The Building Game and Other Discrete Geometric Models for the Self-assmbly of Polyhedral Molecular Structures

by

Daniel C. L. Johnson

B.S., Rensselaer Polytechnic Institute; Troy, NY, 2009

Sc.M., Brown University; Providence, RI, 2012

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2015

This dissertation by Daniel C. L. Johnson is accepted in its present form by The Division of Applied Mathematics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Govind Menon, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Committee Member II, Ph.D., Reader

Date_____ _____

Committee Member III, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter Weber, Dean of the Graduate School

Vitae

Acknowledgements

Abstract of " The Building Game and Other Discrete Geometric Models for the Self-assmbly of Polyhedral Molecular Structures " by Daniel C. L. Johnson, Ph.D., Brown University, May 2015

Contents

Vitae **vii**

Acknowledgments **ix**

1 Introduction **1**
1.1 Scientific Motivation 2
1.1.1 Self-Folding Polyhedra 3
1.1.2 Molecular Cages 3
1.1.3 Viral Capsid Assembly 5
1.1.4 RNA and Protein Folding 5
1.2 Mathematical Work Inspired by Self-Assembly Processes 6
1.2.1 The Building Game 6
1.2.2 Folding 8
1.2.3 Local Rules 9
1.2.4 Dominant Formation Pathways 10
1.2.5 Markov Processes 10
1.2.6 Disconnectivity Graphs 13
1.3 Prior Work 14
1.3.1 Configuration Space 14
1.3.2 Constrained Dynamics 16
1.4 Original Contributions 17

2 The Building Game: Modeling **19**
2.1 The Building Game as a Mathematical Framework for Self-assembly . 20
2.2 Formal Definition 20
2.2.1 Group Actions 22
2.2.2 Building Game Intermediates 24
2.2.3 Group Theoretic Results 29
2.3 Stochastic Modeling Results 32
2.3.1 Stationary Distribution 33
2.3.2 Hitting Times 35

3 The Building Game: Enumeration **41**
3.1 Known Enumerative Results . . . 42
3.2 New Enumerative Results . . . 42
3.2.1 Shellability . . . 42
3.2.2 Bounds and Asymptotics . . . 43
3.3 Computational Methods . . . 47
3.3.1 Hash Functions . . . 48
3.3.2 Data Structures . . . 50

4 Constraint Models and Embedding Intermediates in Space **51**
4.1 Linkages . . . 52
4.2 Geometric Configuration Space . . . 52
4.2.1 Computing the Degrees of Freedom of a 3D Linkage . . . 52
4.3 Cyclohexane Application . . . 55
4.3.1 Sachse Model . . . 55
4.4 Idealized Constraint Model . . . 56
4.4.1 Degrees of Freedom in Ideal Model . . . 57
4.5 Folding Configuration Space . . . 57
4.5.1 Methods . . . 58
4.6 Results . . . 61

5 Processes in Constraint Spaces **63**
5.1 Constrained Dynamics . . . 64
5.1.1 Cyclohexane Dynamics . . . 64
5.1.2 Configurations . . . 64
5.1.3 Finding Intermediate Coordinates . . . 65
5.1.4 Dynamics . . . 66
5.1.5 Transitioning Between Boat Configurations . . . 67
5.1.6 Analogy to Folding Model . . . 69
5.2 Manifold Reflected Brownian Motion . . . 70
5.2.1 Computational Implementation . . . 70
5.2.2 Validation and Test Cases . . . 70
5.2.3 MRBM on Building Game Geometric Configuration Spaces . . . 70

6 Results **71**

List of Tables

List of Figures

1.1 Dodecahedron building game example. 7
1.2 The state space of the cube. 7
1.3 A disconnectivity tree for the 38-atom Lennard-Jones cluster. [from Wales et al] 13

2.1 The building game states of the tetrahedron. 21
2.2 Examples of octahedron states and non-states. 22
2.3 The stabilizer subgroups for various octahedron states. 25
2.4 One Building Game pathway for the Octahedron. 27
2.5 The Building Game combinatorial configuration space of the cube. . . 28
2.6 Degeneracies between two connected cube intermediates. 29

3.1 Building game enumerative results for the Platonic solids. 42
3.2 Building game enumerative results for the Archimedean solids. 43
3.3 Building game enumerative results for the Catalan solids. 43
3.4 Building game enumerative shellability results for the Platonic solids. 44
3.5 Building game enumerative shellability results for the Archimedean solids. 44
3.6 Building game enumerative shellability results for the Catalan solids. 45
3.7 Algorithm for interatively enumerating the Building Game combinatorial configuration space. 48

5.1 Boat to Boat2 Coordinate Transition 68
5.2 Configuration Graphs for Cyclohexane and Folding 69

Chapter One

Introduction

1.1 Scientific Motivation

Self-assembly is a class of formation process in which a product is constructed with out explicit manipulation of its parts. Many—often identical—parts come together by utilizing the dynamics of their environment to create a finished structure. Sometimes such assemblies can be encouraged can be encouraged by manipulating broadly controlled parameters of the assembly environment such as temperature and solution content.

There are many examples of both natural and synthetic self-assembly process that take a wide variety of length scales and serve a plethora of functions. An area of much active research is the self-assembly of RNA, proteins, and viral capsids. These biological processes act on the nano scale and, while it is known that their formation can be aided by a variety of secondary mechanisms such as helper RNA, the formation process is not well understood in general. Synthetic examples of molecular self assembly include the formation of supramolecular cages. These supramolecular structures that can encapsulate a smaller molecule show promise in contributing to a number of medical and scientific fields including drug delivery and nano-scale circuits.

In general, the processes of biological self-assembly are not well understood due to difficulties arising from their small length scale. Conversely, synthetic self-assembly processes often lack the complexity and sophistication of their biological equivalents. The goal of our research is to explore discrete geometric models of self assembly. By using analysis made possible by the simplicity of said models, important properties of the processes are to be identified. Of primary importance is identifying the pathways of formation which consist of a specific order in which unfinished intermediate

states are visited before the assembly is completed. It is thought that these pathways are often robust in the sense that the process follows a very small, and sometimes unique, number of pathways to form the end product. Also of interest is identifying mechanisms by which failed or flawed formations can be avoided or minimized. Possible strategies include the selection of specific precursors, the selection of solution, and the mid-assembly control of experimental parameters.

1.1.1 Self-Folding Polyhedra

In experiment, Pandey et al have been able to form closed polyhedral structures from the self-folding of flat polyhedral nets. Using photolithography, these nets are cut from a two dimensional sheet with great precision. By adding a specific amount of solder at the net's hinges, surface tension from the melted solder causes the net to fold up into the closed polyhedron. Several different polyhedra have been attempted with varying degrees of success. It has been argued that the geometric structure of the polyhedron are largely determinant of the net's propensity to successfully assemble into the polyhedron. In some cases, two distinct polyhedral isomers can be formed from the same net.

1.1.2 Molecular Cages

Self-assembly of molecular cages are the subject of much active research. By isolating molecules of interest inside a molecular cage, targeted application of these molecules is theoretically possible. This has enormous implications in the medical industry. Additionally, by creating a grid-like network of such cages, it may be possible to construct functional electric circuitry at the nano scale.

Fujita et. al. have theorized and subsequently synthesized a family of organometallic cages consisting of metallic connector molecules (M) that each connect four bent ligand molecules (L). With the M molecules as vertices and the L molecules as edges, they can form polyhedral cages. Due to geometric constraints, the number of M and L molecules in a completed cage must satisfy $|M| = k$ and $|L| = 2k$ for $k \in 6, 12, 24, 30, 60$. Each of these five choice of k results in a cage that geometrically resembles a specific Archemedian solid.

In one experiment, two different ligands (L_1 and L_2) with slightly different bend angles were used. As the ratio of $L_1 : L_2$ was varied, each ratio only resulted in the formation of one of the possible cages. More specifically, for ratios $L_1 : L_2 < 0.25$ only $M_{12}L_{24}$ formed and for $L_1 : L_2 > 0.25$ only $M_{24}L_{48}$ was formed. This steep and curious cutoff thus far defied tangible explanation. With the help of our discrete models, we hope to shed light on this phenomena.

In experiments by Liu et al, polyhedral supramolecular cages made from different molecular species were synthesized. Theoretically, tiling type behavior of the molecular species is possible, but it was not observed. Additional experiments in which copies of two types of molecular species could be combined to form two different polyhedral cages were proposed. It is unknown if both potential polyhedral cages would form or if one would be significantly more favorable over the other. Being able to predict the results of such experiments via mathematical analysis and simulation would be a significant contribution toward optimizing strategies for the self-assembly of supramolecular cages.

1.1.3 Viral Capsid Assembly

With forces such as friction playing a much bigger role on their length scale, biological viruses are remarkable at robustly replicating themselves. While this topic has been well studied and a wide variety of mechanisms have been evidenced to contribute, there still is no general understanding of the process in which a virus' capsid is formed. Viral capsids come in many shapes and sizes, but a significant portion of them are icosahedral in structure. Understanding of their pathways of formation may be a key to preventing or slowing down replication.

1.1.4 RNA and Protein Folding

Protein and RNA folding are active fields of research. If we can predict the structure of a RNA or protein based on its amino acid sequence, we will know more about its biological function. As many human and animal disorders are caused by proteins folding abnormally, their cures may lie in the understanding of the folding pathway. While we do not directly consider applications relating to RNA and protein folding, success with the above problems may also provide insights for this complex topic.

Just as in self-assembly, folding constitutes the passage between several intermediate states along a pathway leading to formation of a final structure. In both examples, formation pathways are thought to be nearly unique. The identification of these pathways may allow for the targeted intervention at a specific intermediate in the case of abnormal folding behavior.

1.2 Mathematical Work Inspired by Self-Assembly Processes

1.2.1 The Building Game

First proposed by Zlotnick et al, the Building Game (BG) is a discrete attachment model that simulates the sequential construction of polyhedral structures. The BG can describe the behavior of many two dimensional face molecules interacting in a solution and bonding together to ultimately form polyhedral molecule.

The building game (BG) for a polyhedron $\mathcal{P}$ begins with a single face of $\mathcal{P}$ and iteratively attaches faces to the existing partially formed polyhedron until all faces of $\mathcal{P}$ are present. We denote the set of $\mathcal{P}$'s faces, edges, and vertices as $F(\mathcal{P})$, $E(\mathcal{P})$, and $V(\mathcal{P})$ respectively. A building game **pathway** is a linear ordering $f_1, f_2, f_3, \ldots, f_N$ of the faces of $\mathcal{P}$ such that for $j = 2, \ldots, N$ there exists edges $e_1, e_2, \ldots, e_k \in E(\mathcal{P})$ with $k \geq 1$ satisfying

$$(e_1 \cup \cdots \cup e_k) \subset \left(f_j \cap \left(\bigcup_{i=1}^{j-1} f_i \right) \right)$$

Since the order and location of attachment in the building game can vary, many partially formed polyhedra, called **intermediates**, are possible. Each intermediate x can be represented as $x = \cup_{i=1}^{t} f_i$ where $f_1, \ldots, f_t, \ldots, f_N$ is a BG pathway. For a given polyhedron, we are interested in enumerating all of the distinct intermediates up to rotational equivalence. The **attachment sites** of an intermediate x are the set of faces $\{f_k\}$ such that $f_k \cap x = e_1 \cup e_2 \cup \cdots$ for some edges $e_1, e_2, \ldots \in E(\mathcal{P})$. In other words, the attachment sites are the places in which a new face may join x as part of a valid BG pathway.

Figure 1.1: Dodecahedron building game example.

The **state space** for a particular polyhedron $\mathcal{P}$ is a graph that represents the space of all distinct intermediates and the BG pathways for $\mathcal{P}$. Each node of the state space represents a single intermediate and a connection exists between two nodes if it is possible to construct one of the corresponding intermediates by adding a single face to the other. Each path through the state space, starting at an intermediate with one face and ending at the intermediate with all faces, represents one of the polyhedron's BG pathways.

Figure 1.2: The state space of the cube.

For a state space edge going between an intermediate with k faces to one with $k+1$ faces, the **degeneracy number** is the number of different attachment sites on the k-faced intermediate that will produce the $k+1$-faced intermediate. For example, the state space edge between the cube intermediate with 1 face and the intermediate with 2 faces has degeneracy number 4 since each of the first square's four edges will form the same intermediate when a second square is attached.

As we consider polyhedra with more and more faces, there is a combinatorial explosion in the number intermediates in state space. While the 6-faced cube state space has only 8 vertices and 9 edges, the 20-faced icosahedron state space has 2,649 vertices and 17,241 edges and the 26-faced truncated cuboctahedron state space has 1,525,605 vertices and 17,672,377. We have computed the BG state space for all polyhedra in the Platonic, Archimedean, and Catalan solid classes of up to 30 faces. Due to computational constraints, we do not compute the BG state space for larger polyhedra, but we are exploring the possibility of non-enumerative exploration of the state space of large polyhedra in the future.

The Building Game suffers from the fact that it only allows one end product to form and cannot model physically realistic formation errors. However its simplicity is a feature that aids analysis. We discuss models that do capture formation errors below.

1.2.2 Folding

Another model that we refer to as the Folding Model, was used by Pandey et al to model the folding style self-assembly of meso-scale metallic polyhedra. In this model, a polyhedron's net is sequentially folded up into the finished polyhedron. At each stage one vertex is closed to form a pyramidal structure by a folding move. In analogy with the building game, we can define folding intermediates to be partially folded states and a state spaces where intermediates are connected if one can form the other with a single fold. The state space and corresponding intermediates are shown in figure **??**.

Interestingly, each polyhedron may has multiple nets. In fact, the number of distinct nets grows rapidly with the size of the polyhedron. Certain nets have different folding pathways and measuring the favorability of one of these nets over another introduces an design problem. If nets that fold into the completed polyhedron most reliably can be identified, the efficiency of the self-assembly process can be optimized.

As a feature, this model has been shown to allow folds that do not result in the initially intended polyhedron and can terminate in other polyhedra and blocked states. In figure **??**, the red connections represent folds that can still result in the octahedron while the green connections cannot. Many of the intermediates that cannot fold into the octahedron end up folding into a non-convex boat intermediate

(84).

Like in the building game, the state space of the folding model experiences a combinatorial explosion as the number of faces the polyhedron has increases. Surprisingly, the building game can be used to recover the part of the folding state space which allows folding to the originally intended polyhedron.

Theorem 1. *Given a polyhedron $\mathcal{P}$, there exists a second polyhedron $\mathcal{P}'$ such that the sub-graph of folding state space containing intermediates that can still fold into $\mathcal{P}$ can be recovered from the building game state space for $\mathcal{P}'$.*

1.2.3 Local Rules

A large number of previously studied models for self-assembly can be classifies as *local rules* based approaches. Such models typically consist of a collection of components that can be combined according to a specified grammar. While a basic idea, the variety off possible component and grammars makes this a widely flexible class of models. Schwartz et al used such a model to describe the assembly of viral capsids. Since these capsids are fundamentally composed of proteins that can each assume a number of conformations. By putting a grammar on the ways in which proteins of different conformation can combine, they we able to successfully form a variety of viral capsids in simulation. While this class of models is powerful, we have not actively pursued any such models. In the future, it may be work trying to use a local rules model that is akin to the building game, but will be sufficiently general to allow formation errors as in the folding model.

1.2.4 Dominant Formation Pathways

The concept of formation pathways that occur with overwhelming frequency relative to the myriad of other theoretically possible pathways is a primary focus of our work. To identify the nature of such dominant pathways is to identify the mechanism of self-assembly itself. This knowledge would hopefully enable the formation of the self-assembled products to become more efficient and expedient, or in cases such as biological viruses, inhibit it altogether. There are several means, both quantitative and qualitative, by which we seek to identify and study these dominant pathways.

1.2.5 Markov Processes

We first define a Markov Chain on the Building Game state space and identify the corresponding stationary distribution.

Associated with each intermediate x_j, we have the the combinatorial statistic $E_j \doteq -E(x_j)$ which is negative one times the number of edges in x_j at which two faces meet. Thus, $E_k - E_j$ represents the number of such edges added (or removed) when a face is added to (or removed from) x_j to form x_k. This statistic E_j is an idealization of a bond energy within an intermediate.

We define the Markov chain X_t by the transition rule $P_{jk} \doteq P(X_{t+1} = x_k | X_t = x_j)$, with the heuristic that it should be more likely for an intermediate to add a face in a location in which more new edge connections can be formed.

$$P_{jk} = \begin{cases} \frac{1}{z_j} S_{j,k} e^{-\frac{1}{2}\beta(E_k - E_j)} & j \neq k, j \leftrightarrow k \\ \frac{1}{z_j}\phi_j & j = k \\ 0 & j \neq k, j \not\leftrightarrow k \end{cases}$$

Here the self-transition likelihoods $\{\phi_j\}$ and the thermodynamic β, are parameters that can be chosen later. We define the normalization constants as $z_j \doteq \phi_j + \sum_{k \neq j} S_{jk} e^{-\beta(E_k - E_j)} \mathbb{1}_{j \leftrightarrow k}$.

Since we define Building Game intermediates as rotationally unique from each other, it is useful to think about the rotational symmetry group of each intermediate. For an intermediate x_j, we define r_j to be the order of the rotation group of x_j. By polyhedral group theory arguments, we have proven the following theorem.

Theorem 2. *For two Building Game intermediates x_j and x_k connected in the BG state space, $r_j = S_{jk} C_{jk}$.*

While we indeed have a precise geometric definition of the values C_{jk}, it is most important to not that $C_{jk} = C_{kj}$ is symmetric. This symmetry is useful since the degeneracy number S_{jk} is not itself symmetric in general. By assuming detailed balance, theorem 2 allows us to derive a stationary distribution for X_t.

Theorem 3. *The Markov chain X_t defined by the transition rule P_{jk} admits the unique stationary distribution $\pi_j = \frac{1}{z}\left(\frac{z_j}{r_j}\right) e^{-\beta E_j}$.*

One way to define the concept of a dominant intermediate or pathway is with respect to this stationary distribution. However, dynamics should not be overlooked. One natural question to ask is: which pathway has the highest probability of occurring with respect to our transition probabilities? More formally, we can put a

probability on a pathway $x_0 = x_{n_1}, x_{n_2}, x_{n_3}, ..., x_{n_F} = x_F$ as the following product.

$$P\left(x_{n_1}, x_{n_2}, x_{n_3}, ..., x_{n_F}\right) \doteq \prod_{k=1}^{F} P_{n_{k-1}n_k} \propto \prod_{k=1}^{F} S_{n_{k-1}n_k}$$

Of course, this definition does not allow for backward transitions which can play an important role in the correction of less favorable intermediates, but it provides a probabilistically motivated heuristic.

Since we are interested in the formation of the completed polyhedron beginning from a single face, we can ask about how likely a specific intermediate is to be reached if the state transitions until it reaches the terminal closed polyhedron. If we define the stopping time $\tau_k \doteq \inf\left\{n > 0 : X_n = x_k\right\}$ with τ_F the equivalent stopping time corresponding to the terminal state, we can formulate this question as computing the probability $\rho_k\left(\beta\right) \doteq P\left(\tau_k < \tau_F\right)$. These statistics can be computed exactly via dynamic programming, but it is possible that martingales or other analytic tools would allow a more direct and illuminating derivation.

Markov processes can be used in a similar way for a variety of different discrete models for self-assembly. The folding model tends to describe an irreversible process, so some of the concepts inherent to Markov processes, such as the stationary distribution would not apply. The dynamical aspects, however, may still be appropriately modeled with Markov processes. While we have not explored the topic extensively, it seems as though, due to their modeling and analytic flexibility, Markov processes provide a natural approach to many local rules type models as well.

1.2.6 Disconnectivity Graphs

Due to the scale of many of these self-assembly processes, statistical mechanics provides a framework to think about these models. By finding an analogy with a diffusion process on an implied potential energy surface, we can use statistical mechanical tools to examine the properties of this potential surface. Wales et al introduced a graphical way to represent the structure of a potential surface possessing many local minima and intermediary transition states. Named a *discontinuity graph* (DG), the tree with leaves representing local minima of the potential and other connecting nodes representing transition states that the minima can be reached from is represented in the plane. The vertical height of each node is used to represent the potential energy of that state. Horizontally, the tree is organized to partition these local minima into corresponding funnels. If the graph is truncated at a specific energy level, two minima are reachable via transitions to states strictly below this truncation energy if and only if they remain connected in the truncated disconnectivity tree.

Figure 1.3: A disconnectivity tree for the 38-atom Lennard-Jones cluster. [from Wales et al]

Depicted in figure 1.3 is the disconnectivity graph for a 38-atom Lennard-Jones cluster. This DG is described as having two funnels since there are two low energy local minima (one of which is the global minimum) that are far from each other in the graph and yet not of significantly different energy levels. This suggests that the cluster could be energetically trapped in the non-global minima and it would take a prohibitive amount of time to escape to the global minimum.

While not a quantitative way of assessing a model's pathways, disconnectivity graphs are extremely useful in identifying broad qualitative properties and can be instrumental in gaining insight into a model's dynamics. The use of such graphical and otherwise qualitative techniques should not be overlooked as they can inspire

techniques for more quantitative analysis.

1.3 Prior Work

While models such as the Building Game treat assembly intermediates as idealized structures, the intermediates of the physical application we are trying to model may face a chaotic and volatile range of forces. To more realistically model the ways in which our intermediate might flex and move under these forces, we impose a constraint model in with the rigidity of individual components of an intermediate are assumed, but in which the edges at which they meet are treated like a hinge. This constraint system specifies the ways in which an intermediate has freedom to move if it is able to move at all. A physically motivated way of addressing questions that the discrete models themselves are not adequate to answer is provided by this framework.

1.3.1 Configuration Space

To characterize the freedoms of three-dimensional intermediates composed of rigid two-dimensional faces and connected at hinged edges, we parameterize each face and impose constraints in parameter space. Theoretically, the location and orientation of each face of an intermediate can be parameterized by six parameters: three for translation, three for rotation. This means the configuration of the entire intermediate can be represented by at most $6|F|$ parameters. In practice, however, it is often advantageous for ease of analysis and computational implementation to use more parameters to represent a given configuration. Often times, we use 3 parameters to

represent the location each vertex of each face, even if some of these vertices share common locations. This means that we often have an ambient parameter space of $\mathbb{R}^N$ with $N > 6|F|$.

Upon this parameter space we place three types of constraints. The first removes the configuration's 6 trivial degrees of freedom due to translation and rotation. Since the intermediate is connected, this can be achieved by fixing the parameters of one of the intermediate's faces. Additionally, we enforce a rigidity constraint on each face ensuring that the structure of a constituent face will not change with movement of the larger structure. The final type of constraint ensures that the connections between two faces have hinge-like mobility. Since a shared edge has two vertices belonging to each face, this constraint is imposed by identifying the corresponding vertex locations.

The *configuration space* is defined to be the subset of ambient space $\left\{z \in \mathbb{R}^N : \varphi(z) = 0\right\}$ for which all M of the constraint equations $\varphi : \mathbb{R}^N \to \mathbb{R}^M$ are satisfied. Since it is assumed that the standard configuration of the intermediate given by the model satisfies the constraints, the configuration space is non-empty. The configuration space is an algebraic variety, because the constraint equations can typically be represented as (quadratic) polynomials. The *degrees of freedom* of a particular configuration is taken to be the dimension of the null space of the Jacobian of φ. This is due to the fact that any move in the ambient space that prevents the constraint equations from changing must also be in the configuration space. Interestingly, it is possible for some members of configuration space to have a different number of degrees of freedom than others.

It is important to note that there are many possible parameterizations of a configuration and the choice of which will likely result in a different configuration space.

The selection of which face to constrain in order to remove the trivial degrees of freedom may also affect the structure of the configuration space. While we must be cognizant of these choices, in many cases, the properties we wish to evaluate of the configuration space are independent of the choices.

1.3.2 Constrained Dynamics

Sometimes we are interested in more that simply the number of degrees of freedom a particular configuration has. To explore the configuration space, find connections between different admissible configurations, analyze the configuration space's topology, or compute other relevant statistics, it can be useful to compute dynamics on the constraint space. There are clear computational challenges to computing such dynamics due to the fact that configuration space is often of much smaller dimension than the ambient space it sits in.

Constraint algorithms have been widely studied for use in molecular dynamic simulations: we use the SHAKE algorithm and application specific variants. The SHAKE algorithm is a two stage method in which the laws of motion are first used to evolve the system to an unconstrained configuration $\hat{y}_{n+1}$ based on the previous configuration y_n. In the second stage, Lagrange multipliers are used to correct $\hat{y}_{n+1}$ to a new configuration y_{n+1} which satisfies the constraints.

Assuming connectedness of the configuration space, constraint algorithms can be used to compute various statistics of the configuration space. For example, the d-dimensional volume of a d-dimensional configuration space can be used to measure how much mobility the intermediate has in its configuration space. Additionally, if we have soft constraints, such as preferred angles between faces, we can institute cost

functions on the configuration space with configurations having more favorable angles having a lower cost. Whether physically or artificially motivated, such cost functions can be treated as potential energy functions which can imply specific dynamics. Under this set of dynamics, quantities such as the average or minimum energies may be of interest.

We may also be interested in the behavior of a small part of the configuration as it undergoes constrained dynamics. For instance, suppose we are interested in the tendency for two edges on distinct faces to come together and form a new connection. By looking at the distance between these edges, we can measure how frequently this event happens and the probability that it will occur before another even of interest. This is similar to the exit time problem for diffusion and may provide a physically motivated evaluation tool for the appropriateness of our transition rules we adopt in our models.

1.4 Original Contributions

Chapter Two

The Building Game: Modeling

2.1 The Building Game as a Mathematical Framework for Self-assembly

The Building Game (BG) was first considered by Zlotnick [2] as a model for the assembly of polyhedral viral capsids. In the model, a capsid is idealized as a polyhedron $\mathscr{P}$ with each face treated as a subunit. Assembly proceeds from a singe face with a second face attached to the first along an edge. At each subsequent step of the process, an additional face is added along an edge of one of the already added faces. The process ends when all of $\mathscr{P}$'s face have been added resulting in a completed polyhedron.

A useful way to think about the Building game is as a sequential coloring process. Given a polyhedron with each face painted white, choose a face and paint it black. At each subsequent step, choose a white face that is adjacent to a black face and paint it black. Repeat until all faces are black. Here the black faces represent a face being present at a given step of the assembly process.

–Comparison with approaches in artificial life paper

2.2 Formal Definition

We formalize the Building Game in terms of the group action of the polyhedron's rotation group G acting on subsets of F, the polyhedron's face set. Using the polyhedron's dual graph representation $\mathfrak{G} = (F, E)$ whose nodes are the faces F of the polyhedron and connections E correspond to the pairs of faces sharing an edge we can define the mathematical structures that comprise different building game con-

figurations.

Definition 1. *A Building Game* ***state*** $x \subset F$ *is a non-empty subset of the faces* F *of a polyhedron such that the subgraph* $\mathfrak{G}|_x$ *restricted to the faces in* x *is a single connected component.*

It is often useful to depict building game states with Schlegal diagrams CITE which are two dimensional projections of the three dimensional polyhedron. Figure 2.1 XXX uses Schelegal diagrams to depicts all of the states of the tetrahedron. Since every face of the tetrahedron is adjacent to every other face, any non-empty subset of faces is a building game state. Thus, there are $\binom{4}{k}$ tetrahedron states that have k faces. As pictured in figure 2.2, some subsets of faces are not states. For

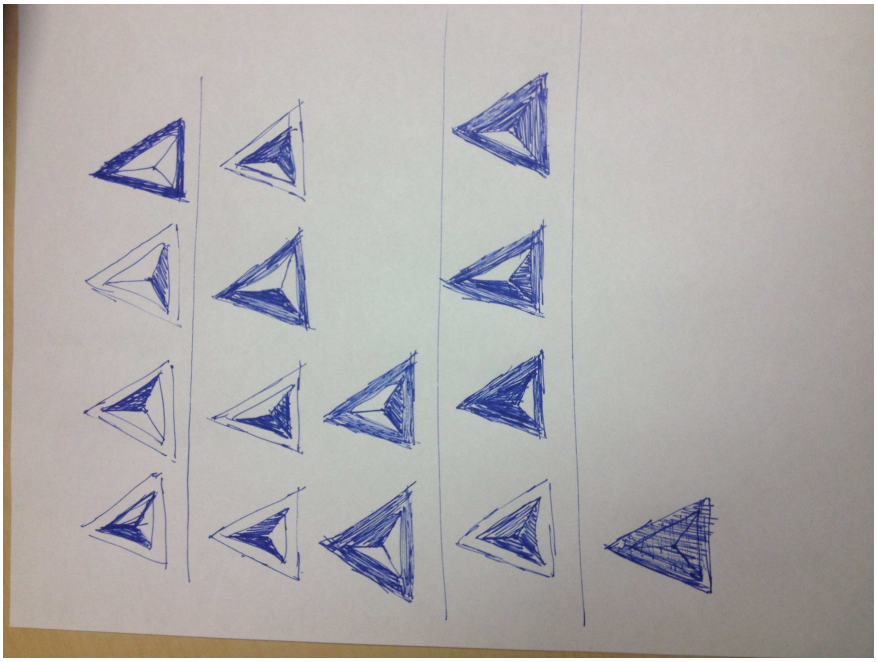

Figure 2.1: The building game states of the tetrahedron.

instance, the subset of octahedron faces with only two faces that are not adjacent is not a state. Similarly, the subset of two faces meeting only at a vertex is not a state as they are not connected through edge adjacency in the graph $\mathfrak{G}$. However, when more faces are added to connect these faces in $\mathfrak{G}$ the subset is indeed a state.

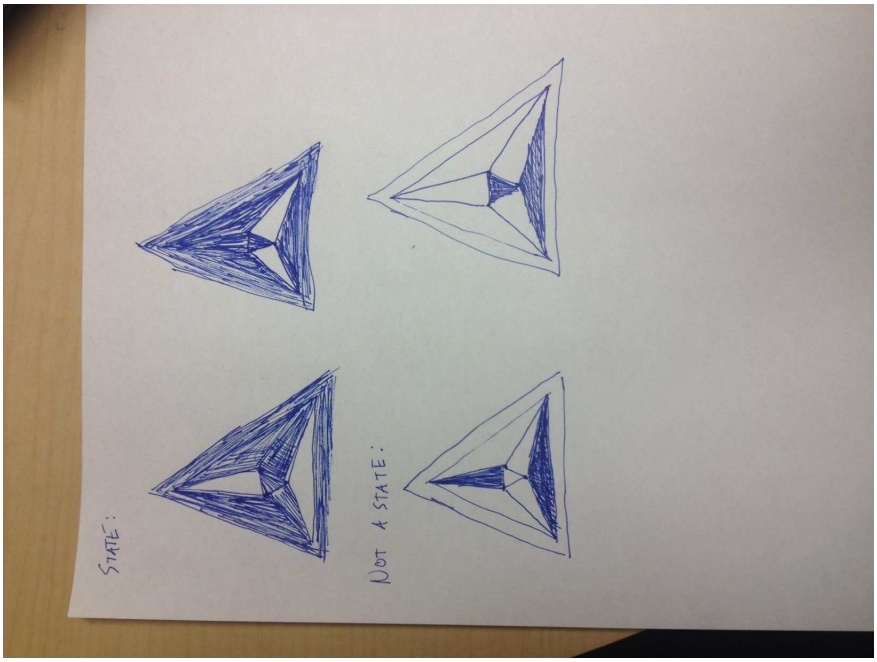

Figure 2.2: Examples of octahedron states and non-states.

2.2.1 Group Actions

It is easy to see that many states are combinatorially equivalent and are just rotations of each other. As in Zlotnick, we group the states into sets that are rotations of each other. However, to do so, we must first build the mathematical infrastructure using group actions.

Definition 2. *A* ***group action*** *of a group G on a set X is a function mapping $G \times X$ to X with $(g, x) \mapsto g.x$ that satisfies: (i) $(gh).x = g.(h.x)$ for $g, h \in G$ and (ii) $e.x = x$ for e the identity element of G.*

We typically use a polyhedron's rotation group or one of its subgroups as G and 2^F, the set of subsets of F, as the set X that G acts on. In this case, the action $g.x$ permutes the faces of the polyhedron according to the element g of the rotation group and results in a new subset of faces. Here, we introduce the group action concepts of orbits and stabilizer subgroups as they play an important role in out

later analyses.

Definition 3. *Let G be a group acting on a set X, the **orbit** of an element $x \in X$ is the subset $G.x \doteq \{g.x : g \in G\}$ of X [1].*

We also use the shorthand notation $[x]$ to refer to the orbit $G.x$ since an orbit can be thought of as an equivalence class under the relation $x \sim \hat{x}$ if there is a $g \in G$ such that $x = g.\hat{x}$.

Definition 4. *For a group G acting on a set X, the **stabilizer subgroup** for an element $x \in X$ is the subgroup $G_x \doteq \{g \in G : g.x = x\}$ of G that fixes x [1].*

Now, we introduce three classical results of group actions that relate orbits and stabilizer subgroups and add a corollary that we will use frequently.

Theorem 4 (Orbit-Stabilizer [1])**.** *Let G be a group acting on a set X, then for any $x \in X, |G.x| = [G : G_x]$*

Theorem 5 (Lagrange [1])**.** *If G is a finite group and $S \leq G$, then $|S|$ divides $|G|$ and $[G : S] = |G|/|S|$*

Lemma 1 (Burnside [1])**.** *Let G be a finite group acting on a set X, then*

$$|X/G| = \frac{1}{|G|} \sum_{g \in G} |X^g|$$

where $|X/G|$ is the number of orbits and $|X^g| = x \in X : g.x = x$

Corollary 1. *Let G be a finite group acting on a set X, then for any $x \in X$*

$$|G.x| = \frac{|G|}{|G_x|}.$$

Proof. This result follows trivially from Lagrange's theorem and the Orbit-Stabilizer theorems. □

2.2.2 Building Game Intermediates

With this framework of group actions, we now formally define the principle unit of the building game.

Definition 5. *A Building Game* ***intermediate*** $[x] \doteq \{g.x : g \in G\}$ *is the orbit of the state* x*.*

Since the orbits of a group action form a partition on the set it acts on, each state belongs to a single intermediate and two states x and y are part of the same intermediate if there is a $g \in G$ such that $y = g.x$. In the case of the tetrahedron, as pictured in figure 2.1, there are only four intermediates since any state with the same number of faces can be rotated to reach the others.

Since we have defined the intermediates to be the orbits of states under the polyhedral group, we are naturally also interested in stabilizer subgroups of building games states.

Definition 6. *The* ***symmetry number*** r_x *of a state* x *is the order of its stabilizer subgroup* $|G_x|$*.*

In figure **??** we see

Theorem 6. *If the states* x *and* $\hat{x}$ *are members of the same intermediate* $[x]$*, they have the same symmetry number. Thus, we extend the notion of a symmetry number to be a property of an intermediate.*

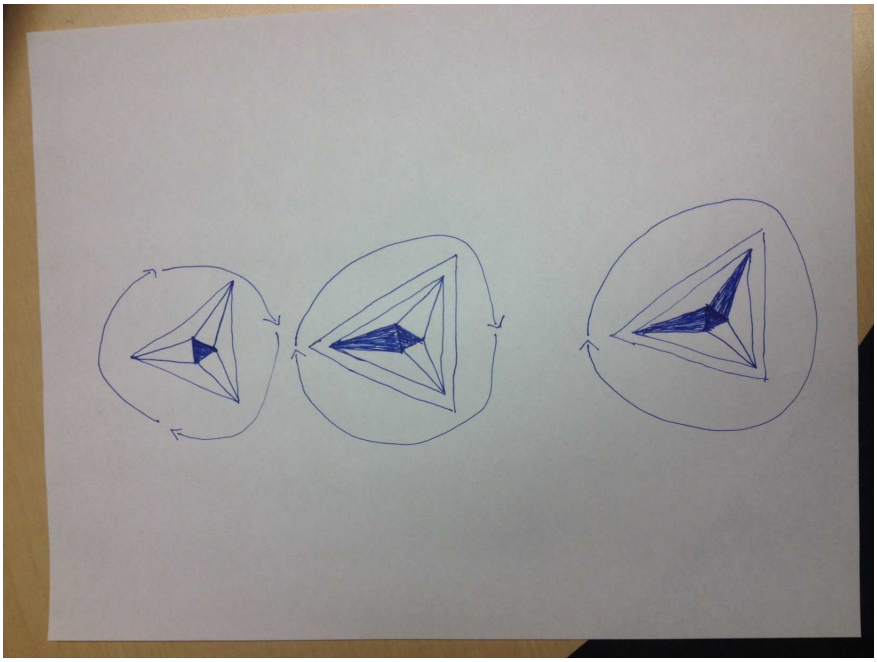

Figure 2.3: The stabilizer subgroups for various octahedron states.

Proof. By Corollary 1 and since $G.x \doteq [x] = [\hat{x}] \doteq G.\hat{x}$, the result follows.

$$\begin{aligned} r_x &= |G_x| && (2.1)\\ &= \frac{|G|}{|G.x|} && (2.2)\\ &= \frac{|G|}{|G.\hat{x}|} && (2.3)\\ &= |G_{\hat{x}}| && (2.4)\\ &= r_{\hat{x}} && (2.5) \end{aligned}$$

□

Since the building game is at its core an attachment model, we are interested in which intermediates can be formed from others by attaching a face to a particular intermediate.

Definition 7. *Two distinct intermediates* $[x]$ *and* $[y]$ *are* ***connected*** *if there exist*

states $x \in [x]$ and $y \in [y]$ such that one of the following holds:

- $\exists f \in y : y = x \cup \{f\}$
- $\exists f \in x : x = y \cup \{f\}$

The act of attaching an additional face is referred to as a forward step in the building game and the removal of a face is a backward step. – Language about forward and backward connection directions?

Lemma 2. *If intermediates $[x]$ and $[y]$ are connected, then for every state $x \in [x]$ there is a state $y \in [y]$ such that $\exists f \in y : y = x \cup \{f\}$ or $\exists f \in x : x = y \cup \{f\}$.*

Proof. Without loss of generality, assume $|x| < |y|$. Since $[x]$ and $[y]$ are connected, let $\hat{x} \in [x]$, $\hat{y} \in [y]$, and $\hat{f} \in \hat{y}$, be such that $\hat{y} = \hat{x} \cup \{\hat{f}\}$. Then, for any $x \in [x]$, pick $g \in G$ such that $x = g.\hat{x}$. By choosing $y = g.\hat{y}$ and $\{f\} = g.\{\hat{f}\}$, we have

$$\begin{aligned} y &= g.\hat{y} && (2.6)\\ &= g.(\hat{x} \cup \{\hat{f}\}) && (2.7)\\ &= g.\hat{x} \cup g.\{\hat{f}\} && (2.8)\\ &= x \cup \{f\} && (2.9)\\ & && (2.10) \end{aligned}$$

and our result is shown. □

Definition 8. *A Building Game **pathway** is a sequence of intermediates $[x^{p_1}], [x^{p_2}], \ldots, [x^{p_N}]$ such that $[x^{p_i}]$ is connected to $[x^{p_{i+1}}]$, $|x^{p_i}| = i$, and $x^{p_N} = F$.*

Figure 2.4 shows a Building Game pathway for the dodecahedron using Schlegel

diagrams. The pathway has 12 intermediates since there must be exactly one intermediate x^{p_i} satisfying $h\left(x^{p_i}\right) = i$ for each $i = 1, 2, \ldots, 12$.

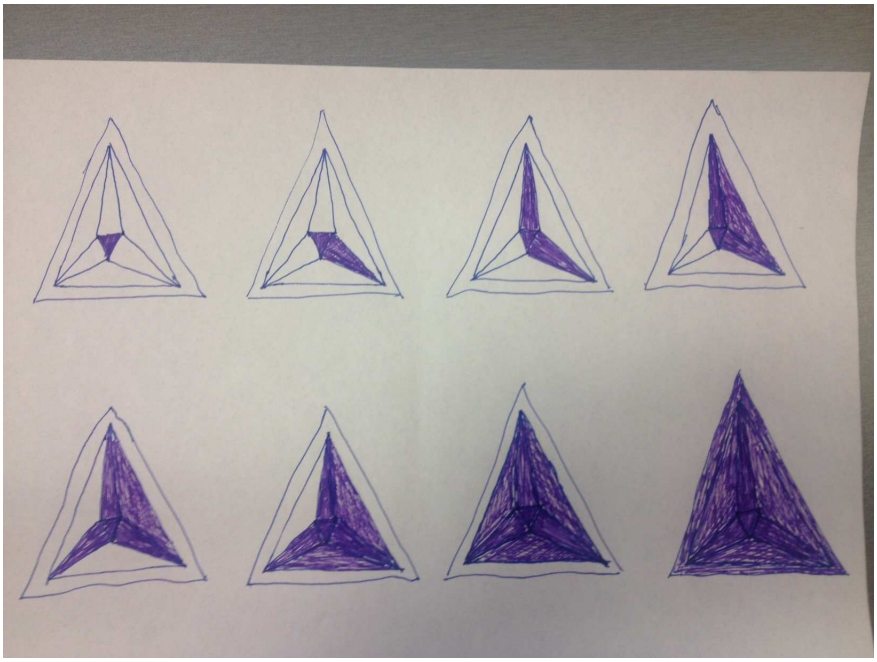

Figure 2.4: One Building Game pathway for the Octahedron.

With many pairs of connected intermediates, we organize these relations in a graph.

Definition 9. *The Building Game* ***combinatorial configuration space*** *for a polyhedron $\mathscr{P}$ is a graph in which the nodes are $\mathscr{P}$'s intermediates and a graph edge exists between two intermediates if and only if they are connected.*

When the intermediates are partitioned by the number of faces they possess, it is natural to arrange the state space as a tiered graph according to this partition. Figure 2.5 shows the Building Game state space for the cube. As seen, each tier has intermediates with the same number of faces and connections thus exist with intermediates that are either in the tier directly above or below them. We can also see that there are three distinct pathways contained in the state space.

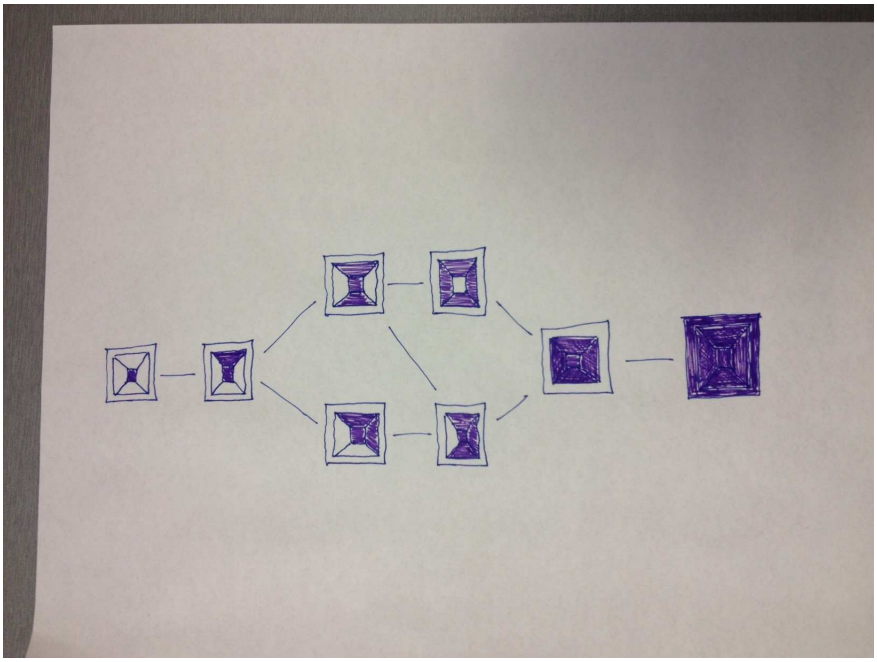

Figure 2.5: The Building Game combinatorial configuration space of the cube.

Interestingly, it is not the case that the addition or removal of each face of an intermediate results in a distinct intermediate.

Definition 10. *For two connected intermediates* $[x^j]$ *and* $[x^k]$*, the set of different faces*

$$F_{jk} \doteq \left\{ f \notin x^j : x^j \cup \{f\} \in [x^k] \right\} \cup \left\{ f \in x^j : x^j \setminus \{f\} \in [x^k] \right\}$$

that can be added or removed from x^j *to get an element of* $[x^k]$ *is called the* ***degeneracy set*** *and the number of such faces* $S_{jk} \doteq |F_{jk}|$ *is called the* ***degeneracy number****.*

It is important to note that in general the degeneracy number is not symmetric, i.e. $S_{jk} \neq S_{kj}$ for some connections $[x^j] \leftrightarrow [x^k]$ in the state space. Figure 2.6 depicts the forward and backward degeneracy numbers for a particular connection. As illustrated, there are two faces that can be added to the first intermediate to form the second. However, removing any of the three faces of the second intermediate will result in the first. Thus the forward degeneracy number is two and the backward

degeneracy number is three.

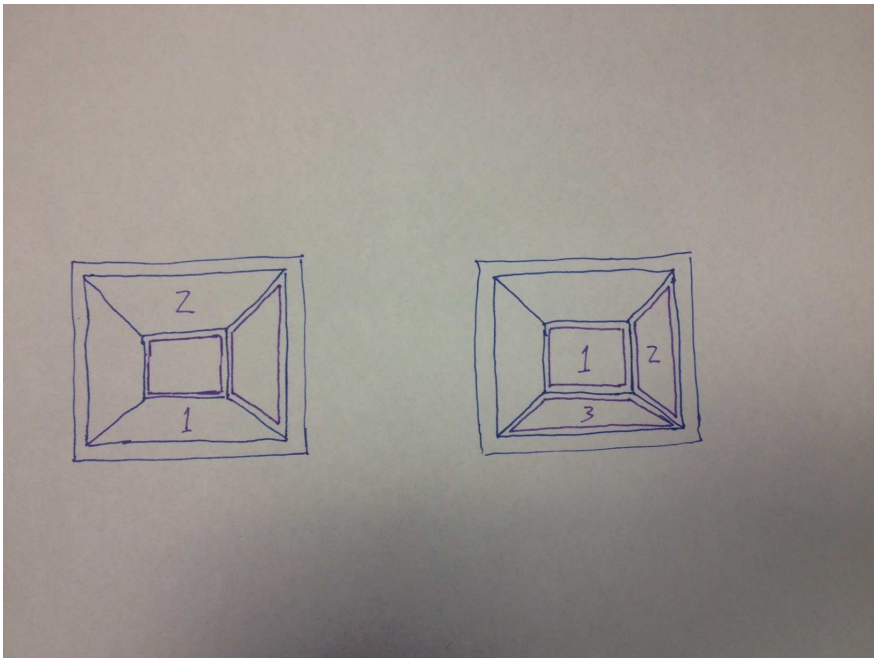

Figure 2.6: Degeneracies between two connected cube intermediates.

Finally, we extend the notion of a pathway to to be a sequence of intermediates of arbitrary lengths, such that the sequence begins at an intermediate with a single face and ends with the completed polyhedron and each pair of successive intrmediates are connected.

Definition 11. *A Building Game* ***reversible pathway*** *is a sequence of intermediates* $[x^{p_1}], [x^{p_2}], \ldots, [x^{p_N}]$ *such that* $[x^{p_i}]$ *is connected to* $[x^{p_{i+1}}]$, $|x^{p_0}| = 1$, *and* $|x^{p_N}| = |F|$.

2.2.3 Group Theoretic Results

Lemma 3. *Let* $g \in G$ *and* $x, y \subset F$. *Then* $g.(x \cup y) = g.x \cup g.y$

Proof. Let $f \in g.(x \cup y)$. Then $g^{-1}.\{f\} \in x \cup y$. Without loss of generality, assume

$g^{-1}.\{f\} \in x$. It follows that $f \in g.x$ and $f \in g.x \cup g.y$ as well. Therefore, $g.(x \cup y) \subset g.x \cup g.y$.

Conversely, suppose $\hat{f} \in g.x \cup g.y$ and without loss pick $\hat{f}$ to be in $g.x$. It follows that $g^{-1}.\{\hat{f}\} \in x, x \cup y$ and subsequently $\hat{f} \in g.(x \cup y)$. From this we have the reverse relation $g.x \cup g.y \subset g.(x \cup y)$ which proves our equality. □

Lemma 4. *Let the Building Game intermediates $[x^j]$ and $[x^k]$ be connected in the combinatorial configuration space. Given $x^j \in [x^j]$, $x^k \in [x^k]$ and $f \in F$ such that $x^k = x^j \cup \{f\}$, the stabilizer subgroups $G_{x^j,\{f\}} = G_{x^k,\{f\}} = G_{x^j,x^k}$ are all equal.*

Proof.

$$G_{x^k,\{f\}} = \{g \in G : g.x^k = x^k, g.\{f\} = \{f\}\} \tag{2.11}$$
$$= \{g \in G : g.(x^j \cup \{f\}) = x^j \cup \{f\}, g.\{f\} = \{f\}\} \tag{2.12}$$
$$= \{g \in G : g.x^j \cup g.\{f\} = x^j \cup \{f\}, g.\{f\} = \{f\}\} \tag{2.13}$$
$$= \{g \in G : g.x^j \cup \{f\} = x^j \cup \{f\}, g.\{f\} = \{f\}\} \tag{2.14}$$
$$= \{g \in G : g.x^j = x^j, g.\{f\} = \{f\}\} \tag{2.15}$$
$$= G_{x^j,\{f\}} \tag{2.16}$$
$$= \{g \in G : g.x^j = x^j, g.\{f\} = \{f\}\} \tag{2.17}$$
$$= \{g \in G : g.x^j = x^j, g.x^j \cup g.\{f\} = x^j \cup \{f\}\} \tag{2.18}$$
$$= \{g \in G : g.x^j = x^j, g.(x^j \cup \{f\}) = x^j \cup \{f\}\} \tag{2.19}$$
$$= \{g \in G : g.x^j = x^j, g.x^k = x^k\} \tag{2.20}$$
$$= G_{x^j,x^k} \tag{2.21}$$

□

Lemma 5. *For connected intermediates $[x]$ and $[y]$, with $x \cup \{f\} = y$ and $x \cup \{\hat{f}\} \doteq$*

$\hat{y} \in [y]$, the stabilizer subgroups $|G_{x,y}|$ and $|G_{x,\hat{y}}|$ satisfy $|G_{x,y}| = |G_{x,\hat{y}}|$.

Proof. If there is an $h \in G_x$ such that $\{f\} = h.\{\hat{f}\}$, the result follows similar to lemma XXX as $|G_{x,y}| = |hG_{x,\hat{y}}h^{-1}| = |G_{x,\hat{y}}|$.

However, if $\{f\} \neq h.\{\hat{f}\}$ for every $h \in G_x$, the result STILL NEEDS PROOF. In the cases we consider, the lemma has been confirmed to be true through brute force computation.

□

Lemma 6. *Let the Building Game intermediates $[x^j]$ and $[x^k]$ be connected in the combinatorial configuration space, the number of orbits $|F_{jk}/G_{x^j}| = |F_{kj}/G_{x^k}|$.*

Proof. Let $G_{x^j}.\{f\} \in F_{jk}/G_{x^j}$. Then, since $f \in F_{jk}$ we know that $x^j \cup \{f\} \in [x^k]$ and thus there is a $g \in G$ such that $g.(x^j \cup \{f\}) = x^k$. Since $g.x^j \in [x^j]$, we know that $g.\{f\} \in F_{kj}$. Then, for any $h \in G_{x^k}$ we see that $hg.(x^j \cup \{f\}) = h.x^k = x^k$. This means that $G_{x^k}.(g.\{f\}) \in F_{kj}/G_{x^k}$.

Now, we similarly assume $G_{x^k}.\{\hat{f}\} \in F_{kj}/G_{x^k}$. Then, $x^k \setminus \{\hat{f}\} \in [x^j]$ and there is a $\hat{g} \in G$ such that $\hat{g}.(x^k \setminus \{\hat{f}\}) = x^j$. Therefore, $\hat{g}.\{\hat{f}\} \in F_{jk}$ as $\hat{g}.x^k \in [x^k]$. So, for every $\hat{h} \in G_{x^j}$ we get $\hat{h}\hat{g}.(x^k \setminus \{\hat{f}\}) = \hat{h}.x^j = x^j$. Thus $G_{x^j}.(\hat{g}.\{\hat{f}\}) \in F_{jk}/G_{x^j}$.

Since we have shown that for every orbit in F_{jk}/G_{x^j} there is a corresponding orbit in $F_{kj}/G_{x^k}|$ and vice versa, the total number of orbits in each set must be the same and our result $|F_{jk}/G_{x^j}| = |F_{kj}/G_{x^k}|$ holds. □

Theorem 7. *For two Building Game intermediates $[x^j]$ and $[x^k]$ are connected in the combinatorial configuration space, $r_k S_{jk} = r_j S_{kj}$.*

Proof. Without loss of generality, assume that $[x^j]$ has one fewer face than $[x^k]$ and pick $x^j \in [x^j]$, $x^k \in [x^k]$ and $f \in F$ such that $x^k = x^j \cup \{f\}$. Then by Burnside lemma, we have the following [1].

$$|F_{jk}/G_{x^j}| = \frac{1}{|G_{x^j}|} \sum_{g \in G_{x^j}} |(F_{jk})^g| \tag{2.22}$$

$$= \frac{1}{|G_{x^j}|} \sum_{\hat{f} \in F_{jk}} |G_{x^j,\{\hat{f}\}}| \tag{2.23}$$

$$= \frac{|G_{x^j,\{f\}}|}{r_j} \sum_{\hat{f} \in F_{jk}} 1 \tag{2.24}$$

$$= \frac{|G_{x^j,\{f\}}| S_{jk}}{r_j} \tag{2.25}$$

Then, by lemmas XXX and YYY we have,

$$\frac{r_j}{S_{jk}} = \frac{|G_{x^j,\{f\}}|}{|F_{jk}/G_{x^j}|} \tag{2.26}$$

$$= \frac{|G_{x^k,\{f\}}|}{|F_{kj}/G_{x^k}|} \tag{2.27}$$

$$\doteq \frac{r_k}{S_{kj}} \tag{2.28}$$

and the result $r_k S_{jk} = r_j S_{kj}$ follows. □

2.3 Stochastic Modeling Results

Since the Building Game is a sequential process with several choices at each step, it is natural to consider it as a stochastic process. If we define a Building Game process that allows faces to be reversibly sequentially added or removed, the process consists of transitions from intermediate to intermediate along connections in the combinatorial configuration space. By specifying a distribution on these transitions,

it will induce a stationary distribution on the state space and provide a framework for compute relevant statistics such as expected formation times.

We define the Markov process X_t by the transition rate matrix Q, with the heuristic that the rate of transition to an intermediate $[x^k]$ from an intermediate $[x^j]$ should be proportional to the number of faces that can be added or removed from $[x^j]$ to reach $[x^k]$. For this reason, we include the degeneracy number S_{jk} as a factor in the transition rate matrix. Furthermore, we model the process using an energetic interpretation in which each intermediate has an energy and to transition between intermediates, an energy barrier $E_{jk} = E_{kj}$ must be overcome.

$$Q_{jk} = \begin{cases} S_{jk} e^{-\beta\left(E_{jk} - E_j\right)} & \text{if } [x^j] \leftrightarrow [x^k] \\ -z_j & \text{if } j = k \\ 0 & \text{else} \end{cases} \tag{2.29}$$

Here, $z_j \doteq \sum_{\ell:\ell\neq j} S_{j\ell} e^{-\beta\left(E_{j\ell} - E_j\right)}$ is the rate at which the process leaves x^j.

2.3.1 Stationary Distribution

Theorem 8. *If the transition rate matrix Q can be decomposed as $Q = DC$ where D is diagonal with each entry of the diagonal positive and C is a symmetric matrix with the non-diagonal entries $C_{jk} > 0$ if and only if $[x^j]$ and $[x^k]$ are connected and $C_{jk} = 0$ if they are not, then X_t has the unique stationary distribution $\pi = \mathrm{diag}\,(D^{-1})$.*

Proof. First, we show Q and π satisfy detailed balance.

$$\pi_j Q_{jk} = \left(\frac{1}{D_{jj}}\right)(D_{jj}C_{jk}) \tag{2.30}$$

$$= C_{jk} \tag{2.31}$$

$$= C_{kj} \tag{2.32}$$

$$= \left(\frac{1}{D_{kk}}\right)(D_{kk}C_{kj}) \tag{2.33}$$

$$= \pi_k Q_{kj} \tag{2.34}$$

Since the combinatorial state space is connected and $Q_{jk}, Q_{kj} > 0$ for all connected intermediates $[x^j]$ and $[x^k]$, the process it trivially aperiodic and positive recurrent. □

Theorem 9. *The Markov process X_t defined by the transition rate matrix Q in equation 2.29 admits the unique stationary distribution $\frac{1}{zr_j}e^{-\beta E_j}$ where $z \doteq \sum_\ell \frac{1}{r_\ell}e^{-\beta E_\ell}$ is the partition function.*

Proof. We take $C_{jk} \doteq \frac{S_{jk}}{zr_j}e^{-\beta E_{jk}}$ and notice that it is symmetric by theorem 7. With $D_{jj} \doteq zr_j e^{\beta E_j}$ we have our partition.

$$Q_{jk} = S_{jk}e^{-\beta\left(E_{jk}-E_j\right)} \tag{2.35}$$

$$= \left(zr_j e^{\beta E_j}\right)\left(\frac{S_{jk}}{zr_j}e^{-\beta E_{jk}}\right) \tag{2.36}$$

$$= D_{jj}C_{jk} \tag{2.37}$$

Thus, by theorem 8, $\pi_j = \frac{1}{D_{jj}} = \frac{1}{zr_j}e^{-\beta E_j}$. □

2.3.2 Hitting Times

While the stationary distribution is an important piece in understanding the nature of our Markov process, other statistics which describe the dynamics are also useful. For instance, we may be interested in the expected time it will take the process to travel from an intermediate $[x]$ to a specific subset A of the combinatorial configuration space. Equation 2.38 provides a mathematical definition for such a stopping time.

$$\tau_j^A \doteq \inf\left\{t \geq 0 : X_t \in A, X_0 = x^j\right\} \tag{2.38}$$

Oftentimes, we choose A to be the intermediate of the fully completed polyhedron. This means the the stopping time will represent the expected formation time from a given intermediate.

Since we are looking at events in which the process first reaches a particular set of intermediates, it is helpful to use X_t's discrete time partner process Y_n which simply records the sequence of intermediates that X_t passes through. Here we formally define Y_n in a recursive manner, such that it is the first intermediate visited after the process X_t leaves Y_{n-1}.

$$Y_n \doteq \arg\min_{y \neq X_t}\{s : s > t, X_s = y, X_t = Y_{n-1}\} \tag{2.39}$$

$$\sim X_{\tau_{Y_{n-1}}^{\{Y_{n-1}\}^C}} \tag{2.40}$$

Using first step analysis, we derive the expected hitting time τ_j^A. Consider the

case of $j \notin A$.

$$E\left[\tau_j^A\right] = E\left[E\left[\tau_j^A | Y_1\right]\right] \tag{2.41}$$

$$= E\left[Exp\left(z_j\right) + \tau_{Y_1}^A\right] \tag{2.42}$$

$$= \frac{1}{z_j} + E\left[\sum_k \tau_{Y_1}^A \mathbb{1}_{Y_1=k}\right] \tag{2.43}$$

$$= \frac{1}{z_j} + \sum_{k:k\neq j} E\left[\tau_k^A\right] P\left(Y_1 = k\right) \tag{2.44}$$

$$= \frac{1}{z_j}\left(1 + \sum_{k:k\neq j} Q_{jk} E\left[\tau_k^A\right]\right) \tag{2.45}$$

$$-\sum_k Q_{jk} E\left[\tau_k^A\right] = 1 \tag{2.46}$$

$$\tag{2.47}$$

Now, clearly for $j \in A$ we trivially have a stopping time of zero:

$$E\left[\tau_j^A\right] = 0. \tag{2.48}$$

$$\tag{2.49}$$

Putting together the two case, we write the solution as a linear system with τ^A the vector of stopping times with each possible initial intermediate.

$$\left(\text{diag}\left(\mathbb{1}_A\right) - \text{diag}\left(\mathbb{1}_{A^c}\right) Q\right) E\left[\tau^A\right] = \mathbb{1}_{A^c} \tag{2.50}$$

$$\tag{2.51}$$

Thus, we have

$$E\left[\tau^A\right] = \left[\left(\text{diag}\left(\mathbb{1}_A\right) - \text{diag}\left(\mathbb{1}_{A^c}\right) Q\right)\right]^{-1} \mathbb{1}_{A^c}. \tag{2.52}$$

$$\tag{2.53}$$

which contains the particular value $E[\tau_j^A]$ that we are interested in.

Further, we can also compute the exact distribution of the stopping times τ^A. First, we define the CDF as follows.

$$\psi_j^A(t) \doteq P\left(\tau_j^A \leq t\right) \tag{2.54}$$

$$\psi_j^A(0) = \mathbb{1}_{j \in A} \tag{2.55}$$

$$\psi_j^A(t) = 0 \forall j \in A \tag{2.56}$$

$$\tag{2.57}$$

Thus, for the case of $j \notin A$ we find

$$\psi_j^A(t) \doteq P\left(\tau_j^A \leq t\right) \tag{2.58}$$

$$= \sum_k P\left(\tau_j^A \leq t | Y_1 = x^k\right) P\left(Y_1 = x^k\right) \tag{2.59}$$

$$= \frac{1}{z_j} \sum_{k:k \neq j} Q_{jk} P\left(Exp(z_j) + \tau_k^A \leq t\right) \tag{2.60}$$

$$= \frac{1}{z_j} \sum_{k:k \neq j} Q_{jk} \int_0^t P\left(\tau_k^A \leq t - s\right) z_j e^{-z_j s} ds \tag{2.61}$$

$$= \sum_{k:k \neq j} Q_{jk} \int_0^t \psi_k^A(t-s) e^{-z_j s} ds \tag{2.62}$$

$$= \sum_{k:k \neq j} Q_{jk} \int_0^t \psi_k^A(r) e^{-z_j (t-r)} dr \tag{2.63}$$

$$e^{z_j t} \psi_j^A(t) = \sum_{k:k \neq j} Q_{jk} \int_0^t e^{z_j r} \psi_k^A(r) dr \tag{2.64}$$

$$e^{z_j t} \frac{d\psi_j^A}{dt} + z_j e^{z_j t} \psi_j^A(t) = \sum_{k:k \neq j} q_{jk} e^{z_j t} \psi_k^A(t) \tag{2.65}$$

$$\frac{d\psi_j^A}{dt} = \sum_k q_{jk} \psi_k^A(t). \tag{2.66}$$

Combining both cases, we get the linear system and solution.

$$\frac{d\psi^A}{dt} = \operatorname{diag}\left(\mathbb{1}_{A^c}\right) Q\psi^A \tag{2.67}$$

$$\psi^A\left(0\right) = \mathbb{1}_A \tag{2.68}$$

$$\psi^A\left(t\right) = e^{\operatorname{diag}(\mathbb{1}_{A^c})Qt}\mathbb{1}_A \tag{2.69}$$

$$\tag{2.70}$$

This is the solution for the CDF of the stopping time τ^A, but we can also compute the PDF explicitly for $t > 0$.

$$p\left(\tau^A = t\right) = \frac{d\psi^A}{dt} \tag{2.71}$$

$$= \operatorname{diag}\left(\mathbb{1}_{A^c}\right) Q\psi^A \tag{2.72}$$

Another hitting time statistic we may be interested in is, given an initial intermediate, what is the probability we will hit a subset of intermediates A before some other disjoint subset of intermediates B.

$$\rho_j^{A,B} \doteq P(\tau_j^A < \tau^B) \tag{2.73}$$

Trivially, we have

$$\rho_j^{A,B} = 1 \text{ if } j \in A \tag{2.74}$$

$$\rho_j^{A,B} = 0 \text{ if } j \in B \tag{2.75}$$

but must compute the case of $j \in (A \cup B)^C$.

$$\begin{aligned}
\rho_j^{A,B} &= P(\tau_j^A < \tau_j^B) && (2.76)\\
&= \sum_k P(\tau_j^A < \tau_j^B, Y_1 = k) && (2.77)\\
&= \sum_k P(\tau_j^A < \tau_j^B | Y_1 = k) P(Y_1 = k) && (2.78)\\
&= \sum_k P(\tau_k^A < \tau_k^B) P(Y_1 = k) && (2.79)\\
&= \frac{1}{z_j} \sum_{k \neq j} \rho_k^{A,B} Q_{jk} && (2.80)\\
0 &= Q\rho^{A,B} && (2.81)
\end{aligned}$$

Again, putting each of these cases together, we get the linear system

$$\left(\text{diag}\left(\mathbb{1}_A\right) + \text{diag}\left(\mathbb{1}_B\right) + \text{diag}\left(\mathbb{1}_{(A\cup B)^c}\right) Q\right) \rho^{A,B} = \mathbb{1}_A \tag{2.82}$$

and our solution is

$$\rho^{A,B} = \left[\text{diag}\left(\mathbb{1}_A\right) + \text{diag}\left(\mathbb{1}_B\right) + \text{diag}\left(\mathbb{1}_{(A\cup B)^c}\right) Q\right]^{-1} \mathbb{1}_A \tag{2.83}$$

Using this result, an insightful choice for A and B would be $A = \{x_k\}$ and $B = x_{|F|}$. Then, the value of $\rho_1^{A,B}$ would correspond to the probability that a

particular intermediate x^k is in a reversible pathway between the intermediate with a single face and the completed polyhedron.

Chapter Three

The Building Game: Enumeration

3.1 Known Enumerative Results

–like polyominos –Endres/Zlotnick –Integer database

3.2 New Enumerative Results

As we consider polyhedra with more and more faces, there is a combinatorial explosion in the number intermediates in state space. While the 6-faced cube state space has only 8 nodes and 9 nodes, the 20-faced icosahedron state space has 2,649 nodes and 17,241 nodes and the 26-faced truncated cuboctahedron state space has 1,525,605 nodes and 17,672,377. Figure **??** details state space sizes of all polyhedra in the Platonic, Archimedean, and Catalan solid classes of up to 26 faces.

Also something about pathway statistics.

Polyhedra Name	$\|F\|$	Intermediates	Connections	Pathways
Tetrahedron	4	5	4	1
Cube	6	9	10	3
Octahedron	8	15	22	14
Dodecahedron	12	74	264	17,696
Icosahedron	20	2,650	17,242	57,396,146,640

Figure 3.1: Building game enumerative results for the Platonic solids.

3.2.1 Shellability

–Def shellability

Polyhedra Name	$\|F\|$	Intermediates	Connections	Pathways
Truncated Tetrahedron	8	29	65	402
Cuboctahedron	14	341	1,636	10,170,968
Truncated Cube	14	500	2,731	101,443,338
Truncated Octahedron	14	556	3,071	68,106,377
Rhombicuboctahedron	26	638,851	6,459,804	16,494,392,631,838,879,380
Truncated Cuboctahedron	26	1,525,605	17,672,377	?
Icosidodecahedron	32	?	?	?
Truncated Dodecahedron	32	?	?	?
Truncated Icosahedron	32	?	?	?

Figure 3.2: Building game enumerative results for the Archimedean solids.

Polyhedra Name	$\|F\|$	Intermediates	Connections	Pathways
Triakis Tetrahedron	12	99	319	38,938
Rhombic Dodecahedron	12	128	494	76,936
Triakis Octahedron	24	12,749	81,297	169,402,670,046,670
Tetrakis Hexahedron	24	50,768	394,278	4,253,948,297,210,346
Deltoidal Icositetrahedron	24	209,676	1,989,549	?
Pentagonal Icositetrahedron	24	345,939	3,544,988	2,828,128,000,716,774,492
Rhombic Triacontahedron	30	?	?	5,266,831,101,345,821,968

Figure 3.3: Building game enumerative results for the Catalan solids.

3.2.2 Bounds and Asymptotics

There is a clear relation between the number of faces in a polyhedron and then number of intermediates it has. However, that relationship also greatly depends on the polyhedral symmetry group. For instance, if you have a polyhedron with a small number of faces and a trivial rotation group consisting only of the identity, every edge-connected subset of the polyhedron's faces will be a distinct intermediate. In agregate, this may mean that the polyhedron has more intermediates than another polyhedron with more faces, yet a larger symmetry group.

An upper bound on the number of intermediates is possible using the theory of group actions. Consider the set of all subsets 2^F of a polyhedron with rotation group G. Trivially $|2^F/G|$ is an upper bound on the number of intermediates since it simply

Polyhedra Name	$\lvert F\rvert$	Intermediates	Connections	Pathways
Tetrahedron	4	5	4	1
Cube	6	8	8	2
Octahedron	8	12	12	14
Dodecahedron	12	53	156	2166
Icosahedron	20	468	1984	105999738

Figure 3.4: Building game enumerative shellability results for the Platonic solids.

Polyhedra Name	$\lvert F\rvert$	Intermediates	Connections	Pathways
Truncated Tetrahedron	8	22	42	174
Cuboctahedron	14	137	470	477776
Truncated Cube	14	248	1002	5232294
Truncated Octahedron	14	343	1466	5704138
Rhombicuboctahedron	26	70836	462149	48399693494788840
Truncated Cuboctahedron	26	?	?	?
Icosidodecahedron	32	?	?	?
Truncated Dodecahedron	32	?	?	?
Truncated Icosahedron	32	?	?	?

Figure 3.5: Building game enumerative shellability results for the Archimedean solids.

relaxes the connectivity requirement for a subset to be a building game state. Using Burnside's lemma, we see that

$$|2^F/G| = \frac{1}{|G|}\sum_{g\in G}|(2^F)^g| \tag{3.1}$$

$$> \frac{|(2^F)^e|}{|G|} \tag{3.2}$$

$$= \frac{|2^F|}{|G|} \tag{3.3}$$

$$= \frac{2^{|F|}}{|G|} \tag{3.4}$$

which is not a particularly good bound in practice. The exact value of $|2^F/G|$ is calcuable with minimal computer assistance. For the cube, the bound is fairly tight, only including the two non-intermediates corresponding to the empty subset of faces, and the non-connected subset consisting of the top and bottom faces. Thus the cube has the bound $|2^F/G| = 10 \geq 8$. In the case of the tetrahedron, the only overcounted

Polyhedra Name	$\|F\|$	Intermediates	Connections	Pathways
Triakis Tetrahedron	12	49	116	5012
Rhombic Dodecahedron	12	68	196	6258
Triakis Octahedron	24	667	2383	15255459
Tetrakis Hexahedron	24	4220	21079	5854799360107
Deltoidal Icositetrahedron	24	?	?	?
Pentagonal Icositetrahedron	24	95127	654537	5607231936129109
Rhombic Triacontahedron	30	97368	697623	6889989896241902854

Figure 3.6: Building game enumerative shellability results for the Catalan solids.

subset of faces is the empty one and the bound is $|2^F/G| = 5 \geq 4$. However, in the case of the icosahedron we have $|2^F/G| \geq \frac{2^{20}}{60} \approx 17476.3 \gg 2649$. Here we use the approximate bound $\frac{2^{|F|}}{|G|}$ which is the largely dominant term in the sum from equation 3.1.

We can get a similar bound on the number of intermediates with a particular number of faces,

$$|\{x \in 2^F : |x| = k\}/G| = \frac{1}{|G|} \sum_{g \in G} |\{x \in 2^F : |x| = k\}^g| \tag{3.5}$$

$$> \frac{|\{x \in 2^F : |x| = k\}^e|}{|G|} \tag{3.6}$$

$$= \frac{|\{x \in 2^F : |x| = k\}|}{|G|} \tag{3.7}$$

$$= \frac{\binom{|F|}{k}}{|G|} \tag{3.8}$$

but again, this is not particularly useful, especially for intermediates with $\sim \frac{|F|}{2}$ faces.

Since the building game is similar in spirit to polyomino enumeration, one might try to assimilate some the techniques used for polyominos. For example, through fairly simple arguments, one can show that $s_m s_n \leq s_{m+n}$ where s_m is the number of unique polyominos with m subunits CITE. This leads to the bound $s_m \leq (const)^m$. Trying to set up such a relation in the building game is sounds initially appealing, but

there is a fundamental diefference between the two growth models that makes this approach futile. In the polyomino case, there is no limit to the number of subunits that can be considered. Importantly, this is not the case for the bulding game since an intermediate can only have $|F|$ faces at most. Thus any such reccurrence relation for the building game will result in a good upper bound for the intermediates with a small number of faces at best.

The formulation of meaningful bounds for the number of building game intermediates with k faces remains an open problem, especially for $k \sim \frac{1}{2}|F|$. At the root of the problem is the difficulty in mathematicallly describing the subsets of F are edge connected. Future approches may incorperate enumeration results for connected subgraphs or Hamiltonian paths since these topics explicitly acknowledge connectedness properties.

From looking at the statistics on number of faces $|F|$ of a polyhedron and the number of intermediates in its commbintorial configuration space, it is natural to want to make statements about the asymptotic growth of thr combinatorial configuration space's size. Unfortunately, when formed in this way, the problem is ill-posed. To discuss assymptotics, we must first specify an infinite class of polyhedra. The Platonic, Archimedean, and Catalan Solid classes that we've worked with thus far are all finite though, so other choices must be consider. One option is to take an existing polyhedron in one of these classes and create an infinite family by describing finer and finer tiling on top of the polyhedron's faces. If designed cafefully each member of the tiled polyhedron family will have the same symmetry group, even as the number of faces grows.

Similar to polyhedra with tiled faces are the icosahedron viral capsids indexed by T-number. This number is related to the number of protein subunits in the virust.

When each subunit is idealized as a polygon, a T-capsid will consist of 12 pentagons and $10(T-1)$ hexagons. Interestingly, this makes most of the icosahedral viral capsid equivalent to the dual of an icosahedron with each face consisting of T trianglular tiles. This would certainly be an interesting an relevant familty of polyhedra to consider, though we leave it as an opn problem.

3.3 Computational Methods

To compute the combinatorial configuration space and enumerate the intermediates for a particular polyhedron, we use a brute force method. Computation begins with first enumerating the intermediates with a single face. This enumeration is then used to compute the intermediates with two faces. This process proceeds iteratively until all intermediates are accounted for. Figure 3.7 outlines the detailed algorithm for this computation.

At each stage of our algorith, we know the set of intermedites that have k faces, which we call A_k, and use this information to compute the set of faces with $k+1$ faces, A_{k+1}. Since all intermediates in A_{k+1} must be formed by adding a single face to an intermediate from A_k, we take each intermediate $[x] \in A_k$ and try adding each face to x that is allowable under the building game rules. This means we look at every face $f \in F$ and check if $f \notin x$ and also that x is edge connected to a face $\hat{f}$ that is in x. For every such faces f, we look at the new $(k+1)$-faced state $y \doteq x \cup \{f\}$. Since we know that $[y]$ is a building game intermediate, it must be represented in A_{k+1}, however before adding y to A_{k+1}, we must verify that there is no $\hat{y}$ already in A_{k+1} such that $y \in [\hat{y}]$.

$A_0 \leftarrow \{\emptyset\}$
$A_1, \ldots, A_{|F|} \leftarrow \{\}$
for $k = 0, \ldots, |F| - 1$ **do**
  for $x \in A_k$ **do**
    for $f \in F \setminus x$ such that $\exists \hat{f} \in x$ with f and $\hat{f}$ sharing an edge. **do**
      NewIntermediate $\leftarrow True$
      for $\hat{y} \in A_{k+1}$ **do**
        if $y \in [\hat{y}]$ **then**
          NewIntermediate $\leftarrow False$
          Add connection $[x] \leftrightarrow [\hat{y}]$ to combinatorial configuration space.
        end if
      end for
      if NewIntermediate $= True$ **then**
        $A_{k+1} \leftarrow A_{k+1} \cup \{y\}$
        Add connection $[x] \leftrightarrow [y]$ to combinatorial configuration space.
      end if
    end for
  end for
end for

Figure 3.7: Algorithm for interatively enumerating the Building Game combinatorial configuration space.

The act of comparing two states y and $\hat{y}$ to check if they are members of the same intrmediate is the task where the majority of computational time is spent. The brute force method of checking if $y \sim \hat{y}$ involves checking if $g.y = \hat{y}$ for each $g \in G$. If a g is found that makes this equality hold, then $[y] = [\hat{y}]$ and the computation terminates and we know that $[y]$ is not a new intermediate. In this case, nothing is added to A_{k+1} but a connection $x \leftrightarrow \hat{y}$ is added in the combinatorial configuration space. Furthermore, by tracking how many time a particular connection $x \leftrightarrow \hat{y}$ is found, the forward degeneracy numbers can be computed.

3.3.1 Hash Functions

In practice, we implement a hash function $\mathfrak{h}$ that maps each state to an integer with the property that if $[y] = [\hat{y}]$, then $\mathfrak{h}(y) = \mathfrak{h}(\hat{y})$. If we can design such a function

and it is computable in significantly less time relative to the brute force method of trying every rotation, the overall computation can be majorly diminished.

When checking if two states y and $\hat{y}$ are members of the same intermediate, we first check if $\mathfrak{h}(y) = \mathfrak{h}(\hat{y})$. If they do not have the same hash, then they cannot be members of the same intermediate and the check is complete. Alternatively, if they do have identical hashes, it is not a guarentee that they are members of the same intermediate, so the brute force rotation method must be used. If the hash is carefully designed then this false positive rate ($\mathfrak{h}(y) = \mathfrak{h}(\hat{y})$ when $[y] \neq [\hat{y}]$) is small and almost all of the brute force calculations will result in a positive match.

In an ideal world, a hash function that also gives the property $[x] = [\hat{x}]$ whenever $\mathfrak{h}(x) = \mathfrak{h}(\hat{x})$ would be best. However we have not been able to find such an $\mathfrak{h}$ that is computable in a relatively reduced amount of time in comparison to the brute force method. We leave the existance of such a hash as a possibility.

Since any hash we choose must a function that maps states that are rotations of each other to the same integer, we look at local connectivity properties. For each face $f_k \in F$ define its local connectivity within y as $\mathfrak{h}_k(y) = |\{\hat{f} \in F : \hat{f} \in y, f_k \cap \hat{f} \in E\}|$, the number of faces adjacent to f_k that are in y. Since this connectivity is preserved by rotations, even though we wont necesarily have $\mathfrak{h}_k(y) = \mathfrak{h}_k(\hat{y})$ when $[y] = [\hat{y}]$, if we take a histogram of all values of $\mathfrak{h}_k$ for $k = 1, \ldots, |F|$, the histogram will be the same for both y and $\hat{y}$. Once a histogram is computed, the final hash $\mathfrak{h}$ is just a simple function that maps histograms to integers.

There are many variants of statistics that this histogram strategy can be used with. We found that taking separate histograms for faces in y and faces not in y provided a hash with less false positives, while only increasing the hash computation

time marginally. Depending on the polyhedron, the implementation of this hash lead to an over all speed up of at least an order of magnitude over the purely brute force method.

3.3.2 Data Structures

Before the computation of the combinatorial configuration space, we hardcode an enumeration $f_1, \ldots, f_{|F|}$ of the faces of our polyhedron. Then, any state x is represented by a binary vector of length $|F|$ with a one in the kth entry if $f_k \in x$ and zero otherwise. With this convention, each rotation $g \in G$ corresponds to a permutation of the indices of x. Each such permutation in the group is precomputed and then applies as necessary when performing a brute force comparison of two states. To track the connectivity structure of the polyhedron, an adjaceny list on faces is stored as a two dimensional array. For each face number, the adjacency list specifies the index of adjacent faces.

Each group of k-faced intermediates A_k is stored as a hash table using the previously described hash function. This allows for order one lookup of intermediates already in the A_k that share the hash of a new proposal intermediate.

Chapter Four

Constraint Models and Embedding Intermediates in Space

4.1 Linkages

4.2 Geometric Configuration Space

4.2.1 Computing the Degrees of Freedom of a 3D Linkage

We consider the linkage of rigid 2-dimensional polygons in $\mathbb{R}^3$ by means of ideal hinges located at the edges of the polygons. In this way, by treating intermediates of our various models as such a linkage, we can compute the non-trivial degrees of freedom (DoF) each intermediate has. Every linkage of this type has 6 trivial degrees of freedom: 3 corresponding to translational movement and 3 corresponding to rotational freedom. While these freedoms do change the orientation of the intermediate, they do not cause the faces to move relative to each other. For this reason, we label them as trivial degrees of freedom and focus on the calculation the remaining, non-trivial degrees of freedom. Since the concept of degrees of freedom exists in many diverse scientific fields, such as mechanical engineering and statistical physics, many different formal definitions of DoF are used in the literature. McCarthy defines degrees of freedom of a mechanical system as follows.

> We derive formulas for the number of parameters needed to specify the configuration of a mechanism, in terms of the number of links and joints and the freedom of movement allowed at each joint. This number is the *degrees of freedom* or *mobility* of the mechanism. Changing the values of these parameters changes the configuration of the mechanism. Thus, if we view the set of all configuration available to a mechanism as a manifold, then the mobility of the mechanism is the dimension of this

manifold.

For our purposes, we define the **degrees of freedom** of a configuration to be the difference between the number of parameters needed to specify a configuration in an ambient parameter space Ω and the number of independent constraints imposed upon the configuration as a function of these parameters. Consequently, the non-trivial degrees of freedom is the number of degrees of freedom minus the six trivial degrees of freedom. If the ambiant parameter space is $\mathbb{R}^N$ and there are M constraint equations given by $\varphi : \mathbb{R}^N \to \mathbb{R}^M$, we find the degrees of freedom to be $N - \text{rank}(\varphi)$ and the non-trivial degrees of freedom to be $N - \text{rank}(\varphi) - 6$. Here we need φ to be a smooth map so that the rank is well defined. Furthermore, we define the **configuration space** to be the subset of parameter space $\{\omega \in \Omega : \varphi(\omega) = 0\}$ upon which the contraints are satisfied.

In our case, the location and orientation of each polygon in the linkage can be represented by a total of 6 parameters: 3 for location and 3 for roational orientation. Thus, the entire configuration of an intermediate x can be represented in the parameter space $\mathbb{R}^{F_x \times 6}$. With constraints corresponding to hinged connections we construct a function φ which is zero if and only if the constraints are satisfied. Thus the configuration space is $\{z \in \mathbb{R}^{F_x \times 6} : \varphi(z) = 0\}$. As we will show below, the constraint equation φ can be constructed as a polynomial which means that the configuration space (as the zero set of this polynomial) is an algebraic variety.

PARAGRAPH ON DIFFERENT DOF VALS AT DIFFERENT CONFIGS OF SAME INTERMEDIATE We refer to the embedded configuration of x such that x is a subset of the original polyhedron as the **canonical configuration**.

To mathematically describe a particular configuration, we specify the locations of the vertices of each face. Thus, the configuration space can be described as a manifold embegged in $\mathbb{R}^{3\times N_x}$ where $N_x \doteq \sum_{f\in(\mathcal{P})} s_f \mathbb{1}_{f\subset x}$ and s_f is the number of sides (and verticies) of face f. By parameterizing the configuration space as enmedding of the ambient space $\mathbb{R}^{3\times N_x}$, we must then identify the corresponding constaint equations as a function of points in ambient space. It is worth noting that while we represent each face with $3\times s_b$ coordinates, only 6 are required to speciafy a face's position and orientation if they are chosen carefully. However, this redundancy will be removed via the constraint equations. Notationally, we refer to the kth vertex of the jth face of x as $v^{jk} = \left(v_x^{jk}, v_y^{jk}, v_z^{jk}\right)$. There are five fundamental types of constraint equations: base constraints, edge length constraints, vertex identification constraints, angle constraints, and 2D face constraint.

The base constaints are put in place to fix one of the intermediate's faces in space to remove the 6 trivial degrees of freedom from the calculation. If f_b is the face we wish to designate as the base we have the following $3 \times s_b$ constraint equations

$$\psi_{base}^{k,x}(\mathbf{v}) = v_x^{b,k} - c_x^{b,k} \tag{4.1}$$

$$\psi_{base}^{k,y}(\mathbf{v}) = v_y^{b,k} - c_y^{b,k} \tag{4.2}$$

$$\psi_{base}^{k,z}(\mathbf{v}) = v_z^{b,k} - c_z^{b,k} \tag{4.3}$$

for $k = 1, \ldots, s_b$ and with $c_\bullet^{b,k}$ a known constant. The corresponding Jacobian calculation for each constraint equations is:

$$\frac{\partial \psi_{base}^{k,\bullet}}{\partial v} = \begin{cases} 1 & \text{if } v = v_\bullet^{b,k} \\ 0 & \text{else} \end{cases}$$

Edge length constraints enforce that the lengths of the edges of each face in an intermediate cannot change. Since there are s_j edges for each face f_j of intermediate x, there are N_x corresponsing edge constraints.

$$\psi_{edge}^{j,k}(\mathbf{v}) = \left|v^{j,k} - v^{j,k-1}\right|^2 - \ell_{j,k}^2 \tag{4.4}$$

$$= \left(v_x^{j,k} - v_x^{j,k-1}\right)^2 + \left(v_y^{j,k} - v_y^{j,k-1}\right)^2 + \left(v_z^{j,k} - v_z^{j,k-1}\right)^2 - \ell_{j,k}^2 \tag{4.5}$$

for all $j : f_j \subset x$ and with the convention that $v^{j,0} \doteq v^{j,s_j}$ and $\ell_{j,k}$ is a known constant. The resulting partial derivatives are

$$\frac{\partial \psi_{edge}^{j,k}}{\partial v} = \begin{cases} 2\left(v_\bullet^{j,k} - v_\bullet^{j,k-1}\right) & \text{if } v = v_\bullet^{j,k} \\ -2\left(v_\bullet^{j,k} - v_\bullet^{j,k-1}\right) & \text{if } v = v_\bullet^{j,k-1} \\ 0 & \text{else} \end{cases}$$

4.3 Cyclohexane Application

4.3.1 Sachse Model

Around the turn of the century, it was thought that cyclohexane's carbon atoms must lie in a plane. A young German assistant, Hermann Sachse, had the idea that allowing the carbons to lie outside the plan could alleviate the angle strain. Inspired by polyhedral geometry, he templates and outlined methods for creating 3D models of the chair and boat configurations his new theory conceptualized. Figure **??** shows a construction of these two models. Despite his best efforts, Sachse's ideas were not accepted by that chemistry community until after his death.

4.4 Idealized Constraint Model

Withe the eclipsing and angle strains in mind, we heuristically define an idealized model of cyclohexane by imposing geometric constraints. Each configuration is represented by the 3-dimensional locations of the center of its carbon atoms and we explicitly parameterize these locations as $v_1, v_2, \ldots, v_6$ where $v_k \in \mathbb{R}^3$. For ease of exposition, $v_0 \doteq v_6$ and $v_{-1} \doteq v_5$ are notationally identified.

First, we require that any two atoms sharing a bond have a known and fixed distance ℓ from each other. Since we can re-scale our coordinate system, we assume $\ell \doteq 1$. Additionally, we assume that the connectivity of the cyclohexane molecule is such that v_k is bonded to v_{k-1} for $k = 1, \ldots, 6$. This gives our first six constraint equations.

$$0 = \varphi^k_{len}(v_{k-1}, v_k) = \|v_k - v_{k-1}\| - 1$$

for $k = 1, \ldots, 6$.

Additionally, we impose constraints representing the angle strain. Since the carbon atoms have the lowest energy when their bonds are at tetrahedral angles, we fix the angle of each set of three adjacent carbons to be at the tetrahedral angle. This is equivalent to the six angle constraint equations

$$0 = \phi^k_{ang}(v_{k-2}, v_{k-1}, v_k) = (v_k - v_{k-1}) \cdot (v_{k-2} - v_{k-1}) + \frac{1}{3}$$

for $k = 1, \ldots, 6$.

4.4.1 Degrees of Freedom in Ideal Model

The first step in answering these questions is determining whether there is any degrees of freedom to each configuration or whether they are rigid. This can be determined by using established theory on rigidity of linkages and degrees of freedom. By computing the Jacobian matrix $J(v)$ of the system of constraint equations, we have the following equation for the degrees of freedom for a constraint satisfying set of coordinates v.

$$DoF(v) = 18 - rank(J(v))$$

When the Jacobian is of full rank 18, none of the constraints are dependent on each other and thus there is no freedom in the linkage. When testing the chair, we found there to be 0 degrees of freedom, meaning that it is impossible to make a transition from the chair to another configuration without breaking one or more of the constraint equations. For the boat, however, one degree of freedom was found. This result shows that it is possible to deform the boat continuously while satisfying the constraints, but it is not informative as to which configurations it can deform to. In particular, we are interested in finding a path to he twist boat or another boat configuration.

4.5 Folding Configuration Space

The number of degrees of freedom measures the rigidity of an intermediate, but some intermediates with the same number of degrees of freedom may have varying degrees of mobility. We seek to quantify an intermediate's mobility by the relative amount of movement the intermediate has for a small movement in its configuration

space. Intermediates with the most mobility will move less than the less mobile intermediates for the same amount of movement in configuration space.

4.5.1 Methods

Each octahedron intermediate is composed of 8 equilateral triangles connected to each other along edges. We make the assumption that these triangles are rigid and cannot be deformed. Furthermore, when two triangles meet at an edge, they move relative to each other as if connected by an ideal hinge. Every configuration has 6 trivial degrees of freedom corresponding to the 3 translation degrees of freedom and the 3 rotational degrees of freedom. Since we are interested in the motion of an intermediate's faces relative to each other, we remove these trivial degrees of freedom by picking a single face to fix in space. Depending on the connectivity of the triangle's edges, an intermediate may still have degrees of freedom. The Octahedron intermediate (83), being a convex polyhedron, is rigid and has no non-trivial degrees of freedom as given by Cauchy's Theorem. However, this is not the case for most of the intermediates.

We formalize the concept of the configuration space, by defining it to be the subset of an ambient parameter space that satisfies constraint equations that correspond to our assumptions. Since the are 8 faces in each intermediate and each face has 3 vertices each with 3 spacial coordinates (x,y,z), we use $\mathbb{R}^{8\times3\times3} = \mathbb{R}^{72}$ as our ambient space. We then have 3 types of constraint equations: base face constraints, rigid face constraints, and hinge constraints. Every admissible configuration of an intermediate will be a point in $\mathbb{R}^{72}$ that satisfies the constraint equations and any admissible movement of a configuration (if possible) is a continuous movement in the subset of $\mathbb{R}^{72}$ where these constraints are satisfied.

Since our constraint equations can all be expressed as polynomials, the configuration space which is the corresponding solution set is an Algebraic variety. The number of degrees of freedom of a configuration is the local dimension of this algebraic variety as a subset of $\mathbb{R}^{72}$. In the case of octahedron intermediates, the number of degrees of freedom is not an informative way of classifying which intermediates are dominant as intermediates $1-11$ have 7 degrees of freedom, $12-33$ have 5, $34-65$ have 3, $66-82$ have 1 and the Octahedron (83) and Boat (84) have 0 degrees of freedom. Thus, a different measure the mobility of a configuration is required to differentiate intermediates. It is important to note that degrees of freedom of in intermediate can theoretically change based on the region of configuration space. However, in practice we have not observed an intermediate's degrees of freedom change in this way.

We define the *canonical configuration* of an intermediate, to be that in which each hinge forms a 180° angle when the hinge does not have a vertex connection at either end and in the case of a closed vertex, the hinge's angle corresponds to the Octahedral angle. In cases where the intermediate cannot form an octahedron, we can use the Boat configuration's angles instead.

Given a particular configuration X, if we wish to make a small move in configuration space, we first find the null space of the Jacobian matrix of the constraint equations. Since this null space corresponds to the directions in which the constraint equations are not changing, the constraints will remain satisfied. The null space can be represented by an orthonormal basis $N \in \mathbb{R}^{72\times d}$, where d is the number of degrees of freedom of the configuration. Then, by taking a small step of size ϵ in each of the d directions we get the configurations $X + \epsilon N_{\cdot,k}$.

We define the *mobility* of a configuration to be the L^2 norm of the gradient of the

configuration space. To approximate this, we measure the mean squared distance between each point of X and $X \pm \epsilon N_{\cdot,k}$ in each of the d directions of N, take the norm, and divide by ϵ. This is given by Equation 4.6 where $\triangle_f$ refers to the fth face of the intermediate and $r(X)$ is the point we are integrating over and $r(X + \sigma\epsilon N_{\cdot,k})$ is the corresponding point in the altered configuration.

To explicitly define the rotation and translation of each face by the movement in configuration space, we use the algorithm given by Arun et al [**?**]. Given two sets of points $P, P' \in \mathbb{R}^{3\times n}$ in 3D space, the algorithm finds the rotation matrix R and translation vector b minimizing the least squares error of $P' \approx RP + b$. By using the three vertices of face f as P and their perturbed values as P', we use this algorithm to find the rotation matrix R_f and translation vector b_f to describe the rigid movement of face f in the configuration space. This leads to the formulation in Equation 4.8.

To make this integral easier to solve, we introduce a change of variables that enables us to integrate in the $x - y$ plane. If the the original triangle f has vertices of a', b', and c', we consider the triangle with vertices at the coordinates $a = (0, 0, 0), b = (|a - b|, 0, 0)$, and $c = (\alpha, \beta, 0)$ as seen in Figure **??**. Here, the choices $\alpha = \frac{|a-b|^2+|a-c|^2-|b-c|^2}{2|a-b|}$ and $\beta = \sqrt{|a - c|^2 - \alpha^2}$ yield a triangle congruent to the original. Using the Arun et al [**?**] algorithm again, we find the rotation matrix S_f and translation vector c_f that gives $[a', b', c'] = S_f[a, b, c] + c_f$. Using this we perform a change of variables and integrate over s and t. This results in an simple double integral over a quadratic function of s and t with constants $u, v, w \in \mathbb{R}^3$ as seen in

Equation 4.10.

$$\mathcal{R}(x) = \frac{1}{\epsilon}\left(\sum_{k=1}^{d}\sum_{f=1}^{8}\int_{\triangle_f}\left|\frac{r(x+\epsilon N_{\cdot,k}) - r(x)}{2\sqrt{3}}\right|^2\right)^{\frac{1}{2}} \tag{4.6}$$

$$= \frac{1}{2\sqrt{3}\epsilon}\left(\sum_{k=1}^{d}\sum_{f=1}^{8}\int_{\triangle_f}|R_f r + b_f - r|^2\right)^{\frac{1}{2}} \tag{4.7}$$

$$= \frac{1}{2\sqrt{3}\epsilon}\left(\sum_{k=1}^{d}\sum_{f=1}^{8}\int_{\triangle_f}|(R_f - I)\, r + b_f|^2\right)^{\frac{1}{2}} \tag{4.8}$$

$$= \frac{1}{2\sqrt{3}\epsilon}\left(\sum_{k=1}^{d}\sum_{f=1}^{8}\int_{0}^{\beta}\int_{\frac{\alpha}{\beta}t}^{|a-b|+\frac{\alpha-|a-b|}{\beta}t}\left|(R_f - I)\left(S_f\begin{bmatrix}s\\t\\0\end{bmatrix} + c_f\right) + b_f\right|^2 dsdt\right)^{\frac{1}{2}} \tag{4.9}$$

$$= \frac{1}{2\sqrt{3}\epsilon}\left(\sum_{k=1}^{d}\sum_{f=1}^{8}\int_{0}^{\beta}\int_{\frac{\alpha}{\beta}t}^{|a-b|+\frac{\alpha-|a-b|}{\beta}t}|u + sv + tw|^2\, dsdt\right)^{\frac{1}{2}} \tag{4.10}$$

Unfortunately, the particular choice of base face affects the mobility. To rectify this bias, we compute the mobility with each of the eight faces as the base and take the average. This gives the final mobility in Equation 4.12.

$$\mathcal{R} = \frac{1}{8}\sum_{g=1}^{8}\mathcal{R}_g \tag{4.11}$$

$$= \frac{1}{16\sqrt{3}\epsilon}\sum_{g=1}^{8}\left(\sum_{k=1}^{d}\sum_{f=1}^{8}\int_{0}^{\beta}\int_{\frac{\alpha}{\beta}t}^{|a-b|+\frac{\alpha-|a-b|}{\beta}t}|u + sv + tw|^2\, dsdt\right)^{\frac{1}{2}} \tag{4.12}$$

4.6 Results

Mobility was computed for all octahedral intermediates in their canonical configuration. The perturbation parameter $\epsilon = 10^{-6}$ was selected, but the mobility showed to be robust to both smaller and larger values. Figure **??** outlines the results. From these calculations, it is clear that while mobility is highly correlated to degrees of freedom, mobility gives additional information about an intermediate's configuration space. Interestingly, all nets had the same mobility. Similarly, the symmetric pairs

(14,16), (15,18), (19,20), (34,38), and (36,39) have matching mobility values. Of the intermediates with 5 degrees of freedom, 19 and 20 had the lowest mobility and 17 had the highest. As for intermediates with 3 degrees of freedom, 35 was the least mobile and 37 was the most mobile.

Intermediate	Mobility
1	0.20518
2	0.20518
3	0.20518
4	0.20518
5	0.20518
6	0.20518
7	0.20518
8	0.20518
9	0.20518
10	0.20518
11	0.20518
12	0.18376
13	0.18313
14	0.18249
15	0.18171
16	0.18249
17	0.18634
18	0.18171
19	0.18102
20	0.18102
21	0.18255
22	0.18485
34	0.14487
35	0.14350
36	0.14530
37	0.14973
38	0.14487
39	0.14530
66	0.07954
83	0.00000

Chapter Five

Processes in Constraint Spaces

5.1 Constrained Dynamics

5.1.1 Cyclohexane Dynamics

Cyclohexane is a molecule composed of six carbon atoms and twelve hydrogen atoms. The carbon atoms are connected in a ring with two hydrogen atoms attaching to each carbon. Since each carbon has four bonds, the energetically preferred bond spacing is at tetrahedral angles. While this preference dictates much of the cyclohexane structure, there are several structurally distinct conformations the molecule can take.

Of the many forces acting on the cyclohexane molecule, we focus on three. **Eclipsing strain** refers to the force between the carbon atoms that prevents them from getting too close to each other. This imposes a preferred distance between each pair of bonded atoms. Second, **angle strain** corresponds to the carbon atom's four bonds trying to spread apart from each other. As mentioned before, tetrahedral angles are preferred. Finally, **steric crowding** is similar to eclipsing strain, but the eclipsing in this case is between the hydrogen atoms bonded to the carbons.

5.1.2 Configurations

Due to the aforementioned forces, there are a variety of frequently observed configurations with geometries that alleviate these strains to varying amounts.

The lowest energy configuration is the **chair**. With the absence of both angle and eclipsing strain, the chair only has a small amount of steric strain.

Another well studied configuration is the **boat**. With two sets of parallel carbons

arranged in a planar rectangular structure, the remaining two carbon atoms sit at opposite ends slightly above the plain. The structure resembles the bow and stern of a boat. Despite having little angle strain, it does have some steric crowding between the hydrogen atoms at the ends of the boat as well as some eclipsing stain.

Besides the chair and boat, there are a few intermediate configurations that cyclohexane assumes as when it transitions between the chair and boat. The aptly named **twist boat** is similar to the boat, but some of the carbons are rotated from the rest of the molecule. Due to this twisting, the twist boat actually has a lower energy than the boat as the hydrogen atoms at the bow and stern are no longer aligned. Interestingly, the boat can transition via the twist boat to a second boat configuration in which the bow and stern atoms are different, but with an otherwise indistinguishable geometry. As part of the boat's transition to the chair, the **twist chair** is another distinguished configuration. Having large angle and eclipsing strains, the twist boat represents somewhat of an energy barrier between the chair and boat.

5.1.3 Finding Intermediate Coordinates

Since any configuration has an infinitude of equivalent configurations given by rotations and translations, we fix the first three atoms v_1, v_2, v_3, at positions c_1, c_2, c_3 that satisfy the length and angle constraints. These account for nine of the twenty-one total constraint equations.

We wish to examine each of the four cyclohexane configurations to verify if they satisfy these constraints, but to do so, an explicit representation of the coordinates

of each configuration is required. Since we know from Sachse's model that the boat and chair have a special polyhedral representation, geometry enables us to find such coordinates. We pick the $c_1 = \left(-\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right)$, $c_2 = (0,0,0)$, $c_3 = \left(\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right)$, and solve for the following coordinates.

$$
\begin{aligned}
v_1^{(chair)} &= \left(-\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right) & &= v_1^{(boat)} = \left(-\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right) \\
v_2^{(chair)} &= (0,0,0) & &= v_2^{(boat)} = (0,0,0) \\
v_3^{(chair)} &= \left(\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right) & &= v_3^{(boat)} = \left(\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right) \\
v_4^{(chair)} &= \left(\sqrt{\frac{2}{3}}, \sqrt{\frac{2}{3}}, 2\sqrt{\frac{1}{3}}\right) & &= v_4^{(boat)} = \left(\sqrt{\frac{2}{3}}, \sqrt{\frac{2}{3}}, 2\sqrt{\frac{1}{3}}\right) \\
v_5^{(chair)} &= \left(0, \sqrt{\frac{2}{3}}, \sqrt{3}\right) & v_5^{(boat)} &= \left(0, \frac{5}{3}\sqrt{\frac{2}{3}}, \frac{5}{3}\sqrt{3}\right) \\
v_6^{(chair)} &= \left(-\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right) & &= v_6^{(boat)} = \left(-\sqrt{\frac{2}{3}}, 0, \sqrt{\frac{1}{3}}\right)
\end{aligned}
$$

It is easily verified that both the boat and chair coordinates satisfy all of the constraint equations exactly. As for the twist boat and twist chair, we do not have a precise definition of their coordinates other than they exist somewhere in transition between the boat and chair.

5.1.4 Dynamics

As we are interested in transitions between the four configurations, we can use our constraint model to examine such movements. Is it possible to start with the chair

coordinates and continuously deform them to the boat coordinates without ever breaking any of the constraints? Is there a similar deformation between different boat configurations that satisfies the constraints? If so, how hard is it to find such a path?

5.1.5 Transitioning Between Boat Configurations

Under the conjecture that it is indeed possible to transition between two boat configurations, some numerical and visual experimentation was used to find coordinates to a second boat configuration, **boat2**. Even though it was easily verified that boat2 satisfies all of the constraints and was equivalent to the first boat by translation and rotation up to a permutation of the carbon atoms, it was not clear how to find a path between the two boats. Common in molecular dynamics simulations, a constrained dynamics SHAKE-type scheme was used to explore the admissible paths away from the boat configuration.

The method is a two stage scheme in which we first step to a new configuration in the direction of boat2 and second enforce the constraints with Lagrange multipliers to get the updated configuration. For mathematical simplicity, we represent the coordinates $v_1, \ldots, v_6$ as a single point $x \in \mathbb{R}^{18}$ where $(v_j)_k = x_{3j+k}$.

It is very important to make the estimate $\hat{x}^{n+1}$ of the next configuration satisfy the constraints reasonably well, as it will make the correction step easier. Rather than just defining the naive update of

$$\hat{x}^{n+1} = x^n + \frac{1}{2}(\Delta t)^2 \left(x^{boat2} - x^n\right),$$

Figure 5.1: Boat to Boat2 Coordinate Transition

we seek a smarter method for making the estimate $\hat{x}^{n+1}$. Since we know the that the boat has one degree of freedom, its null space has one dimension. If we step in the direction of null-space $\nu^n \in \mathbb{R}^{18}$, the constraints should still be close to being satisfied. Thus, we use the following estimate.

$$\hat{x}^{n+1} = x^n + \frac{1}{2}(\Delta t)^2 \left[\nu^n \cdot \left(x^{boat2} - x^n \right) \frac{\nu^n}{\|\nu^n\|} \right]$$

To enforce our constraints on this prediction, we use Lagrange multipliers. We define the update to be

$$x^{n+1} = \hat{x}^{n+1} + (\Delta t)^2 \sum_{k=1}^{21} \lambda_k^{n+1} \nabla \varphi_k \left(x^n \right)$$

and find λ^{n+1} such that x^{n+1} satisfies the constraint equations. To do this, we plug x^{n+1} into the constraint equations φ and use Newton iteration to find λ^{n+1}.

Using this scheme, we were able to simulate the transition between the boat and boat2 configurations. While maintaining the constraints up to arbitrary precision, the carbon coordinates x were updated along a path that lead from the boat to boat2. This serves as confirmation that it is possible to maintain the constraints given by our model and transition between boat and twist boat configurations. The continuous change in carbon coordinates during this transition can be seen in figure 5.1.

This algorithm also enabled us to approximate coordinates for the twist boat as they were taken to be midway between the boat and boat2. Also, by relaxing the constraints to only enforce fixed lengths between carbons and allowing some angle strain, a path from the Chair to the twist boat was found. Similarly, this path allowed for the approximation of the twist chair coordinates.

5.1.6 Analogy to Folding Model

Transition between the different configurations of cyclohexane can be compared to the transitioning within the folding model configuration space. Just as each folding intermediate is represented by a node in a graph, the Cyclohexane intermediates can similarly be organized. In both graphs an edge would mean that the intermediates at either end could transition into each other. As seen in figure 5.2, to transition between the chair and the boat, the molecule must first become a twist-chair and then a twist boat before it can finally become a boat. Similarly, to transition between the boat and boat2, the twist boat intermediate must first be visited. This relationship is mirrored by the boat and octahedron folding intermediates.

Figure 5.2: Configuration Graphs for Cyclohexane and Folding

5.2 Manifold Reflected Brownian Motion

5.2.1 Computational Implementation

5.2.2 Validation and Test Cases

5.2.3 MRBM on Building Game Geometric Configuration Spaces

Self-intersection Boundary

Fixing the Center of Mass

Fixing Rotations

BG Test Cases

Chapter Six

Results

Bibliography

[1] J Rotman. *An Introduction to the Theory of Groups*. Springer-Verlag, New York, NY, fourth edition, 1995.

[2] A Zlotnick. An equilibrium model of the self assembly of polyhedral protein complexes. *Journal of Molecular Biology*, 241:59–67, 1994.