Discrete Attachment Models for Self-Assembly of Supramolecular Polyhedral Structures

Daniel Johnson
Division of Applied Mathematics
Brown University

June 25, 2013

Contents

1 Introduction

We are primarily interested in understanding different self-assembly processes by using discrete geometrical attachment models. Applications of interest include metal-ligand coordination spheres, molecular cages, and viral capsid assembly. Each involve the progressive formation of a structure from a set of smaller, more basic building blocks. We hope to uncover patterns that these

processes share and those that dictate a successful assembly. What are the possible pathways of assembly? When there are many pathways, are some dominant over others, and why?

1.1 Scientific Motivations

Sun et al. theorized five metal-ligand coordination sphereswe refer to as Fujita supermolecules [15]. Seen in Figure 1, each is composed of n metallic M molecules each connecting four of the $2n$ bent ligand L molecules. The

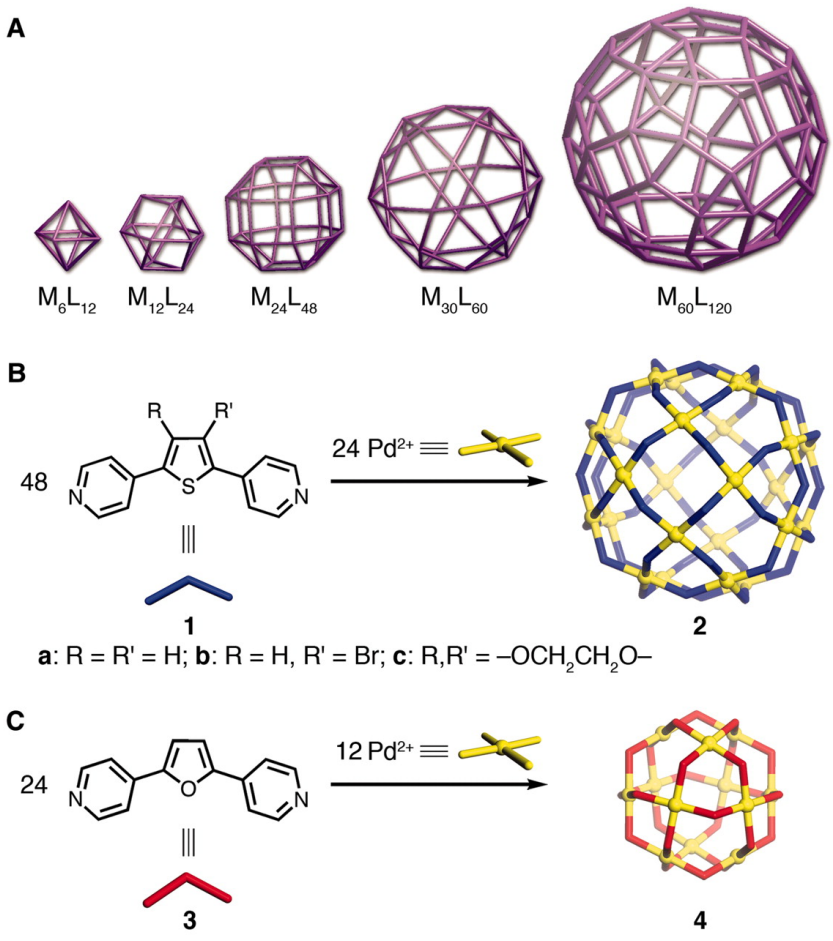

Figure 1: (A) Polyhedral representations of the five Fujita molecules (B,C) Molecular structure of $M_{24}L_{48}$ and $M_{12}L_{24}$ [15]

entire structure is referred to as the M_nL_{2n} supermolecule. Interestingly, geometric reasons only allow for $n \in \{6, 12, 24, 30, 60\}$ and thus five different supermolecules.

As yet, only the first three of Fujita's supermolecules have been synthesized in laboratory experiments: M_6L_{12}, $M_{12}L_{24}$, and $M_{24}L_{48}$. Our main focus is the mathematical explanation of a phenomenon observed in an experiment involving two different types of ligand molecules, each with slightly different bend angle; call them L^1 and L^2. For certain ratios $L^1 : L^2$ in the experiment, only $M_{12}L_{24}$ were successfully synthesized and for other ratios, only $M_{24}L_{48}$ were synthesized. Most interestingly, there was no ratio found that resulted in the formation of both the $M_{24}L_{48}$ and $M_{12}L_{24}$. This is indicative of a phase transition or some other, similarly steep process. Us-

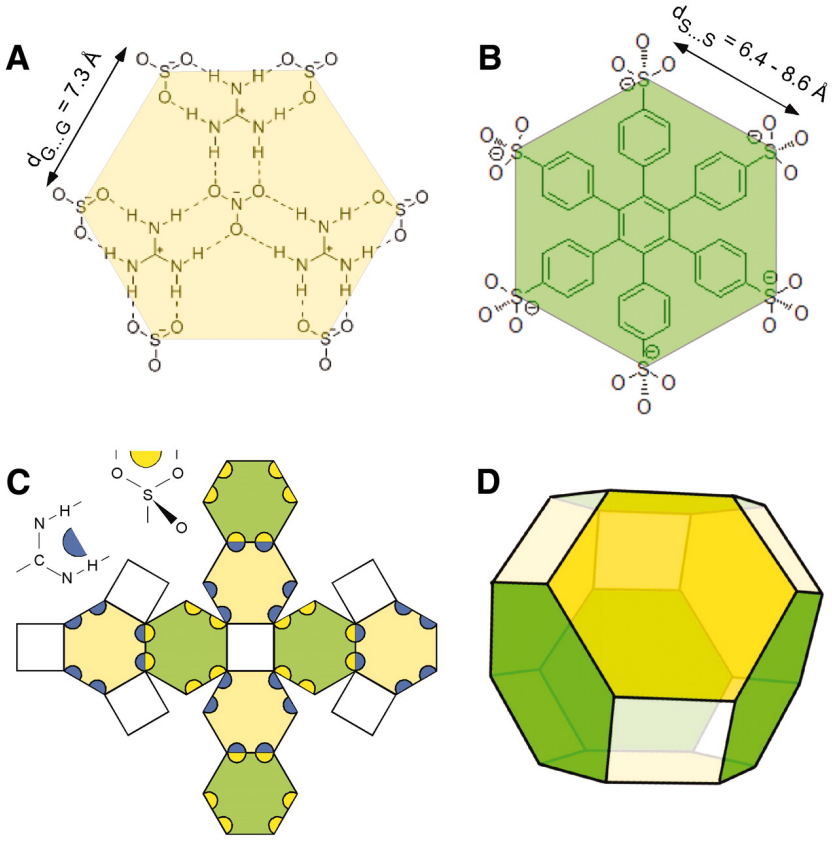

Figure 2: Ward's quasi-truncated octahedron molecular cage [10]

ing discrete attachment models, we hope to uncover the true nature of this phenomena.

Another self-assembly application of interest is the formation of supramolecular cages. Figure 2 depicts a molecular cage synthesized by Liu et al. using two types of hexagonal shaped molecules [10]. The cage is shaped like a truncatd octahedron, with 8 hexagonal faces and 6 square hols, and thus was names the quasi-truncated octahedron (qTo). In experiments, the qTo was shown to be able to cage a several different types of smaller molecule.

While the modeling of the qTo's formation is an interesting question in itself, there are a few other related questions that our discrete attachment models should provide incite toward. Since one of the qTo's two hexagonal molecule types can bond to itself, it would be theoretically possible for a honeycomb-like tiling to occur in experiments, but this was not observed. Perhaps our models can explain why the qTo structure is a more favorable formation to such tiling. Additionally, if one considers a different set of building block molecules, different polyhedral structures should be possible. For

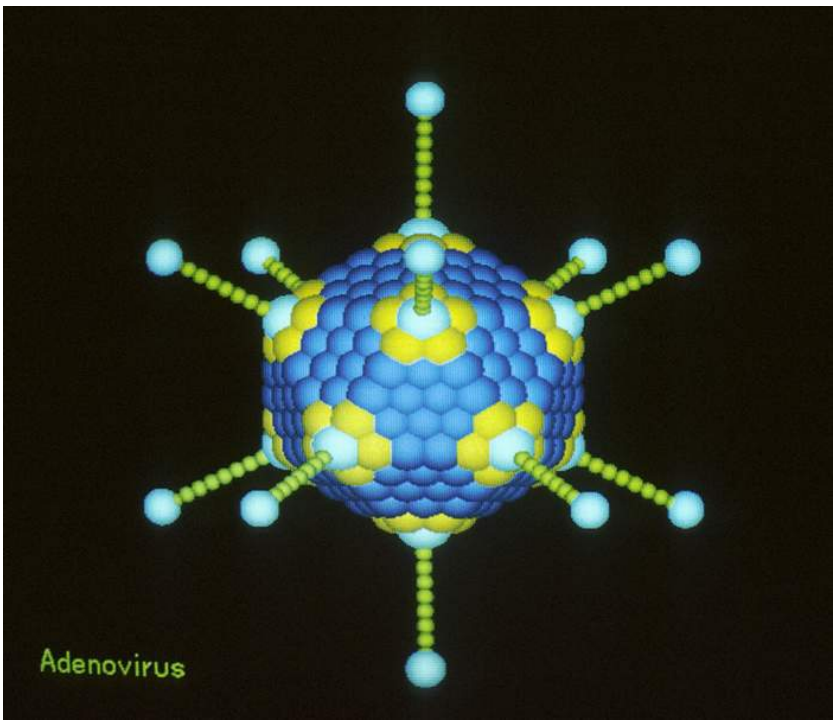

Figure 3: The icosahedral structure of the Adenovirus

example, if square molecules are combined with hexagonal molecules, there are three Archimedean solids that can be formed: truncated octahedron, truncated cuboctahedron, and truncated icosidodecahedron with hexagonal, octagonal, and decogonal holes respectively. Though no lab experiments investigating these possibilities have been performed, perhaps our attachment models could shed light on strategies for synthesizing each polyhedral structure.

Another notable self-assembly process is the formation of viral capsids. The majority of biological virus capsids have an icosahedral structure (Figure 3) and their formation process is not well understood. While there is evidence that suggests some viral capsids are aided by scafolding protein during formation, discrete attachment models may still be useful and because of the vital implications to the medical community they are the subject of much research.

1.2 Mathematical Motivations

While much of our work is motivated by these thoroughly scientific questions, there are several connections to theoretical mathematics that are pertinent in their own right. The discrete geometric nature of the attachment models we consider lends itself to questions regarding polyhedral construction formalisms such as shellability. Furthermore, what kinds of specimen can our models produce? This provides a natural enumeration problem. After identifying all rotationally unique species, we also seek to understand their connection to each other in the context of the model. The state space of our model is represented by a graph with the related states connected. How can we mathematically determine which states are the most important by examining this graph? Does the graph have special properties? What kind of metrics are reasonable to define on the graph?

1.3 Similar Work

The class of discrete attachment models we consider specify (1) the basic substructures used to assembly the desired structure, and (2) rules that specify the ways in which the substructures may combine during assembly. The study of such models is decades old [3]. Several papers have examined a local rules approach to viral capsid assembly in which the subunits each have a specific conformation and the configurations in which these conformations

combine is specified by a list of templates [1, 13, 7, 6]. Other work has focused more on the feasibility of such attachment models by conducting large-scale physical experiments that are easily observed [8, 2, 9]. One of the models we consider, which puts minimal constraints on the ways in which substructures can combine was first studied by Zlotnick [16, 4].

While Pandey et al. and Gracias et al. [12, 5] consider a fundamentally different model of self-assembly, our analysis will be of the same nature. First the model's state space is enumerated and organized into a graph and then various tool are used to decipher which states are most important, including energy functions and Markov processes.

1.4 Research Directions

Our work currently has two major directions. First, we have done extensive computation and analysis relating to a discrete attachment model called the building game. Secondly, we do a structuarl analysis of Fujita supermolecules composed of the two types of ligand trying to identify the source of the transition phenomenon. Since we both consider physical self-assembly models and corresponding mathematical structures, we hope that our work will be of interest to computational chemists and combinatorialists alike.

Thus far, the building game work has been presented as a poster at the NSF Building Engineered Complex Systems grant conference and as an hour long talk at the Brown University Graduate Student Statistics Seminar. Currently, a paper focusing on the geometric and combinatorial aspects of the building game is in preparation for submission to *Experimental Mathematics*.

2 The Building Game

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 verticies 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.

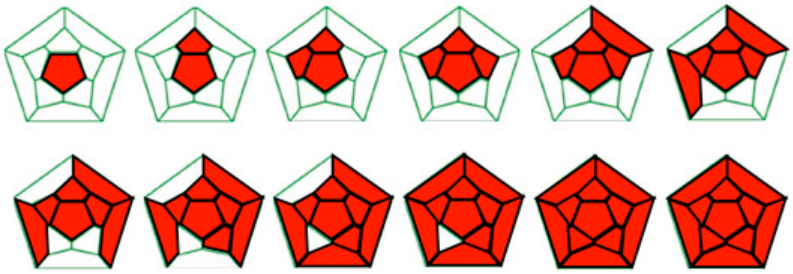

Figure 4: Dodecahedron building game example.

The **configuration space** for a particular polhedron $\mathcal{P}$ is a graph that represents the space of all distinct intermediates and the BG pathways for $\mathcal{P}$. Each node of the configuration 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 configuration space, starting at an intermediate with one face and ending at the intermediate with all faces, represents one of the polyhedron's BG pathways.

For a configuration 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 configuration 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 configuration space. While the 6-faced cube configuration space has only 8 vertices and 9 edges, the 20-faced

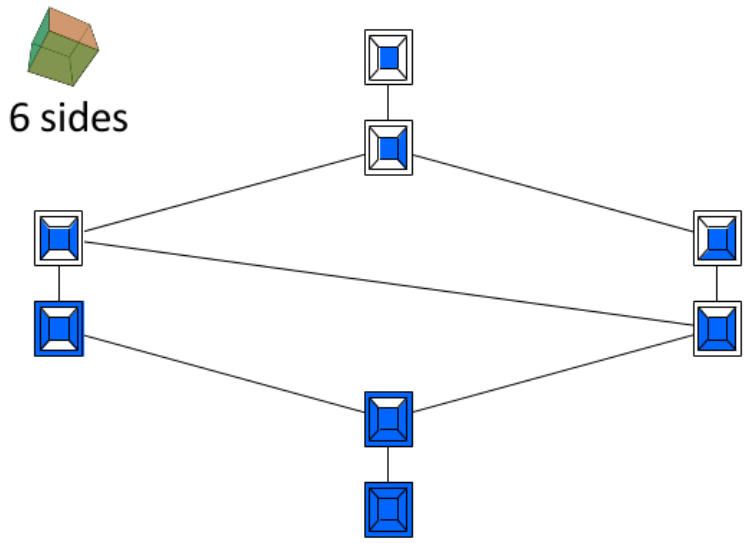

Figure 5: Configuration space of the cube

icosahedron configuration space has 2,649 vertices and 17,241 edges and the 26-faced truncated cuboctahedron configuration space has 1,525,605 vertices and 17,672,377. Figure 8 details configuration space sizes of all polyhedra in the Platonic, Archimedean, and Catalan solid classes of up to 26 faces. Computational constraints do not currently allow us to compute the building game configuration space for the remaining polyhedra in these classes.

2.1 Computation

The configuration space is computed sequentially. Given all of the rotationally unique intermediates that have $t-1$ faces, we compute the rotationally unique intermediates with t faces. The majority of the computation is spent determining if a given intermediate is rotationally equivalent to another. The indices of all possible rotations are precomputed and this comparison reduces to simply checking if two (often binary) vectors are identical. Code for computing the building game configuration space was written in C++ and an outline of the algorithm is in figure 9.

Because there may be hundreds of potential rotations that must be checked for each comparison of two intermediates, a hash function was developed that

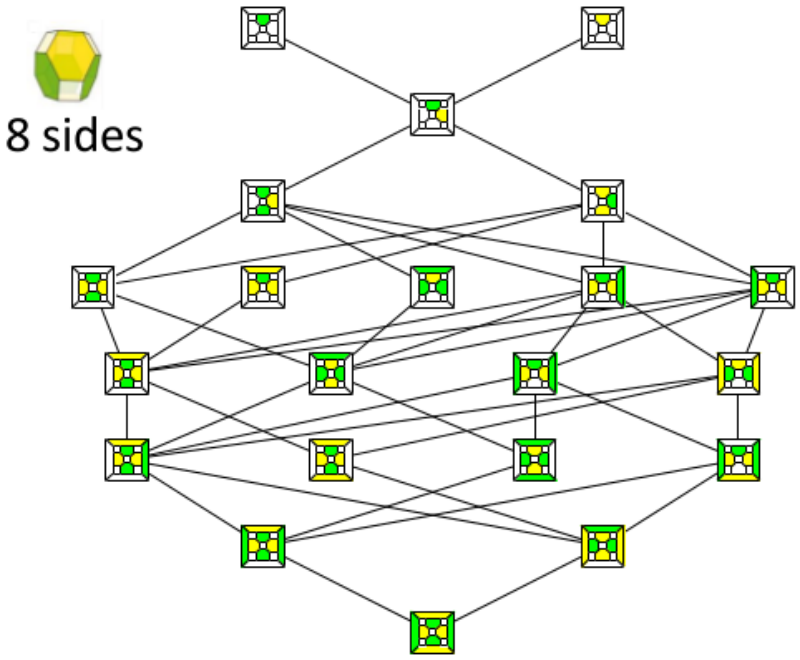

Figure 6: Configuration space of the quasi-truncated octahedron

takes an intermediate and returns an integer such that any two rotationally equivalent intermediates are mapped to the same integer. However, the hash function is not one to one. Two intermediates with the same hash may not be equivalent. The hash, a serialized histogram of an intermediate's connectivity, is relatively inexpensive to compute and has the potential to eliminate the majority of comparisons required as no two intermediates with different hashes can be equal. In the case when two intermediates are found to have the same hash, the standard method of checking all possible rotations must still be carried out.

Computational time needed to run our algorithm is heavily dependent on the number of intermediates, which, unfortunately, seems to grow exponentially in the number of faces F a polyhedron has. The gives a worst-case computational complexity of $O\left(|F|^{\frac{3}{2}}2^{|F|}\right)$ under the reasonable assumption that $\frac{|E|}{2|F|} \ll |F|$. While it seems unlikely that an algorithm with improved worst case performance is possible, there is hope for some improvement. Though not an intrinsically parallel problem, there may be ways to significantly re-

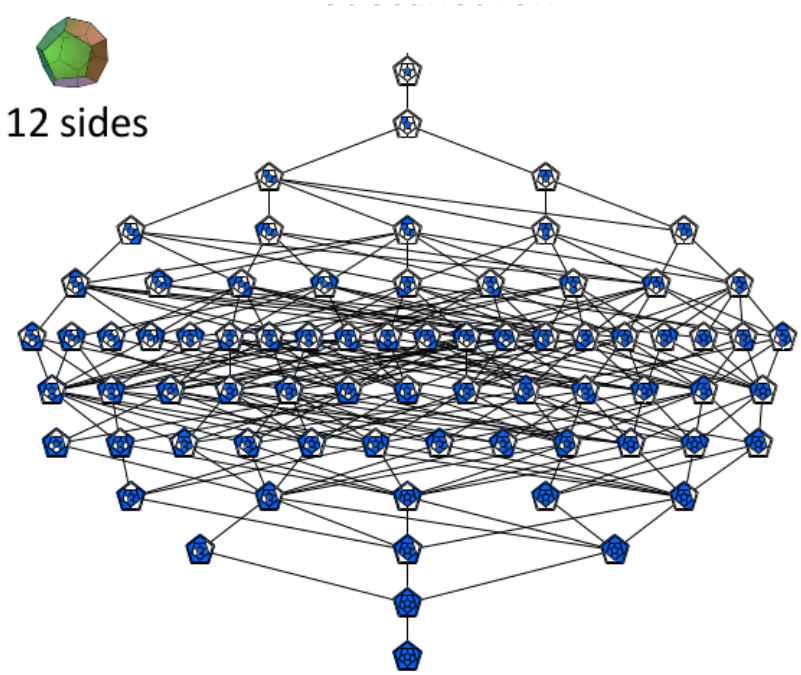

Figure 7: Configuration space of the dodecahedron

duce the wall-clock computational time by parallelizing the algorithm. It may also be possible to improve the hash to further lessen the number of long form comparisons that must be computed.

2.2 Pathway Enumeration

In the example of the icosahedron, there are 549 intermediates that have 13 faces. How can we quantify the relative importance of each of these intermediates? One natural method is to count the number of distinct building game pathways that go through each intermediate. There are 3 BG pathways for the cube, 17,696 for the dodecahedron, and after that presumably a huge number for polyhedra with more faces. While we have not yet fully explored this problem, it will provide and engaging computatinoal problem and the results should be interesting.

Shape Name	Class	Faces	Edges	Vertices	Intermediates	Connections
Tetrahedron	Platonic	4	6	4	5	4
Cube	Platonic	6	12	8	9	10
Octahedron	Platonic	8	12	6	15	22
Dodecahedron	Platonic	12	30	20	74	264
Icosahedron	Platonic	20	30	12	2,650	17,242
Truncated Tetrahedron	Archimedean	8	18	12	29	65
Cuboctahedron	Archimedean	14	24	12	341	1,636
Truncated Cube	Archimedean	14	36	24	500	2,731
Truncated Octahedron	Archimedean	14	36	24	556	3,071
Rhombicuboctahedron	Archimedean	26	48	24	638,851	6,459,804
Truncated Cuboctahedron	Archimedean	26	72	48	1,525,605	17,672,377
Icosidodecahedron	Archimedean	32	60	30	?	?
Truncated Dodecahedron	Archimedean	32	90	60	?	?
Truncated Icosahedron	Archimedean	32	90	60	?	?
Snub Cube	Archimedean	38	60	24	?	?
Rhombicosidodecahedron	Archimedean	62	120	60	?	?
Truncated Icosidodecahedron	Archimedean	62	180	120	?	?
Snub Dodecahedron	Archimedean	92	150	60	?	?
Triakis Tetrahedron	Catalan	12	18	8	99	319
Rhombic Dodecahedron	Catalan	12	24	14	128	494
Triakis Octahedron	Catalan	24	36	14	12,749	81,297
Tetrakis Hexahedron	Catalan	24	36	14	50,768	394,278
Deltoidal Icositetrahedron	Catalan	24	48	26	209,676	1,989,549
Pentagonal Icositetrahedron	Catalan	24	60	38	345,939	3,544,988
Rhombic Triacontahedron	Catalan	30	60	32	?	?
Disdyakis Dodecahedron	Catalan	48	72	26	?	?
Triakis Icosahedron	Catalan	60	90	32	?	?
Pentakis Dodecahedron	Catalan	60	90	32	?	?
Deltoidal Hexecontahedron	Catalan	60	120	62	?	?
Pentagonal Hexecontahedron	Catalan	60	150	92	?	?
Disdyakis Tricontahedron	Catalan	120	180	62	?	?

Figure 8: Table of polyhedra in the Platonic, Archimedean, and Catalan solid classes ordered by number of building game intermediates.

2.3 Energy Landscapes

Another approach to determining the class of dominant intermediates is to think of the configuration space as an energy landscape. By defining energy functions on both the configuration space's nodes (intermediate energies) and connections (transitions energies) using physically motivated heuristics, we hope to recover the relevant behaviors of the self-assembly process.

To roughly approximate free energy, we define E_j the energy of intermediate x_j to be the number of edges in the intermediate that connect two adjacent faces. We formalize this energy as

$$E_j = \#\left\{e \in E\left(\mathcal{P}\right) : e = f_e^1 \cap f_e^2 \text{ s.t. } f_e^1, f_e^2 \subset x_j\right\}$$

where $f_e^{(1)}, f_e^{(2)} \in F$ refer to the two faces of the polyhedron $\mathcal{P}$ that edge e

1. For $t = 1$ to F
   (a) For each intermediate x_k^{t-1} with $t-1$ faces
      i. For each face j that can be added to x_k^{t-1} resulting in the intermediate $x_{k,j}^t$
         A. Check all rotations of $x_{k,j}^t$ to see if it is equivalent to an intermediate already in the list of t-faced intermediates
         B. If so: connect x_k^{t-1} and $x_{k,j}^t$ in the configuration space
         C. If not: add $x_{k,j}^t$ to the configuration space and connect it to x_k^{t-1}

Figure 9: Algorithm for computing the building game configuration space

joins.

Transitions energies are defined using a simple function that seems to work well, but is not physically motivated. To ensure the transition barrier has a higher energy than the two intermediates it sits between, we use

$$E_{j,k} = \begin{cases} \max(E_j, E_k) + \alpha & \text{if } x_j \text{ connected to } x_k \text{ in configuration space} \\ \infty & : \text{else} \end{cases}$$

where α is a constant and roughly on the same order as the average value of $|E_j - E_k|$.

With energies defined on both nodes and connections in the configuration space graph, we construct the Markov chain X_t on the configuration space with the transition rule

$$p_{jk} \doteq P(X_{t+1} = x_k \mid X_t = x_j) = \frac{d_{jk}}{\gamma} e^{-\beta(E_{jk} - E_j)}$$

where d_{jk} is the degeneracy number between intermediates x_j and x_k, and γ is a normalization constant. Furthermore, we see that X_t has the stationary distribution

$$\pi_j \doteq \pi(x_j) \doteq \frac{1}{z} e^{-\beta E_j}$$

since the following detailed balance equation holds.

$$\pi_j p_{jk} = \frac{d_{jk}}{z\gamma} e^{-\beta E_{jk}} = \pi_k p_{kj}$$

It is important to note that the stationary distributions are Boltzmann by design due to the physical scale of these self-assembly processes.

Protein folding is often framed in the context of energy landscapes and a 'folding funnel' refers to the shape of the energy landscape that encourages the protein to fold in an optimal manor. Is there a similar type of 'funnel' for our self-assembly processes? What does the energy landscape say about which intermediates are most dominant? Analysis of the spectrum of the transition operator should provide a nice way to quantify each intermediate's importance, though this work is ongoing.

3 Fujita Ligand Preference Model

The curious affect of altering the ratio of two different ligand types on the formation of Fujita supermolecules is of principal interest to us. However, the building game will provide little insight as to the root of this phenomenon. To try and discover the true cause we are developing a model that looks at the $M_{12}L_{24}$ and $M_{24}L_{48}$ supermolecules from a more structural engineering inspired viewpoint. Calling the two types of ligands L^1 and L^2 with preferred bend angles of ϕ^1 and ϕ^2 we hope to examine which combinations of the two type will allow successful formation of each supermolecule and which combinations are not physically realistic. Referring to the model as the $M_n L_k^1 L_{2n-k}^2$ model, we first seek to enumerate all rotationally unique ways in which the two kinds of molecules can appear in both of the complete supermolecules. While not yet completed, this task will be very similar to the building gamer intermediate enumeration. Secondly we seek to quantify how feasible each of these ligand configurations are.

In analyzing the physical plausibility of a particular configuration, we seek to find the minimum value of a cost function G that can be physically achieved. There are two aspects that constitute a well formed supermolecule: the ligands are not bent too much more or less than their preferred angle and at each metallic connector the four attached ligands come in a angles that are nearly orthoganal to each other. For this reason, we define the cost function to be of the form

$$G(\chi, \mathcal{P}) = \inf_{\mathbf{m},\boldsymbol{\ell}} \left[\lambda g_m(\mathbf{m}, \boldsymbol{\ell}; \chi, \mathcal{P}) + (1-\lambda) g_\ell(\mathbf{m}, \boldsymbol{\ell}; \chi, \mathcal{P})\right]$$

where $\chi \in \{1,2\}^{2n}$ represents the ligand configuration of interest, $\mathbf{m} \in \mathbb{R}^{n\times 3}$ represent the 3-dimensional positions of the n M molecules, $\boldsymbol{\ell} \in \mathbb{R}^{2n\times 3}$ repre-

sent the 3-dimensional positions of the $2n$ L molecules, g_ℓ is the cost function on ligand angles and g_m is the cost function on angles of attachment to the M molecules.

We define the cost function g_ℓ to be

$$g_\ell\left(\mathbf{m}, \boldsymbol{\ell}; \chi, \mathcal{P}\right) = \sum_{e \in E(\mathcal{P})} f_\ell\left(\phi_{\chi_e} - \angle\left(m_e^1 \ell_e m_e^2\right)\right)$$

and cost function g_m as

$$g_\ell\left(\mathbf{m}, \boldsymbol{\ell}; \chi, \mathcal{P}\right) = \sum_{v \in V(\mathcal{P})} \sum_{k=1}^{4} f_m\left(\frac{\pi}{2} - \angle\left(\ell_v^k m_v \ell_v^{k+1}\right)\right) + f_m\left(\frac{\pi}{2} - \angle\left(\ell_v^k \eta_v \ell_v^{k+1}\right)\right)$$

where

$$\eta_v \doteq m_v + \frac{\sum_{k=1}^{4}\left(\ell_v^k - m_v\right)}{\left|\sum_{k=1}^{4}\left(\ell_v^k - m_v\right)\right|}$$

is a unit vector starting from m_v going in the average direction of the four connected ligands $\ell_v^{1:4}$. The functions $f_\ell, f_m : \mathbb{R} \to \mathbb{R}$ are even, convex, and satisfy $f(0) = 0$. Currently, we use

$$f_\ell = f_m = \begin{cases} \rho\left[\frac{1}{(x^2-\epsilon^2)^\nu} + \frac{1}{\epsilon^{2\nu}}\right] & : x \in (-\epsilon, \epsilon) \\ \infty & : x \notin (-\epsilon, \epsilon) \end{cases}$$

Since there is no simple closed form solution to minimization problem, we must use optimization algorithms to compute the values of G for the various ligand configurations. Currently, off-the-shelf algorithms from the scipy.optimize Python library are being used, but to ensure proper convergence and perhaps to enhance computational efficiency, we may need to code optimization schemes manually. The current code has not yet been fully tested, let alone used on the actual polyhedra on interest, but we hope to start computing results soon.

Acknowledgments

Supported by NSF grants DMS 07-48482 and EFRI 10-22638

References

[1] B Berger, P Shor, L Tucker-Kellogg, and J King. Local rule-based theory of virus shell assembly. *Proceedings of the National Academy of Sciences*, 91:7732–7736, 1994.

[2] J. Bishop, S. Burden, E. Klavins, R. Kreisberg, W. Malone, N. Napp, and T. Nguyen. Programmable parts: a demonstration of the grammatical approach to self-organization. In *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 3684–3691, 2005.

[3] M Eden. A two-dimensional growth process. In Neyman [11], pages 223–239.

[4] D Endres, M Miyahara, P Moisant, and A Zlotnick. A reaction landscape identifies the intermediates critical for selfassembly of virus capsids and other polyhedral structures. *Protein Science*, 14:1518–1525, 2005.

[5] D Gracias, R Kaplan, J Klobusicky, G Menon, and S Pandey. Building polyhedra by self-folding: theory and experiment. *pre-print*, 2013.

[6] N Grayson. *Models of molecular self-assembly for RNA viruses and synthetic DNA cages*. PhD thesis, University of York, 2012.

[7] M Hagan and D Chandler. Dynamic pathways for viral capsid assembly. *Biophysical Journal*, 91:42–54, 2006.

[8] Kazuo Hosokawa, Isao Shimoyama, and Hirofumi Miura. Dynamics of self-assembling systems: Analogy with chemical kinetics. *Artif. Life*, 1(4):413–427, January 1994.

[9] E. Klavins, R. Ghrist, and D. Lipsky. A grammatical approach to self-organizing robotic systems. *Automatic Control, IEEE Transactions on*, 51(6):949–962, 2006.

[10] Y Liu, C Hu, A Comotti, and M Ward. Supramolecular archimedean cages assembled with 72 hydrogen bonds. *Science*, 333:436–440, 2011.

[11] J Neyman, editor. *Fourth Berkeley symposium on mathematical statistics and probability*, volume 4, Berkely, CA, 1961. University of California Press.

[12] S Pandey, M Ewing, A Kunas, S Ngyen, D Gracias, and G Menon. Algorithmic design of self-folding polyhedra. *Proceedings of the National Academy of Sciences*, 108:19885–19890, 2011.

[13] R Schwartz, P Shor, P Prevelige Jr., and B Berger. Local rule simulation of the kinetics of virus capsd self-assembly. *Biophysical Journal*, 75:2626–2636, 1998.

[14] J Snell, editor. *Topics in contemporary probability and its application.* CDC Press, Boca Raton, FL, 1995.

[15] Q Sun, J Iwasa, D Ogawa, Y Ishido, S Sato, T Ozeki, Y Sei, K Yamaguchi, and M Fujita. Self-assembled M24L48 polyhedra and their sharp structural switch upon subtle ligand variation. *Science*, 328:1144–1147, 2010.

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