THE UNIVERSITY OF CHICAGO

DATA-DRIVEN DESIGN OF SELF-ASSEMBLING SOFT MATERIALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE PRITZKER SCHOOL OF MOLECULAR ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
YUTAO MA

CHICAGO, ILLINOIS
AUGUST 2022

To my family, for their invaluable supports.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to first thank my research advisor, Prof. Andrew Ferguson, for his help and guidance through my research works. He has always been very patient in explaining technical concepts that I am not familiar with, and we have always had very productive discussions on the technical challenges I have encountered in my research. Besides technical aspects, he has also helped me a lot in improving my scientific communication skills. I still could remember how struggling I was when I was making my first presentation slides for the group meeting and how nervous I was when I gave that presentation, and Prof. Ferguson has been very nice and patient in offering me suggestions on how to improve my presentation skills. I benefited a lot from all these advices and my presentation skills have improved significantly. In addition, his suggestions on revising the drafts of the papers have been invaluable in helping me improving my scientific writing skills. He is truly an incredible and great mentor, and I would not be able to get where I am today without his guidance. I sincerely believe that the skills which I have developed under his guidance would greatly benefit me throughout my future career.

I would also like to thank my thesis committee members, Prof. Andrew Ferguson, Prof. Paul Nealey and Prof. Juan de Pablo, for their excellent suggestions on my thesis and their time and efforts on the thesis defense process.

Moreover, I would like to thank the alumni and current members of the Ferguson Lab for their tremendous help: Dr. Hythem Sidky, Dr. Brandon Peters, Dr. Andrew Long, Dr. Gregory Hart, Prof. Jiang Wang, Prof. Rachael Mansbach, Dr. Bryce Thurston, Dr. Wei Chen, Dr. Mingfei Zhao, Dr. Siva Dasetty, Kirill Shmilovich, Xinran Lian, Nicholas Rego, Shiqi Chen, Walter Alvarado, Max Topel, Mike Jones, Yifeng Tang, Nicholas Herringer, Kate Johnson, Nikša Praljak, Joseph Aulicino, Melody Leung and Rohan Kapoor. Special thanks to Dr. Andrew Long for his helpful suggestions on the colloidal self-assembly projects and to Dr. Brandon Peters, Dr. Mingfei Zhao and Dr. Siva Dasetty for their helpful advices on the peptide self-assembly work.

Furthermore, I would like to thank my friends in UIUC and UChicago. A lot of the time it could be stressful in graduate school, and chatting or playing video/computer games with my friends have been a good way to relax myself.

Last but very importantly, I would like to offer my thanks to my parents. Although being thousands of miles apart, they provide me with invaluable encouragement and supports through my pursuit of the Ph.D. degree. Their strong supports have helped me get through some of the most difficult periods in graduate school, and I truly owe a lot to them.

Examining retrospectively, I believe that every stage during the past five years in graduate school, no matter whether it is the success of an approach or failure of a concept, has been an invaluable treasure in my life. I would like to sincerely thank everyone who helps me along the way.

ABSTRACT

Self-assembly refers to a process in which initially disordered systems spontaneously form ordered structures over time driven by the interactions between the building blocks. The self-assembly of soft materials systems, such as colloids and peptides, has drawn a lot of research interests. When their components and inter-molecular interactions are carefully calibrated, these systems could self-assemble into interesting nanostructures with practical applications, such as colloidal crystalline lattices or peptide nanorods. A major area of study in the self-assembly is then the rational design of the building blocks and interactions between them to direct these systems to self-assemble into target nanostructures. This is the so-called "inverse design" problem, where we are given a target nanostructure and would like to figure out the optimal building block and interactions that could lead to the target structure. In our works, we explore and develop various inverse design techniques for anisotropic colloids and peptides. We begin by employing molecular simulation techniques, such as molecular dynamics and Monte Carlo simulations, to characterize the ability of a particular building block design to form the target structure. Then we use modern optimization and machine learning algorithms to find the optimal design that could lead to maximum ability of forming the target structure. Based on this general methodology, we have developed (1) inverse design protocols that optimize anisotropic colloids to self-assemble into target crystalline lattices with omnidirectional optical bandgaps by combining molecular simulation with stochastic optimization algorithm and (2) high-throughput screening pipeline to find optimal peptides that could self-assemble into vesicular structures as chassis materials for synthetic cells using molecular simulation and Bayesian optimization.

CHAPTER 1

INTRODUCTION

1.1 Overview of the Self-assembly of Soft Matter Systems

The self-assembly of soft matter systems is a popular research area since it provides a way for the "bottom-up" fabrication of nanostructures that are difficult to fabricate in a "top-down" manner. For example, an interesting class of nanostructures are "photonic crystals" which are composed of dielectric materials arranged in a periodic lattice whose spatial periodicity roughly matches the wavelength of visible light[2]. Traditional top-down methods to fabricate photonic crystals include holographic lithography[3] and direct laser writing[4] that essentially drill the bulk dielectric materials using laser to periodically modulate the dielectric constant. These top-down approaches are expensive, tedious and have trouble with making 3D photonic crystals[5]. However, because the sizes of colloids are naturally on the scale of desired wavelength, the self-assembly of colloids into periodic arrays provides a more straightforward and cheap bottom-up approach to fabricate photonic crystals[5]. Similarly, the self-assembly of synthetic and natural polymers provides a bottom-up approach to synthesize monolayers as water-repellent coating[6] or micelles and vesicles as drug carriers[7]. Due to the versatility of soft materials, their chemical and physical properties could be easily modified. This gives researchers extensive flexibility in the control of the self-assembly behaviors of soft matter systems, as these building blocks could be modified in a controllable way to guide their self-assembly toward specific directions. For example, recent advances in experimental techniques have enabled modifications of the surfaces of colloidal particles, such as grafting DNA oligomers[8–11] or attaching interactive patches by glancing angle deposition[12–14] and contact layer lithography[15]. These surface-modified colloidal particles have proven their abilities to spontaneously form various nanoscale colloidal crystalline structures[8,12]. Peptides are another important building block in self-assembly. By changing the constituent amino acids, they are able to self-assemble into various structures such as nanosheets[16], nanospheres[17], nanorods[18,19] and nanotubes[20]. In particular, the nanorods formed by A_9K peptide could potentially serve as antibiotic by breaking bacterial membrane[19]. A specific class of peptides that has drawn a lot of research focuses is elastin-like polypeptide (ELP), which are synthetic biopolymers that share a lot of structural characteristics with intrinsically disordered proteins such as tropoelastin. In recent experiments, some specially designed ELPs have been shown to self-assemble into large and stable vesicles[21,22] that may sustain more osmotic stress than conventional lipid vesicles and become ideal candidates as drug carriers[23].

A specific research focus in the self-assembly of soft matter systems is the inverse design of the building blocks. Given a target structure, the inverse design strategy tries to find the optimal building block that could self-assemble into the target structure. With the help of modern machine learning and optimization algorithms, this approach could accelerate the design of materials relative to Edisonian trial-and-improvement and avoid traps associated with flawed intuition. The general approach is to first quantify the ability of building block to form the target structure and then use machine learning and optimization algorithms to find the optimal design of building blocks that maximizes this ability. For example, some efforts have been made to optimize the isotropic interactions between spherical colloids, such as the functional form of potential or parameters in the potential, so that the colloidal particles interacting through the optimal potential could self-assemble into various target 2D or 3D crystalline structures[24–28]. Many researches have also been conducted to perform inverse design of peptides[29], such as using deep representational learning and Bayesian optimization to identify optimal amino acid sequence capable of assembling 1D nanoaggregates with good stacking of the electronically active π-cores[30].

Two key components in the inverse design are the quantitative measurement of the “goodness” of a specific building block for desired self-assembly behavior and the optimization algorithm to maximize the “goodness” with respect to the building block design parameters.

The former is usually done by running molecular simulation to simulate the self-assembly behavior of the building blocks and then computing a thermodynamic or kinetic objective function that quantitatively characterizes the quality of the self-assembly behavior from the simulation trajectory. Since the self-assembly behavior is directly affected by the building block design parameters, a numerical optimization algorithm could be employed to optimize the objective function with respect to those parameters. If the functional form of the objective function is not available (usually referred as a "black-box" optimization problem), genetic optimization algorithms could be employed, or a machine learning algorithm could first be used to build a predictive model for the objective function and numerical optimization could be performed over this proxy model. In the next two sections, we describe the molecular simulation techniques, machine learning methods and optimization algorithms that are usually used in inverse design strategies.

1.2 Molecular Simulation Techniques

In general, molecular simulation uses numerical algorithms to generate microscopic realizations of a system governed by a particular Hamiltonian. These snapshots of configurations can then be used to evaluate the thermodynamic and kinetic properties of the system. Generally speaking, molecular simulation techniques can be divided into two categories: molecular dynamics and Monte Carlo simulations.

In a molecular dynamics simulation, the system configurations, which are completely described by the positions $\vec{\boldsymbol{r}}$ and momenta $\vec{\boldsymbol{p}}$ of particles, are generated by integrating the Hamiltonian equation of motion forward in time:

$$\begin{aligned}\dot{\vec{\boldsymbol{r}}} &= \frac{\partial \mathcal{H}}{\partial \vec{\boldsymbol{p}}} \\ \dot{\vec{\boldsymbol{p}}} &= -\frac{\partial \mathcal{H}}{\partial \vec{\boldsymbol{r}}}\end{aligned} \tag{1.1}$$

Various numerical methods exist to integrate equation 1.1 numerically forward in time[31]. For example, a well-known numerical algorithm is the Verlet algorithm[32], which is based on the following scheme:

$$
\begin{aligned}
\vec{\boldsymbol{r}}(t+\Delta t) &= \vec{\boldsymbol{r}}(t) + \frac{\vec{\boldsymbol{p}}(t)}{m}\Delta t + \frac{1}{2}\vec{\boldsymbol{a}}(t)\Delta t^2 \\
\vec{\boldsymbol{p}}(t+\Delta t) &= \frac{\vec{\boldsymbol{p}}(t)}{m} + \frac{\vec{\boldsymbol{a}}(t)+\vec{\boldsymbol{a}}(t+\Delta t)}{2}\Delta t
\end{aligned}
\tag{1.2}
$$

It could be shown that the Verlet algorithm is time-reversible and volume-preserving, so it approximately conserves energy in the long term. After integrating the equation of motion for T steps, the desired property of the system could be evaluated as a time-average over simulation trajectory:

$$
\langle A \rangle = \frac{1}{T}\sum_{t=1}^{T} A(\vec{\boldsymbol{r}}(t), \vec{\boldsymbol{p}}(t)) \tag{1.3}
$$

The Verlet algorithm essentially describes a system that follows Hamiltonian equation of motion, which conserves the number of particles N, the total energy E and the volume V. Therefore, it simulates the system in a NVE ensemble. However, in many practical situations such as membrane simulation, the system is better modeled to be under constant pressure and temperature (NPT ensemble) or constant volume and temperature (NVT ensemble). There exist various extensions to the basic Verlet algorithm to simulate system in other ensembles. For example, the Nose-Hoover thermostat[33] is an algorithm that is based on an extended Hamiltonian:

$$
\mathcal{H}_{\text{Nose}} = \sum_i \frac{\boldsymbol{p}_i^2}{2m_i s^2} + U(\boldsymbol{r}^N) + \frac{p_s^2}{2Q} + L\frac{\ln(s)}{\beta} \tag{1.4}
$$

where s and p_s are the generalized coordinate and momentum of the extra degree of freedom corresponding to the thermal reservoir and $\beta = \frac{1}{k_B T}$ is the inverse temperature. Integrating the equation of motion derived from the extended Hamiltonian would generate configurations $\{\boldsymbol{r}^N(t), \boldsymbol{p}^N(t)\}$ of the original system according to NVT ensemble. The Berendsen barostat[34] aims to keep the system pressure at a target pressure P_{target} by scaling the particle position and simulation box length at each step by a factor $\eta^{\frac{1}{3}}$

$$\eta^{\frac{1}{3}} = \left(1 - \frac{\gamma \Delta t}{\tau_p}(P_{\text{target}} - P(t))\right)^{\frac{1}{3}} \tag{1.5}$$

where γ is the isothermal compressibility, Δt is the time step and τ_p is the pressure coupling time constant and $P(t)$ is the instantaneous pressure.

Monte Carlo simulations generate the microscopic realizations in a different way. It tries to directly sample system configurations $\{\boldsymbol{r}^N, \boldsymbol{p}^N\}$ from the corresponding equilibrium probability distribution over these variables. In a NVT ensemble, this probability distribution is given by

$$P(\boldsymbol{r}^N, \boldsymbol{p}^N) = \frac{1}{Z} e^{-\beta \mathcal{H}(\boldsymbol{r}^N, \boldsymbol{p}^N)} \tag{1.6}$$

where Z is the partition function, and the ensemble average of an observable is given by

$$\langle A \rangle = \frac{1}{Z} \int A(\boldsymbol{r}^N, \boldsymbol{p}^N) e^{-\beta \mathcal{H}(\boldsymbol{r}^N, \boldsymbol{p}^N)} d\boldsymbol{r}^N d\boldsymbol{p}^N \tag{1.7}$$

The Hamiltonian is $\mathcal{H} = U(\boldsymbol{r}^N) + \mathcal{K}(\boldsymbol{p}^N)$ and $\mathcal{K}(\boldsymbol{p}^N)$ is a quadratic function of the momenta, so usually in equation 1.7 the momenta part could be analytically integrated out and we are interested in

$$\langle A \rangle = \frac{1}{Z} \int A(\boldsymbol{r}^N) e^{-\beta U(\boldsymbol{r}^N)} d\boldsymbol{r}^N \tag{1.8}$$

which is an ensemble average according to the marginal probability density

$$P(\boldsymbol{r}^N) = \frac{1}{Z} e^{-\beta U(\boldsymbol{r}^N)} \tag{1.9}$$

The Monte Carlo approach tries to directly sample $\{\boldsymbol{r}_1^N, ..., \boldsymbol{r}_L^N\}$ from equation 1.9 and then use the sample average to approximate the ensemble average:

$$\langle A \rangle \approx \frac{1}{L} \sum_{j=1}^{L} A(\boldsymbol{r}_j^N) \tag{1.10}$$

Equation 1.9 is a probability density where we don't know the normalization constant Z, since in practical cases it is impossible to calculate analytically. Markov Chain Monte Carlo (MCMC)[35] is a technique to sample from a probability density without knowing the normalization constant. It proceeds by constructing a Markov Chain of microstates $\{\boldsymbol{r}^N\}$ whose steady state distribution is equation 1.9. In order for the Markov Chain to have this steady state distribution, its transition probability $T(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N)$ between microstates must satisfy the detailed balance condition:

$$P(\boldsymbol{r}_t^N)T(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N) = P(\boldsymbol{r}_{t+1}^N)T(\boldsymbol{r}_t^N|\boldsymbol{r}_{t+1}^N) \tag{1.11}$$

One way to construct a transition probability that satisfies detailed balance condition is by first sampling a new state $\boldsymbol{r}_{t+1}^N$ given current state $\boldsymbol{r}_t^N$ from a proposal distribution $Q(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N)$ and then defining the transition probability as

$$T(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N) = Q(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N) \min\left(1, \frac{P(\boldsymbol{r}_{t+1}^N)Q(\boldsymbol{r}_t^N|\boldsymbol{r}_{t+1}^N)}{P(\boldsymbol{r}_t^N)Q(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N)}\right) \tag{1.12}$$

The latter part in equation 1.12 can be interpreted as the acceptance probability of the proposed state $\boldsymbol{r}_{t+1}^N$ given the state $\boldsymbol{r}_t^N$. The algorithm is summarized in algorithm 1.

Note that the normalization constant Z in equation 1.9 cancels out when evaluating the transition probability in equation 1.12. The Markov chain constructed in this way is guaranteed to have unique steady state distribution according to equation 1.9, and for sufficiently large L the samples generated (once the chain has stabilized) should be distributed according to the steady state distribution. The choice of proposal distribution Q is arbitrary; common

Algorithm 1 MCMC

```
Initialize  $\boldsymbol{r}_0^N$
while $t \leq L$ do
    Sample  $\tilde{\boldsymbol{r}}_{t+1}^N \sim Q(\boldsymbol{r}_{t+1}^N|\boldsymbol{r}_t^N)$
    Set  $a = \min\left(1, \frac{P(\tilde{\boldsymbol{r}}_{t+1}^N)Q(\boldsymbol{r}_t^N|\tilde{\boldsymbol{r}}_{t+1}^N)}{P(\boldsymbol{r}_t^N)Q(\tilde{\boldsymbol{r}}_{t+1}^N|\boldsymbol{r}_t^N)}\right)$
    Sample  $u \sim \text{Unif}[0, 1]$
    if $u \leq a$ then
        $\boldsymbol{r}_{t+1}^N \leftarrow \tilde{\boldsymbol{r}}_{t+1}^N$
    else
        $\boldsymbol{r}_{t+1}^N \leftarrow \boldsymbol{r}_t^N$
    end if
    $t \leftarrow t + 1$
end while
```

choices include uniform distribution with a step size δ or a multivariate Gaussian with diagonal covariance matrix $\sigma^2\mathbb{I}$. There are some more advanced Monte Carlo methods that aim to increase the sampling efficiency, such as parallel tempering method[36] and cluster methods[37,38].

1.3 Machine Learning and Optimization Algorithms

Generally speaking, machine learning algorithms could be divided into two categories: supervised and unsupervised learning algorithms. In supervised learning, we are provided with some inputs $\boldsymbol{x}$ and outputs $\boldsymbol{y}$. The task of supervised learning is to find an optimal function mapping from the inputs to the outputs. Depending on the type of outputs, this could be a regression problem ($\boldsymbol{y}$ is continuous) or classification problem ($\boldsymbol{y}$ is discrete). On the other hand, unsupervised learning algorithms try to find patterns in the unlabeled data. Two major areas in unsupervised learning include clustering and dimensionality reduction.

There exist numerous algorithms for performing supervised learning, ranging from linear regression to modern deep neural networks. Many of these algorithms have been widely applied to build fast predictive and classification models for properties of materials. For example, Sanchez-Lengeling et al. has implemented a Gaussian Process Regression model to enable fast prediction of the solubility parameters of compounds[39]. Leslie et al. proposed mismatch kernels based on a tree data structure to perform support vector machine classification of proteins for several benchmark tasks and demonstrated good performance[40]. Yang et al. employed doc2vec model from natural language processing to embed proteins into a vector space on which they performed Gaussian Process Regression on benchmark tasks and nonlinear dimensionality reduction to evaluate the performance of protein embedding[41]. Ye et al. trained a graph neural network model to accurately predict the stabilities of composite crystals[42]. The AlphaFold model developed by DeepMind has done an amazing job in predicting the 3D folding structures of proteins[43]. With the help of supervised learning, researchers can not only predict the material properties, but can also perform high-throughput screening to search for optimal building blocks that have desired properties. For example, Shmilovich et al.[30] trained variational autoencoders to embed π-conjugated peptides into a low-dimensional vector space over which Gaussian Process Regression were constructed to support Bayesian optimization discovery of new π-conjugated peptides that can spontaneously form nanostructures with emergent optoelectronic properties.

Unsupervised learning algorithms have also been helpful in understanding the behaviors of molecular systems. A key aspect in unsupervised learning is dimensionality reduction technique which tries to find a meaningful low-dimensional representation of high-dimensional data. This is particularly useful in molecular simulation, where snapshots of system configurations live in a very high dimensional space ($3N$-dimensional space where N is the number of particles in the system) and projecting the snapshots into a low-dimensional space could provide more insights into the time-evolution of system. The most well-known dimensionality reduction technique is the principal component analysis (PCA) which tries to project data into a lower-dimensional hyperplane that maximizes the variance after projection. PCA has been applied to understand the conformational change of proteins and lipids in molecular dynamics simulations[44,45]. A limitation of PCA is that it could only find lower dimensional representations that consist of linear combination of original features. However, usually the configurations generated in a molecular simulation occupy a nonlinear manifold in $3N$-dimensional configuration space due to the coupling between the degrees of freedom. In this case, linear dimensionality reduction algorithms such as PCA may be inadequate to find a meaningful lower-dimensional representation. Thus, many nonlinear dimensionality reduction techniques have become popular, such as locally linear embedding, Isomap[46], Laplacian Eigenmap[47] and diffusion map[48]. These nonlinear dimensionality reduction techniques all aim to find a low-dimensional representations that capture the manifold structure of the data in high-dimensional space. These nonlinear techniques have been successfully applied to the analysis of molecular simulation trajectories, such as probing the folding pathway of antimicrobial peptide[49] and defining reaction coordinates for the construction of free energy surface[50]. Deep-learning approaches, such as variational autoencoders, have also been applied to find meaningful low-dimensional representation of peptides[30]. Besides dimensionality reduction, clustering techniques such as K-means and K-medoids have also been widely utilized to study the transition between different metastable states in molecular system. Usually, a clustering algorithm is first applied to the simulation trajectories to identify meaningful metastable states by grouping conformations into clusters based the similarities between pairs of conformations, and then the transition between the metastable states could be analyzed[51]. For example, Beauchamp et al.[52] proposed to use Ward clustering to identify appropriate clusters for conformations of 14 different proteins and then used a Markov State Model to understand the protein folding dynamics.

Finally, another major area in machine learning that is often helpful for materials research is black-box optimization, which is usually used in combination with supervised and unsupervised learning. Usually, researchers would like to optimize the chemical and physical properties of materials (for example, ability of colloids to self-assemble into desired crystalline lattice). However, when considering the target properties as functions of the features of materials, we usually do not have the information about the functional form and the derivatives of the target functions but still want to optimize them. This is called a black-box optimization problem since the target functions act like “black boxes”: we could evaluate their values for a given set of features of materials by simulation or experiment, but we could not evaluate their derivatives. This requires the use of optimization algorithms that do not need information about gradient or Hessian. A variety of such algorithms have been developed. For example, genetic algorithms are stochastic optimization algorithms that are based on guided random walks over input space to direct the random walks to regions with high probability of optimal target function values[53]. Such algorithms have been successfully applied in the optimization of copolymer self-assembly[54]. Bayesian optimization is another commonly used black-box optimization algorithm which relies on building a Gaussian Process Regression model for the target function[55]. It has been used in the search of optimal nanoporous materials[56,57].

1.4 Outline of the Thesis Work

In this thesis, we have employed modern machine learning and black-box optimization algorithms to tackle some of the challenging tasks in the data-driven design of soft materials. The outlines of each of the subsequent chapters are summarized below.

- In chapter 2, we describe our work that combines molecular simulation and stochastic optimization algorithm to find optimal design parameters in anisotropic patchy colloids that enable them to self-assemble into open crystalline lattices. The method is based on generating the free energy surface that quantifies the stability of achievable self-assembled structures of colloidal system by molecular simulation, and then sculpting the free energy surface to maximize the stability of target structure by stochastic optimization algorithms. We demonstrate the success of our inverse design strategy in designing patchy colloids to self-assemble into colloidal pyrochlore and cubic diamond lattices, which are highly sought-after nanostructures with useful optical properties. This chapter is primarily based on the paper: **Y. Ma** and A.L. Ferguson "Inverse design of self-assembling colloidal crystals with omnidirectional photonic bandgaps" *Soft Matter* 15 8808-8826 (2019).

- In chapter 3, we describe a follow-up work to chapter 2 in which we perform inverse design on colloidal clusters composed of anisotropic patchy colloids to enable those clusters to self-assemble into open crystalline lattices. The patchy colloid model is simpler than the one considered in chapter 2 and is more amenable to experimental realization. Also, we use a simpler objective function that is computationally cheaper to evaluate but still captures the thermodynamic stability of target structure. We demonstrate our inverse design protocol in designing colloidal clusters to self-assemble into cubic diamond lattice with omnidirectional bandgap. This chapter is based on the paper: **Y. Ma**, J. Aulicino, and A.L. Ferguson "Inverse design of self-assembling diamond photonic lattices from anisotropic colloidal clusters" *J. Phys. Chem B* 125 9 2398-2410 (2021).

- In chapter 4, we describe our work on building a high-throughput screening protocol for finding optimal elastin-like polypeptides that could self-assemble into stable vesicles. Here we use molecular simulation to generate potential of mean force profile that quantifies the thermodynamic stability of self-assembled vesicles and use Bayesian optimization to find peptide sequence that maximizes this stability. This chapter is based on the works reported in: (1) B. Sharma, **Y. Ma**, A.L. Ferguson, and A.P. Liu "In search of a novel chassis material for synthetic cells: Emergence of synthetic peptide compartment" *Soft Matter* 16 10769 (2020); (2) B. Sharma, **Y. Ma**, H.L. Hiraki, B.M. Baker, A.L. Ferguson, and A.P. Liu "Facile formation of giant elastin-like polypeptide vesicles as synthetic cells" *Chem. Commun.* 57 13202-13205 (2021).

- In chapter 5, we summarize our projects and suggest potential future directions to continue our work.

CHAPTER 2
INVERSE DESIGN OF SELF-ASSEMBLING COLLOIDAL CRYSTALS WITH OMNIDIRECTIONAL PHOTONIC BANDGAPS

2.1 Abstract

Open colloidal lattices possessing omnidirectional photonic bandgaps in the visible or near-visible regime are attractive optical materials whose realization has remained elusive. We report the use of an inverse design strategy called landscape engineering that rationally sculpts the free energy landscape of self-assembly using evolutionary algorithm to discover anisotropic patchy colloids capable of spontaneously assembling pyrochlore and cubic diamond lattices possessing complete photonic bandgaps. We validate the designs in computer simulations to demonstrate the defect-free formation of these lattices via a two-stage hierarchical assembly mechanism. Our approach demonstrates a principled strategy for the inverse design of self-assembling colloids for the bottom-up fabrication of desired crystal lattices. This chapter is mainly based on the work reported in: **Y. Ma** and A.L. Ferguson "Inverse design of self-assembling colloidal crystals with omnidirectional photonic bandgaps" *Soft Matter* 15 8808-8826 (2019).

2.2 Introduction

The self-assembly of colloidal nanoparticles provides a powerful tool for forming many complex structures, such as colloidal aggregates[58–62], multi-shell clusters[63], helical structures[64] and crystals[8,24,65–71]. The assembly of open colloidal lattices has drawn particular attention[8,24,65–71] because many of these structures posses complete photonic bandgaps and are therefore useful as 3D photonic crystals with omnidirectional bandgaps[69,72–74].The optical properties of a colloidal crystal are dictated by the organization of colloids within the crystal lattice[75]. The size of the colloids, refractive index contrast between the colloids and voids, and specific pattern of refractive index changes due to the packing of colloids all dictate the photonic properties of the crystal[75]. Whereas hexagonal close packed (hcp) and face-centered cubic (fcc) lattices are most easily assembled from isotropic colloidal spheres, they do not possess complete photonic bandgaps that forbid passage of photons with particular energies in all directions. It is for this reason that more exotic open lattices such as pyrochlore[69,72,74], diamond[76,77] and inverse opal[75,78,79] that do possess complete bandgaps have attracted much attention. These crystals have desirable applications in the manipulation of photons in optical wave guiding[80] and in optical computing[81].

Many techniques have been explored to synthesize colloidal crystals by bottom-up self-assembly[8,24,65–71]. For example, van Driel and coworkers have synthesized single crystalline silicon inverse opal using close-packed silica colloidal spheres as a template[79]. Damman and coworkers have assembled colloidal lattices by vertical deposition on curved surfaces and demonstrated that the optical properties of the resulting colloidal crystal could be manipulated by the surface curvature without introducing crystal defects[82]. Crocker and coworkers have employed two differently-sized spherical colloids functionalized with complementary DNA oligomers to fabricate "double diamond" (B32) colloidal crystals isomorphic to the NaTl Zintl phase[83]. Pine and coworkers have co-assembled tetrahedral colloidal clusters and colloidal spheres using complementary DNA binding to fabricate a colloidal $MgCu_2$ crystal[8]. In a recent work, Pine and coworkers also have employed compressed tetrahedral colloidal clusters with a combination of adhesive interactions and steric interlocking to realize cubic diamond lattice[84]. Grzybowski and coworkers have assembled diamond-like colloidal lattices from nearly equally-sized oppositely-charged nanoparticles[70,85]. In the context of the assembly of open crystal structures, patchy colloids functionalized with anisotropic and directional surface interactions have emerged as a promising means for their fabrication by self-assembly[12,66–68,71,75,78]. This class of building blocks is experimentally attractive as they are based on simple spherical colloids that can be flexibly functionalized through anisotropic surface patterning techniques[9,15,86–88]. For example, Chen et al.[89] have experimentally demonstrated the self-assembly of triblock patchy colloids into metastructures by a step-wise control of ion concentration in solution, and Morphew et al.[66] have computationally investigated the hierarchical self-assembly of triblock patchy colloids into body-centered cubic and cubic diamond crystals.

A primary challenge in patchy colloid self-assembly is the design of the anisotropic interactions to favor the assembly of the desired crystal lattice. Of particular concern is the existence of competing crystal structures with similar free energies that can frustrate defect-free assembly of the target lattice. For example, the pyrochlore lattice (also known as cubic tetrastack)[74,90] can be viewed as a tetrahedral network of corner-sharing tetrahedral clusters[72], where the tetrahedral clusters occupy the voids of a cubic diamond lattice[8,90]. However, the pyrochlore lattice has a closely-related analogue known as the hexagonal tetrastack lattice, which differs from the pyrochlore lattice by the orientations of adjacent layers but has similar free energy[65,90,91]. Similarly, cubic diamond lattice (also known as conventional diamond or c-diamond) and hexagonal diamond lattice (also known as h-diamond or Lonsdaleite) are structurally similar crystals with similar free energies that differ in the stacking of subsequent layers[68]. In each case, care must be taken in the patchy particle design to favor one polymorph over the other[65,67,68,90]. As such, a primary concern in our design protocol is to engineer anisotropy into the particle interaction to break the degeneracy between the desired lattice structures and closely related analogues (i.e., pyrochlore vs. hexagonal tetrastack, cubic diamond vs. hexagonal diamond).

A number of inverse design techniques have been proposed for optimizing the interactions between colloidal particles[8,24–27,66,67,90,92–98] to favor the formation of the desired target crystal. For example, Truskett and coworkers have used inverse design strategies to design isotropic pairwise potentials that favor the formation of various two-dimensional and three-dimensional colloidal crystals[25,96–98]. Torquato and coworkers also have employed inverse statistical mechanical strategies to design isotropic potentials that favor various colloidal lattices[26,27,95] as well as anisotropic potentials that favor the formation of various two-dimensional colloidal lattices[92]. Lyubartsev and Laaksonen[99] and Mungan et al.[100] have deduced optimal interaction potentials from structural correlation functions. Cohn and Kumar[28] have employed linear programming to determine isotropic potentials leading to the desired configuration as ground state. Dijkstra and coworkers have employed an isotropic repulsive pairwise potential to favor the formation of a pyrochlore-like colloidal crystal[24]. Escobedo[93] has used anisotropic particles and potentials to form different colloidal compounds. Romano and Sciortino have employed asymmetric patterning to robustly assemble pyrochlore and disfavor hexagonal tetrastack lattice[90]. Morphew et al.[66] have used a basin-hopping optimization method to design the potentials that favor the formation of three-dimensional cubic diamond lattice and BCC lattice via colloidal molecules. Glotzer and coworkers have proposed the introduction of angular potentials or charge repulsion to favor cubic over hexagonal diamond[67]. Pine and coworkers have designed isotropic DNA-grafted colloidal clusters and singlet colloids to realize a colloidal $MgCu_2$ lattice[8]. In the absence of some means to break the degeneracy between competing polymorphs it is typically necessary to seed the system with a fragment of the desired crystal structure[67] to robustly assemble the target crystal. Failing to break the degeneracy through one or other of these strategies risks the uncontrolled fabrication of hybrid lattices[66].

In this work, we employ a recently developed inverse design protocol, termed landscape engineering[101], to systematically discover patchy colloid building blocks capable of spontaneously assembling into a pyrochlore lattice and a cubic diamond lattice formed from tetrahedral clusters. The approach iteratively sculpts the free energy surface of the self-assembling colloids using evolutionary algorithm to update the placement and strength of the colloidal patches to stabilize the target lattice over all competing polymorphs. We target pyrochlore and cubic diamond as 3D lattices possessing complete photonic bandgaps that have proven frustratingly elusive to fabrication via self-assembly[66,68]. We show that the colloidal designs predicted by landscape engineering spontaneously nucleate and grow defect-free photonic lattices of the desired crystal polymorphs in a two-stage hierarchical assembly mechanism. Since we conduct inverse design over the free energy surface rather than the potential energy surface, the interaction potentials discovered by landscape engineering are not those that would have been expected by energy minimization or zero-temperature optimization of the target lattice, which demonstrates the importance of incorporating many-body and entropic effects into the particle design. The interaction potentials discovered by landscape engineering are not those that would have been expected by energy minimization or zero-temperature optimization of the target lattice, demonstrating the importance of many-body and entropic effects in particle design.

The anisotropic potentials employed in this work are relatively simple and generic, but may be considered as simplified and idealized models of inter-particle interactions that may be experimentally realized through advanced surface-patterning techniques[8–15,86,102]. For example, the patchy colloid model considered in this work can be considered as a simplified representation of nanodot-decorated nanoparticles of the sort realized by Bae et al.[15] and Wang et al.[86] through regions of titania or propyl methacrylate whose interaction strengths depend on the specific materials properties. In a similar vein, Zhang et al.[103] used colloidal masks to fabricate anisotropic nanoparticles decorated with nanodots on opposite poles. We might also consider our models to be idealized representations of colloids whose surfaces are functionalized with localized patches of complementary DNA oligomers with defined sequence and specificity[8–11,102]. The Kern-Frenkel model[104] is one of the most popular computational models employed to simulate patchy particle assembly[66,90] and can be considered a simplified model for patchy particles with surface interaction patches deposited via glancing angle deposition[13,14,105]. Accordingly, our computational patchy particle model and others like it are intended as simplified idealizations of experimentally-realizable interparticle interactions. It is the primary goal of the present work to employ such potentials to expose the fundamental principles governing assembly, provide new insight into the thermodynamic, kinetic, and morphological processes underpinning assembly, and demonstrate a new methodology for the rational design of patchy colloids programmed to self-assemble into desired aggregates. In doing so, we aim to provide new understanding and precepts for the experimental design of self-assembling colloidal lattices. Romano and Sciortino have previously proposed the use of asymmetric Kern-Frenkel type patchy colloids to form pyrochlore lattice[90]. The present work considers a different patchy particle model with defined isotropic surface interactions that may be considered a simplified representation of nanodot-decorated nanoparticles[15,86]. Moreover, we design the anisotropic interaction potentials using a systematic and automated inverse design protocol. Accordingly, this work reports a new automated inverse design strategy for the fabrication of desired colloidal lattices, and reduces this to practice in the design of two patchy particle building blocks capable of spontaneously self-assembling pyrochlore and cubic diamond lattices with omnidirectional photonic bandgaps.

2.3 Methods

2.3.1 Self-Assembling Patchy Colloid Model

Pyrochlore Lattice

The pyrochlore lattice can be viewed as a tetrahedral network of corner-sharing tetrahedra. An illustration of the pyrochlore lattice is given in figure 2.1a. Every particle (i.e. vertex of tetrahedron) in the pyrochlore lattice exists in a staggered local configuration (figure 2.1b) where its six nearest neighbors are rotated by 60° around it. A competing crystal structure with similar free energy is the hexagonal tetrastack lattice in which 75% of particles exist in staggered local configuration and 25% of particles exist in eclipsed local configuration[90] (figure 2.1c). In order to favor the pyrochlore lattice against hexagonal tetrastack lattice, we decorate three "B" patches (blue patches) forming an equilateral triangle on the north pole of the central sphere ("A" particle) and three "D" patches (purple patches) forming an equilateral triangle on the south pole. The "D" patches are rotated by 60° degree around the central axis with respect to the "B" patches. As we shall see, the "B" patches serve as the interaction sites for directing individual patchy colloids to form tetrahedral clusters, and "D" patches serve as the interaction sites for directing tetrahedral clusters to form corner-sharing network of tetrahedra while maintaining the staggered local configuration of the vertices of tetrahedra in this network. This model is illustrated in figure 2.1d-f.

The patches and the central particle are treated as a single rigid body building block, and the interactions within the same rigid body are ignored. The diameters of central particle and the patches are chosen as $\sigma_A = 5\sigma$ and $\sigma_B = \sigma_D = \sigma$. The masses are chosen as $m_A = 125m$ and $m_B = m_D = m$. The interactions between "B"-"B" and "D"-"D" patches on different patchy colloids are modeled by Lennard-Jones potential:

$$U_{\mathrm{LJ}}^{ii}(r) = 4\varepsilon_i \left[\left(\frac{\sigma_i}{r} \right)^{12} - \left(\frac{\sigma_i}{r} \right)^{6} \right] \quad \text{for } i \in \{\text{B,D}\} \tag{2.1}$$

where ε_i is the well depth, or the "interaction strength", of particle i, and σ_i is its diameter. The interactions between "A"-"X", where "X" $\in$ {"A","B","D"}, and between "B"-"D" particles on different patchy colloids are modeled by Weeks-Chandler-Andersen (WCA) potential[106] to incorporate excluded-volume effects:

$$U_{\mathrm{WCA}}^{ij}(r) = \begin{cases} 4\varepsilon_{ij} \left[\left(\frac{\sigma}{r-\Delta} \right)^{12} - \left(\frac{\sigma}{r-\Delta} \right)^{6} \right] + \varepsilon_{ij} & \text{if } r < 2^{\frac{1}{6}}\sigma + \Delta_{ij} \\ 0 & \text{if } r \geq 2^{\frac{1}{6}}\sigma + \Delta_{ij} \end{cases} \tag{2.2}$$

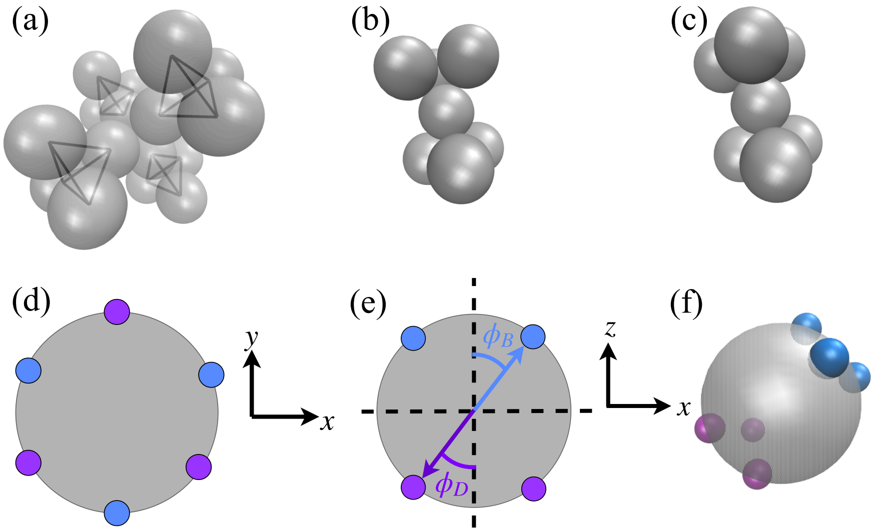


Figure 2.1: Pyrochlore lattice. (a) A unit cell of the pyrochlore lattice. (b) An illustration of the staggered tetrahedral configuration that forms the fundamental motif of the lattice. The central particle forms a regular tetrahedron with its three nearest neighbors above and similarly with its three nearest neighbors below. The two tetrahedra are rotated 60° relative to one another to form a staggered configuration. (c) An illustration of the eclipsed local configuration. Schematic (d) top-down and (e) side views of the anisotropic patchy colloid building block to be optimized by landscape engineering. The blue "B" patches with interaction strengths ε_B define the vertices of an equilateral triangle on the north pole of the patchy colloid at a polar angle of ϕ_B. The purple "D" patches with interaction strengths ε_D lie at the vertices of an analogous south pole equilateral triangle at a polar angle $\phi_D = \phi_B$ and a relative azimuthal rotation of 60°. The 60° azimuthal rotation between the north and south pole patches energetically favors the staggered tetrahedral configuration (b) over the eclipsed (c). (f) Three dimensional rendering of the patchy colloid where the central particle is made transparent to show the staggered orientation between "B" and "D" patches. Landscape engineering is employed to optimize $\{E_B, \phi_B, E_D, \phi_D\}$ to promote the two-stage hierarchical self-assembly of the pyrochlore lattice.

$\Delta_{ij} = \frac{\sigma_i+\sigma_j}{2} - 1$ shifts the potential to act on the surfaces of particles i and j. $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ is given by the Lorentz-Berthelot mixing rule. Since there are three "B" patches and three "D" patches on each patchy colloid, we specify the total interaction strength of each of these two species: $E_i = 3\varepsilon_i$ where $i \in \{\text{B,D}\}$, and evenly distribute it among all three patches of each species.

We design the particles to assemble the pyrochlore lattice via a two-step hierarchical assembly mechanism. The "B" patches possess a stronger interaction strength and are activated during the high temperature phase of the assembly process at T_{high} to direct the patchy colloids to assemble into tetrahedral clusters. The "D" patches, which have a weaker interaction strength, are then activated during the temperature cooling process down to T_{low} and direct the assembly of the tetrahedral clusters into the pyrochlore lattice. The two-stage hierarchy in interaction strengths maps to a two-stage hierarchy in structure that has previously been exploited for the fabrication of hierarchically structured materials[66,89,107,108]. In the present work, the high-temperature assembly process can be conceived as producing tetrahedral building blocks from spherical colloids, and the low-temperature assembly process as directing the assembly of the tetrahedra into a pyrochlore lattice. Thermal decoupling between the two levels of the hierarchy is asserted in the relation between E_B and E_D as:

$$E_D = \frac{T_{\text{low}}}{T_{\text{high}}} E_B. \tag{2.3}$$

Since the "B" and "D" patches both mediate the formation of tetrahedra, their polar angles are related as:

$$\phi_D = \phi_B. \tag{2.4}$$

As mentioned above, their azimuthal angles are mutually rotated by 60° to favor staggered local configurations (figure 2.1d). An illustration of the desired two-stage assembly process is shown in figure 2.2.

The "D" patches on the south pole act as the low-temperature counterparts of "B" patches on the north pole and the desired self-assembled tetrahedral motif in both cases is the same. At the high temperature T_{high} we seek to optimize the interaction strength and polar angle of the "B" patches at the north pole $\{E_B, \phi_B\}$ to favor the assembly of tetrahedral clusters relative to all competing structures. We solve this inverse design problem using the landscape engineering[101] approach described below. We then obtain the optimal solution for the "D" patches on the south pole through equations 2.3 and 2.4. In doing so we assume

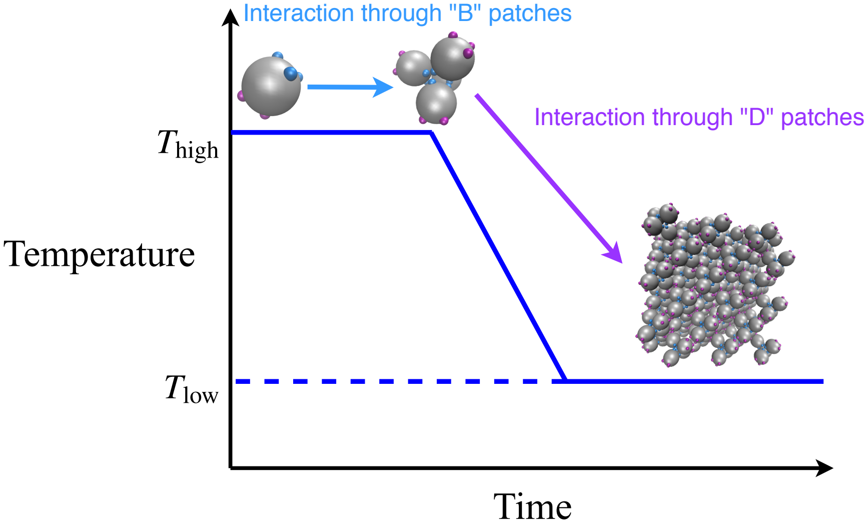

Interaction through "B" patches
Interaction through "D" patches
T_{high}
Temperature
T_{low}
Time

Figure 2.2: An illustration of the two-stage hierarchical assembly mechanism for the pyrochlore lattice. During the high temperature phase at T_{high}, the more strongly interacting "B" patches direct the formation of tetrahedral clusters while the more weakly interacting "D" patches are effectively inert. During the cooling process to T_{low}, the "B" patches lock the patchy colloids into the self-assembled tetrahedra and the "D" patches direct the tetrahedral clusters to assemble into the pyrochlore lattice.

that the optimal solution for the "D" patches at the low temperature is identical to that for the "B" patches at the high temperature, with the interaction strength just appropriately scaled by the temperature ratio. The motivation for this equivalence is that the "B" and "D" interfaces are structurally identical, but we do note that the former is formed from constituent monomers whereas the latter is formed from constituent tetrahedra, so the multibody and entropic interactions during the assembly process may differ. Nevertheless, the assumption of this equivalence simplifies the inverse design problem for the pyrochlore lattice by reducing it to a single optimization. As we will show, the assumption turns out to be a good one as it leads to the successful assembly of defect-free crystals. As discussed later, in the case of cubic diamond this symmetry is absent and we must independently optimize the two poles of the colloidal building blocks within two separate optimization protocols.

Our computational model for the anisotropic interaction patches is deliberately a simple and generic potential. Experimentally, such directional and specific patches with particular

$\{E_B, \phi_B, E_D, \phi_D\}$ might be realized by advanced surface-patterning techniques[8–15,86,102]. For example, the central colloidal spheres may be functionalized by DNA oligomers with tunable interaction strengths and specificities[9], or patterned with regions of titania or propyl methacrylate whose interaction strengths depend on the specific materials properties[15,86].

We perform our simulation in reduced units, where $\sigma = 1$, $\varepsilon = \varepsilon_A = 1$, and $m = 1$. Using these units, we specify $\sigma_A = 5$, $\sigma_i = \sigma = 1$ for $i \in \{B, D\}$, $m_A = 125$ and $m_i = m = 1$ for $i \in \{B, D\}$. We may define a mapping between our reduced units and real units. For example, consider central particle "A" of diameter $\sigma_A = 5\sigma = 1$ μm and density $\rho_A = 1$ g/cm^3, and an energy scale of $\varepsilon = 1$ k_BT at $T = 298$ K. From these fundamental units, we can derive the temperature in real units as $T = T^* \frac{\varepsilon}{k_B}$ and time in real units as $t = t^* \sigma \sqrt{\frac{m}{\varepsilon}}$, where T^* and t^* are temperature and time in reduced units. In our two-stage assembly process, we use $T^*_{\text{high}} = 0.8$ at high temperature and $T^*_{\text{low}} = 0.3$ at low temperature, which correspond to $T_{\text{high}} = 238.4$ K and $T_{\text{low}} = 89.4$ K in real units, respectively. Also, we use a step size of $dt^* = 0.005$ in our simulations, which corresponds to $dt = 0.36$ μs in real units.

Cubic Diamond Lattice

The fundamental motif of the cubic diamond lattice formed from tetrahedral clusters is dimers of tetrahedra in staggered configurations (figure 2.3a)[66,67,90]. (For short, we will henceforward simply refer to this lattice as cubic diamond except when it is unclear to do so.)

These staggered tetrahedral dimers come together to form chair-like rings (figure 2.3b) within the cubic diamond crystal. A competing structure sharing similar free energy is the hexagonal diamond lattice that contains tetrahedral dimers in both staggered and eclipsed (figure 2.3c) configurations. The hexagonal diamond lattice consists of 25% chair-like rings and 75% boat-like rings (figure 2.3d)[68]. To favor cubic over hexagonal diamond, we employ a similar patchy colloid design to that for pyrochlore. Three "B" patches with interaction

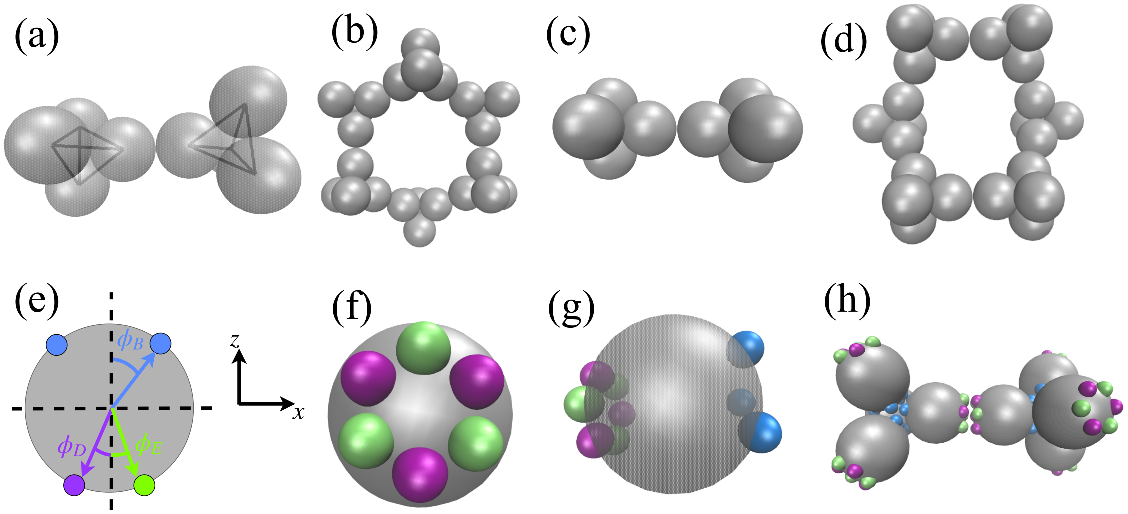


Figure 2.3: Cubic diamond lattice. (a) Staggered dimer of tetrahedral clusters. (b) Chair-like ring of tetrahedral clusters. (c) Eclipsed dimer of tetrahedral clusters. (d) Boat-like ring of tetrahedral clusters. (e) Schematic side view of the anisotropic patchy colloid. The blue "B" patches with interaction strengths ε_B define the vertices of an equilateral triangle on the north pole of the patchy colloid at a polar angle of ϕ_B. The purple "D" and the lime "E" patches with interaction strengths ε_D and ε_E lie at the vertices of a analogous south pole equilateral triangles at a polar angle $\phi = \phi_D = \phi_E$. The "D" patches are azimuthally aligned with the "B" patches and the "E" patches rotated by 60°. The (f) bottom and (g) side views of a three dimensional rendering of the patchy colloid. In (g) the central particle is made transparent to show the relative orientations of the "B", "D", and "E" patch types. (h) Illustration of a staggered dimer formed by two tetrahedral clusters of patchy colloids. Landscape engineering is employed to optimize $\{E_B, \phi_B, E_D, E_E, \phi = \phi_D = \phi_E\}$ to promote the two-stage hierarchical self-assembly of the cubic diamond lattice.

strength ε_B are placed in an equilateral triangle on the north pole at a polar angle of ϕ_B to favor the high temperature formation of tetrahedral clusters (figure 2.3e-g). The south pole must be functionalized with two kinds of patches, "D" patches and "E" patches, at a polar angle $\phi = \phi_D = \phi_E$ in order to preferentially stabilize the staggered dimer relative to the eclipsed one. The "D" patches (purple) are aziumthally aligned with the "B" patches (blue), and the "E" patches (lime) are rotated by 60°. The "B"-"B" interaction is still modeled by Lennard-Jones potential as in equation 2.1. The "D"-"E" interaction is also modeled by Lennard-Jones potential with the interaction strength and range given by the Lorentz-Berthelot mixing rules:

$$\begin{aligned} U_{\mathrm{LJ}}^{DE}(r) &= 4\varepsilon_{DE}\left[\left(\frac{\sigma_{DE}}{r}\right)^{12} - \left(\frac{\sigma_{DE}}{r}\right)^{6}\right] \\ \varepsilon_{DE} &= \sqrt{\varepsilon_D \varepsilon_E} \\ \sigma_{DE} &= \frac{\sigma_D + \sigma_E}{2} \end{aligned} \tag{2.5}$$

All other interactions are given by WCA potential defined in equation 2.2. In this way, the "D"-"E" attractive interactions induce the contact dimer formed by two tetrahedra to stabilize a mutual rotation of 60° and favor the staggered dimer over the eclipsed one figure 2.3(h). Our use of this design with two patch types in an alternating ring with attractive interactions between unlike patches is motivated by the need to induce a 60° rotation between the two tetrahedra at the dimer interface.

The optimal parameters $\{E_B, \phi_B\}$ for the "B" patches will be the same as those for pyrochlore since in both cases the high temperature assembly process into tetrahedral clusters is identical. The remaining goal is to optimize the parameters for "D" and "E" patches $\{E_D, E_E, \phi = \phi_D = \phi_E\}$ through landscape engineering. Without loss of generality, we choose to optimize the parameters for "D" and "E" patches at $T^*_{\mathrm{high}} = 0.8$, and then scale down the interaction strengths to match the low temperature phase $T^*_{\mathrm{low}} = 0.3$ by a factor of $T^*_{\mathrm{low}}/T^*_{\mathrm{high}}$. We specify $\sigma_D = \sigma_E = 1$ and $m_D = m_E = 1$ in reduced units. The mapping between reduced units and real units is defined in the same manner as section 2.3.1.

2.3.2 Landscape Engineering

We follow a previously developed optimization procedure, called landscape engineering[101], to optimize the design parameters of patchy colloid. Hereafter we will refer to a set of design parameters as a "candidate". The whole procedure consists of the following steps: Starting from a group of initial candidates, we first conduct Langevin dynamics simulations of the self-assembly of each candidate and use diffusion map nonlinear dimensionality reduction to construct a low-dimensional embedding of the various self-assembled configurations observed over the course of the simulation. Next, on this low-dimensional diffusion map space, we select biasing centers and perform umbrella sampling[109,110] for each candidate in order to enhance the sampling of accessible configurations. From the results of umbrella sampling, for each candidate, we construct its self-assembly free energy surface which contains information about the stability of all accessible configurations. We then locate the target structure on the free energy surface and define a fitness metric based on the relative stability between the target structure and the nearest competitor. Finally, we use an evolutionary algorithm, called Covariance Matrix Adaptation Evolution Strategy (CMA-ES)[111], to propose new and improved candidates based on the fitness values of old candidates. These steps are repeated until the proposed new candidates stabilize around an optimal candidate. A flowchart of this procedure is shown in figure 2.4.

As described in the previous section, the parameters being optimized in the patchy colloid models for assembling pyrochlore lattice and cubic diamond lattice are $\{E_B, \phi_B\}$ and $\{E_D, E_E, \phi\}$, respectively. In case of pyrochlore lattice, our goal is to optimize $\{E_B, \phi_B\}$ that favor the formation of tetrahedral cluster at $T^*_{\text{high}} = 0.8$, and then the parameters for the "D" patches are obtained through equations 2.3 and 2.4. In this case, we conduct landscape engineering on particles possessing only "B" patches on the central sphere such that

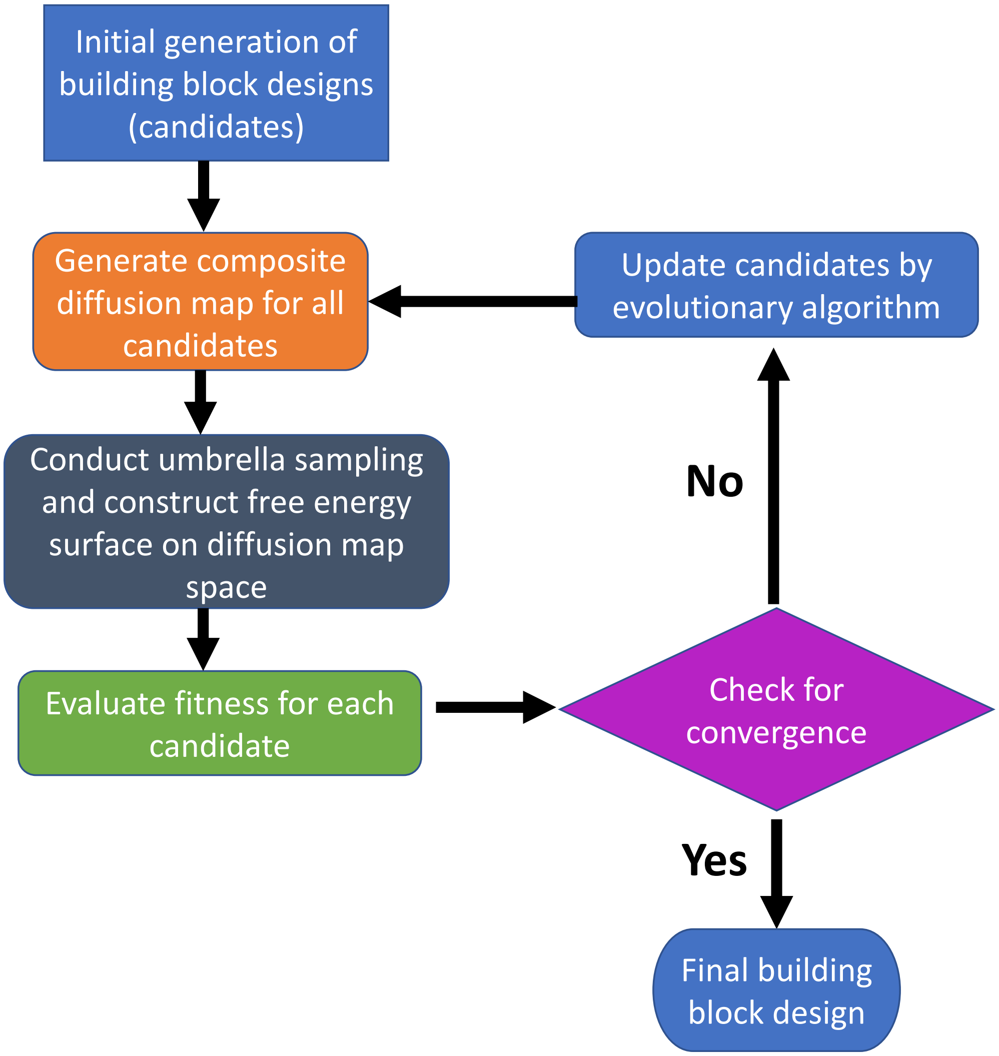

Initial generation of building block designs (candidates)
Generate composite diffusion map for all candidates
Update candidates by evolutionary algorithm
Conduct umbrella sampling and construct free energy surface on diffusion map space
No
Evaluate fitness for each candidate
Check for convergence
Yes
Final building block design

Figure 2.4: Flowchart of the landscape engineering inverse design procedure.

we temporarily ignore the “D” patches during this optimization. This approach is valid if the interaction strengths of the “D” patches are sufficiently weak compared to that of the “B” patches to be considered thermally decoupled. After the optimization is complete we place “D” patches on the opposite pole with parameters given by equations 2.3 and 2.4.

In case of the cubic diamond lattice, our goal is to optimize $\{E_D, E_E, \phi\}$ to favor dimer formation. During this optimization, we only put “D” and “E” patches on the central sphere and temporarily ignore the “B” patches. This assumption is a warranted if it can be assumed that the “B” and “D”,“E” interactions are thermally decoupled and the tetrahedra formed by the "B" interactions can be considered to be rock-like building blocks at the low-temperature at which "D" and "E"-mediated assembly into the cubic diamond lattice proceeds. After the optimization is complete, the "B" patches are added back on the opposite pole with $\{E_B, \phi_B\}$ taken from the pyrochlore optimization, and the interaction strengths of "D" and "E" patches are scaled down from $T^*_{\text{high}} = 0.8$ at which the optimization was conducted down to $T^*_{\text{low}} = 0.3$.

We now briefly discuss each step of the landscape engineering procedure (figure 2.4).

Langevin Dynamics Simulation

For each candidate, we need to estimate the corresponding accessible configurations. To do this, we employ Langevin dynamics simulation using HOOMD-blue[112,113]. For each simulation, we initialize the system with 64 randomly placed and oriented patchy colloids in a cubic simulation box with side length $L = 52.52\sigma$. Taking a patchy colloid to be a sphere with radius corresponding to the sum of the radii of the central particle and its surface patches, the corresponding volume fraction of patchy colloids in the system is $\varphi = 0.05$. We use Langevin dynamics integrator with $T^* = 0.8$ and step size $dt^* = 0.005$. We evolve the system for 3.5×10^7 steps and track the cluster formed by one randomly-selected tagged colloid every 3500 steps. This results in a total of 10^4 snapshots per simulation. We perform three independent Langevin dynamics simulations for each candidate.

Diffusion Maps

Diffusion map[48,114] is a widely-used nonlinear dimensionality reduction technique that has previously been applied to study of time evolution of molecular systems[49,50,59,115,116]. In the study of self-assembly process, it can provide a dynamically meaningful low-dimensional representation of the assembly process[59]. In our case, the diffusion map embeds the N self-assembled aggregates $\{x_i\}_{i=1}^{N}$ observed from the molecular simulations onto a low-dimensional manifold. The algorithm starts from constructing the Gaussian kernel matrix based on the pairwise distances between aggregates:

$$A_{ij} = e^{-\frac{d_{ij}^2}{2\epsilon}} \tag{2.6}$$

where d_{ij} is the pairwise distance between aggregate i and aggregate j, and ϵ is a Gaussian bandwidth. The definition of d_{ij} will be provided below. From this kernel matrix, a stochastic matrix representing the random walk over the data set is defined by row-normalizing the kernel matrix:

$$M_{ij} = \frac{A_{ij}}{\sum_j A_{ij}} \tag{2.7}$$

where M_{ij} may be interpreted as the probability of hopping from aggregate i to aggregate j in a time step $\Delta t = \epsilon$. The eigenvectors $\{\vec{\psi}_i\}_{i=1}^N$ of the stochastic matrix M are the discrete approximations of the eigenfunctions of the backward Fokker-Planck operator which describes a diffusion process over the data set[48,114]. The eigenfunctions associated with large eigenvalues describe the "slow" modes of the diffusion process, while the eigenfunctions associated with small eigenvalues describe the "fast" modes. The long-time behavior of the system is captured by top few eigenfunctions. Since the matrix M is Markovian, its top eigenvalue is $\lambda_1 = 1$ and the associated eigenvector is the trivial eigenvector $\vec{\psi}_1 = \vec{1}$[117]. The diffusion map nonlinear embedding into a d-dimensional space is achieved by projecting each self-assembled aggregate observed over the course of the simulations $\{x_i\}_{i=1}^N$ into the top d nontrivial eigenvectors:

$$x_i \rightarrow \begin{bmatrix} \vec{\psi}_2(i) & \vec{\psi}_3(i) & \dots & \vec{\psi}_{d+1}(i) \end{bmatrix}^T . \tag{2.8}$$

An appropriate choice of d is defined by a gap in the eigenvalue spectrum[59,101]. In all cases in the present work a gap was identified after the third eigenvalue λ_3 informing two-dimensional embeddings into the two leading non-trivial eigenvectors $\{\vec{\psi}_2, \vec{\psi}_3\}$.

The key to construct the diffusion map is to appropriately define the pairwise distance d_{ij} in equation 2.6. We need a way to compare the structural similarities between aggregates formed by the patchy colloids. To do this, we employ the graph-based approach described in ref[101], which is a modification of the Isorank algorithm[118]. This graph-based approach transforms the task of comparing structural similarities between aggregates into the task of comparing similarities between the graphs representing the aggregates. It first represents each aggregate by a graph G whose nodes correspond to the colloids within the aggregate and whose edges are weighted by the Euclidean distances between those colloids. In case of two aggregates with different number of colloids (i.e. two graphs with different number of nodes), the algorithm augments the smaller graph with $|N_i - N_j|$ ghost nodes. The algorithm then employs a greedy approach to find the pseudo-optimal alignment between two graphs by seeking an alignment H_{min} that minimizes the L_1 distance between two graphs G_i and G_j:

$$H_{\text{min}} = \arg\min \frac{\sum_{m,n} |(H^T G_i H)(m,n) - G_j(m,n)|}{|N_j|(|N_j| - 1)} \tag{2.9}$$

and the "distance" (i.e. structural similarity) between aggregate i and aggregate j is:

$$d_{ij} = \frac{\sum_{m,n} |(H_{\text{min}}^T G_i H_{\text{min}})(m,n) - G_j(m,n)|}{|N_j|(|N_j| - 1)}. \tag{2.10}$$

Importantly, this graph-based distance measure is invariant to rotation, translation, and particle permutation (i.e., particle relabeling) of the self-assembled aggregates. It serves as a good similarity metric between aggregates since it captures both the local fluctuation within clusters of the same shape (e.g. tetrahedra with small fluctuations of intra-cluster particle distance) and the global difference between clusters of different shapes (tetrahedra vs trimers).

During each Langevin dynamics simulation conducted for each candidate in a generation, we record the aggregates comprising a single tagged colloid. Then, we collect together the aggregates sampled from all such simulations for all candidates within a generation. From this group of aggregates, we construct a single composite diffusion map. By generating a composite diffusion map for all candidates in a generation, we obtain a unified low-dimensional embedding within which to construct and compare free energy surfaces.

Umbrella Sampling – Hamiltonian Monte Carlo

Having constructed the low-dimensional diffusion map space, we now need to construct the free energy surface for each candidate in terms of the diffusion map coordinates $\{\vec{\psi}_i\}_{i=2}^{d+1}$. To avoid the possible kinetic trapping due to potentially high free energy barrier, we combine umbrella sampling[109,110,119] with Hamiltonian Monte Carlo[120–122] to efficiently sample configuration space by applying biasing potentials within the collective variables (i.e., leading eigenvectors $\{\vec{\psi}_i\}$) determined by diffusion maps. We then reweight these data to estimate the unbiased free energy surfaces governing the self-assembly of each candidate building block. In brief, we tile the d-dimensional diffusion map embedding with harmonic biasing potentials:

$$W(\vec{\psi}, \vec{\psi}^*) = \frac{1}{2}(\vec{\psi} - \vec{\psi}^*)^T K(\vec{\psi} - \vec{\psi}^*) \tag{2.11}$$

where $\vec{\psi}^*$ is the d-dimensional harmonic center and K is a $d \times d$ dimensional diagonal matrix whose elements are the strengths of harmonic potential along each dimension. We conduct an independent biased simulation under each biasing potential and efficiently sample configuration space within the Hamiltonian Monte Carlo (HMC) framework using a NVE integrator to propose a trial move under the unbiased Hamiltonian. The volume-preserving and time-reversible properties of NVE integrator (e.g. Verlet or leapfrog algorithm) ensure that detailed balance could be maintained. The initial translational and angular momenta are drawn from the Maxwell-Boltzmann distribution. The acceptance probability of the trial move from the old state $\{\{q\}^{\text{old}}, \{p\}^{\text{old}}, \vec{\psi}^{\text{old}}\}$ to the new $\{\{q\}^{\text{new}}, \{p\}^{\text{new}}, \vec{\psi}^{\text{new}}\}$ is dictated by the Metropolis-Hasting criterion:

$$P_{\mathrm{acc}}(\mathrm{old} \rightarrow \mathrm{new}) = \min\left(1, \frac{e^{-\beta(U(\{q\}^{\mathrm{new}})+K(\{p\}^{\mathrm{new}})+W(\vec{\psi}^{\mathrm{new}},\vec{\psi}^{*}))}}{e^{-\beta(U(\{q\}^{\mathrm{old}})+K(\{p\}^{\mathrm{old}})+W(\vec{\psi}^{\mathrm{old}},\vec{\psi}^{*}))}}\right) \tag{2.12}$$

where $\beta = (k_B T)^{-1}$, $U(\{q\})$ is the potential energy associated with the particle positions $\{q\}$, $K(\{p\})$ is the kinetic energy associated with the particle velocities $\{p\}$, and $W(\vec{\psi}, \vec{\psi}^{*})$ is the artificial biasing potential defined by equation 2.11. Importantly, the HMC NVE trial move proposal does not require the calculation of biasing forces on the particles due to the artificial biasing potentials. The diffusion map does not provide an explicit differentiable expression for the collective variables as a function of the particle positions, meaning that analytical expressions for these biasing forces are unavailable. Should these forces be desired to perform, for example, biased molecular dynamics, techniques such as SandCV exist to estimate these forces by approximate interpolation and basis function expansions within the low-dimensional embedding[123], or the CVs themselves could be estimated using artificial neural network approaches such as MESA that provide the necessary derivatives through automatic differentiation[124–126].

We perform umbrella sampling simulations for each candidate around each harmonic biasing potential. In each case the system is initialized from the snapshot that is closest to the center of the harmonic biasing potential in the diffusion map embedding. Next, the aggregate formed by the tagged colloid is frozen, and the system is relaxed using the Fast Inertial Relaxation Engine[127] (FIRE) algorithm until the energy of the system converges within a tolerance of 0.1ε. During the relaxation, only WCA potential is enabled. After the relaxation, the aggregate formed by the tagged colloid is unfrozen and the full Hamiltonian comprising all Lennard-Jones and WCA potentials are enabled. During the first three generations of optimization for tetrahedron, we set the harmonic constant of the biasing potentials to be 2500ε. In later iterations we relax this to 25ε. Each Hamiltonian Monte Carlo loop is conducted at $T^{*} = 0.8$ and comprises 3500 steps of NVE integration with step size $dt^* = 0.005$. We perform 16000 Monte Carlo loops by equilibrating the system for the first 7000 loops and collecting data for the remaining 9000 loops. All molecular dynamics calculations are performed using HOOMD-blue[112,113]. After conducting umbrella sampling on the diffusion map space, we use BayesWHAM[128] algorithm to reconstruct the maximum a posteriori (MAP) estimate of the unbiased free energy surface for each candidate supported in the basis of the diffusion map collective variables by reweighting the biased umbrella sampling data.

Covariance Matrix Adaptation Evolution Strategy

Having constructed the free energy surface for each candidate in a generation, we then employ an objective function to define their relative fitnesses. The free energy surface is first coarse-grained by its inherent structures[129] by partitioning it into the basins of attraction for local free energy minima detected by steepest descent. The free energy of the inherent structure associated with the target self-assembled aggregate βF_{target} is compared with the lowest free energy inherent structure of a competitor aggregate $\beta F_{\text{competitor}}$. The fitness of each candidate colloidal building block is defined as the free energy gap:

$$\Delta\beta F = \beta F_{\text{target}} - \beta F_{\text{competitor}}. \tag{2.13}$$

Minimization of this objective function seeks to make the target structure the global free energy minimum on the self-assembly free energy surface and also open up a free energy gap between the nearest metastable competing structure. This topography can carry kinetic benefits in mitigating kinetically-trapped configurations and increasing both the yield and the rate of assembly of the target aggregate.

Having evaluated the fitness values for all candidates in a generation, we then propose new candidates by Covariance Matrix Adaptation Evolutionary Strategy[111](CMA-ES), which is a derivative-free algorithm for non-convex optimization problem. By stochastically seeding multiple walkers to probe the local topography based on running estimates of the local covariance matrix, CMA-ES has demonstrated good robustness and convergence rates on a variety of optimization problems and rugged landscapes[111,130]. Moreover, the CMA-ES could be formulated as a combination of natural gradient descent and step-size control on the expected fitness based on sampling distribution[131]. Based on the fitness values of candidates in generation g, the algorithm first selects top μ candidates. Next, based on these top μ candidates, the algorithm updates the estimate of covariance matrix C and step size σ. Then, it proposes a new generation $(g+1)$ of P candidates by:

$$\mathbf{x}^{(g+1)} = \langle \mathbf{x}^{g} \rangle_{\mu} + \sigma^{g} \mathcal{N}(0, C^{g}) \tag{2.14}$$

where $\mathbf{x}$ is the vector of design parameters characterizing a candidate, and $\langle \mathbf{x}^{g} \rangle_{\mu}$ is the mean value of the top μ candidates in generation g. We set $\mu = 3$, so CMA-ES will select top three candidates to propose the next generation. The optimization for the self-assembly of tetrahedral clusters for pyrochlore and cubic diamond proceeds in the two-dimensional design space $\mathbf{x} = [E_B,\ \phi_B]^T$, and for the self-assembly of the tetrahedral clusters into dimers required by the cubic diamond lattice in the three-dimensional space $\mathbf{x} = [E_D,\ E_E,\ \phi = \phi_D = \phi_E]^T$. $\mathcal{N}(0, C^{g})$ is a k dimensional multivariate Gaussian random vector with mean 0 and covariance matrix C^{g}, where k is the dimensionality of the design space. If the standard deviation of each parameter in $\mathbf{x}$ dips lower than 1 $k_B T$ at $T = 298$ K in the interaction strengths and 1° in the polar angle, we declare the CMA-ES to have converged and terminate the optimization. Otherwise, a new generation of candidates is proposed, and we repeat the whole optimization procedure for the new generation: conducting Langevin dynamics simulations for each new candidate, generating composite diffusion map for the new candidates, performing umbrella sampling on the composite diffusion map space, constructing the free energy surfaces and evaluating fitness values for each new candidate.

2.4 Results

We now proceed to describe our results for the inverse design of patchy colloids by landscape engineering to spontaneously nucleate and grow defect-free photonic lattices by a two-stage hierarchical assembly mechanism.

2.4.1 Inverse Design of Self-Assembling Pyrochlore Lattice

Optimization of Tetrahedral Aggregate Formation

We first apply landscape engineering to perform inverse design of patchy colloids to assemble pyrochlore lattice. As described in section 2.3.1, the pyrochlore lattice may be viewed as a tetrahedral network of corner-sharing tetrahedra. The optimization of the anisotropic patchy colloid design in figure 2.1 proceeds in the four-dimensional design space $\{E_B, \phi_B, E_D, \phi_D\}$ defining the polar angle and interaction strength of the more strongly interacting "B" patches on the north pole that mediate high-temperature assembly of monomers into tetrahedra at $T^*_{\text{high}} = 0.8$ and the more weakly interacting "D" patches on the south pole that direct assembly of the pre-assembled tetrahedra into the pyrochlore lattice at $T^*_{\text{low}} = 0.3$. By the symmetry of the design, we may first optimize "B" patches at a reduced temperature of $T^* = 0.8$ and then obtain the corresponding parameters for "D" patches by equations 2.3 and 2.4. Thus, we optimize $\{E_B, \phi_B\}$ at $T^* = 0.8$ and the target structure is the tetrahedron.

To initialize the optimization, we generate 10 initial candidates from a multivariate Gaussian distribution centered around (15.41ε, $30.0°$) with an initial covariance matrix of $C_0 = \text{diag}(5, 5)$ (i.e., a diagonal matrix with main diagonal vector (5,5)) and initial step size 1. This relatively large choice of initial covariance matrix and step size was made to favor early exploration of the design space. As detailed in section 2.3.2, for each candidate in each generation we perform unbiased Langevin dynamics simulations of assembly, construct composite diffusion maps, perform biased umbrella sampling – Hamiltonian Monte Carlo simulations, estimate self-assembly free energy landscapes, and evaluate the relative fitness of each candidate. The evolution of fitness values $\Delta\beta F$ and parameters $\{E_B, \phi_B\}$ as a function of generation is shown in figure 2.5. In the 16th generation, the parameters converge to $E_B = 15.54\varepsilon$ and $\phi_B = 30.44°$ within standard deviations of 1 k_BT at $T = 298$ K and 1°.

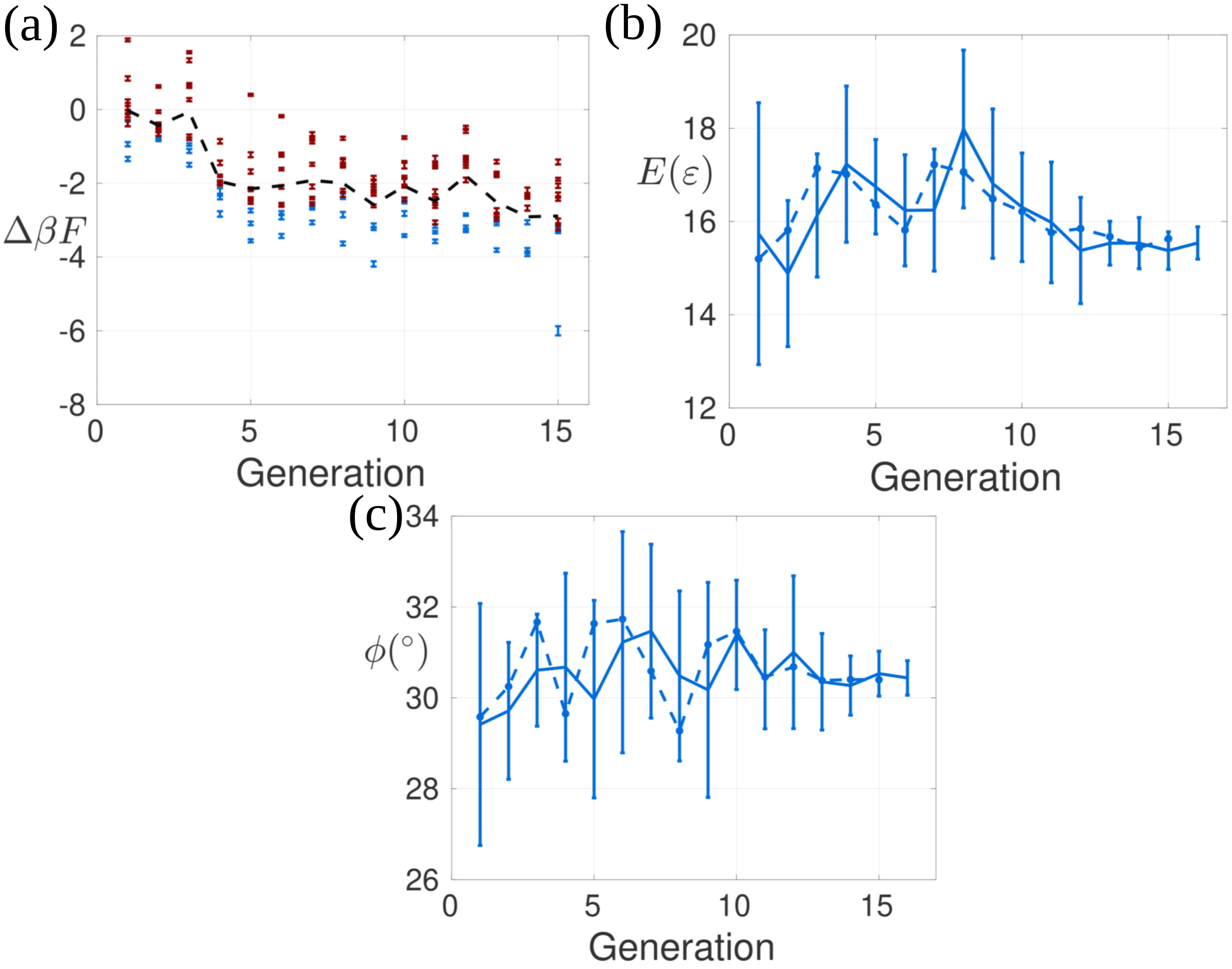


Figure 2.5: Landscape engineering of the pyrochlore patchy colloid. (a) The fitness values $\Delta\beta F$ for all candidates in each generation. Error bars are estimated from the standard deviation of the fitness value of each candidate. The blue points correspond to the μ = 3 best candidates selected by CMA-ES in each generation, and the red points to those less fit candidates that are discarded. The black dashed line corresponds to the boundary between them. Evolution of (b) interaction strength E_B and (c) polar angle ϕ_B of the "B" patches as a function of generation. The solid line corresponds to the mean value among all candidates in each generation, and the dashed line corresponds to the mean value of the μ = 3 best candidates in each generation. The optimization converges after 16 generations to $E_B = 15.54\varepsilon$ and $\phi_B = 30.44°$.

To see how the free energy surfaces change across generations, we select the best candidate from each generation and generate a composite diffusion map for all such candidates to provide a common set of collective variables that we can use to compare their free energy surfaces. The result is shown in figure 2.6. Here we compare the free energy surfaces of the best candidates in generations 1, 7 and 15. In particular, figure 2.6 (d)-(f) show the free energy surfaces of these candidates in the composite diffusion map space, and figure 2.6 (a)-(c) show the partition of design space into the Voronoi cells around the candidates in each of these generations. Panels (a)-(c) show that CMA-ES draws the initial distribution of candidates down into the optimum of the fitness landscape in $\Delta\beta F$ over the course of the 16-generation optimization. Panels (d)-(f) show that the free energy surface is sculpted such that the tetrahedron is preferentially stabilized with respect to all competitors. In the 1st generation the tetrahedron is the most stable aggregate but the monomer and dimer are also very stable, lying, respectively, just +1 k_BT and +2 k_BT higher in free energy. The trimer lies at +4 k_BT. In the 7th generation, the stability of the monomer relative to the tetrahedron is decreased to nearly +3 k_BT, but that of the dimer and trimer now lie at +2 k_BT. Finally in the 15th generation, the relative stabilities of the dimer and trimer are decreased to +2.6 k_BT and +4 k_BT, respectively, and the monomer lies at +2.5 k_BT, making the tetrahedron at least 2.5 k_BT more stable than all of its competitors. The net effect of the landscape engineering approach can be seen to have maximized the free energy gap (relative stability) between tetrahedron and all competing aggregates.

High-Temperature Assembly of Tetrahedra

Landscape engineering discovers the optimized parameters for the "B" patches of $E_B = 15.54\varepsilon$ and $\phi_B = 30.44°$. We now proceed to verify that this design leads to the self-assembly of tetrahedral aggregates in high yield. We perform four independent unbiased Langevin dynamics simulations at $T^* = 0.8$ for 2×10^6 reduced time units for patchy colloids

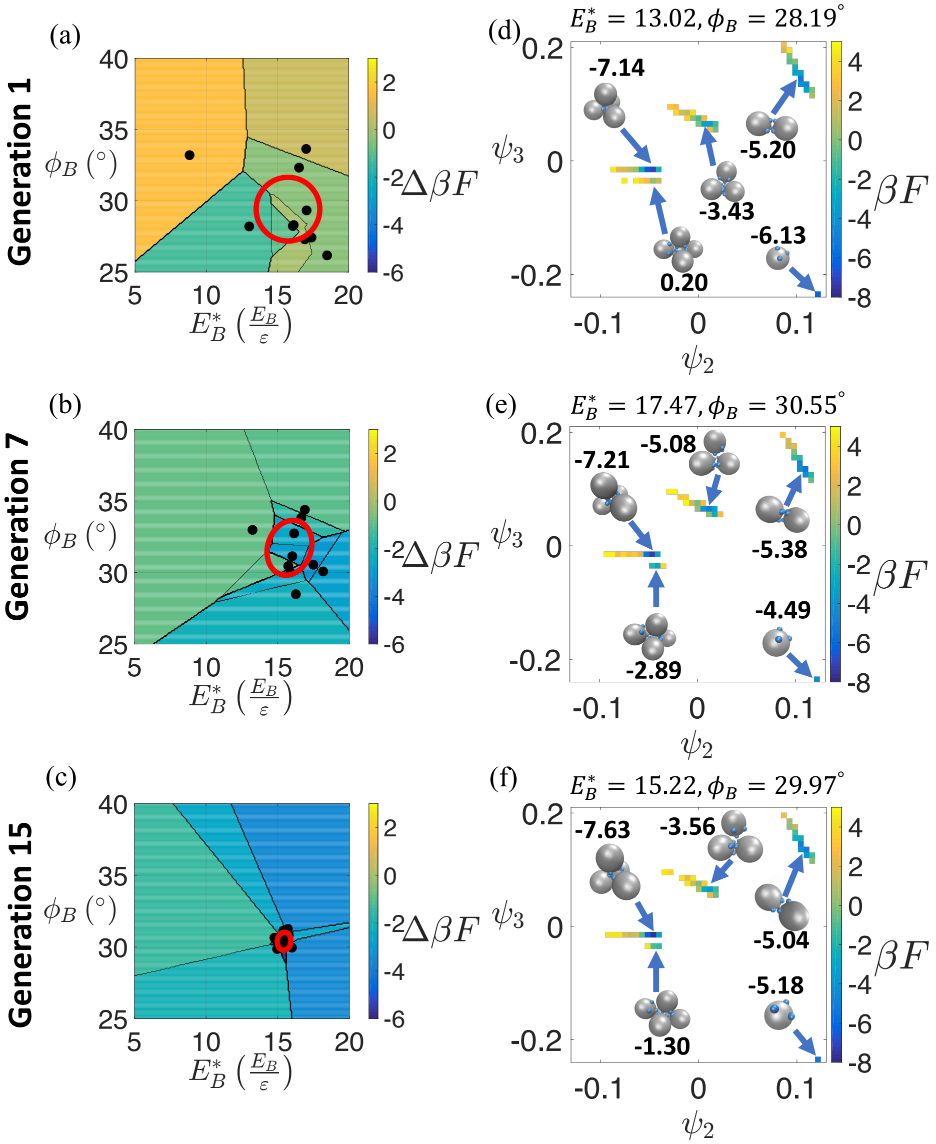


Figure 2.6: Landscape engineering sculpting of the self-assembly free energy landscape for tetrahedral cluster formation. (a)-(c) Distribution of candidates within the $\{E_B, \phi_B\}$ design space in generations 1, 7 and 15. The candidates are represented by black dots. The red circle represents the CMA-ES covariance matrix from which the candidates in the current generation are sampled. For visualization purposes, we partition design space into Voronoi cells around each candidate and color each cell by the fitness $\Delta\beta F$ of the corresponding candidate. (d)-(f) Free energy surfaces of the best candidates in generations 1, 7 and 15 in the composite diffusion map space spanned by the leading two diffusion map collective variables $\{\psi_2, \psi_3\}$. The particular values of $\{E_B, \phi_B\}$ pertaining to each candidate are listed above each panel. Representative aggregates from the local free energy minima are projected onto the low-dimensional embedding. The values of the local free energy minimum associated with each aggregate are displayed next to the representative structures.

decorated with "B" patches employing the optimal design parameters. All simulations are initialized with 512 randomly placed and oriented particles in a cubic box of side length $L = 105.54\sigma$, corresponding to a volume fraction of $\varphi = 0.05$. The solid colored lines in figure 2.7 show the temporal yield of tetrahedral aggregates as a function of time. Assuming monomers are depleted according to simple first-order kinetics, we can fit an expression for the tetrahedral yield of form $y(t) = b\left(1 - e^{-kt}\right)$, where $y(t)$ is the fraction of colloids residing within tetrahedra, t is time, $b = (96.2 \pm 0.3)\%$ is the equilibrium fraction of colloids forming tetrahedral clusters, and $k = (63.4 \pm 6.7)\ \mathrm{s}^{-1}$ is the best-fit first-order rate constant.

It is also instructive to compare the assembly kinetics to that for a patchy colloid design employing the same interaction strength but an empirical patch angle based on the tetrahedral geometry. A polar angle of $\phi_B = 35.26°$ corresponds to the case where the attractive patches point directly towards the neighboring particles within an ideal tetrahedral aggregate. This can be considered the patch angle arising from zero-temperature energy minimization of an isolated tetrahedral cluster[60,132]. We note that direct application of the landscape engineering approach at $T = 0$ K may present challenges in sampling the configurational energy landscape at absolute zero and would require the use of an alternative sampling technique to molecular dynamics such as simulated annealing or basin hopping. The dashed lines in figure 2.7 present the tetrahedral yield for particles with "B" patches of $E_B = 15.54\varepsilon$ and $\phi_B = 35.26°$. Fitting of the first-order kinetic model yields values of $b = (87.2 \pm 2.3)\,\%$ and $k = (38.6 \pm 6.0)\ \mathrm{s}^{-1}$, demonstrating that the optimal design discovered by landscape engineering exhibits both higher asymptotic yield and faster assembly kinetics. Analysis of the simulation trajectories shows that the $\sim 5°$ larger polar angle for the empirical geometric design results in the formation of many clusters larger than tetrahedra. This can be understood as the larger polar angle of the patches enabling promiscuous interactions between the particles comprising a tetrahedral cluster and outsider particles, whereas the smaller polar angle optimized through landscape engineering disfavors the formation of

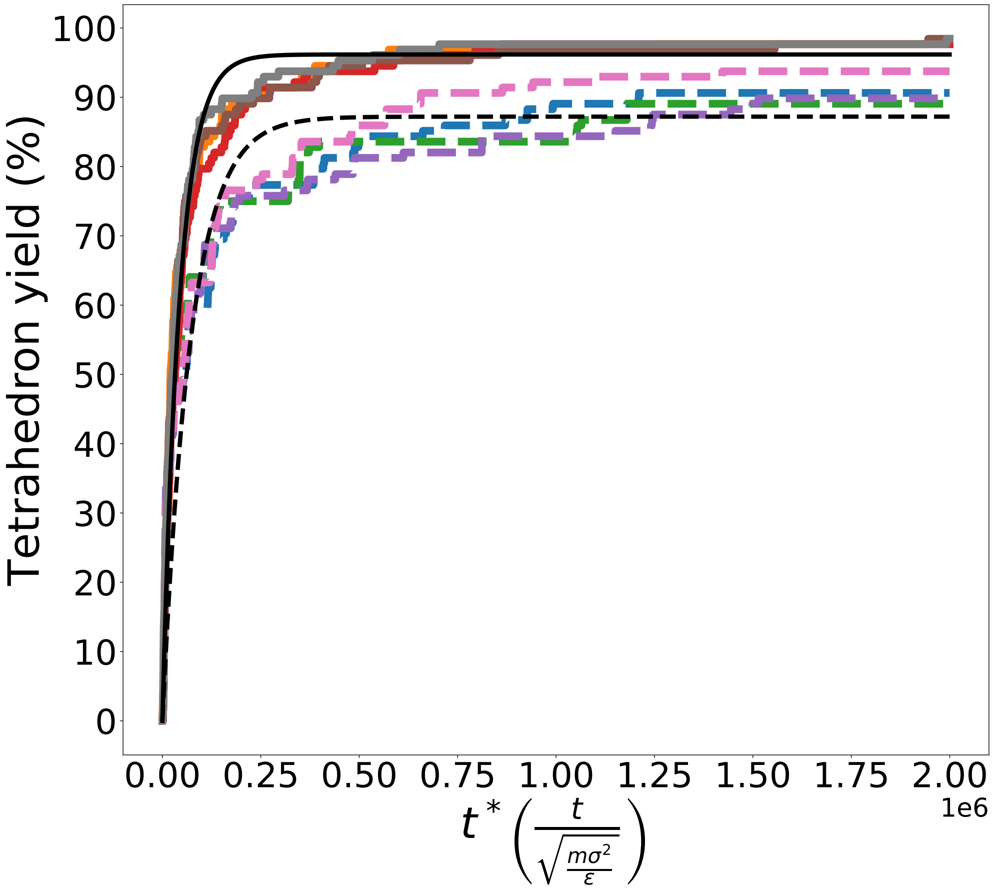


Figure 2.7: Yield of tetrahedral clusters as a function of simulation time at $T^* = 0.8$ in unbiased Langevin dynamics simulations. Each colored line corresponds to an independent simulation. The four colored solid lines correspond to patchy colloids decorated with "B" patches of the optimal design $\{E_B = 15.54\varepsilon, \phi_B = 30.44°\}$ deduced by landscape engineering. The four colored dashed lines correspond to "B" patches with $\{E_B = 15.54\varepsilon, \phi_B = 35.26°\}$ employing the same interaction strength but a polar angle corresponding to the zero-temperature energy minimum of an isolated tetrahedral cluster. The black solid and dashed lines are fits of the corresponding data to first-order kinetics. Landscape engineering discovers an improved particle design exhibiting better yield and assembly rate beyond that derived from purely geometric considerations.

these large aggregates to improve assembly rate and yield.

The landscape engineering optimization was conducted at a volume fraction $\varphi = 0.05$, but it is of interest to assess the robustness of this design in mediating high-yield tetrahedral assembly at other volume fractions. Langevin dynamics simulations conducted at volume fractions over the range $\varphi = 0.025$-0.1 reveal the tetrahedral yield to remain stable and high at 95% or better for up to two-fold increases and decreases in the volume fraction away from that at which the optimization was conducted. Very high volume fractions risk trapping within kinetically arrested glassy states, whereas very low volume fractions introduce strong entropic driving forces disfavoring assembly. At either of these extremes we anticipate that re-optimization under the volume fraction of interest would be required to maintain high tetrahedral yields

Two-Stage Hierarchical Assembly of Pyrochlore Lattice

After obtaining $\{E_B = 15.54\varepsilon, \phi_B = 30.44°\}$ as the optimal design parameters for "B" patches, we obtain the optimal design for the "D" patches according to equations 2.3 and 2.4:

$$\begin{aligned} &\phi_D = \phi_B = 30.44° \\ &E_D = \frac{T^*_{\text{low}}}{T^*_{\text{high}}} E_B = \frac{0.3}{0.8} \times 15.54\varepsilon = 5.83\varepsilon. \end{aligned} \tag{2.15}$$

We then decorate the patchy colloids with the optimal north pole "B" patches and south pole "D" patches to arrive at the final landscape engineering design of the patchy colloids. We validate the capacity of the design to achieve two-stage hierarchical assembly of pyrochlore lattice by locating 512 randomly placed and oriented patchy colloids in a cubic simulation box of side length $L = 105.04\sigma$, corresponding to a volume fraction of $\varphi = 0.05$. The first stage of assembly proceeds by a high-temperature hold at which the system is evolved at $T^*_{\text{high}} = 0.8$ for 2×10^6 reduced time units to allow for the formation of tetrahedral clusters from the colloidal monomers. The second stage of assembly is effected by a two-stage cooling

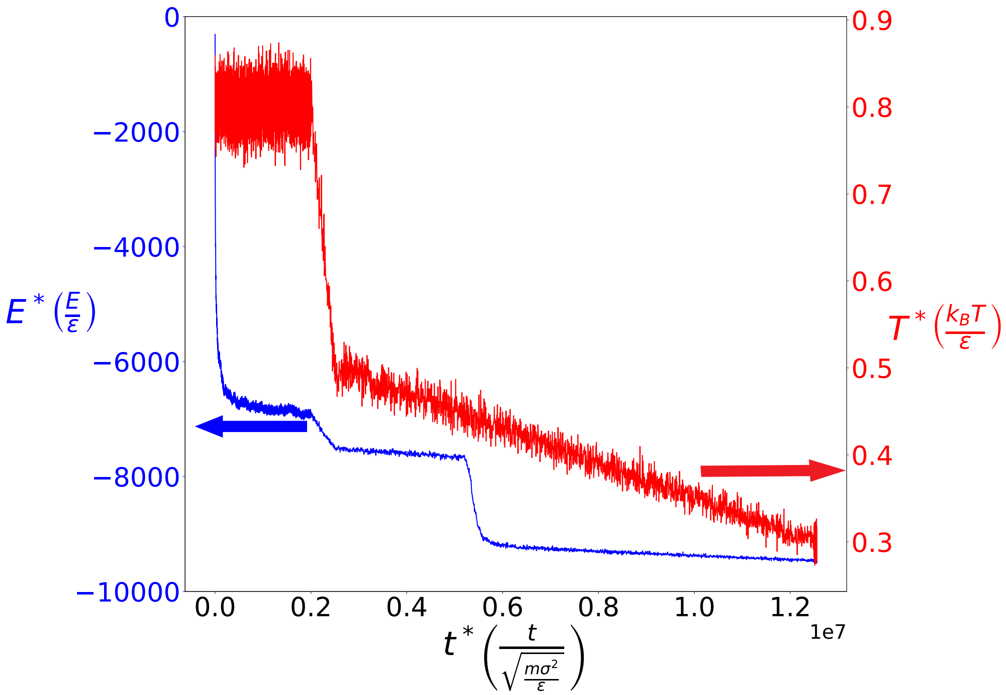


Figure 2.8: Evolution of system potential energy and temperature for two-stage hierarchical assembly of pyrochlore lattice. The two horizontal arrows indicate which of the two axes – potential energy or temperature – pertain to each curve on this double y-axis plot.

protocol to favor nucleation of the pyrochlore lattice whereby the system is rapidly cooled from $T^*_{\text{high}} = 0.8$ to $T^*_{\text{intermediate}} = 0.5$ for 5×10^5 reduced time units and then slowly cooled from $T^*_{\text{intermediate}} = 0.5$ to $T^*_{\text{low}} = 0.3$ for 1×10^7 reduced time units. Finally, the system is subjected to a low-temperature hold at $T^*_{\text{low}} = 0.3$ for another 5×10^4 reduced time units to gather statistics on the terminal crystal. The evolution of system potential energy and temperature is presented in figure 2.8. Nucleation of the pyrochlore lattice occurs during the second slow cooling phase at around $T^* = 0.45$ as indicated by the sudden drop in potential energy corresponding to the latent heat of crystallization.

At the end of the $T^*_{\text{high}} = 0.8$ high-temperature assembly stage the yield of tetrahedral clusters is 97.7% corresponding to the formation of 125 tetrahedral clusters mediated by interactions between the north pole "B" patches. The radial distribution function $g(r)$ between the geometric centers of the tetrahedral clusters demonstrates that they behave

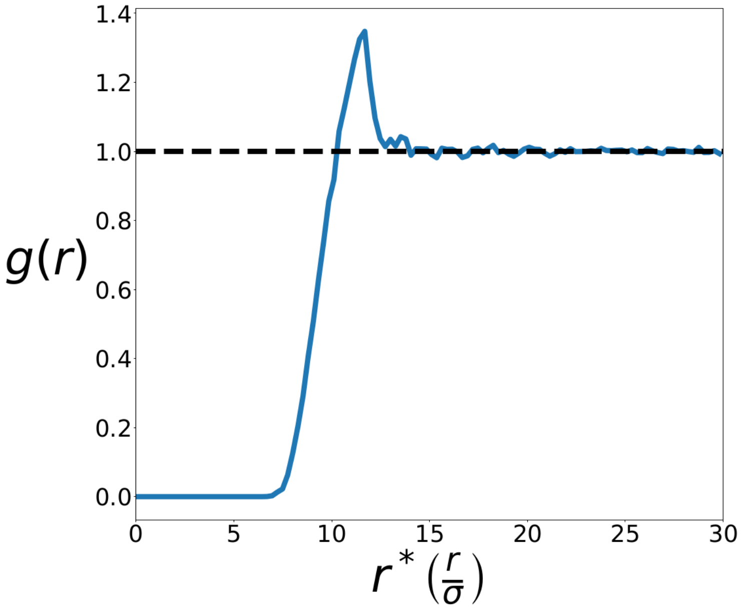


Figure 2.9: Radial distribution function between the geometric centers of tetrahedral clusters at the end of the high-temperature assembly stage of pyrochlore lattice.

effectively as an ideal gas (figure 2.9). The small correlation peak at $r^* \approx 12$ indicating the presence of very weak structural correlations between the tetrahedra is largely attributable to the weak interactions between the south pole "D" patches, but numerical simulations in which the "D"-"D" interactions are turned off show that ~20% of the correlation peak can be attributed to effective entropic attractions driven by excluded volume interactions. This validates the design expectation that interactions between "D" patches should be thermally decoupled from that of the "B" patches.

During the slow cooling process we observe nucleation and subsequent growth of a pyrochlore lattice (figure 2.10a). We note that sufficiently slow cooling rates are necessary to assure a single nucleation event and production of defect-free crystal. In experimental realizations employing orders of magnitude more colloids than our simulations it can be quite challenging to achieve defect-free crystals. We anticipate that very slow cooling rates, possibly coupled with programmed temperature oscillations to heal defects, may be required to obtain high-fidelity periodic crystal lattices. At the end of the $T^*_{\text{low}} = 0.3$ low-temperature hold, we perform structural characterization of the crystal. To do so we compute the radial distribution function $g(r)$ between the centers of mass of patchy colloids (figure 2.10b) and Steinhardt bond order parameters[133,134] $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ (figure 2.10c) and $q_4(i)$ (figure 2.10d), where "$*$" denotes the complex conjugate. The complex vector $\vec{q_l}(i)$ is a $(2l+1)$ dimensional vector whose non-normalized elements are:

$$q_{lm}(i) = \frac{1}{N_b(i)} \sum_{k=1}^{N_b(i)} Y_{lm}(\hat{r}_{ik}) = \frac{1}{N_b(i)} \sum_{k=1}^{N_b(i)} Y_{lm}(\theta_{ik}, \phi_{ik}) \tag{2.16}$$

where $N_b(i)$ is the number of nearest neighbors of particle i, k loops over all such nearest neighbors, $\hat{r}_{ik}$ is the unit displacement vector from particle i to particle k, $\{\theta_{ik}, \phi_{ik}\}$ are the polar and azimuthal angles that $\hat{r}_{ik}$ makes with respect to a specific coordinate system, and Y_{lm} are the spherical harmonics. A nearest neighbor is defined as a particle lying within a cutoff distance $d_{\text{cut}} = 6.0\sigma$, where this threshold is calibrated to cover the first peak in the radial distribution function at $r_{\text{peak},1} \approx 5.25\sigma$ (figure 2.10b). When computing the inner product $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$, we normalize each vector to have unit l^2-norm. It can be shown that the inner product, $\vec{q_l}^*(i) \cdot \vec{q_l}(j)$, between two particles i and j is real and independent of the coordinate system. An outline of the proof is given here. By the addition theorem of spherical harmonics, given two unit vectors $\hat{r} = (\theta, \phi)$ and $\hat{r}' = (\theta', \phi')$, we have the following identity:

$$\sum_{m=-l}^{l} Y_{lm}(\theta, \phi) Y_{lm}^*(\theta', \phi') = \frac{2l+1}{4\pi} P_l(\cos\gamma) \tag{2.17}$$

where γ is the angle between $\hat{r}$ and $\hat{r}'$ and P_l is the Legendre polynomial. Let us choose some arbitrary coordinate system, and focus on two particles i and j. The complex inner product is evaluated as:

$$\begin{aligned}
\vec{q}_l^*(i) \cdot \vec{q}_l(j) &= \sum_{m=-l}^{l} q_{lm}^*(i) q_{lm}(j) \\
&= \frac{1}{N_b(i)N_b(j)} \sum_{m=-l}^{l} \sum_{k=1}^{N_b(i)} \sum_{k'=1}^{N_b(j)} Y_{lm}(\theta_{ik}, \phi_{ik})^* Y_{lm}(\theta_{jk'}, \phi_{jk'}) \\
&= \frac{1}{N_b(i)N_b(j)} \sum_{k=1}^{N_b(i)} \sum_{k'=1}^{N_b(j)} \sum_{m=-l}^{l} Y_{lm}(\theta_{ik}, \phi_{ik})^* Y_{lm}(\theta_{jk'}, \phi_{jk'}) \\
&\propto \sum_{k=1}^{N_b(i)} \sum_{k'=1}^{N_b(j)} P_l(\cos \gamma_{\{ik,jk'\}})
\end{aligned} \tag{2.18}$$

where $\gamma_{\{ik,jk'\}}$ is the angle between $\hat{r}_{ik}$ and $\hat{r}_{jk'}$. Since these angles are independent of the coordinate system, we have proved that the inner product is real and independent of the coordinate system. The parameter $q_l(i)$ is defined using the relation:

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2} \tag{2.19}$$

The parameter $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ defined between nearest neighbor pairs $\{i, j\}$ has been shown to be able to distinguish between pyrochlore lattice and hexagonal tetrastack lattice[90] by aggregating bond angle information between pairs of particles separated by up to three bonds. The parameter $q_4(i)$ provides a more localized structural characterization by averaging over only the nearest neighbors of particle i and providing a way to tell whether a particle exists in a locally staggered or eclipsed configuration. Taken together $g(r)$, $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$, and $q_4(i)$ allow us to determine whether the system adopts the radial and angular order expected for a pure pyrochlore lattice, and whether it contains crystal defects or is a mixture of the pyrochlore and hexagonal tetrastack polymorphs.

The radial distribution function computed over the final snapshot of the simulation at $T^*_{\text{low}} = 0.3$ possesses exactly the characteristic peaks of a pyrochlore lattice (figure 2.10b).

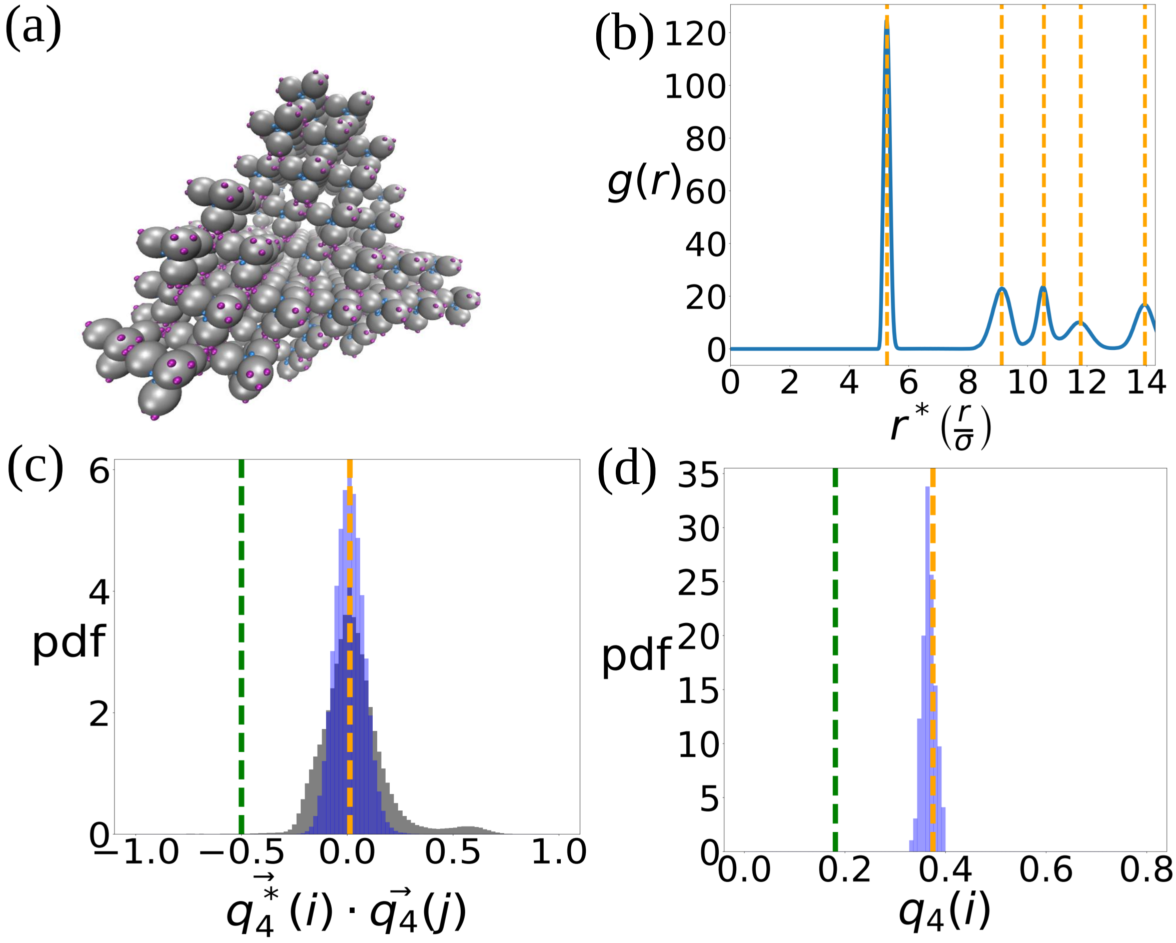


Figure 2.10: Structural characterization of the self-assembled pyrochlore lattice. (a) Snapshot of terminal crystal lattice structure. (b) Radial distribution function $g(r)$ between the patchy colloid centers of mass. (c) The distribution of $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ computed between all patchy colloid nearest-neighbor pairs (gray) and restricted to crystalline colloid pairs (blue) defined as those in which each partner possesses six bonded nearest neighbors. (d) The distribution of $q_4(i)$ for crystalline colloids. In all panels the orange dashed lines represent the expected peak positions for an ideal pyrochlore lattice and the green dashed lines represent those for an ideal hexagonal tetrastack lattice.

The distributions of $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ (figure 2.10c) and $q_4(i)$ (figure 2.10d) also both possess peaks at the positions expected for the ideal pyrochlore lattice, and lack those expected for the hexagonal tetrastack. Specifically, the pyrochlore lattice will possess only one peak at 0.0123 in $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ and one peak at 0.375 in $q_4(i)$[90]. In contrast, the hexagonal tetrastack will show two peaks at 0.0123 and -0.5 in $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ and two peaks at 0.375 and 0.181 in $q_4(i)$[90]. The gray bars in figure 2.10c show the distribution of $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ for all nearest neighbor pairs of patchy colloids. Due to finite-size effects, in some nearest neighbor pairs the constituent colloids lie on the boundary of the crystal structure and do not have exactly six bonded nearest neighbors. These finite-size effects cause the second small peak at $\vec{q_4}^*(i) \cdot \vec{q_4}(j) \approx 0.5$ to emerge. The blue transparent bars show the distribution $\vec{q_4}^*(i) \cdot \vec{q_4}(j)$ among nearest neighbor pairs in which both colloids are constrained to have six bonded nearest neighbors (i.e., crystalline colloids). Here we see that the blue transparent bars are indeed centered around the characteristic peak in ideal infinite pyrochlore lattice and the small shoulder disappears. Similarly, the distribution of $q_4(i)$ in figure 2.10d restricted to six-neighbor crystalline colloids possesses a single peak centered on the pyrochlore result. Taken together, figure 2.10 demonstrates that the colloids do spontaneously form a defect-free pyrochlore lattice, and not a mixture of the pyrochlore and hexagonal tetrastack polymorphs.

Having characterized the final structure, we then proceed to compute the band structure of the corresponding periodic crystal using the MIT Photonic Bands (MPB) software[135]. We create an infinite periodic pyrochlore lattice from the primitive lattice vectors $\vec{R_1} = (0, \frac{a}{2}, \frac{a}{2}), \vec{R_2} = (\frac{a}{2}, 0, \frac{a}{2}), \vec{R_3} = (\frac{a}{2}, \frac{a}{2}, 0)$ where a is the lattice constant. The positions of the four basis particles reported in the basis of the primitive lattice vectors (i.e., (l,m,n) denotes a position vector $l\vec{R_1} + m\vec{R_2} + n\vec{R_3}$) are given in table 2.1. The lattice constant a is related to the first peak in radial distribution function (nearest-neighbor distance) $r_{\text{peak},1}$ by $a = \frac{4}{\sqrt{2}} r_{\text{peak},1}$. We estimate the nearest-neighbor distance from the radial distribution function of final configuration (figure 2.10b) to be $r_{\text{peak},1} \approx 5.25\sigma$. The radius of the colloidal

Basis particle index	Position
1	$\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$
2	$\left(0, \frac{1}{2}, \frac{1}{2}\right)$
3	$\left(\frac{1}{2}, 0, \frac{1}{2}\right)$
4	$\left(\frac{1}{2}, \frac{1}{2}, 0\right)$

Table 2.1: Positions of basis particles of the pyrochlore lattice in the basis of primitive lattice vectors $\vec{R_1} = (0, \frac{a}{2}, \frac{a}{2}), \vec{R_2} = (\frac{a}{2}, 0, \frac{a}{2}), \vec{R_3} = (\frac{a}{2}, \frac{a}{2}, 0)$ where a is the lattice constant. The tuple (l,m,n) denotes a particle position vector $l\vec{R_1} + m\vec{R_2} + n\vec{R_3}$.

particles is that of "A" spheres $\sigma_A/2 = 2.5\sigma$ and the dielectric constant is set to $\epsilon_r = 12.0$ corresponding to the value for silicon. The medium is taken to be air. We use a $16 \times 16 \times 16$ grid to discretize the primitive unit cell to compute the band structure along the high-symmetry lines in the first Brillouin zone. We have verified that our results have converged with respect to the grid spacing. The resulting photonic band structure is shown in figure 2.11. The band structure shows the opening of an indirect bandgap between the second and third bands with a width-to-midgap ratio (ratio between the bandgap width and the midgap frequency) of 4.63%.

2.4.2 Inverse Design of Self-Assembling Cubic Diamond Lattice

Optimization of Dimer Formation

Following the success in pyrochlore assembly, we then apply our landscape engineering approach to design a new patchy colloid to assemble cubic diamond lattice via tetrahedral clusters. As described in section 2.3.1 and illustrated in figure 2.3, the cubic diamond lattice comprises tetrahedral clusters arranged in staggered dimers. The high-temperature assembly of patchy colloids into tetrahedra proceeds in exactly the same fashion as for pyrochlore, and we adopt $E_B = 15.54\varepsilon$ and $\phi_B = 30.44°$ as the optimal design solution for the "B" patches. The design problem then reduces to optimization of the interaction strengths and polar an-

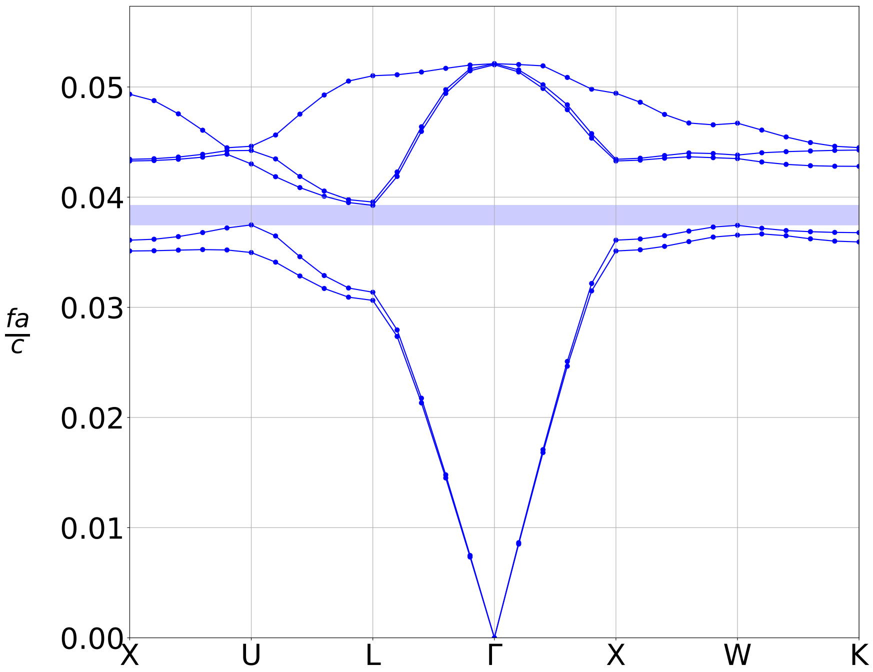


Figure 2.11: Photonic band structure of the self-assembled pyrochlore lattice. The y-axis reports the dimensionless frequency $\frac{fa}{c}$, where f is the frequency, a is lattice constant, and c is the speed of light in vacuum. The x-axis labels the corners of the irreducible region of first Brillouin zone in canonical order. For a lattice constant of $a = 2.97$ μm corresponding to our choice of length scale, the bandgap lies within the frequency range of $3.74 < f < 3.94$ THz and wavelength range of $76.2 < \lambda < 80.3$ μm, placing the bandgap in the infrared regime of electromagnetic spectrum. For a lattice constant of $a = 27.0$ nm the bandgap lies within the frequency range of $411 < f < 433$ THz and wavelength range of $692 < \lambda < 729$ nm, placing the bandgap in the visible regime.

gles of the south pole "D" and "E" patches to mediate the low-temperature assembly of the pre-assembled tetrahedral aggregates into cubic diamond lattice. We choose to optimize the parameters for "D" and "E" patches at $T^*_{\text{high}} = 0.8$, and then scale down the interaction strengths to match the low temperature phase $T^*_{\text{low}} = 0.3$ by a factor of $T^*_{\text{low}}/T^*_{\text{high}}$. Thus, we optimize $\{E_D, E_E, \phi = \phi_D = \phi_E\}$ at $T^* = 0.8$ and the target structure is a staggered dimer.

We initialize the optimization by generating 10 initial candidates from a multivariate Gaussian distribution centered around $(6.67\varepsilon, 6.67\varepsilon, 26.60°)$ with an initial covariance matrix of $C_0 = \text{diag}(5, 5, 5)$ and an initial step size 1. The evolution of $\{E_D, E_E, \phi\}$ and the fitness

$\Delta\beta F$ over the landscape engineering generations are shown in figure 2.12. In the 18th generation, all parameters have converged to $E_D = 10.02\varepsilon$, $E_E = 11.64\varepsilon$ and $\phi = 26.68°$ within standard deviations of 1 $k_B T$ at $T = 298$ K and 1°.

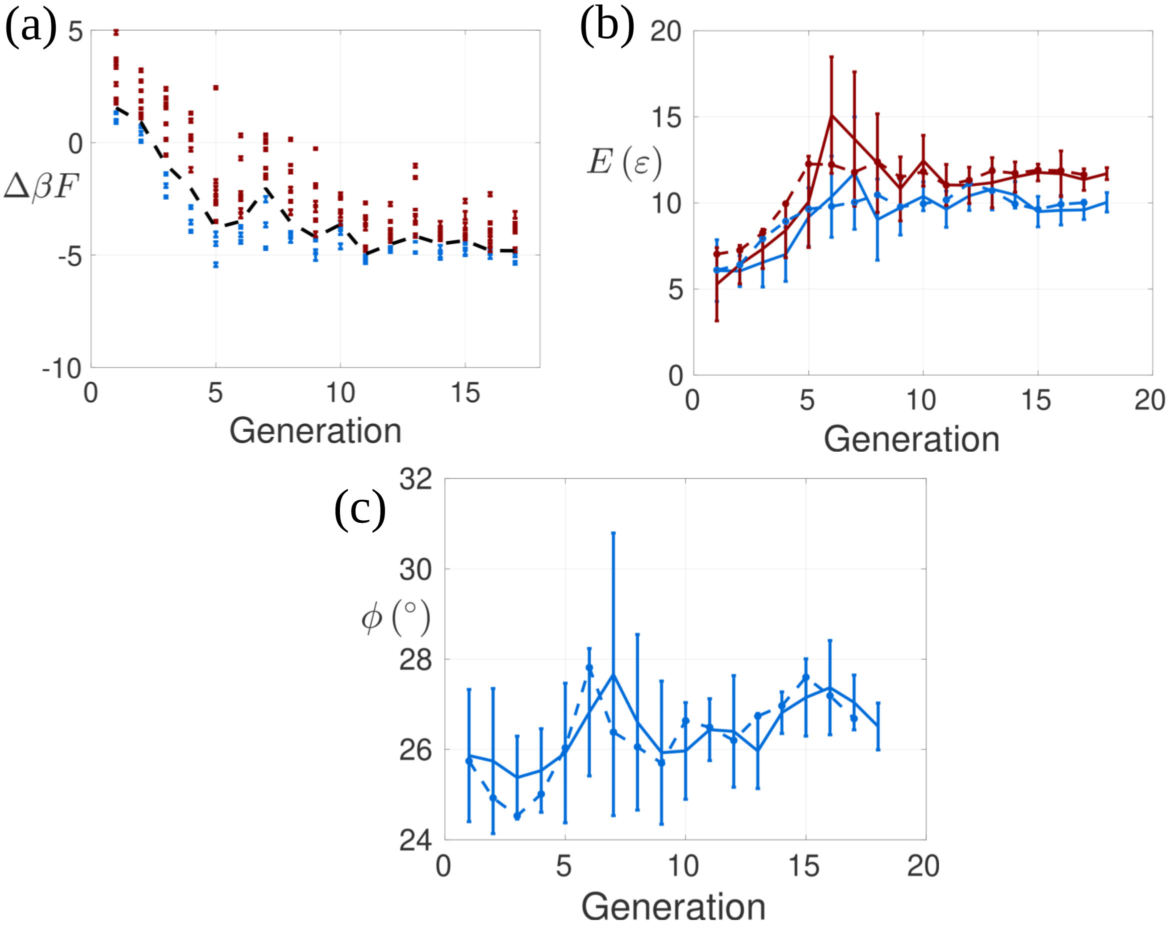
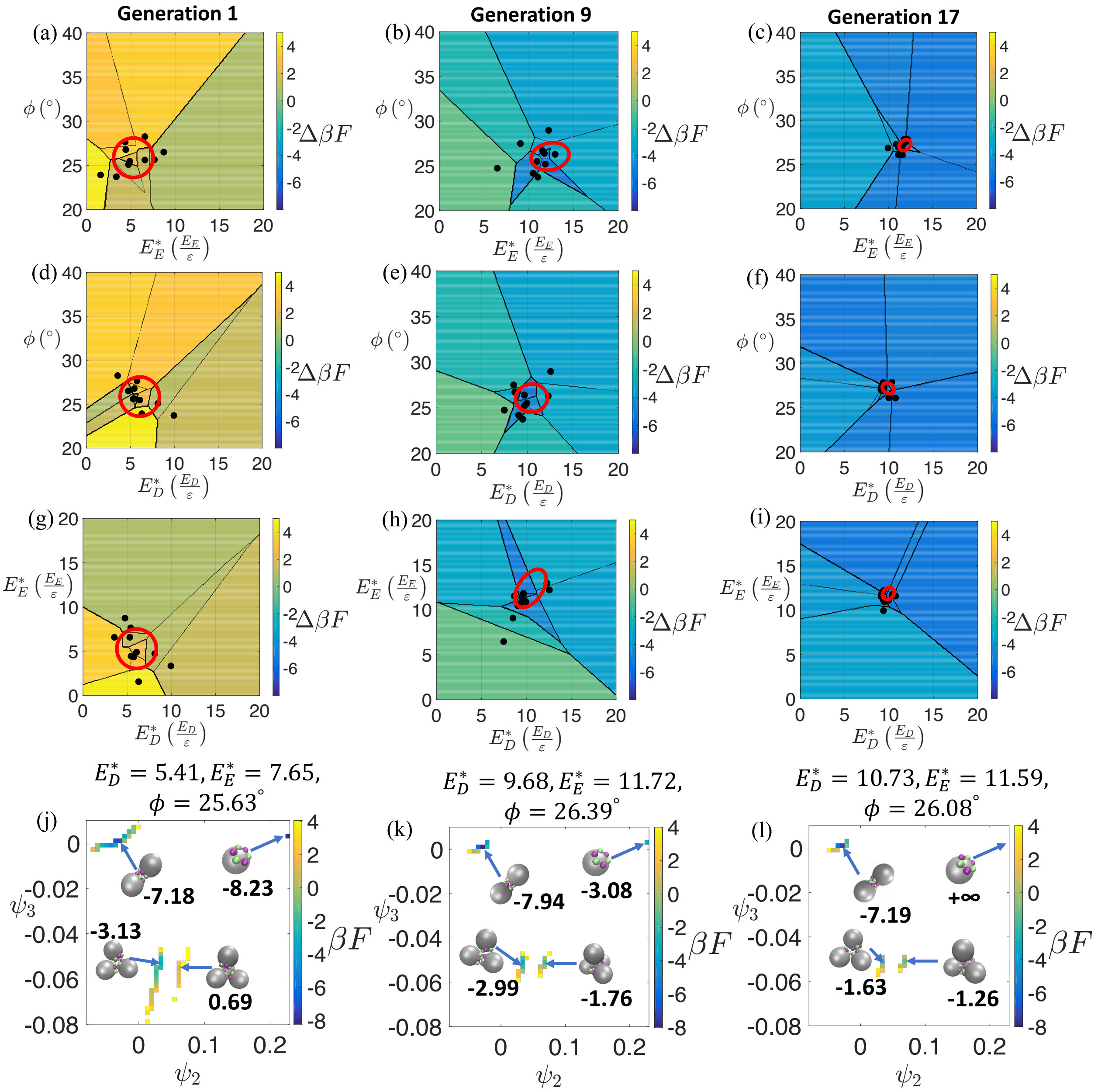


Figure 2.12: Landscape engineering of the cubic diamond patchy colloid. (a) The fitness values $\Delta\beta F$ for all candidates in each generation. Error bars are estimated from the standard deviations in the fitness values corresponding to each candidate. The blue points correspond to the $\mu = 3$ best candidates selected by CMA-ES in each generation, and the red points to those less fit candidates that are discarded. The black dashed line corresponds to the boundary between them. Evolution of (b) interaction strengths E_D (blue) and E_E (red) and (c) polar angle $\phi = \phi_D = \phi_E$ as a function of generation. The solid line corresponds to the mean value among all candidates in each generation, and the dashed line corresponds to the mean value of the $\mu = 3$ best candidates in each generation. The optimization converges after 18 generations to $E_D = 10.02\varepsilon$, $E_E = 11.64\varepsilon$ and $\phi = 26.68°$.

The distribution of candidates within the design space and free energy surfaces for the best candidates in generations 1, 9 and 17 are presented in figure 2.13 to show how landscape engineering changes the design and assembly properties of the building block over the course of the optimization. In figure 2.13a-i we partition the design spaces $\{\phi, E_E\}$, $\{\phi, E_D\}$ and $\{E_E, E_D\}$ by the Voronoi cells around the candidates in generations 1, 9 and 17, and we color each Voronoi cell by the fitness value of the corresponding candidate. Despite the relatively poor initial guesses for E_D and E_E, CMA-ES was able to efficiently move the mean and shrink the variance of subsequent generations of candidates to converge to the optimum of the $\Delta\beta F$ fitness landscape. In figure 2.13j-l we show that the self-assembly free energy surfaces are driven towards a topography in which the staggered dimer is preferentially stabilized relative to all competing aggregates. In the 1st generation, the monomer is the most stable aggregate lying (-1) k_BT lower in free energy than the dimer. The trimer and tetrahedron are each less stable than the dimer, lying, respectively, +4 k_BT and +8 k_BT higher in free energy. In the 9th generation, the landscape engineering protocol has successfully rendered the dimer the most stable aggregate on the landscape, with the monomer, trimer, and tetrahedron lying, respectively, +5 k_BT, +5 k_BT, and +6 k_BT higher in free energy. In the 17th generation, the dimer has been even further stabilized, with the trimer and tetrahedron each lying +6 k_BT higher in free energy, and the monomer rendered completely unstable within the sampling resolution of our simulations.

High-Temperature Assembly of Dimers

We verify the optimal landscape engineering design of $E_D = 10.02\varepsilon$, $E_E = 11.64\varepsilon$ and $\phi = 26.68^{\circ}$ by performing four Langevin dynamics simulations at $T^* = 0.8$ for 2×10^6 reduced time units for patchy colloids decorated with “D” and “E” patches. Simulations are initialized with 512 colloidal monomers with random positions and orientations in a cubic simulation box with side length $L = 105.04\sigma$ corresponding to a volume fraction of $\varphi = 0.05$. The yield of staggered dimers as a function of time for the four runs is presented in figure



$E_D^* = 5.41, E_E^* = 7.65,$
$\phi = 25.63^\circ$

$E_D^* = 9.68, E_E^* = 11.72,$
$\phi = 26.39^\circ$

$E_D^* = 10.73, E_E^* = 11.59,$
$\phi = 26.08^\circ$

Figure 2.13: Landscape engineering sculpting of the self-assembly free energy landscape for the formation of staggered dimers. (a)-(i) Distribution of candidates within the $\{\phi, E_E\}$, $\{\phi, E_D\}$ and $\{E_E, E_D\}$ design space in generations 1, 9 and 17. The candidates are represented by black dots. The red circle represents the CMA-ES covariance matrix from which the candidates in the current generation are sampled. For visualization purposes, we partition design space into Voronoi cells around each candidate and color each cell by the fitness $\Delta\beta F$ of the corresponding candidate. (j)-(l) Free energy surfaces of the best candidates in generations 1, 9 and 17 in the composite diffusion map space spanned by the leading two diffusion map collective variables $\{\psi_2, \psi_3\}$. The particular values of $\{E_D, E_E, \phi\}$ pertaining to each candidate are listed above each panel. The values of the local free energy minimum associated with each aggregate are displayed next to the representative structures.

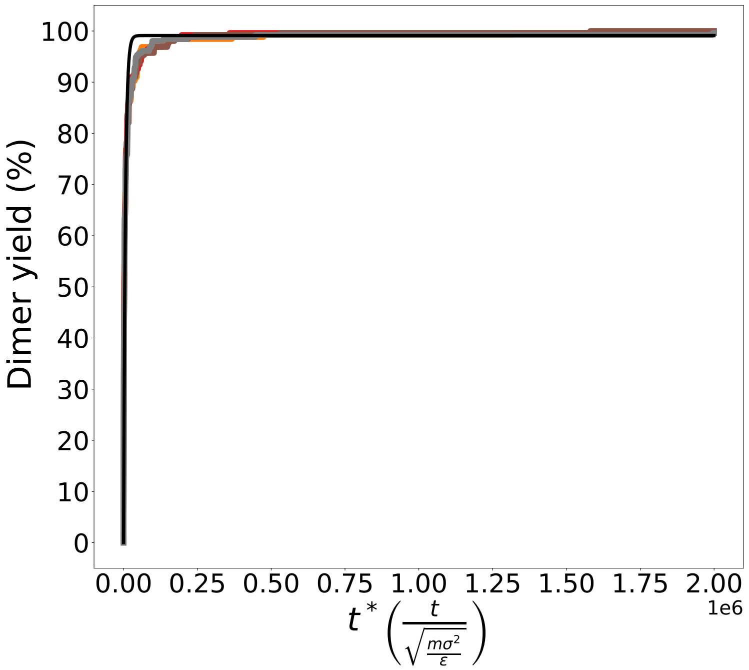


Figure 2.14: Yield of staggered tetrahedral dimers as a function of simulation time at $T^* = 0.8$ in unbiased Langevin dynamics simulations. Each colored line corresponds to an independent simulation at the optimal design $\{E_D = 10.02\varepsilon, E_E = 11.64\varepsilon, \phi = 26.68°\}$ deduced by landscape engineering. The solid black line is the fit to first-order kinetics.

2.14. Fitting the first-order kinetic model for the dimer yield $y(t) = b\left(1 - e^{-kt}\right)$ results in best-fit constants of $k = (459.9 \pm 29.1)\text{s}^{-1}$ and $b = (99.1 \pm 0.1)\,\%$, demonstrating that this design produces staggered dimers with nearly quantitative yield.

Two-Stage Hierarchical Assembly of Cubic Diamond Lattice

Landscape engineering furnished $\{E_D = 10.02\varepsilon, E_E = 11.64\varepsilon, \phi = 26.68°\}$ as the optimal values of design parameters for "D" and "E" patches at $T^*_{\text{high}} = 0.8$. We proportionally scale these interaction strengths by a factor of $T^*_{\text{low}}/T^*_{\text{high}} = 0.3/0.8$ in order to thermally decouple the "D" and "E" interactions from the "B" interactions such that they direct assembly of tetrahedral clusters into cubic diamond lattice at the second, low-temperature stage of assembly. This results in optimal "D" and "E" patch designs of $\{E_D = \frac{T_{\text{low}}}{T_{\text{high}}}10.02\varepsilon = 3.76\varepsilon,$ $E_E = \frac{T_{\text{low}}}{T_{\text{high}}}11.64\varepsilon = 4.36\varepsilon, \phi = 26.68°\}$. We test our design in simulations of 512 randomly

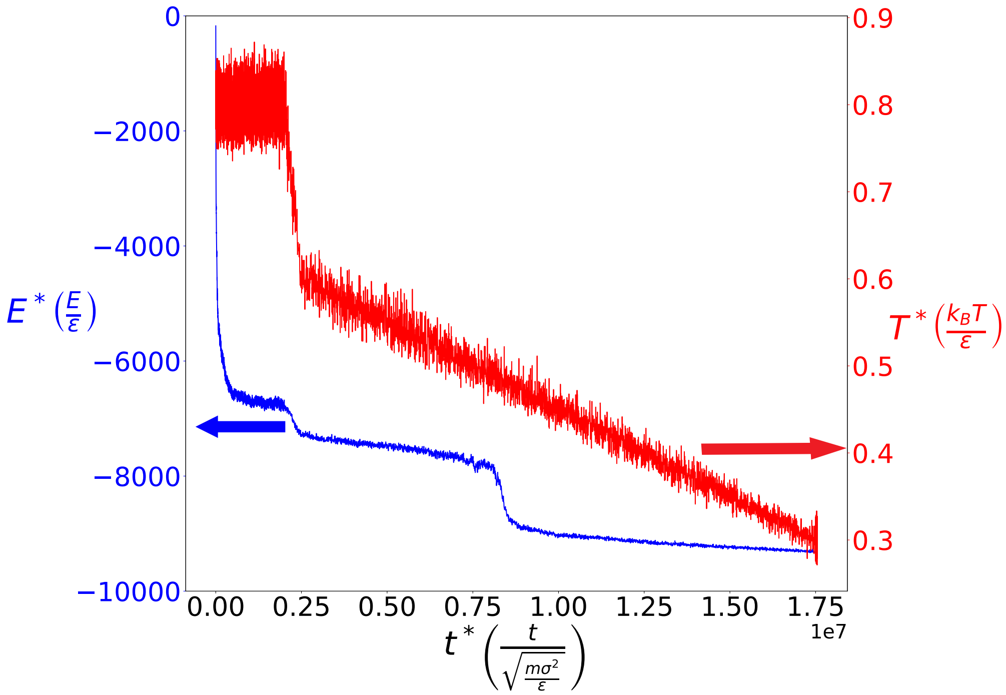


Figure 2.15: Evolution of system potential energy and temperature for two-stage hierarchical assembly of cubic diamond lattice. The two horizontal arrows indicate which of the two axes – potential energy or temperature – pertain to each curve on this double y-axis plot.

placed and oriented colloids in a cubic simulation box with side length $L = 132\sigma$, corresponding to a volume fraction of $\varphi = 0.025$. We first evolve the system at high temperature $T^*_{\text{high}} = 0.8$ for 2×10^6 reduced time units, then quickly cool the system to $T^*_{\text{intermediate}} = 0.6$ for 5×10^5 reduced time units, then slowly cool the system down to $T^*_{\text{low}} = 0.3$ for 1.5×10^7 reduced time units, and finally equilibrate the system at $T^*_{\text{low}} = 0.3$ for 5×10^4 reduced time units to gather statistics. The evolution of system potential energy and temperature versus simulation time is shown in 2.15. Nucleation of the cubic diamond lattice occurs at around $T^* = 0.48$ as indicated by the sudden drop in potential energy.

At the termination of the $T^*_{\text{high}} = 0.8$ high-temperature hold the yield of tetrahedral clusters is 95%. The radial distribution function between the geometric centers of tetrahedral cluster demonstrates that they behave as an effective ideal gas with only a small correlation peak due to weak "D" and "E" patch interactions and effective entropic attractions driven by excluded volume interactions (figure 2.16). A snapshot of the structure formed at the end

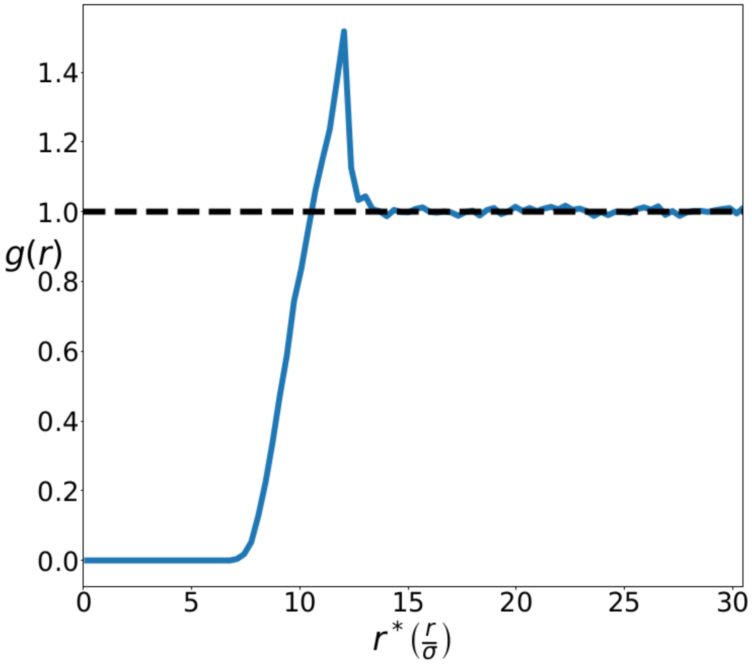


Figure 2.16: Radial distribution function between the geometric centers of tetrahedral clusters at the end of the high-temperature assembly stage of cubic diamond.

of the $T^*_{\text{low}} = 0.3$ low-temperature hold is presented in figure 2.17a. The radial distribution function (figure 2.17b), the distribution of $\vec{q_3}^*(i) \cdot \vec{q_3}(j)$ (figure 2.17c), and distribution of $\vec{q_3}(i)$ (figure 2.17d) between the geometric centers of tetrahedral clusters all show peaks at precisely the expected locations for cubic diamond lattice, and no peaks at the locations for hexagonal diamond lattice. In calculating the Steinhardt bond order parameters, a pair of tetrahedral clusters are defined as nearest neighbors if their geometric centers lie within a cutoff distance $d_{\text{cut}} = 13.0\sigma$ calibrated to cover the first peak in the radial distribution function at $r_{\text{peak},1} \approx 12.05\sigma$ (figure 2.17b). In figure 2.17c the gray bars correspond to $\vec{q_3}^*(i) \cdot \vec{q_3}(j)$ computed for all pairs of tetrahedral clusters, and the blue bars correspond to the values computed for pairs of crystalline tetrahedral clusters defined as those in which each cluster has four bonded nearest neighbors. It is clear that the finite-size effect causes the distribution of gray bars to spread out, but the blue bars are centered on the expected peak location for ideal cubic diamond lattice. The value of $q_3(i)$ is the same for the ideal cubic and hexagonal diamond lattices, so this measure possesses no discriminatory power between the two but does show the $q_3(i)$ distribution of the self-assembled lattice to be located in exactly the expected location (figure 2.17d). This structural characterization verifies that the tetrahedral clusters have assembled a defect-free cubic diamond lattice instead of a mixture of the diamond and hexagonal polymorphs.

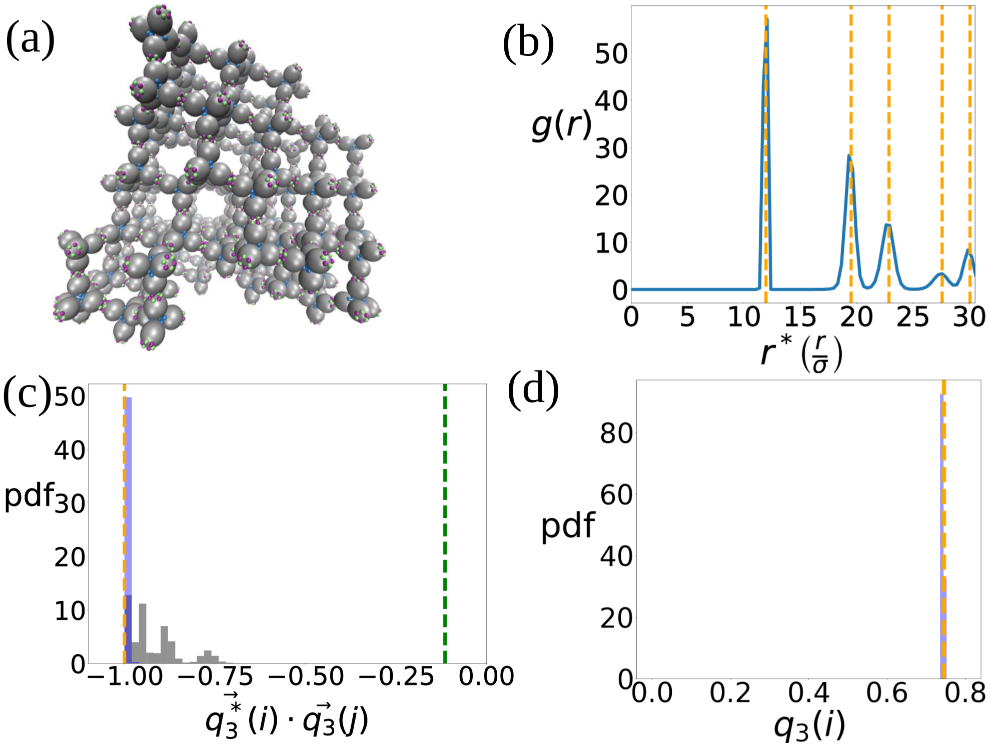


Figure 2.17: Structural characterization of the self-assembled cubic diamond lattice. (a) Snapshot of terminal crystal lattice structure. (b) Radial distribution function $g(r)$ between the geometric centers of the tetrahedral cluster. (c) The distribution of $\vec{q}_3^*(i) \cdot \vec{q}_3(j)$ computed between all pairs of tetrahedral clusters (gray) and restricted to pairs of crystalline tetrahedral clusters (blue) defined as those in which each partner has four bonded nearest neighbors. (d) The distribution of $q_3(i)$ for pairs of crystalline tetrahedral clusters. In all panels the orange dashed lines represent the expected peak positions for an ideal cubic diamond lattice and the green dashed lines represent those for an ideal hexagonal diamond lattice. In the last panel the orange and green dashed lines are coincident.

The photonic band structure of the assembled cubic diamond lattice is determined by defining an infinite periodic lattice with primitive lattice vectors $\vec{R_1} = (0, \frac{a}{2}, \frac{a}{2}), \vec{R_2} = (\frac{a}{2}, 0, \frac{a}{2}), \vec{R_3} = (\frac{a}{2}, \frac{a}{2}, 0)$, where a is the lattice constant. The eight basis particles within the $\{\vec{R_1}, \vec{R_2}, \vec{R_3}\}$ basis are given in table 2.2. The lattice constant is related to the first peak in radial distribution function between colloidal geometric centers $r_{\text{peak},1}$ as $a = \frac{4r_{\text{peak},1}}{\sqrt{2}} \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}$, where we take $r_{\text{peak},1} \approx 5.25\sigma$ as estimated for pyrochlore. Using the same parameters as for the pyrochlore calculation, we employ MPB[135] to obtain the band structure in figure 2.18. We observe an indirect bandgap between the second and third bands with a width-to-midgap ratio of 7.47%.

Basis particle index	Position
1	$(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8})$
2	$(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(\frac{3}{8}, -\frac{1}{8}, -\frac{1}{8})$
3	$(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(-\frac{1}{8}, \frac{3}{8}, -\frac{1}{8})$
4	$(\frac{1}{8}, \frac{1}{8}, \frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(-\frac{1}{8}, -\frac{1}{8}, \frac{3}{8})$
5	$(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(\frac{1}{8}, \frac{1}{8}, \frac{1}{8})$
6	$(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(-\frac{3}{8}, \frac{1}{8}, \frac{1}{8})$
7	$(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(\frac{1}{8}, -\frac{3}{8}, \frac{1}{8})$
8	$(-\frac{1}{8}, -\frac{1}{8}, -\frac{1}{8}) + \frac{\sqrt{3}}{\sqrt{3}+\sqrt{2}}(\frac{1}{8}, \frac{1}{8}, -\frac{3}{8})$

Table 2.2: Positions of basis particles of the cubic diamond lattice of tetrahedral clusters in the basis of primitive lattice vectors $\vec{R_1} = (0, \frac{a}{2}, \frac{a}{2}), \vec{R_2} = (\frac{a}{2}, 0, \frac{a}{2}), \vec{R_3} = (\frac{a}{2}, \frac{a}{2}, 0)$ where a is the lattice constant. A tuple (l,m,n) denotes a particle position vector $l\vec{R_1} + m\vec{R_2} + n\vec{R_3}$.

2.5 Conclusions

In this work we have demonstrated an automated data-driven strategy for the inverse design of colloidal particles capable of spontaneous self-assembly into periodic crystals. This approach combines molecular simulation, enhanced sampling, and nonlinear dimensionality reduction to efficiently estimate self-assembly free energy landscapes, and the use of evolutionary algorithms to rationally sculpt the topography of the landscape to stabilize desired

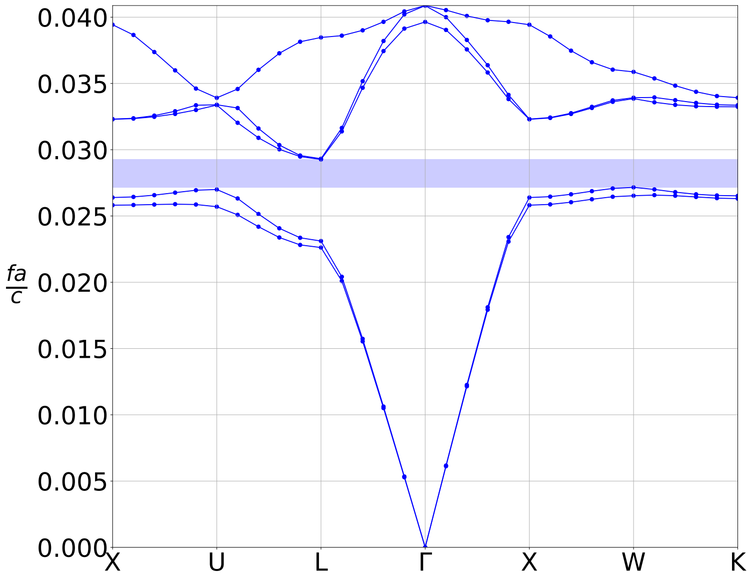


Figure 2.18: Photonic band structure of the self-assembled cubic diamond lattice. The y-axis reports the dimensionless frequency $\frac{fa}{c}$, where f is the frequency, a is lattice constant, and c is the speed of light in vacuum. The x-axis labels the corners of the irreducible region of first Brillouin zone in canonical order. For lattice constant of $a = 1.63$ μm corresponding to our choice of length scale, the bandgap lies within the frequency range of $4.95 < f < 5.32$ THz and wavelength range of $56.4 < \lambda < 60.6$ μm, placing the bandgap to be in the infrared regime of electromagnetic spectrum. For a lattice constant of $a = 16.3$ nm the bandgap lies within the frequency range of $495 < f < 532$ THz and wavelength range of $564 < \lambda < 605$ nm, placing the bandgap in the visible regime.

aggregates by manipulation of the building block design parameters. We demonstrated the technique in the successful design of anisotropic patchy colloids to self-assemble pyrochlore and cubic diamond lattices of tetrahedral clusters as highly sought-after optical materials possessing omnidirectional photonic bandgaps. Our approach presents a principled and constructive means to reverse engineer the optimal building block design. This systematic approach can accelerate design relative to Edisonian trial-and-improvement and avoid traps associated with flawed intuition. The approach can be straightforwardly generalized to arbitrary particle designs and lattice structures by identifying the self-assembled aggregates and interfaces to be stabilized and the design variables to be manipulated.

We adopted a relatively simple and generic model of our patchy particles that introduces anisotropy through the precise placement of specific and attractive patches. The interaction potentials we employed (Lennard-Jones and WCA) were deliberately simple in form but sufficient to capture the essential physics of assembly. These patchy particle potentials[60] and similar anisotropic Kern-Frenkel models[66,90] can be considered rude models of the anisotropy introduced by current experimental fabrication techniques such as glancing angle deposition[12–14], grafting of complementary DNA oligomers[8–11,102], contact layer lithography[15] and colloidal fusion[136]. In follow-on work it would be of interest to employ more realistic potentials designed to more closely mimic experimentally-realizable interaction potentials and particle designs[8,11], and incorporate the limits of fabrication robustness and precision by considering polydispersity in the building block ensemble[66,90]. Also, the rational design strategy used in the current work may be extended to design patchy colloids decorated with nanodots that may form helical structures[64], which are fundamental building blocks for chiral photonic crystals[137]. Possessing omnidirectional photonic bandgaps and the capacity to circularly polarize light, these materials have potential applications as chiral beamsplitters and components of optical computers[137].

CHAPTER 3

INVERSE DESIGN OF SELF-ASSEMBLING DIAMOND PHOTONIC LATTICES FROM ANISOTROPIC COLLOIDAL CLUSTERS

3.1 Abstract

Colloidal nanoparticles with anisotropic interactions are promising building blocks for the fabrication of complex functional materials. A challenge in the self-assembly of colloidal particles is the rational design of geometry and chemistry to program the formation of a desired target structure. We report an inverse design procedure integrating Langevin dynamics simulations and evolutionary algorithms to engineer anisotropic patchy colloidal clusters to spontaneously assemble into a cubic diamond lattice possessing a complete photonic band gap. This work is a follow-up to the work in chapter 2 in which we consider a more simplified model of the patchy colloids that is more amenable for experimental realization. The combination of a tetrahedral cluster geometry and optimized placement of a single type of anisotropic interaction patch results in a colloidal building block predicted to assemble a cubic diamond lattice with around 82% yield. This design represents an experimentally viable colloidal building block capable of high fidelity assembly into a cubic diamond lattice. This chapter is based on the work reported in: **Y. Ma**, J. Aulicino, and A.L. Ferguson "Inverse design of self-assembling diamond photonic lattices from anisotropic colloidal clusters" *J. Phys. Chem B* 125 9 2398-2410 (2021).

3.2 Introduction

In chapter 2, we developed an inverse design protocol that sculpts the free energy surface governing the self-assembly of patchy colloids and used this approach to find optimal design parameters that favor the formation of pyrochlore and cubic diamond lattices. This design strategy relied on relatively intricate placement of patches on the surface of spherical patchy colloids requiring multiple specific patch types and a two-stage hierarchical assembly mechanism. Using this approach, we discovered patchy colloid designs capable of assembling defect-free pyrochlore and cubic diamond lattices, but the relatively complex design of the particles placed them at the very edge of what is experimentally achievable even with state-of-the-art fabrication techniques.

In this chapter, we followed up on the work in chapter 2 employing a simpler inverse-design protocol and a simplified design space more amenable to experimental realization. We also employed a simpler and less computationally expensive objective function. We target the cubic diamond lattice and restrict our designs to a single patch type placed upon pre-assembled clusters of spherical patchy colloids within a rigid tetrahedral tetramer. Many surface-patterning techniques have been developed recently to decorate the surfaces of colloids with functional materials[15,86,138] and recent experimental advances have realized the fabrication of colloidal clusters[62,139–141]. The geometry of the tetrahedral colloidal cluster compensates for the loss of design flexibility associated with restricting ourselves to a single patch type. Using this strategy we report a design for an experimentally-realizable anisotropic patchy colloidal cluster that exhibits in excess of 82% yield of the open cubic diamond lattice.

The remainder of this chapter is structured as follows. In the next section we describe our computational model for the colloidal particles, Langevin dynamics simulations, and our inverse design strategy based on evolutionary algorithm. In the following section we describe the results of our inverse design approach and the validation that tetrahedral tetramers composed of patchy colloids with optimal patch design can self-assemble into the target cubic diamond lattice with high fidelity. Finally, we present our conclusions and outlook for future work.

3.3 Methods

3.3.1 Cubic Diamond Lattice

In chapter 2, we have described the basic motif of cubic diamond lattice and its analogue, the hexagonal diamond lattice (figure 2.3 (a)-(d)). The cubic diamond lattice is composed solely of rings of staggered dimers (chair-like rings, figure 2.3(b)) while the hexagonal diamond contains 25% chair-like rings and 75% boat-like rings (figure 2.3(d)). Compared to the cubic diamond lattice, the hexagonal diamond lattice possesses smaller photonic band gap occurring at higher band indices[142], making it less desirable for optical applications than cubic diamond lattice. The hexagonal diamond lattice possesses a very similar free energy to the cubic diamond lattice, making it hard to thermodynamically favor the cubic diamond lattice over the hexagonal diamond lattice[68]. The similar stabilities of competing polymorphs of open lattices has been a principal challenge in the bottom-up assembly of defect-free crystals with desirable band structures.

3.3.2 Anisotropic Patchy Colloid Building Blocks

Advanced experimental techniques have enabled fabrication of colloidal clusters with high yield and fidelity[8,62,139–141] and the surfaces of the constituent colloids forming these colloidal clusters can also be anisotropically functionalized to program the hierarchical assembly into more complex structures[8,89]. In our prior work (chapter 2), we used an inverse design strategy known as landscape engineering to discover a design for spherical patchy colloids to assemble a cubic diamond lattice by a two-stage hierarchical process: (i) the high-temperature assembly of groups of four patchy colloids into tetrahedral tetramers followed by (ii) the low-temperature assembly of these tetrahedral tetramers into a cubic diamond lattice. The two-stage hierarchical assembly mechanism and required rigidity of the tetrahedral tetramers necessitated a relatively complex colloid design possessing nine patches of three different types (figure 2.3 (e)-(h)) with different interaction potentials and patch-patch specificities. In the present work, we greatly simplify the design problem in two ways. First, we adopt as the fundamental building block a patchy tetrahedral tetramer as opposed to a spherical colloid and optimize the single-step assembly of tetrahedral tetramers into the cubic diamond lattice. Second, we functionalize the colloids using only a single patch type. We previously used three patch types and employed complementarity between patch types to stabilize the staggered (Figure 2.3a) over the eclipsed (Figure 2.3c) dimer configurations to favor the cubic diamond lattice over the hexagonal one. In the present work, we show that by adjusting the protrusion of the interaction patches above the surface of the colloid we may exploit excluded volume interactions to preferentially stabilize the staggered dimer configuration using only a single patch type. These two simplifications are motivated by experimental advances in the fabrication of rigid colloidal clusters ("colloidal molecules") with quite complex geometries, including the tetrahedral tetramer[62,139–141,143,144] and sophisticated surface-patterning techniques to precisely functionalize the surfaces of colloids with anisotropic interaction patches composed of organic, inorganic, or biological materials[15,86,138,145]. Conceptually, we reduce the complexity in the anisotropic interaction patches (i.e., going from three patch types to one) at the expense of increased complexity in the colloid shape (i.e., tetrahedral colloidal building blocks rather than spherical ones) to design a building block that is more readily accessible to existing experimental techniques.

In our model, each tetramer is treated as a tetrahedral assembly of four spherical colloids (type "A" particles) that move as a rigid body. Figure 3.1a illustrates a tetrahedral tetramer dimer in a staggered configuration wherein the base of one tetrahedral tetramer is azimuthally rotated through 60° with respect to the other along the axis connecting their centers of mass. The surface of each spherical colloid is functionalized with three anisotropic interaction patches (type "B" particles) in an equilateral triangle arrangement as illustrated in Figure 3.1b. The "B" patches are located at a polar angle ϕ_B from the pole of each "A" colloid. The placement of each patch on each colloid in the tetrahedral cluster is identical such that the tetrahedral tetramer is tetrahedrally symmetric. The “B” interaction patches are represented as Lennard-Jones spheres on the surface of the “A” colloid. The degree of protrusion of the “B” patches is quantified by the protrusion ratio $\alpha_B = d_{AB}/R_A$, where d_{AB} is the distance between the center of the “B” patch and the center of the “A” colloid and R_A is the radius of the “A” colloid. A protrusion ratio of $\alpha_B = 1$ indicates that the center of the “B” sphere is coincident with the surface of the “A” colloid, a value of $\alpha_B = (1 + R_B/R_A)$ indicates that the “B” sphere lies tangent upon (i.e., “kisses”) the “A” colloid, and a value of $\alpha_B = (1 - R_B/R_A)$ indicates that the “B” patch is buried just below the surface of the “A” colloid. Controlling the protrusion ratio of the patch enables us to stabilize the staggered dimer of tetrahedral tetramers using a single patch type by favoring interlocking configurations of the patches at the interface as shown in Figure 3.1c. Figure 3.1d presents a schematic drawing of the colloidal particle architecture illustrating ϕ_B and α_B.

As in chapter 2, we model the patch-patch (“B”-“B”) interactions between the spherical patches with a Lennard-Jones potential,

$$U_{\mathrm{LJ}}^{BB}(r) = 4\varepsilon_B \left[\left(\frac{\sigma_B}{r}\right)^{12} - \left(\frac{\sigma_B}{r}\right)^{6} \right], \tag{3.1}$$

where r is the center of mass distance between the “B” spheres, ε_B is the well depth controlling the interaction strength, and σ_B is the patch diameter. Following our previous work we choose the colloid to be five times larger than the surface patches such that $\sigma_A = 5\sigma$ and $\sigma_B = \sigma$. The colloid-colloid (“A”-“A”) and colloid-patch (“A”-“B”) interactions are treated by a surface-shifted Weeks-Chandler-Andersen (WCA) potential[106] to model excluded-volume

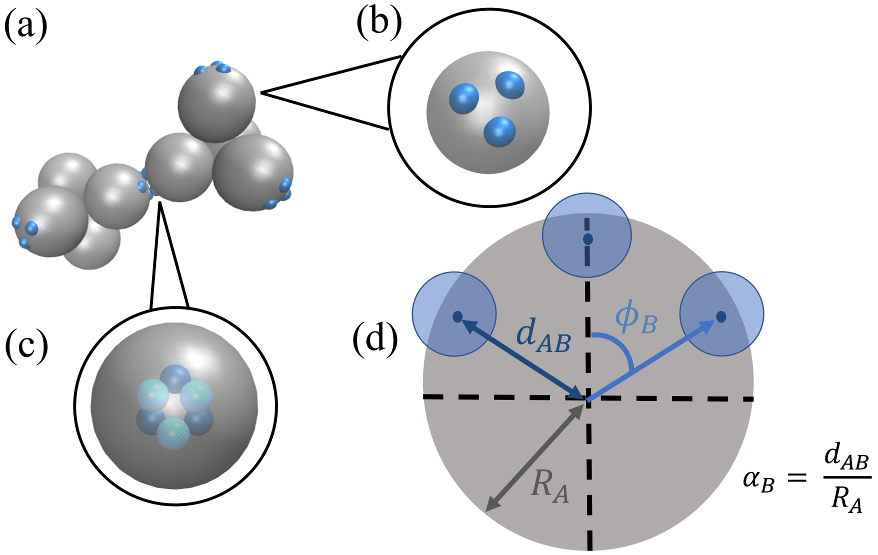


Figure 3.1: Computational model of the patchy colloid tetrahedral tetramers. (a) A staggered dimer of tetrahedral tetramers showing the colloids ("A" particles, grey) functionalized with anisotropic surface patches ("B" particles, blue). (b) A zoomed-in view of a single spherical patchy colloid belonging to one of the tetrahedral tetramers. The interaction patches are modeled as Lennard-Jones spheres placed in an equilateral triangle configuration at a tunable surface depth. (c) A zoomed-in view of the staggered dimer interface between the two tetrahedral tetramers along the axis connecting their centers of mass. The dark blue spheres represent the surface patches in one of the tetrahedral tetramers and the light blue spheres to those in the other. An interlocked configuration of the surface patches favors the staggered dimer configuration. (d) A schematic diagram of a patchy colloid illustrating the polar angle ϕ_B and protrusion ratio α_B of the surface patches. The transparent blue circles represent the patches and the dark blue dots represent the centers of patches. d_{AB} is the distance between the center of colloid and the center of patch and R_A is the radius of colloid. The protrusion ratio α_B is defined as $\alpha_B = d_{AB}/R_A$. All molecular renderings in this figure and throughout the paper are constructed using Visual Molecular Dynamics (VMD)[1].

interactions,

$$U_{\mathrm{WCA}}^{ij}(r) = \begin{cases} 4\varepsilon_{ij}\left[\left(\frac{\sigma}{r-\Delta_{ij}}\right)^{12} - \left(\frac{\sigma}{r-\Delta_{ij}}\right)^{6}\right] + \varepsilon_{ij} & \text{if } r < 2^{\frac{1}{6}}\sigma + \Delta_{ij}, \\ 0 & \text{if } r \geq 2^{\frac{1}{6}}\sigma + \Delta_{ij}, \end{cases} \tag{3.2}$$

where ε_A is the well depth controlling the interaction strength for the "A"-"A" interaction, $\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$ is given by the Lorentz-Berthelot mixing rule, and $\Delta_{ij} = (\sigma_i + \sigma_j)/2 - \sigma$ shifts the potential to act between the surfaces of particles i and j. The assembly of four "A" colloids and 12 "B" patches comprising the tetrahedral tetramer is treated as a rigid body and interactions between particles in the same rigid body are neglected.

The geometry and interaction potential of the system is fully defined by the six parameters $\{\varepsilon_A, \varepsilon_B, \sigma_A, \sigma_B, \alpha_B, \phi_B\}$ defining the interaction strength, size and relative arrangement of the "A" and "B" particles. Since only the relative strength of the "A" and "B" interactions is meaningful – the absolute values can be scaled by modulating temperature – we reduce the parameter space by eliminating ε_A from consideration and considering only the relative value of ε_B. Similarly, only the relative values of σ_A and σ_B are meaningful, with the absolute values corresponding to a global rescaling in the size of the particles. In this work, we follow our previous work and fix the relative ratio of the particle size as $\sigma_A = 5\sigma_B = 5\sigma$. We achieve good results under this choice, but, in principle, we could also consider changing the size and/or shape of the patch. As such, the inverse design problem is defined over the three-dimensional design space defining the interaction strength, polar angle, and protrusion ratio of the "B" patch $\{\varepsilon_B, \phi_B, \alpha_B\}$. The design strategy seeks to optimize both the chemistry (i.e., interaction strength) and geometry (i.e., polar angle and protrusion ratio) of the anisotropic surface patches to favor the staggered dimer configuration and promote spontaneous defect-free assembly of a cubic diamond lattice.

3.3.3 Optimization Objective Function

A direct computational approach to optimizing $\{\varepsilon_B, \phi_B, \alpha_B\}$ would randomly place tetrahedral tetramers within a simulation box, gently anneal the system to induce nucleation and growth of a crystal, and then modify $\{\varepsilon_B, \phi_B, \alpha_B\}$ to maximize yield of the cubic diamond lattice at the termination of the annealing procedure. This direct optimization is inefficient, however, due to the need for very slow cooling rates in order to avoid kinetic traps and reliably estimate the thermodynamic yield of cubic diamond crystals[67,146]. Instead, we define a proxy optimization problem in which $\{\varepsilon_B, \phi_B, \alpha_B\}$ are optimized to favor the formation of staggered dimers between colloidal monomers (i.e., isolated "A" spheres functionalized with "B" patches) at a fixed temperature. This problem is simpler and faster since we do not perform explicit slow temperature ramping during the optimization and directly focus on optimizing the colloid-colloid interface to favor an interlocking patch conformation. We show later during temperature ramping simulation for pre-assembled tetramers that transferring the optimal design found for monomeric patchy colloid to the tetrahedral tetramer does indeed result in quite high-yield cubic diamond crystals and provides *post hoc* validation of our more efficient proxy optimization.

We evaluate the quality of a particular $\{\varepsilon_B, \phi_B, \alpha_B\}$ triplet by conducting Langevin dynamics simulations of the assembly of colloidal monomers (Section 3.3.4) and computing the fraction of aggregates that exist as staggered dimers at equilibrium. We define a geometric criterion under which a dimer between colloidal monomers i and j should be classified as staggered based on the planar angle θ_{ij} and the dihedral angle Ψ_{ij} between the constituent colloids (Figure 3.2). The angle θ_{ij} is defined as,

$$\begin{aligned} \cos(\theta_i) &= \hat{\xi}_i \cdot \hat{r}_{ij} \\ \cos(\theta_j) &= \hat{\xi}_j \cdot \hat{r}_{ji} \\ \theta_{ij} &= \max(\theta_i, \theta_j), \end{aligned} \tag{3.3}$$

where $\hat{\xi}_i$ and $\hat{\xi}_j$ denote the unit orientation vectors of each particle pointing from the center of mass of the "A" colloid to the centroid of the three "B" patches and $\hat{r}_{ij}$ denotes the unit displacement vector from i to j (Figure 3.2a). The planar angle θ_i measures the angle between $\hat{\xi}_i$ and $\hat{r}_{ij}$ and θ_j measures the angle between $\hat{\xi}_j$ and $\hat{r}_{ji}$. Defining θ_{ij} as maximum of these returns the larger deviation of either partner in the dimer from a face-to-face alignment in which $\hat{\xi}_i$ and $\hat{\xi}_j$ are antiparallel.

The dihedral angle Ψ_{ij} is defined as,

$$\Psi_{ij} = \min_{m \in \text{patch}_i, n \in \text{patch}_j} (\psi_{ij}^{mn}), \tag{3.4}$$

where ψ_{ij}^{mn} defines the relative rotation between each of the three patches m on colloid i and the three patches n on colloid j. We compute ψ_{ij}^{mn} by finding the centroid of three patches on colloid i denoted by c_i and the centroid of three patches on colloid j denoted by c_j, then calculate ψ_{ij}^{mn} as the angle between the plane through $\{m, c_i, c_j\}$ and the plane through $\{c_i, c_j, n\}$. Denoting the vector from m to c_i as $\vec{b}_1$, the vector from c_i to c_j as $\vec{b}_2$, and the vector from c_j to n as $\vec{b}_3$, the dihedral angle ψ_{ij}^{mn} is computed as,

$$\begin{aligned} \vec{n}_1 &= \frac{\vec{b}_1 \times \vec{b}_2}{\left\|\vec{b}_1 \times \vec{b}_2\right\|}, \\ \vec{n}_2 &= \frac{\vec{b}_2 \times \vec{b}_3}{\left\|\vec{b}_1 \times \vec{b}_2\right\|}, \\ \cos\left(\psi_{ij}^{mn}\right) &= \vec{n}_1 \cdot \vec{n}_2. \end{aligned} \tag{3.5}$$

The dihedral angle Ψ_{ij} is defined as the minimum over the nine ψ_{ij}^{mn}, which – assuming a small value of θ_{ij} and therefore relatively cofacial dimer alignment – quantifies the minimum rotational dihedreal between the colloids in the dimer pair required to align the patches on each colloid into an eclipsed configuration. Figure 3.2c provides a schematic illustration for the case in which $\Psi_{ij} = \psi_{ij}^{11'} = \psi_{ij}^{22'} = \psi_{ij}^{33'}$.

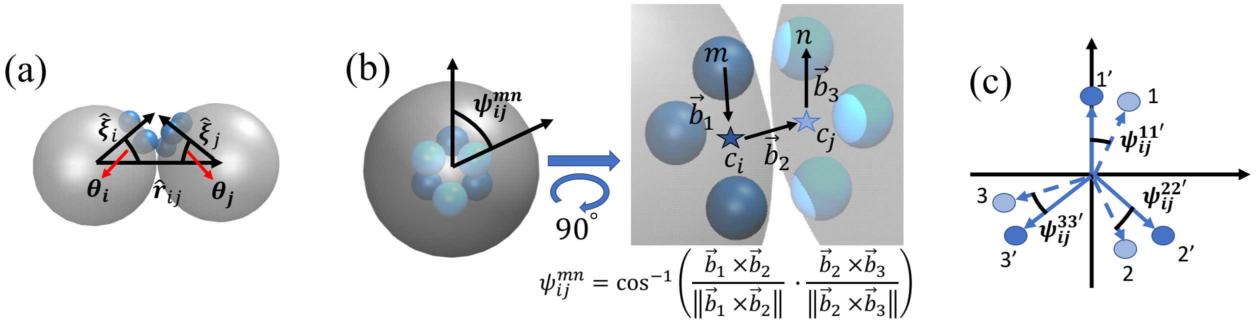


Figure 3.2: Geometry of self-assembled dimers of colloid monomers. (a) The angle $\theta_{ij} = \max(\theta_i, \theta_j)$ measures the larger deviation of either partner dimer i or j from a face-to-face alignment in which the unit orientation vectors $\hat{\xi}_i$ and $\hat{\xi}_j$ linking the center of the colloid to the pole containing the patch are antiparallel and collinear with the unit vector $\hat{r}_{ij}$ linking the colloidal centers. (b) The dihedral angle Ψ_{ij} measures the minimum dihedral rotation required within a cofacial dimer pair to align the patches on each colloid into an eclipsed configuration. The value of Ψ_{ij} is taken as the minimum over all nine ψ_{ij}^{mn} defining the dihedral angles between the three patches m on colloid i and the three patches n on colloid j. The ψ_{ij}^{mn} are computed by first finding the centroid c_i of patches on colloid i and the centroid c_j of patches on colloid j and then computing the angle between the plane through $\{m, c_i, c_j\}$ (spanned by $\vec{b}_1$ and $\vec{b}_2$) and the plane through $\{c_i, c_j, n\}$ (spanned by $\vec{b}_2$ and $\vec{b}_3$). (c) Schematic diagram of $\Psi_{ij} = \min_{m \in \text{patch}_i, n \in \text{patch}_j}(\psi_{ij}^{mn})$ considering a particular interfacial arrangement of the three light blue "B" patches on colloid i and three dark blue "B" patches on colloid j. Assuming θ_{ij} is small such that the colloids are approximately cofacial $\Psi_{ij} = \psi_{ij}^{11'} = \psi_{ij}^{22'} = \psi_{ij}^{33'}$ defines the minimum azimuthal rotation required to achieve an eclipsed configuration.

We classify a dimer as staggered if ($0° \leq \theta_{ij} \leq 5°$) and ($55° \leq \Psi_{ij} \leq 60°$). Enforcing a low threshold on θ_{ij} ensures that the patches are approximately face-to-face aligned (i.e. $\hat{\xi}_i$ and $\hat{\xi}_j$ are nearly antiparallel) and that the value of Ψ_{ij} is meaningful. An ideal staggered dimer would possess ($\theta_{ij} = 0°$, $\Psi_{ij} = 60°$). The 5° threshold in both θ_{ij} and Ψ_{ij} is motivated by the range of the observed distribution of these angles in the ensemble of stable staggered dimers resulting from favorable $\{\varepsilon_B, \phi_B, \alpha_B\}$ choices.

Having defined a criterion by which counts the number of staggered dimers, we define the objective function to be maximized as the equilibrium fraction of staggered dimers among all self-assembled aggregates,

$$\begin{aligned} f(\varepsilon_B, \phi_B, \alpha_B) &= \left\langle \frac{N_{\text{staggered dimer}}}{N_{\text{aggregates}}} \right\rangle \\ &= \frac{1}{Z} \int e^{-\beta U(\boldsymbol{r}^N, \boldsymbol{\Omega}^N; \varepsilon_B, \phi_B, \alpha_B)} \frac{N_{\text{staggered dimer}}(\boldsymbol{r}^N, \boldsymbol{\Omega}^N; \varepsilon_B, \phi_B, \alpha_B)}{N_{\text{aggregates}}(\boldsymbol{r}^N, \boldsymbol{\Omega}^N; \varepsilon_B, \phi_B, \alpha_B)} d\boldsymbol{\Omega}^N d\boldsymbol{r}^N \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \frac{N_{\text{staggered dimer}}(\boldsymbol{r}_i^N, \boldsymbol{\Omega}_i^N; \varepsilon_B, \phi_B, \alpha_B)}{N_{\text{aggregates}}(\boldsymbol{r}_i^N, \boldsymbol{\Omega}_i^N; \varepsilon_B, \phi_B, \alpha_B)} \end{aligned} \tag{3.6}$$

where $\boldsymbol{r}^N$ and $\boldsymbol{\Omega}^N$ denote the positions and orientations of the N patchy colloidal monomers in the simulation, $\boldsymbol{r}_i^N$ and $\boldsymbol{\Omega}_i^N$ denote their positions and orientations in frame i of simulation trajectory, $U(\boldsymbol{r}^N, \boldsymbol{\Omega}^N; \varepsilon_B, \phi_B, \alpha_B)$ is the potential energy of the system, $\beta = 1/k_B T$ is the reciprocal temperature, $N_{\text{staggered dimer}}$ is a function returning a count of staggered dimers according to the criterion defined above for a particular system configuration, and $N_{\text{aggregates}}$ is a function returning a count of aggregates of all sizes (monomers, staggered and non-staggered dimers, trimers, tetramers, pentamers, etc.). The ensemble average in the second line is approximated by a time-average in the third line that is evaluated over n frames from the equilibrated production portion of the Langevin dynamics simulation. We evaluate this objective function by running unbiased Langevin dynamics simulations, whereas the objective function in chapter 2 (equation 2.13) is evaluated by running many umbrella sampling simulations to construct the free energy surface. Although it is more rigorous to evaluate the relative thermodynamic stabilities of self-assembled structures by constructing the free energy surface, the objective function defined here requires less simulations to calculate and the results of our study show that its optimization could guide the colloidal clusters toward desired self-assembly behavior.

3.3.4 Langevin Dynamics Simulations

We evaluate the objective function in Equation 3.6 by conducting Langevin dynamics simulations of the self-assembly of patchy colloidal monomers in HOOMD-blue[112,113]. We initialize each simulation from a random dispersion of $N = 64$ monomeric patchy colloids with a particular combination of $\{\varepsilon_B, \phi_B, \alpha_B\}$ design parameters and observe the distribution of self-assembled aggregates that spontaneously form. Importantly, by running many-body simulations of assembly we seek to both maximize the thermodynamic yield of the desired staggered dimers relative to all competing aggregates and also assure their kinetic accessibility. We perform our simulation in reduced units, where $\sigma = 1$, $\varepsilon = \varepsilon_A = 1$, and $m = 1$. Using these units, we specify $\sigma_A = 5$, $\sigma_B = \sigma = 1$, $m_A = 125$ and $m_B = m = 1$. The relative mass of the colloid and patch is scaled in proportion to size but these choices could be tuned based on the relative densities of the colloidal "A" particle (e.g., silica, silicon, polystyrene) and "B" patches (e.g., metal, polymer). We perform simulations in a cubic simulation box of side length $L = 52\sigma$, corresponding to a $\varphi = 0.05$ volume fraction of colloidal monomers. The equations of motion are numerically propagated for 1×10^8 steps using a Langevin dynamics integrator with a step size of $dt^* = 0.005$ and temperature of $T^* = 0.8$. The first 5×10^7 steps are discarded for equilibration and frames are saved every 1×10^4 steps over the remaining 5×10^7 step production period to evaluate $f(\varepsilon_B, \phi_B, \alpha_B)$ using Equation 3.6. We verify that the equilibration period is sufficiently long such that the system energy and aggregation numbers of various aggregates (monomers, staggered dimers, eclipsed dimers,

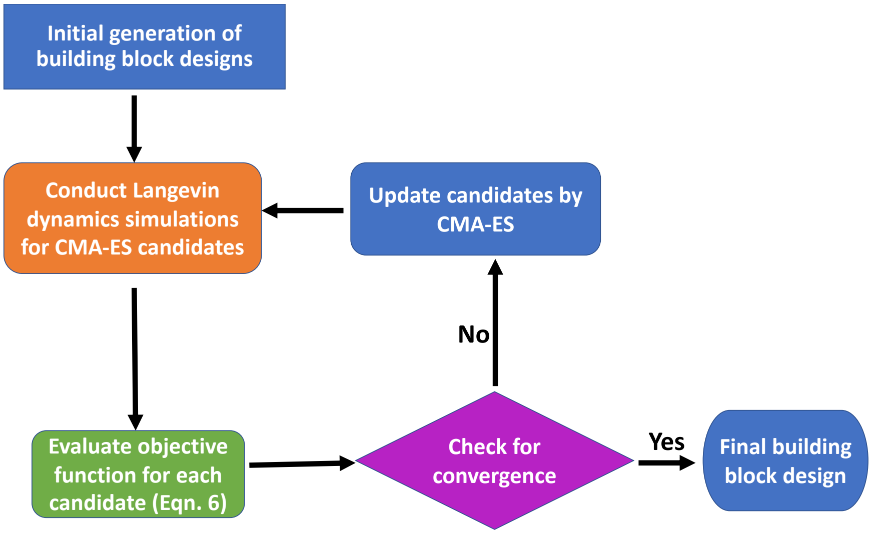

Initial generation of building block designs
Conduct Langevin dynamics simulations for CMA-ES candidates
Update candidates by CMA-ES
No
Evaluate objective function for each candidate (Eqn. 6)
Check for convergence
Yes
Final building block design

Figure 3.3: Flow diagram of the optimization procedure combining Covariance Matrix Adaptation Evolution Strategy (CMA-ES) and Langevin dynamics simulations.

trimers, etc.) fluctuate around stable mean values over the production period. We perform three independent simulations for each candidate and pass the mean value of the objective function to the evolutionary optimization algorithm (CMA-ES in section 2.3.2 of chapter 2). An illustration of the inverse design pipeline is shown in figure 3.3.

By performing the evaluation of the objective function at a single temperature of T^* = 0.8, we optimize assembly of staggered dimers at this temperature. Once the optimal design is determined, we transfer the patch design to the tetrahedral tetramers and perform slow temperature annealing from a high temperature state point at which the tetrahedral tetramers are fully dispersed to a low temperature state at which the system is fully assembled. Since ε_B is the only tunable energy scale in our reduced unit calculations, the optimal ε_B discovered at $T^* = 0.8$ may be arbitrarily rescaled to modulate the assembly temperature.

A mapping between reduced units and real units can be made by specifying the size σ_A and density ρ_A of the "A" colloid and the energy scale ε. The temperature and time in real units (T, t) are then related to corresponding quantities in reduced units (T^*, t^*) as

$T = T^* \frac{\varepsilon}{k_B}$ and $t = t^* \sigma \sqrt{\frac{m}{\varepsilon}}$. For example, adopting $\sigma_A = 5\sigma = 1$ μm, $\rho_A = 1$ g/cm^3, and ε $= 0.4$ k_BT at $T = 298$ K means that the reduced temperature of $T^* = 0.8$ corresponds to $T = 95$ K, the reduced time step of $dt^* = 0.005$ to $dt = 0.05$ μs, and the total length of our simulations to $t = 5$ s.

3.4 Results and Discussion

We first report the determination of the optimal patch parameters determined using our CMA-ES optimization and Langevin dynamics protocol to maximally favor the assembly of staggered dimers of colloidal monomers. We then transfer this patch design to tetrahedral tetramers and report our validation of their capacity to assemble into a cubic diamond lattice in slow temperature annealing simulations.

3.4.1 Determination of Optimal Patch Design

We commenced the optimization loop in figure 3.3 by seeding it with $M = 12$ initial building block designs with parameters $\{\varepsilon_B, \phi_B, \alpha_B\}$ sampled from a multivariate Gaussian distribution with mean $\langle \boldsymbol{x}^0 \rangle = (3.33\varepsilon,\ 20.00°,\ 0.85)^T$ and covariance $C^0 = \mathrm{diag}(1.11, 10.00, 0.01)$. Hereafter, we call each particular set of parameters $\{\varepsilon_B, \phi_B, \alpha_B\}$ a “candidate” in the parameter space. The initial step size was set to $\sigma^0 = 1.0$ to favor early exploration of the design space and mitigate possible trapping in a local optimum. For each candidate, we ran three independent Langevin dynamics simulations and passed the average of the objective function value (Equation 3.6) obtained from each simulation to the optimizer. Subsequent CMA-ES generations were seeded based on the top $\mu = 3$ of the $M = 12$ candidates.

The evolution of $\{\varepsilon_B, \phi_B, \alpha_B\}$ over the course of CMA-ES generations are presented in figure 3.4a-c and the evolution of the mean value of objective function over the top $\mu = 3$ candidates for each generation is shown in figure 3.4d. In conducting the optimization we constrained the protrusion ratio to lie in the range $\alpha_B = [0.8,1.0]$ by penalizing the objective function to $f = (-\infty)$ for candidates outside of this range. This prevented the unphysical situations of the patch detaching from the colloid ($\alpha_B > 1.0$) or completely sinking below the surface ($\alpha_B < (1 - \sigma_B/\sigma_A) = 0.8$). The optimizer converges to a value of α_B inside this range and these constraints are inactive in the later generations of the optimization. The optimization is terminated at generation $g = 28$ at which point the standard deviations $\{0.239\varepsilon, 0.300°, 0.005\}$ in the three design variables proposed for generation $g = 29$ fall below the prescribed convergence thresholds of $\{0.33\,\varepsilon, 1.0°, 0.01\}$ and the mean design is declared the converged solution $\{\varepsilon_B^{\text{opt}}, \phi_B^{\text{opt}}, \alpha_B^{\text{opt}}\} = \{8.18\varepsilon, 19.6°, 0.907\}$. Under the reduced to real unit mapping defined in Section 3.3.4, the optimal interaction strength corresponds in real units to $\varepsilon_B^{\text{opt}} = 3.3\ k_BT$ at $T = 298$ K.

Inspection of the optimization time courses show that the interaction strength ε_B climbs from its initial starting point of 3.33ε to more than double and reach its terminal plateau at 8.18ε by around generation $g = 20$ (Figure 3.4a). The protrusion ratio α_B undergoes some exploratory fluctuations before increasing slightly from it starting point of 0.85 to settle down to its terminal optimum of 0.907 (Figure 3.4c). The polar angle ϕ_B reaches an optimum of 19.6° that is changed very little from the initial guess of 20.0°, but the large fluctuations in the early generations show that the algorithm does broadly explore a variety of angles before converging (Figure 3.4b). The mean value of objective function evaluated over the top $\mu = 3$ candidates undergoes large fluctuations in the early generations but by generation $g = 15$ approaches and then asymptotes to a high plateau that nearly quintuples the fraction staggered dimers among aggregates from $\langle f^0 \rangle_\mu = 0.17$ to $\langle f^{28} \rangle_\mu = 0.83$.

Analysis of the assembly trajectories allow us to rationalize the behavior of the optimizer from a structural perspective. The polar angle ϕ_B is constrained to lie in the vicinity of 20° in order to admit interlocked colloidal interfaces between the tetrahedral tetramers (Figure 3.1). At this angle, the patches can maximize favorable contacts via short range attractive Lennard-Jones interactions (Equation 3.1) in an interlocked configuration wherein

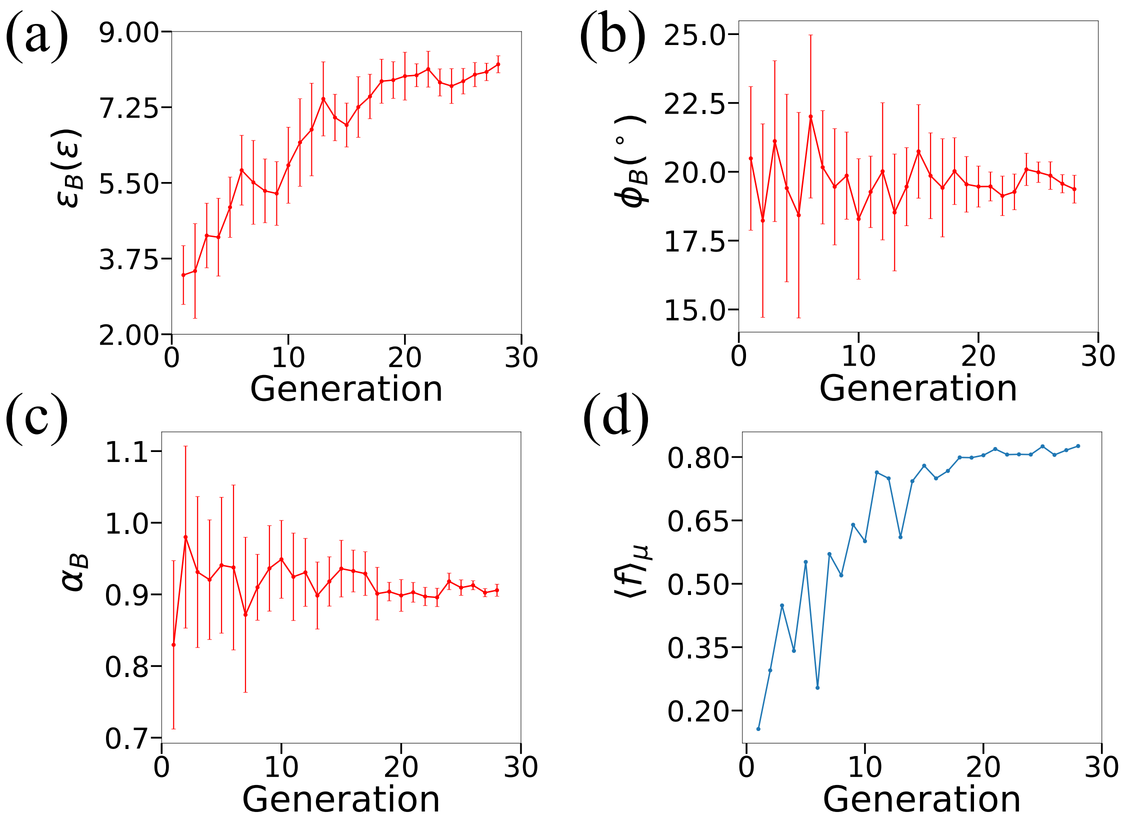


Figure 3.4: Evolution of design parameters and objective function over the CMA-ES optimization course. Evolution of the (a) interaction strength ε_B, (b) polar angle ϕ_B, and (c) protrusion ratio α_B over the 28-generation optimization. The lines and error bars correspond, respectively, to the mean and standard deviation of each parameter over the M = 12 candidates in each generation. The optimization converges at generation $g = 28$ to an optimum of $\{\varepsilon_B^{\text{opt}}, \phi_B^{\text{opt}}, \alpha_B^{\text{opt}}\} = \{8.18\varepsilon, 19.6°, 0.907\}$. (d) The mean value of objective function (Equation 3.6) evaluated over the top $\mu = 3$ candidates in each generation reporting the fraction of staggered dimers among self-assembled aggregates. The terminal value of the mean objective function value reaches $\langle f^{28} \rangle_\mu = 0.83$ corresponding to 83% staggered dimers.

each patch interacts with two nearest neighbors. Smaller angles prevent a tight interlocking due to insufficient free volume between the patches and lager angles spread the patches too far apart to admit two nearest neighbor contacts. Larger protrusion ratios α_B and stronger interaction strengths ε_B would appear to offer increasing energetic stabilization of the staggered dimer but this process must be viewed in the context of alternative accessible assembly pathways and states. This process is limited by the fact that too large protrusion ratios and interaction strengths make the triplet of "B" patches too accessible and strongly bound to multiple interaction partners, thereby favoring the formation of large aggregates that can outcompete the staggered dimer. Furthermore, the interaction strength cannot be so strong as to prevent mutual rearrangements and relaxations of bound particles thereby preventing irreversible aggregation and kinetic trapping into a glass[147].

3.4.2 Validation of Optimal Patch Design

The optimal patch design programs 83% of colloidal monomers to assemble into staggered dimers. We now proceed to validate that this same patch design can also induce the robust assembly of tetrahedral tetramers into a cubic diamond lattice.

Slow Temperature Annealing Assembly of Optimal Tetramers

We conduct Langevin dynamics simulations of $N = 512$ initially randomly placed tetrahedral tetramers decorated with the optimal patch design $\{\varepsilon_B^{\text{opt}}, \phi_B^{\text{opt}}, \alpha_B^{\text{opt}}\} = \{8.18\varepsilon, 19.6°, 0.907\}$. We recall that these tetrahedral tetramers comprise a rigid cluster of four "A" colloids each decorated with three "B" patches as a simplified model of an experimentally-realizable "colloidal molecule"[62,139–141,143,144] functionalized by surface patterning techniques to induce anisotropic patchy interactions[15,86,138,145]. Simulations were conducted in a cubic box with side length $L = 204.08\sigma$, corresponding to a tetrahedral tetramer volume fraction of $\varphi = 0.05$. We perform a high-temperature equilibration of the system at $T^*_{\text{high}} = 4.0$ ($T_{\text{high}} =$

476.8 K, under the real unit mapping defined in Section 3.3.4) for 1×10^8 steps with a step size of $dt^* = 0.005$ ($t^* = 5 \times 10^5$; $t = 5.0$ s). Under these high temperature conditions the attractive "B" patch interactions are insufficient to promote aggregation and the tetrahedral tetramers behave effectively as an ideal gas. We then perform slow temperature annealing of the system under a linear ramp down to $T^*_{\text{low}} = 2.0$ ($T_{\text{low}} = 238.4$ K) over the course of 2×10^9 steps ($t^* = 1 \times 10^7$; $t = 101$ s; $\Delta T/t = 2.36$ K/s). Finally, we conduct a 1×10^7 step ($t^* = 5 \times 10^4$; $t = 0.5$ s) hold at $T^*_{\text{low}} = 2.0$ ($T_{\text{low}} = 238.4$ K) over which we collect data on the self-assembled structure.

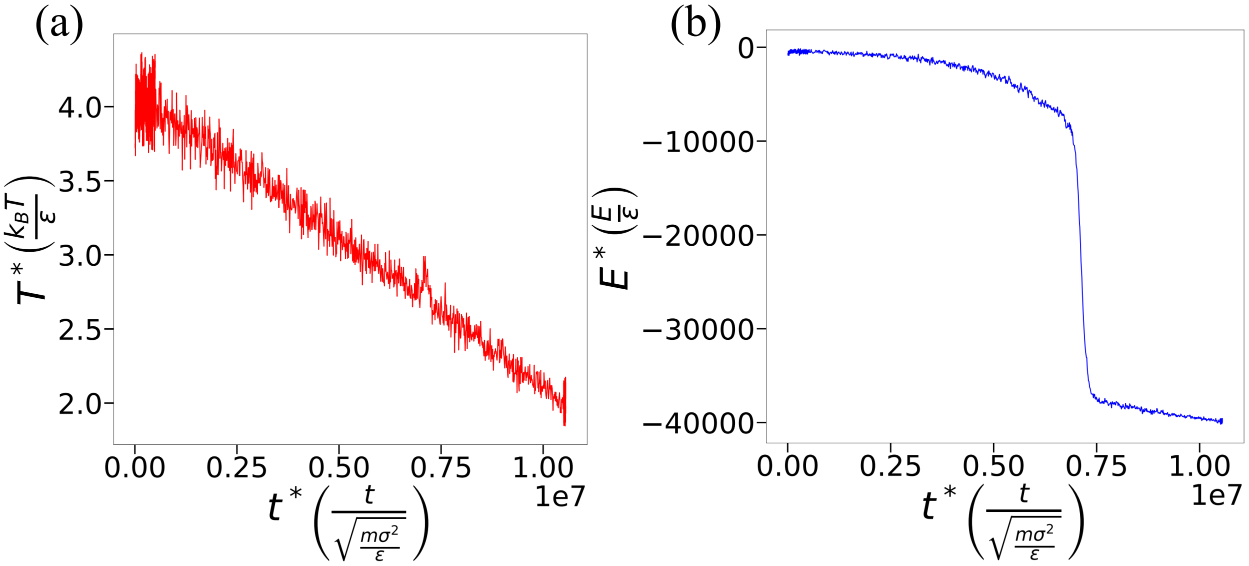


Figure 3.5: Slow temperature annealing induction of tetrahedral tetramer self-assembly. (a) Commencing from an equilibrated high-temperature effective ideal gas of tetrahedral tetramers at $T^*_{\text{high}} = 4.0$, a linear temperature ramp down to $T^*_{\text{low}} = 2.0$ is executed over the course of 2×10^9 integration steps of $dt^* = 0.005$. (b) The potential energy over the course of the cooling run undergoes a precipitous drop at $t^* \approx 0.7 \times 10^7$ corresponding to a temperature of $T^* \approx 2.75$ that marks the assembly transition. The small temperature spike at $t^* \approx 0.7 \times 10^7$ is attributable to the latent heat of fusion released by the self-assembly process.

The plots of temperature and potential energy over the course of the annealing run are presented in Figure 3.5. The system undergoes an assembly transition from the initial dispersion of isolated tetrahedral tetramers marked by the precipitous drop in potential energy at $t^* \approx 0.7 \times 10^7$ (Figure 3.5b) corresponding to a temperature of $T^* \approx 2.75$ (Figure 3.5a). The small temperature peak in Figure 3.5a at $t^* \approx 0.7 \times 10^7$ can be attributed to the latent heat of fusion released upon assembly. A slow cooling schedule was adopted to favor a single nucleation event of the most thermodynamically favored polymorph and avoid kinetic traps. Assembly commences at a higher temperature ($T^* \approx 2.75$) than that at which the patch design was optimized ($T^* = 0.8$), indicating that the optimization of the patches was conducted well below the phase boundary for assembly. As detailed above, regardless of the temperature at which the optimization was conducted, the transition temperature may be tuned by rescaling $\varepsilon_B^{\text{opt}}$.

Characterization of the Self-assembled Lattice

We now analyze the structure of the self-assembled lattice produced in the slow temperature annealing to assess the yield of the desired cubic diamond lattice. The self-assembled crystal produced at the end of the low temperature hold is presented in Figure 3.6a. We characterize the structure by computing the radial distribution function of the geometric centers of tetramers (Figure 3.6b) and the inner product of the Steinhardt bond-order parameters (defined in equation 2.16) $\vec{q_3}(i)^* \cdot \vec{q_3}(j)$ between nearest-neighbor pairs of geometric centers of tetramers. Nearest-neighbors are defined according to a cut-off distance $d_{cut} = 12.0\sigma$ covering the first peak in the radial distribution function.

The radial distribution function shows sharp peaks at $r^* = 11.25\sigma$, 18.45σ, 21.55σ, 26.05σ and 28.35σ corresponding to the locations of the first five characteristic peaks expected for a cubic diamond lattice (Figure 3.6b). The hexagonal diamond lattice, however, possesses a nearly indistinguishable peak fingerprint that differs only in a weak splitting of the second and fourth peaks. Instead we turn to the $\vec{q_3}(i)^* \cdot \vec{q_3}(j)$ Steinhardt bond-order parameter analysis that is better able to distinguish these two polymorphs by also incorporating angular information[68]. The distribution of $\vec{q_3}(i)^* \cdot \vec{q_3}(j)$ computed over all nearest-neighbor pairs

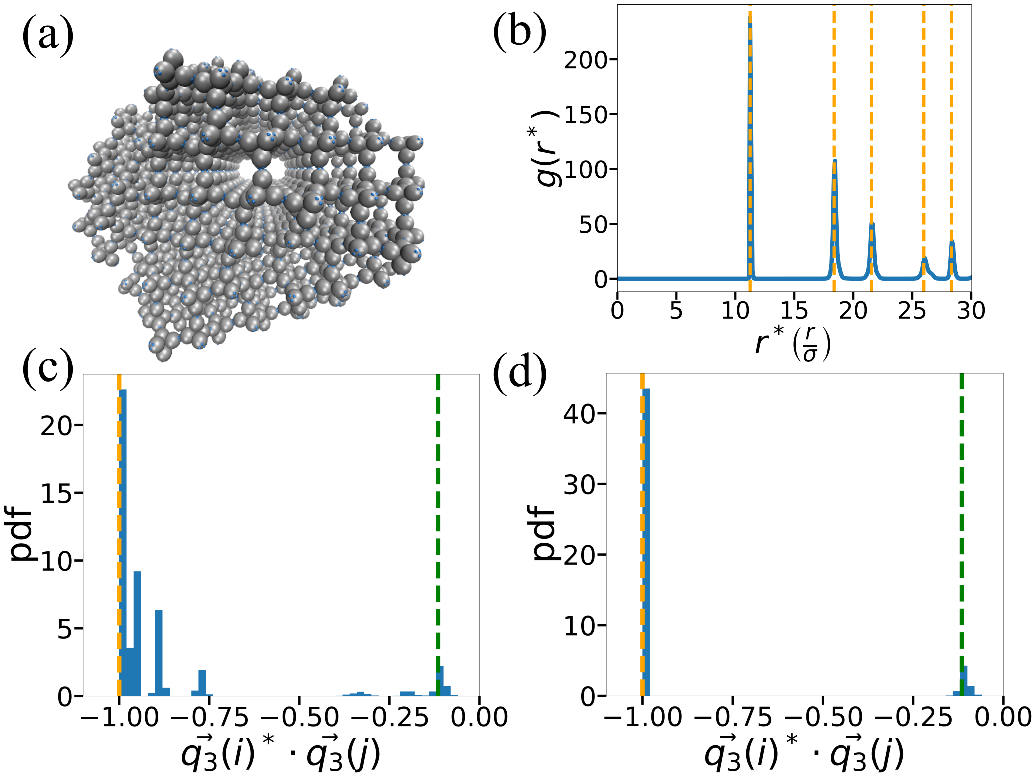


Figure 3.6: Characterization of the self-assembled crystal lattice. (a) A snapshot of the terminal self-assembled lattice. (b) Radial distribution function of the geometric centers of tetramers in the terminal lattice. The orange dashed lines correspond to characteristic peak positions of an ideal cubic diamond lattice. (c) Distribution of the inner product of Steinhardt bond-order parameters $\vec{q_3}(i)^* \cdot \vec{q_3}(j)$ between nearest-neighbor pairs of geometric centers of tetramers. The orange dashed line corresponds to the characteristic peak of cubic diamond lattice and the green dashed line corresponds to the extra peak in hexagonal diamond lattice. (d) Reproduction of panel c considering only tetramers possessing four nearest neighbors to exclude tetramers at the boundary of the finite-sized crystal.

(Figure 3.6c) exhibits a strong primary peak at (-1) and a smaller peak at (-0.115). The additional peaks between (-1) and (-0.115) are attributable to finite-size effects due to computing $\vec{q_3}(i)^* \cdot \vec{q_3}(j)$ over tetramers on the boundary of the final structure that do not possess exactly four nearest-neighbors. Excluding these boundary particles by restricting the calculation to nearest-neighbor pairs possessing exactly four nearest-neighbors eliminates these ancillary peaks (Figure 3.6d). An ideal cubic diamond lattice should possess a single peak in $\vec{q_3}(i)^* \cdot \vec{q_3}(j)$ at (-1). A hexagonal diamond lattice should possess an additional peak at (-0.115) with a magnitude one third of that of the (-1) peak[68].

We quantify the proportions of tetrahedral tetramers within cubic and hexagonal diamond environments following an approach suggested by Romano et al.[68] First, we classify a tetramer i as solid-like if it has four nearest neighbors with $\vec{q_3}(i)^* \cdot \vec{q_3}(j) \in [-1, -0.87) \cup [-0.3, 0.1)$ – where the former range identifies cubic diamond neighbors and the latter range hexagonal diamond neighbors – and each of its nearest neighbors also has four nearest neighbors. Second, we classify a solid-like tetramer i as living in a cubic diamond environment if $\vec{q_3}(i)^* \cdot \vec{q_3}(j) \in [-1, -0.87)$ for all four neighbors j. To provide a statistical estimate of these fractions, we performed 10 independent temperature annealing simulations using the method described at the beginning of this section and measured the fraction of solid-like tetramers living in a cubic diamond environment in self-assembled crystal at the termination of the low-temperature hold. We compute a mean cubic diamond fraction of 58% with a 95% confidence interval of (53%, 63%).

The root of the mixed cubic/hexagonal character of the self-assembled lattice is the small free energy difference between the two polymorphs. A short-ranged attractive model of tetramers developed by Romano et al. calculated the cubic phase to be only marginally more stable than the hexagonal phase by only 0.02 k_BT in a short-range patchy particle model[68]. In analogous work, triblock patchy colloids designed by Rao et al.[146] were observed to form a mixed pyrochlore/hexagonal tetrastack lattice upon slow temperature annealing as a result of a similarly marginal stability of the pyrochlore polymorph. In the present work, we sought to break the degeneracy between the desired cubic diamond polymorph relative to the undesired hexagonal diamond by engineering the geometry and interactions of the three "B" patches to favor a staggered interface between tetrahedral tetramers over the eclipsed interface. Although we were able to achieve 83% selectivity for the formation of staggered dimers over all competing aggregates in our simulations of colloidal monomer aggregation, this only translated to a 58% selectivity for the cubic diamond under our temperature annealing protocol.

Boosting the Cubic Diamond Fraction

We experimented with a number of ways to boost the cubic diamond fraction of the self-assembled lattice. First, we explored the sensitivity of the observed cubic diamond fraction to modifications of the patch size σ_B. It is valuable to assess the robustness of our optimal design to variations in patch size and we also reasoned that small changes could potentially elevate the cubic diamond fraction. Following the same cooling schedule described in section 3.4.2, we measured the fraction of solid-like tetramers in a cubic diamond environment for self-assembled crystals produced by our optimal particle design but now with patch sizes of $\sigma_B = 0.90\sigma$, 0.95σ, 1.05σ, 1.10σ, 1.20σ, 1.30σ and 1.40σ. For each σ_B, we performed three independent cooling simulations. The resulting cubic diamond fractions with 95% confidence intervals for those patch sizes are illustrated in Figure 3.7. Within the error bars of our calculations, the observed selectivity for cubic diamond is robust to variations in the patch size over the range $\sigma_B = 0.95\text{-}1.20\sigma$. A degradation in the observed fraction is observed outside this range at $\sigma_B = 0.90\sigma$, 1.30σ, and 1.40σ. This result indicates that our optimal design lies within a relatively flat-topped optimum with respect to perturbations in σ_B and provides post hoc validation that fixing $\sigma_B = 1.0\sigma$ produces good assembly behaviors. It is conceivable that augmenting our design space to explicitly include σ_B could potentially open

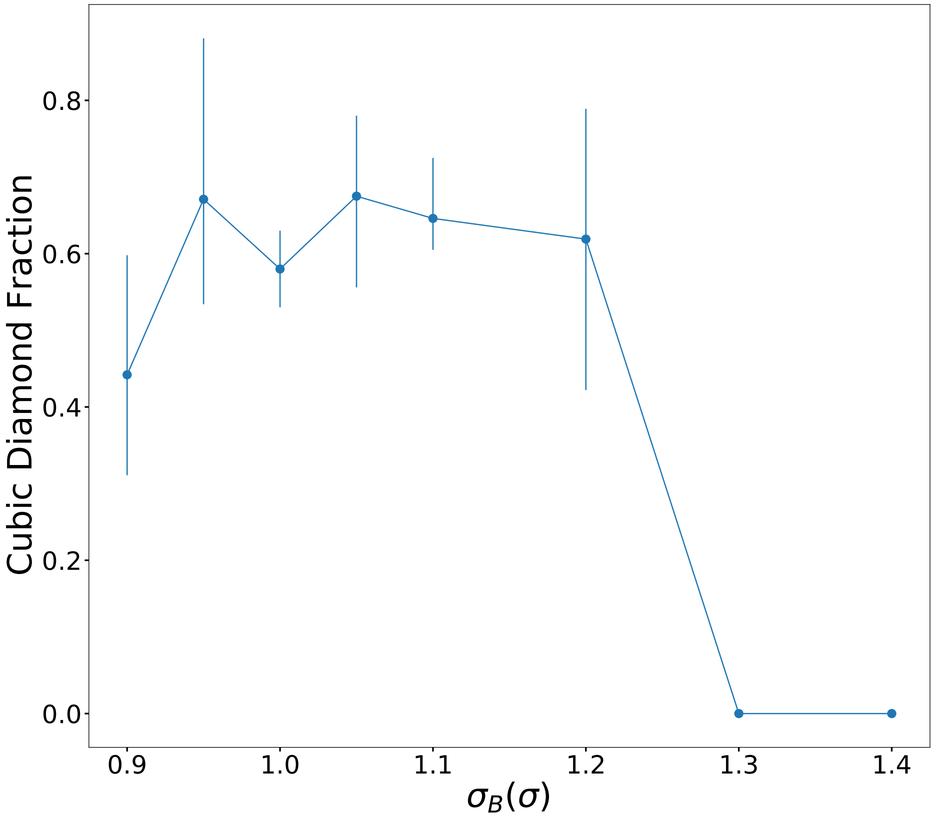


Figure 3.7: Sensitivity to patch size σ_B of the observed cubic diamond fraction produced by the optimal particle design $\{\varepsilon_B^{\text{opt}}, \phi_B^{\text{opt}}, \alpha_B^{\text{opt}}\} = \{8.18\varepsilon, 19.6°, 0.907\}$ after slow temperature annealing. The selectivity for cubic diamond within the self-assembled crystal is robust to perturbations in the patch size over the range σ_B = 0.95-1.20σ. Error bars represent 95% confidence intervals.

up directions in the 4D space of $\{\varepsilon_B, \phi_B, \alpha_B, \sigma_B\}$ along which significant improvements in the cubic diamond fraction might be observed, but the present result indicates that this cannot be achieved by modulating σ_B alone. Second, we explored the use of cubic diamond seeds to promote nucleation of the desired polymorph. Specifically, we introduced a small rigid seed of cubic diamond lattice composed of 18 tetrahedral tetramers, performed 10 independent cooling simulations for the seeded system using the same cooling schedule. This proved to be a quite successful strategy, with the cubic diamond fraction of the terminal crystal in the seeded system achieving 82% with a 95% confidence interval of (74%, 89%). This suggests that a combination of slow annealing and initial seeding may be combined to exploit the small separation in the stability of the cubic over hexagonal polymorphs to induce the assembly of defect-free cubic diamond lattices.

Band Structure Calculation

Finally, we compute the band structure of an ideal cubic diamond lattice formed by our designed tetrahedral tetramers. This calculation verifies that the patches do not disrupt the band structure of the underlying cubic diamond lattice of tetrahedral tetramers and that if a defect-free crystal lattice of these particles can be achieved, it will possess a complete band gap. We used the MIT Photonic Bands (MPB) package[135] to compute the band diagram along the corners of the irreducible region of the first Brillouin zone employing a $16 \times 16 \times 16$ grid to discretize the unit cell. For the cubic diamond lattice composed of tetramers, the lattice constant a is related to the nearest neighbor distance between colloids r_{nn} by $a = (2\sqrt{2} + \frac{4}{3}\sqrt{3})r_{\mathrm{nn}}$. We set $r_{\mathrm{nn}} = 5.05\sigma$, corresponding to the the first peak of the radial distribution function between patchy colloids ("A" particles) in the final structure. Assuming a relative permittivity between the colloidal particles and the medium of $\epsilon_r = 12.0$ corresponding to that between silicon and air, we compute the band structure in Figure 3.8a. The cubic diamond lattice does indeed possess a complete photonic band gap between the second and third bands with a ratio between gap size and midgap frequency of $\Delta\omega/\omega_m = 13.8\%$. Under the real unit mapping described in Section 3.3.4 with $\sigma_A = 5\sigma = 1$ μm, the corresponding lattice constant of $a = 5.2$ μm places the band gap in the frequency range $39 < \nu < 45$ THz and wavelength range of 6.7 $\mu\mathrm{m} < \lambda < 7.7$ μm, situating the band gap around the near-infrared regime of the electromagnetic spectrum.

We performed a corresponding calculation for the ideal hexagonal diamond lattice in which the lattice constant a is related to r_{nn} as $a = (2 + \frac{2}{3}\sqrt{6})r_{\mathrm{nn}}$. The resulting band structure assuming the same silicon/air relative permittivity of $\epsilon_r = 12.0$ is presented in Figure 3.8b. At this relative permittivity, the hexagonal diamond lattice does not possess a complete photonic band gap. By examining the band structure as a function of relative permittivity, the hexagonal diamond lattice does support a complete band gap between the fourth and fifth bands, but the band only opens for relative permittivity in excess of ϵ_r = 14.0, corresponding to the approximate relative permittivity between silicon-germanium alloy and air (Figure 3.8c). The cubic diamond lattice possesses a substantially larger ratio between gap size and midgap frequency $\Delta\omega/\omega_m$ at all values of ϵ_r, making the cubic diamond a more attractive photonic crystal than the hexagonal analogue.

3.5 Conclusions

We have performed inverse design of a patchy tetrahedral colloidal cluster that spontaneously assembles into a cubic diamond lattice with high fidelity. We stabilize the open lattice structure through the geometry and anisotropic interaction potentials of the colloidal cluster building blocks and promote the cubic diamond polymorph over the competing hexagonal diamond by rational engineering of the strength, positioning, and protrusion of the interaction patches through an iterative optimization strategy. In chapter 2, we considered a patchy spherical colloidal building block comprising nine patches of three different types and interaction complementarities that was capable of defect-free assembly into a cubic diamond lattice in slow cooling simulations. In this work, we greatly simplified the design space to a single interaction patch type upon a tetrahedral colloidal cluster that is more representative of experimentally realizable designs and amenable to existing experimental fabrication techniques. Colloidal clusters, including the tetrahedral tetramer, have been produced by a variety of experimental techniques[141], including controlled surface-nucleation of colloids onto seeds[148], advanced encapsulation emulsion techniques[8,140,149], depletion interactions[139] and crystal-templated fabrication[62]. Anisotropic interaction patches can be functionalized onto colloids using techniques such as contact area lithography[15,86,138], glancing angle deposition[12–14], grafting of DNA oligomers[8–11,102] and surface-patterning with polymeric or

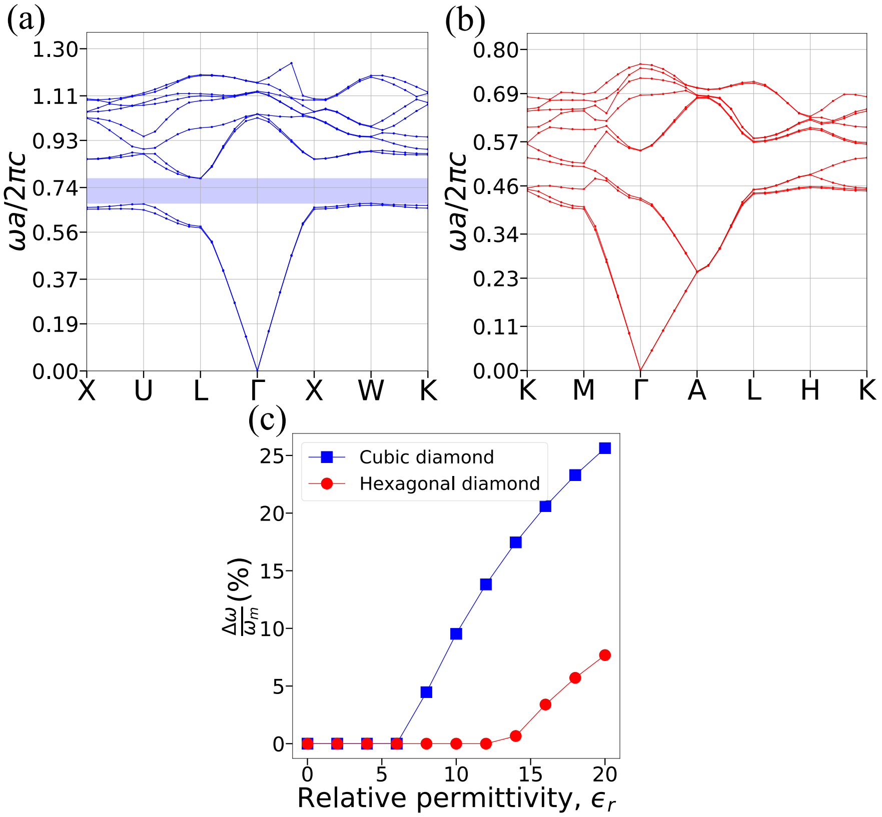


Figure 3.8: Photonic band structure of diamond lattices of the designed tetrahedral tetramers. (a) Computed photonic band diagram for the cubic diamond lattice of tetrahedral tetramers at a relative permittivity between the colloidal particles and the medium of $\epsilon_r = 12.0$. The x-axis traverses the corners of irreducible region in the first Brillouin zone. The y-axis reports the dimensionless frequency $\omega a/2\pi c$, where ω is the angular frequency, a is the lattice constant, and c is the speed of light in vacuum. The shaded bar denotes the complete band gap between the second and the third bands. (b) Computed photonic band diagram for the hexagonal diamond lattice of tetrahedral tetramers at $\epsilon_r = 12.0$. (c) Dependence of the ratio between gap size $\Delta\omega$ and midgap frequency ω_m as a function of relative permittivity ϵ_r for the cubic and hexagonal diamond lattices.

metallic patches[145]. We demonstrated the assembly of a cubic diamond lattice with 82% yield using seeded slow temperature annealing simulations of our optimal design. The ideal cubic diamond lattice composed of these colloidal particles was computationally verified to possess a complete photonic band gap. It is hoped that this computational work may guide the experimental fabrication of self-assembling building blocks to realize this material in the laboratory.

We see multiple avenues for potential future work. First, we would like to expand the design space to incorporate additional design variables within the optimization that may further promote robust assembly of the target cubic diamond lattice. The present work defined a three-dimensional optimization problem in the interaction strength, polar angle, and protrusion ratio of the surface patches, but the design strategy could be straightforwardly extended to include, for example, the size and shape of the patches to explore larger, smaller, and potentially non-spherical geometries. One might also allow for more elaborate expansions of the design space such as allowing for potentially non-isotropic patch interactions[90], deviations from the idealized tetrahedral tetramer geometry to a compressed or other imperfect geometries[84], or changes in the surface chemistry and interaction of the colloids to mimic, for example, polymer adsorption and depletion effects[150]. Second, the present study can be conceived as defining the single optimal patch design within the defined design space. Determining the influence of polydispersity in the design of the colloidal particles is important in understanding its impact on assembly fidelity and placing bounds on acceptable variabilities and imperfections in particle synthesis. We envisage a comprehensive follow-on study on the influence of polydispersity in the optimizable (i.e., patch interaction strength, polar angle, and protrusion ratio) and fixed (e.g., patch size, shape, and interaction strength; size, shape, and relative arrangement of the colloids comprising the tetrahedral tetramer) design variables in which we conduct ensembles of additional simulated annealing calculations to sample this multidimensional parameter space and explore these effects. One could also conceive of a more sophisticated design strategy where the optimization is performed such that the particle design parameters are random variables drawn from pre-defined distributions representing the anticipated polydispersity[151]. The terminal designs discovered by this strategy are likely to be inferior in assembly performance relative to the single best design discovered in the absence of polydispersity, but superior in terms of robustness to imperfections in the particle designs.

CHAPTER 4
HIGH-THROUGHPUT SCREENING OF ELASTIN-LIKE POLYPEPTIDES FOR VESICLE SELF-ASSEMBLY

4.1 Abstract

Giant lipid vesicles have been used extensively as a synthetic cell model to recapitulate various life-like processes, including in vitro protein synthesis, DNA replication, and cytoskeleton organization. Cell-sized lipid vesicles are mechanically fragile in nature and prone to rupture due to osmotic stress, which limits their usability. Recently, peptide vesicles have been introduced as a synthetic cell model that would potentially overcome the aforementioned limitations. Peptide vesicles are robust, more stable than lipid vesicles and can withstand harsh conditions including pH, thermal, and osmotic variations. In this work, we aim to use molecular simulation and machine learning techniques to perform high-throughput screening of diblock elastin-like polypeptides to search for candidate peptides that could form stable vesicular structures. Our approach could provide a systematic way of screening promising elastin-like polypeptides for specific self-assembly behavior. This chapter is based on the work reported in the following papers: (1) B. Sharma, **Y. Ma**, A.L. Ferguson, and A.P. Liu "In search of a novel chassis material for synthetic cells: Emergence of synthetic peptide compartment" *Soft Matter* 16 10769 (2020); (2) B. Sharma, **Y. Ma**, H.L. Hiraki, B.M. Baker, A.L. Ferguson, and A.P. Liu "Facile formation of giant elastin-like polypeptide vesicles as synthetic cells" *Chem. Commun.* 57 13202-13205 (2021).

4.2 Introduction

Synthetic cells are engineered biological or polymeric membranes that mimic one or many functions of a biological cell. They have a wide range of applications in understanding the fundamentals of biological cells as well as smart drug delivery[152,153]. Traditionally, lipid bilayers (liposomes) have been used as the chassis materials for synthetic cells since they resemble the natural component and morphology of biological cells[153]. However, cell-sized lipid vesicles are rather fragile[154] and sensitive to osmotic pressure[155], making their bottom-up synthesis a challenging task. Therefore, there exist opportunities to explore alternative chassis materials for the synthesis of artificial cells.

Recently, a class of materials called elasin-like polypeptides (ELPs) have drawn a lot of research interests. ELPs are synthetic biopolymers that share structural characteristics with intrinsically disordered proteins such as tropoelastin. The general structure of ELP polymers is $(\mathrm{VPGXG})_\mathrm{n}$, where V is Valine, P is Proline, G is Glycine and X can be any guest residue except Proline. ELPs are intrinsically disordered polymers that exhibit temperature-triggered phase transition: below certain lower critical solution temperature (LCST), ELP stays in a random coil configuration; as the temperature rises above LCST, ELP transforms to β-spiral configurations[23,156]. The guest residue X is one of the key component in determining the LCST; usually the LCST decreases as the hydrophobicity of the guest residue increases[23]. It is for this reason that amphiphilic diblock and multiblock ELPs have drawn a lot of research interests because their self-assembly behavior could be manipulated by controlling the guest residues in each block and thus controlling the LCST of each block. Micelles are the most common self-assembled structures from amphiphilic diblock ELPs, where the hydrophobic block has lower LCST and the hydrophilic block has higher LCST. At the temperature between these LCSTs, the hydrophobic blocks associate with each other to form the core of micelles while the hydrophilic blocks form micelle corona[157]. However, several recent experiments have indicated that amphiphilic diblock and triblock ELPs could self-assemble into large vesicular structures that are stable under extreme conditions such as extreme pH or temperature[158]. For example, Vogele et al. have utilized glass bead method to direct a diblock ELP with glutamic acid as the hydrophilic guest residue and phenylalanine as the hydrophobic guest residue to form giant vesicles[22]. Schreiber et al. compared the vesicular structures formed by two kinds of diblock ELPs with the same length but different guest residues and conclude that guest residue composition could be a key factor in modulating the vesicle stability[159]. Frank et al. demonstrated the formation of giant ELP vesicles by using solvent evaporation method[21].

Although there are some experimental efforts in producing self-assembled vesicles from ELPs, the current experiments choose ELP candidates by physically and chemically inspired intuition. Therefore, there exist opportunity to build a more systematic approach for the high-throughput screening of candidate ELPs that could form stable vesicular structures. Recent advances in machine learning have provided novel tools for the prediction of chemo-physical properties of peptides and the rational design of peptides to optimize those properties. For instance, kernel regressions[160,161], support vector machines (SVM)[162] and artificial neural networks[163] have been deployed for the classification of peptides, prediction of peptide properties and design of novel peptide sequences that suit particular purposes. Leslie et al. proposed mismatch kernels based on a tree data structure to perform SVM classification of proteins for several benchmark tasks and demonstrated good performance[40]. Lee et al. used SVMs to identify and discover new membrane-active and antimicrobial α-helical peptides[164]. Zhou et al. used Gaussian Process Regression (GPR) model with custom kernel to predict the antimicrobial abilities of various pentadecapeptides[160]. Thurston and Ferguson proposed a quantitative structure–property relation model to perform extensive screening over π-conjugated oligopeptides and selected promising candidates that can self-assemble into nanoaggregates with desired optoelectronic properties[165]. Yang et al. employed doc2vec model from natural language processing to embed proteins into a vector space on which they performed Gaussian process regression (GPR) on benchmark tasks and nonlinear dimensionality reduction to evaluate the performance of protein embedding[41]. Shmilovich et al. trained variational autoencoders to embed π-conjugated peptides into a low-dimensional vector space over which GPR model is constructed to support Bayesian optimization discovery of new π-conjugated peptides that can spontaneously form nanostructures with emergent optoelectronic properties[30]. In general, machine learning techniques can assist the discovery of novel promising materials by building predictive or generative models from existing data, thus making it a powerful tool to complement and guide simulation and experiment in the rational design of peptides that may be suitable as a novel chassis materials for synthetic cells.

In this work, we have proposed a high-throughput screening procedure that combines molecular dynamics simulation, free energy calculation and Bayesian optimization to select out optimal peptide from a library of diblock ELPs that could form stable vesicular structures. To the best of our knowledge, this is the first attempt to use modern machine learning techniques to perform high-throughput screening of diblock ELPs for the vesicle self-assembly. We envisage that this framework could be easily extended to the rational design of multiblock ELPs for various tasks.

4.3 Methods

4.3.1 Molecular Modeling of ELP Vesicle

In this work, we primarily considered diblock amphiphilic ELP sequences as our search space where one block contains a hydrophilic guest residue and the other block contains a hydrophobic guest residue. The hydrophobicity scales of amino acids were taken from the Kyte-Doolittle scale[166]. Because an entire vesicle is too large to simulate, we instead focused on a zoomed-in region of the vesicle which could be approximated as a planar bilayer (figure 4.1). We used PyMol[167] software to construct an all-atom representation of the ELP chain and then used Martini force field version 2.2[168] to coarse-grain it. When coarse-graining the ELP, we set the secondary structure of the hydrophilic block (i.e. block with a hydrophilic

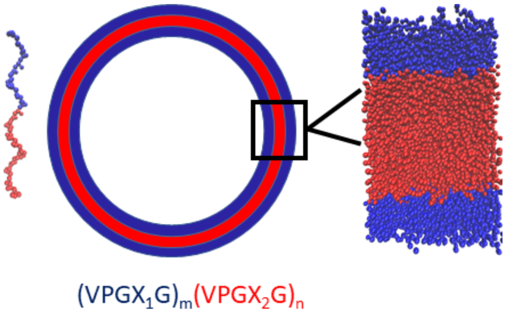


Figure 4.1: Illustration of ELP vesicle and bilayer. The hydrophobic block is shown in red and the hydrophilic block is shown in blue. The ELP bilayer (right) represents an approximate zoomed view of the vesicle.

guest residue) to random coil and the secondary structure of the hydrophobic block to β-turn as experimentally observed for many ELPs[23]. We inserted a 10×10 gird of diblock ELP chains separated by 0.1 nm to create the upper layer and another 10×10 grid of ELP chains to create the lower layer. The ELP chains in the lower layer were flipped upside down so that the hydrophobic blocks of these two layers were close to each other to form the middle part of the bilayer. We then added coarse-grained water molecules and ions that were also modeled by Martini force field to the regions above and below the bilayer. A complete initial setup of the bilayer is illustrated in figure 4.2. After setting up the initial bilayer, we first ran steepest descent energy minimization to relax the system. After energy minimization, we performed 10 ns NVT equilibration at 300 K. Then we performed 10 ns NPT equilibration at 300 K and 1 bar. Finally, we performed 200 ns NPT production simulation at 300 K and 1 bar to prepare the system for subsequent umbrella sampling simulations which are detailed in section 4.3.2. For all simulations, the temperature was controlled by a stochastic velocity re-scaling algorithm[169]. For NPT equilibration run, Berendsen barostat[34] was used while for NPT production run Parrinello-Rahman barostat[170] was used. The coupling type of pressure coupling was semi-isotropic, which was isotropic in x- and y- directions (lateral directions of bilayer) but different in z-direction (normal direction of bilayer), and the time constant for pressure coupling was 12 ps. The step size for all simulations was set to 20 fs.

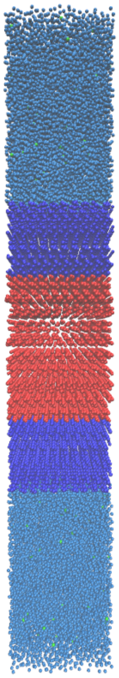


Figure 4.2: An illustration of initial bilayer setup. The dark blue beads represent hydrophilic blocks, the red beads represent hydrophobic blocks, the light blue bead are coarse-grained water and the green beads are coarse-grained ions.

All molecular dynamics simulations were performed using Gromacs 2019[171].

4.3.2 Quantitative Characterization of Stability of Diblock ELP Vesicle

The thermodynamic stability of self-assembled structures of peptides and proteins could be estimated by evaluating the free energy of dissociation of the structure. One way to estimate this free energy is to pull out one molecule from the self-assembled structure and mapping out the potential of mean force (PMF) of this dissociation process (figure 4.3). The PMF reflects the free energy profile along a specific pulling coordinate and the free energy difference between the starting (bounded) and ending (free) state could reflect the stability of the self-assembled structures. For example, Lemkul and Bevan have performed molecular dynamics simulation to determine the PMF of extracting constituent peptide from Alzheimer's amyloid protofibril to assess the stabilities of these fibrils[172]. Sevgen et al. used

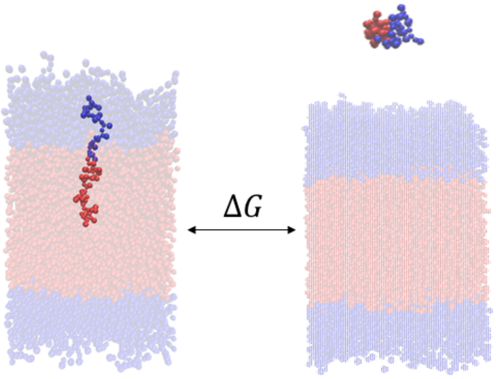


Figure 4.3: Illustration of the free energy of dissociation ΔG. The bilayer is made transparent to emphasize the pulled ELP chain.

similar technique to estimate the stability of micelles formed by block copolymers composed of OES and PEG blocks[173].

Due to the possible existence of many local free energy minima that the system could be trapped in, enhanced sampling techniques are usually used to facilitate sufficient sampling of all relevant configurations to ensure good estimate of free energy profile[174]. There exist many techniques to perform enhanced sampling, including metadynamics[175], adaptive force bias[176] and umbrella sampling[110]. Since enhanced sampling methods usually introduce biasing force and potential to help the system escape local free energy minima, the resulting trajectories need to be post-processed to obtain unbiased estimate of the free energy profile and numerous analysis techniques, such as Multistate Bennett Acceptance Ratio (MBAR)[177] and Weighted Histogram Analysis Method (WHAM)[178], have been proposed to perform that task. Here, we use a combination of umbrella sampling and WHAM to construct the PMF.

After creating and equilibrating the ELP bilayer as detailed in section 4.3.1, we randomly chose a peptide chain from the bilayer as the chain to be pulled out. We chose two reaction coordinates, $\mathrm{z_{head}}$ and $\mathrm{z_{tail}}$, as described in figure 4.4a. $\mathrm{z_{head}}$ is the z-component of displacement from the COM of the bilayer to the COM of hydrophilic block of chosen chain and $\mathrm{z_{tail}}$ is the z-component of displacement from the COM of the bilayer to the COM of hydrophobic block of the chosen chain. The pulling process was divided into three stages (figure 4.4b): In the first stage, the relative displacement between bilayer and hydrophilic block of the chosen chain was constrained by high harmonic potential and the hydrophobic block of the chosen chain was gradually pulled away from the bilayer. In the second stage, both hydrophilic and hydrophobic blocks were pulled away from the bilayer to extract the chosen chain out of the bilayer. This process ensures that the chosen chain does not undergo sudden conformation change when it comes out of the bilayer and avoids hysteresis in PMF due to insufficient sampling of orthogonal degrees of freedom[179]. In the final stage, the hydrophilic block was pulled out while the hydrophobic block was constrained by harmonic potential. The purpose of the last stage is to allow the peptide chain to relax and estimate the equilibrium chain length in bulk solvent. After all of the pulling simulations were completed, we placed

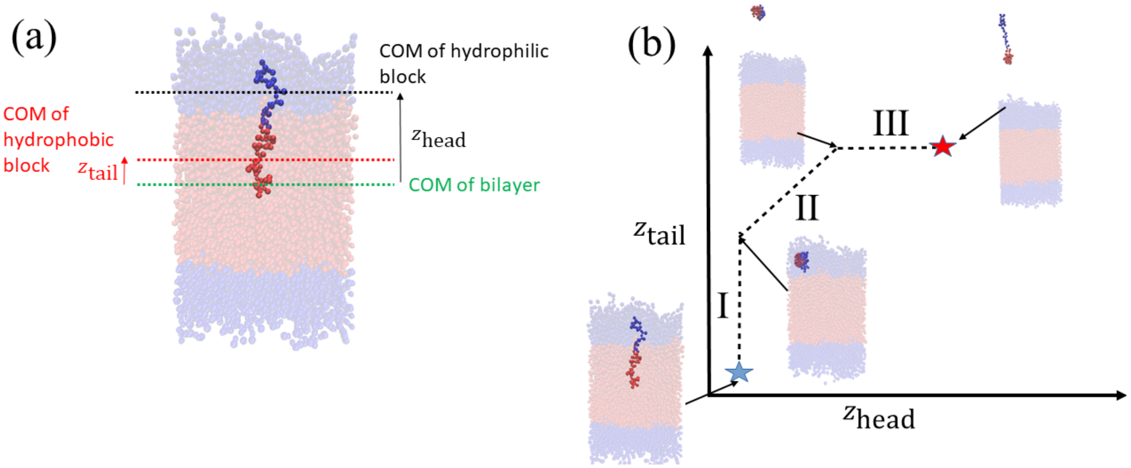


Figure 4.4: Illustration of reaction coordinates and the path over which the PMF is constructed. (a) Definition of reaction coordinates. The bilayer is made transparent to highlight the opaque chosen peptide chain. The hydrophobic blocks are shown in red and the hydrophilic blocks are shown in dark blue. The green dashed line represents the COM of the bilayer, the red dashed line represents the COM of the hydrophobic block of the chosen chain and the black dashed line represents the COM of the hydrophilic block of the chosen chain. (b) Illustration of the path over which the PMF is constructed. Solvents are not shown in all plots for clarity. Stage I is the stage where the hydrophilic block is constrained and hydrophobic block is pulled up. Stage II is the stage where both blocks are pulled up to extract the chain out of the bilayer. Stage III is the stage where the hydrophilic block is again pulled up to estimate the equilibrium chain length in bulk solvent.

equidistant centers for harmonic potential along the path and selected out snapshots from the pulling simulations closest to each center as the starting configuration for each umbrella sampling simulation. We then introduced harmonic potential of the form

$$W(\{z_{\text{head}}, z_{\text{tail}}\}, \{z^*_{\text{head}}, z^*_{\text{tail}}\}) = \frac{1}{2}k_1(z_{\text{head}} - z^*_{\text{head}})^2 + \frac{1}{2}k_2(z_{\text{tail}} - z^*_{\text{tail}})^2 \tag{4.1}$$

where $\{z^*_{\text{head}}, z^*_{\text{tail}}\}$ are the centers for the harmonic potential, and performed 100 ns umbrella simulations around each center under T = 300 K and P = 1 bar. We fine-tuned the spacing between adjacent umbrella windows and the strength of harmonic constraints to ensure good overlaps of histograms from each umbrella sampling simulation. The histograms of reaction coordinates $\{z_{\text{head}}, z_{\text{tail}}\}$ recorded during the umbrella sampling simulations were then processed by WHAM to reconstruct the unbiased PMF. The free energy of dissociation(ΔG) is taken as the difference between the minimum free energy when the peptide chain is embedded in the bilayer and the minimum free energy when the chain exists in bulk solvent:

$$\Delta G = G_{\text{bounded}} - G_{\text{solvent}} \tag{4.2}$$

In all pulling simulations, the time step was set to 20 fs. In stage I pulling, a harmonic potential of the form 4.1 was applied to the chosen chain. Both spring constants (k_1 and k_2) were chosen to be 15,000 kJ mol^{-1} nm^{-2}. The harmonic center z^*_{head} was fixed at the starting value of z_{head} and z^*_{tail} was gradually increased from the starting value with a rate of 5×10^{-5} nm ps^{-1} until it reaches z^*_{head}. In stage II pulling, both spring constants were set to 20,000 kJ mol^{-1} nm^{-2} and both harmonic centers were increased from their starting values with a rate of 5×10^{-5} nm ps^{-1} until the chain leaves the upper layer. In stage III pulling, both spring constants were set to 5000 kJ mol^{-1} nm^{-2}. z^*_{tail} was kept fixed at the starting value of z_{tail} and z^*_{head} was increased from the starting value with a rate of 5×10^{-5} nm ps^{-1} for 80 ns. In all pulling simulations, temperature was controlled by stochastic velocity re-scaling and pressure was controlled by Parrinello-Rahman barostat.

For all subsequent umbrella sampling simulations, the harmonic centers were kept fixed and we dynamically adjusted the spring constants based on whether the chain is embedded in the bilayer (larger spring constants) or exists in bulk solvent (smaller spring constants). In each umbrella sampling simulation, we first performed 40 ns NPT equilibration with stochastic velocity re-scaling thermostat and Berendsen barostat followed by a 100 ns production run with stochastic velocity re-scaling thermostat and Parrinello-Rahman barostat. All simulations were performed using Gromacs 2019[171] and the WHAM analysis was done using the program developed in Grossfield Lab[180].

4.3.3 Gaussian Process Modeling and Bayesian Optimization of ΔG

Having characterized the free energy of dissociation ΔG of single peptide chain from preassembled bilayer, we want to minimize this ΔG (i.e. making it as negative as possible) with respect to the ELP sequence so as to maximize the stability of vesicle. This could be again viewed as a black-box optimization problem, where the target function ΔG is related to the inputs (ELP sequence) in an unknown way, and the only thing we could do is to run simulations for any given input to calculate the target function value. The CMA-ES algorithm mentioned ins previous chapters is primarily designed to handle black-box optimization of continuous numeric inputs. However, in the current setting, the inputs are peptide sequences which are essentially strings of amino acids. Since the inputs are not numeric, we seek other way to perform the black-box optimization.

Bayesian optimization is a widely used black-box optimization technique[55]. It is based on building a surrogate predictive model for the target function that is better suited for optimization over the search space. It then employs an “acquisition function” calculated based on the surrogate model to guide the search toward most promising candidates. Since the surrogate model is an approximation of the target function, it has some uncertainty on the prediction of target function values and the acquisition function would propose promising candidates under this uncertainty. The most commonly used surrogate model is the Gaussian Process Regression(GPR) model[55]. The GPR model assumes the target function $f(x)$ is the realization of a Gaussian Process over its inputs x with mean 0 and covariance function given by a kernel K that acts on pair of inputs. That is, given a set of inputs $X = \{x_1, .., x_n\}$, the target functions $\vec{f} = \{f(x_1), f(x_2), ..., f(x_n)\}$ follow the multivariate Gaussian distribution:

$$\vec{f} \sim \mathcal{N}(\vec{0}, K(X, X)) \tag{4.3}$$

where $K(X, X)$ is the $n \times n$ Gram matrix whose components are $K(x_i, x_j)$. Now, given a set of training data $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ where each $y_i = f(x_i) + \varepsilon_i$ is noisy observation of $f(x_i)$ and ε_i are independent errors following $\mathcal{N}(0, \sigma_i^2)$ distributions, we could get the joint distribution of training $\vec{y}$ and the target function values $\vec{f^*}$ at m testing points $X^* = \{x_1^*, ..., x_m^*\}$:

$$\begin{bmatrix} \vec{y} \\ \vec{f^*} \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} K(X, X) + \Sigma & K(X, X^*) \\ K(X^*, X) & K(X^* X^*) \end{bmatrix}\right) \tag{4.4}$$

where $\Sigma = \text{diag}(\sigma_1^2, ..., \sigma_n^2)$. Therefore, the posterior predictive distribution of $\vec{f^*}$ given the training data is

$$\vec{f^*} | \mathcal{D}, X^* \sim \mathcal{N}\left(\vec{\mu}, \text{cov}(\vec{f^*})\right) \tag{4.5}$$

where $\vec{\mu}$ and $\text{cov}(\vec{f^*})$ are given by

$$\begin{aligned} \vec{\mu} &= K(X^*, X)[K(X, X) + \Sigma]^{-1}\vec{y} \\ \text{cov}(\vec{f^*}) &= K(X^*, X^*) - K(X^*, X)[K(X, X) + \Sigma]^{-1}K(X, X^*) \end{aligned} \tag{4.6}$$

Usually, the kernel function contains some parameters $\vec{\theta}$, and these parameters are optimized by maximizing the log-likelihood of training data[181],

$$\begin{aligned} l(\mathcal{D};\vec{\theta}) &= \log p(\vec{y}|X;\vec{\theta}) \\ &= -\frac{1}{2}\vec{y}^T[K(X,X;\vec{\theta})+\Sigma]^{-1}\vec{y} - \frac{1}{2}\log|K(X,X;\vec{\theta})+\Sigma| - \frac{n}{2}\log 2\pi. \end{aligned} \tag{4.7}$$

The key component in GPR is the kernel function K. It needs to be positive definite, meaning that for any set of inputs $X = \{x_1, ..., x_n\}$ the Gram matrix $K(X, X)$ must be positive semidefinite. Moreover, in our case the inputs are amino acid strings, so that we need kernels that operate on string data. Many string kernels have been proposed for peptide and protein data. For example, the Hamming distance kernel measures the minimum number of substitutions to make two equal length strength the same. The weighted degree kernel[182] computes similarity between two equal length sequences by counting the co-occurrences of k-mers at corresponding positions of two sequences. However, many of the proposed kernels only operate on strings with fixed length. In our case, the number of hydrophilic and hydrophobic blocks could vary so the ELP sequences could have varying lengths. Thus, we need a string kernel that could handle amino acid sequences with different lengths. In 2013, Giguère et al. proposed a string kernel, termed the “generic string kernel”, that could serve this purpose[161]. The generic string kernel first featurizes each amino acid into a d-dimensional vector of chemical properties:

$$\vec{\psi}(a) = \begin{bmatrix} \psi_1(a) & \psi_2(a) & \cdots & \psi_d(a) \end{bmatrix} \tag{4.8}$$

As such, a string of l amino acids $(a_1, \cdots, a_l)$ could be encoded into a $d \times l$ matrix:

$$\Psi(a_1, \cdots, a_l) = \begin{bmatrix} \vec{\psi}(a_1), \cdots, \vec{\psi}(a_l) \end{bmatrix} \tag{4.9}$$

Now, given two amino acid strings x and x', the generic string kernel with parameters

$\{L, \sigma_p, \sigma_c\}$ gives:

$$GS(x, x'; L, \sigma_p, \sigma_c) = \sum_{l=1}^{L} \sum_{i=0}^{|x|-l} \sum_{j=0}^{|x'|-l} e^{-\frac{(i-j)^2}{2\sigma_p^2}} e^{-\frac{||\Psi(x[i:i+l]) - \Psi(x'[j:j+l])||^2}{2\sigma_c^2}} \tag{4.10}$$

In other words, the generic string kernel look at all substrings of maximum size L and compute the similarity between these substrings (second exponent inside triple-summation) conjugated with a position-dependent term (first exponent) that decays exponentially. The parameter L controls the maximum length of substrings to be compared, and σ_p and σ_c control the penalties due to shifts in positions and difference in chemical properties of substrings, respectively. The parameters $\{\sigma_p, \sigma_c\}$ could be optimized by taking the derivative of log-likelihood (equation4.7) and performing a gradient-based optimization. The parameter L is found by trying out values from 1 to the length of shortest ELP in the training data and using the value that maximize the log-likelihood. We use the BLOSUM62 matrix[183] as the featurization of amino acids.

Having defined the GPR model, we can then proceed to the Bayesian optimization. The basic procedure of Bayesian optimization is summarized in algorithm 2. The key part is to maximize the acquisition function $u(x|\mathcal{D})$, which guides our search under the uncertainty of the predictions on true target function given by GPR model. There are many ways to define the acquisition function. A common choice is the expected improvement (EI)[184] which is

Algorithm 2 Bayesian Optimization

```
Initialize   Training data $\mathcal{D}_0$
for $t = 1 \cdots T$ do
    Build GPR model based on $\mathcal{D}_{1:t-1}$
    Find $x_t = \arg\max u(x|\mathcal{D}_{1:t-1})$
    Evaluate the (noisy) target function $y_t = f(x_t) + \varepsilon_t$
    Augment training data $\mathcal{D}_{1:t} = \mathcal{D}_{1:t-1} \cup \{(x_t, y_t)\}$
end for
Output $x^*$ with minimum target function value in $\mathcal{D}_{1:T}$
```

defined as:

$$\begin{aligned} EI(x|\mathcal{D}) &= \mathbb{E}[\max(f^{\dagger} - \xi - f, 0)] \\ &= (f^{\dagger} - \mu(x) - \xi)\Phi\left(\frac{f^{\dagger} - \xi - \mu(x)}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{f^{\dagger} - \xi - \mu(x)}{\sigma(x)}\right) \end{aligned} \tag{4.11}$$

where $f(x) = \Delta G(x)$ is the predicted target function value following the posterior predictive distribution at x (equation 4.5) with mean $\mu(x)$ and standard deviation $\sigma(x)$ given by the GPR model trained on $\mathcal{D}$. $f^{\dagger}$ is the minimum target function value in $\mathcal{D}$ and ξ is a hyperparameter controlling the exploration-exploitation trade-off: higher value of ξ tends to favor regions in input space with high posterior variance $\sigma(x)^2$ and lower ξ tends to favor input space with lower posterior mean $\mu(x)$[57]. An intuitive way of interpreting equation 4.11 is that it represents the expected amount of reduction from previous minimum: $max(f^{\dagger} - \xi - f, 0)$ is equal to the reduction $f^{\dagger} - \xi - f$ only if $f < f^{\dagger} - \xi$. The candidate with the greatest expected amount of reduction then becomes the next one to consider since we focus on minimizing the target function.

At each iteration of Bayesian optimization, we trained a GPR model based on the currently explored peptides $\mathcal{D}_{1:t}$. Then, for all the unexplored peptides, we could get the posterior predictive distribution over their ΔG values by equation 4.5 and evaluate the acquisition function (equation 4.11) for each of them. Then we selected one single peptide with the maximum acquisition function value as the next peptide to explore. It should be noted that there exist choices to propose multiple candidates to be explored next during one iteration of Bayesian optimization[30] in order to maximally utilize available computing resources. However, in our case we already maximally utilized available computing resources to parallelize the umbrella sampling simulations in section 4.3.2. Proposing multiple candidates at the same time would not speed up the process since the simulations for one peptide have already consumed all available computing resources.

We performed the Bayesian optimization and keep monitoring the coefficient of determination (R^2) of GPR models during the fitting process (by leave-one-out cross validation). We also monitored the Bhattacharyya distance[185] between posterior Gaussian distributions (equation 4.5) returned by successive GPR models on a testing set that are not used during training. The Bhattacharyya distance is a statistical distance that measures the similarity between two probability distributions and is defined as:

$$D_B(p, q) = -\ln\left(\int \sqrt{p(x)q(x)}dx\right) \tag{4.12}$$

These two metrics help to identify whether the GPR models given by the Bayesian optimization have saturated and the optimization could be stopped[186]. A schematic of the high-throughput screening pipeline is shown in figure 4.5.

4.4 Results

We gathered a search space of 168 candidate ELPs of the form $(\mathrm{VPGX_1G})_5(\mathrm{VPGX_2G})_\mathrm{n}$, where $\mathrm{X_1}$ is a hydrophilic guest residue, $\mathrm{X_2}$ is a hydrophobic guest residue and $n = 4, 5$. This choice for the sizes of hydrophilic and hydrophobic blocks was made because many experimental works on forming ELP vesicles use nearly equal number of hydrophilic and hydrophobic blocks[21,187]. In these experiments, usually longer ELP chains are used (for example, Schreiber et al.[187] tested diblock ELPs with more than 70 blocks). Since longer chains require much longer simulation time to equilibrate, we decided to try shorter chains with nearly equal number of hydrophilic and hydrophobic blocks to keep the relative size of hydrophilic and hydrophobic blocks similar to experiments, and made the assumption that the trends of ΔG that we see for shorter chains reflect the trends of ΔG for longer ones that are usually considered in experiments.

We started the screening process by generating an initial set of 20 candidates randomly selected from the search space. For each candidate in the initial training set, we performed

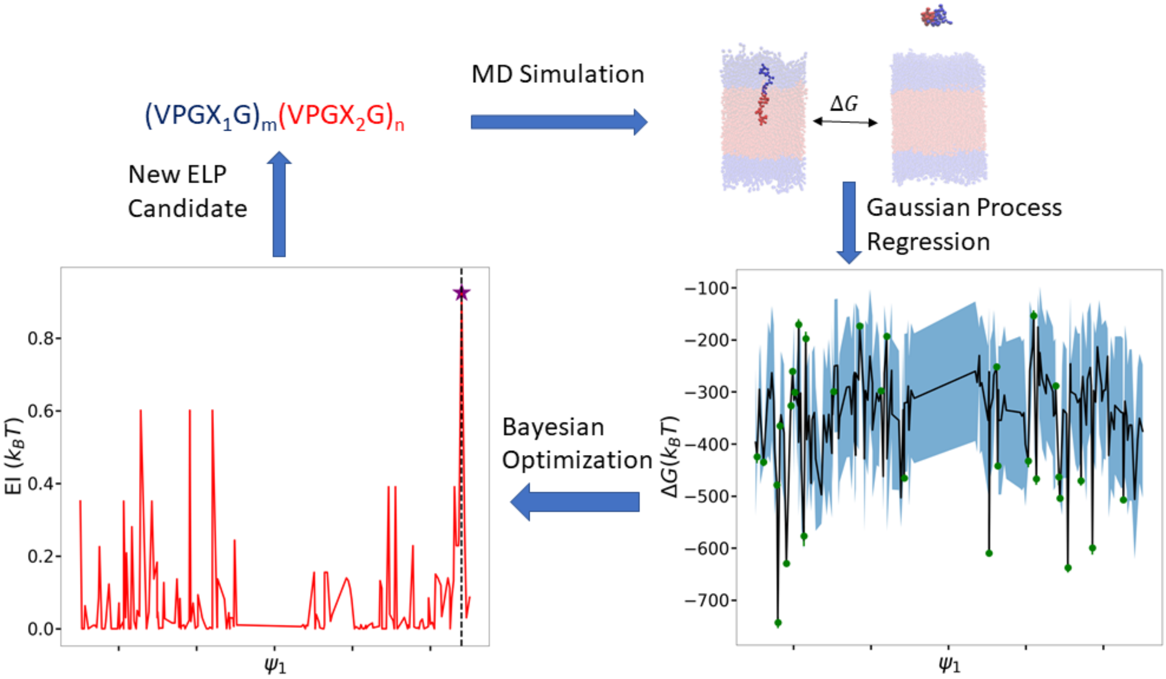


Figure 4.5: Pipeline of high-throughput screening. The top left panel shows a diblock ELP where X_1 is a hydrophilic guest residue and X_2 is a hydrophobic residue. The lower right panel shows an example of GPR model for ΔG. The green dots are training data. The black line and blue shaded area are posterior mean and variance of the prediction, respectively. The lower left panel shows an example of expected improvement acquisition function. The black dashed line and the purple star shows the candidate with maximum acquisition function value, which is the one that is selected to evaluate next. The x-axes, ψ_1, in the plots of GPR model and acquisition function represent the projection of all ELPs onto a one-dimensional space by multidimensional scaling.

umbrella sampling simulations as described in section 4.3.2. After we obtained the two-dimensional PMF in terms of reaction coordinates z_{head} and z_{tail}, we further projected the PMF onto the one-dimensional reaction coordinate z_{com}, which describes the z-component of the relative displacement between the COM of the entire pulled peptide and COM of bilayer using the projection algorithm described in reference 128. This projection of PMF onto an one-dimensional reaction coordinate makes the visualization of PMF easier. In figure 4.6, we show an example of original two-dimensional PMF and one-dimensional projected PMF for the sequence $(VPGYG)_5(VPGCG)_4$. As expected, the value of PMF gradually rises when the chain is pulled from the bilayer out to the bulk solvent. Once the pulled chain

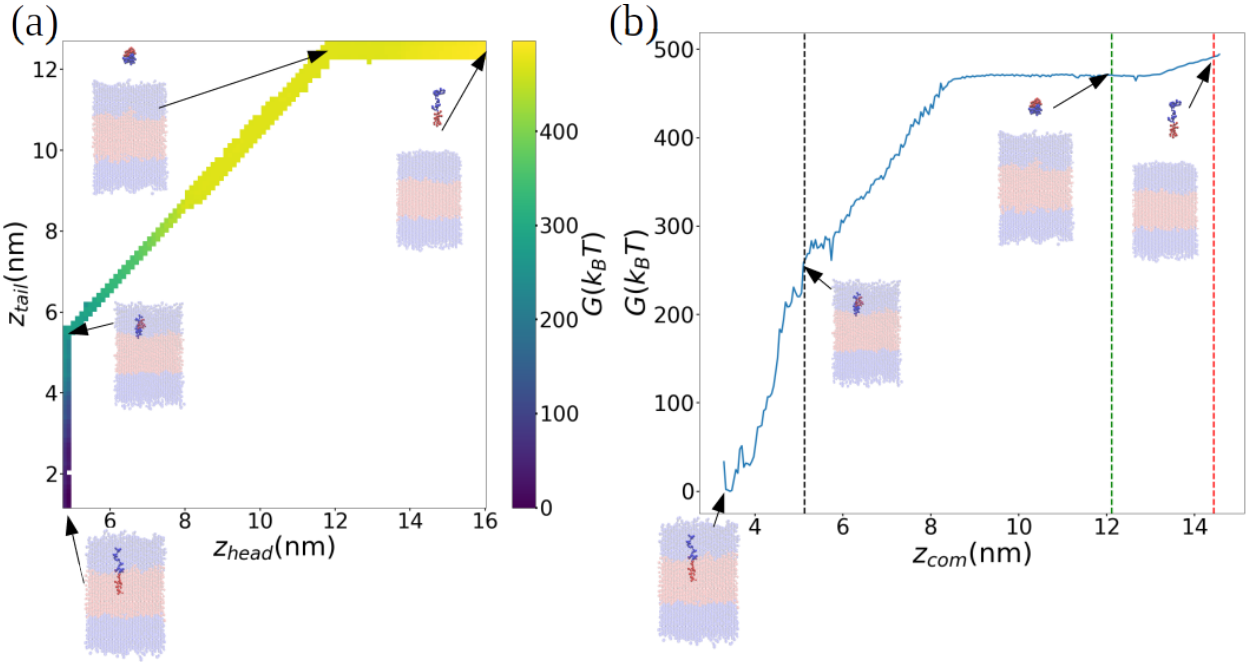


Figure 4.6: Example PMF for $(\text{VPGYG})_5(\text{VPGCG})_4$ system with snapshots to identify the stage of the pulling process. The hydrophilic $(\text{VPGYG})_5$ blocks are shown in blue and the hydrophobic $(\text{VPGCG})_4$ blocks are shown in red. The bilayer is made transparent to highlight the opaque pulled chain. Solvents are not shown for clarity. (a) Two-dimensional PMF in the reaction coordinates $\{\text{z}_{\text{head}}, \text{z}_{\text{tail}}\}$ that are directly used in the harmonic potential during umbrella sampling. (b) The projection of the two-dimensional PMF onto the one-dimensional reaction coordinate z_{com}. The black, green and red dashed lines represent the ends of each of the three stages of pulling as defined in figure 4.4.

enters the bulk solvent, the PMF becomes flat then rises again once the hydrophilic block is pulled away from the hydrophobic block. The ΔG value for this specific ELP sequence is calculated as the difference between the free energy minimum when the pulled chain is within the bilayer and the free energy minimum when the pulled chain is in bulk solvent, which is $\Delta G = -462.9 \pm 16.5 k_B T$ where the error is obtained by block analysis.

We show the prediction of ΔG for all peptides in the search space made by the terminal GPR model in figure 4.7. We project each peptide onto a one-dimensional space using multi-dimensional scaling[188] based on the Jukes-Cantor distance between biological sequences[189]. We monitored the evolution of R^2 of all GPR models and the Bhattacharyya distance be-

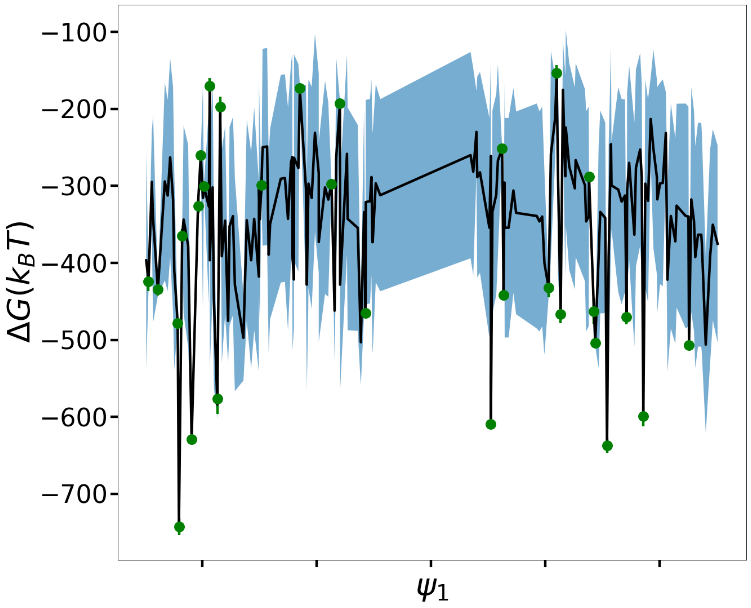


Figure 4.7: Demonstration of the GPR model. The black solid line is the mean of posterior predictive distribution and the blue shaded area represents one standard deviation. The green dots are the training data. The x-axis, ψ_1, represents the projection of all ELPs onto a one-dimensional space by multidimensional scaling.

tween successive GPR models. The plots of R^2 and the Bhattacharyya distance are shown in figure 4.8. Both quantities converge after six generations. After running up to 10 iterations, we assert that the Bayesian optimization has converged and have explored 30 candidates (roughly 18% of entire search space) during the search. We ranked the explored candidates in increasing order of their ΔG values obtained in simulation, and the top 10 candidates are shown in table 4.1. We also ranked all 168 ELPs in increasing order of their predicted ΔG values made by the terminal GPR model and show top 10 ELPs in table 4.2. From table 4.1 and table 4.2, we could see that the Bayesian optimization proposes the sequence (VPGHG)$_5$(VPGVG)$_5$ as the candidate that could form the most stable self-assembled vesicle, which has not been tested in experiments before. The experimentally verified sequence (VPGHG)$_5$(VPGLG)$_4$[187] was ranked among the top ten candidates by the terminal GPR model. Therefore, our approach could discover previously unknown sequences, while having some agreements with current experimental result. For future work, we would work with our

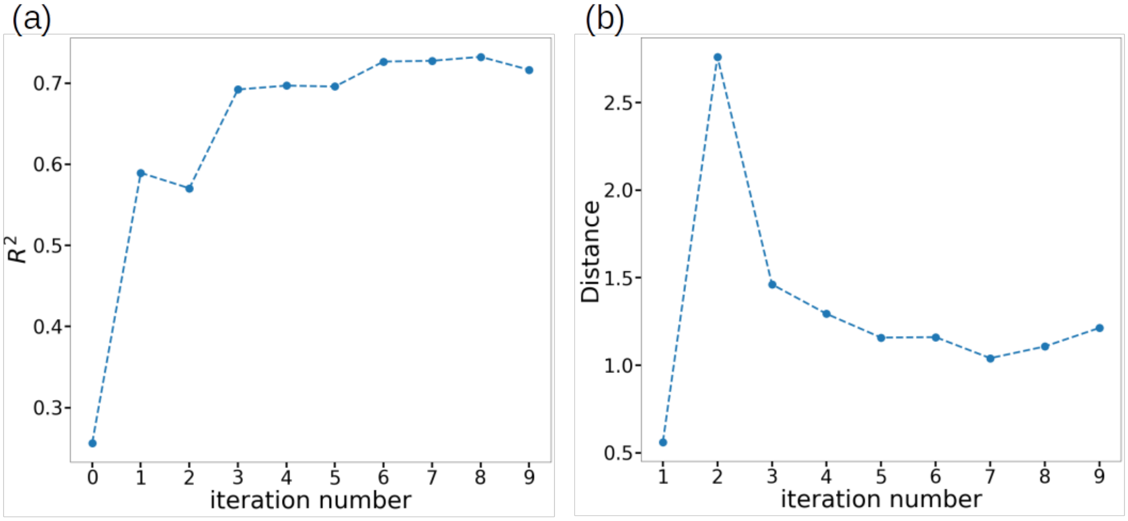


Figure 4.8: Plots of the R^2 values of GPR models and the Bhattacharyya distance between successive GPR models in the Bayesian optimization.

ELP sequence	Computed $\Delta G(k_B T)$	Discovery Iteration	Previously Known?
H_5V_5	-742.57 ± 10.73	2	N
H_5F_4	-637.01 ± 9.94	5	N
H_5F_5	-629.27 ± 3.55	6	N
H_5V_4	-609.56 ± 3.69	3	N
H_5L_4	-599.24 ± 12.87	0	Reference 187
H_5L_5	-576.68 ± 19.70	1	N
H_5C_4	-506.8101 ± 4.91	4	N
Y_5F_4	-503.64 ± 4.60	0	N
Y_5V_5	-478.12 ± 1.94	0	N
H_5A_4	-470.43 ± 9.14	7	N

Table 4.1: Table of top 10 candidates among 30 explored diblock ELPs and their computed ΔG values. Abbreviation $(\mathrm{X}_1)_\mathrm{m}(\mathrm{X}_2)_\mathrm{n}$ stands for $(\mathrm{VPGX}_1\mathrm{G})_\mathrm{m}(\mathrm{VPGX}_2\mathrm{G})_\mathrm{n}$. Uncertainties are measured via block averaging.

ELP sequence	Predicted $\Delta G(k_BT)$	Computed $\Delta G(k_BT)$
H_5V_5	-738.13 ± 10.52	-742.57 ± 10.73
H_5F_4	-636.64 ± 9.81	-637.01 ± 9.94
H_5F_5	-629.20 ± 3.55	-629.27 ± 3.55
H_5V_4	-609.90 ± 3.68	-609.56 ± 3.69
H_5L_4	-594.76 ± 12.56	-599.24 ± 12.87
H_5L_5	-584.03 ± 18.61	-576.68 ± 19.70
H_5C_4	-506.85 ± 4.90	-506.8101 ± 4.91
H_5I_4	-506.14 ± 114.87	N/A
Y_5F_4	-503.58 ± 4.60	-503.64 ± 4.60
H_5C_5	-503.07 ± 56.46	N/A

Table 4.2: Table of top 10 candidates among all 168 ELPs. The predicted ΔG is the prediction of ΔG and its standard deviation estimated by the terminal GPR model. The computed ΔG is the actual ΔG value (and its uncertainty) obtained from MD simulation if the corresponding candidate has been explored during Bayesian optimization. The abbreviation of ELPs follows the same convention as in table 4.1.

experimentalist collaborators (Dr. Bineet Sharma and Prof. Allen Liu) at the University of Michigan to verify the newly discovered sequences.

4.5 Discussion and Conclusion

We have developed a high-throughput screening procedure for discovering optimal diblock amphiphilic elastin-like polypeptide that could form stable vesicles. Our approach uses molecular simulation to obtain a quantitative measurement of the stability of the self-assembled vesicle, and then employs Bayesian optimization to discover the best diblock ELP that maximizes this stability. The optimization procedure converges after 10 iterations and we have explored 30 peptides (roughly 18% of the entire search space). The explored ELPs are ranked according to their computed vesicle stabilities. In addition, all peptides in the search space are ranked by the predicted vesicle stabilities by the final GPR model. We have discovered some previously unknown ELPs that could form very stable vesicles and predicted vesicle stability in agreement with the results for experimentally tested ELP. We will be working with our experimentalist collaborators in Prof. Allen Liu's group at the University of Michigan to test the top candidates suggested by our high-throughput screening procedure.

For future work, we would expand the search space to longer chains. In current experiments, the diblock ELPs usually contain dozens of hydrophobic and hydrophilic blocks that enable relatively thick vesicles to be formed[21,22,159,187]. These larger vesicles structurally resemble the compartments of biological cells[152] and enable interesting bioactivities, such as compartmentalized peptide synthesis[22], to happen inside. Besides diblock ELPs, triblock ELPs with hydrophilic blocks on two ends and hydrophobic blocks in the middle have also been experimentally used to form vesicles[190]. Therefore, in future work it would be interesting to expand the search space to include larger diblock and triblock ELPs with more than ten hydrophilic and hydrophobic blocks to match experimental interests. In principle, this is feasible for the current GPR model because the generic string kernel[161] could operate on amino acid strings of arbitrary lengths, but the evaluation of the kernel on long amino acid strings could be very expensive. Some alternative strategies are discussed in section 5.2 in chapter 5. Another potential challenge is that the umbrella sampling simulations might become inefficient due to increasing number of degree of freedom and the potential existence of high free energy barriers in degree of freedom orthogonal to the chosen reaction coordinates[179]. Thus, it would interesting to explore alternative enhanced sampling techniques, such as adaptive biasing force[191] or alchemical free energy calculations[192], to try to overcome the shortcomings of umbrella sampling when considering larger ELP systems.

CHAPTER 5
SUMMARY AND FUTURE WORK

5.1 Summary of Work

In summary, we have demonstrated the use of machine learning and black box optimization algorithms in the development of novel inverse design and high-throughput screening protocols for soft materials design. Throughout our work, we have formulated three material inverse design and high-throughput screening tasks as optimization problems and implemented suitable black box optimization algorithms for each task.

In chapter 2, we considered the inverse design of patchy colloids capable of self-assembling into open crystalline lattices. Colloidal crystals possess periodic relative permittivity whose periodicity roughly matches the wavelength of light, and thus they possess complete photonic bandgaps and allow researchers to control the flow of light inside these crystals[2,5]. An active research topic in colloidal crystal is the inverse design problem, which considers how to optimally tune the properties of colloids such that they could self-assemble into a target lattice[193]. In this work, we have developed a novel inverse design protocol to optimize the properties of patchy colloids for desired self-assembly behavior. We performed Monte Carlo simulations of the self-assembly process of patchy colloids and utilized a popular stochastic black-box optimization algorithm, CMA-ES[130], to maximize the free energy gap between target structure and competing structures based on the free energy surface calculated from simulation trajectories. CMA-ES algorithm is a derivative-free optimization algorithm which does not require the gradient information of the target function and has been widely used in various soft materials inverse design problems[54,101]. The patchy colloids with the optimal design parameters demonstrated the ability to self-assemble into high-quality open crystalline lattices in molecular dynamics simulation.

In chapter 3, we considered the inverse design of pre-assembled tetrahedral colloidal clusters capable of self-assembling into cubic diamond lattice. This work was a follow-up to the work in chapter 2 where we considered a simplified version of the model for patchy colloid that is more amenable for experimental realization. CMA-ES is applied to maximize the fraction of target structure in equilibrium, which is an easier objective function to evaluate than the free energy surface considered in chapter 2. Although we make these simplifications, the CMA-ES optimization still could converge to a set of design parameters that direct the tetrahedral colloidal clusters to self-assemble into a mixture of cubic and hexagonal diamond lattices in which the cubic diamond lattice structure is dominant. By further introducing a small initial seed, which is a common practice in colloidal self-assembly[67,194], we could significantly enhance the fraction of cubic diamond lattice formed by the self-assembly of colloidal clusters with optimal design parameters.

In chapter 4, we have developed a high-throughput screening protocol to search for diblock amphiphilic elastin-like polypeptides (ELPs) that could form stable vesicles. Elastin-like polypeptides are block co-polymers whose motifs are derived from tropoelastin. When carefully designed, they could self-assemble into vesicles that are more robust to osmotic pressure and membrane stretching than lipid vesicles[21], thus serving as good candidates for forming the compartments of synthetic cells. Our approach of high-throughput screening is to use enhanced sampling simulation to measure the stability of self-assembled ELP vesicles, and then use black box optimization to search for the sequence in search space that maximizes this stability. Quantitatively, the stability of the vesicle is measured by the free energy ΔG of dissociation of the vesicle. We implemented a string kernel to perform Gaussian Process Regression and Bayesian optimization to find the ELP in the search space that maximizes this free energy gap. After around 10 iterations, our Bayesian optimization has converged and we have identified some good diblock ELPs that have not been experimentally explored before. We would then work with our experimentalist collaborators to verify that these ELPs could self-assemble into stable vesicles.

To conclude, recent developments in machine learning and black-box optimization have opened new opportunities for the inverse design and high-throughput screening of materials. We have explored some of the recently developed machine learning and optimization methods and demonstrated their success in several challenging problems.

5.2 Future Directions

We foresee several possible future directions to continue our work.

First, for the inverse design of patchy colloids, currently we assume that the patch properties could be controlled precisely and there is no imprecision or fluctuation in the patch properties in individual patchy colloids. In experiments, usually there would be some polydispersity in patch properties due to the unavoidable imprecision in the manufacturing process[66,84,89,90,136]. For example, Chen et al.[89] have manufactured triblock patchy colloids with an uncertainty of less than 5° in patch size. Morphew et al.[66] have simulated the self-assembly of triblock patchy colloids whose patch sizes are drawn from pre-defined Gaussian distributions. In order to take polydispersity in patch properties into account, one direct modification to our method could be to draw patch properties from pre-defined Gaussian distributions whose standard deviation is determined by experimental precision. Then, we could apply the inverse design protocols developed in chapter 2 and 3 to perform optimization over the mean values of the patch properties. At every iteration of the optimization, for each patchy colloid in the system we would construct its surface patches by sampling from the Gaussian distributions with pre-defined standard deviations and the mean values at that iteration and proceed to the molecular simulation. Another way of incorporating polydispersity into account is to use a black-box optimization algorithm that directly takes the uncertainty of input variables into account. For example, Oliveira et al.[151] have proposed a modified Bayesian optimization framework that takes uncertainty in the measurements of input variables into account and minimize the expected regret. It is thus possible to replace the CMA-ES algorithm used in chapter 2 and 3 by such approaches[151]. Another possible future direction of investigation would be to change the way of measuring the goodness of self-assembly. Currently we break the self-assembly process into hierarchical stages and optimize the self-assembly at each stage separately. This requires us to have a priori knowledge of the kind of desired intermediate structures at each stage in order to form the final target structure (e.g. tetrahedral network of tetrahedron for pyrochlore lattice or dimers of tetrahedron for cubic diamond lattice). However, there might be situations in which this a priori knowledge is not available or we would like the target structure to be formed in a single stage. This would require a more direct way to estimate the quality of self-assembled structure. A possible way of doing so is to characterize the local environments of each colloid in the structure and measure the quality by the fraction of colloids whose local environment matches that of the target structure. This fraction would then be the target function to be maximized using black-box optimization. There exist multiple ways of characterizing the local environments of a particle. For example, one could use the Steinhardt bond order parameter defined in equation 2.16 in section 2.4 of chapter 2 and compare it with the value obtained from the desired crystal[68]. However, this would still require a priori knowledge of the degree of spherical harmonics (l) that is suitable for this identification. It is well known that $l = 3$ could distinguish cubic vs hexagonal diamond[67,68] and $l = 4$ distinguishes between pyrochlore and hexagonal tetrastack[65,90]. However, there is no general rule on how to choose l to distinguish one specific lattice from its competitor. The network graph analysis method proposed by Reinhart et al.[195] might provide an alternative and automatic way of characterizing the local environment of particles without a priori knowledge.

For the high-throughput screening of elastin-like polypeptides, we could expand the search space to larger diblock and triblock ELP chains. Current experiments on multiblock ELP self-assembly usually use much longer chains with dozens of hydrophilic and hydrophobic blocks[187,190]. The large vesicles formed by long ELP chains may support RNA transcription or protein expression inside[22], making their functionalities more closely mimic those of natural cells. Due to the increase in the degree of freedom, it might be necessary to explore alternative ways of performing the enhanced sampling[174] to evaluate the stability of self-assembled vesicles of larger chains. Adaptive biasing force[191] or alchemical free energy calculations[192] could potentially be used. Also, as the chain length increases, the evaluation of the string kernel defined in equation 4.10 of chapter 4 becomes more expensive[161]. It might be helpful to first train a neural network (e.g. autoencoder) to embed all peptides onto an Euclidean space and perform Bayesian optimization over this embedding. The GPR model in this case would consist of kernel functions defined in Euclidean space, such as radial basis kernel or Matérn kernel[181], that are less computationally expensive to evaluate. Such approaches have been used by Shmilovich et al.[30] to perform Bayesian optimization over three-dimensional embedding of peptides provided by a variational autoencoder. Yang et al.[41] have also employed doc2vec model originated in natural language processing to find embeddings of proteins with varying lengths.

REFERENCES

[1] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 1996.

[2] Steven G. Johnson, Attila Mekis, Shanhui Fan, and John D. Joannopoulos. Molding the flow of light. *Computing in Science and Engineering*, 3(6):38–47, 2001.

[3] M. Campbell, D. N. Sharp, M. T. Harrison, R. G. Denning, and A. J. Turberfield. Fabrication of photonic crystals for the visible spectrum by holographic lithography. *Nature*, 404(6773):53–56, 3 2000.

[4] Markus Deubel, Martin Wegener, Artan Kaso, and Sajeev John. Direct laser writing and characterization of "Slanted Pore" photonic crystals. *Applied Physics Letters*, 85(11):1895–1897, 9 2004.

[5] Zhongyu Cai, Zhiwei Li, Serge Ravaine, Mingxin He, Yanlin Song, Yadong Yin, Hanbin Zheng, Jinghua Teng, and Ao Zhang. From colloidal particles to photonic crystals: advances in self-assembly and their emerging applications. *Chemical Society Reviews*, 50(10):5898–5951, 5 2021.

[6] Woo Kyung Cho, Sangjin Park, Sangyong Jon, and Insung S. Choi. Water-repellent coating: formation of polymeric self-assembled monolayers on nanostructured surfaces. *Nanotechnology*, 18(39):395602, 10 2007.

[7] Jennifer E. Gagner, Wookhyun Kim, and Elliot L. Chaikof. Designing protein-based biomaterials for medical applications. *Acta Biomaterialia*, 10(4):1542–1557, 4 2014.

[8] Étienne Ducrot, Mingxin He, Gi Ra Yi, and David J. Pine. Colloidal alloys with preassembled clusters and spheres. *Nature Materials*, 16(6):652–657, 2017.

[9] Yu Wang, Yufeng Wang, Dana R. Breed, Vinothan N. Manoharan, Lang Feng, Andrew D. Hollingsworth, Marcus Weck, and David J. Pine. Colloids with valence and specific directional bonding. *Nature*, 491(7422):51–55, 2012.

[10] Wenyan Liu, Miho Tagawa, Huolin L. Xin, Tong Wang, Hamed Emamy, Huilin Li, Kevin G. Yager, Francis W. Starr, Alexei V. Tkachenko, and Oleg Gang. Diamond family of nanoparticle superlattices. *Science*, 351(6273):582–586, 2 2016.

[11] W. Benjamin Rogers and John C. Crocker. Direct measurements of DNA-mediated colloidal interactions and their quantitative modeling. *Proceedings of the National Academy of Sciences*, 108(38):15687–15692, 9 2011.

[12] Qian Chen, Sung Chul Bae, and Steve Granick. Directed self-assembly of a colloidal kagome lattice. *Nature*, 469(7330):381–384, 2011.

[13] Amar B. Pawar and Ilona Kretzschmar. Patchy particles by glancing angle deposition. *Langmuir*, 24(2):355–358, 2008.

[14] Amar B. Pawar and Ilona Kretzschmar. Multifunctional patchy particles by glancing angle deposition. *Langmuir*, 25(16):9057–9063, 2009.

[15] Changdeuck Bae, Jooho Moon, Hyunjung Shin, Jiyoung Kim, and Myung M. Sung. Fabrication of Monodisperse Asymmetric Colloidal Clusters by Using Contact Area Lithography (CAL). *Journal of the American Chemical Society*, 129(46):14232–14239, 11 2007.

[16] Ian W. Hamley, Jessica Hutchinson, Steven Kirkham, Valeria Castelletto, Amanpreet Kaur, Mehedi Reza, and Janne Ruokolainen. Nanosheet Formation by an Anionic Surfactant-like Peptide and Modulation of Self-Assembly through Ionic Complexation. *Langmuir*, 32(40):10387–10393, 10 2016.

[17] Kazunori Matsuura, Kazuya Murasato, and Nobuo Kimizuka. Artificial peptide-nanospheres self-assembled from three-way junctions of β-sheet-forming peptides. *Journal of the American Chemical Society*, 127(29):10148–10149, 2005.

[18] John B. Matson, R. Helen Zha, and Samuel I. Stupp. Peptide self-assembly for crafting functional biological materials. *Current Opinion in Solid State and Materials Science*, 15(6):225–235, 12 2011.

[19] Cuixia Chen, Fang Pan, Shengzhong Zhang, Jing Hu, Meiwen Cao, Jing Wang, Hai Xu, Xiubo Zhao, and Jian R. Lu. Antibacterial activities of short designer peptides: A link between propensity for nanostructuring and capacity for membrane destabilization. *Biomacromolecules*, 11(2):402–411, 2 2010.

[20] Julia Y. Rho, Henry Cox, Edward D.H. Mansfield, Sean H. Ellacott, Raoul Peltier, Johannes C. Brendel, Matthias Hartlieb, Thomas A. Waigh, and Sébastien Perrier. Dual self-assembly of supramolecular peptide nanotubes to provide stabilisation in water. *Nature Communications*, 10(1):4708, 2019.

[21] Thomas Frank, Kilian Vogele, Aurore Dupin, Friedrich C. Simmel, and Tobias Pirzer. Growth of Giant Peptide Vesicles Driven by Compartmentalized Transcription–Translation Activity. *Chemistry - A European Journal*, 26(72):17356–17360, 12 2020.

[22] Kilian Vogele, Thomas Frank, Lukas Gasser, Marisa A. Goetzfried, Mathias W. Hackl, Stephan A. Sieber, Friedrich C. Simmel, and Tobias Pirzer. Towards synthetic cells using peptide-based reaction compartments. *Nature Communications*, 9(1):1–7, 9 2018.

[23] Duc H.T. Le and Ayae Sugawara-Narutaki. Elastin-like polypeptides as building motifs toward designing functional nanobiomaterials. *Molecular Systems Design and Engineering*, 4(3):545–565, 6 2019.

[24] Harini Pattabhiraman, Guido Avvisati, and Marjolein Dijkstra. Novel Pyrochlore-like Crystal with a Photonic Band Gap Self-Assembled Using Colloids with a Simple Interaction Potential. *Physical Review Letters*, 119(15):157401, 10 2017.

[25] Beth A. Lindquist, Ryan B. Jadrich, and Thomas M. Truskett. Communication: Inverse design for self-assembly via on-the-fly optimization. *Journal of Chemical Physics*, 145(11):111101, 2016.

[26] Salvatore Torquato. Inverse optimization techniques for targeted self-assembly. *Soft Matter*, 5(6):1157–1173, 2009.

[27] É Marcotte, F. H. Stillinger, and Salvatore Torquato. Communication: Designed diamond ground state via optimized isotropic monotonic pair potentials. *Journal of Chemical Physics*, 138(6):61101, 2013.

[28] Henry Cohn and Abhinav Kumar. Algorithmic design of self-assembling structures. *Proceedings of the National Academy of Sciences*, 106(24):9570–9575, 6 2009.

[29] Rahul Upadhya, Shashank Kosuri, Matthew Tamasi, Travis A. Meyer, Supriya Atta, Michael A. Webb, and Adam J. Gormley. Automation and data-driven design of polymer therapeutics. *Advanced Drug Delivery Reviews*, 171:1–28, 2021.

[30] Kirill Shmilovich, Rachael A. Mansbach, Hythem Sidky, Olivia E. Dunne, Sayak Subhra Panda, John D. Tovar, and Andrew L. Ferguson. Discovery of Self-Assembling π-Conjugated Peptides by Active Learning-Directed Coarse-Grained Molecular Simulation. *Journal of Physical Chemistry B*, 124(19):3873–3891, 2020.

[31] Daan Frenkel and Berend Smit. Chapter 4 - Molecular Dynamics Simulations. In Daan Frenkel and Berend Smit, editors, *Understanding Molecular Simulation (Second Edition)*, pages 63–107. Academic Press, San Diego, second edi edition, 2002.

[32] Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, 7 1967.

[33] Shuichi Nosé. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics*, 81(1):511–519, 7 1984.

[34] H. J.C. Berendsen, J. P.M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 8 1984.

[35] W. K. Hastings. Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[36] Robert H. Swendsen and Jian Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 57(21):2607–2609, 11 1986.

[37] Ulli Wolff. Collective Monte Carlo Updating for Spin Systems. *Physical Review Letters*, 62(4):361–364, 1 1989.

[38] Stephen Whitelam and Phillip L. Geissler. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. *Journal of Chemical Physics*, 127(15):154101, 2007.

[39] Benjamin Sanchez-Lengeling, Loïc M. Roch, José Darío Perea, Stefan Langner, Christoph J. Brabec, and Alán Aspuru-Guzik. A Bayesian Approach to Predict Solubility Parameters. *Advanced Theory and Simulations*, 2(1):1800069, 2019.

[40] Christina S. Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.

[41] Kevin K. Yang, Zachary Wu, Claire N. Bedbrook, and Frances H. Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 8 2018.

[42] Weike Ye, Chi Chen, Zhenbin Wang, Iek Heng Chu, and Shyue Ping Ong. Deep neural networks for accurate predictions of crystal stability. *Nature Communications*, 9(1):3800, 2018.

[43] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W.R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[44] Charles C. David and Donald J. Jacobs. Principal component analysis: A method for determining the essential dynamics of proteins. In Dennis R Livesay, editor, *Methods in Molecular Biology*, volume 1084, pages 193–226. Humana Press, Totowa, NJ, 2014.

[45] Pavel Buslaev, Valentin Gordeliy, Sergei Grudinin, and Ivan Gushchin. Principal Component Analysis of Lipid Molecule Conformational Changes in Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation*, 12(3):1019–1028, 2016.

[46] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[47] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[48] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 5 2005.

[49] Andrew L. Ferguson, Athanassios Z. Panagiotopoulos, Ioannis G. Kevrekidis, and Pablo G. Debenedetti. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chemical Physics Letters*, 509(1-3):1–11, 2011.

[50] Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *Journal of Chemical Physics*, 134(12):124116, 2011.

[51] Jianyin Shao, Stephen W. Tanner, Nephi Thompson, and Thomas E. Cheatham. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, 11 2007.

[52] Kyle A. Beauchamp, Robert McGibbon, Yu-Shan Lin, and Vijay S. Pande. Simple few-state models reveal hidden complexity in protein folding. *Proceedings of the National Academy of Sciences*, 109(44):17807–17813, 10 2012.

[53] Xin-She Yang. Preface. In Xin-She Yang, editor, *Nature-Inspired Optimization Algorithms (Second Edition)*, pages xv–xvi. Academic Press, second edi edition, 2021.

[54] Jian Qin, Gurdaman S. Khaira, Yongrui Su, Grant P. Garner, Marc Miskin, Heinrich M. Jaeger, and Juan J. De Pablo. Evolutionary pattern design for copolymer directed self-assembly. *Soft Matter*, 9(48):11467–11472, 2013.

[55] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 1 2016.

[56] Aryan Deshwal, Cory M. Simon, and Janardhan Rao Doppa. Bayesian optimization of nanoporous materials. *Molecular Systems Design and Engineering*, 6(12):1066–1086, 2021.

[57] Ke Wang and Alexander W. Dowling. Bayesian optimization for chemical products and functional materials. *Current Opinion in Chemical Engineering*, 36:100728, 6 2022.

[58] Alex W. Wilber, Jonathan P.K. Doye, Ard A. Louis, Eva G. Noya, Mark A. Miller, and Pauline Wong. Reversible self-assembly of patchy particles into monodisperse icosahedral clusters. *Journal of Chemical Physics*, 127(8):085106, 8 2007.

[59] Andrew W. Long and Andrew L. Ferguson. Nonlinear machine learning of patchy colloid self-assembly pathways and mechanisms. *Journal of Physical Chemistry B*, 118(15):4228–4244, 4 2014.

[60] Zhenli Zhang and Sharon C. Glotzer. Self-assembly of patchy particles. *Nano Letters*, 4(8):1407–1413, 8 2004.

[61] Eric Lauga and Michael P. Brenner. Evaporation-driven assembly of colloidal particles. *Physical Review Letters*, 93(23):238301, 12 2004.

[62] James T. McGinley, Yifan Wang, Ian C. Jenkins, Talid Sinno, and John C. Crocker. Crystallated Colloidal Clusters Exhibit Directional DNA Interactions. *ACS Nano*, 9(11):10817–10825, 11 2015.

[63] Alexander J. Williamson, Alex W. Wilber, Jonathan P.K. Doye, and Ard A. Louis. Templated self-assembly of patchy particles. *Soft Matter*, 7(7):3423–3431, 3 2011.

[64] Ruohai Guo, Jian Mao, Xu Ming Xie, and Li Tang Yan. Predictive supracolloidal helices from patchy particles. *Scientific Reports*, 4(1):7021, 2014.

[65] Nathan A. Mahynski, Lorenzo Rovigatti, Christos N. Likos, and Athanassios Z. Panagiotopoulos. Bottom-Up Colloidal Crystal Assembly with a Twist. *ACS Nano*, 10(5):5459–5467, 2016.

[66] Daniel Morphew, James Shaw, Christopher Avins, and Dwaipayan Chakrabarti. Programming Hierarchical Self-Assembly of Patchy Particles into Colloidal Crystals via Colloidal Molecules. *ACS Nano*, 12(3):2355–2364, 3 2018.

[67] Zhenli Zhang, Aaron S. Keys, Ting Chen, and Sharon C. Glotzer. Self-assembly of patchy particles into diamond structures through molecular mimicry. *Langmuir*, 21(25):11547–11551, 12 2005.

[68] Flavio Romano, Eduardo Sanz, and Francesco Sciortino. Crystallization of tetrahedral patchy particles in silico. *Journal of Chemical Physics*, 134(17):174502, 5 2011.

[69] Antti Pekka Hynninen, Job H.J. Thijssen, Esther C.M. Vermolen, Marjolein Dijkstra, and Alfons Van Blaaderen. Self-assembly route for photonic crystals with a bandgap in the visible region. *Nature Materials*, 6(3):202–205, 2 2007.

[70] Alexander M. Kalsin, Marcin Fialkowski, Maciej Paszewski, Stoyan K. Smoukov, Kyle J.M. Bishop, and Bartosz A. Grzybowski. Electrostatic self-assembly of binary nanoparticle crystals with a diamond-like lattice. *Science*, 312(5772):420–424, 2006.

[71] Eva G. Noya, Carlos Vega, Jonathan P.K. Doye, and Ard A. Louis. The stability of a crystal with diamond structure for patchy particles with tetrahedral symmetry. *Journal of Chemical Physics*, 132(23):234511, 6 2010.

[72] Angel Garcia-Adeva. Band gap atlas for photonic crystals having the symmetry of the kagomé and pyrochlore lattices. *New Journal of Physics*, 8(6):86–86, 6 2006.

[73] Martin Maldovan and Edwin L. Thomas. Diamond-structured photonic crystals. *Nature Materials*, 3(9):593–600, 2004.

[74] T. T. Ngo, C. M. Liddell, M. Ghebrebrhan, and J. D. Joannopoulos. Tetrastack: Colloidal diamond-inspired structure with omnidirectional photonic band gap for low refractive index contrast. *Applied Physics Letters*, 88(24):241920, 6 2006.

[75] Luis González-Urbina, Kasper Baert, Branko Kolaric, Javier Pérez-Moreno, and Koen Clays. Linear and nonlinear optical properties of colloidal photonic crystals. *Chemical Reviews*, 112(4):2268–2285, 4 2012.

[76] Alexander Moroz. Metallo-dielectric diamond and zinc-blende photonic crystals. *Physical Review B*, 66(11):115109, 9 2002.

[77] Martin Maldovan, Chaitanya K. Ullal, W. Craig Carter, and Edwin L. Thomas. Exploring for 3D photonic bandgap structures in the 11 f.c.c. space groups. *Nature Materials*, 2(10):664–667, 2003.

[78] Geoffrey I.N. Waterhouse and Mark R. Waterland. Opal and inverse opal photonic crystals: Fabrication and characterization. *Polyhedron*, 26(2):356–368, 2007.

[79] Alvaro Blanco, Emmanuel Chomski, Serguel Grabtchak, Marta Ibisate, Sajeev John, Stephen W. Leonard, Cefe Lopez, Francisco Meseguer, Hernan Miguez, Jessica P. Mondla, Geoffrey A. Ozin, Ovidiu Toader, and Henry M. Van Driel. Large-scale synthesis of a silicon photonic crystal with a complete three-dimensional bandgap near 1.5 micrometres. *Nature*, 405(6785):437–440, 2000.

[80] M. Loncar, Theodor Doll, J. Vuckovic, and Axel Scherer. Design and fabrication of silicon photonic crystal optical waveguides. *Journal of Lightwave Technology*, 18(10):1402–1411, 10 2000.

[81] Renju Rajan, Padmanabhan Ramesh Babu, and Krishnamoorthy Senthilnathan. The Dawn of Photonic Crystals: An Avenue for Optical Computing. In Alexander Vakhrushev, editor, *Theoretical Foundations and Application of Photonic Crystals*, chapter 6, pages 120–132. InTech, 4 2018.

[82] Branko Kolaric, Sylvain Desprez, Fabian Brau, and Pascal Damman. Design of curved photonic crystal using swelling induced instabilities. *Journal of Materials Chemistry*, 22(32):16205–16208, 2012.

[83] Yifan Wang, Ian C. Jenkins, James T. McGinley, Talid Sinno, and John C. Crocker. Colloidal crystals with diamond symmetry at optical lengthscales. *Nature Communications*, 8:14173, 2017.

[84] Mingxin He, Johnathon P. Gales, Étienne Ducrot, Zhe Gong, Gi Ra Yi, Stefano Sacanna, and David J. Pine. Colloidal diamond. *Nature*, 585(7826):524–529, 9 2020.

[85] Kyle J.M. Bishop, Nicolas R. Chevalier, and Bartosz A. Grzybowski. When and why like-sized, oppositely charged particles assemble into diamond-like crystals. *Journal of Physical Chemistry Letters*, 4(9):1507–1511, 2013.

[86] Likui Wang, Linhua Xia, Gang Li, Serge Ravaine, and X. S. Zhao. Patterning the surface of colloidal microspheres and fabrication of nonspherical particles. *Angewandte Chemie - International Edition*, 47(25):4725–4728, 2008.

[87] Amar B. Pawar and Ilona Kretzschmar. Fabrication, assembly, and application of patchy particles. *Macromolecular Rapid Communications*, 31(2):150–168, 2010.

[88] Gi Ra Yi, David J. Pine, and Stefano Sacanna. Recent progress on patchy colloids and their self-assembly. *Journal of Physics Condensed Matter*, 25(19):193101, 2013.

[89] Qian Chen, Sung Chul Bae, and Steve Granick. Staged Self-Assembly of Colloidal Metastructures. *Journal of the American Chemical Society*, 134(27):11080–11083, 7 2012.

[90] Flavio Romano and Francesco Sciortino. Patterning symmetry in the rational design of colloidal crystals. *Nature Communications*, 3:975, 2012.

[91] D. Zeb Rocklin and Xiaoming Mao. Self-assembly of three-dimensional open structures using patchy colloidal particles. *Soft Matter*, 10(38):7569–7576, 2014.

[92] D. Chen, G. Zhang, and S. Torquato. Inverse Design of Colloidal Crystals via Optimized Patchy Interactions. *Journal of Physical Chemistry B*, 122(35):8462–8468, 2018.

[93] Fernando A. Escobedo. Optimizing the formation of colloidal compounds with components of different shapes. *Journal of Chemical Physics*, 147(21):214501, 2017.

[94] Huikuan Chao and Robert A. Riggleman. Inverse design of grafted nanoparticles for targeted self-assembly. *Molecular Systems Design and Engineering*, 3(1):214–222, 2018.

[95] Mikael C. Rechtsman, Frank H. Stillinger, and Salvatore Torquato. Optimized interactions for targeted self-assembly: Application to a honeycomb lattice. *Physical Review Letters*, 95(22):228301, 11 2005.

[96] Avni Jain, Jeffrey R. Errington, and Thomas M. Truskett. Dimensionality and design of isotropic interactions that stabilize honeycomb, square, simple cubic, and diamond lattices. *Physical Review X*, 4(3):31049, 9 2014.

[97] Avni Jain, Jeffrey R. Errington, and Thomas M. Truskett. Inverse design of simple pairwise interactions with low-coordinated 3D lattice ground states. *Soft Matter*, 9(14):3866–3870, 2013.

[98] Avni Jain, Jonathan A. Bollinger, and Thomas M. Truskett. Inverse methods for material design. *AIChE Journal*, 60(8):2732–2740, 2014.

[99] Alexander P. Lyubartsev and Aatto Laaksonen. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Physical Review E*, 52(4):3730–3737, 10 1995.

[100] Muhittin Mungan, Chorng-Haur Sow, Susan N. Coppersmith, and David G. Grier. Determining pair interactions from structural correlations. *Physical Review B*, 58(21):14588–14593, 12 1998.

[101] Andrew W. Long and Andrew L. Ferguson. Rational design of patchy colloids via landscape engineering. *Molecular Systems Design & Engineering*, 3(1):49–65, 2018.

[102] Yufeng Wang, Yufeng Wang, Xiaolong Zheng, Étienne Ducrot, Jeremy S. Yodh, Marcus Weck, and David J. Pine. Crystallization of DNA-coated colloids. *Nature Communications*, 6(1):7253, 11 2015.

[103] Gang Zhang, Dayang Wang, and Helmuth Möhwald. Decoration of microspheres with gold nanodots - Giving colloidal spheres valences. *Angewandte Chemie - International Edition*, 44(47):7767–7770, 2005.

[104] Norbert Kern and Daan Frenkel. Fluid-fluid coexistence in colloidal systems with short-ranged strongly directional attraction. *Journal of Chemical Physics*, 118(21):9882–9889, 2003.

[105] Qian Chen, Erich Diesel, Jonathan K. Whitmer, Sung Chul Bae, Erik Luijten, and Steve Granick. Triblock Colloids for Directed Self-Assembly. *Journal of the American Chemical Society*, 133(20):7725–7727, 5 2011.

[106] John D. Weeks, David Chandler, and Hans C. Andersen. Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids. *The Journal of Chemical Physics*, 54(12):5237–5247, 6 1971.

[107] Matthew N. O'Brien, Matthew R. Jones, and Chad A. Mirkin. The nature and implications of uniformity in the hierarchical organization of nanomaterials. *Proceedings of the National Academy of Sciences*, 113(42):11717–11725, 10 2016.

[108] Insung S Choi, Ned Bowden, and George M Whitesides. Macroscopic, Hierarchical, Two-Dimensional Self-Assembly. *Angew. Chem. Int. Ed*, 38(20), 1999.

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

[110] Johannes Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(6):932–942, 11 2011.

[111] Nikolaus Hansen, Sibylle D. Müller, and Petros Koumoutsakos. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, 11(1):1–18, 3 2003.

[112] Joshua A. Anderson, Chris D. Lorenz, and A. Travesset. General purpose molecular dynamics simulations fully implemented on graphics processing units. *Journal of Computational Physics*, 227(10):5342–5359, 2008.

[113] Jens Glaser, Trung Dac Nguyen, Joshua A. Anderson, Pak Lui, Filippo Spiga, Jaime A. Millan, David C. Morse, and Sharon C. Glotzer. Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Computer Physics Communications*, 192:97–107, 2015.

[114] Boaz Nadler, Stéphane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[115] Lilia V. Nedialkova, Miguel A. Amat, Ioannis G. Kevrekidis, and Gerhard Hummer. Diffusion maps, clustering and fuzzy Markov modeling in peptide folding transitions. *The Journal of chemical physics*, 141(11):114102, 9 2014.

[116] Wenwei Zheng, Mary A. Rohrdanz, and Cecilia Clementi. Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. *Journal of Physical Chemistry B*, 117(42):12769–12776, 2013.

[117] Andrew L. Ferguson, Athanassios Z. Panagiotopoulos, Pablo G. Debenedetti, and Ioannis G. Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences*, 107(31):13597–13602, 8 2010.

[118] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 9 2008.

[119] Andrew L. Ferguson, Athanassios Z. Panagiotopoulos, Pablo G. Debenedetti, and Ioannis G. Kevrekidis. Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide. *Journal of Chemical Physics*, 134(13):135103, 2011.

[120] Dan Wei, Junjie Song, and Chuan Liu. Hybrid Monte Carlo Micromagnetics. *IEEE Transactions on Magnetics*, 54(11):216–222, 2018.

[121] Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, 2011.

[122] M. A. Gonzalez, E. Sanz, C. McBride, J. L.F. Abascal, C. Vega, and C. Valeriani. Nucleation free-energy barriers with Hybrid Monte-Carlo/Umbrella Sampling. *Physical Chemistry Chemical Physics*, 16(45):24913–24919, 2014.

[123] Behrooz Hashemian, Daniel Millán, and Marino Arroyo. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *Journal of Chemical Physics*, 139(21):214101, 12 2013.

[124] Wei Chen and Andrew L. Ferguson. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry*, 39(25):2079–2102, 2018.

[125] Wei Chen, Aik Rui Tan, and Andrew L. Ferguson. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *Journal of Chemical Physics*, 149(7):72312, 2018.

[126] Wei Chen, Hythem Sidky, and Andrew L. Ferguson. Nonlinear discovery of slow molecular modes using state-free reversible VAMPnets. *Journal of Chemical Physics*, 150(21):214114, 2019.

[127] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 2006.

[128] Andrew L. Ferguson. BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *Journal of Computational Chemistry*, 38(18):1583–1605, 7 2017.

[129] Naoko Nakagawa and Michel Peyrard. The inherent structure landscape of a protein. *Proceedings of the National Academy of Sciences*, 103(14):5279–5284, 4 2006.

[130] Nikolaus Hansen and Andreas Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 6 2001.

[131] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Theoretical Foundation for CMA-ES from Information Geometry Perspective. *Algorithmica*, 64(4):698–716, 12 2012.

[132] Alex W. Wilber, Jonathan P.K. Doye, Ard A. Louis, and Anna C.F. Lewis. Monodisperse self-assembly in a model with protein-like interactions. *Journal of Chemical Physics*, 131(17):175102, 11 2009.

[133] Paul J. Steinhardt, David R. Nelson, and Marco Ronchetti. Bond-orientational order in liquids and glasses. *Physical Review B*, 28(2):784–805, 7 1983.

[134] Wolfgang Lechner and Christoph Dellago. Accurate determination of crystal structures based on averaged local bond order parameters. *Journal of Chemical Physics*, 129(11):114707, 2008.

[135] Steven Johnson and John Joannopoulos. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Optics Express*, 8(3):173, 2001.

[136] Mingxin He, Johnathon P. Gales, Xinhang Shen, Min Jae Kim, and David J. Pine. Colloidal Particles with Triangular Patches. *Langmuir*, 37(23):7246–7253, 6 2021.

[137] Qun Li Lei, Ran Ni, and Yu Qiang Ma. Self-Assembled Chiral Photonic Crystals from a Colloidal Helix Racemate. *ACS Nano*, 12(7):6860–6870, 2018.

[138] Marlous Kamp, Bart De Nijs, Marjolein N. Van Der Linden, Isja De Feijter, Merel J. Lefferts, Antonio Aloi, Jack Griffiths, Jeremy J. Baumberg, Ilja K. Voets, and Alfons Van Blaaderen. Multivalent Patchy Colloids for Quantitative 3D Self-Assembly Studies. *Langmuir*, 36(9):2403–2418, 3 2020.

[139] Rebecca W. Perry and Vinothan N. Manoharan. Segregation of "isotope" particles within colloidal molecules. *Soft Matter*, 12(11):2868–2876, 2016.

[140] In Seong Jo, Joon Suk Oh, Shin Hyun Kim, David J. Pine, and Gi Ra Yi. Compressible colloidal clusters from Pickering emulsions and their DNA functionalization. *Chemical Communications*, 54(60):8328–8331, 2018.

[141] Etienne Duguet, Anthony Désert, Adeline Perro, and Serge Ravaine. Design and elaboration of colloidal molecules: An overview. *Chemical Society Reviews*, 40(2):941–960, 2011.

[142] Hui Chen, Weiyi Zhang, and Zhenlin Wang. Comparative studies on photonic band structures of diamond and hexagonal diamond using the multiple scattering method. *Journal of Physics Condensed Matter*, 16(6):741–748, 1 2004.

[143] Weiya Li, Hervé Palis, Rémi Mérindol, Jérôme Majimel, Serge Ravaine, and Etienne Duguet. Colloidal molecules and patchy particles: Complementary concepts, synthesis and self-assembly. *Chemical Society Reviews*, 49(6):1955–1976, 2020.

[144] Yu Wang, Yufeng Wang, Xiaolong Zheng, Gi Ra Yi, Stefano Sacanna, David J. Pine, and Marcus Weck. Three-dimensional lock and key colloids. *Journal of the American Chemical Society*, 136(19):6866–6869, 2014.

[145] Rachelle M. Choueiri, Elizabeth Galati, Héloise Thérien-Aubin, Anna Klinkova, Egor M. Larin, Ana Querejeta-Fernández, Lili Han, Huolin L. Xin, Oleg Gang, Ekaterina B. Zhulina, Michael Rubinstein, and Eugenia Kumacheva. Surface patterning of nanoparticles with polymer patches. *Nature*, 538(7623):79–83, 10 2016.

[146] Abhishek B. Rao, James Shaw, Andreas Neophytou, Daniel Morphew, Francesco Sciortino, Roy L. Johnston, and Dwaipayan Chakrabarti. Leveraging Hierarchical Self-Assembly Pathways for Realizing Colloidal Photonic Crystals. *ACS Nano*, 14(5):5348–5359, 5 2020.

[147] George M. Whitesides and Bartosz Grzybowski. Self-assembly at all scales. *Science*, 295(5564):2418–2421, 2002.

[148] Adeline Perro, Etienne Duguet, Olivier Lambert, Jean Christophe Taveau, Elodie Bourgeat-Lami, and Serge Ravaine. A chemical synthetic route towards "Colloidal molecules". *Angewandte Chemie - International Edition*, 48(2):361–365, 2009.

[149] Vinothan N. Manoharan, Mark T. Elsesser, and David J. Pine. Dense packing and symmetry in small clusters of microspheres. *Science*, 301(5632):483–487, 2003.

[150] Joel Henzie, Michael Grünwald, Asaph Widmer-Cooper, Phillip L. Geissler, and Peidong Yang. Self-assembly of uniform polyhedral silver nanocrystals into densest packings and exotic superlattices. *Nature Materials*, 11(2):131–137, 2012.

[151] Rafael Oliveira, Lionel Ott, and Fabio Ramos. Bayesian optimisation under uncertain inputs. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1177–1184. PMLR, 2020.

[152] Pasquale Stano. Is research on “synthetic cells” moving to the next level? *Life*, 9(1):3, 12 2019.

[153] Yuval Elani, Robert V. Law, and Oscar Ces. Vesicle-based artificial cells: Recent developments and prospects for drug delivery. *Therapeutic Delivery*, 6(5):143–145, 2015.

[154] Lisa Sercombe, Tejaswi Veerati, Fatemeh Moheimani, Sherry Y. Wu, Anil K. Sood, and Susan Hua. Advances and challenges of liposome assisted drug delivery. *Frontiers in Pharmacology*, 6(DEC):286, 2015.

[155] Joshua A. Jackman, Jae Hyeok Choi, Vladimir P. Zhdanov, and Nam Joon Cho. Influence of osmotic pressure on adhesion of lipid vesicles to solid supports. *Langmuir*, 29(36):11375–11384, 9 2013.

[156] Stefan Roberts, Michael Dzuricky, and Ashutosh Chilkoti. Elastin-like polypeptides as models of intrinsically disordered proteins. *FEBS Letters*, 589(19):2477–2486, 9 2015.

[157] M. Hamed Misbah, Luis Quintanilla, M. Alonso, and J. Carlos Rodríguez-Cabello. Evolution of amphiphilic elastin-like co-recombinamer morphologies from micelles to a lyotropic hydrogel. *Polymer*, 81:37–44, 12 2015.

[158] Diana Juanes-Gusano, Mercedes Santos, Virginia Reboto, Matilde Alonso, and José Carlos Rodríguez-Cabello. Self-assembling systems comprising intrinsically disordered protein polymers like elastin-like recombinamers. *Journal of Peptide Science*, 28(1):e3362, 1 2022.

[159] Andreas Schreiber, Lara G. Stühn, Matthias C. Huber, Süreyya E. Geissinger, Ashit Rao, and Stefan M. Schiller. Self-Assembly Toolbox of Tailored Supramolecular Architectures Based on an Amphiphilic Protein Library. *Small*, 15(30):1900163, 7 2019.

[160] Peng Zhou, Xiang Chen, Yuqian Wu, and Zhicai Shang. Gaussian process: An alternative approach for QSAM modeling of peptides. *Amino Acids*, 38(1):199–212, 1 2010.

[161] Sébastien Giguère, Mario Marchand, François Laviolette, Alexandre Drouin, and Jacques Corbeil. Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC Bioinformatics*, 14(1):82, 12 2013.

[162] Ernest Y. Lee, Gerard C.L. Wong, and Andrew L. Ferguson. Machine learning-enabled discovery and design of membrane-active peptides. *Bioorganic & Medicinal Chemistry*, 26(10):2708–2718, 6 2018.

[163] Morten Nielsen and Ole Lund. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics*, 10(1):296, 9 2009.

[164] Ernest Y. Lee, Benjamin M. Fulan, Gerard C. L. Wong, and Andrew L. Ferguson. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proceedings of the National Academy of Sciences*, 113(48):13588–13593, 11 2016.

[165] Bryce A. Thurston and Andrew L. Ferguson. Machine learning and molecular design of self-assembling -conjugated oligopeptides. *Molecular Simulation*, 44(11):930–945, 7 2018.

[166] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 5 1982.

[167] Schrödinger. LLC. The PyMOL Molecular Graphics System, Version 1.5.0.4, 2012.

[168] Djurre H. De Jong, Gurpreet Singh, W. F.Drew Bennett, Clement Arnarez, Tsjerk A. Wassenaar, Lars V. Schäfer, Xavier Periole, D. Peter Tieleman, and Siewert J. Marrink. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697, 1 2013.

[169] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *Journal of Chemical Physics*, 126(1):014101, 1 2007.

[170] M. Parrinello and A. Rahman. Crystal structure and pair potentials: A molecular-dynamics study. *Physical Review Letters*, 45(14):1196–1199, 10 1980.

[171] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindah. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 9 2015.

[172] Justin A. Lemkul and David R. Bevan. Assessing the stability of Alzheimer's amyloid protofibrils using molecular dynamics. *Journal of Physical Chemistry B*, 114(4):1652–1660, 2 2010.

[173] Emre Sevgen, Moshe Dolejsi, Paul F. Nealey, Jeffrey A. Hubbell, and Juan J. De Pablo. Nanocrystalline Oligo(ethylene sulfide)- b-poly(ethylene glycol) Micelles: Structure and Stability. *Macromolecules*, 51(23):9538–9546, 12 2018.

[174] Rafael C. Bernardi, Marcelo C.R. Melo, and Klaus Schulten. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1850(5):872–877, 5 2015.

[175] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, 9 2011.

[176] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *Journal of Chemical Physics*, 115(20):9169–9183, 11 2001.

[177] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics*, 129(12):124105, 9 2008.

[178] Benoît Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282, 9 1995.

[179] Fangqiang Zhu and Gerhard Hummer. Convergence and error estimation in free energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry*, 33(4):453–465, 2 2012.

[180] Alan Grossfield. The Weighted Histogram Analysis Method, Version 2.0.10, 2003.

[181] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, 8 2018.

[182] Robert Gentleman. Current Topics in Computational Molecular Biology. *Journal of the American Statistical Association*, 99(466):560–560, 2004.

[183] S Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 11 1992.

[184] Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimisation*, 2(2):117–129, 1978.

[185] A. Bhattacharyya. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics*, 7(4):401–406, 1946.

[186] Garrett Beatty, Ethan Kochis, and Michael Bloodgood. The Use of Unlabeled Data Versus Labeled Data for Stopping Active Learning for Text Classification. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 287–294. IEEE, 1 2019.

[187] Andreas Schreiber, Matthias C. Huber, and Stefan M. Schiller. Prebiotic Protocell Model Based on Dynamic Protein Membranes Accommodating Anabolic Reactions. *Langmuir*, 35(29):9593–9610, 6 2019.

[188] Michael C. Hout, Megan H. Papesh, and Stephen D. Goldinger. Multidimensional scaling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):93–103, 1 2013.

[189] S. JEFFERY. Evolution of Protein Molecules. *Biochemical Society Transactions*, 7(2):452–453, 4 1979.

[190] Laura Martín, Emilio Castro, Artur Ribeiro, Matilde Alonso, and J. Carlos Rodríguez-Cabello. Temperature-triggered self-assembly of elastin-like block co-recombinamers: The controlled formation of micelles and vesicles in an aqueous medium. *Biomacromolecules*, 13(2):293–298, 2 2012.

[191] Jeffrey Comer, James C Gumbart, Jérôme Hénin, Tony Lelièvre, Andrew Pohorille, and Christophe Chipot. The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *The Journal of Physical Chemistry B*, 119(3):1129–1151, 1 2015.

[192] Tai Sung Lee, Bryce K. Allen, Timothy J. Giese, Zhenyu Guo, Pengfei Li, Charles Lin, T. Dwight McGee, David A. Pearlman, Brian K. Radak, Yujun Tao, Hsu Chun Tsai, Huafeng Xu, Woody Sherman, and Darrin M. York. Alchemical binding free energy calculations in AMBER20: Advances and best practices for drug discovery. *Journal of Chemical Information and Modeling*, 60(11):5595–5623, 11 2020.

[193] Mikael C. Rechtsman, Frank H. Stillinger, and Salvatore Torquato. Self-assembly of the simple cubic lattice with an isotropic potential. *Physical Review E*, 74(2):021404, 8 2006.

[194] A. Cacciuto, S. Auer, and Daan Frenkei. Onset of heterogeneous crystal nucleation in colloidal suspensions. *Nature*, 428(6981):404–406, 3 2004.

[195] Wesley F. Reinhart, Andrew W. Long, Michael P. Howard, Andrew L. Ferguson, and Athanassios Z. Panagiotopoulos. Machine learning for autonomous crystal structure identification. *Soft Matter*, 13(27):4733–4745, 7 2017.