Modeling Autonomous Supramolecular Assembly

Meera Sitharam

Abstract Supramolecular assembly is often a remarkably robust, rapid and spontaneous process, starting from a small number of monomeric types. Although, the process occurs widely in nature and is increasingly important in healthcare and engineering, it is poorly understood. Icosahedral viral shell assembly is one such outstanding example. We sketch the experimental roadblocks that necessitate mathematical and computational modeling of assembly, and list the types of experimental data available for model validation, thereby defining the models' input and output, and framing the scope of model predictions. We isolate the various factors, specifically *configurational and combinatorial entropy* that influence spontaneous supramolecular assembly, pinpointing the modeling challenges and motivating the use of *multiscale* models. We then survey existing modeling paradigms for the modeling different scales, emphasizing the newest models and paradigms developed by the author's group, geared towards not only predicting, but also intuitively explaining, analyzing and engineering assembly processes. The models leverage geometric and algebraic characteristics unique to molecular assembly (as opposed to folding), and permit provable performance guarantees together with some level of forward and backward analysis as well as a desired level of precision and refinability of prediction.

1 Motivation

Understanding supramolecular assembly is useful for many practical applications. Rational drug design is a vast area of study and requires understanding the site-specific assembly or docking of ligands with proteins and other biomolecules. Similarly, nanoscale self-assembly of materials is a vast area of study in nanotechnol-

Meera Sitharam
University of Florida, Computer and Information Sciences and Engineering, P.O. Box 32611-6120, e-mail: sitharam@cise.ufl.edu

ogy. Many viral capsids form by self-assembly of an icosahedral shell from nearly identical coat protein monomers enclosing genomic material. Understanding how to disrupt assembly permits us to target this part of the viral lifecycle using drugs and vaccines The pathophysiology of viral infections includes other parts of the the viral lifecycle that involve site-specific docking and assembly. Understanding how to encourage assembly can help engineer effective viral vectors that are used as transport for gene therapy or potentially for bacteriophage virus therapy to attack specific bacteria.
Scope. In this paper, we are interested only in structures formed by direct, autonomous assembly rather than structures assembled with the aid of extraneous chaperones or scaffolding molecules that do not end up as part of the assembled structure. Furthermore, we are not interested in structures formed by multistage assembly; i.e, by various deformation and/or folding processes subsequent to the assembly of an initial structure.

1.1 Limitations of Experimental Data and Modeling Motivation

Supramolecular assembly is a rapid, economical process driven by weak interactions and non-covalent binding between the constituent molecular components. The assembly takes place spontaneously at room temperature, in solution, or in a lipid bilayer membrane. Available types of experimental data on supramolecular assembly include:

- X-ray crystallography for details of relatively large assembled structures (often possessing nontrivial symmetries, as in the case of icosahedral viruses);
- cryo-electron microscopy and stochiometry studies of various approximate subassembly intermediate structures and their sizes;
- primary sequence or even NMR spectroscopy structure of the starting monomers and smaller (sub)assemblies;
- calorimetric studies to determine dissociation energies for (sub)assemblies;
- in vitro systems to measure concentrations of various subassembly intermediates;
- selective mutagenesis of starting monomers, and its effect in encouraging or disrupting assembly; and
- mining a comparable database of all of the above types of data for assembly systems classified by various similarity criteria, for example, by structural or biological similarity of viruses.

Despite the above types of experimental data and exploration capabilities, supramolecular assembly processes are poorly understood partly because of their remarkable rapidity, spontaneity and robustness. Spontaneity makes it difficult to control in vitro, rapidity makes it difficult to get snapshots, and robustness (multiple pathways and insensitivity to individual interactions of the constituent molecules) makes it difficult to isolate *crucial combinations of assembly-driving interactions* - from among a combinatorial explosion of possible combinations. In addition, many of these ex-

perimental methods are labor and resource-intensive, making blind alleys extremely expensive.

This generates a strong motivation to go beyond guesswork guided by theoretical first principles alone, and develop effective mathematical and computational models for supramolecular assembly that can inform further experimentation. On the other hand, the necessity to validate model predictions using the available experimental data and within the prevailing experimental capabilities - frames the scope of our models, and defines their inputs, outputs and tuning parameters.

1.2 Prediction Tasks

Based on the previous discussion, we focus on models for the following types of prediction tasks.

- *Input:* the 3D configurations of the rigid components of the starting monomers, and the inter-component interactions (Section 2 describes how they are formally specified).*Output:* prediction of the terminal assembly structures and their concentrations (or probabilities).
- *Input:* as in the previous item, plus a 3D configuration of final assembly. *Output:* prediction of those atoms or monomers that are crucial for the assembly process to terminate in the given input assembly configuration.
- *Input:* as in the previous item. *Output:* prediction of minimal atomic alterations that would significantly increase probability of the assembly process terminating in the given input assembly configuration.
- *Input:* as in the previous item, additionally more than one choice of final assembly configuration. *Output:* prediction of key events such as specific intermediate subassembly configuration choices during assembly that determine which one of the final assembly configuration results.

These types of predictions cannot be made by theoretical first principles, combinatorial experimentation (trying various possibilities), and guesswork alone, even with the help of known data on similar assemblies and biological knowledge about evolutionarily conserved structures. In addition, for larger assemblies, these predictions cannot be made by direct application of standard methods such as Monte Carlo or Molecular Dynamics mixed with informatics style approaches for mining existing knowledge for similar assemblies.

1.3 The Methods of This Paper

The methods emphasized in this paper begin with isolating and abstracting crucial factors influencing assembly, thus motivating a *multiscale* model of assembly.

One such factor influencing supramolecular assembly at the *nanoscale* is *configurational entropy* of small assemblies at inter-monomeric interfaces, driven by weak forces and non-covalent binding. The exact computation of configurational considered a notoriously difficult problem in chemical theory and computational chemistry.

This paper describes our new modeling paradigm towards the judicious approximation of configurational entropy suited to a specific type of prediction that can be validated by mutagenesis experiments. The paradigm consists of two aspects. The first aspect is the generation of an *atlas* of the configuration space using classical Thom-Whitney stratification from algebraic-geometry. The second aspect is our new theory of *convexification*, i.e, choosing parameters by which the regions of the stratification can be represented as convex regions. Both aspects are implemented as a prototype software EASAL (*efficient atlasing and search of assembly landscape*). Recent mutagenesis validation of predictions of crucial interactions for the assembly of AAV2 (Adeno Associated Virus) were based on an approximation of interface configurational entropy obtained by EASAL.

Another crucial factor influencing assembly at the *microscale* is *combinatorial entropy* in the formation of larger assemblies from smaller subassembly intermediates, especially when symmetries are present. This, too, is difficult to model or compute. Traditional approaches have been primarily based on simplified geometric approximations of the assembly constituents, and local assembly rules, together with statistical mechanics simulation heuristics that incorporate kinetics as well. Most of these methods do not provide performance guarantees, nor facilitate backward analysis of the computational model's input-output function; nor are they suited to providing intuitive, mechanistic explanations and predictions.

This paper details our approach for combinatorial entropy at the microscale, using algorithms with performance guarantees (or even generating functions), for counting assembly pathways with desired features, especially in the presence of symmetry.

1.4 Organization

In Section 2 we discuss factors influencing assembly, motivating a multiscale model of assembly. In Section 3 we discuss the crucial *nanoscale* factor that influences supramolecular assembly namely *interface configurational entropy* at inter-monomeric interfaces. We give a brief sketch of the literature tracing the long and distinguished history of the notoriously difficult problem of configurational entropy computation. We then describe our modeling paradigm for approximating interface configurational entropy: generation of an atlas of the configuration space using stratified convexification, implemented as a prototype software EASAL. In Section 4 we discuss the crucial *microscale* factor of that *combinatorial entropy*, which influences the number of pathways to formation of larger assemblies from smaller subassembly intermediates, especially when symmetries are present. We briefly survey tradi-

tional approaches based on local assembly rules and statistical mechanics simulation heuristics that incorporate kinetics. We describe our approach for computing combinatorial entropy at the microscale, using algorithms (or even generating functions), for counting assembly pathways. In Section 5 we briefly present recent mutagenesis validation of predictions of crucial interactions for the assembly of AAV2 (Adeno Associated Virus), based on an approximation of interface configurational entropy obtained by EASAL. We conclude by highlighting the remaining challenges in Section 6.

2 Multiscale Model based on Factors Influencing Assembly

First we describe an assembly system, i.e, the typical input to an assembly process. This is followed by a discussion of the key factors that influence assembly that highlights the challenges of the above prediction tasks and motivates a multiscale assembly model.

2.1 Assembly System

An input to a computational model of an assembly process is an *assembly system* consisting of the following.

- A collection of of *monomers* drawn from a small set of *monomeric types* (often just a single type). Each monomeric type is specified as a collection of *rigid molecular components*; a rigid component is in turn specified as the set of positions of the centers of their constituent *atoms*, in a local coordinate system. In many cases, an *atom* could be the representation for the average position of a *collection of atoms in an amino acid residue*. Note that an assembly *configuration* is given by the positions and orientations of the entire set of n rigid molecular components in an assembly system, relative to one fixed component. Since each rigid molecular component has 6 degrees of freedom, a configuration is a point in $6(n-1)$ dimensional Euclidean space.
- The pairwise component of the potential energy function of the assembly system, specified as a sum of potential energy (also called enthalpy) terms between pairs of constituent atoms i and j in two different rigid components of the assembly system. The weak interactions between the rigid molecular components is captured by this potential energy function. The pairwise potential energy terms are, in turn, specified using pairwise *Lennard-Jones* and *Hard-Sphere pairwise potential energy functions*. The pairwise Lennard-Jones term is typically present only for selected pairs of atoms, i and j, one from each component, while the Hard-Sphere potentials apply to all other pairs. Both are functions of the distance $d_{i,j}$ between i and j; The former function is typically discretized to take different constant values on 3 intervals for the distance value

$d_{i,j}$: $(0, l_{i,j}), (l_{i,j}, u_{i,j}), and (u_{i,j}, \infty)$. Typically, $l_{i,j}$ is the so-called Van der Waal or steric distance given by "forbidden" regions around atoms i and j. And $u_{i,j}$ is a distance where the attractive (electrostatic or other weak) forces between the two atoms is no longer strong (typically these forces decay as the reciprocal of some power of $d_{i,j}$). Intuitively, the interval $(0, l_{i,j})$ is where the repulsive force dominates, and $(l_{i,j}, u_{i,j})$ is where the attractive force and repulsive forces are balanced, and $(u_{i,j}, \infty)$ is where neither force is strong. Over these 3 intervals respectively, the Lennard-Jones potential assumes a very high value $h_{i,j}$, a small value $s_{i,j}$, and a medium value $m_{i,j}$. All of these *bounds* for the intervals for $d_{i,j}$, as well as the values for the Lennard-Jones potential on these intervals are *specified constants* as part of the input to the assembly model. These constants are specified for each pair of atoms i and j, i.e., the subscripts are necessary. The middle interval is called the *well*. The Hard-Sphere potentials are defined solely by the Van der Waal's forbidden distance, $l_{i,j} = u_{i,j}$.

- A non-pairwise component of the potential energy function in the form of *global potential energy* terms that capture the tethers between the rigid components within a monomer, as well as other global potential energy terms that implicitly represent the solvent (water or lipid bilayer membrane) effect [23, 24, 16]. These are specified using discrete values over intervals of the distances or angles between pairs of entire rigid components (as opposed to pairs of atoms).

It is important to note that all the above potential energy terms are *functions of the assembly configuration*.

Observe that an assembly system can alternatively be represented as a set of rigid molecular components drawn from a small set of types, together with *assembly constraints*, in the form of distance and angle intervals. These constraints define *feasible* configurations (where the pairwise inter-atoms distances are larger than $l_{i,j}$, and any relevant tether and implicit solvent constraints are satisfied). The set of feasible configurations is called the *assembly configuration space*. The *active constraint* regions of the configuration space are regions where at least one of the Lennard-Jones inter-atom distances lies in the well, i.e, the interval $(l_{i,j}, u_{i,j})$.

Note that for the prediction tasks given above, the input to the assembly model assembly consists of an assembly system, optionally along with one or more final assembly configurations.

2.2 Factors, Challenges, Multiscale

The assembly configuration space can be partitioned into regions with constant potential energy. The *free energy* of such a region is related to the probability of finding the assembly system in a configuration in the region and is dependent on both its potential energy (inversely) and on the log volume of the region (directly). The former is constant over the region, as defined and is easy to compute for our model. Roughly speaking, the latter represents the *configurational entropy* of the region.

We refer the reader to [18, 43] for a succinct exposition of the relationship between these properties.

2.2.1 Nanoscale: Interface Configurational Entropy

At equilibrium, the configuration space (complex of constant potential energy regions) partitions into potential energy *basins* representing *equilibrium* configurations.

The potential energy computation for these configurations is immediate. and the challenge is to to compute the configurational entropy, i.e., volumes of these basins to determine the *stability or binding affinity* for these equilibrium configurations.

The dimension as well as geometric and topological complexity of a potential energy basin corresponding to an equilibrium assembly configuration make the computation of the basin volume challenging. If the volume is determined by sampling, it takes time *exponential* in the dimension, and each rigid component in the assembly system punishingly adds 6 to this dimension. Already for small, *interface assembly systems* that are associated with specific types of interfaces between rigid molecular components, this *interface configurational entropy* computation, at the *nanoscale* is thus highly challenging.

2.2.2 Microscale: Combinatorial Entropy

For larger, microscale assemblies, this type of direct configurational entropy computation is impossible. Instead, they are treated as being recursively assembled as an interface assembly system, from small number of stable intermediate subassemblies [36]. This recursive assembly is usually represented as an *assembly tree* whose leaves are the rigid molecular components of the assembly system, the root is the final large assembly configuration, and the internal nodes are the intermediate subassembly configurations. The overall entropy of a configuration space region of the large assembly C is a combination of:

- the entropies of its small number of constituent equilibrium subassemblies C_i;
- the *interface configurational entropy* of the assembly of the C_i's to form C;
- the *combinatorial entropy* at the *microscale* that arises from the number of different collections of subassemblies C_i that assemble to form C, heavily influenced by the symmetries of C [38, 26, 7]; and finally
- the *microscale kinetics* that interrelate the stability and binding affinity of different interface assembly configurations, with the concentrations of the constituent subassembly configurations.

The potential energy basins corresponding to equilibrium configurations of the large assembly system C as well as their stability and binding affinity are again determined by the geometry and topology of the initial partition into constant potential energy regions as well as *microscale kinetics*.

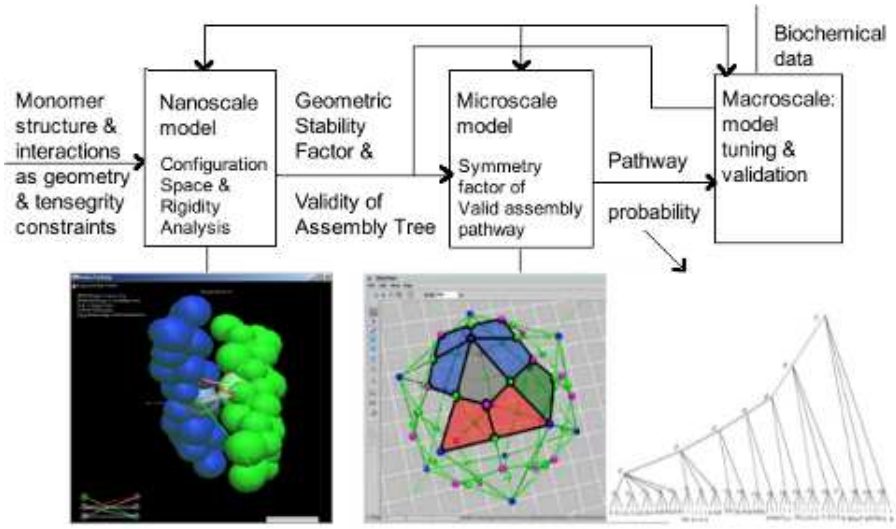

Fig. 1 Multiscale assembly model scales shown; Left:combinatorial entropy using nanoscale interface assembly system of 2 rigid molecular components with pair potentials; Mid: Large T=1 viral, microscale assembly shown as polyhedron; Right: whose combinatorial entropy is given using (recursive) Assembly Trees

The above discussion isolates the factors influencing assembly as: *potential energy*; *interface configurational entropy and nanoscale kinetics*; *combinatorial entropy and microscale kinetics*. This motivates a 3-scale model for assembly (see Figure 1).

3 Nanoscale Models: Interface Configurational Entropy

We discuss 3 types of models that attempt to capture the following related properties of interface configuration space regions for small assemblies: free energy, partition function (relative probability), stability, binding affinity, configurational entropy. We refer the reader again to [18, 43] for understanding the exact relationship between these properties.

3.1 Stability based on Extent of Rigidity

In [36], rigidity was roughly equated with being an equilibrium assembly configuration (i.e, low energy representative configuration of a potential energy basin) and a further shortcut was used to quantify stability of an equilibrium assembly configuration, namely, the number of Lennard-Jones pairs that had to be removed in order to

degenerate into a flexible configuration with many small rigid subassemblies. This shortcut also been suggested by Ileana Streinu in a personal communication. However, as mentioned in the previous section, even for small assemblies, the bottleneck in computing the stability and binding affinity of equilibrium configurations is the computation of the volume of the high dimensional potential energy basin corresponding to the equilibrium configuration, possessing a complicated geometry and topology. This rigidity-based approximation of the volume is too coarse to be effective, as demonstrated for example in trying to determine crucial interactions for AAV2 assembly as in Section 5, for which a different method had to be used. In particular, for that example, even a straightforward PCA or Eigenvalue based method outperformed the rigidity based method.

3.2 Traditional Methods for Configurational Entropy and Free Energy

There has been a long and distinguished history of configurational entropy and free energy computation methods [18, 2, 14, 15, 13, 19, 12, 33, 20], many of which use as input the configuration trajectories of Molecular Dynamics or Monte Carlo simulations.

All configurational entropy computations reduce to computing cartesian volumes of constant potential energy regions of a configuration space, as mentioned earlier [18, 43]. Even methods that directly compute *partition integrals* (i.e, probabilities of a configuration being in a region of the configuration space) or directly compute free energy (e.g. the Mining Minima method [12]) must effectively compute volumes of configuration space regions since free energy gradients (binding affinities) are effectively based on *entropy differences* between the configuration space regions that correspond to ”before” and ”after” assembly. Again, the picture of configurational entropy as volume computation for configuration space regions clarifies the intrinsic nature of the two mutually compounding challenges that *any* method will have to overcome: dimensionality and topological/geometric complexity.

As mentioned earlier, accurate computation of volumes of configuration space regions cannot escape exponential dependence on dimension as long as the computation is achieved by counting samples explicitly. Sampling is often the only way to compute volume of constant potential energy regions, since they are typically semi-algebraic sets (i.e., sets of configurations satisfying systems of quadratic inequalities, since distance is a quadratic function of the cartesian configuration). Such semi-algebraic sets have high geometric and topological complexity, going beyond just the inherent nonlinearity, even in relatively low dimensional scenarios. In addition, during sampling, Jacobian computations are necessary to map from the ”free” internal coordinates to constant potential energy regions of the cartesian configuration space. Such Jacobian computations are necessary since since Lennard-Jones and Hard sphere pair potentials are both dependent on inter-atom distances, which depend quadratically (not linearly) on the cartesian coordinates of a configuration.

For instance, with 2 rigid molecular components, the dimension of the cartesian configuration space is just 6. However, when each component has tens of atoms, the active constraint regions induced by any standard potential landscape are complexes of nested boundaries of different (effective) dimensions. It is due to this reason that one cannot guarantee ergodicity of Monte Carlo sampling nor give any reasonable bounds on the number of rejected samples. For both Monte Carlo and Molecular Dynamics, uniform sampling can only be claimed in the limit, or "if run for sufficiently long, or starting from sufficiently many initial configurations." This also causes problems for many entropy computation methods that rely on principal component analyses of the covariance matrices from a trajectory of samples in internal coordinates, followed by a quasiharmonic [2] or nonparametric (such as nearest-neighbor-based) [14] estimates. Such methods generally *overestimate* the volumes of configuration space regions with high geometric or topological complexity, even when hybridized with higher order mutual information [15], and nonlinear kernel methods, such as the Minimally Coupled Subspace approach of [13]. Ab initio methods such as [33] based on geometric algebras (Lie algebra, Grassman-Cayley algebra etc. common in robotics) are used to give bounds or to approximate configurational entropy without relying on Monte Carlo or Molecular Dynamics sampling. However, it is not clear how to extend them beyond restricted assembly systems such as a chain or loop of rigid molecular components each consisting of at most 3 atoms, where each component is noncovalently bound to the each neighboring component at exactly 2 sites.

There has been some research on infering the topology of the configuration space [11, 39, 22, 30] starting from Monte Carlo and Molecular Dynamics samples, and using the topology to guide dimensionality reduction, [42].

3.3 Approximations of Configurational Entropy via Atlas of Configuration space

As mentioned earlier, geometric constraints on inter-atom distances and angles can be extracted from our potential energy function. A recent paper by the author introduces the notion of an *atlas* of a configuration space, which consists of two ingredients. The first ingredient is a *stratification* of the configuration space into *active constraint regions* (next subsection). The second ingredient is a representation of each active constraint region by carefully chosen parameters that make the region convex (following subsection).

3.3.1 Stratification, Active Constraint Regions

Consider an assembly configuration space $\mathscr{A}$ of k rigid components, defined by a system A of assembly constraints. The configuration space has dimension $m \leq 6(k-1)$, the number of internal degrees of freedom of the configurations since a rigid

object in Euclidean 3-space has 6 rotational and translational degrees of freedom. For $k = 2$, m is at most 6 and in the presence of two tether constraints, it is at most 4.

A *Thom-Whitney stratification* of the configuration space $\mathscr{A}$ (see Figure 2) is a partition of the space into regions grouped into strata X_i of $\mathscr{A}$ that form a filtration $\emptyset \subset X_0 \subset X_1 \subset \ldots \subset X_m = \mathscr{A}$, $m = 6(k-1)$. Each X_i is a union of nonempty closed *active constraint regions* R_Q where $m-i$ inequality constraints $Q \subseteq A$ are *a*ctive, meaning equality is attained and they are independent. Each active constraint set Q is itself part of at least one, and possibly many, hence l-indexed, nested chains of the form $\emptyset \subset Q_0^l \subset Q_1^l \subset \ldots \subset Q_{m-i}^l = Q \subset \ldots \subset Q_m^l$. These induce corresponding reverse nested chains of active constraint regions $R_{Q_j^l}$: $\emptyset \subset R_{Q_m^l} \subset R_{Q_{m-1}^l} \subset \ldots \subset R_{Q_{m-i}^l} = R_Q \subset \ldots \subset R_{Q_0^l}$ Note that here for all l, j, $R_{Q_{m-j}^l} \subseteq X_j$ is closed and j dimensional.

We represent the active constraint system for a region, by an *active constraint graph* whose vertices represent the participating atoms (at least 3 in each rigid component) and edges representing the active constraints between them. Between a pair of rigid components, there are only a small number of possible active constraint graph isomorphism types since there are at most 12 contact vertices.

There could be regions of the stratification of dimension j whose number of active constraints exceeds $6(k-1)-j$, i.e. the active constraint system is overconstrained, or whose active constraints are not all independent. Dependent constraints diminish the set of realizations. For entropy calculations, these regions should be tracked explicitly, but in the present paper, we do not consider these special regions in the stratification. Our regions are obtained by choosing any $6(k-1)-j$ independent active constraints.

3.3.2 Convex Representation of Active Constraint Region and Atlas

A new theory of Convex Cayley Configuration Spaces **(CCCS)** recently developed by the author [37] gives a clean characterization of active constraint graphs whose configuration spaces are convex when represented by a specific choice of so-called *Cayley parameters* i.e., distance parameters between pairs of atoms that are inactive in the given active constraint region (see Figure 3). Such active constraint regions are said to be *convexifiable*, and the corresponding Cayley parameters are said to be its *convexifying* parameters.

The *Atlas* of an assembly configuration space is a stratification of the configuration space into convexifiable regions. In [27], we have shown that *molecular assembly configuration spaces with 2 rigid molecular components have an atlas.* The software EASAL (Efficient Atlasing and Search of Assembly Landscapes) efficiently finds the stratification, incorporates provably efficient algorithms to choose the Cayley parameters [37] that convexify an active constraint region, efficiently computes bounds for the parametrized convex regions [8], and converts the parametrized configurations into standard cartesian configurations [29].

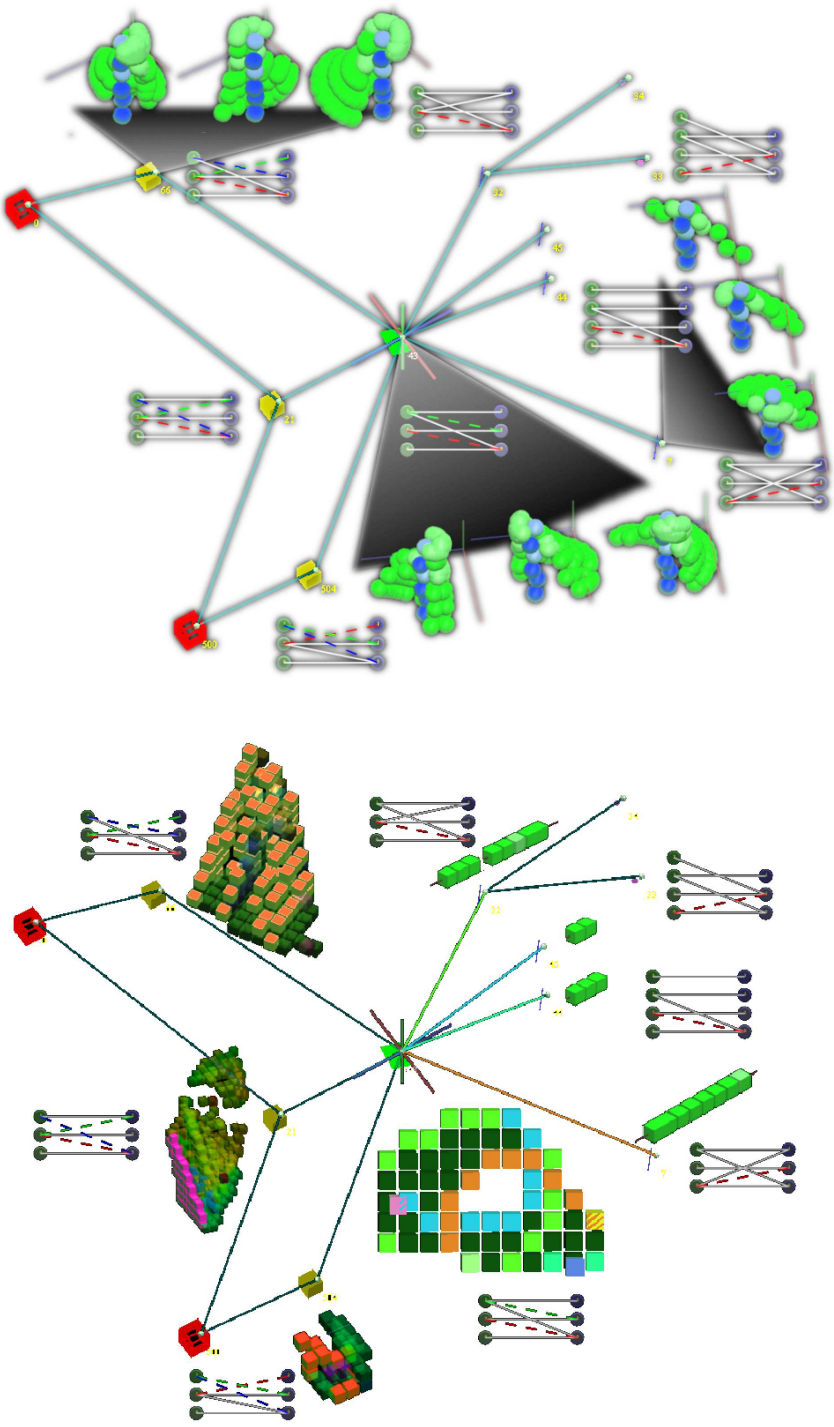

Fig. 2 Top: atlas portion, with active constraint regions labeled by their active constraint graphs (dark edges); the regions are shown as sweeps around a stationary reference molecule. Bottom: active constraint regions with convexifying Cayley parameters (light edges), which decrease with dimension, as edges are added to the active constraint graph; note intersection with the complement of a convex subregion in the center. Edges are successively added to the active constraint graphs for the child and descendant atlas regions as more constraints become active.

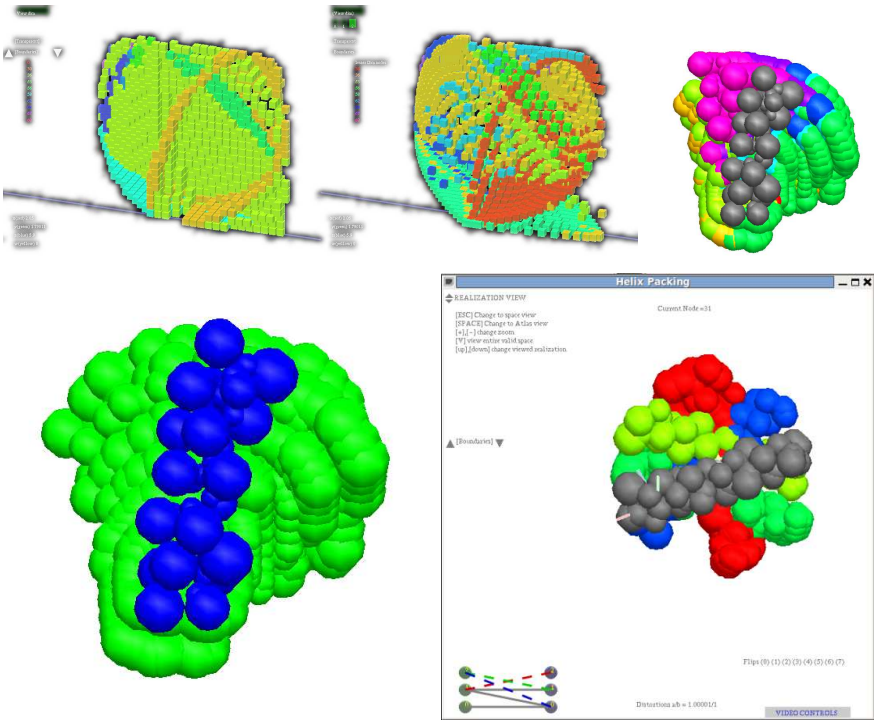

Fig. 3 Top Left: atlas region showing interiors and boundaries sampled in its convexifying Cayley parameters; boundary/child regions sampled in their own Cayley parameters and mapped back to the parent region's Cayley parameters (*note increase in samples*). Top Right: boundary/child regions sampled in their own Cayley parameters shown as sweeps around grey reference (toy) helix. Bottom Left: union of boundary regions sampled in parent's Cayley parameters, shown as sweep around blue reference helix (*notice (b) is bigger*) Bottom Right: sweep of one of the boundary regions sampled in parent's Cayley parameters is shown in red around gray reference helix; the sampling *misses the other colored configurations* in the same boundary region, obtained by sampling in its own Cayley parameters.

The key point is that EASAL is *tailored for assembly and leverages its unique properties*; in particular, even simple folding configuration spaces, (e.g., the classic cycloheptane or cyclooctane) do not have atlases.

3.3.3 EASAL-based Approximations of Configurational Entropy

There are many natural ways to approximate configurational entropy. Their efficacy depends on the particular application where they are used. We give one example here that we used for determining crucial constraints as in Section 5. The potential energy basins of an interface assembly system are centered around the configurations in the zero-dimensional active constraint regions of the configuration space atlas. These regions cannot be found by EASAL without finding the higher dimensional regions of the atlas. Furthermore, each distinct configuration in such a region is rigid and could be considered an equilibrium assembly configuration with its own potential

energy basin. Any configuration in a basin satisfies at least $6(n-1)$ of the input constraints (for n rigid molecular components), i.e, the corresponding inter-atomic distances fall within their respective Lennard-Jones wells. The number of copies of one of the configurations in a basin is the number of higher dimensional regions of the atlas whose active constraint graphs are subgraphs of the the active constraint graph of the given configuration. This is an approximate measure of the size or volume of a potential energy basin (configurational entropy associated with that basin).

4 Microscale Model: Combinatorial Entropy

As mentioned in Section 2, the computation of combinatorial entropy requires both (a) a count of assembly trees (defined in Section 2) (see Figure 4) weighted by the combined probability of their constituent stable subassemblies (in turn obtained from their free energies); and (b) microscale kinetics as described in Section 2.

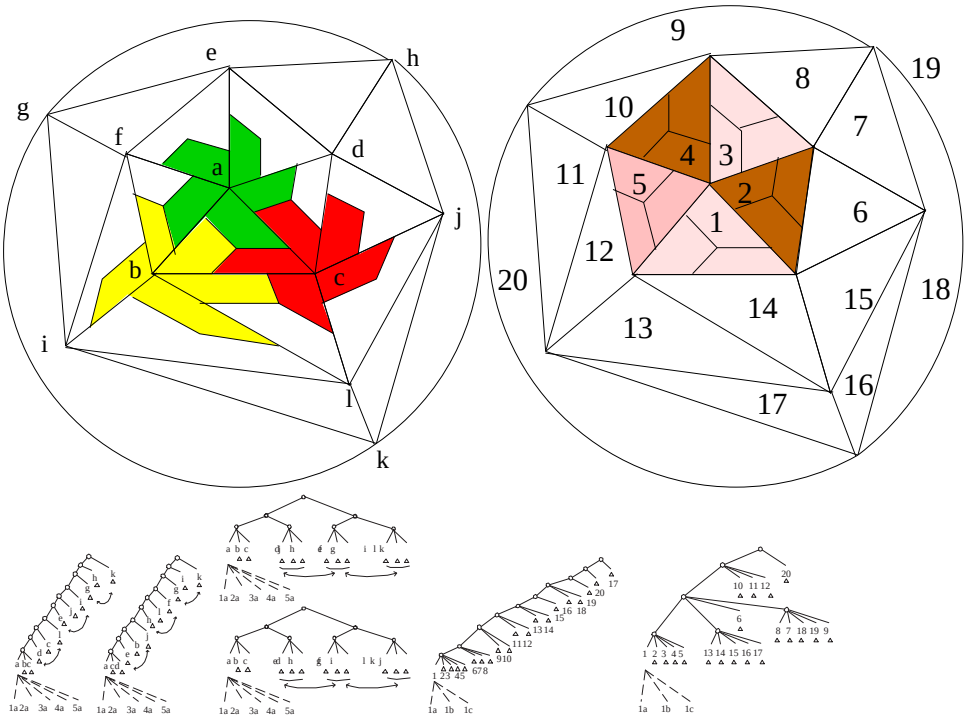

Fig. 4 Top: icosahedral assemblies shown as 12 pentamers, or 20 trimers; Bottom: Assembly Trees based on (left) sequential addition of pentamers, and starting with a trimer of pentamers, with bottom level triangles representing pentamers, or (right) sequential addition of trimers starting with a pentamer, with bottom level triangles representing trimers.

4.1 Combinatorial Entropy via Simplified Assembly Components and Local Rules

The assembly model [35] combines both (a) and (b) above, based on the "local-rules" theory of [4, 5, 6, 34]. In addition, differentiation of these models from other similar [32, 44, 25, 31, 17] are given in [35]. The [35] model relies crucially on the following. (i) Full-blown dynamic simulation (their approach has no static analogy for analyzing successful assembly trees alone); (ii) Simple polygonal representations of monomers an explicitly specified set of stable configurations for the subassemblies; (iii) Simplified geometric interactions between monomeric types explicitly and procedurally specified as local rules. The above type of assembly model provided just the necessary level of detail to answer the kinds of questions about concentrations of subassembly configurations. However, forward and reverse analysis are difficult as are intuitive explanations as to what sets of local rules and stable subassemblies are likely to result in a given final assembly.

4.2 Combinatorial Entropy via Assembly Trees and Orbits (Pathways)

To compute the weighted sum of assembly trees for an assembly configuration A as given in (a) above, it is useful to analyze orbits of assembly trees under the action of automorphism group G of A, or the polyhedral graph corresponding to A. Similarly it is useful to analyze orbits of subassemblies A' under the action of G. The induced action of G on an assembly tree is obtained by the G acting on each subassembly occuring as a node in the tree. Two trees in the same orbit under this induced action represent the same assembly process (properties such as the combined probabilities of its constituent subassemblies - are the same), hence we call each orbit under this action an *assembly pathway*. Similarly the orbit of a subassembly A' under G is called an *assembly type.* More generally, if any finite group G acts on a finite set S, there is an induced action of G on the set of assembly trees for S. Even if all assembly trees have equal probability of occuring, not all assembly pathways have equal probability of occuring, since the corresponding orbits have different sizes depending on their stabilizer subgroup in G. In [38], we formulated various questions about probabilities of pathways with various properties. In [26, 7] we answered some of these questions; specifically, in BoSiVi, we gave explicit generating functions for counting all pathways with a given orbit size.

5 Validation of EASAL Prediction of AAV2 Crucial Interactions

The results in this section have appeared in [41]. We started from simplified potential energy landscapes designed from known X-ray structure of AAV2 coat protein monomers and interfaces [1, 10, 28] (data provided by Mavis Agbandje-Mckenna's lab, see Figure 2). For each of the 3 interfaces (2-fold, 3-fold and 5-fold), we determined the pairs of interacting atoms that are conserved in related viruses (10-20 pairs for each interface). These were used as the candidate interactions for the crucial interactions. For the mutagenesis experiment in Mckenna's lab, these candidate interactions were disabled one by one, by mutating one of the atoms in the pair. The effect of the mutation on assembly efficacy was determined by measuring concentration of successfully assembled viral shells via cryo-electron microscopy. *This experiment [3] took at least 2 years.*

For EASAL's predictions, we treated monomers as single rigid components in the interface assembly systems. We used Lennard-Jones potentials for the above pairs of interacting atoms and hard spheres for the sterics of the remaining atoms. No solvent effects were considered. For each interface, for each of its interactions, the approximate interface configurational entropy was computed as described in Section 3, when the specific interaction was dropped. We called this the *sensitivity* of that interaction. In fact, for each of the interfaces we generated a new atlas and computed the above quantity for more than one assembly system obtained from different pairs of participating multimers - see below for a detailed description. The rationale was that the same interface drives assembly of different types of multimer-pairs during the formation of larger intermediate subassemblies. We obtained a cumulative sensitivity ranking for each interaction, over all of the relevant interface assembly systems for that interaction. *This computation took 1 week.*

The tabulated results for Dimer and Pentamer interfaces are given in the two Tables 1. The atom numbers in the first two columns are standard numbering used in the cited papers. In some cases, Atom1 interacts with more than one partner Atom2. Mutagenesis disables all interactions in which a mutated atom participates. Atom 1 and Atom 2 give the residue In both cases, the highest ranked interactions (the corresponding atom pair names are given) output by EASAL indicate that assembly is most sensitive to these interactions. They were validated by mutagenesis, resulting in assembly disruption (the "Confirmed" column). Note that blank entries in the "Confirmed" column indicate that mutagenesis was not performed to disable those interactions, i.e, it is as yet unknown whether EASAL's predictions are correct.

Pentamer interface with participating Multimers

During the formation of larger assembly intermediates two multimers (as opposed to monomers) could assemble across the same interface. We obtained a new pentamer interface atlas for a monomer and a dimer. While the weak-force interactions remain

Table 1 Sensitivity ranking: Dimer (top), Pentamer(bottom) Interfaces

atom1	atom2	Confirmed
P293	W694, P696	Yes[3]
R294	E689, E697	Yes[3, 40]
E689	R298	Yes[3]
W694	P293, Y397	Yes[3]
P696	P293	Yes[3]
Y720	W694	Yes[3]

atom1	atom2	Confirmed
N227	Q401	Yes[40]
R389	Y704	
K706	N382	
M402	Q677	Yes[3]
K706	N382	
N334	T337,Q319	
S292	F397	Yes[40]

the same, the number of hard-sphere sterics increases and changes the interface configuration space significantly. Factoring this into the rankings, we found two other crucial interactions for the pentamer interface: S292-F397 and N227-Q401. Both of them were confirmed by assembly disruption through mutagenesis, and have been included in the above tables.

Note concerning the Trimer interface: We could not obtain useful sensitivity rankings for the trimer interface due to heavy influence of sterics caused by interdigitation). This tallied with the fact that mutagenesis of the any of the trimer interface interactions could not disrupt assembly. We do not believe that assembly of the AAV2 shell is sensitive to any of the trimer interactions. We conjecture that the assembly proceeds primarily by dimeric and pentameric interface interactions. Trimers interdigitate and contribute to stability of the capsid after the assembly is complete.

6 Conclusions and Open Questions

We defined the scope of assembly models based on the type of experimental data available for validation. We gave factors influencing assembly, and motivated a multiscale model. We surveyed traditional models and new modes for both the nanoscale and the microscale and highlighted the issues that are still outstanding. We then gave an example of model prediction that could be experimentally validated.

Open Questions on Configurational Entropy

At the moment the exact computation of configurational entropy, i.e., volumes of atlas regions is done by sampling, which, as mentioned, does not escape the exponential time dependence on dimension. However, for convexified regions, faster methods for example based on [9] may help with volume computation for atlas regions. Another unresolved issue is that kinetics influence the stucture of equilibrium potential energy basins, which we have not taken into account.

Open Questions on Combinatorial Entropy

The generating function in [7] for counting pathways with the same orbit size does not extend to pathways with a given property, not even those whose intermediate subassemblies are stable. Sacrificing the generating function for an algorithm opens the field to matroid basis-exchange type algorithms, provided stable subassemblies can be defined appropriately. We gave a randomized counting algorithm [36] based on matroid basis exchange, for counting all assembly trees with stable subassemblies. However what is needed is to count pathways (i.e, tree orbits) with stable subassemblies. Furthermore, the question is open how to combine microscale kinetics with the above type of orbit counting.

References

1. M. Agbandje-McKenna, A.L. Llamas-Saiz, F. Wang, P. Tattersall, and M.G. Rossmann. Functional implications of the structure of the murine parvovirus, minute virus of mice. *Structure*, 6:1369–1381, 1998.
2. I. Andricioaei and M. Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *The Journal of Chemical Physics*, 115(14):6289, 2001.
3. A. Bennett. N/a. Unpublished manuscript, N/A 2012.
4. B. Berger, P. Shor, J. King, D. Muir, R. Schwartz, and L. Tucker-Kellogg. Local rule-based theory of virus shell assembly. *Proc. Natl. Acad. Sci. USA*, 91:7732–7736, 1994.
5. B. Berger and P.W. Shor. On the mathematics of virus shell assembly. 1994.
6. B. Berger and P.W. Shor. Local rules switching mechanism for viral shell geometry. *Technical report, MIT-LCS-TM-527*, 1995.
7. M. Bóna, M. Sitharam, and A. Vince. Tree orbits under the action of the icosahedral group and enumeration of macromolecular assembly pathways. *Bull. Math. Bio, Special Issue on Algebraic Biology, to appear*, 2011.
8. U. Chittamuru. Sampling configuration space of partial 2-trees in 3d. 2011.
9. M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM*, 38:1–17, January 1991.
10. E. Padron, V. Bowman, N. Kaludov, L. Govindasamy, H. Levy, P. Nick, R. McKenna, N. Muzyczka, J. A. Chiorini, T. S. Baker, and M. Agbandje-McKenna. Structure of adeno-associated virus type 4. *Journal of Virology*, 79:5047–58, 2005.
11. D. Gfeller, D. Morton D. Lachapelle, P. De Los Rios, G. Caldarelli, and F. Rao. Uncovering the topology of configuration space networks. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 76(2 Pt 2):026113, 2007.

12. M.S. Head, J.A. Given, and M.K. Gilson. Mining minima, Direct computation of conformational free energy. *The Journal of Physical Chemistry A*, 101(8):1609–1618, 1997.
13. U. Hensen, O.F Lange, and H. Grubmller. Estimating absolute configurational entropies of macromolecules: The minimally coupled subspace approach. *PLoS ONE*, 5(2):8, 2010.
14. V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *Journal of Computational Chemistry*, 28(3):655–668, 2007.
15. V. Hnizdo, J. Tan, B.J. Killian, and M.K. Gilson. Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *Journal of Computational Chemistry*, 29(10):1605–1614, 2008.
16. W. Im, M. Feig, and C.L. Brooks. An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. *Biophysical Journal*, 85(5):2900–2918, 2003.
17. J.E. Johnson and J.A. Speir. Quasi-equivalent viruses: a paradigm for protein assemblies. *J. Mol. Biol.*, 269:665–675, 1997.
18. M. Karplus and J.N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.
19. B.J. Killian, J. Yundenfreund Kravitz, and M.K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *The Journal of chemical physics*, 127(2):024107, 2007.
20. B.M King, N.W. Silver, and B. Tidor. Efficient calculation of molecular configurational entropies using an information theoretic approximation. *The Journal of Physical Chemistry B*, 0(ja):null, 0.
21. T-C. Kuo. On Thom-Whitney stratification theory. *Mathematische Annalen*, 234:97–107, 1978. 10.1007/BF01420960.
22. Z. Lai, J. Su, W. Chen, and C. Wang. Uncovering the properties of energy-weighted conformation space networks with a hydrophobic-hydrophilic model. *International Journal of Molecular Sciences*, 10(4):1808–1823, 2009.
23. T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins*, 35(2):133–152, 1999.
24. T. Lazaridis. Effective energy function for proteins in lipid membranes. *Proteins*, 52(2):176–192, 2003.
25. C.J. Marzec and L.A. Day. Pattern formation in icosahedral virus capsids: the papova viruses and nudaurelia capensis β virus. *Biophys*, 65:2559–2577, 1993.
26. M. Bóna and M. Sitharam. Influence of symmetry on probabilities of icosahedral viral assembly pathways. *Computational and Mathematical Methods in Medicine: Special issue on Mathematical Virology, Stockley and Twarock Eds*, 2008.
27. A. Ozkan and M. Sitharam. Easal: Efficient atlasing and search of assembly landscapes. In *Proceedings of BiCoB*, 2011.
28. E. Padron, R. McKenna, N. Muzyczka, N. Kaludov, J. A. Chiorini, and M. Agbandje-McKenna. Structurally mapping the diverse phenotype of adeno associatedvirus serotype 4. *Journal of Virology*, 80:11556–570, 2006.
29. J. Peters, J. Fan, M. Sitharam, and Y.Zhou. Elimination in generically rigid 3d geometric constraint systems. In *Proceedings of Algebraic Geometry and Geometric Modeling*, pages 27–29, Nice, September 2004. Springer Verlag,1-16,2005.
30. D. Prada-Gracia, J. Gmez-Gardees, P. Echenique, and F. Falo. Exploring the free energy landscape: From dynamics to networks and back. *PLoS Comput Biol*, 5(6):e1000415, 06 2009.
31. D Rapaport, J Johnson, and J Skolnick. Supramolecular self-assembly: molecular dynamics modeling of polyhedral shell formation. *Comp Physics Comm*, 1998.
32. V S Reddy, H A Giesing, R T Morton, A Kumar, C B Post, C L Brooks, and J E Johnson. Energetics of quasiequivalence: computational analysis of protein-protein interactions in icosahedral viruses. *Biophys*, 74:546–558, 1998.

33. G.S. Chirikjian. Chapter four - modeling loop entropy. In Michael L. Johnson and Ludwig Brand, editors, *Computer Methods, Part C*, volume 487 of *Methods in Enzymology*, pages 99 – 132. Academic Press, 2011.
34. R Schwartz, PE Prevelige, and B Berger. Local rules modeling of nucleation-limited virus capsid assembly. *Technical report, MIT-LCS-TM-584*, 1998.
35. R Schwartz, PW Shor, PE Prevelige, and B Berger. Local rules simulation of the kinetics of virus capsid self-assembly. *Biophysical journal*, 75:2626–2636, 1998.
36. M Sitharam and M Agbandje-Mckenna. Sampling virus assembly pathway: Avoiding dynamics. *Journal of Computational Biology*, 13(6), 2006.
37. M. Sitharam and H. Gao. Characterizing graphs with convex cayley configuration spaces. *Discrete and Computational Geometry*, 2010.
38. M Sitharam and M. Bóna. Combinatorial enumeration of macromolecular assembly pathways. In *Proceedings of the International Conferecnce on bioinformatics and applications*. World Scientific, 2004.
39. G Varadhan, Y J Kim, S Krishnan, and D Manocha. Topology preserving approximation of free configuration space. *Robotics*, (May):3041–3048, 2006.
40. P Wu, W Xiao, T Conlon, J Hughes, M Agbandje-McKenna, T Ferkol, T Flotte, and N Muzyczka. Mutational analysis of the adeno-associated virus type 2 (AAV2) capsid gene and construction of AAV2 vectors with altered tropism. *Journal of virology*, 74(18):8635–47, September 2000.
41. R. Wu, A. Ozkan, A. Bennett, M. Agbandje-Mckenna, and M. Sitharam. Robustness measure for aav2 is correctly predicted by configuration space atlasing using easal. In *Proceedings of ACM Bioinformatics and Comptutational Biology, Orlando*, 2012.
42. Y. Yao, J. Sun, X. Huang, G.R. Bowman, G. Singh, M. Lesnick, L.J. Guibas, V.S Pande, and G. Carlsson. Topological methods for exploring low-density states in biomolecular folding pathways. *The Journal of chemical physics*, 130(14):144115, 2009.
43. Huan-Xiang Zhou and Michael K Gilson. Theory of free energy and entropy in noncovalent binding. *Chemical Reviews*, 109(9):4092–107, 2009.
44. A Zlotnick. To build a virus capsid: an equilibrium model of the self assembly of polyhedral protein complexes. *J. Mol. Biol.*, 241:59–67, 1994.