A Statistical Model of Short Carbohelicenes

By

Matt Anderson

A Thesis

Submitted to the Department of Integrated Science

in Partial Fulfillment of the Requirements

for the Degree

Bachelor of
Science

McMaster University

McMaster University BACHELOR OF SCIENCE (2018) Hamilton, Ontario (Integrated Science)

TITLE: A Statistical Model of Short Carbohelicenes

AUTHOR: Matthew Anderson (McMaster University)

SUPERVISOR: Professor An-Chang Shi

NUMBER OF PAGES: viii, 38

Abstract

Carbohelicenes are molecules composed of multiple adjoined benzene rings that form a helical structure due to a repulsive steric interaction between overlapping rings. These molecules and their more general helicene derivatives are of interest for their useful electronic and structural properties. To advance understanding of the molecule, a statistical mechanics model based upon 2 supposedly significant interactions is proposed. In the model, the energy of a molecular conformation depends upon a bending potential between rings and a distance potential between every sixth-ring pair. Predicted equilibrium conformations are compared to the accepted 3D structures and the free energy curve between equilibrium states is computed for a 6 ringed helicene ([6] helicene). Equilibrium conformations are calculated with a Monte Carlo Metropolis procedure simulation, while a variation on this sampling method, the umbrella sampling scheme is used for the free energy curve calculation. The proposed model effectively predicts the conformation of [6] helicene. The model predicted planar molecules for [4] and [5] helicene, and non-physical conformations for larger helicenes. The non-physical conformations were conformations that folded into themselves. They are attributed to the lack of a hard-core interaction preventing atoms from occupying space too close to one another. The free energy calculation yielded different free energies for the two equilibrium states, but the curve had the general expected form and was of the correct order of magnitude. The next step in developing this model is to include a hard-core interaction.

Acknowledgements

I would like to thank my supervisor Professor An-Chang Shi for his support and guidance during this project. Pointing me in the right direction more than once and being readily available to meet when I needed the help was invaluable for this project. I really appreciated the opportunity to practice my presentation with Dr. Shi on two separate occasions before the Synthesis symposium. I am also thankful for the opportunity to get exposure to the field of computational physics and develop my skills in this field. I'd like to thank my peers in Integrated Science for the support we gave to each other as we strived to achieve our best this year. Special mention to Jonathan for your help with some of my code issues, Varsha for the support when I felt like things were falling apart, and to Jordan and Jamie for their listening to my practice presentation of my thesis and the suggestions they provided. I would like to thank my family and friends for the support and the interest they've shown in what I'm doing.

Contents

List of figures

1 Introduction

Research into the properties and applications of helicenes is an active field (Gingras, 2013). The unique chiroptical properties of helicenes and its two enantiomeric states are exploitable properties that have been hypothesized to have applications in molecular sensing, liquid crystal technology, molecular machinery, and data storage to name a few (Gingras, 2013, Feringa 2001). These exploitable properties are fundamentally a result of the structure of helicene. Despite the importance of the structure, there has been no attempt made to develop a model of the molecule. Instead, semi-empirical quantum calculation methods have been used to compute the stability of helicenes (Sehnal et. al, 2009, Grimme and Peyerimhoff, 1996, Janke, et al. 1996). Semi-empirical methods offer little insight into the molecule and lack the same predictive power a matured model could have.

The goal of this thesis is to develop a statistical model for helicenes. Helicenes will be modelled as a chain of benzene rings where a bending potential and a "6th ring" potential define the energy of a conformation. This model will be tested by calculating the free energy profile along the reaction coordinate between the right and left-handed conformations. The height of the energy barrier between the two states (known as the racemization barrier), the conformation of the transition state, the number of peaks along the reaction coordinate will all be compared to existing data and results from previous experiments to assess the validity of the model. Thus, this thesis will aim to put forward a theoretical model of helicenes.

1.1 Helicenes

Helicenes are a class of molecules that are composed of benzene rings adjoined to form a ring that is pushed apart into a 3-D helical shape due to steric interactions between the ends (Chen and Shen, 2017). The simplest type is a carbohelicene, a helicene composed solely of benzene rings (Gingras, 2013). The smallest possible number of rings for which helicity is observed and thus qualifying as a helicene is 4 adjoined benzene rings. The case of 2 adjoined benzene rings is called naphthalene and 3 rings, phenanthrene, and they are both planar molecules (PubChem, 2004). The convention by which a helicene of a certain number of rings is referred to is to put the number in square brackets preceding the name of the helicene. For

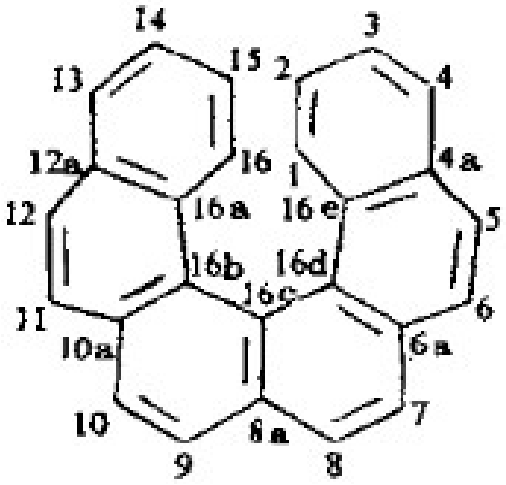

Figure 1: A flattened schematic of [6]Helicene depicting the standard numbering system for carbon atoms in the molecule. (Martin, 1974)

example, the helicene with 6 rings and without additional functional groups is referred to as [6]Helicene. This naming convention will be used throughout the remainder of this review.

1.1.1 Structure

The base unit of the helicene molecule is the benzene ring (Martin, 1974). A benzene ring is a molecule consisting of 6 carbon atoms at the points of a symmetrical hexagon. These six atoms forming 3 double "pi" bonds and 3 single bonds, resulting in what is known as a pi system where the double bonds are essentially delocalized and all bond lengths between carbon atoms equal; a mix between double and single bond lengths. This is a relevant point to note as in a helicene molecule there is significant distortion of this pi system due to the torsion on the molecule (Gingras, 2013). This manifest itself in the structure as lengthened bond lengths between carbon atoms in the 'interior ring' of the helix and shortened bonds in the 'exterior ring' of the helix (Martin, 1974). This is due to the 'exterior ring' of the helix is composed of 3 of the sides of the benzene rings in contrast to 1 for the 'interior ring' (see fig.1 above), and so even though the inner ring has a smaller perimeter, there are much fewer bonds to cover that distance. The bond lengths for [6]helicene as an example are ~1.47 Angstroms for inner bonds, and 1.334 Angstroms for outer ring bonds (Martin, 1974). It is important to note however that these are averages and the degree of bond length distortion varies from ring to ring in the molecule (Chen and Shen, 2017).

The helicity of helicenes is one of their characteristic structural traits. It is well documented that their helicity is a direct result of the steric repulsive interaction between the terminal rings (Ravat et. al. 2017, Ben Yahia et al. 2017). To complete one full turn of the helix requires 7 rings. However, due to the deformation of the rings the 7th ring only partially overlaps with the first (Ben Yahia et al. 2017). Helicenes also posses C2 symmetry about the axis perpendicular to the axis of helicity, resulting in chirality of the molecule despite no molecular centre of chirality (Ben Yahia et al. 2017).

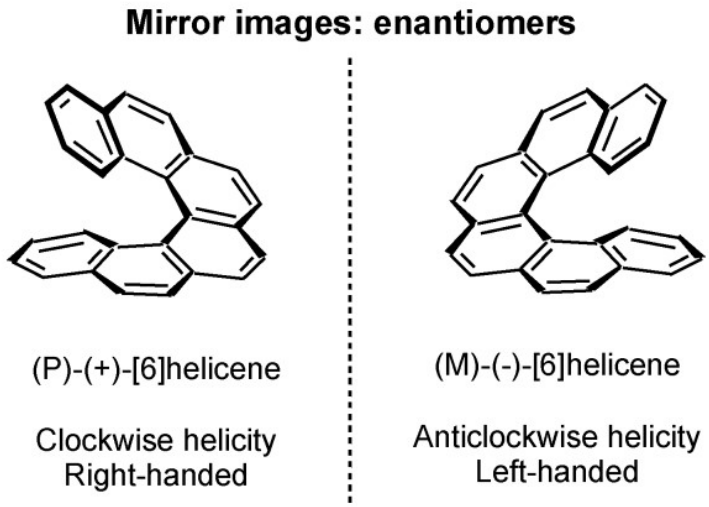

Figure 2: 3D Visualization of [6]helicene illustrating left and right-handed conformations (Gingras, 2013)

1.1.2 Configurational stability

Configurational stability, generally, is a measure of whether a system will stay in it's current configuration (free energy minima) or transition to other configurations. In the case of enantiomers, it can be used to determine how long an enantiomerically pure mixture will endure under defined conditions (Ravat et al., 2017). It is related to the energy barrier to transition to the other enantiomer, a higher barrier would mean greater stability and less frequent transitions, which means a slower rate of racemization (Reist et al., 1995). For helicenes, as a general rule, increasing number of rings will increase this energy barrier up until [9]Helicene where the free energy of racemization plateaus at 44.1 kcal/mol (Ravat et al., 2017). Values for other helicenes can be viewed in table 1.

The implication of the observation that the free energy of racemization plateaus is that helicenes are in fact a flexible molecule (Ravat et al., 2017). It is believed that the transition from one helicene enantiomer to another happens via a conformational pathway, rather than chemically through bond breaking or what is called an 'internal double Diers-Alder adduct (Martin, 1974; Ravat et al., 2017). This conformational pathway involves the stretching and torsion of bonds that allow the terminal ends of the molecule to pass by each other (Martin, 1974). As a result, it may be necessary to include a description of this molecular deformation in a model of helicene to accurately determine the energy barriers. Especially since it is noted that the relatively low energy barrier indicate that [n]Helicenes are much more flexible than is generally believed (Martin, 1974; Ravat et al., 2017), a rigid model would likely predict higher energy barriers than what is observed in experiment.

It should be noted however, that the configurational stability of helicenes can be improved beyond the 44.1 kcal/mol of [9]Helicene through the addition of functional groups at various positions (Ravat et al., 2017). Addition of functional groups can result in additional steric hindrance that increases the free energy of the conformational pathway. Although as Richard
Martin discusses in The Helicenes (1974), the position is absolutely crucial to whether or not there will even be significant effect, citing that the addition of bromine or methyl groups at the 2 and 15 position will barely change the conformation while 2 methyl groups at the 1,16a position will cause significant change in the conformation. In table 1 on the next page is a summary of the Energy barriers of racemization for various helicenes.

Entry	Helicenes[a]	E_a (kcal/mol)	T (K)	$t_{1/2}$ (min)
1	[5]	24.6	293	62.7 (57 °C)
2	1-Me[5]	38.7	473	/
3	[6]	36.2	300	13.4 (221.7 °C)
4	1-Me[6]	43.8	542	231
5	1,1′-DiMe[6]	44.0	543	444
6	2,2′-DiMe[6]	39.5	513	222
7	[7]	41.7	542	13.4 (295 °C)
8	[8]	42.4	543	3.1 (293.2 °C)
9	[9]	43.5	543	12.3 (293.5)
10	Thia[6][b]	22	298	13
11	1-Aza[5]	21.7	298	/

[a][*n*]=[*n*]helicene, Me=methyl; [b]3,6,9-trithia[6]helicene

Figure 3: Table of free energies of racemization for various helicenes. Taken from Structures and Properties of Helicenes (Chen and Shen, 2017)

1.2 Calculation Methods

In this section, we explain the theory of the computation method we use to calculate free energy differences between helicene conformations. In statistical physics, a free energy calculation takes the form:

$$F = -kTln(Z) \tag{1.1}$$

Where Z is the canonical partition function, given as the sum of all Boltzmann factors of all microstates corresponding to the macrostate of interest.

$$Z = \sum_i e^{-\beta E_i} \tag{1.2}$$

Alternatively, in integral form over a continuous space of microstates:

$$Z = \int e^{-\beta E(q)} dq \tag{1.3}$$

Where β is the usual thermodynamics beta defined as:

$$\beta = \frac{1}{k_b T} \tag{1.4}$$

Given this, we can write the free energy difference between a macrostate of interest and a reference macrostate as simple difference of equation 1.1:

$$\Delta F = F_i - F_r = -kTln\left(\frac{Z_i}{Z_r}\right) \tag{1.5}$$

Where index i denotes the macrostate of interest and r the reference macrostate. This equation can be expanded using the definition of Z and brought to the following form:

$$\Delta F = -kTln\langle \exp(-\Delta E\beta)\rangle \tag{1.6}$$

With ΔE given by:

$$\Delta E = E_i - E_r \tag{1.7}$$

Full derivation available in appendix C.1. This expression of the free energy difference between two systems agrees with that given by Torrie and Valleau in their 1977 paper on calculating the free energy differences via the umbrella sampling method. In this form, the difference of free energy is written as an ensemble average over the reference systems (Torrie and Valleau, 1977). Generally, to calculate ensemble averages in statistical physics a Monte Carlo simulation using the Metropolis procedure is sufficient (Metropolis et. al., 1953, Chib and Greenberg 1995). However, as will be discussed, in the case of the free energy, difficulties arise and modification to the methodology is necessary. In the following sections, we will describe Monte Carlo methods, the Metropolis-Hastings algorithm, the difficulty involved in calculating free energy differences and the method of umbrella sampling to overcome these difficulties.

1.2.1 Monte Carlo method

A Monte Carlo method, generally, is a stochastic calculation method for approximating the answer to integrodifferential equations (Metropolis and Ulam, 1949). The method was proposed in 1949 by Metropolis and Ulam. In solving our problem involving the expectation value of a distribution, i.e. the Boltzmann distribution, we are interested in a specific class of Monte Carlo Methods known as a Markov Chain Monte Carlo method [cite Hastings]. For the statistical physics problems we are concerned with, it is specifically the Metropolis procedure that is often used as the method's transition algorithm results in sampling converging to sampling from the Boltzmann distribution (Metropolis et. al., 1953). We will describe the method in depth in the following section and include some of the generalizations of the method described by Hastings in 1970.

1.2.2 Metropolis-Hastings Algorithm

The Metropolis algorithm was first introduced by Metropolis et. al. in 1953. In their paper they describe their method of randomly sampling from a Markov Chain they create, which converges to the Boltzmann distribution (as desired) over the course of sampling. They use this method for more efficiently calculating an expectation value integral (view equation 1.8 for example), than randomly sampling a uniform distribution.

$$\bar{F} = \frac{\int F \exp(-E\beta)\ d^{2N}p\, d^{2N}q}{\int \exp(-E\beta)\ d^{2N}p\, d^{2N}q} \tag{1.8}$$

The motivation for this method lies in that the expectation value is being calculated from a known probability distribution P(X), where X is a particular state. In this case it is the Boltzmann distribution, but as discussed by Hastings [1970], similar methods could be applied for any probability distribution. By nature of expectation value calculations, the largest contributions to the integral are those for which P(X) is large. Thus, preferentially sampling the parts of the integral where P(X) is large and avoiding those for which P(X) is arbitrarily small is advantageous for calculating a good estimate while conserving computation resources (Metropolis et. al., 1953).

The algorithm of Metropolis et. al [1953] is as follows. Given the current state of the system X, a random trial displacement is generated. The energy of this new trial state is then computed and compared against the previous state's energy. If the energy is lower the new state is accepted, however if lower the relative probability of the two states is used:

$$\frac{P_2}{P_1} = \frac{e^{-E_2\beta}}{e^{-E_1\beta}} = e^{-\Delta E\beta} \tag{1.9}$$

If the relative probability is greater than a random number on the interval [0,1] then it is accepted. Regardless of whether the trial displacement passed or was rejected, the current state is used in the calculation of the expectation value. The calculation of which takes the following form (Metropolis et. al., 1953):

$$\bar{F} = \left(\frac{1}{M}\right) * \sum_{j}^{M} F_j \tag{1.10}$$

Where M is the total number of iterations considered in the calculation, and F_j is the value of F of the j th sample. In practice, collecting samples for the expectation value calculation is done after it is believed that the Markov chain has converged to the target distribution (Cowles and Carlin 1996, Chib and Greenberg 1995). In this way, the calculated value is independent of the initial conditions of the simulation as the Markov Chain has reached its equilibrium distribution; the distribution from which it is desired to sample from. However, statistically determining convergence of the Markov chain is a difficult task as discussed by Cowles and Carlin [1996]. A rudimentary method noted by Cowles and Carlin [1996] that should be sufficient for our purposes is to analyze the output of the algorithm, and when the difference between outputs over a series of outputs has become negligible then convergence of the chain can be assumed.

1.2.3 The Free Energy Dilemma

The calculation of free energy differences as an ensemble average provides a unique problem not present in other ensemble average calculations. The nuance to this particular problem is that the high energy regions of phase space over which the integral is computed do contribute significantly

to the ensemble average in contrast to other ensemble average calculations. This becomes more apparent when the equation 1.6 for free energy differences is re-expressed by bringing the negative sign inside the logarithm to invert it, producing the following equations (Beveridge 1989):

$$\Delta F = kTln\left(\frac{\int e^{\Delta E\beta} e^{-E_i\beta} dX^n}{\int e^{-E_i\beta}\, dX^n}\right) \tag{1.11}$$

Which can further be re-written as:

$$\Delta F = kTln\left(\int e^{\Delta E\beta}\, P_i(X^n)\, dX^n\right) \tag{1.12}$$

Where $P_i(X^n)$, the Boltzmann probability function for the system of interest, is given by:

$$P_i(X^n) = \frac{e^{-E_i(X^n)\beta}}{Z} \tag{1.13}$$

Since the exp(ΔEβ) factor in the integrand of equation 1.12 grows rapidly with increasing energy of the system of interest, regions of phase space where the energy of the system of interest is high have significant contributions to the integral. This is fundamentally detrimental to the Metropolis procedure, which samples according to the Boltzmann distribution and so preferentially samples the low energy conformations while rarely sampling the high energy calculations (Beveridge 1989). In Torrie and Valleau's 1977 paper, they describe Boltzmann sampling as an inefficient means of exploring the relevant conformation space. While Kastner [2011] simply refers to direct Boltzmann sampling as "infeasible" for obtaining a profile of the free energy. In most cases, a conventional simulation of practical length will simply never have adequate sampling of the high energy regions (Beveridge 1989).

In 1977, Torrie and Valleau propose the umbrella sampling method as a solution for calculating free energy differences. This is an importance sampling method in which sampling is performed on an alternative distribution and then the probabilities related back to the distribution of interest after sampling has been completed. In this way sparse sampling of important but improbable regions of the Boltzmann distribution can be overcome. This method will be discussed in full in the following section.

1.2.4 Umbrella Sampling method

The Umbrella sampling method is a modification on the Metropolis procedure. In this procedure, the standard Boltzmann factor weighting of conformations is adjusted with the addition of another weighting function (Torrie and Valleau, 1977). In practice, this is usually achieved by adding a potential to the energy function that depends on the reaction coordinate, thus changing the Boltzmann factor (Kastner, 2011). Adjusting the Boltzmann factor governing the transition of states in the simulation has the effect of changing the Markov Chain of the simulation, and thus the equilibrium distribution it will converge to. Sampling of the computational experiment is now performed on this alternative distribution, generally referred to as sampling distribution π (Torrie

and Valleau, 1977). The goal is that 2 or more regions of interest are more probable in the sampling distribution π and so are more frequently and uniformly sampled during the simulation than with the regular Boltzmann distribution. The key to this method is that the results of sampling experiments on multiple different π distributions can be combined into a global distribution (Kumar et al. 1992). The methods of which will be discussed shortly.

An important question then becomes, can we find a sampling distribution that will result in complete and uniform sampling of the region of phase space of interest? In principle this can be achieved with a biasing potential equal to the free energy profile of the system of interest. However, it is exactly this free energy profile for which we are searching (Kastner, 2011). Torrie and Valleau [1977] admit in their paper this is the largest drawback of the umbrella method, the lack of direct and straightforward method of determining the correct energy bias to use. To resolve this issue there are 2 methods used: the use of a series of harmonic potentials in which the reaction coordinate is broken into multiple windows of sampling, and the adaptive method in which a choice of biasing function is made iteratively better with each prediction of the free energy until sampling is uniform (Kastner, 2011). We will be concerned with the former method.

The harmonic potentials method makes use of biasing potentials of the following form, where indices i denote which window it is for (Kastner, 2011):

$$w_i(\varepsilon) = \frac{K}{2}\left(\varepsilon - \varepsilon_i^{ref}\right)^2 \tag{1.14}$$

This function works to keep sampling close to the reference point, ε_i^{ref}, by increasing the energy as the system gets away from the reference point. Each reference point will be unique for each window and are generally evenly spaced out along the reaction coordinate in practice (Kastner, 2011). The choice of K will determine how tight sampling will be around the reference point. The value of K must be great enough that the bias potential pushes sampling over the energy barrier, while being small enough that sampling is not too tight around the reference point (Kastner, 2011).

The results of sampling from multiple windows can be combined after sampling into a single probability distribution via the weighted histogram analysis method (WHAM) (Kumar et al. 1992). This method assigns a weight to the probability distribution of each window. These weights are chosen to minimize the statistical error in the global distribution (Kastner, 2011, Kumar et al. 1992) The following set of equations are used to determine the weights and the unified distribution (Kastner, 2011):

$$P^u(\varepsilon) = \sum_i^{windows} p_i(\varepsilon) P_i^u(\varepsilon) \tag{1.15}$$

Where, P^u, is the normalized frequency of finding the system in a small vicinity of ξ, and P_i^u, the normalized frequency of finding the system in a small vicinity in the window i. The weight for window i is p_i.

$$p_i = N_i \exp[-\beta w_i(\varepsilon) + \beta F_i] \tag{1.16}$$

Where N_i is the number of sampling points in the window I, and F_i given by the following equation:

$$F_i = -\frac{1}{\beta}\ln\left(\int P^u(\varepsilon)\, exp[\beta w_i(\varepsilon)]\, d\varepsilon\right) \tag{1.17}$$

These three equations are interdependent. With F_i depending on P^u which depends on p_i, which in turn depends on F_i. As a result, a guess must be used and then the calculation of these three equations must be iterated until they converge (Kastner, 2011).

Finally, upon computation of $P^u(\xi)$, the free energy profile along the reaction coordinate can be computed using the following equation (Kastner, 2011):

$$A(\varepsilon) = -\frac{1}{\beta}\ln\left(P^u(\varepsilon)\right) \tag{1.18}$$

1.3 Existing computational studies of Helicenes

There are a limited number of studies involving the theoretical calculation of the free energies of racemization of helicenes. All 3 studies found involve quantum chemistry calculations including density functional theory (DFT) calculations, self-consistent field theory (SCF) methods, and semi-empirical approximations of SCF: AM1, MNDO, PM3 [cite all 3]. The relevance of these studies for the purposes of this project is in their discussion of the transition state and the pathway of racemization.

Janke, et al. [1996] computed the pathway and energy barrier of racemization for 5-9 ringed helicenes. They found that for the unsubstituted helicenes, the transition state for all but [9]helicene were symmetric about a plane through their centre, known as c_s-symmetry. This symmetry is best illustrated in the following figure:

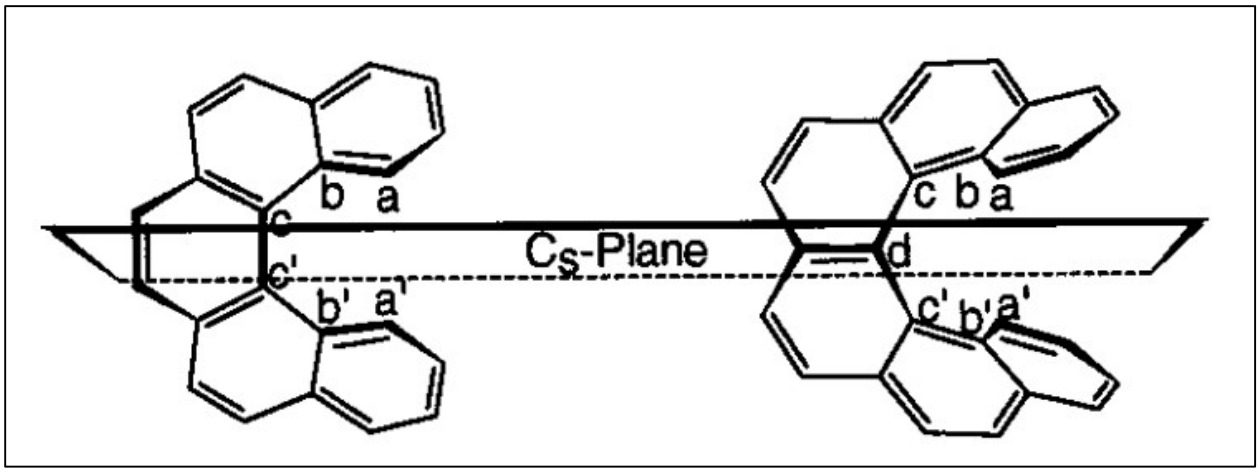

Figure 4: Helicenes show Cs - plane symmetry in the transition state [1996]

In the case of [9]helicene, the c_s-structure did resemble the transition state – it was a saddle point, but ultimately not the highest point on the reaction coordinate (Janke, et al. 1996). Instead they found that the reaction coordinate to have two symmetric peaks corresponding to two antisymmetric transition states with the c_s-structure between them.

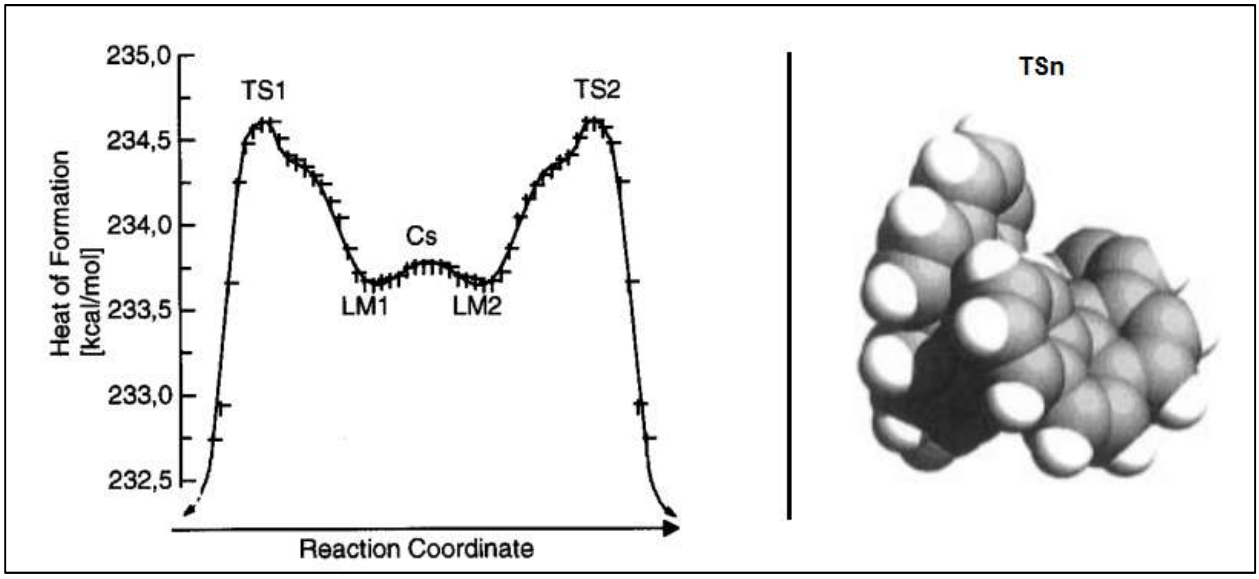

Figure 5: On the left, a plot illustrating the relation of the Cs structure in relation to the transition states for [9]helicene. On the right, a space-filling rendition of the transition state. TS1 and TS2 are mirror images of one another. [1996]

These results agree with calculations performed by Grimme and Peyerimhoff [1994]. In Grimme and Peyerimhoff's study [1994] they explore multiple methods semi empirical SCF methods of computing the racemization energies. They provide transition states for 4, 5, 6, and 8 helicene, which are visually summarized in figure 6 below.

In addition to the transition states, Grimme and Peyerimhoff [1994] discuss some structural characteristics of Helicene that result in the observed racemization barrier. The increasing energy barrier with n is a result of an increase in non-bonded atom interactions between the terminal rings. The observed plateau of the energy barrier with n > 6 is attributed to the flexibility of the benzene ring for twisting, which results in the distance between terminal in the transition state to have no further decrease.

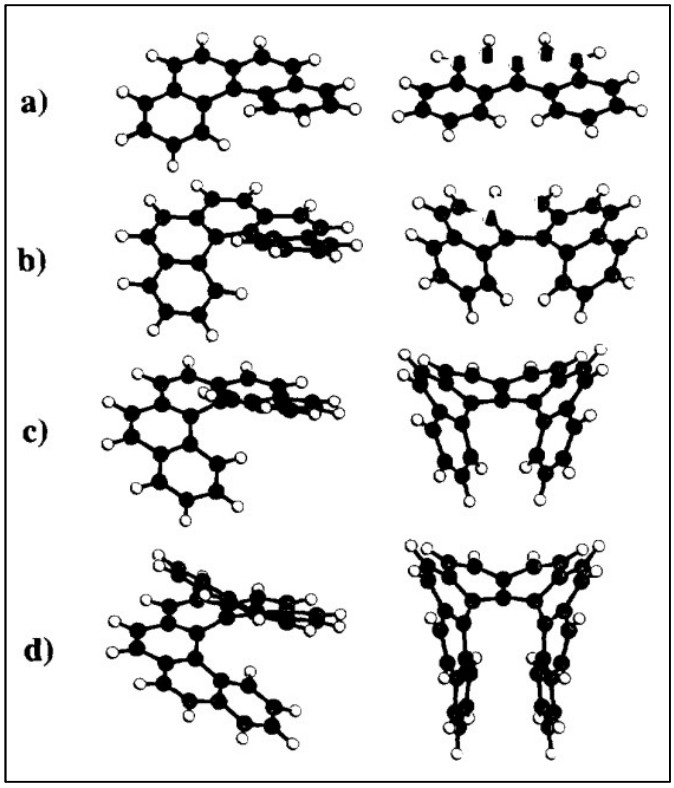

Figure 6: On the left, the equilibrium state. On the right, the corresponding transition state. a) [4]helicene b) [5]helicene c) [6]helicene d) [8]helicene [1994]

In 2009, Sehnal et. Al., investigated the synthesis of long helicenes. In their analysis they computed the racemization energy and the racemization pathway of [11]helicene. The transition pathway for [11]helicene has the same form as [9]helicene. The c_s-structure is at an approximate local minimum between two mirror image transition states. However, in the case of [11]helicene, the results of Sehnal et. Al. show that the difference in energy between the c_s-structure and the transition states is much greater than for [9]helicene as determined by Janke, et al. [1996]. The increasing energy difference between the c_s-structure and transition state with increasing ring count, as seen from [9]helicene to [11]helicene is a possible new avenue to explore. Sehnal et. Al. [2009] also comment that the conformational process between the two equilibrium states is complex.

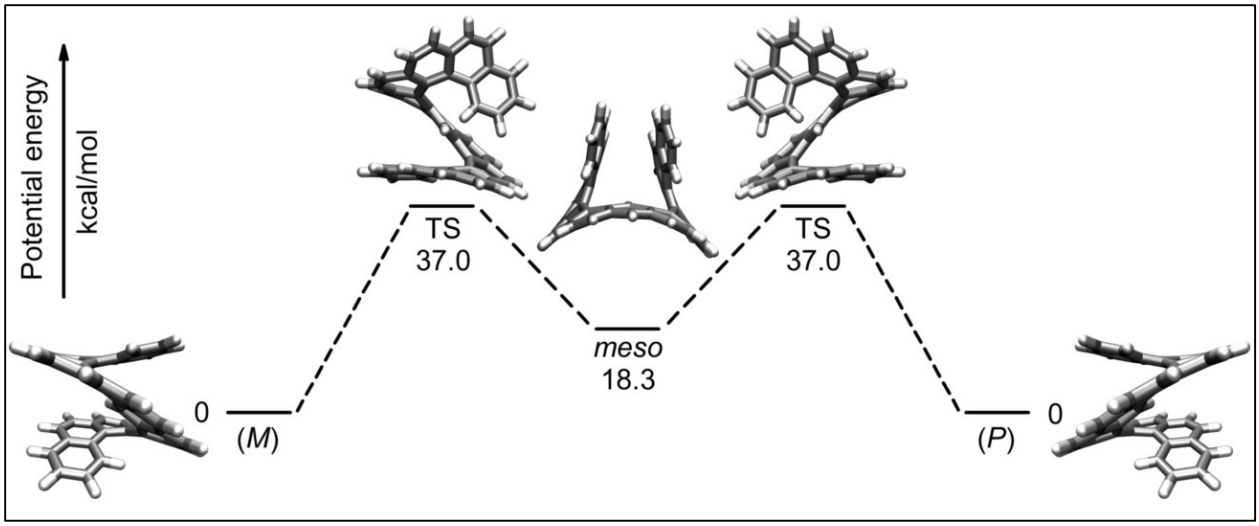

Figure 7: The calculated pathway of racemization by Sehnal et. al. for [11]helicene.

To summarize, these three studies provide information on the geometry of the transition state and its relation to the c_s-symmetric plane. For [n]helicenes with n < 9, the c_s-structure was the transition state, while for [9] and [11]helicene the c_s-structure was essentially a minimum between two mirror image transition states. This minimum was much deeper for [11]helicene. These discussed results will prove useful to compare to. The effectiveness of the model proposed in this Thesis project will be evaluated by its ability to predict the correct transition structure and number of peaks along the reaction coordinate in addition to the value of the energy of racemization.

1.4 Implications

The aim of this research is to test a model of helicene based on the bending potential between rings and the repulsive interaction between every sixth ring. The terminal ring steric interaction is generally referenced as the sole source of the helicity of helicenes (Ravat et. al. 2017, Ben Yahia et al. 2017). While it is clear from the structures of naphthalene and phenanthrene that it is energetically favourable for adjoined benzene rings to lie in a plane without bending. Thus, if these the two interactions considered in the model are not enough to account for the behaviour of helicene (i.e. produce the proper free energy profile over the reaction coordinate between stable states), then there are other unconsidered interactions that are important. The goal of this research is to determine if these two interactions sufficiently describe the molecule so the question of whether other interactions must be considered first before expanding the model to handle more complicated cases (such as the addition of functional groups).

2 Methods

2.1 The Model of Helicene

Helicenes were modelled with a coarse-grained model that tracks the positions and orientation of the benzene rings that constitute the molecule. This model approximates the shape of each benzene ring as a perfect hexagon with a carbon atom at each vertex. The model does not allow for bond lengths or the shape of the hexagon to vary. The model only allows for the bending of two adjacent rings by some angle, A. Thus, a conformation is uniquely defined by the set of angles between all adjacent rings. However, this set of all angles is not enough to compute the energy of a conformation, of which one energy term depends on the absolute distance between every sixth ring.

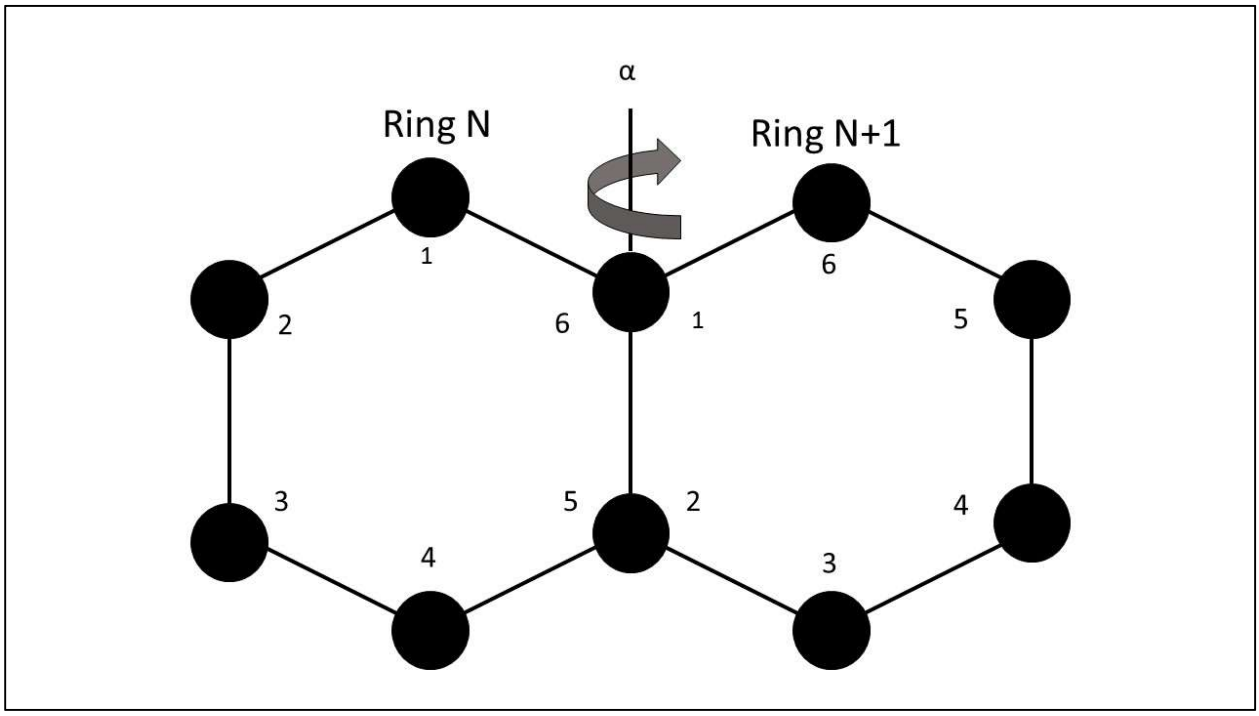

Figure 8: A visual description of the model for Helicene. 6 carbon atoms at the vertices of a symmetric hexagon, carbons 5 and 6 corresponding to carbons 2 and 1 respectively on the next ring. The rings may bend at the joint between the two rings by an angle α. Note that carbon 1 of a ring is always carbon 6 in the adjacent ring, likewise for the 6th ring down the chain (the rings circle around every 6 rings)

The two major contributing factors to the structure of helicene are the well documented steric interaction between the terminal rings (Chen and Shen, 2017) and the bending potential between rings. The bending potential is the repulsive interaction between rings as they bend towards each other at the joint. The structures of benzene chains of only two and three rings, known as naphthalene and phenanthrene, are planar and have slight thermal oscillations (Cruickshank, 1957). This suggests the energy of an angle between rings is at a minimum at 0 radians and

increases with increasing angle. The energy contribution to a conformation due to the angles between rings can then be expressed harmonically as:

$$E = \sum_i \alpha {A_i}^2 \tag{2.1}$$

Where, α is the bending energy parameter, A the angle, and I the index for the position of the angle.

The interaction between terminal rings and overlapping rings were modelled with an inverse distance potential. This potential is applied for every sixth ring pair or if none exists, the two end rings. Two different methods of computing the distance were tested. The first computed the distance from the midpoints of the overlapping rings. The second method computed the distance from the first carbon atom of the N^{th} ring to the sixth carbon atom of the (N+6) ring. The motivation for the second method is that these are overlapping carbon atoms when the rings circles around. In addition, it is documented that the steric interaction between hydrogens on these carbons is much more significant than the interaction between hydrogens bonded to carbons in the 5 and 2 positions (Martin, 1974). Both methods use the following general formula for the energy of the sixth ring interaction:

$$E = \sum_j \frac{\beta}{D_j} \tag{2.2}$$

Where, β is the distance energy parameter, D the distance between 6^{th} ring pairs, and J the index referring to which pair. The total energy of a conformation is then expressed as:

$$E_{total} = \sum_i \alpha {A_i}^2 + \sum_j \frac{\beta}{D_j} \tag{2.3}$$

To compute this energy and have a full description of the conformation of the molecule all the positions of carbon atoms, i.e. vertices of the hexagonal rings, must be computed. This procedure is outlined in appendix A3.

2.2 Preliminary tests of Model and Code

Two preliminary tests of the model were conducted to ensure the simulation was functioning as expected. The first test was a Metropolis Monte Carlo simulation of the molecule in the absence of both potentials. In this test the calculation of the energy step was omitted from the procedure and the energy of each conformation set to zero. This has the effect that the simulation will accept every randomly generated conformation. This test was done for chains of three different lengths: 6, 9, and 12 rings, for 100 000 iterations each. The root mean square (RMS) end-end distance and average angles were computed for each simulation.

The second test was a Metropolis Monte Carlo simulation where the sixth ring interaction potential was omitted and only the bending potential included in the energy function. This test was done for chains of 6 and 12 rings. The bending potential parameter was varied from 1 to 1,000,000 in powers of 10. The test was performed 5 times for each chain length and bending potential value. The number of iterations performed was determined during testing and set to be the number required for the 5 independent tests to converge to the same RMS end-to-end distance value within a small error.

2.3 Determining Equilibrium States

Equilibrium states of the model were examined using the Metropolis procedure in Monte Carlo experiments. Equilibrium state properties can be determined by averaging the results of sampling after the Markov Chain being sampled converges to its steady state solution. For the purposes of this study, convergence was assumed for the last 1000 steps of a 100 000 step simulation. The equilibrium properties of interest depended on the averaged equilibrium structure. To find the averaged equilibrium structure, the set of final angles was averaged over the 1000 steps. The set of all angles includes all the degrees of freedom of the conformation molecule and therefore a satisfactory means for averaging the equilibrium structure.

For each averaged equilibrium structure computed the set of carbon atom positions, set of averaged angles, z-axis end to end distance, ratio of accepted conformations, and the gyration tensor were output to text files. The set of carbon atom positions for the averaged equilibrium structure is determined by evaluating the set of averaged angles with the molecule construction algorithm outlined in Appendix A.3. Z-axis end to end distance is evaluated by taking the difference in the z-component of the positions of the 1st position carbon of the first ring and the 6th position carbon of the last ring. These carbons positions are selected as they are overlapping positions for adjacent rings, on the interior of the ring, and the carbon position pair for which the 6th ring pair interaction potential is computed. The gyration tensor is evaluated from the set of carbon atom positions for the averaged equilibrium structure. The components of the gyration tensor are the second moments of position, which are determined via the following equations.

$$\bar{r} = \begin{bmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{bmatrix} = \frac{1}{N} \sum_{i}^{carbons} \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} \tag{2.4}$$

$$T_{mn} = \frac{1}{N} \sum_{i}^{carbons} (r_m{}^i - \bar{r}_m{}^i)(r_n{}^i - \bar{r}_n{}^i) \tag{2.5}$$

Equation 1 is used to compute the average position of the equilibrium structure. This is necessary for performing the required change of coordinates to the coordinate system where the average position of the structure is zero. In this coordinate system the components of the gyration tensor are computed as the averaged sum of products between components of positions. Equation 2 is

the expression for components of the gyration tensor in the coordinate system that carbon positions are given in from the simulation.

Equilibrium conformations for helicenes ranging in length from 4 rings up to 42 rings were computed. A variety of energy parameter values were tested with [6]helicene, of which a value of 1000 in energy units of kT for both the bending and distance energy parameters was determined to be used for all testing. This was selected by qualitative comparison of the equilibrium structure predicted by the model to that which is documented. Qualitative comparison of the predicted structures to the true structures served as one of the two methods used for evaluating the accuracy of the predictions. This qualitative comparison was performed by plotting the carbon positions in molecular visualization software. Then visually comparing the structure to the true equilibrium structure.

In addition to qualitative comparison, the gyration tensor was diagonalized and the asphericity and acylindricity of each predicted equilibrium structure computed. The diagonal components of the diagonalized gyration tensor are used to define the asphericity and acylindricity:

$$b = \lambda_z^2 - \frac{1}{2}(\lambda_x^2 + \lambda_y^2) \tag{2.6}$$

$$c = \left(\lambda_y^2 - \lambda_x^2\right) \tag{2.7}$$

Where b is the asphericity and c the acylindricity. The diagonal components of the gyration tensor were computed numerically with the Wolfram alpha computation engine.

Lastly, the two equilibrium states are mirror images of each other. Thus, a set of angles defining one equilibrium state should give the mirror image equilibrium state when it is inverted. This assumption was tested by taking the set of angles from a simulation of [6]helicene, inverting the set, and performing a new simulation using the inverted set as the starting location. The simulation was run for the standard number of steps and averaged the same as all other simulations. Change in angle from the starting set to the final averaged set were compared to determine how close the inverted set was to the mirror image equilibrium state.

2.4 Determining Free Energy Barriers

The free energy barrier for [6] helicene was computed using the umbrella sampling method. The distance along the z-axis between the terminal rings was selected as the reaction coordinate. This is a natural choice for the reaction coordinate for this molecule. The ends of the molecule must slide past each other and their z-positions invert to move from one equilibrium state to the mirror equilibrium state. The reaction coordinate was divided into 20 evenly spaced windows of sampling. Each window was sampled for 5 and 10 million iterations. During the simulation, successful conformations had their reaction coordinate value output to a file. Post simulation the files containing the list of reaction coordinate values were processed to create new files with the counts of reaction coordinates in bins along the reaction coordinate. Total number of 50, 100,

200, and 400 bins along the reaction coordinate were used. These files contained the probability of a certain width of reaction coordinate in a particular window and formed the set of biased probability distributions. The weighted histogram analysis method (WHAM) was used to determine the unbiased probability distribution from the set of biased probability distributions. The convergence algorithm used to iteratively solve the WHAM equations is outlined in appendix A.4. With the unbiased probability distribution, the relative free energy of a width of reaction coordinate can be evaluated via:

$$A(\xi) = -\beta \ln\big(P(\xi)\big) \tag{2.8}$$

Where ξ is the width of reaction coordinate, and P(ξ) the unbiased probability for the width of reaction coordinate ξ.

3 Results and Discussion

3.1 Preliminary test results

In the absence of a potential, the simulation sampled many conformations at random without any bias to the sampling. For all helicene lengths tested, the RMS end-to-end distance approached 1.41. This result was generally unaffected by whether the distance was measured from the midpoint of the end rings or the 1 and 6 carbon positions of the end rings. In a typical random walk of a chain, the RMS end-to-end distance will increase proportionally to the root of the number of units in the chain:

$$\mathrm{R_{ee}} \propto \sqrt{\mathrm{N}} \tag{3.1}$$

The observed RMS end-to-end distance can be explained in that the helicene chain does not do a true random walk. In a random walk, at every step, the direction in which the next unit is placed is random and independent of previous steps. While in the absence of a potential, the next ring in the helicene chain is free be at any angle relative to the previous, the geometry of how the rings are adjoined to each other forces them to circle around in a ring. A change in angle at one joint between rings only changes the plane through which the ring arcs. For example, an angle of π for one of the joints would reverse the direction of the arc (i.e. the direction new rings are added onto the chain). Combined with the fact that these angles are chosen at random, and for the most part cancel out, the center of the arc will not change. As a result, the rings will almost always lie on an arc sharing the same radius as the original arc, only lying on a rotated plane. For example, in the case where the center of the arc does change, as in a helix, the angles between units are uniform. Thus, it would make sense that the RMS approaches roughly the same value for the chains of varying length, because all the chains are roughly bound within the same radius. This radius being determined by the geometry of the rings.

The second test, a Metropolis Monte Carlo simulation including only the bending energy potential yielded an exponential decay relationship between the RMS end-to-end distance and the strength

of the bending energy parameter. The relationship between the average RMS end-to-end distance of 5 trials per energy parameter value and the logarithm base 10 of the energy parameter value is given by:

$$< R_{ee} >= 1.3507e^{-1.033\xi} \tag{3.2}$$

Where R_{ee} is the RMS end-to-end distance, and ξ the logarithm base 10 of the energy parameter. The fit of this equation to the data is close, with an r-squared value of 0.9962. Qualitatively, this behaviour is expected. The energy parameter multiplies the energy of a conformation, and thus, also the energy difference between two conformations. As a result, it becomes less probable to sample the states with larger angles, due to the exaggeration of the energy difference by the large parameter. Qualitatively, the exponential nature of this decay can be explained as a result of entropy. An exponentially increasing number of states are available as the end-to-end distance is increased. Therefore, to restrict sampling to the few states with small end-to-end distances, a proportionally small probability factor must be applied to the vast number of states with an end-to-end distance that is large. Thus, as the fraction of small to large end-to-end distance states decreases exponentially as the criteria for small gets smaller, the probability bias for those states must increase exponentially. The results of this test are summarized in the plot below:

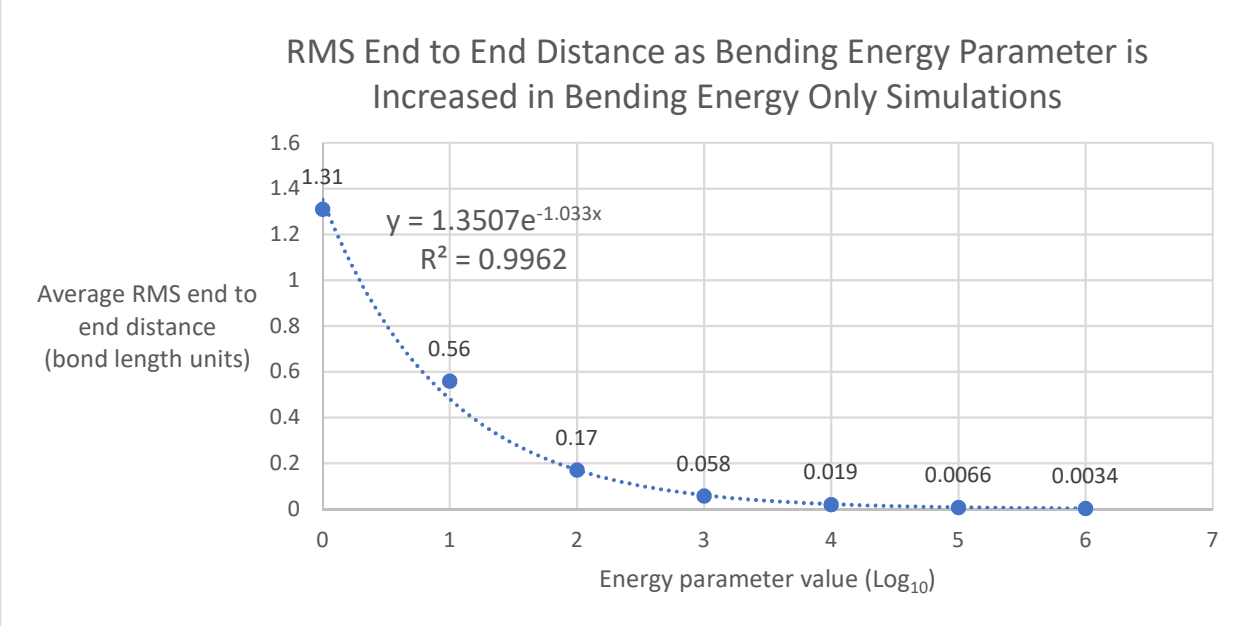

Figure 9: A plot of the RMS end-to-end distance as the bending energy parameter is increased for [6]Helicene. The exponential trend line shows that the RMS decreases by a factor of e for every order of magnitude larger the energy parameter becomes.

These preliminary results show that the model and methods used are functioning as they are expected to.

3.2 Equilibrium States

3.2.1 Acylindricity and Asphericity

Metropolis Monte Carlo simulation found that the model was effective at predicting the equilibrium state for [6] helicene. For longer helicenes, the model predicted non-physical equilibrium states, and for shorter helicenes it predicted planar molecules.

The acylindricity and asphericity are good proxy measures to judge if the model is behaving as expected. A helicene should be imagined as an inflexible rod. The acylindricity should remain relatively constant and zero, since extension of the chain should not change the cylindrical nature of the helix. However, asphericity should increase as the length of the chain increases, since the width of the helix becomes arbitrarily small compared to the length of the helix.

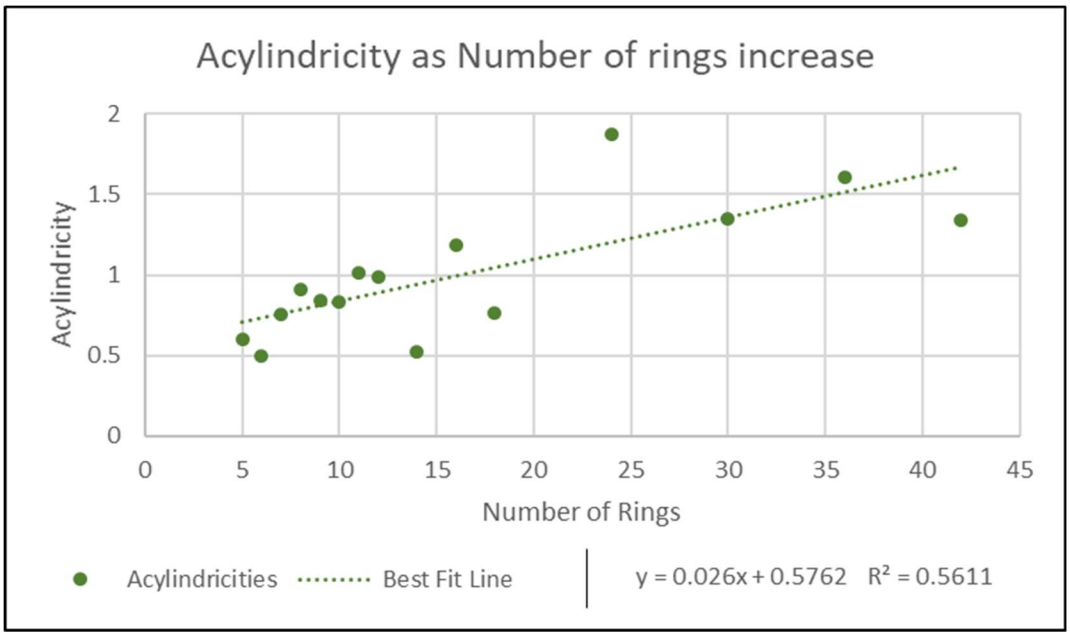

Figure 10: Acylindricity of the helicenes predicted by the model as the number of rings in the chain increases.

The acylindricity behaves approximately as expected. There is a small increase in the acylindricity as the number of rings increases, however, the acylindricity may be reaching a plateau with the longer helicenes. The sampling of longer helicenes is poor, and the trend of the longest 4 helicenes shows a decreasing trend. It would be expected that as the length of the chain increases any fluctuation in the x-y position of the chain contributing to the acylindricity would be increasingly negligible, and the acylindricity trend towards zero.

The asphericity behaves contrary to what is expected of it. As opposed to increasing it decays as the length of the chain increases. This result is understandable in the context with the qualitative visual inspection of the predicted structures. Long chains folded into themselves, resulting in the increase in sphericity.

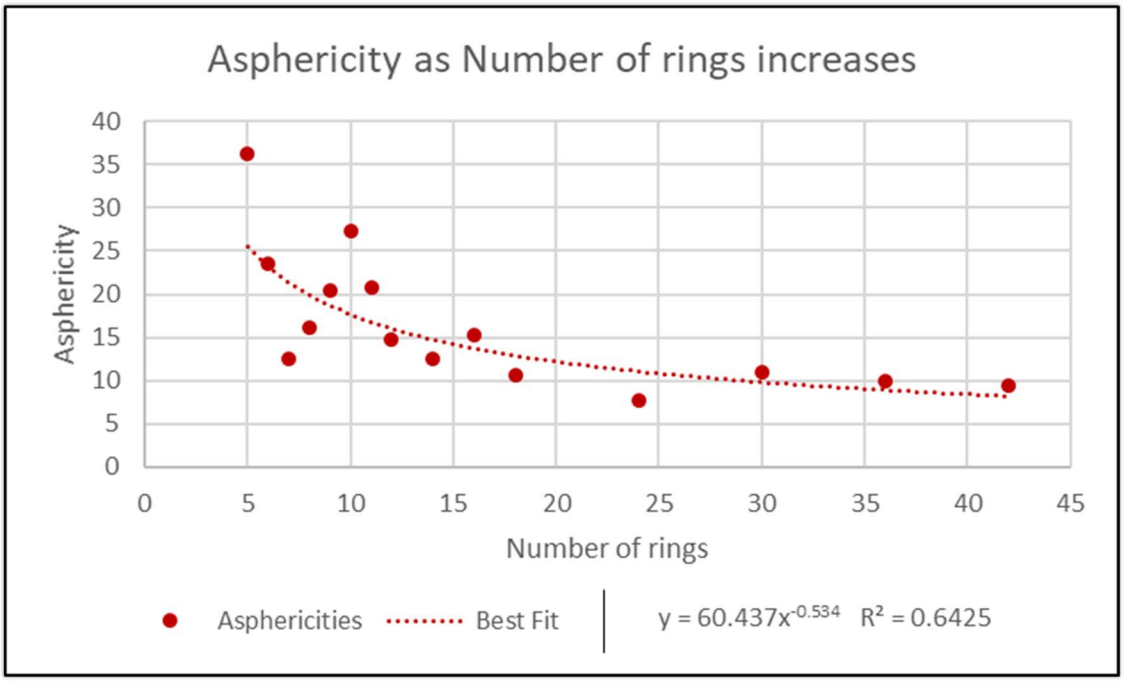

Figure 11: Plot of the asphericity of helicenes predicted by the model as their number of rings increases.

3.2.2 Equilibrium structure visualization

Visual inspection reveals that the model is accurate when predicting the equilibrium conformation of [6] helicene and properly predicts that the inverse set of angles of an equilibrium state produces another equilibrium state.

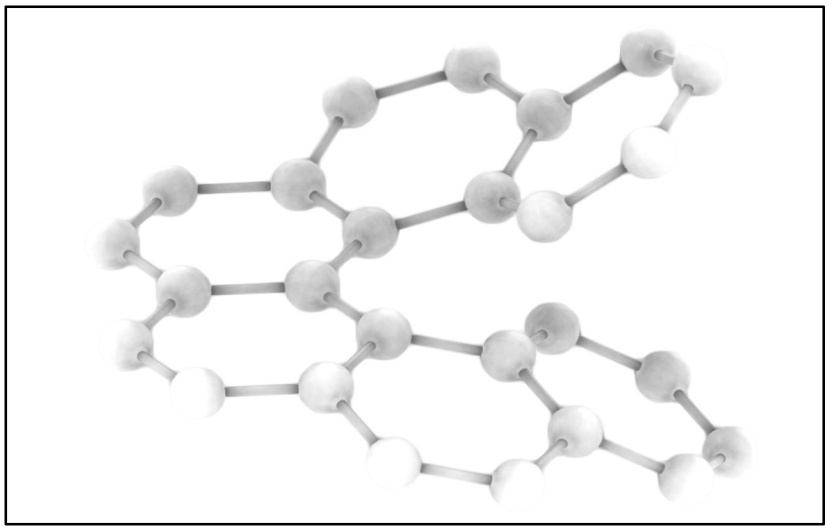

Figure 12: Predicted equilibrium conformation for [6] helicene. The characteristic separation between the terminal rings was properly predicted.

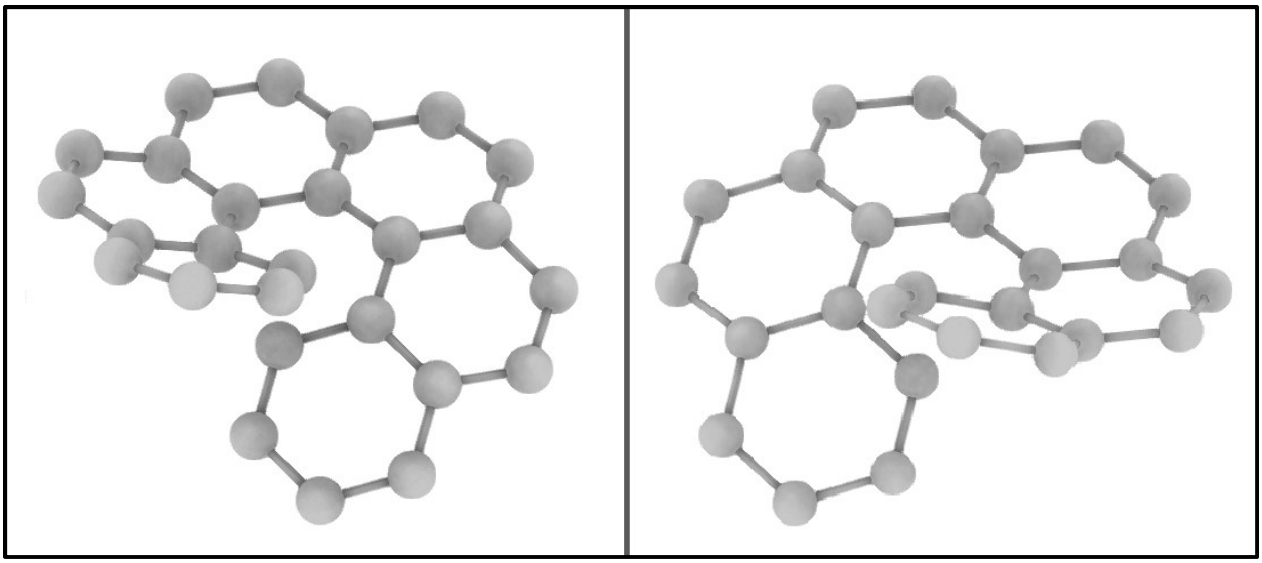

Figure 13:Predicted right and left-handed equilibrium structure. Structure on right produced by simulation starting with the inverted set of angles of the left structure

While the predictions for [6]helicene had a good fit, visual inspection of the longer helicenes reveals deviations from reality. The predicted structures of the long chains were non-physical conformations in which the chain travelled through itself.

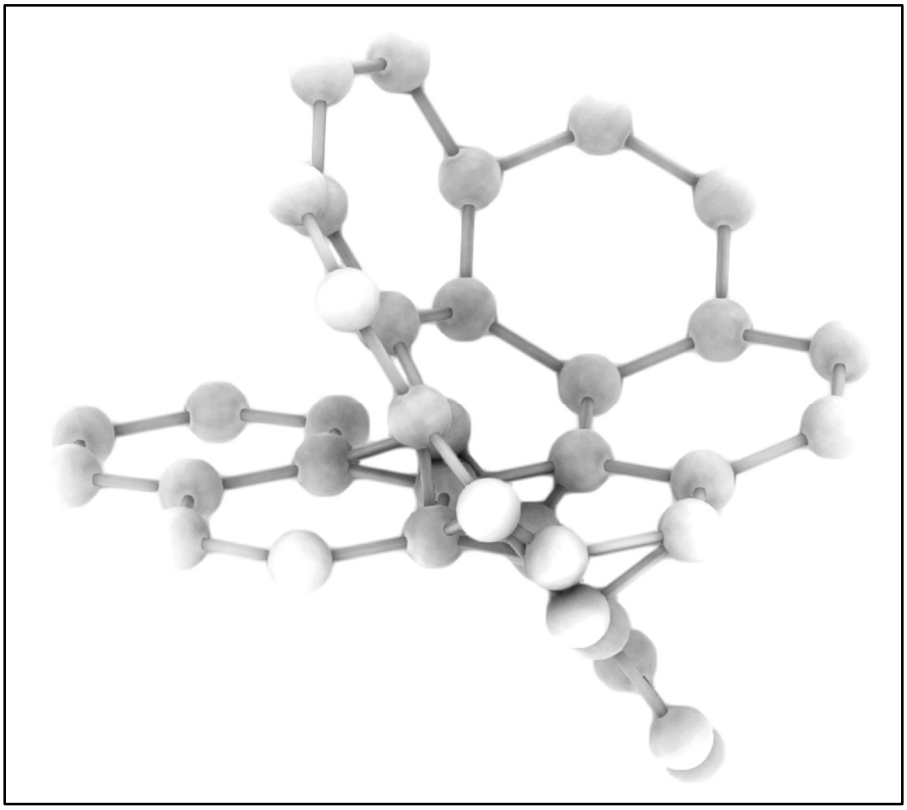

Figure 14: Predicted structure of [9] helicene. [9] helicene exemplifies the behaviour of the chain travelling through itself observed in the longer chains.

In figure 14, the carbon atoms of the two rings that pass through are very close to each other. However, the 6th ring pairs are well separated. The model did not include any hard-core interaction preventing any carbon from being within a small distance of any other carbon atom. As a result, the model is allowing some non-physical states to exist where the chain can travel through itself. Some of these non-physical states being conformations in which the 6th ring pairs are well separated and the average angle low, resulting in a small energy for the non-physical conformation. The effect of the lack of a hard-core interaction is increasingly apparent for longer helicenes. The very long helicenes are predicted to be ordered and condensed by the model.

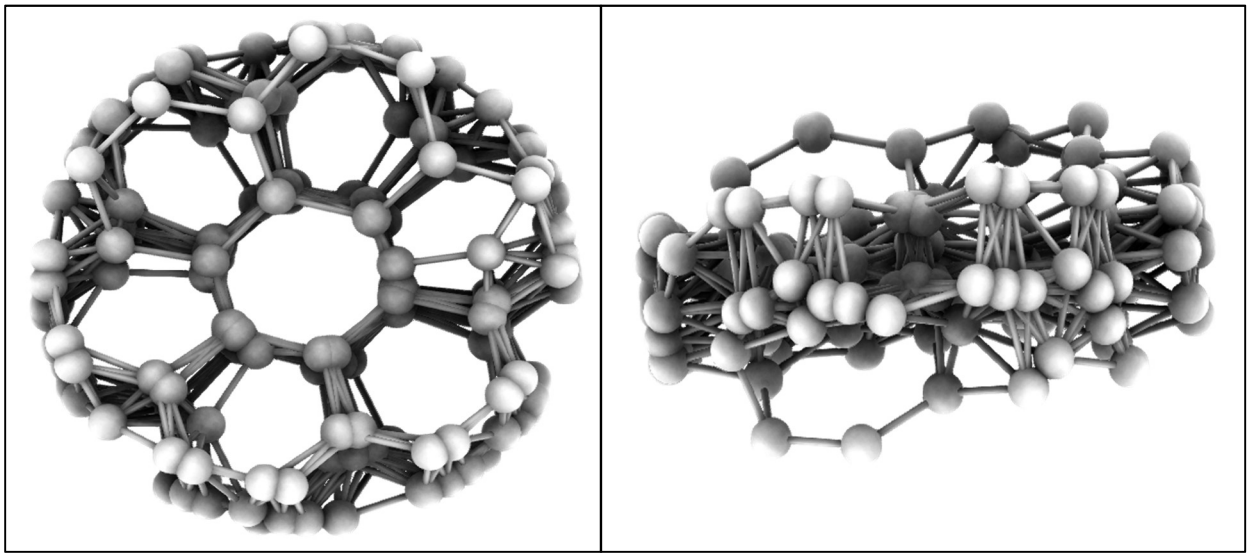

Figure 15: Top view (left) and side view (right) of [42] helicene. [42] helicene exhibits a condensed structure and maintains an ordered structure

It is significant that the very long helicenes, despite not having verticality, have an ordered structure. It is further evidence that the lack of hard-core interactions allows for alternative non-physical low-energy states. Furthermore, a very specific set of angles is required to produce the helical shape associated with the true equilibrium state, while condensed state depicted in figure 15 is highly degenerate. The degeneracy of the predicted equilibrium conformation of [42] helicene was inferred from the fact that each of 3 different simulations of [42] helicene yielded a different set of angles for the equilibrium conformation. This contrasts with the predictions for [6] helicene in which the same set of angles were produced every time.

A possible explanation for the apparent degeneracy of the very long helicenes in the model is due to limitations in the implementation of the simulation. The Metropolis Monte Carlo procedure implemented only used small changes in angle between steps. In addition, the length of the simulation was only 100 000 steps. The combination of a limited sample time and no method for traversing large energy barriers in phase space renders it highly unlikely the simulation would travers from the first local energy minima it reaches to another, let alone the global energy minima. As a result, it is unlikely that the computed equilibrium structure for the very long

helicenes are the true equilibrium structure the model predicts, rather they are metastable states predicted by the model.

For the very short helicenes, [4],[5] helicene, the predicted structures are planar when using the same combination of energy parameters as were used for all other helicene lengths. The simulation could be made to predict some verticality when the distance energy parameter was made orders of magnitude larger than the angle parameter. This is no solution however, the strengths of interactions considered are independent of the length of a helicene. A possible explanation is the approximation made in the model that the distance for the 6^{th} ring pair interaction potential is measured from carbon atom to carbon atom. Whereas, the steric interaction is mostly a result of the hydrogen or functional group attached to the ring (Martin 1974, Gringas). In the case of [4] and [5] helicene, these hydrogens/functional groups extend from the ring inward into the space between the two terminal rings in the same plane as the rings. It is the proximity of these hydrogens/functional groups that results in the steric interaction. Thus, by measuring the distance between two carbons, 1 from each terminal ring, the steric interaction for

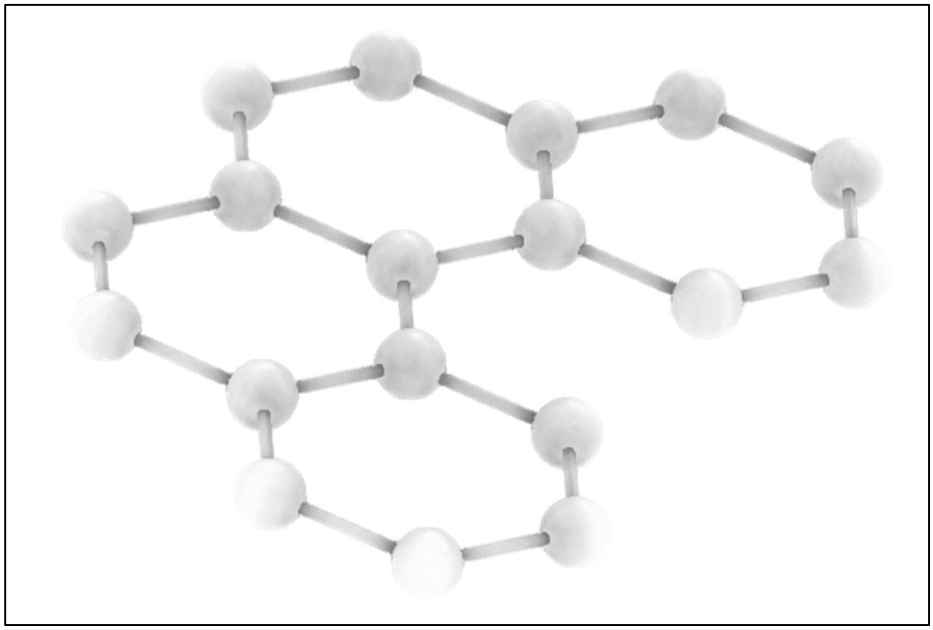

Figure 16: The predicted structure of [4]helicene when using the same energy parameter combination as for all other helicene simulations. The predicted structure exhibits extremely planar topology in contrast the slight verticality the true structure exhibits

[4] and [5] helicene is not being accurately modelled in comparison to how well it is modelled for [6] helicene. In the case of [6] helicene, the steric interaction is also between hydrogens/functional groups. However, another approximation made in the model, that the rings are rigid and circle around perfectly every 6 rings, places the two carbons from the terminal rings right on top of each other.

3.3 Umbrella Sampling results

The equilibrium z-axis distances of the two equilibrium states of [6] helicene are approximately -0.9, and 0.9. In the sampling experiment the reference reaction coordinate varied from positive to negative 1.2, to sample the free energy slightly past the equilibrium state. This choice of reaction coordinate was unnecessarily large. The z-distance was never sampled at a value of greater than 1 even when using an arbitrarily large bias parameter. Thus, it was concluded that there existed a geometrical limit of 1, to the distance a conformation could attain along the z-axis between the two carbon atoms on the terminal rings.

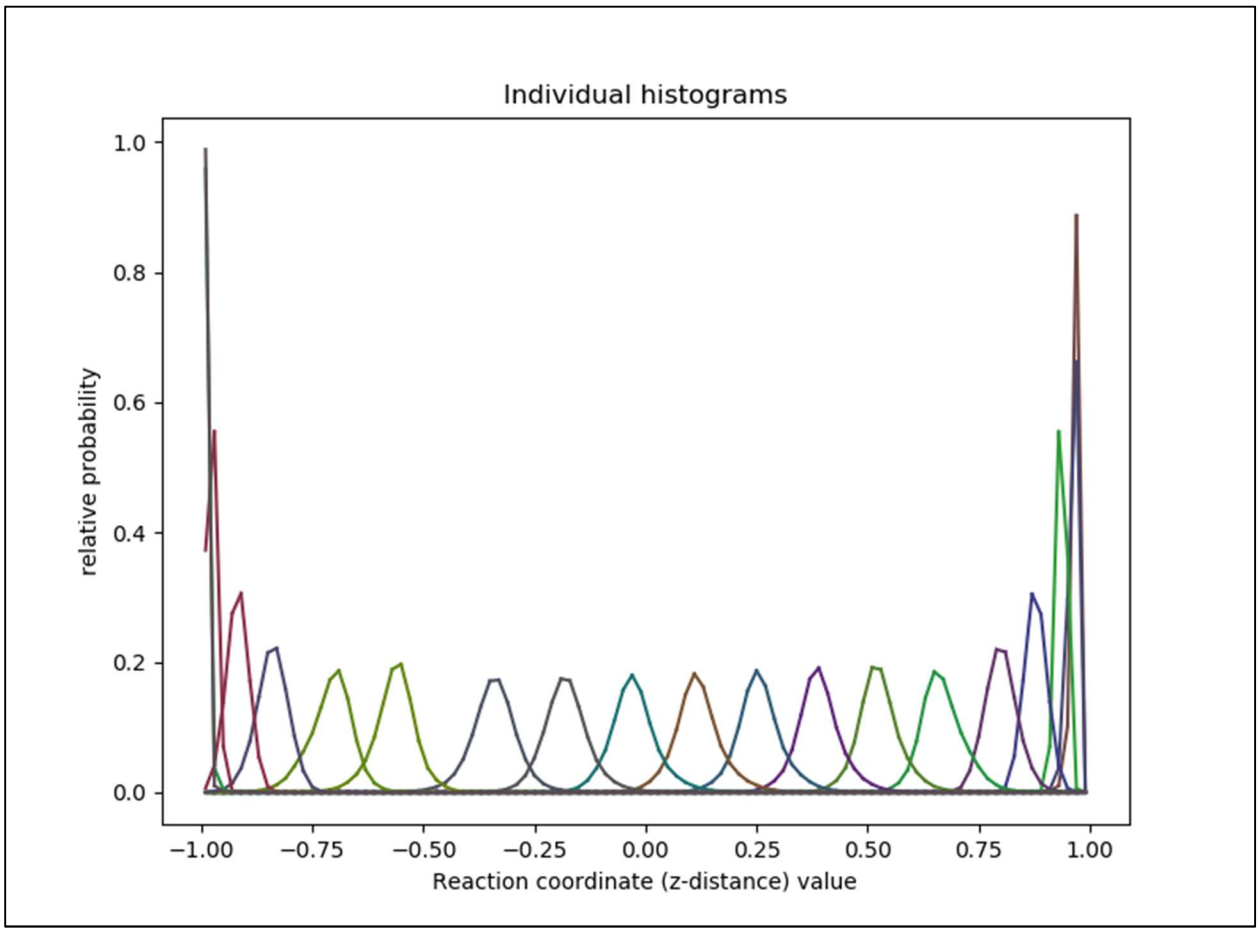

Figure 17: Plot of all 20 histograms corresponding to the probability distributions in individual windows. There is a fair degree of overlap between windows everywhere except around reaction coordinate value of -0.5

The sampling experiment had a good spread over the entire reaction coordinate. There was significant overlap between individual windows except around reaction coordinate value of -0.5. However, the calculated free energy curve was smoothest around this value of reaction coordinate indicating the relatively smaller frequency of sampling at -0.5 was not an issue.

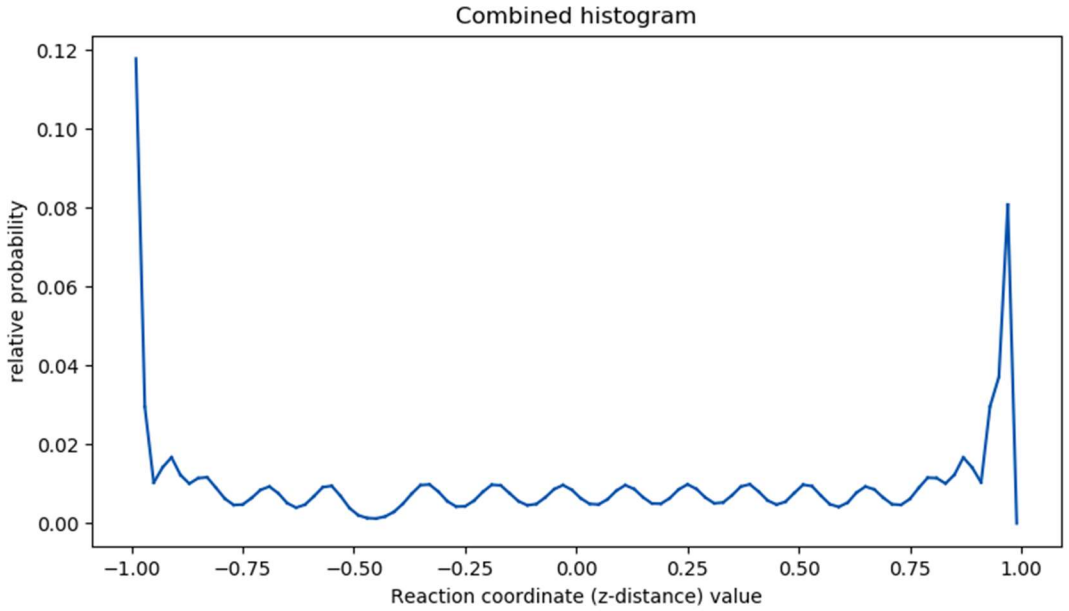

Figure 18: Combined histogram compiled from all 20 window histograms. Apart from the ends the distribution of sampling is relatively even. From -0.25 to 0.75 the curve exhibits the same periodic amplitude.

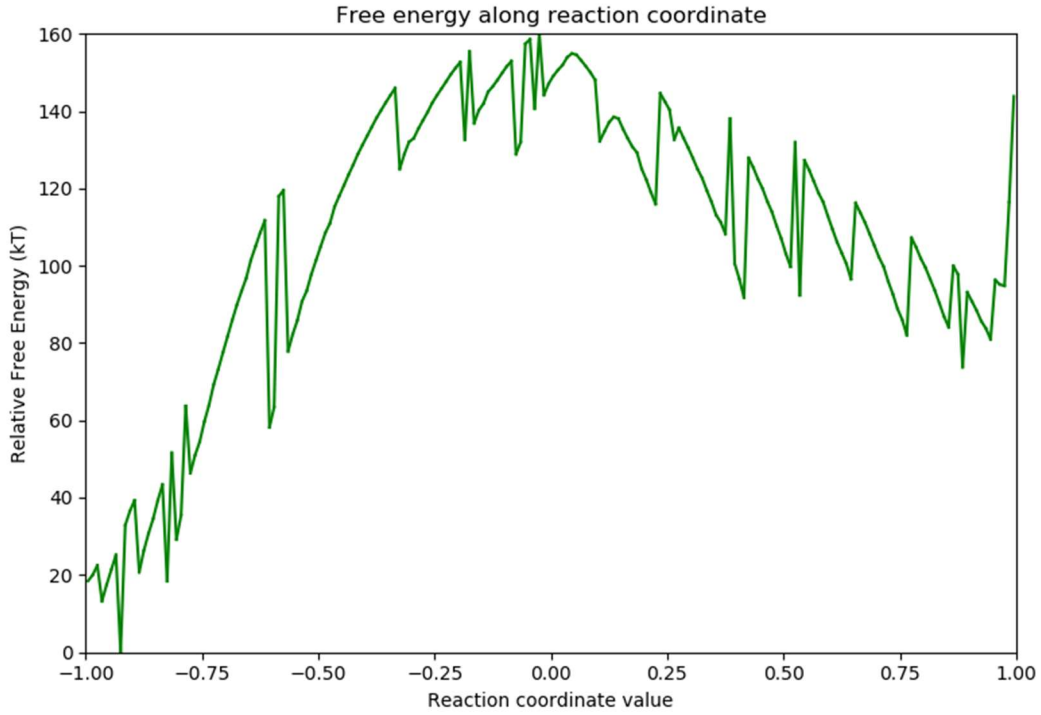

Figure 19: Plot of the free energy along the reaction coordinate for simulation using 10 million iterations per window and 200 bins along the reaction coordinate. Reaction coordinate of approximated -0.91 and 0.91 correspond to the equilibrium states

The calculated free energy curve between the two mirror image equilibrium states is not as expected. The two mirror image equilibrium states should not have a free energy difference between them. The produced curve has approximately an 80 k_bT/mol difference in free energy between the two equilibrium states. Although, the general scale of 10s of k_bT/mol is the same scale as the measured energy barrier of racemization of ~61 kT/mol (Chen and Shen, 2017). In addition, the presence of many local minimums is consistent with the idea that a lack of hardcore interactions allows for non-physical conformations with low energy. However, the jaggedness of the plot may also be a symptom of too little sampling.

Despite the difference in calculated free energy between the equilibrium states, both are still local minima. As well the general shape of a W, holds. The energy peaks directly between the two states at a reaction coordinate of approximately 0. The free energy peak is expected at 0 as it is the point where the ends rings are closest together and thus a very high energy region due to the inverse distance potential between the end rings.

It was observed that the WHAM equations would either converge in a very short number of iterations or reach a minimum difference between each iteration and not improve from step to step. Within 3-7 iterations, the WHAM equations would converge to within 10^{-4} or less for most calculations, but often never converge smaller than 10^{-6}.

4 Summary

Helicenes are a chain of benzene rings characterized by their helical structure (Gringas, 2013). A statistical mechanics model was proposed to describe the conformation of the molecule. The model described the energy of a conformation of helicene to depend upon the bending between rings, and an inverse distance potential between every 6^{th} ring pair. The molecule was represented as a series of rigid hexagons in a Monte Carlo simulation used to calculate the equilibrium conformation and the free energy profile along the reaction coordinate. Equilibrium structures were computed for helicenes of lengths 4 to 42 using the Metropolis procedure. The model predicted the structure for [6] helicene well, but for longer chains it predicted non-physical structures in which the chain moves through itself. The addition of a hardcore interaction to prevent any carbon atom from being too close to any other carbon atom should prevent the simulation from sampling non-physical conformations and produce more realistic and accurate structures. The free energy curve along the reaction coordinate between the equilibrium states of [6] helicene was computed using the umbrella sampling method. The free energy curve had the peak expected between the equilibrium states and was in the correct order of magnitude for energies, but the calculation predicts different free energies for the equilibrium states. The next step for developing this model is to include the hardcore interaction and observe the effect it has. As well to repeat the free energy calculation.

Bibliography

Ben Yahia, M., Tounsi, M., Aouaini, F., Knani, S., Ben Yahia, M. and Ben Lamine, A. (2017). A statistical physics study of the interaction of [7]-helicene with alkali cations (K+ and Cs+): new insights on microscopic adsorption behavior. *RSC Advances*, 7(71), pp.44712-44723.

Beveridge, D. (1989). Free Energy Via Molecular Simulation: Applications To Chemical And Biomolecular Systems. *Annual Review of Biophysics and Biomolecular Structure*, 18(1), pp.431-492.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4), pp.327-335.

Cowles, M. and Carlin, B. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434), p.883.

Chen, C. and Shen, Y. ed., (2017). Structures and Properties of Helicenes. In: Helicene Chemistry. Berlin Heidelberg: Springer, pp.19-44.

Cruickshank, D. (1957). A detailed refinement of the crystal and molecular structure of naphthalene. *Acta Crystallographica*, 10(8), pp.504-508.

Feringa, B. (2001). In Control of Motion: From Molecular Switches to Molecular Motors†. *Accounts of Chemical Research*, 34(6), pp.504-513.

Gingras, M. (2013). One hundred years of helicene chemistry. Part 1: non-stereoselective syntheses of carbohelicenes. *Chem. Soc. Rev.*, 42(3), pp.968-1006.

Gingras, M. (2013). One hundred years of helicene chemistry. Part 3: applications and properties of carbohelicenes. *Chem. Soc. Rev.*, 42(3), pp.1051-1095.

Gingras, M., Félix, G. and Peresutti, R. (2013). One hundred years of helicene chemistry. Part 2: stereoselective syntheses and chiral separations of carbohelicenes. *Chem. Soc. Rev.*, 42(3), pp.1007-1050.

Grimme, S. and Peyerimhoff, S. (1996). Theoretical study of the structures and racemization barriers of [n]helicenes (n = 3–6, 8). *Chemical Physics*, 204(2-3), pp.411-417.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1), p.97.

Janke, R., Haufe, G., Würthwein, E. and Borkent, J. (1996). Racemization Barriers of Helicenes: A Computational Study. *Journal of the American Chemical Society*, 118(25), pp.6031-6035.

Kästner, J. (2011). Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6), pp.932-942.

Kumar, S., Rosenberg, J., Bouzida, D., Swendsen, R. and Kollman, P. (1992). THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8), pp.1011-1021.

Martin, R. (1974). The Helicenes. *Angewandte Chemie International Edition in English*, 13(10), pp.649-660.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, 44(247), p.335.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6), pp.1087-1092.

PubChem (2004). PHENANTHRENE. [online] Pubchem.ncbi.nlm.nih.gov. Available at: https://pubchem.ncbi.nlm.nih.gov/compound/phenanthrene#section=Top [Accessed 15 Nov. 2017].

Ravat, P., Hinkelmann, R., Steinebrunner, D., Prescimone, A., Bodoky, I. and Juríček, M. (2017). Configurational Stability of [5]Helicenes. *Organic Letters*, 19(14), pp.3707-3710.

Reist, M., Testa, B., Carrupt, P., Jung, M. and Schurig, V. (1995). Racemization, enantiomerization, diastereomerization, and epimerization: Their meaning and pharmacological significance. *Chirality*, 7(6), pp.396-400.

Torrie, G. and Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2), pp.187-199.

Appendices

A Code Algorithms

A.1 Rotation Algorithm

The rotation algorithm is a core component of the build algorithm. The Rotation algorithm takes an arbitrary axis of rotation and rotates a given vector a specified angle about that axis of rotation. The algorithm accomplishes this by rotating the frame such that the given rotation axis lies upon the z-axis in the rotated frame and then applies the basic z-axis rotation matrix before rotating the frame back to the original coordinate system. This can be summarized in the following matrix equation:

$$V' = {R_x}^{-1}{R_y}^{-1}R_z(\theta)R_yR_xV \tag{A.1}$$

Where V' is the vector after the rotation, R denotes a rotation matrix with the lower index denoting the axis of rotation. In this equation only R_Z was shown to depend on theta and this is because the other 4 rotation matrices do not depend on the angle of rotation but instead on the rotation axis vector. The rotations about the x and y axes serve the purpose of rotating the frame such that the rotation axis is on the z-axis, thus the angle of rotation for these matrices depends on the rotation axis vector. The components of these 4 rotation matrices depending on the rotation axis can be derived as follows.

The rotation algorithm used begins with the basic rotation about the x-axis that aligns the projection of the rotation vector on the y-z plane to the z-axis. The magnitude of the projection of the rotation axis onto the y-z plane is given by:

$$d = \sqrt{b^2 + c^2} \tag{A.2}$$

Where b is the y component and c the z component of the rotation axis vector. This magnitude can then be used along with b and c to define the cosine and sine of the rotation angle. The desired rotation angle is that that rotates the projection to the z-axis, thus the sine and cosine of that angle can be determined via Pythagorean theorem to give:

$$\cos\theta = \frac{c}{d} \tag{A.3}$$

$$\sin\theta = \frac{b}{d} \tag{A.4}$$

A basic rotation about the x-axis in the positive sense is given by:

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \tag{A.5}$$

And using the relations for cosine and sine given in equations A.3 and A.4 it can be written as:

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{c}{d} & -\frac{b}{d} \\ 0 & \frac{b}{d} & \frac{c}{d} \end{bmatrix} \tag{A.6}$$

With the corresponding inverse matrix to be used when reverting to the original coordinate system:

$$R_x{}^{-1} = R_x{}^{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{c}{d} & \frac{b}{d} \\ 0 & -\frac{b}{d} & \frac{c}{d} \end{bmatrix} \tag{A.7}$$

In the case the rotation axis vector lay on the x-axis, i.e. d = 0, this rotation about the x-axis is omitted and only the y-axis rotation is performed.

After the x-axis rotation, the rotation axis will lie solely in the z-x plane and the frame must be rotated about the y-axis to align the rotation axis to the z-axis. Noting that the length along the z-axis is now d and applying Pythagorean theorem as before the following relations are given:

$$\cos\theta = \frac{d}{u} \tag{A.8}$$

$$\sin\theta = \frac{a}{u} \tag{A.9}$$

Where u is the magnitude of the rotation axis, a is the x-coordinate of the rotation axis vector and d the magnitude of the projection of the rotation axis vector onto the y-z plane. A rotation in the direction from x-axis to z-axis is the negative sense y-axis rotation as determined by the right-hand rule and is given by:

$$R_y = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \tag{A.10}$$

With the cosine and sine relations from equation A.8 and A.9, the y-axis rotation and its inverse rotation can be written:

$$R_y = \begin{bmatrix} \frac{d}{u} & 0 & -\frac{a}{u} \\ 0 & 1 & 0 \\ \frac{a}{u} & 0 & \frac{d}{u} \end{bmatrix} \quad and \quad R_y{}^{-1} = \begin{bmatrix} \frac{d}{u} & 0 & \frac{a}{u} \\ 0 & 1 & 0 \\ -\frac{a}{u} & 0 & \frac{d}{u} \end{bmatrix} \tag{A.11}$$

Lastly a rotation about a z-axis in the positive sense is given by:

$$R_z = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (A.12)$$

Thus, in full, the expression for a rotation about an arbitrary axis of rotation:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{c}{d} & \frac{b}{d} \\ 0 & -\frac{b}{d} & \frac{c}{d} \end{bmatrix} \begin{bmatrix} \frac{d}{u} & 0 & \frac{a}{u} \\ 0 & 1 & 0 \\ -\frac{a}{u} & 0 & \frac{d}{u} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{d}{u} & 0 & -\frac{a}{u} \\ 0 & 1 & 0 \\ \frac{a}{u} & 0 & \frac{d}{u} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{c}{d} & -\frac{b}{d} \\ 0 & \frac{b}{d} & \frac{c}{d} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad (A.13)$$

In practice, this equation is applied as a series of calculation steps handling one matrix at a time alternating between two sets of variables for the coordinates. Two sets of variables are used to overcome the issue that the calculation of x, y, and z will depend on their values of the previous step in the calculation (i.e. you cannot change y and then use that changed y in the calculation for z).

A.2 Sampling Algorithm

The Sampling algorithm is the process of sampling the next point of conformation space. Sampling involves 3 main steps. These include selecting at random the angle to adjust, applying a random change, determining pass or fail of new state, and the 4th minor step is reverting change if necessary.

A key process in the sampling algorithm is the generation of random numbers. Random numbers on a uniform integer distribution were generated using the Mersenne Twister algorithm from the <random> C++ library. Random numbers generated with the algorithm were converted onto an interval of 0 to 1 by dividing by the maximum possible random integer. When a random number is referred to throughout the remainder of this section it refers to the random number on the interval 0 to 1 produced in this way.

The first sampling step is selecting at random the angle to adjust. This is accomplished in the following equation where Position is an integer and so the equation is truncated down to the lowest integer:

$$Position = R * (N - 1 - \xi) \quad (A.14)$$

Where R represents the random number, N is the number of rings in the chain, and ξ, a very small number. This formula takes advantage of the truncation of integers and that indices for the array of angles begin at 0. The purpose of ξ is for the case when R = 1. The largest desired number is N – 2, so ξ ensures that for R = 1 the number is slightly smaller than N – 1 and so becomes truncated down to N – 2.

The second step is applying a change of random magnitude to the selected angle. The maximum possible change is manipulated with a parameter, α, and is chosen so that the number of trial

moves that pass are in the range of 10 - 20%. The change must be equally probable to induce a negative change to the angle and the following formula is used to create a random change on the interval -α to α.

$$A = A + \alpha(2 * R - 1) \quad (A.15)$$

Where A is the angle, R the random number and α the parameter.

After applying the random change, the angle must be checked to be within the interval -π to π. While angle is periodic, the energy function depends on the square of the angle and so all angles must be in terms of the principal angle. This is done by checking for a value greater than or lower than π and if true, correspondingly subtracting or adding 2π.

The final step is determining if the trial passes or is rejected. First the energy is computed. The energy function as discussed in the methods depends on 2 potentials of the general form:

$$E_s = \sum_i \alpha {A_i}^2 + \sum_j \frac{\beta}{D_j} \quad (A.16)$$

Where E is the energy of a state, α is an angle parameter, A is the angle, β is the distance potential parameter, D is the distance, index I is for the angles, and index J is for all 6th ring interactions. The angle sum can be computed directly from the array of angles, however to get the 6th ring interaction potential term, the conformation the angles represent must be built (i.e. all the carbon positions computed). Details of this procedure are in the Build Algorithm section of the appendix.

After the energy is computed it is compared to the energy of the previous state. If lower, it is automatically accepted. If the energy is higher, a probability of passing is produced by taking the energy difference and computing the Boltzmann factor associated with this energy difference. The probability factor takes the form:

$$P = e^{-\frac{\Delta E}{kT}} \quad (A.17)$$

Where ΔE is the energy difference between current and previous state, and kT the usual Boltzmann constant – temperature factor. This probability factor is then compared to a random number. If the probability factor is greater, then the trial state is accepted, else it is rejected.

In the case the step is rejected, the angle and the calculated energy need to be reverted to the previous state. In practice this is achieved by storing the position and old angle of the changed angle and when reversion is required the angle at the stored position is set to the old angle and the energy is set to the energy of the previous state.

A.3 Molecule Construction Algorithm

The molecule build algorithm is the method in which all the positions of the carbon atoms in the molecule are computed. The method begins with the first ring and prescribes the first two carbons of the ring in fixed positions, and from here computes the rest using the array of angles defining the conformation. The calculation of each ring's carbon positions is the repeatable unit of the method. Aside from the first ring, all other rings are added onto the chain in the same way, repeating the same procedure.

Calculating the positions of the carbons in the first ring is the initialization step. The positions of the first two carbon atoms are given as (0,0,0) and (1,0,0) and another vector, called the perpendicular vector, is given as (0,1,0). The perpendicular vector is a vector lying in the plane of the ring that is orthogonal to the vector between the two given carbon atoms, which is called the orientation vector. The orientation vector is in the direction of carbon one to carbon two. The cross product of the perpendicular vector and the orientation vector will yield a normal vector to the plane of the ring. This normal vector is used as an axis of rotation to rotate the orientation vector by sixty degrees four times. Each time the orientation vector is rotated its value is added to the value of the most recently calculated carbon position. For example, the first time it is rotated its value is added to the position of the second carbon atom, producing the position for the third carbon atom. This process accomplishes the task of placing the remaining 4 carbon atoms at the proper positions: the vertices of the hexagon. After this is completed the positions for the first ring have been computed and this ring can be used as the base for the general procedure of adding another ring.

The general procedure for adding a ring shares many of the steps in the initialization process but is generalized to include rotation of the plane of the ring and receive the starting points and the perpendicular vector from the previous ring. By nature of the way that the rings are adjoined, carbons one and two of the next ring are the same as carbons six and five respectively of the previous ring (view figure 20 for clarification). These starting points are then used to calculate the orientation vector in the same way as in the initialization step. The perpendicular vector can also be calculated by taking the vector from carbon 3 to 5 in the previous ring and normalizing it.

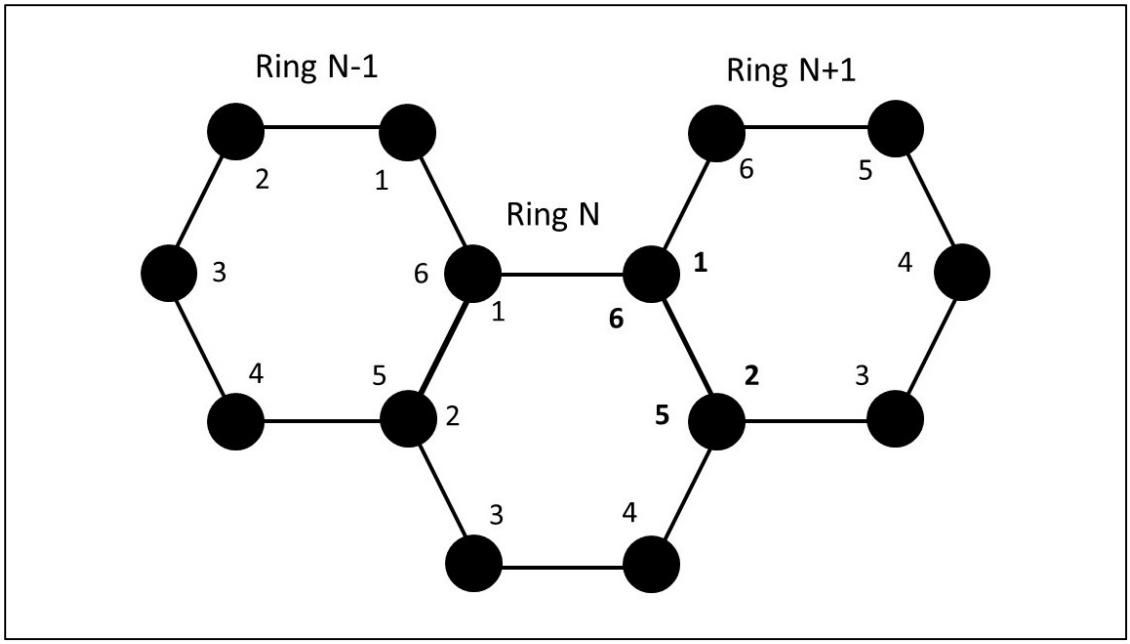

Figure 20: A diagram showing the numbering system for carbons in a ring. The number labels for shared carbons between rings N and N+1 have been bolded for clarity.

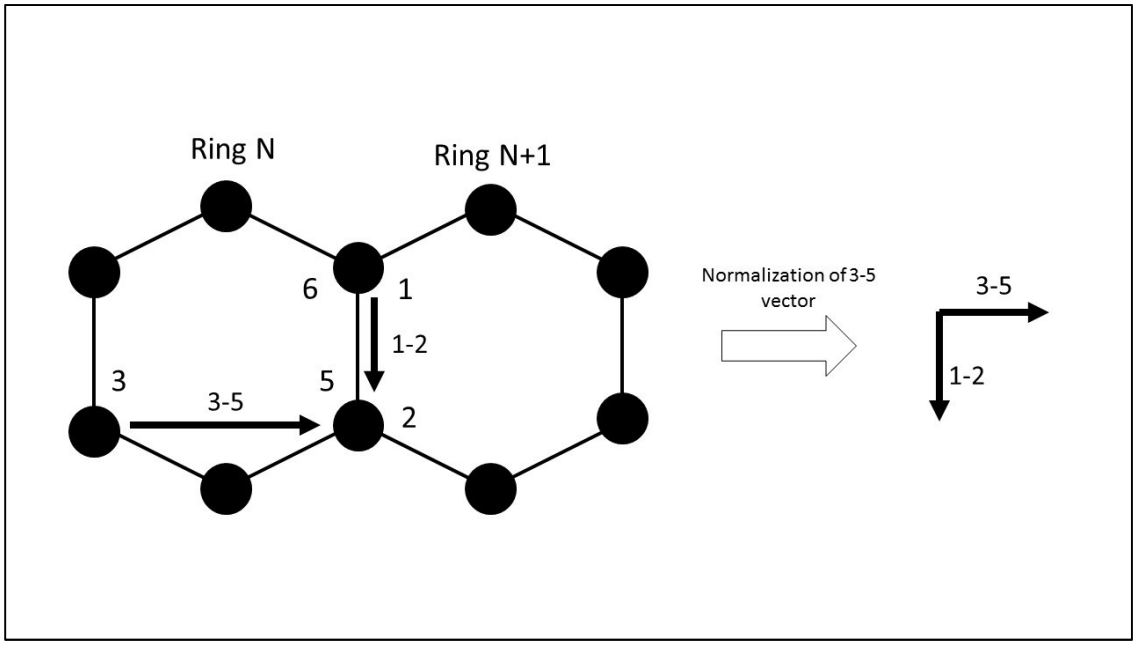

Figure 21: A visual description of how the perpendicular vector and orientation vector for the new ring are determined from the old ring. These two vectors span the plane of the new ring.

However, before computing the normal vector for the new ring, the perpendicular vector is rotated about the orientation vector by the amount prescribed for that joint of the chain. This adjust the plane of the new ring and creates the proper angle between rings N and N+1 as given in the array of angles. The remaining 4 carbons of the new ring are computed with the same procedure as the initialization method. A normal to the plane of the ring is created by taking the

cross product of the orientation and the perpendicular vector. The orientation vector is then rotated around the normal vector 60 degrees in the positive sense 4 times sequentially. Every time adding its value to the last calculated carbon atom position to calculate the next carbon position. After this has been completed all 6 carbon atom positions of the ring will have been computed and lying on the vertices of the hexagon that lies in the plane that is at the prescribed angle to the plane of the previous ring. The procedure can then be repeated for another ring using this newly calculated ring as the reference until all rings in the chain have been computed.

A.4 WHAM Convergence algorithm

The unbiased probability distribution is determined by iteratively recalculating the following two equations until convergence has been achieved.

$$P(\xi) = \sum_{i}^{windows} \frac{n_i(\xi)}{N_i e^{\beta(F_i - w_i(\xi)}} \qquad (A.18)$$

$$F_i = -\frac{1}{\beta} \ln \left[\sum_{\xi}^{bins} P(\xi) e^{[-\beta w_i(\xi)} \right] \qquad (A.19)$$

The calculation begins with an estimation of the values of F_i. Estimates of the following form are used:

$$F_i^{est} = b^2 \alpha \qquad (A.20)$$

Where b is the bin width, and α, the strength parameter for the biasing potential function. With this estimate the global distribution is calculated via eqn. 1. and then the corresponding F_i values are calculated. This process is iterated until the maximum difference between any of the probabilities for a specific ξ bin, P(ξ), is below a specified convergence threshold value. Values of P(ξ) are normalized before comparison to the previous step's value of P(ξ).

B Python Plotting Scripts

B.1 Plotting Multiple histograms

```
# code to plot multiple histograms together with data taken from a text file

import numpy as np
import matplotlib.pyplot as plt

directory = "C:/thesisdata/umbrellasampling/"
file = "multiplehisto.txt"

data = np.loadtxt(directory+file,dtype=np.float64, unpack='true', delimiter=",")

x = data[0,:]
colors = [0.1, 0.1, 0.1]
for k in range(1, len(data[:,3]), 1):
    y = data[k,:]
    colors[0] = np.random.rand()
    colors[1] = np.random.rand()
    colors[2] = np.random.rand()
    sumed = colors[0] + colors[1] + colors[2]
    colors[0] = colors[0] / sumed
    colors[1] = colors[1] / sumed
    colors[2] = colors[2] / sumed
    usecolor = tuple(colors)
    print(colors)
    for i in range(0, len(x), 2):
        plt.plot(x[i:i+2], y[i:i+2], color=usecolor)
        plt.plot(x[i+1:i+3], y[i+1:i+3], color=usecolor)
plt.title('Individual histograms')
plt.ylabel('relative probability')
plt.xlabel('Reaction coordinate (z-distance) value')
plt.show()
```

B.2 Plotting single histograms

```
#plotting free energy and reaction coordinate

import numpy as np
import matplotlib.pyplot as plt

directory = "C:/thesisdata/umbrellasampling/"
file = "freeenergy6100bin.txt"

x, y = np.loadtxt(directory+file, dtype=np.float64, unpack='true', delimiter=",")

#plt.scatter(x, y, s=4, c='red')
for i in range(0, len(x), 2):
    plt.plot(x[i:i+2], y[i:i+2], 'g-')
    plt.plot(x[i+1:i+3], y[i+1:i+3], 'g-')
plt.axis([-1, 1, -0.01, 300])
plt.title('Relative probability of sampling along reaction coordinate from all
windows')
plt.ylabel('relative probability')
plt.xlabel('reaction coordinate value')
#plt.title('Free energy curve for [6]helicene')
plt.show()
```

C Derivations

C.1 Free energy difference expression

The following expression will be derived for a free energy difference:

$$\Delta F = -kTln\langle \exp(-\Delta E\beta)\rangle \qquad (C.1)$$

To begin, free energy is expressed as the following function of the partition function:

$$F = -k_b Tln(Z) \qquad (C.2)$$

A difference is expressed as:

$$\Delta F = F_1 - F_2 \qquad (C.3)$$

$$\Delta F = -k_b Tln(Z_1) + k_b Tln(Z_2) \qquad (C.4)$$

$$\Delta F = -k_b Tln\left(\frac{Z_1}{Z_2}\right) \qquad (C.5)$$

The partition function is the sum of the Boltzmann factors over all states of the system:

$$Z = \int e^{-\beta E} \qquad (C.6)$$

$$\Delta F = -k_b Tln\left(\frac{\int e^{-\beta E_1}}{\int e^{-\beta E_2}}\right) \qquad (C.7)$$

Introducing a factor of unity into the numerator:

$$1 = e^0 = e^{-\beta E_2 + \beta E_2} \quad (C.8)$$

$$\Delta F = -k_b T ln\left(\frac{\int e^{-\beta E_1} e^{-\beta E_2 + \beta E_2}}{\int e^{-\beta E_2}}\right) \quad (C.9)$$

$$\Delta F = -k_b T ln\left(\frac{\int e^{-\beta E_1 + \beta E_2} e^{-\beta E_2}}{\int e^{-\beta E_2}}\right) \quad (C.10)$$

Since,

$$-\beta E_1 + \beta E_2 = -(E_1 - E_2) = -\Delta E \quad (C.11)$$

$$\Delta F = -k_b T ln\left(\frac{\int e^{-\beta \Delta E} e^{-\beta E_2}}{\int e^{-\beta E_2}}\right) \quad (C.12)$$

The factor inside the logarithm is an ensemble average over system 2 of the following quantity:

$$A = e^{-\beta \Delta E} \quad (C.13)$$

$$< A > = \int \frac{A e^{-\beta}}{Z} \quad (C.14)$$

Recalling equation N for Z, equation NN simplifies to:

$$\Delta F = -kT ln\langle \exp(-\Delta E \beta) \rangle \quad (C.15)$$

As desired.