takes in the 2D drawing and the shape which would minimize the energy in 3D.

Skeletal drawings of organic molecules show intended torsional angles through the direction of turns formed in the drawing. Bonds which are drawn close to

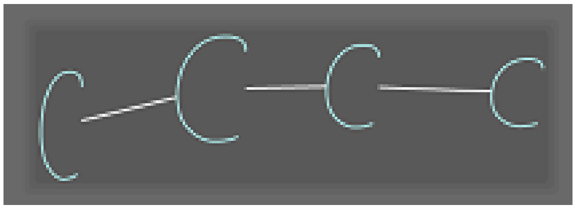


Figure 8: A drawing of butane such as a beginner chemistry student user might make. Here all three atom angle turns are in the same direction and could be interpreted as syn-periplanar even though the carbons are drawn in a mostly straight line.

each other in the molecule and which have been drawn as parallel are usually intended to be parallel in the final structure. Conversely, when bonds are drawn as forming two consecutive turns in the same direction, this conveys an intention to form a syn-periplanar torsion - a local energy maximum. This simplistic dichotomy of planar options at torsions breaks down upon the introduction of rings. Since drawings of rings are locally considered to be viewed from “above the ring” which makes the ring members appear planar when they usually are not, bond angle directions can convey alternative interpretations such as gauche torsions, or axial and equatorial positioning of side branches.

The development of force field terms to solve problems of this nature is significantly different from the terms that solve the aforementioned stereochemistry problems. Whereas with stereochemistry problems, the terms needed to penalize some number of geometric alternates so that their energies are no longer as low as the specified conformation, the terms for these notational intentions need to penalize all minima to prevent those conformations. New energy minima must be created where there were none previously and those minima must adhere to the cues placed in the drawing.

There is also a question as to the degree of certainty in the intent of the diagram. For instance, a novice chemistry student might draw diagram such as Figure8 where a strict analysis of the turn directions would indicate a syn-periplanar confomation. However, it seems much more likely that the alignment was made in ignorance and repeated attempts to draw the diagram would result in the conformation being chosen apparently by chance.

Also within this classification is the problem of drawings for rings and other structures made from alternate or multiple viewing angles. While perhaps the most common way to draw cyclohexane is as a hexagon, drawing it from a “side” viewing angle allows the chemist to denote a “chair” or “boat” conformation as shown in Figure 9. These side viewing angles also allow for explicit designation

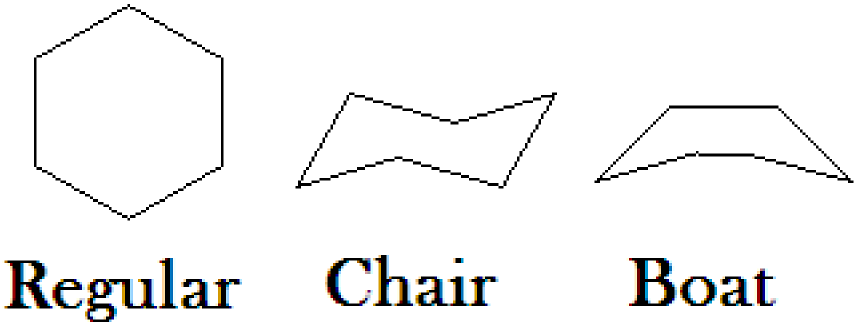


Figure 9: Three diagrams of cyclohexane. The chair and boat angles depict specific conformations of the molecule.

of axial and equatorial connectivity of atoms adjacent to the ring without adding new notation conventions in a way not possible with the overhead view.

3 Example of Augmentation Term

Of the force field augmentations needed, we have developed potential solutions for the R/S and Z/E stereochemistry terms. The Z/E term is presented here for insight into the nature of the problem.

As previously explained, the nature of electrons shared in a double bond prevents the atoms at the end of that bond from freely rotating. If we use CIP rules (explained in Appendix A) to order the atoms adjacent to those on each side of the double bond the high priority atoms can either be on the same side or opposite sides of the line passing through the double bond. The two stereoisomers are displayed in Figure 10 and have comparable energies in an unmodified force field.

3.1 Definitions of Useful Terms

Here we will define a new force field term in terms of four atoms. The atoms at the ends of the double bond are termed A and B. Similarly, neighbors of A which are not B and neighbors of B which are not A are termed A' and B' respectively. Figure 11 shows one of the four selections of A, A', B, and B' for a given double bond. Instead of defining this relationship for only the highest priority A' and B', the term is summed for all four permutations.

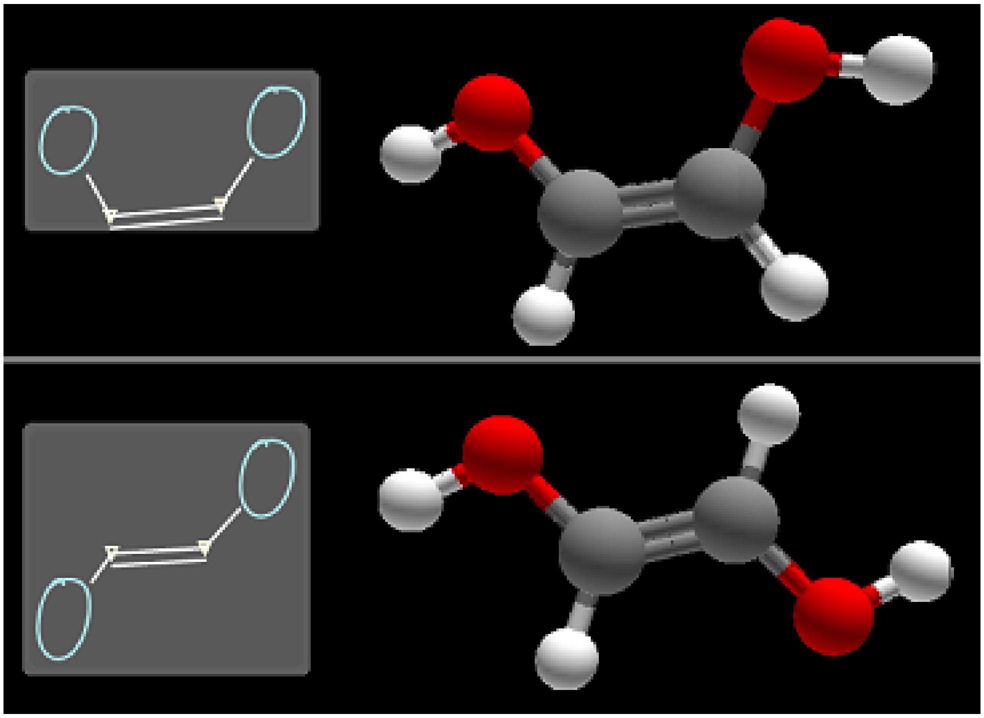


Figure 10: The conformation chosen, in particular the placement of the oxygen atoms, depends on the configuration drawn in the diagram.

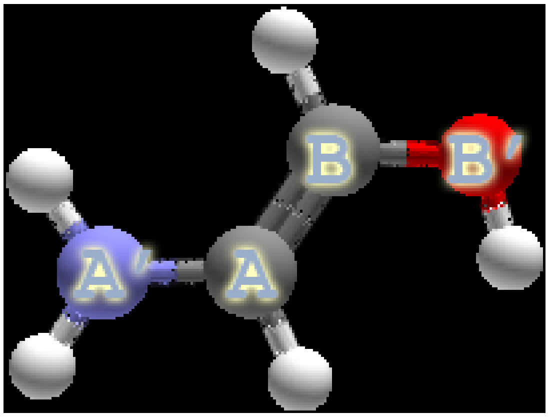


Figure 11: The labeling of atoms for the Z/E stereochemistry terms.

$$\vec{CROSSA} = (\frac{\vec{AB}}{\|AB\|} \times \frac{\vec{AA'}}{\|AA'\|})$$

$$\vec{CROSSB} = (\frac{\vec{BA}}{\|BA\|} \times \frac{\vec{BB'}}{\|BB'\|})$$

Figure 12: Definitions of $\vec{CROSSA}$ and $\vec{CROSSB}$

$$NRG_{ZE} = \begin{cases} K * (\vec{CROSSA} \cdot \vec{CROSSB})^2 & \text{if } \vec{CROSSA} \cdot \vec{CROSSB} \text{ has the wrong sign.} \\ 0 & \text{otherwise.} \end{cases}$$

Figure 13: Z/E Penalty Term

In Figure 12 we define two vectors, $\vec{CROSSA}$ and $\vec{CROSSB}$ which will be useful for determining if the 3D conformation does not fit the 2D drawing. When $\vec{CROSSA} \cdot \vec{CROSSB}$ is positive, A' and B' are on opposite sides of the $\overline{AB}$ line. Conversely, the $\vec{CROSSA} \cdot \vec{CROSSB}$ is negative, A' and B' are on the same side of the $\overline{AB}$ line. To determine whether a penalty term should be added then requires us to look at the original molecule drawing and check the turn directions (left or right) of $\angle A'AB$ and $\angle ABB'$. If the turn directions are the same, A' and B' are on the same side of the bond line. If they are different, they are on opposite sides of the bond line.

3.2 Formulation of Term

Once we have detected that a penalty should be applied, $\vec{CROSSA}$ and $\vec{CROSSB}$ are again useful for determining the magnitude of the penalty. Figure 13 shows the value of the penalty. Here K is a constant used to control the magnitude of the penalty and NRG_{ZE} (Z/E Energy) is the penalty. Intuitively, $\vec{CROSSA}$ and $\vec{CROSSB}$ are normals to the planes formed by the atoms in the double bond and A' and B' respectively. Since $\vec{CROSSA}$ and $\vec{CROSSB}$ are normalized, their dot product is the cosine of the angle between the vectors. When a given atom such as say B' is on the plane dividing the sides, the normals are perpendicular and the dot product is zero. As B' moves to one side or the other, the dot product increases until A, A', B, and B' are co planar and the dot product is one. A cross-section of the energy function is shown in Figure 14.

This differs from the use of constrained minimizations in molecular mechanics packages because we are not holding a specific feature such as a torsional angle fixed, but allow the atoms to move within the confines of preserving the configuration. As such, we do not need to set hard constraints which may prove to be inaccurate. So long as the configuration is correct, there is no penalty

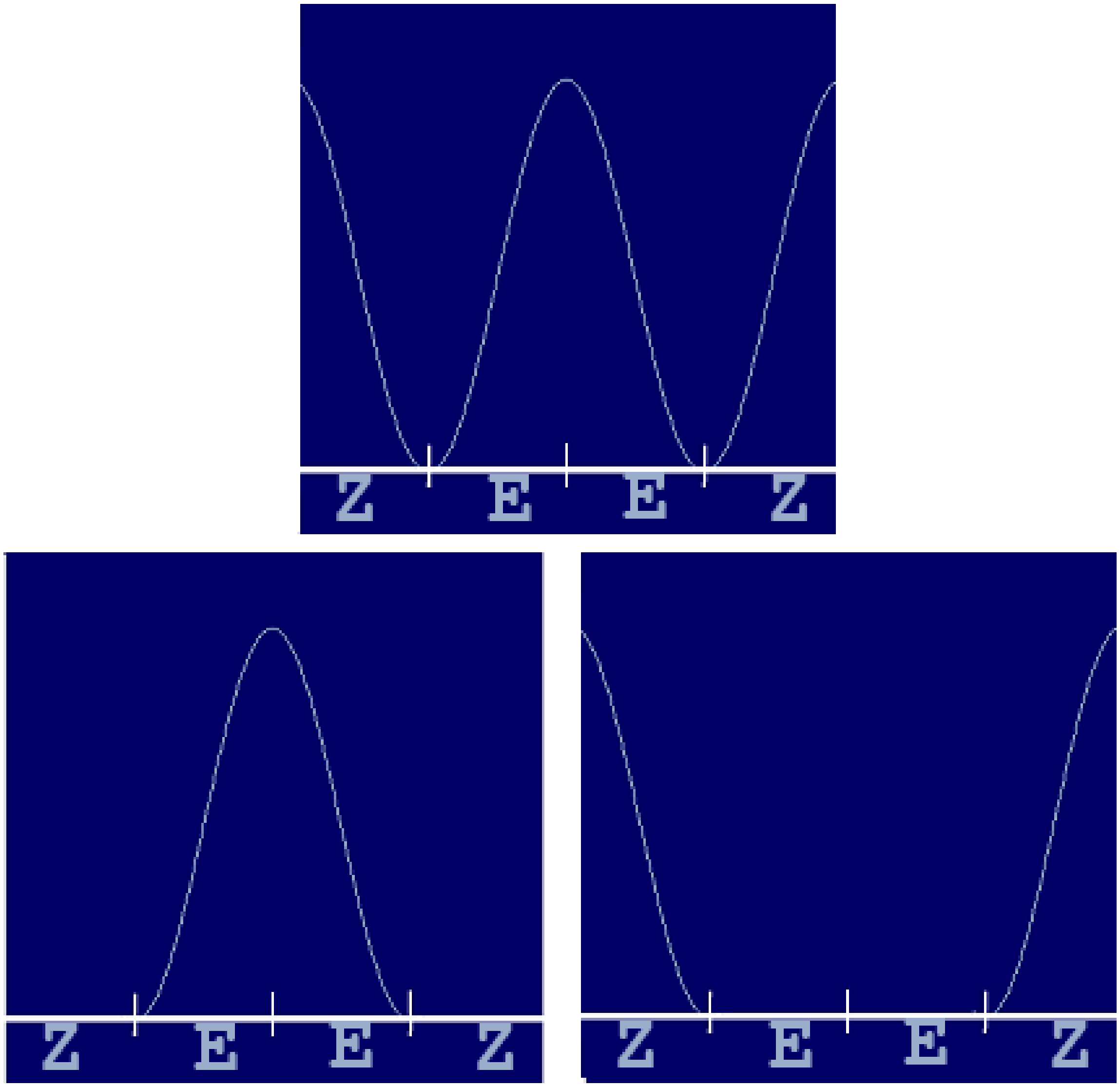


Figure 14: Variations of the Z/E penalty term value as B' is rotated around the bond axis. The graph on the top shows the function $\cos^2\theta$. The bottom left and right graphs show our Z and E (respectively) variants of the equation. Here the energy is zero when B' is on the correct side of the plane and increases as it gets farther onto the wrong side.

value added to the energy function in the minimization.

3.3 Derivative of Term

The partial derivatives of NRG_{ZE} with respect to the x, y, and z coordinates of A, A', B, and B' can be calculated directly from the formula and appear below. It is important to note that our formulation satisfies the constraint of continuity because at the points where the function itself is discontinuous, i.e. when A' or B' is on the dividing plane, both the function and the derivative are zero. For the purposes of this analysis, we are only considering the derivative at points where the $NRG_{ZE} > 0$.

The derivative of NRG_{ZE} with respect to the x,y, and z coordinates of A,B,A', and B' is as follows. Let $DP = (\vec{CROSSA} \cdot \vec{CROSSB})$. The derivative $\frac{dNRG_{ZE}}{dDP} = 2K * DP * dDP$ is a common multiplier for all component derivatives. Then,

$$\frac{d}{dA'}DP = [\frac{d}{dA'}(\vec{CROSSA}) \cdot \vec{CROSSB}] + [\vec{CROSSA} \cdot \frac{d}{dA'}(\vec{CROSSB})]$$

3.3.1 Derivatives for A' and B'

Since A' does not appear in $\vec{CROSSB}$, we know that the right bracketed term is 0 and

$$\frac{d}{dA'}DP = [\frac{d}{dA'}(\vec{CROSSA}) \cdot \vec{CROSSB}]$$

Expanding the derivative of the normalized $\vec{AA'}$ vector we get that

$$\begin{aligned}\frac{d}{dA'}(\frac{\vec{AA'}}{\|AA'\|}) &= \vec{AA'} \cdot \frac{d}{dA'}(\frac{1}{\|AA'\|}) + \frac{1}{\|AA'\|}\frac{d}{dA'}(\vec{AA'}) \\ &= \vec{AA'} \cdot (\frac{-1 * \frac{d}{dA'}(\|AA'\|)}{\|AA'\|^2}) + \frac{1}{\|AA'\|}\frac{d}{dA'}(\vec{AA'}) \\ &= \frac{1}{\|AA'\|}(\frac{d}{dA'}(\vec{AA'}) - \frac{\vec{AA'}}{\|AA'\|} * \frac{d}{dA'}(\|AA'\|))\end{aligned}$$

At this point, we need to consider the x,y,and z components individually.

$$\frac{d}{dA'_x}(\vec{AA'}) = (1,0,0)$$

$$\begin{aligned}\frac{d}{dA'_x}(\|AA'\|) &= \frac{d}{dA'_x}((A'_x - A_x)^2 + (A'_y - A_y)^2 + (A'_z - A_z)^2)^{\frac{1}{2}} \\ &= \frac{1}{2\|AA'\|} * 2(A'_x - A_x) = \frac{A'_x - A_x}{\|AA'\|}\end{aligned}$$

The derivatives with respect to y and z follow naturally. In total,

$$\frac{d}{dA'_x}DP = [\frac{\vec{AB}}{\|AB\|} \times [\frac{1}{\|AA'\|}((1,0,0) - \frac{\vec{AA'}}{\|AA'\|} \cdot \frac{A'_x - A_x}{\|AA'\|})]] \cdot [\frac{\vec{BA}}{\|BA\|} \times \frac{\vec{BB'}}{\|BB'\|}]$$

Similarly, for B'

$$\frac{d}{dB'}DP = \vec{CROSSA} \cdot \frac{d}{dB'}(\vec{CROSSB})$$

$$\frac{d}{dB'}(\vec{CROSSB}) = \frac{\vec{BA}}{\|BA\|} \times \frac{d}{dB'}(\frac{\vec{BB'}}{\|BB'\|})$$

$$\frac{d}{dB'}(\frac{\vec{BB'}}{\|BB'\|}) = \frac{1}{\|BB'\|}(\frac{d}{dB'}(\vec{BB'}) - \frac{\vec{BB'}}{\|BB'\|} * \frac{d}{dB'}(\|BB'\|))$$

3.3.2 Derivatives for A and B

With the atoms adjacent to the π-bond, neither term of the dot product derivative is reduced to zero.

$$\frac{d}{dA}DP = [\frac{d}{dA}(\vec{CROSSA}) \cdot \vec{CROSSB}] + [\vec{CROSSA} \cdot \frac{d}{dA}(\vec{CROSSB})]$$

Following the other derivatives, we get that

$$\frac{d}{dA}(\vec{CROSSB}) = \frac{d}{dA}(\frac{\vec{BA}}{\|BA\|}) \times \frac{\vec{BB'}}{\|BB'\|}$$

$$\frac{d}{dA}(\frac{\vec{BA}}{\|BA\|}) = \frac{1}{\|BA\|}(\frac{d}{dA}(\vec{BA}) - \frac{\vec{BA}}{\|BA\|} * \frac{d}{dA}(\|BA\|))$$

$$\frac{d}{dA_x}(\vec{BA}) = (1, 0, 0)$$

$$\frac{d}{dA_x}(\|BA\|) = \frac{A_x - B_x}{\|BA\|}$$

$$\frac{d}{dA}(\vec{CROSSA}) = \frac{\vec{AB}}{\|AB\|} \times \frac{d}{dA}(\frac{\vec{AA'}}{\|AA'\|}) + \frac{d}{dA}(\frac{\vec{AB}}{\|AB\|}) \times \frac{\vec{AA'}}{\|AA'\|}$$

$$\frac{d}{dA}(\frac{\vec{AB}}{\|AB\|}) = \frac{1}{\|AB\|}(\frac{d}{dA}(\vec{AB}) - \frac{\vec{AB}}{\|AB\|} * \frac{d}{dA}(\|AB\|))$$

$$\frac{d}{dA_x}(\vec{AB}) = (-1, 0, 0)$$

$$\frac{d}{dA_x}(\|AB\|) = \frac{B_x - A_x}{\|AB\|}$$

$$\frac{d}{dA}(\frac{\vec{AA'}}{\|AA'\|}) = \frac{1}{\|AA'\|}(\frac{d}{dA}(\vec{AA'}) - \frac{\vec{AA'}}{\|AA'\|} * \frac{d}{dA}(\|AA'\|))$$

$$\frac{d}{dA_x}(\vec{AA'}) = (-1, 0, 0)$$

$$\frac{d}{dA_x}(\|AA'\|) = \frac{A'_x - A_x}{\|AA'\|}$$

Similarly, for B

$$\frac{d}{dB}DP = [\frac{d}{dB}(\vec{CROSSB}) \cdot \vec{CROSSA}] + [\vec{CROSSB} \cdot \frac{d}{dB}(\vec{CROSSA})]$$

$$\frac{d}{dB}(\vec{CROSSA}) = \frac{d}{dB}(\frac{\vec{AB}}{\|AB\|}) \times \frac{\vec{AA'}}{\|AA'\|}$$

$$\frac{d}{dB}(\frac{\vec{AB}}{\|AB\|}) = \frac{1}{\|AB\|}(\frac{d}{dB}(\vec{AB}) - \frac{\vec{AB}}{\|AB\|} * \frac{d}{dB}(\|AB\|))$$

$$\frac{d}{dB}(\vec{CROSSB}) = \frac{\vec{BA}}{\|BA\|} \times \frac{d}{dB}(\frac{\vec{BB'}}{\|BB'\|}) + \frac{d}{dB}(\frac{\vec{BA}}{\|BA\|}) \times \frac{\vec{BB'}}{\|BB'\|}$$

$$\frac{d}{dB}(\frac{\vec{BA}}{\|BA\|}) = \frac{1}{\|BA\|}(\frac{d}{dB}(\vec{BA}) - \frac{\vec{BA}}{\|BA\|} * \frac{d}{dB}(\|BA\|))$$

$$\frac{d}{dB}(\frac{\vec{BB'}}{\|BB'\|}) = \frac{1}{\|BB'\|}(\frac{d}{dB}(\vec{BB'}) - \frac{\vec{BB'}}{\|BB'\|} * \frac{d}{dB}(\|BB'\|))$$

$$\frac{d}{dB_x}(\vec{BB'}) = (-1, 0, 0)$$

$$\frac{d}{dB_x}(\|BB'\|) = \frac{-B'_x + B_x}{\|BB'\|}$$

$$\frac{d}{dB_x}(\vec{BA}) = (-1, 0, 0)$$

$$\frac{d}{dB_x}(\|BA\|) = \frac{A_x - B_x}{\|BA\|}$$

4 Conformational Search

4.1 Molecule Construction

The secondary proposed contribution of this thesis is an algorithm for constructing molecule conformations in a timely fashion using the previously mentioned modified force field. Traditionally, research into optimizing conformations has focused on correctness at the cost of computational time and has started with the input of an existing 3D model to optimize (Finn *et al.*, 1996), (Smellie *et al.*, 1995). In this task, we have only the output of our modified force field to guide the initial and final placement of atoms and must complete in a time small enough to make molecule sketching a viable input technique for 3D structures.

4.2 Gradient-Based Methods

A common tool to find in molecular modeling packages is a gradient descent, conjugate gradient, or simulated annealing algorithm for making fine adjustments to the structure of the model. The user first designs the molecule so that the structure matches the user's overall intention. Then these algorithms can be applied to bring the already approximately correct model to a close local energy minimum. Depending on the algorithm and parameters used, the search can make major structure changes in the process, however they do not produce any assurance as to the optimality of the output, indeed local minima are frequently produced. What they do offer is speed of computation and therefore form the backbone of our conformations generation exploration.

4.3 Current ChemPad Algorithm

Presently, molecule construction in ChemPad is performed via the addition of atoms into locally optimal locations in an order based on heuristics to predict the importance of the atom on the overall shape of the molecule. Algorithm 4.1 details this process.

Algorithm 4.1: BUILD(*molecule*)

```
procedure BUILD($Molecule$)
  while $Molecule.Atoms.Count < TotalAtoms$
    do
      $a \leftarrow HighestPriorityAtom()$
      $parent \leftarrow AtomAlreadyInMoleculeAdjacentTo(a)$
      $a.Position \leftarrow parent.Position + OutboundDirection(parent) *$
        $BondLengthEquilibirum(a, parent)$
      $Molecule.Atoms.Add(a)$
      $LimitedConjugateGradient(Molecule, a)$
```

Here HighestPriorityAtom() selects an atom that is adjacent to an atom already in the molecule which is heuristically chosen to have the highest priority. The parent is an atom adjacent to a. OutboundDirection() is a unit vector which would place a as far as possible from other atoms already connected to parent. In other words, it's the vector which maximizes the distance to the closest existing vector to another neighbor of parent. Note that while this makes a certain degree of sense because atom nuclei do not like to be squished together, the final placement of the atom is often far from this initial placement. Finally, a limited conjugate gradient algorithm is applied where the existing atoms are largely held fixed and only the new atom and close neighbors are allowed to move.

4.4 Problems With the Naïve Algorithm

While this algorithm performs well on straight chains molecules, problems begin to arise as soon as one atom is put in an unfortunate position. For instance, in determining the shape of a ring, the atom which closes the ring may not be able to satisfy constraints with only the limited conjugate gradient. This problem can be overcome to a certain degree by occasionally checking for high strain in the molecule and allowing a more thorough conjugate gradient to be run, but the problem becomes more complex as more rings enter the molecule. Unfortunately, molecules with many interlocking rings are very useful for organic chemists and therefore the algorithm must accommodate this difficulty. (Crippen, 1977),(Crippen & Havel, 1978) developed a distance geometry technique for determining ring shapes by simultaneously solving for the positions of all the atoms in the ring. However, integrating a distance geometry library into ChemPad did not produce results as good as those produced with gradient descent-based algorithms.

In the absence of rings, this problem still arises sometimes and is most painfully detectable in atoms critical to determining the stereochemistry of the molecule. Since the augmentation force field terms are not properly defined until all of the participant atoms are included in the molecule, atoms can be put in painfully wrong positions early and later face the difficulty of needing to "push through" the other atoms to get to the right positions. This is further complicated in cases where a given atom may be part of two or more stereochemistry terms.

Therefore, the decision of how to prioritize atoms for addition to the structure can significantly affect the conformation produced. Currently, we sort the atoms by their participation in major structure components such as rings and long chains. Implicit hydrogen atoms occur last in the order and have little impact on the major structure features. Preliminary investigations into using the order the user drew each sketch component indicates a disposition towards drawing important structures first. However, temporal order should probably not be the only indictor of priority.

4.5 A Few Discrete Options

A few potential solutions to the problem of misplacing an atom early revolve around limited search over a few discrete options. One possibility would be to identify situations where multiple similar minima exist for a specific atom and note the one chosen. If later on a great deal of strain is found in the molecule, the algorithm can backtrack to this decision point and take another track. While this would not account for cases where an atom needs to sit far from its specific minimum to satisfy other constraints, such as is the case with cyclobutane, it would specifically overcome the ring and stereochemistry problems mentioned earlier.

Another such possibility to explore in this thesis is the early identification of "critical components" of the molecule such as distinct rings and stereocenters. These components would be built in isolation and initially bound only to the constraints within their individual structures. The components could then be assembled together thereby reducing the number of optimization steps performed and reducing optimization errors at points not involved in multiple structures.

4.6 Using the Structure Cues

Given that cues to the final 3D structure are present in the molecule sketch, it would behoove us to leverage this information rather than only trying to satisfy the constraints defined by it. While the template and heuristic method described in section 2.3 and (Tenneson, 2005a) fails to address the chemical needs of the conformation, its techniques for utilizing the structure cues in the drawing could still be used to help choose likely positions for individual atoms in our algorithm before limited optimization is applied. By moving the atoms to places implied by the structure cues, optimization time can be reduced and we are less likely to need to backtrack on the discrete options mentioned in the previous section.

4.7 Other Solutions

Some more heavy-weight potential solutions exist that we have yet to apply to the problem. A commercial non-linear systems solver could potentially find global energy minima, but at the expected expense of unreasonable computation time. Likewise, Molecular Dynamics or a similar Monte Carlo system could be used to simulate the molecule under constant dynamic motion and minima can be found much like they are in some protein folding systems (Clote & Backofen, 2000). This too is expected to be unfeasible due to computation time. While these systems are not expected to be useful for the user interface, molecular dynamics solutions could provide an interesting reference point for the quality of produced answers.

5 Pedagogy

Although not directly part of the future work in this proposal, it is noteworthy that preliminary versions of this work have already proven valuable as an educational tool which has been used by hundreds of Brown University chemistry students. ChemPad has been part of Prof. Matthew Zimmt's introductory organic chemistry course spring semesters 2005 - 2007 as both a part of lectures and as a tool available for students to use outside of class time. ChemPad's value as an educational tool comes primarily from its ability to help students learn to visualize molecules as 3D structures and not just as the 2D drawings in their textbooks. Students can use ChemPad to quickly prototype molecules

Figure 15: Students working in the ChemPad lab.

and investigate their 3D properties. Additionally, ChemPad contains a number of pedagogically-inspired visualizations concentrating on 3D thinking concepts students often have difficulty grasping.

So far, we have had more than 250 Brown University student users of ChemPad in our mobile Tablet PC "ChemPad Lab" (seen in Figure 15) and over a thousand students who have seen ChemPad used in lectures to both present material and to field questions. The results of a user study on its effectiveness as a educational tool as well as more details on its deployment at Brown can be found in (Tenneson, 2007), (Tenneson, 2005b), (Tenneson, 2005a). ChemPad is also available as a free download from its website and feedback emails on its effectiveness (and bugs) have been received from several teachers using it in their high school and college chemistry courses.

6 Evaluation

In idealistic terms, this work will be successful if a completed version could be given to a room of chemists and they could use it without once complaining that the output doesn't "look right". However, the reality is that the documented list of augmentation terms does not cover the entirety of molecule sketch notations and conventions, only an important subset. Additionally, a conformation algorithm which ensures a globally optimal solution is computationally unfeasible for a user interface technique. Therefore, our evaluation guidelines will need to be accommodating of these limitations.

The evaluation criteria, in order of importance, are as follows:

1. Stereochemistry Correctness: Producing a molecule conformation with a stereochemistry error is unacceptable since the produced molecule is not the molecule the user drew.

2. Speed: To be reasonably useable as a user interface technique, the algorithm must complete within a short period of time. As a rule, we would like to be able to handle sixty atom molecules or smaller in less than a minute.

3. Conformational Correctness: Much less absolute is the criteria of conformational correctness. Conformations produced should be considered "reasonable" interpretations of the drawn input by the user.

4. Ease of Use: At a more holistic level, the produced interface needs to be natural for trained chemists. Something that can be picked up and used effectively after only a short demo.

Admittedly, the scope of this project is limited so as to actually produce a working system. As such, we expect our algorithms to fail when presented with input in the following categories:

- Undefined Force Field Terms: Since our technique builds upon an existing force field, it can only be as powerful as that particular force field. Many notational correct molecule drawings have undefined energies due to lack of experimentally-based parameters in the force field data set.

- Undefined Notations: Additional molecule sketch notations such as group shorthand e.g. H_3C or Et, formal charges, lone pairs, etc. are not part of our input mechanism and therefore not expected to be handled by our algorithm. Implementing these notations would require additional force field augmentation terms and, in some cases, the use of a different force field.

- Large Molecules: Our target molecules are those such as would be found in an organic chemistry textbook - on the order of sixty atoms or smaller. Particularly large molecules such as proteins are well beyond the scope of this work.

In formal terms, a test set of molecule diagrams will be assembled which represent the range of molecule diagram features we wish to accommodate. These test cases will be drawn from sets such as those in (Wang *et al.*, 2004), (Wang *et al.*, 2006) and contain additional cases of interest to this thesis. Output conformations generated using the test set will be judged as reasonably correct or incorrect by domain experts.

Informally, evaluation will also be performed in an ongoing fashion by members of the Brown Chemistry Department. We have been working closely with Professor Matthew Zimmt on this project for several years receiving quick feedback on the work being done and the shortcomings left to address. We intend to continue this relationship with Zimmt and other professors interested in using the work in their classrooms over the duration of the project.

7 Recap of Proposed Work

This document proposes a thesis centered on the task of creating a pen-based user interface for 3D molecule models of organic compounds such as would be found in a college organic chemistry course. In particular, this thesis work shall produce:

- A molecular-mechanics force field augmented to produce minima only at model conformations which satisfy both the notation of the molecule diagram sketch and chemical feasibility. Specifically, solutions for the following notational problems will be found:

  - R/S Stereochemistry
  - Z/E Stereochemistry
  - Cis/Trans Ring Structures
  - Syn-periplanar Torsions
  - Alternate View Angles including Ring Notations such as Chair/Boat Conformations and Axial/Equitorial Positions of Groups

- A conformation building algorithm which quickly creates molecule models satisfying the constraints of the aforementioned force field equation. This algorithm must run on a modern Tablet PC at speeds suitable for a user interface technique.

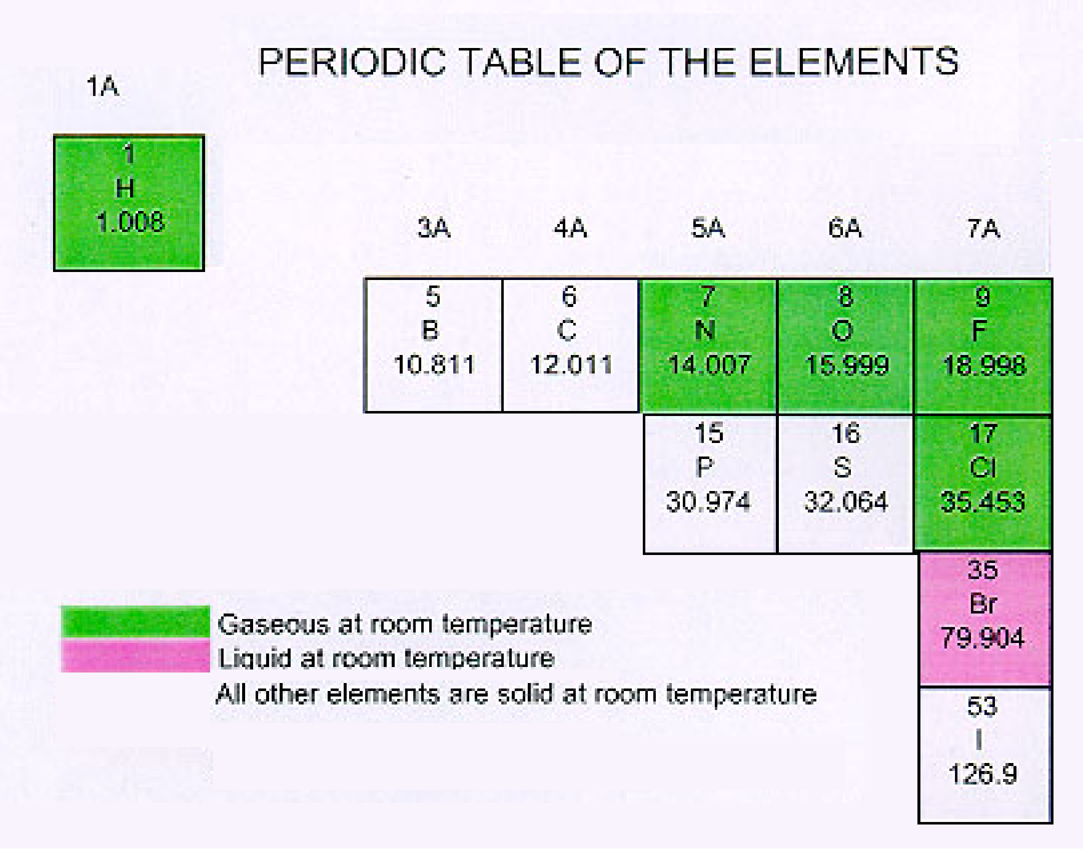
A Chemistry Primer

This primer is intended to give an introduction to chemistry for computer scientists. It will focus specifically on concepts useful to understanding this thesis proposal and should not be considered a primer suitable for chemistry students.

A.1 The Nature of Atoms and Bonds

According to atomic theory, the basic unit of matter is the atom; a single nucleus of positrons and neutrons surrounded by a shell of moving electrons. The number of positrons in the atom determines many properties of the atom and thus atoms with different numbers of positrons, or different atomic numbers, are considered different elements. The atomic numbers and weights of some elements important to this work can be found in Figure 16 which is a subsection of the periodic table of the elements.

The positive charges in atom nuclei attract negatively charged electrons and tend to be most stable when the number of positrons and electrons are equal. These atoms are considered to be neutrally charged. Two atom nuclei can also share pairs of electrons between them which encourage the atoms to stay close to each other. This relationship is called a covalent bond, or bond, for the purpose of this document. If more than two pairs of electrons are shared between atoms this is considered to be a "double-bond" and three pairs makes a "triple-bond". Atoms so bonded together into structures are called molecules. Molecules are dynamic structures constantly changing as they collide into each other and the forces within the molecule are challenged to keep the molecule together. For this document, we will be focusing on stable molecules which may contort their shapes under outside influence, yet keep maintain their bonds as a rule.



PERIODIC TABLE OF THE ELEMENTS

1A	3A	4A	5A	6A	7A
1 H 1.008					
	5 B 10.811	6 C 12.011	7 N 14.007	8 O 15.999	9 F 18.998
			15 P 30.974	16 S 32.064	17 Cl 35.453
					35 Br 79.904
					53 I 126.9

🟩 Gaseous at room temperature
🟪 Liquid at room temperature
All other elements are solid at room temperature

Figure 16: Elements of the periodic table used in this work.

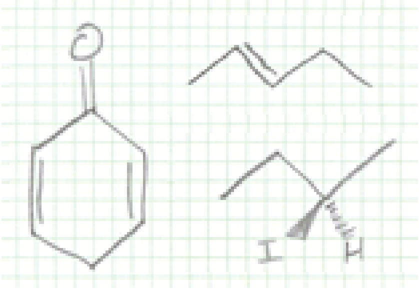


Figure 17: Examples of molecule diagrams drawn on paper.

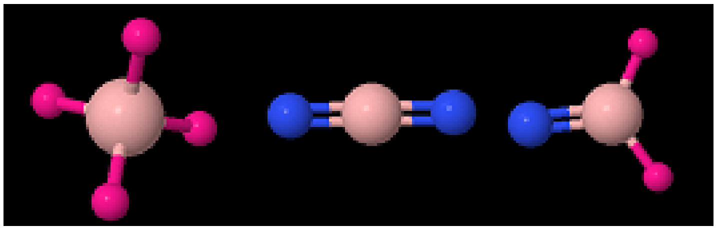


Figure 18: Hybridizations of carbon. From left to right: sp^3 tetrahedral, sp linear, sp^2 trigonal-planar.

A.2 Molecule Diagrams

A typical molecule drawing consists mostly of atoms and bonds connected in a graph-like fashion. Atoms are represented by their element symbol from the periodic table and bonds are shown by drawing lines between atoms sharing electron pairs. If two atoms double or triple bonded, two or three lines are drawn between them respectively. Symbols next to atoms can designate the presence of electron lone pairs, formal charges on the atoms, or additional chemical cues to the nature of the molecule although those are outside the context of this paper. Examples of some simple molecule drawings can be seen in Figure 17.

Within the context or organic molecules, chains of carbons are so frequently used, that designating a letter "C" at each bond intersection becomes cumbersome. Therefore, the C's in carbon backbones are often dropped and implied by the change of direction in the bonds. Similarly, hydrogen atoms which consist of a single proton and electron "fill in" the gaps of organic molecules providing the right number of electrons to make the rest of the atoms neutrally charged–unless otherwise denoted. Therefore, hydrogen atoms and their bonds are usually dropped from the diagrams and are considered implicit except where their presence or absence is needed to show a specific property of the molecule.

A.3 Structure of Atoms and Hybridization

The properties of shared of electrons between bonded atoms defines the shapes a given molecule will assume. A given pair of electrons in a bond should not be considered to be floating over the entirety of the molecule (in most cases), but as moving in a region between the two atom nuclei holding those atoms together. When additional electrons are shared to form a higher order bond, these electrons do not join the region directly between the atoms because the electrons would push against each other too much. Instead, the electrons are said to take on a different hybridization, or shape. The sp^2 hybridization explains the electron sharing in a double bond by sharing the extra electrons in orbitals "above" and "below" the atom. Similarly, the sp hybridization explains the electron sharing in a triple bond by adding an additional pair of electrons being shared to the "left" and "right" of the atom. The sp^3 hybridization is used to show electrons shared in single bonds where the electrons fall only directly between the atoms. Examples of atoms in these hybridizations appear in Figure 18.

A.4 Stereochemistry and algorithms

When examining the shape of sp^3 hybridized atoms and their bonds to neighboring atoms, a tetrahedral shape emerges. If one were label these neighboring atoms #1,#2,#3,#4, there would be exactly two ways to do this that are not rotationally equivalent. In other words, there is a chirality, or handedness at these sp^3 atoms when they have four different structures attached to them. Such atoms are called stereocenters. Determining the handedness of a stereocenter first requires the neighboring groups to be ordered according to the CIP rules described later in this section. Once the groups have been labeled #1,#2,#3, and #4, the algorithm for determining the chirality is to put the lowest priority (#4) group on the positive z-axis. Then, from the viewpoint of the z-axis looking in the positive direction (#4 is directly behind the central atom from this viewpoint), determine if the groups #1,#2,#3 form a clockwise or counter-clockwise turn. A clockwise turn is termed R and the counter-clockwise turn is termed S.

Similarly, the rigid structure of double bonds causes another kind of stereochemistry for the atoms adjacent to those in the double-bond. In this case, the CIP rules are to number the groups, but only on each side of the bond. Thus, you will have a #1 and #2 on each side of the bond rather than an overall #1,#2,#3,#4. In cases, where a double bonded atom only has one other neighbor, that neighbor receives the #1 designation and no atom receives the #2. If you were to consider a plane passing through the double bond roughly perpendicular to all the bonds formed with the atoms in the double bond and their neighbors, you would find two possible configurations. Either the #1 atoms are both on the same side of this plane, or they are on different sides of this plane. The configuration with the #1's on the same side is termed Z and the other is termed E.

A.5 CIP

The Cahn-Ingold-Prelog (CIP) rules are used to determine the priorities of atoms attached to a stereocenter for the purposes of naming the stereochemistry. Priorities go to atoms with higher atomic numbers. When there is a tie for atomic number, the tie is broken by looking at the atomic numbers of the neighboring atoms. Things get more complicated when the neighboring atoms also make a tie for atomic numbers. The CIP rules define a recursive search along the branches with the highest atomic number until a difference is found. The rules are further complicated with special cases to handle double bonds and atoms with the same atomic number, but different atomic weights. An algorithm for computing CIP priorities based on string comparisons is documented in (Tenneson, 2005a) and another for finding all CIP priorities in a molecule simultaneously is available in (Labute, 1996)

B Molecular Mechanics Primer

This primer is intended to give an introduction to molecular mechanics for computer scientists. Molecular mechanics concerns itself with formulating molecule structure energies efficiently on a computer. A number of different molecular mechanics systems, or force fields have been developed over the years to accommodate different types of molecules and calculations. A force field contains an equation defining the energy of a conformation and a data set of constants for the equation terms. While the formula terms are general in their definitions, the constants are used to approximate the interactions of specific types of atoms in specific environments. They are produced from data fitting techniques applied to results from lab experiments in molecule structure.

Where equations are shown in this section, the equations shown are for the AMBER (Ponder & Case, 2003) and GAFF (Wang *et al.*, 2004) force fields in particular although the formulations are usually very similar for other force fields. Good explanations of molecular mechanics concepts for computer scientists wishing to implement a force field can be found in (Heath *et al.*, 2005).

A force field equation typically contains terms for bond length, angles between atoms, torsional angles of bonds, improper torsional angles for certain atoms, and Van der Waal interactions between atoms separated by at least 4 bonds. Additional terms may be present to account for more complex phenomena.

B.1 Atom Types

Although atoms are defined in chemistry by their atomic numbers and atomic weights alone, for molecular mechanics, atoms are also differentiated by additional features such as the connectivity neighborhoods they inhabit. For this reason, there are 17 different versions of, or atom types for, carbon within the GAFF force field. Alternatively, there is only one atom type for fluorine. These different atom types can be thought of as the different pieces in a plastic ball-and-stick modeling kit, since they define the idealized shape the atoms take under different conditions. However, instead of defining only the idealized shape local at that atom, they define the shape as a function of nearby atom types and define the strength of those shapes - how easily they give way. In the following force field terms, whenever a constant is determined by the atoms engaged in the term, it is the atom types, rather than the atomic numbers that determine the constants.

B.2 Bond

$$\sum_{\text{bonds}} K_r (r - r_{eq})^2$$

The distance between bonded atoms achieves equilibrium where the force of the pull by electrons shared by the atoms exactly equals the force of the repulsion of the atom nuclei. The bond length term measures the amount of energy required for two bonded atoms to have a distance different from this equilibrium. In the formulation, K_r is the empirically determined force constant, r_{eq} is the empirically determined equilibrium bond length, and r is the current bond length. K_r and r_{eq} depend on the specific atoms (atom types) in the bond and the order of the bond and can be found in the parameter sets of the force field.

B.3 Angle

$$\sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2$$

or

$$\sum_{\text{a}\in\text{atoms}} \sum_{\text{b1,b2}\in\text{a.neighbors}} K_\theta (\theta - \theta_{eq})^2$$

Similarly, there is an equilibrium state for the angles formed by a given atom and any two of its neighboring atoms. These atom angles represent the specific hybridization of the atom electrons. The angle term represents the amount of energy required to "bend" these bonds into non-idealized conformations. Here K_θ is the empirically determined force constant and θ_{eq} is the empirically determined equilibrium constant. θ is the existing angle between the atoms. K_θ and θ_{eq} both depend on the three atoms engaged in the angle.

B.4 Torsion

$$\sum_{\text{dihedrals}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$$

or

$$\sum_{\text{b}\in\text{bonds}} \sum_{\text{a}\in\text{b.atom1.neighbors}} \sum_{\text{c}\in\text{b.atom2.neighbors}} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]$$

Torsional angles, or dihedral angles, are the angles formed by four atoms connected in a three bond chain. The angle measured ϕ is that of $\angle ABC$ in the plane perpendicular to the center bond. Here both bond atoms have the point B and the other two atoms have the points A and C. The constants are n the periodcity of the term, γ the equilibrium, and V_n the force constant.

One way to think of the torsion term is to think of this as the energy representing the repulsion of atoms that are close together but not already accounted for by the bond and angle terms. The bond term defines the interactions of atoms one bond apart and the angle term handles the atoms two bonds apart.

From an implementation perspective, the torsional term is the most difficult to complete correctly. First, one should check closely if the V_n term has been pre-divided by 2 in the data set as it is in AMBER and GAFF. Furthermore, a means to calculate the torsional angle itself is not intuitive, although good references appear in (Rainey, 2003), (Bekker, 1996). Finally, for gradient descent algorithms, the analytical derivative the term with respect to each coordinate for each atom is a non-trivial problem due to singularities in the most straight forward solutions. (Blondel & Karplus, 1996) finally provides a derivative provably without singularities.

B.5 Improper Torsion

$$\sum_{\text{improper}} \frac{V_n}{2}[1+\cos(n\phi-\gamma)]$$

or

$$\sum_{\text{a}\in\text{atoms}} \sum_{\text{b,c,d}\in\text{a.neighbors}} \frac{V_n}{2}[1+\cos(n\phi-\gamma)]$$

Although the energy function of the improper torsion is the same of the torsion term, improper torsions are defined over three atoms all connected by one bond to a fourth. In particular, the improper torsion term is defined for sp^2 atom types and is zero for all other cases. Here the torsion angle is formed in the order $\angle BACD$ even though C and D are not bonded to each other.

B.6 Steric

$$\sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right]$$

or calculated differently

$$\sum_{i<j} 4*\rho*(\frac{\sigma}{R_{ij}}^{12} - \frac{\sigma}{R_{ij}}^{6}) + \frac{Q_i * Q_j}{\epsilon} * Rij$$

While the other terms define the interactions of atoms connected by three or less bonds, the steric term defines the interactions of each atom on atoms more than three bonds away. These atom pairs are attract each other gently, but push apart strongly as the atoms begin to invade each others' electron shells. The measured term here is R_{ij} the current distance between atoms i and j. The constants are ρ the well depth of i and j, σ the hard sphere radius of i and j (the Van der Waal radii averaged under the Lorentz-Berelot combining rule), ϵ the effective dialectric constant for i and j and Q_i and Q_j electrostatic constants for i and j respectively. While the electrostatic term at the end is not strictly steric strain, it is a relationship between distant atoms and therefore considered here.

B.7 Others

While this concludes the force field terms present in GAFF, additional terms are present in other force fields to represent complex relationships that occur in other classes of molecules that are not represented in this set of terms. For instance, AMBER contains a term which represents the interaction of hydrogen bonds while MM2/MM3 contains a stretch-bend term and MM4 (Nevins *et al.*, 1996) contains even more terms to compensate for spectroscopic frequencies.

References

Bekker, Hendrik. 1996. *Molecular dissociation induced by electron transfer to multicharged ions.* Ph.D. thesis, University of Groningen.

Blondel, Arnaud, & Karplus, Martin. 1996. New Formulation for Derivatives of Torsion Angles and Improper Torsion Angles in Molecular Mechanics: Elimination of Singularities. *Journal of Computational Chemistry*, **17**(9), 1132–1141.

Bryfczynski, Sam. 2006. *OrganicPad Molecule Creator Application.* http://people.clemson.edu/∼sbryfcz/481/index_files/OrganicPad.htm.

Butler, Declan. 2005. Electronic notebooks: A new leaf. *Nature*, **436**(July), 20–21.

Cieplak, T., & Wisniewski, J.L. 2001. A New Effective Algorithm for the Unambiguous Identification of the Stereochemical Characteristics of Compounds During Their Registration in Databases. *Molecules*, **6**, 915–926.

Clark, David E., Jones, Gareth, & Willett, Peter. 1994. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *Journal of Chemical Information Computer Sciences*, **34**, 197–206.

Clote, Peter, & Backofen, Rolf. 2000. *Computational Molecular Biology, An Introduction.* John Wiley and Sons, Ltd. Chap. Structure Prediction.

Crippen, Gordon M. 1977. A Novel Approach to Calculation of Conformation: Distance Geometry. *Journal of Computational Physics*, **24**, 96–107.

Crippen, Gordon M., & Havel, Timothy F. 1978. Stable Calculation of Coordinates from Distance Geometry. *Acta Crystallographica*, **A34**, 282–284.

Edwards, B., & Chandran, V. 2000. Machine Recognition of Hand-Drawn Circuit Diagrams. *Pages 3618–3621 of: ICASSP '00. Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6.

Finn, Paul W., Halperin, Dan, Kavraki, Lydia E., an Rajeev Motwani, Jean-Claude Latombe, Sheton, Christian, & Venkatasubramanian, Suresh. 1996. Geometric Manipulation of Flexible Ligands. *In: Proceedings of the First ACM Workshop on Applied Computational Geometry.*

Forsberg, Andrew S., Dieterich, Mark, & Zeleznik, Robert C. 1998. The Music Notepad. *Pages 203–210 of: ACM Symposium on User Interface Software and Technology.*

Gennari, Leslie, Kara, Levent Burak, & Stahovich, Thomas F. 2005. Combining Geometry and Domain Knowledge to Interpret Hand-Drawn Diagrams. *Computers and Graphics*, **29**(4), 547–562.

Heath, Allison, Kavraki, Lydia, & Shehu, Amarda. 2005. *Energy Functions.* Connexions Web Module 11449 - http://cnx.rice.edu.

Igarashi, Takeo, Matsuoka, Satoshi, & Tanaka, Hidehiko. 1999. Teddy: a sketching interface for 3D freeform design. *Pages 409–416 of: SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques.* New York, NY, USA: ACM Press/Addison-Wesley Publishing Co.

Karpenko, Olga A., & Hughes, John F. 2006. Implementation details of SmoothSketch: 3D free-form shapes from complex sketches. *Page 51 of: SIGGRAPH '06: ACM SIGGRAPH 2006 Sketches.* New York, NY, USA: ACM Press.

Labahn, George, MacLean, Scott, Marzouk, Mirette, Rutherford, Ian, & Tausky, David. 2006. A preliminary report on the mathbrush pen-math system. *Pages 162–178 of: Proceedings of Maple Conference 2006.* Maplesoft.

Labute, P. 1996. An Efficient Algorithm For The Determination Of Topological RS Chirality. *Journal of the Chemical Computing Group*, November.

LaViola, Jr., Joseph J., & Zeleznik, Robert C. 2004. MathPad^2: a system for the creation and exploration of mathematical sketches. *Pages 432–440 of: SIGGRAPH '04: ACM SIGGRAPH 2004 Papers.* New York, NY, USA: ACM Press.

m. c. schraefel, Hughes, G., Mills, H., Smith, G., Payne, T., & Frey, J. 2004. Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment. *In: Proceedings of CHI 2004.*

Nealen, Andrew, Sorkine, Olga, Alexa, Marc, & Cohen-Or, Daniel. 2005. A Sketch-Based Interface for Detail-Preserving Mesh Editing. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, **24**(3), 1142–1147.

Nevins, Neysa, Chen, Kuohsiang, & Allinger, Norman L. 1996. Molecular Mechanics (MM4) Calculations on Alkenes. *Journal of Computational Chemistry*, **17**(5-6), 669–694.

NIH. *NIH Guide to Molecular Modeling.* http://cmm.cit.nih.gov/modeling/guide_documents/.

Ouyang, Tom, & Davis, Randall. 2006. Recognition of Hand Drawn Chemical Diagrams. *In: Second Annual CSAIL Student Workshop.*

Ponder, J.W., & Case, D.A. 2003. Force fields for protein simulations. *Advances in Protein Chemistry*, **66**, 27–85.

Rainey, J.K. 2003. *Collagen structure and preferential assembly explored by parallel microscopy and bioinformatics.* Ph.D. thesis, University of Toronto.

Smellie, Andrew, Kahn, Scott, & Teig, Steven. 1995. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *Journal of Chemical Information and Computer Sciences*, **35**, 285–294.

Tenneson, Dana. 2005a (May). *ChemPad: A Pedagogical Tool For Exploring Handwritten Organic Molecules.* M.Phil. thesis, Brown University, Providence, RI.

Tenneson, Dana. 2005b. *Report On The Development of ChemPad for Teaching Organic Chemistry Students to Visualize Three-Dimensional Molecular Structures.* Tech. rept. Brown University.

Tenneson, Dana. 2007. ChemPad: Visualizing Molecules in Three Dimensions. *In: Workshop on the Impact of Pen Technologies in Education Monograph 2007 (title t.b.a).* Purdue University Press. In Press.

Tenneson, Dana, & Becker, Sascha. 2005. ChemPad: generating 3D molecules from 2D sketches. *Page 87 of: SIGGRAPH '05: ACM SIGGRAPH 2005 Posters.* New York, NY, USA: ACM Press.

Tenneson, Dana, & Maloney, Chris. 2007. *ChemPad3 Beta Software.* http://www.chempad.org. Software Download.

Wang, Cheuk-San. 2000. *Determining Molecular Conformation from Distance or Density Data.* Ph.D. thesis, Massachusetts Institute of Technology.

Wang, Junmei, Wolf, Romain M., Caldwell, James W., Kollman, Peter A., & Case, David A. 2004. Development and testing of a general AMBER force field. *Journal of Computational Chemistry*, **25**(9), 1157–1174.

Wang, Junmei, Wang, Wei, Kollman, Peter A., & Case, David A. 2006. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, **25**(2), 247–260.

Weiner, Scott J., Kollman, Peter A., Case, David A., Singh, U. Chandra, Ghio, Caterina, Alagona, Guiliano, Jr., Salvatore Profeta, & Weiner, Paul. 1984. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *Journal of the American Chemical Society*, **106**, 765–784.

Zeleznik, Robert, & Miller, Timothy. 2006. Fluid Inking: augmenting the medium of free-form inking with gestures. *Pages 155–162 of: GI '06: Proceedings of the 2006 conference on Graphics interface*. Toronto, Ont., Canada, Canada: Canadian Information Processing Society.

Zeleznik, Robert, Miller, Timothy, Holden, Loring, & LaViola, Jr., Joseph J. 2004. *Fluid Inking: Using Punctuation to Allow Modeless Combination Of Marking and Gesturing.*

Zeleznik, Robert C., Herndon, Kenneth P., & Hughes, John F. 1996. SKETCH: An Interface for Sketching 3D Scenes. *Pages 163–170 of:* Rushmeier, Holly (ed), *SIGGRAPH 96 Conference Proceedings*. Addison Wesley.