Thesis Proposal: Designing an information-driven approach for targeted colloidal self-assembly

Shannon Moran

February, 2018

Broad thesis: Information can be encoded and retrieved from colloidal materials to guide their behavior.

Committee chair: Prof. Sharon Glotzer
Committee members:
Prof. Ronald Larson
Prof. Robert Ziff
Prof. Xiaoming Mao (Physics)

Note: Total length must be less than 15 pages of text. Includes figures, excludes title page, list of references, and CV.

Contents

List of Figures

Preliminary Exam: Project Summary

Paragraph 1: Motivation

Paragraph 2: Where there are openings in the conversation about the role of information in self-assembly

Paragraph 3: 2 sentences each on the projects proposed

Paragraph 4: Concluding thoughts on future work

1 Introduction and motivation - 1 pg

When we think about the major challenges facing materials science, we are fundamentally faced with this idea of inversely designing materials. That is, I decide that I want to create a material that behaves your sweat-wicking shirt under one condition, stiffens under another, and when given a particular stimulus can reconfigure its structure. Currently, if I wanted to make a material like that for you, I'd naively take materials that have each of those properties and figure out how I could get them to work together. Or, I'd look for novel materials that have properties close to those of the material I want to make. We would call this designing the material.

This is inefficient. In the inverse design problem, we take the properties we want and create the materials that will give us those properties. Machine learning and materials science are coming together on the active front of this research. However, being able to predict, even perfectly, what can be made from existing materials by definition limits us to the set of materials that currently exist. This is a known challenge in materials science– how do we probabilistically explore phase space outside of phase space where we have data? While intriguing in its own right, that is not the topic of the thesis proposed here.

Instead, we might think about this from a fundamental physics point of view. If we want to make complex materials that have embedded stimuli responses, or assemble into a specific target structure, we must give the building blocks of such complex materials some amount of direction. We can think about this amount of direction as an amount of *information*.

This is not to say that we are looking to have building blocks act as storage devices, as in [1]. In that work, each building block is a cluster of multiple particles in whose arrangement can be stored a "high density" of information.

Similarly, in a recent proposal between our group and those of Marke Bathe (MIT), Mawgwi Bawendi (MIT), and Oleg Gang (Columbia), we proposed a biosynthetic, high-density storage structure composed of DNA nanocubes (Figure **??**). Within these nanocubes, information could be stored in the different dies intercolated into the frame of the cube, into quantum dots placed into the frames, or even in the shape of the frames themselves. If we them move a level higher, we can imagine storing additional information in the order of these nanocubes relative to one another.

However, using these and solutions like them for high-density storage requires us being able to write, read, and store information into these formats. Fundamentally, these three challenges are predicated upon the ability to specifically place blocks where they need to go ("write"). Current methods include sonic and laser tweezers (manual), specific DNA interactions (energetic), or incremental addition (kinetic). How do we compare between these methods, though?

Here, I propose that the ability of building blocks to form a target structure can be distilled down into a concept of *information*.

This is not a new concept. In our group, we are comfortable with the concept that a target structure is the result of a minimization of free energy. In systems devoid of inter-particle forces, this then reduces down to a maximization of entropy.

Statistical mechanical "entropy" shares its name with information "entropy" in communications theory. While this directly came about because of the form similarity between the two, much energy since has been devoted to developing frameworks connecting the two. Jaynes,

in the 1950s, spent two long articles trying to reconcile the two. Books, and multiple articles, have been dedicated to explaining why these concepts are similar.

While much time has been spent developing the theory, very little time has been spent directly leveraging this concept for embedding information in systems governed by statistical mechanical ensembles– such as colloidal-scale self-assembly.

Key line from Simons proposal: "A coherent framework of thermodynamic and non-equilibrium processes seen through information theoretic eyes could lead to new theories for encoding information in matter– which would allow for the design of novel materials and novel material behavioral control."

1.1 Prior thoughts on this matter

Goal: Predictively and reliably encode information in materials systems. (I know this is too broad)

Problem statement:

- *Engineering*: Designing functional, reconfigurable materials will require some method of storing information in a material– we might call this "memory"
- *Science*: Understanding how to "store information" in a material that embeds a response is a fundamental problem; further understanding the linkage between information-theoretic entropy and thermodynamic entropy concepts (accessible states, heat) could further our understand of dynamics of biological, etc processes

Ideas:

- Reconfigurable systems may be the key for unlocking adaptive material applications
- At a particle level, we can think of this as building pluripotent building blocks that contain some response to a stimulus
- At a system level, we can think of this as having metastable configurations in response to some stimulus
- At an assembly-level, we can think of this as building blocks that can be engineered to form specifically-ordered arrays

2 Literature Background - 2-3pg

2.1 What is information?

Let's first look at information in the context of communications theory.

The *information*, I, we get from an event happening is given by:

$$I(p) = -\log(p)_b \tag{1}$$

where p is the probability of an event happening and b is the base. Base 2 is commonly used in information theory, and forms the unit of information. For instance, the unit of base 2 information is a bit, base 3 are trits, base 10 are Harleys, and base e are nats.

In 195X, Claude Shannon also extended this concept by introducing the concept of *information entropy*. In this context, entropy is the average (expected) amount of information gained from a given event. Specifically, for an event with n different outcomes this can be written as:

$$\text{Entropy} = \sum_{i=1}^{n} p_i \log p_i \tag{2}$$

For a discrete random variable X with $p(x)$, the entropy can be written as:

$$H(X) = \sum_{x} p(x) \log p(x) \tag{3}$$

Entropy does not range from 0 to 1. The range is set based on the number of possible outcomes n, i.e. $- \leq \text{Entropy} \leq \log(n)$. Entropy is equal to 0 (minimum entropy) when one of the probabilities is 1 and the rest are 0's. Entropy is $\log(n)$ (maximum entropy) when all the probailities have equal values of 1/n.

In the case of designing specific outcomes for an event, then, we want to minimize the entropy along each leg of the pathway leading to an event. Put another way, we want to maximize the probability that the event will proceed down the pathway we want it to.

However, the concept of "information" in this context is then counter-intuitive. Information in communication refers to how likely an event is. When a rare event happens, we gain more "information" from that event. However, in the context of designing specific outcomes, we are not looking to read out bits of information once an event has happened. We are looking to design the likelihood of an event occurring.

In the words of MIT professor Cèsar Hidalgo, "It is hard for us humans to separate information from meaning because we cannot help interpreting messages." We face the same problem here– by saying that a pathway has more information than another, we are implicitly saying that it is a rarer event than a lower-information pathway.

Counter-intuitively, in designing pathways for self-assembly, then, we are looking to design minimum-information pathways. **However, in aligning with our intuition from self-assembly, this means we are looking to maximize the entropy of an assembly pathway.**

However, we can use the concept of *mutual information* in defining how much information is stored in an interaction in an intuitive manner. (See notes on the Brenner paper below.) Mutual information $I(X;Y)$ is a global measure of interaction specificity in systems with many distinct species. It quantifies how predictive the identity of a lock x_i is to the identity of key y_i found bound to it.

2.1.1 How does this tie into statistical mechanics?

2.2 What is information, in the context of self-assembly?

Let's first look at a paper from the Brenner group, the "Information capacity of specific interactions' [2]. Their main thesis is that specific binding interactions have energetics that

allow binding to occur with measurable probability. Thus, we can measure the relative information in different types of binding. This is more in line with the communication theory view of information (rare events giving more information) than it is with the materials view of information, in which high information events imply high probability of a desired event happening.

Our group, and many others in the materials community, are looking to engineering materials to control their structures, behaviors, etc. A common method of engineering these materials is by tailoring the interactions between their components through chemistry, shape, etc. By understanding how much *assembly information* can be contained in these interactions, we can:

1. Compare the efficacy of different types of interactions in delivering desired behavior(s)
2. Theoretically predict the efficacy of new types of interactions

Let's take the example of a lock and key system. ADD SECTION ON BRENNER PAPER FROM LIT REVIEW LAST YEAR

Any of Jacobs' papers that talk about this?: uses connectivity graphs

2.3 Where have folks applied directed self-assembly? Why do we care about it?

Glotzer, Kotov - self-assembly

Mirkin - experimental, DNA-directed

Kamien - kirigami

Glotzer, Desmaine - folding

Frenkel, Jacobs - pathway design

Wales - pathway designs, disconnectivity graphs

2.4 Already proposed work: Semiconductor Synthetic Biology

Include work done on the NSF grant proposal: Using DNA-mediated assembly to store information in nanoparticle arrays.

Specifically, this is an example of 1b): addressable complexity, then trying to engineer how to get the particles to where they should go in the most energetic and information/complexity-efficient manner possible.

2.5 Where the conversation is in the literature

Information as a measure of the likelihood of a particular configuration being preferred. In 2015, a review article in *Nature Physics* (which has since been cited 255 times) reviewed the state of the art on applying information theoretic entropy– i.e. Shannon entropy– as a way of understanding non-equilibrium thermodynamics [3]. In this work, they

investigate information entropy as a placeholder for non-equilibrium entropy production. This entropy production gives an overall likelihood of a configuration (one which minimizes the non-equilibrium free energy of a system while maximizing the non-equilibrium entropy). However, this method applies to the overall structure, or the overall likelihood of a structure being the preferred structure.

"Addressable complexity" seeks to engineer pathways for particular particles to reach their destination. Low free energies of a target structure, however, do not guarantee efficient assembly. There are a number of ways addressing this problem in the literature. One way of forcing systems into assembly is to design a free energy landscape that minimizes such meta-stable traps [4]. Competition between degenerate structures of equivalent potential energy was reported for clusters of six attractive spherical colloids, where symmetry breaking leads to higher rotational entropy of the less symmetric conformation, resulting in lower free energy [5].

Taken from [6]: A well-known example of addressable complexity– that is, specific binding– can be found in "one-pot" DNA self-assembly of DNA tiles, which use the hybridization of complementary DNA sequences to construct complex structures consisting of hundreds of subunits from a single soup of monomers [7] (5). Simulation results have shown that such one-pot self-assembly can succeed with highly simplified model subunits that lack the molecular details f DNA tiles, suggesting that similar design strategies should be widely applicable [8] (6). In the work by Jacobs *et al*, they had particles with designed interactions between one another. They represented the target bonds by a graph, G. However, this model is based on the assumption that "designed interactions in the target structure are typically much stronger than any incidental associations between sub-units that should not be connected in the final assembly". This is a fine assumption for their proof of concept, but is not valid in real-world system. As a concrete example, protein-folding is perhaps the most well-explored biological system that assembles due to specific interactions [9]. However, one of the major challenges to solving the protein-folding problem are competing "cross-talk" interactions [CITATIONS NEEDED; chaperoned folding and assembly Chakrabarty 2017].

In later work, Jacobs et al addressed this oversight and accounted for incidental interactions in addition to designed interactions [10].

Low energies may not guarantee efficient assembly compartmentalized, multi-stage assembly grannemana and baserga 2004

Talk outline: 1. self-assembly kinetics can be rationally designed– leverage thermodynamics 2. evolution has already selected for optimal assembly pathways in complex biomolecules

Specific binding interactions can be tailored to lead to target structures. However, there is a delicate balance of specificity required. On the over-specified side, we have bonds that are specific to their intended neighbor with probability 1.0. On the under-specified side, we have non-specific interaction patches that will bond to any other patch with probability $1/n$, where n is the total number of patches in the system.

In work from our group, Eric Jankowski sought to generate energy-minimizing configurations for such patchy particles [11] in a process he called "bottom-up building block analysis", or BUBBA. Cluster Monto Carlo (cMC) and LAcMC methods are relatively poor methods for finding potential energy minima formed from patchy particles with disparate interaction energies due to their tendency to become trapped in metastable configurations as well as the low degeneracy of potential energy-minimizing configurations (Q). BUBBA effectively

searches a subset of the configuration space for energy-minimizing configurations. Jankowski predicted that that BUBBA would be useful for evaluating many different particles for self-assembly "propensity".

Partition functions encode all the thermodynamics of a system, but for most systems of practical importance they cannot cannot be calculated exactly. This is due to many indistinguishable degenerate states. In the cases where small numbers of distinguishable configurations comprise a majority of a partition functions' weight, as is the case for systems at low temperatures and for many anisotropic building blocks with disparate interactions, BUBBA is a particulalry effective method for generating partition functions that have been heretofore inaccessible. This allows us to ask "What structures are thermodynamically favored for this building block at any temperature?" to be answered independently of assembly kinetics. [12]

Both thermodynamic and kinetic barriers to assembling target structures.

Key problem, taken from [13]: Self-assembly holds promise for creating new materials and devices because of its inherent parallelism, allowing many building blocks to simultaneously organize using preprogrammed interactions. An important trend in nanoparticle and colloid science is the synthesis of particles with unusual shapes and/or directional (??patchy??) interactions, whose anisotropy allows, in principle, assemblies of unprecedented complexity. However, patchy particles are more prone to long relaxation times during thermodynamically driven assembly, and there is no a priori way of predicting which particles might be good assembly candidates. Here we demonstrate a new conceptual approach to predict this information using sequences of intermediate clusters that appear during assembly. **Unfortunately, when an equilibrium solution or simulation of patchy particles fails to generate an ordered pattern it is not always obvious whether the culprit is thermodynamics or kinetics.** Recently there have been studies that attempt to quantify kinetic trapping through fluctuation-dissipation ratios (21,22) and through the interplay between specific and nonspecific interactions (3,5,23) but these methods do not provide predictive capabilities for thermodynamically stable structures. The fact that both thermodynamics and kinetics can prevent a system of particles from self-assembling is particularly trou- blesome for experimentalists that search parameter space via trial-and-error because experiments that fail to assemble do not provide information about how assembly might be improved.

We are ultimately searching for rational design of building blocks optimized for self-assembly that focuses on assembly pathway engineering: identifying the traps that occur as a system assembles so they may be circumvented. As systems self-assemble we hypothesize that the thermodynamically stable intermediate clusters that arise hold information about their ability to order. These sequences of intermediate clusters are assembly pathways and we propose a methodical analysis of them to predict the degree to which a system of building blocks will assemble a target pattern, which we refer to as the building block?s assembly propensity for the pattern. We foresee assembly pathway engineering proceeding as a collaboration among structural identification, kinetic measurements, and the assembly pathway analysis described here. [13]

2.6 Motivation 1: DNA assembly, DNA tiles, DNA cubes

Taken from intro of [8]:
The observation by Ke et al. [Science 338, 1177 (2012)] that large numbers of short, predesigned DNA strands can assemble into three-dimensional target structures came as a great surprise, as no colloidal self-assembling system has ever achieved the same degree of complexity. That failure seemed easy to rationalize: the larger the number of distinct building blocks, the higher the expected error rate for self-assembly. The experiments of Ke et al. have disproved this argument. Here, we report Monte Carlo simulations of the self-assembly of a DNA brick cube, comprising approximately 1000 types of DNA strand, using a simple model. We model the DNA strands as lattice tetrahedra with attractive patches, the interaction strengths of which are computed using a standard thermodynamic model. We find that, within a narrow temperature window, the target structure assembles with high probability. Our simulations suggest that misassembly is disfavored because of a slow nucleation step. As our model incorporates no aspect of DNA other than its binding properties, these simulations suggest that, with proper design of the building blocks, other systems, such as colloids, may also assemble into truly complex structures.

Need to include Winfree, Rutherford papers in here.

Include work from NSF proposal.

2.7 Motivation 2: Protein folding

Steal references.

3 Description of proposed research (7-8 pg, 2-3 pg per aim)

"Information" are those factors that impact the yield and kinetics of self-assembly (thermodynamics of the free energy landscape and the kinetics of the path to get to a target structure from a given starting point).

Specifically, as we look to both understand how nature governs self-assembly into target structures, we need a language to understand this. Nature is very good at already picking an optimized route through a free energy landscape [14].

3.1 What I want to accomplish in my thesis

1) Define a measure of pathway information.

- We already have ways of measuring how good a particular bond is
- Are there particular bonds/connections that are the most important to get correct to enable forming the desired final structure?
-

2) Use that measure of pathway information to design ideal pre-cursors for target structures.

3) Attempt to use machine learning to predict ideal pre-cursors for given target structures.

- [15]: Nonlinear Machine Learning of Patchy Colloid Self-Assembly Pathways and Mechanisms out of the Furguson group

4) Why is it important we find the "most important" pathway points? from [16]
How then can self-folding origami be folded with a minimal number of actuators? A lesson can be drawn from similar glassy landscape search problems in models of protein folding (e.g., Levinthal?s paradox [17, 19, 20, 41]) and related NP-hard satisfiability (SAT) problems [21, 42] that vary from the Traveling Salesman Problem to Sudoku [43]. A common element in these satisfiability problems is that random seeding of the search for the global minimum leads to repeated backtracking after reaching local minima, both in the context of computer algorithms (as the DPLL algorithm for k-SAT [21]) or for physical dynamics (as in protein folding) [42]. However, careful seeding of the search - e.g., if the right boxes are filled in first in Sudoku [43] or if the right parts of the protein are folded first - can greatly reduce or even eliminate backtracking [21] before reaching the global minimum. Correct seeding is even more critical for origami since folding is assumed to happen at ?zero temperature? (e.g., without any noise or fluctuations). As a result, the structure cannot backtrack out of a local minimum as in the case of non-zero temperature SAT problems [42].

This reference also has a really good introduction section relating origami and self-assembly [16].

3.2 Paper 1, Role of particle shape on the emergent behavior of active systems: 2 pg

2-3 pages of work already complete

Relation to proposed thesis topic: If we define information broadly as any quality of a building block that impacts the emergent behavior of a system of those particles (in this case, force direction and shape), then we can argue this work is looking at a few aspects of information in active systems.

3.3 Paper 2 - Defining information as it pertains to colloidal systems - 2.5 pg

Need to figure out what angle I want to approach this from:

- Pathway engineering?
- "Smart"particle building blocks?
- Algorithmically-designed interactions/patterned structures?
- ... something else?

For each aim:

- Goal and significance - paper introduction-esque, why is this worked needed in the conversation (citations)
- Hypothesis
- Approach, methods, analysis to be used (including relevant citations) - methods, how I'll go about this and what methods I'll need to develop

For this one, need to understand:

- When I say "come up with a metric for information", what does this mean?
-

3.4 Paper 3 - Applying that definition of information to nets (folding systems) - 2.5 pg

Need to figure out what angle I want to approach this from:

- Pathway engineering?
- "Smart"particle building blocks?
- Algorithmically-designed interactions/patterned structures?
- ... something else?

For each aim:

- Goal and significance - paper introduction-esque, why is this worked needed in the conversation (citations)
- Hypothesis
- Approach, methods, analysis to be used (including relevant citations) - methods, how I'll go about this and what methods I'll need to develop

3.5 Paper 4 - Using machine learning to design information-rich starting structures: 1 pg

4 Time table

See Figure 1 for key tasks and milestones through 2020, based on the projects outlined in the above sections.

5 Conclusions and potential impact

Long, long-term goal: Instead of simply observing emergent behavior as an outcome of collective motion of individuals, we could instead engineer such behavior as a quantifiable outcome of the interaction of an information-rich network of agents.

			2017				2018				2019				2020	
Project	**Status**	**Description**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**
Active shapes	100%	Data collection														
		Analysis														
		Writing														
		Submit for publication														
Defining information		Object 1														
		Object 2														
		Object 3														
		Object 4														
Applying information		Object 1														
		Object 2														
		Object 3														
		Object 4														
Machine learning		Object 1														
		Object 2														
		Object 3														
		Object 4														
Thesis		Data meeting														
		Defense														

Figure 1: Key milestones and tasks from Preliminary Exam through target defense date.

References

[1] Carolyn L. Phillips, Eric Jankowski, Bhaskar Jyoti Krishnatreya, Kazem V. Edmond, Stefano Sacanna, David G Grier, David J. Pine, and Sharon C. Gltozer. Digital colloids: reconfigurable clusters as high information density elements. *Soft Matter*, 10:7468–7479, 2014.

[2] Miriam H. Huntley, Arvind Murugan, and Michael P. Brenner. Information capacity of specific interactions. *Proceedings of the National Academy of Sciences*, 113(21):5841–5846, May 2016.

[3] Juan M. R. Parrondo, Jordan M. Horowitz, and Takahiro Sagawa. Thermodynamics of information. *Nature Physics*, 11:131–139, 2015.

[4] David J. Wales. Atomic clusters with addressable complexity. *The Journal of Chemical Physics*, 146(054306), 2017.

[5] Guangnan Meng, Natalie Arkus, Michael P. Brenner, and Vinothan N. Manoharan. The free-energy landscape of clusters of attractive hard spheres. *Science*, 327:560–563, 2010.

[6] William M. Jacobs, Aleks Reinhardt, and Daan Frenkel. Communication: Theoretical prediction of free-energy landscapes for complex self-assembly. *The Journal of Chemical Physics*, 142(021101), 2015.

[7] Yonggang Ke, Luvena L. Ong, William M. Shih, and Peng Yin. Three-dimensional structures self-assembled from dna bricks. *Science*, 338(6111):1177–1183, 2012.

[8] Aleks Reinhardt and Daan Frenkel. Numerical evidence for nucleated self-assembly of dna brick structures. *Physical Review Letters*, 112(23), Jun 2014.

[9] Ken A. Dill. Folding proteins: finding a needle in a haystack. *Current Opinion in Structural Biology*, 3:99–103, 1993.

[10] William M. Jacobs, Aleks Reinhardt, and Daan Frenkel. Rational design of self-assembly pathways for complex multicomponent structures. *Proceedings of the National Academy of Sciences*, 112(20):6313–6318, May 2015.

[11] Eric Jankowski and Sharon C. Glotzer. A comparison of new methods for generating energy-minimizing configurations of patchy particles. *The Journal of Chemical Physics*, 131(104104), 2009.

[12] Eric Jankowski and Sharon C. Glotzer. Calculation of partition functions for the self-assembly of patch particles. *The Journal of Physical Chemistry B*, 115:14321–14326, 2011.

[13] Eric Jankowski and Sharon C. Glotzer. Screning and designing patchy particles for optimized self-assembly propensity through assembly pathway engineering. *Soft Matter*, 8(2852), 2012.

[14] William M. Jacobs and Eugene I. Shakhnovich. Structure-based prediction of protein-folding transition paths. *Biophysical Journal*, 111:925–936, 2016.

[15] Andrew W. Long and Andrew L. Ferguson. Nonlinear machine learning of patchy colloid self-assembly pathways and mechanisms. *The Journal of Physical Chemistry B*, 118:4228–4244, 2014.

[16] Menachem Stern, Matthew Pinson, and Arvind Murugan. The difficulty of folding self-folding origami. *arXiv*, 1703.04161v1, 2017.