# Data-Driven Analysis of Experimental Design Spaces for Colloidal Synthesis and Assembly

Huat Thart Chiang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2025

*Reading Committee:*
Lilo D. Pozzo, Chair
Zachary Sherman
David Bergsman

Program Authorized to Offer Degree:
Chemical Engineering

University of Washington

## Abstract

Data-Driven Analysis of Experimental Design Spaces for Colloidal Synthesis and Assembly

Huat Thart Chiang

Chair of the Supervisory Committee:

Lilo D. Pozzo

Department of Chemical Engineering

Colloidal nanomaterials offer diverse functionalities driven by their structural properties, requiring precise control over synthesis and assembly. Due to the complexity of the experimental design space, a self-driving lab approach, which combines artificial intelligence (AI) with autonomous experimentation, is a powerful method used to navigate it. In this approach, an AI agent selects and evaluates experiments in a closed loop, progressively improving its understanding of the design space as data is collected. Some ways to improve self-driving labs include improving the distance metric to enhance the system's ability to guide synthesis. In this work, the amplitude-phase distance metric is introduced, which captures shape differences in functional datasets (e.g., UV-Vis Spectroscopy, SAXS). Its performance is compared to the Euclidean distance metric, and key differences are observed in nanoparticle structural differentiation (e.g., between nanospheres and nanorods) and the balance between the exploitation and exploration of the design space. Further advancements in self-driving labs include using multiple characterization methods (UV-Vis, SAXS, TEM), minimizing reliance on literature for design space definition, and employing interpretable AI to extract experimental insights. These improvements are demonstrated with another self-driving lab tested with silver nanoplate synthesis, where the derived design rules align with established knowledge. In addition to inorganic nanoparticle synthesis, this work also explores engineering stimuli responsive protein assemblies, which was done by modifying the RhuA protein with light and chemically responsive molecules. Struc-

tural analysis via a Monte Carlo-based SAXS fitting method reveals light-controlled assembly into tubes or sheets, influenced by solution ionic strength. This efficient modeling approach supports future exploration of metastable structures using in situ SAXS combined with AI-guided light sequencing. Finally, we explore DNA-mediated assembly of lipid encapsulated nanoparticles where high-throughput experimentation is used to identify the effect of design variables on the final assembly.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

First, I would like to express my gratitude to my advisor, Dr. Lilo Pozzo, for the support, guidance, patience and encouragement throughout the course of my PhD. Thank you for being so supportive during my PhD, for believing in me, and for teaching me how to think like a scientist. It is also truly a privilege to experience and learn from so many conferences, beamline trips, and scattering courses. I am very grateful to have been your student. I would also like to thank the members of my dissertation committee Dr. David Bergsman, Dr. Zachary Sherman, Dr. David Beck, and Dr. James De Yoreo for their valuable feedback, constructive criticism, and generous contributions of time and knowledge. I am also thankful for the center for the science of synthesis across scales, where I have enjoyed and learned so much from all the collaborations, meetings, and retreats.

The Pozzo Research group is a fantastic place to perform research, and I am grateful to all the past and present members group, especially Kacper, Kiran, Moeez, and Naomi, who have been a wonderful source of collaboration, encouragement, and inspiration during this journey. You have greatly supported me throughout my PhD, and I am thankful to have interacted with each one of you. My experience as a PhD student would not have been the same without your support, and I look forward to learn about all your achievements in the future. I also would like to acknowledge the members of the Sherman Research Group who have been a great source of support and inspiration. Next, I want to thank my parents, brothers and cousins. Your love, encouragement, and belief in me have been the foundation of all that I have achieved. Finally, I would like to thank my girlfriend Kaung Su for supporting me all the way through my PhD journey. Thank you for standing by me every step of the way.

# DEDICATION

To the reader, I hope you find what you are looking for.

# Chapter 1

# Introduction

## 1.1 Objectives

Controlling the structure or assembly of colloidal nanoparticles is important because the structure of these particles often gives rise to unique functionalities which are not present in the material's bulk form. However, this is often challenging due to the large and complex experimental design spaces of colloidal synthesis or assembly, where several variables can have complex effects on the final structure. One powerful method to navigate this large and complex design spaces is by using data-driven methods and high-throughput experimentation. In this thesis, data-driven methods will be used to explore the experimental design space of several different colloidal systems. Chapter 1 will focus on the theoretical background of colloidal synthesis and assembly, such as the commonly used methods to synthesize and assemble colloidal particles as well as applications of nanomaterials. Chapter 2 is about the experimental methods that were used for the work in this thesis, such as the machine learning methods and the characterization techniques. Chapter 3 is about the effect of a distance metric on closed-loop retrosynthesis which is using a data-driven algorithm together with a liquid handling robot and UV-Vis Spectroscopy to optimize the properties of gold nanoparticles. In the chapter, it is shown that the choice of the distance metric has a significant impact on the performance of the closed-loop retrosynthesis system. Because of this, the Amplitude-Phase distance metric is introduced, which primarily considers the shape of the spectroscopic curve when assigning a distance. At the end of the chapter, the limitations of a closed-loop retrosynthesis are discussed, which can be summarized as the

requirement of multi modal characterization, knowledge generation, and the reduction of reliance on the literature. These limitations are then addressed in Chapter 4 where a new data-driven method is introduced and tested on the synthesis of silver nanoplates. The method first uses UV-Vis spectroscopy to determine regions in the design space where plate-like particles are most likely to be synthesized. By randomly sampling in these regions of interest and by using small-angle x-ray scattering to obtain high quality information on the size and dispersity of the samples, information on how each reagent affects the structure of the nanoparticles was obtained. The findings obtained from the large amounts of collected data agreed with the mechanisms described in the literature of the well-studied system of silver nanoplate synthesis. In Chapter 5, we transition from nanoparticle synthesis to protein self-assembly using a model protein called RhuA. To form the assemblies, either a $\beta$-cyclodextrin (host) or azobenzene (guest) molecule is covalently attached to the protein, which forms chemically and light responsive protein assemblies when mixed together. To characterize the structure of the protein assembly, small angle x-ray scattering is used and scattering data is obtained from the assemblies. We also present a method to efficiently analyze scattering curves of large biomolecular assemblies and use it on the data collected on our system. The results show that the RhuA protein assembles into tube or sheet like structures which is consistent with data obtained from microscopy. Finally in Chapter 6, the DNA-mediated assembly of lipid encapsulated silica nanoparticles in investigated. The objective is to determine how DNA surface mobility on the nanoparticle surface affects the final structure of the assembly. In this chapter, high-throughput experimentation is used to rapidly screen structures and the effect of design variables such solution ionic strength and DNA concentration is identified.

## 1.2 Inorganic Nanoparticles

Inorganic nanoparticles are nanomaterials that are composed of metals, metal oxides, or nonmetallic materials. Because of their small size (1-100 nm), they contain special properties such as chemical, optical, physical, or biological, that are not present in the bulk material. For example, metallic nanoparticles like gold or silver possess a special property called surface plasmon resonance which allows for strong light absorption and scattering at specific wavelengths. This property is useful for applications in medicine, catalysis, and sensors. Silica nanoparticles have special properties such as surface tunability (e.g., making them hydrophobic or hydrophilic) which allows modifications to be easily made on their surface. In addition,

their optical transparency, and thermal and chemical stability make them especially useful for industrial applications such as protective coatings. Another useful property of inorganic nanoparticles is their high surface area-to-volume ratio. Mesoporous silica nanoparticles are spherical particles that contain an ordered arrangement of pores inside them which results in an extremely high surface area. This is useful for applications in drug delivery, where drugs can be delivered in the pores, or bioimaging where high concentrations of contrast agents can be delivered to targeted locations in the pores [2]. Another popular kind of inorganic nanoparticle is the quantum dot (QD). These semiconductor nanoparticles (e.g., CdSe, CdS) have optical and electronic properties due to quantum mechanical effects, which depend on the nanoparticle's structure. In short, when a QD is exposed to light, an electron in the valence band is excited into the conduction band. When it returns to the valence band, it will release its energy as light of a specific wavelength. The wavelength of the emitted light is tunable by the size of the QD, where larger QDs usually emit light with longer wavelengths and smaller QDs emit shorter wavelengths. One application of quantum dots is in quantum dot light-emitting diode displays. Because QDs emit light of tunable wavelengths, they can be used as light sources for the individual pixels in displays, which increases the power efficiency and highly pure colors in the display [3].

### 1.2.1 Inorganic Nanoparticle Synthesis

Inorganic nanoparticles can be synthesized by many methods such as physical, chemical, photochemical, or mechanical methods. The synthesis is also frequently categorized into either *top down* or *bottom up* approaches. The *top down* approach refers to the breaking down of bulk materials into smaller nanoparticles and generally involve physical or mechanical methods. The *bottom up* approach involves the nucleation and growth of the nanoparticles from atomic precursors, and are mostly associated with chemical methods [4]. In this thesis, all the nanoparticles studied are synthesized using chemical methods. This method involves several reagents, each of which serves a specific purpose. The first is a precursor which is the source of atoms that make up the nanoparticle. In the case of silver nanoparticle synthesis, this precursor is often silver nitrate, a metal salt. The second reagent is the reducing agent, which reduces the precursor. It achieves this by driving electrons to the precursor to form atoms, which are insoluble in the solution. Some common reducing agents are ascorbic acid, tannic acid, or sodium borohydride. Once the precursor is

reduced into atoms (e.g., Ag), it will aggregate into clusters, eventually forming the nanoparticle. The strength of the reducing agent controls the rate in which atoms are released in the solution, and influences the final nanoparticle structure [5]. Another important reagent is the stabilizer, which are bulky molecules such as polymers or surfactants that adsorb onto the surface of the nanoparticle, preventing additional growth and inducing a stabilizing effect through steric or electrostatic forces [6]. Stabilizers can also be used to direct the growth of nanoparticle into desired morphologies (e.g., plates, rods) by having preferential adsorption onto specific facets, such as the case of polyvinylpyrrolidone, which preferentially binds to the Ag(111) facet of silver nanoplates (i.e., the axial direction) [7]. Finally, while not required, many syntheses involve the use of seeds, which are small nanoparticle precursors. These nanoparticles act as catalysts by serving as nucleation sites for the precursor atoms to grow on, and also influence the structure of the nanoparticle [5]. A similar procedure can be performed for the synthesis of gold nanorods [8].

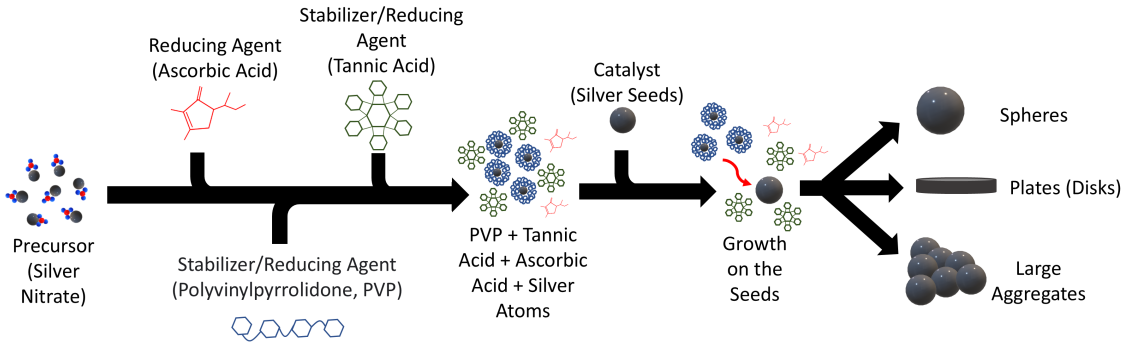



**Figure 1.1:** The chemical synthesis of silver nanoparticles using a precursor and several reducing agents and stabilizers.

Another powerful method to synthesize inorganic nanoparticles is the sol-gel method, which is commonly used for the synthesis of silica nanoparticles. To synthesize silica nanoparticles, the precursor, tetra orthosilicate (TEOS), is mixed with alcohol and water to undergo a hydrolysis reaction. The hydrolysis reaction breaks down the precursor into silanol groups (Si-OH). The condensation reaction then occurs between two silanol groups to form siloxane bonds (Si-O-Si). These bonds results in particle aggregation and the formation of the gel network. The pH of the solution greatly affects the reaction rate of the reactions and can be tuned to obtain silica nanoparticles of different sizes and dispersities. In addition, it can be used control the surface charge of the nanoparticles, where a pH lower than the point of zero charge will result in silica

nanoparticles with a positive charge and a pH higher than that will result in a negative charge. During the aging step, further polymerization reactions occur and the silica particles are physically rearranged into the gel structure. The final step is drying where all the water and alcohol groups evaporate, resulting in a solid powder which contains the silica nanoparticles. The nanoparticles can be redispersed in a solvent to obtain a colloidal dispersion [9]. Mesoporous silica nanoparticles are a special kind of nanoparticle that contains an ordered array of pores, which greatly increases the surface area of the nanoparticle. These pores can be created with the sol-gel method by using surfactants like CTAB or pluronics. Surfactants self-assemble into different structures based on the presence of inorganic particles and experimental conditions, such as the cubic, hexagonal or the lamellar phase. These ordered phases can serve as templates for the silica nanoparticles to condense on which results in mesoporous silica nanoparticles [9].

## 1.3 Colloidal Self Assembly

### 1.3.1 Inorganic Nanoparticle Self Assembly

The assembly of inorganic colloidal particles gives rise to structures with various functionalities and applications, such as, catalysis, sensor, and display technology. For instance, arranging nanoparticles into a 3D crystal structure, with an interparticle spacing larger than the diameter of the nanoparticle, optimizes their surface area while minimizing the overall volume of the crystal. This characteristic is particularly advantageous for applications in catalysis. A 3D assembly of nanoparticles, such as platinum, palladium, or cobalt, enhances the exposure of reactants to active sites, improving reaction efficiency [10]. In addition, the self-assembled structure imposes mechanical stability on the nanoparticles, preventing aggregation and enabling their reusability.

The assembly of plasmonic nanoparticles, such as gold and silver, plays a crucial role in sensor applications like surface-enhanced Raman spectroscopy (SERS). This technique leverages electromagnetic enhancement, which occurs when localized surface plasmon resonances (LSPR) in metallic nanostructures generate intense electromagnetic fields that amplify Raman signals. Furthermore, chemical enhancement, facilitated by charge transfer between the metal surface and the molecule, also amplifies the Raman signal, enhancing the overall sensitivity of the technique. By assembling plasmonic nanoparticles and confining

them to a small surface area, the Raman scattering signals of molecules adsorbed onto their surfaces are greatly enhanced, enabling the detection of extremely low molecule concentrations. This can be used for identifying substances such as toxins, biomarkers, and pesticide residues, as well as for studying drug delivery mechanisms [11].

The self-assembly of quantum dots (QD) plays an important role in quantum-dot light-emitting diodes (QLED), a display technology. Quantum dots are semiconductor nanoparticles that emit narrow emission spectra, which allows for highly saturated colors in displays. The emission spectra can be tuned by changing the size or composition of the quantum dot, meaning that a wide range of colors can be created. In a QLED display, self-assembled monolayers of quantum dots are used in the pixels of the display and play an important role in the performance of the device. For example, a uniform monolayer ensures efficient charge transport and balance within the device, and minimizes defects that prevent electron leakage. Additionally, a smooth and consistent QD layer is essential for uniform light emission, as it reduces scattering losses and enhances light extraction. Proper assembly also contributes to the stability of the device by preventing QD aggregation, which can otherwise lead to a loss of emission and degradation of overall performance [12] [13].

All the applications that were discussed demonstrate the importance of the self-assembly of inorganic nanoparticles. It is clear that obtaining control over the self-assembly process of inorganic nanoparticles has led to advancements in several technologies, which underscores its importance. Future chapters of this thesis will explore methods to control the self-assembly of inorganic nanoparticles.

### 1.3.2 Protein Self Assembly

The assembly of biomolecules, such as proteins, is ubiquitous in nature and is responsible for most of the processes that make life possible. Proteins are made up of a chain of amino acid monomers that fold into three-dimensional structures depending on their sequence. The three-dimensional structure of a protein is important for the protein to function properly. For example, enzymes or antibodies must have the correct structure to be able to bind to other molecules and perform their function. In addition, many proteins are composed of an assembly of smaller protein subunits. It is important for proteins to assemble properly since it is well-known that errors or defects in the assembly can lead to severe consequences. For example, neuro

degenerative diseases such as Alzheimer's and Parkinson's originate from the misfolding of the Amyloid protein that undergoes uncontrolled aggregation, resulting in the death of neurons [14].

As previously discussed, many examples of complex protein assemblies already exist in nature, and it has been a long-standing goal to synthetically design protein assemblies which could lead to several advancements in medicine and biotechnology. This is extremely challenging because proteins are made up of a sequence of tens to hundreds of amino acids, and these amino acids dictate the three-dimensional structure of the folded protein. To design a self-assembling protein, a balance between intermolecular forces (e.g., van der Waals, electrostatic, hydrogen bonding, steric) must be present in specific locations on the surface of the protein, so that they can assemble, but not aggregate uncontrollably. Recently, several successful tools have been developed to navigate this near-infinite design space. These tools are based on artificial intelligence (AI) and trained on data from the protein data bank (PBD). The PDB is a large comprehensive database that contains the amino acid sequence of a protein as well as its 3D structure (i.e., locations of all the atoms). The structure is experimentally solved with X-ray crystallography or Cryo-EM, and there are millions of entries in the databank. An example of a tool is Alphafold 2, released in 2021, which uses a deep neural network with transformer based architectures to predict the resulting 3D structure of an amino acid sequence [15]. This is extremely useful because it allows scientists to quickly obtain the predicted structure of different amino acid sequences, which would take a lot of effort and time to determine experimentally. In addition, it was one of the first major applications of AI in science, leading to several other applications.

Since the development of Alphafold, new tools have been developed. One useful tool for the design of protein assemblies is RFDiffusion, which is a generative model used to create new proteins that have specific function [16]. Diffusion models were first developed for image generation, where they are trained on vast amounts of images, and slowly learn how to recreate similar, realistic images. By implementing text encoding, the user can specify special characteristics of the image to be generated. Similarly, scientists can use RFDiffusion to generate proteins with special structural or functional motifs (e.g., binding sites). This has several useful applications such as designing vaccines made out of a self-assembling protein scaffold and antibody complex, which is more potent, stable, modifiable, and cheaper than other conventional vaccines [17]. Another example is designing proteins that bind to snake venom toxins, neutralizing them before they cause complications and induce damage to the human body [18].

### 1.3.3 Intermolecular Forces

Intermolecular forces are essential for colloidal self-assembly and synthesis, enabling the stability of colloidal systems. Although generally weaker than intramolecular forces, they significantly influence colloidal behavior. These forces are either attractive or repulsive, and achieving a balance between them is key to controlling the structure and function of a colloidal system. The fundamental intermolecular forces are discussed in this section.

#### Van der Waals

Van de Waals forces are generally attractive forces between molecules, but can be repulsive in some cases such as in heterogeneous systems. This force originates from the movement of electrons in a molecule resulting in permanent or transient dipoles. More specifically, the van der Waals force is composed of the dipole-dipole (Keesom), dipole-induced dipole (Debye), and induced dipole-induced dipole (London Dispersion) interactions. The Keesom force occurs between two molecules of permanent dipoles. The Debye force occurs when a molecule with a permanent dipole induces a dipole in another neutrally charged molecule. This force is weaker than the Keesom force, but can still be significant. Finally, the London Dispersion force exists in all molecules, regardless of their polarity. It originates from the oscillations of the orbital electrons in a molecule which induces temporary dipoles in nearby molecules, resulting in the attraction of the two molecules. The sum of the three forces makes up the van der Waals force [19]. The van der Waals interaction potential can be calculated with the Hamaker constant, which summarizes the interactions between a pair of macroscopic objects like colloidal particles. It is calculated by the pairwise summation of all the energies between two materials. A full derivation can be found in the literature [20], but a simplified equation for the van der Waals pair potential between two spheres of equal radii is:

$$\Phi_{vdw} = -\frac{Aa}{12S_0} \tag{1.1}$$

Where $A$ is the Hamaker constant, $a$ is the radius of the sphere, and $S_0$ is the surface-to-surface distance between the two spheres under the condition: $S_0 << a$.

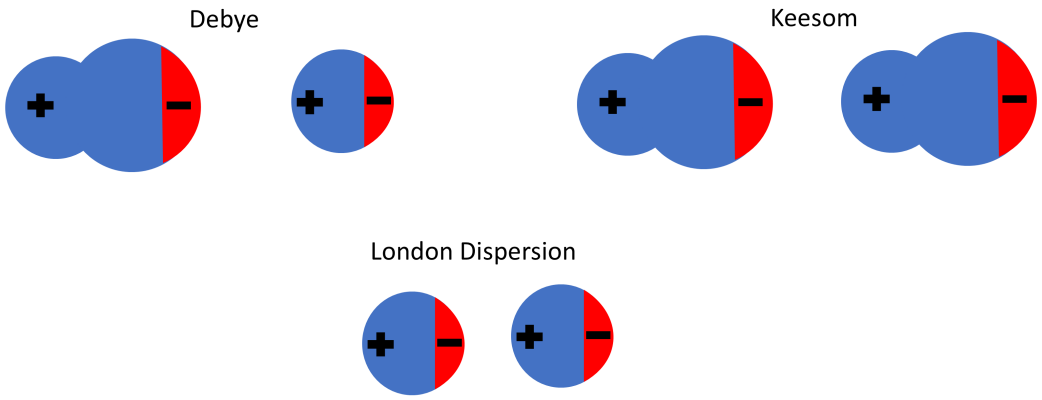



**Figure 1.2:** The van der Waals interaction, which is made up of the Debye, Keesom, and London Dispersion forces.

### Electrostatic

Electrostatic interactions occur between two charged particles and can be attractive or repulsive depending on the charge. It originates from Coulomb's law which describes the electrostatic potential between two point charges. The electrostatic potential depends on several factors such as the dielectric constant of the solvent, the presence of ions in the solution, pH, and the surface charge density. In colloids, the electrostatic interaction originates from interactions between the electrical double layer of two particles, which is a layer of ions that form around colloidal particles. If a colloidal particle has a surface charge, counter ions from the solution will adsorb onto the surface of the particle. The first layer of counter ions is known as the Stern later, which is mostly immobile. The second layer is the diffuse layer which is composed of both counter and co-ions surrounding the Stern layer. The ions in this layer are more mobile and the ion concentration decreases as they move further away from the colloidal particle. Several mathematical models have been developed to model the electrical double layer such as the Helmholtz Model which models the Stern layer with immobile counter ions, or the Gouy-Chapman model which models the diffuse layer with mobile counter ions. The Stern model combines the two models and is commonly used to model the electrical double layer as a Stern layer of immobile counter ions on the surface of the particle as well as a diffuse layer of mobile counter ions further away from the surface.

The size of the electric double layer affects the electrostatic interaction between colloidal particles, and several factors can affect the size of the double layer. One factor is the ionic strength of the medium which

is related to the electrolyte concentration and valence of ions in the medium. Higher ionic strengths will reduce the Debye length, which describes the thickness of the electric double layer. The collapse of the double layer results in reduced electrostatic interactions. The pH of the solution also affects the electrostatic force by altering the ionization state of surface functional groups on the colloids, which can change the size of the double layer. Finally, the surface charge density of the particle will affect the size of the electrical double layer since it controls the magnitude of the electric field.

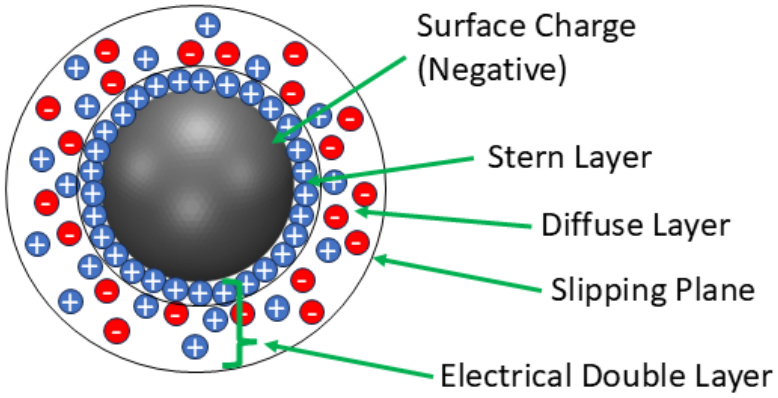



**Figure 1.3:** The electrical double layer surrounding a negatively charged colloidal particle. It is composed of the stern and the diffuse layer.

To model the electrostatic potential of spherical colloidal particles, the Derjaguin approximation is used to calculate the electrostatic interactions between curved surfaces. This approximation is shown in Equation (1.2). The surface potential of the particles is denoted by $\psi_0$ in volts, and $\kappa$ is the inverse Debye length, expressed in $\mathrm{m}^{-1}$. $k$ is the Boltzmann constant, $1.38 \times 10^{-23}$ J/K, and $T$ is the absolute temperature in kelvins. The relative permittivity (dielectric constant) of the medium is represented by $\epsilon$, while $\epsilon_0$ is the permittivity of free space, with a value of $8.854 \times 10^{-12}$ F/m. Equation (1.2) is only valid for the case where $S_0 << a$ and for particles with a constant surface density.

$$\Phi_E = \pi \epsilon \epsilon_0 a \psi_0 e^{-\kappa S_0} \tag{1.2}$$

The Debye length $\kappa$ is a measure of the thickness of the electrical double layer, and it can be calculated with Equation (1.3), for the case of a symmetric electrolyte of charge $z$. The elementary charge is denoted

by $e$ and has units of coulombs, $z$ is the valence of the electrolyte, $N_{AV}$ is Avogadros number, and $C$ is the concentration of electrolyte in Molarity.

$$\kappa = \sqrt{\frac{2000e^2z^2N_{AV}^2C}{\epsilon\epsilon_0RT}} \tag{1.3}$$

Equation (1.3) can be simplified for the case of deionized water at room temperature and this is shown in Equation (1.4) [20].

$$\kappa^{-1} = \frac{0.304}{|z|\sqrt{C}} \tag{1.4}$$

**Steric**

Steric forces are interaction forces between colloidal particles that are repulsive. This force comes from the adsorption of bulky ligands or polymers on the surface of a colloidal particle which creates a protective layer surrounding the particle. When two particles with protective layers approach each other, the layers interact with each other leading to the steric repulsive interaction. This repulsion can occur due to several reasons. The first reason is the overlap of electron clouds between the protective layers would repel each other. The second is from osmotic pressure which occurs when water molecules are expelled between the two layers. This imbalance in osmotic pressure drives water back between the two layers, separating them from each other. The last reason is elastic repulsion which occurs when the ligands or polymers in the protective layer are compressed when the particles approach each other. If the natural state of the ligands or polymers is uncompressed, this would lead to elastic repulsion between the particles. Steric repulsion is an important method to stabilize colloidal particles because it is mostly insensitive to changes in pH, temperature, and solution ionic strength. This makes it an ideal method to stabilize colloidal particles that undergo a wide range of conditions. A common method of inducing steric forces in colloidal systems is the use of polyethylene glycol (PEG), which is a neutrally charged water soluble polymer. When modified with a thiol molecule (PEG-thiol), it can be used to functionalize gold nanoparticles, increasing its colloidal stability [21]. Other functionalization methods can be used to attach PEG to particles of different kinds of materials.

While the steric force is repulsive, the addition of bulky ligands or polymers to a solution of colloidal

particles can also cause aggregation under certain conditions. A well-known effect is bridging flocculation which occurs when a long polymer attaches itself onto two or more colloidal particles, bringing the particles together and causing aggregation. Bridging flocculation is a common method of water purification, since most solid particles in contaminated water (e.g., clay, dirt) are negatively charged. By adding a cationic polymer to the solution, bridging flocculation occurs causing the separation of the solid particles from the liquid phase. This is advantageous because flocculation occurs without the addition of ions to the solution (to reduce electrostatic repulsion) which is important when the objective is purified water. Despite this, it is well known that adding an excess concentration of cationic polymer to the solution can have the opposite effect of electrostatically stabilizing the solid particles when the polymer wraps itself around each particle. Therefore, the concentration of polymer added to the solution can determine whether the attractive or repulsive force is enhanced [20]. Another attractive effect that occurs when high concentrations of polymers are added to a colloidal suspension is known as the depletion effect. This effect occurs when the concentration of unadsorbed polymers in the solution is high. When two colloidal particles approach each other, the polymers are expelled from the space between the particles, causing an imbalance of osmotic pressure. This results in the movement of water between the colloidal particles to the bulk solution (i.e., high polymer concentration), creating an attractive force between the colloidal particles.

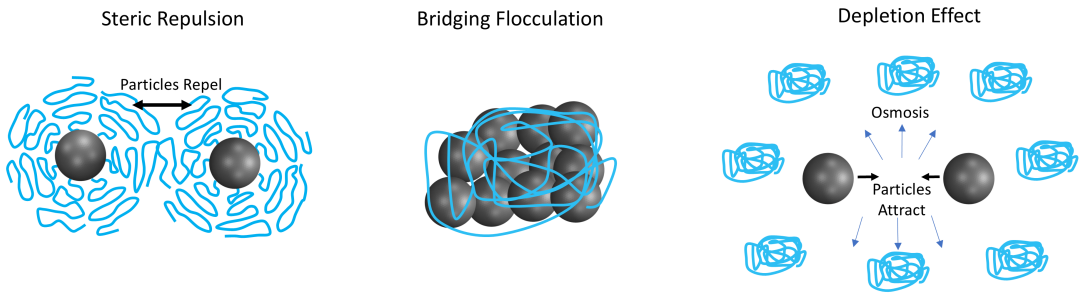



**Figure 1.4:** The steric force is a repulsion force between particles coated in bulky ligands such as PEG. Under certain conditions, the addition of bulky ligands can cause aggregation by bridging flocculation or by the depletion effect.

### Hydrophobic

The hydrophobic interaction results in a repulsive force between a hydrophobic surface and water. This is because hydrophobic surfaces are incapable of bonding with water with methods such as hydrogen bonding or ionic interactions. The hydrophobic force between colloidal particles is attractive and causes them to aggregate in an aqueous solution, minimizing the system's free energy. In addition, this attractive force is long-ranged and relatively strong. The origin of the hydrophobic force is entropic. This is because water molecules tend to organize themselves in an ordered cage-like structure around hydrophobic molecules, which increases entropy. To minimize the system's free energy, hydrophobic molecules must aggregate to minimize the total surface area exposed to water, minimizing the system's free energy. In colloid science, the hydrophobic force can be used to induce aggregation of colloidal particles through hydrophobic ligands such as alkanes. For example, the use of thiol-modified alkanes can be used to functionalize gold nanoparticles causing them to aggregate [21].

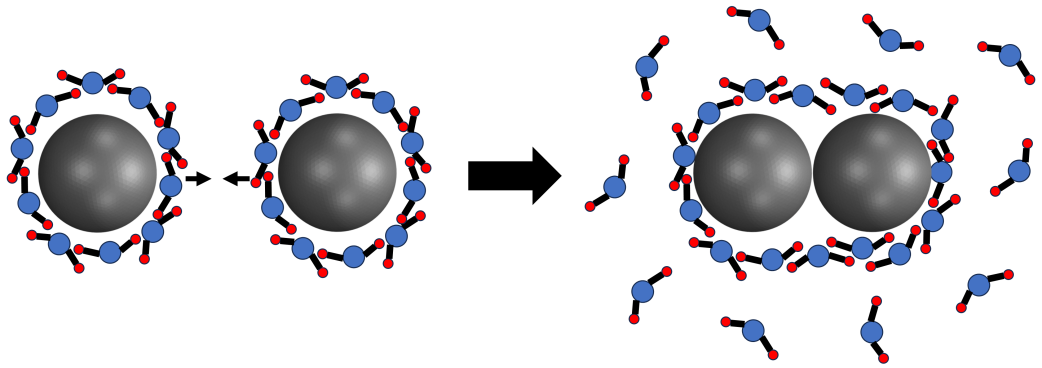


**Figure 1.5:** The hydrophobic force is an entropic force caused by water molecules forming an ordered cage like structure around hydrophobic particles. To reduce the amount of surface area exposed to the water, the particles will aggregate.

### DLVO Theory

Derjaguin, Landau, Verwey and Overbeek (DLVO) theory can be used to predict the colloidal stability of electrocratic colloids, which are one of the most basic colloidal systems. The theory describes the total interaction energy of electrocratic colloidal systems as the sum of the attractive van der Waals interaction and the electrostatic repulsion. The total interaction energy has units of joules, but is commonly reported in

units of kT, an energy unit composed of the Boltzmann constant multiplied by temperature. The interaction energy is usually plotted as a function of the distance between particles, and features of the curve can be used to determine the stability of a colloidal system. Often, the curve contains a "potential energy barrier" which is the maximum height of the curve and represents the energy that must be surpassed for aggregation to occur. As a general "rule of thumb", if an energy barrier is less than a few kT, aggregation will occur rapidly. Therefore, the higher the energy barrier, the higher the stability of the colloidal system.

Several equations can be used to calculate the interaction energies in the system. The total interaction energy can be described with Equation (1.5), where $\Phi_T$ represents the total interaction potential in joules, which is the sum of $\Phi_A$ (the van der Waals attraction potential in joules) and $\Phi_R$ (the electrostatic repulsion potential in joules).

$$\Phi_T = \Phi_A + \Phi_R \tag{1.5}$$

The van der Waals attraction can be calculated with Equation (1.6) for spherical particles in a binary system, where the Hamaker constant, $A_{212}$, is expressed in joules and describes the material-dependent strength of van der Waals forces between a particle of material 2 in a medium of material 1. $a$ is the radius of the particles in meters, and $S_0$ is the separation distance between particle surfaces in meters.

$$\Phi_A = -\frac{A_{212}a}{12S_0} \tag{1.6}$$

The electrostatic repulsion is shown in Equation (1.7). The surface potential of the particles is denoted by $\psi_0$ in volts, and $\kappa$ is the inverse Debye length, expressed in $\mathrm{m}^{-1}$. $k$ is the Boltzmann constant, $1.38 \times 10^{-23}$ J/K, and $T$ is the absolute temperature in kelvins. The relative permittivity (dielectric constant) of the medium is represented by $\epsilon$, while $\epsilon_0$ is the permittivity of free space, with a value of $8.854 \times 10^{-12}$ F/m.

$$\Phi_R = \pi\epsilon\epsilon_0 a\psi_0 e^{-\kappa S_0} \tag{1.7}$$

The equation for the inverse Debye length is shown in Equation (1.8) for the case of a symmetric electrolyte in deionized water at room temperature. $C$ represents the concentration of electrolyte in the solution in mol/L and $z$ is the valence of the electrolyte. [20].

$$\kappa^{-1} = \frac{0.304}{|z|\sqrt{C}} \tag{1.8}$$

The DLVO interaction energies are then plotted shown in Equation (1.5).

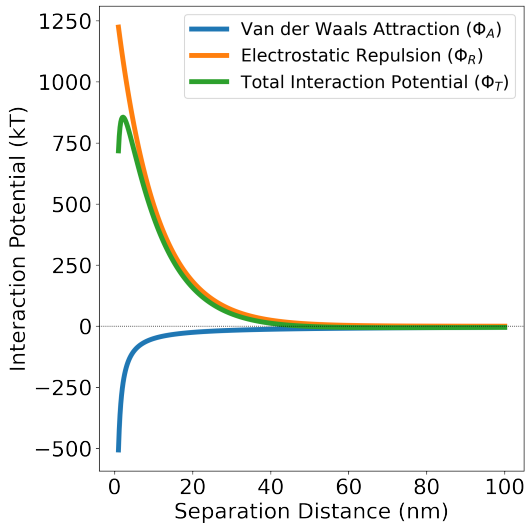



**Figure 1.6:** The DLVO interaction energies.

### 1.3.4 Ligand-Mediated Assembly

#### DNA

DNA-mediated assembly is a popular method for assembling colloidal particles due to its programmability and tunability. This programmability comes from the numerous combinations of nucleotide base pairs that can be used to form a DNA strand. DNA assembly works based on the complementary base pairing of the nucleotides by hydrogen bonds, specifically adenine (A) with thymine (T) and cytosine (C) with guanine (G). C and G bind with three hydrogen bonds while A and T bind with two. DNA can be single or double-stranded, and two single strands of complementary base pairs will bind via hydrogen bonds to form a double strand if it is below the strand's melting temperature. This means that a DNA double-strand can easily be divided into single strands simply by heating it to a temperature higher than its melting temperature. The length of the DNA strand and the C and G content determine the melting temperature of the strand. The directionality of DNA is also an important property of DNA. Nucleotides (A, C, T, G) are composed of a

phosphate group, a pentose sugar, and a nitrogenous base. In terms of directionality, the phosphate group is often referred to as the 5' side and the third carbon of the sugar molecule is the 3'. In a single DNA strand, covalent bonds form between the 5' end of one nucleotide to the 3' end of another one. DNA directionality means that a double strand can only form between single strands that are antiparallel.

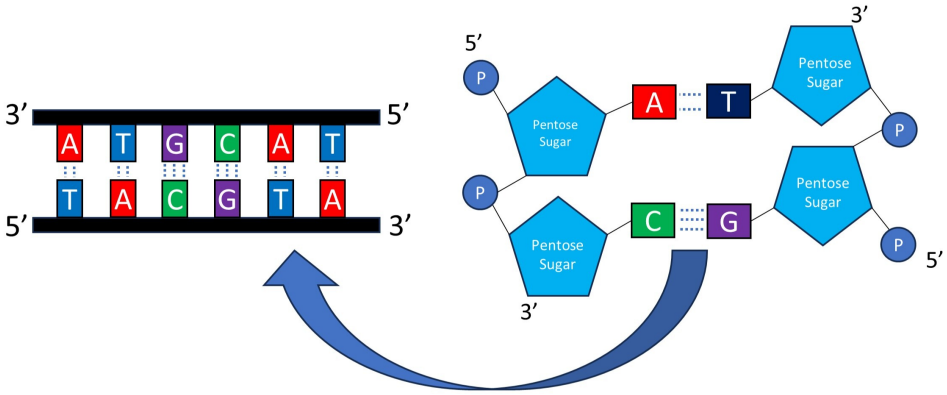



**Figure 1.7:** Figure showing the programmability and directionality of DNA.

DNA can also be chemically modified with other molecules, many of which are commercially available. These modifications are made on either the 5' or 3' end of the DNA strand and is commonly used to attach DNA molecules to colloidal particles. Common modifications are the addition of a thiol group to the beginning of a DNA strand which can covalently bond to materials such as gold [22]. Another common modification is the addition of a cholesterol molecule, which can bind to hydrophobic regions such as the inner part of lipid membranes [23]. In addition to its ease of chemical modification, DNA is a popular method for assembling colloidal particles because of its ability to form high-quality 3D colloidal crystals. In the literature, there are several examples that use DNA to create thermodynamic colloidal crystals that are made out of gold nanoparticles [22] [24]. In short, the method involves functionalizing a gold nanoparticle with thiol-modified DNA, adding another DNA strand that has complementary base pairs to the first strand on both of its ends, and then performing a thermal protocol which usually involves heating the system to the melting point of the DNA and then slowly cooling. It is hypothesized that crystallization occurs because of the gradual cooling from the thermal protocol, where DNA strands can frequently de- and rehybridize, rearranging the positions of the colloidal nanoparticle into the thermodynamic crystal structure [24]. Despite this, it is well documented that other factors such as DNA length, complementary DNA length, DNA

concentration, solution ionic strength, and particle size determine the degree of crystallization as well as the structure of the crystal [25]. The programmability of DNA as well as it ease of chemical modification make DNA a powerful tool for assembling colloidal particles.

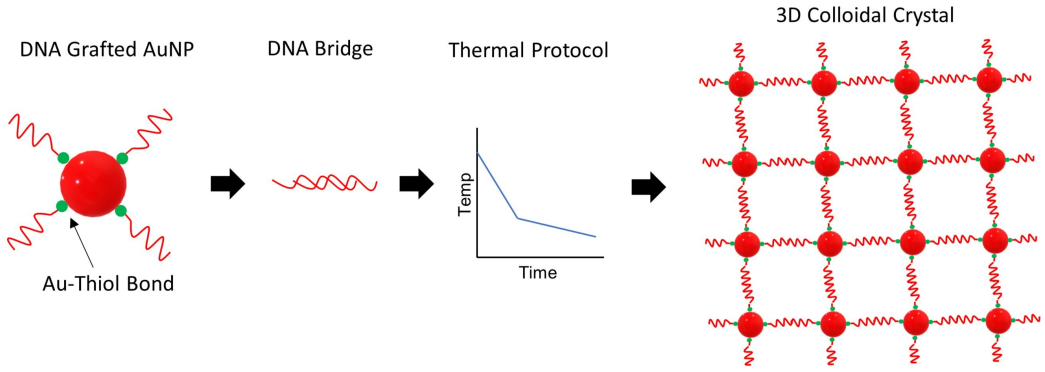



**Figure 1.8:** The process of assembling gold nanoparticles (AuNP) into 3D colloidal crystals using DNA.

### Host-Guest Interactions

Host-guest interactions refer to the non-covalent bonding between two distinct molecules, leading to the formation of unique complexes. In these complexes, guest molecules are typically encapsulated or incorporated within the host molecules, linking them together using physical bonds. This type of interaction is frequently used to create supramolecular systems because of its responsiveness to external stimuli like light, pH, temperature, etc. [26]. In addition, the structure of the host is often designed to complement the size and shape of the guest, leading to high selectivity. One kind of host-guest complex used in this thesis is the light-responsive azobenzene (guest) and $\beta$-cyclodextrin (host) complex, which is used to assemble the RhuA protein in light-sensitive supramolecular assemblies. In short, $\beta$-cyclodextrin has a high affinity to (E)-azobenzene, but not to (Z)-azobenzene. The (E)-isomer is thermodynamically more stable than the (Z)-isomer, so the formation of the complex is the most stable form. Exposure to UV light triggers a conformational change from (E)-azobenzene to (Z)-azobenzene, leading to the separation of the complex, while visible light reverses this process [27]. The light-sensitive azobenzene/cyclodextrin complex is only one of the many kinds of host-guest interactions that can be used to assemble colloidal particles.

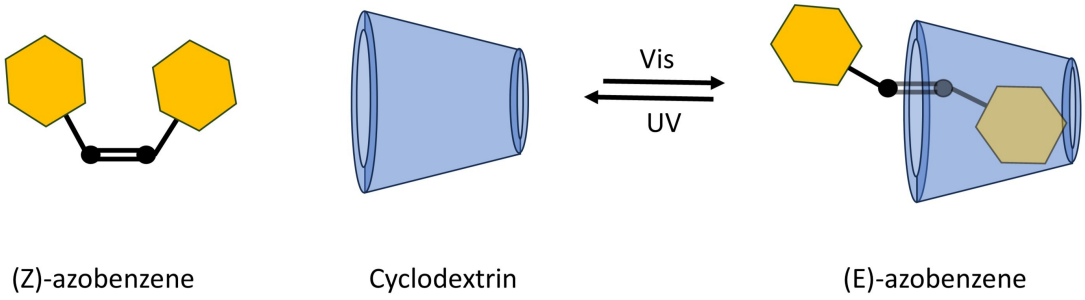



**Figure 1.9:** The light responsive host-guest interaction between azobenzene and cyclodextrin.

# Chapter 2

# Methods

## 2.1 Artificial Intelligence and Machine Learning Methods

### 2.1.1 Gaussian Process

Gaussian Processes (GP) are supervised machine learning models that are effective in making predictions in small to medium datasets. A GP works by fitting linear combinations of kernels to data points. Kernels, a hyperparameter of a GP, are functions (e.g., Gaussian, exponential, quadratic) used to measure the covariance or the similarity between two data points. The kernels are placed in arbitrary points in the parameter space, and the weighted sum of these kernels makes up an estimation function which is used to make predictions of the outcomes of the input parameter set. The weight coefficients are sampled from a Gaussian distribution, which results in a distribution of different estimator functions called Gaussian Process Priors. A loss function, which is composed of the sum of a similarity term and a regularization term, is minimized to obtain the best estimation function. The similarity term is composed of the difference between the actual training data points and the estimation function's predictions of the points. The regularization term accounts for the smoothness of the function which is important to prevent overfitting. The loss function is minimized by changing the weight coefficients of the kernels and a parameter that affects the smoothness of the fit, obtaining an estimation function that best represents the data [28]. The advantages of GPs are that they are nonlinear and nonparametric. Nonparametric regression is advantageous because it does not assume any underlying correlations in the data, in contrast to parametric models such as linear regression. The drawback

of nonparametric models is that they are less accurate than parametric ones when there is a known analytical relationship between the data points. Another drawback is that they are hard to interpret. It is easy to understand how a linear regression model works but not a GP, which is important when trying to gain scientific insight from a dataset. Finally, the computational time used to fit a GP tends to grow cubically compared to the dataset size, which limits its application to small datasets [29].

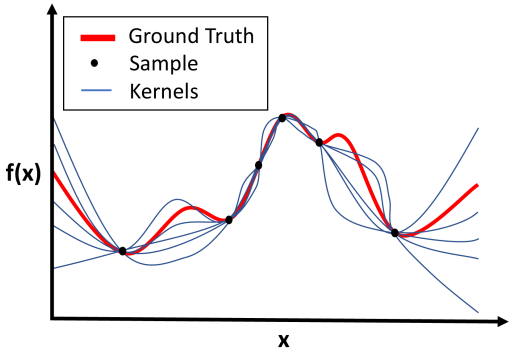



**Figure 2.1:** A Gaussian Process works by using a series of kernels to estimate the ground truth.

### 2.1.2 Bayesian Optimization

Bayesian Optimization (BO) is an optimization algorithm that uses the information from all previous samples to determine which experiments to perform next, allowing it to find the optimal solution within a few samples. This optimization technique is composed of two parts: the surrogate model, and the acquisition function. The surrogate model, most commonly a GP, is an estimation function that is used to predict the outcomes of possible experiments. The acquisition function is what determines which experiments to perform next. One function that is commonly used is the probability of improvement which samples based on which outcomes have the highest probability of having any kind of improvement over the current best value. Another function, the expected improvement, samples where the greatest improvement is expected. To obtain optimal results efficiently, these acquisition functions must balance exploration and exploitation, which is a problem encountered by all optimization algorithms. Exploitation is the search for promising regions where an optimum is likely to be located. Repeatedly exploiting risks results in the confinement in local maxima which will lead to small improvements in the best value, until no improvements occur at all. Exploration is the search of regions that have not been sampled before. This ensures that information

on the whole parameter space is collected and that there are no other solutions better than the predicted optimal value (i.e., global maximum). Repeatedly exploring will also lead to poor performance because the search will be focused on different areas of the parameter space. For optimal performance, the algorithm must balance the amount of exploration and exploitation, which is often difficult since this is controlled by a hyperparameter (i.e., a parameter determined by the user's experience or from literature). BO works by creating a loop with two main steps. With the GP as a surrogate model and a chosen acquisition function, the first step of BO is to define the GP prior, which is made by fitting the kernel functions to the initial data points (in the first iteration, the prior is also the posterior). After more experiments are conducted, the prior is updated and becomes the posterior, from which the acquisition function is used to determine which experiments to perform next. This loop is repeated until a stopping criterion is met, which is usually convergence or performing a predetermined number of iterations/experiments [30].

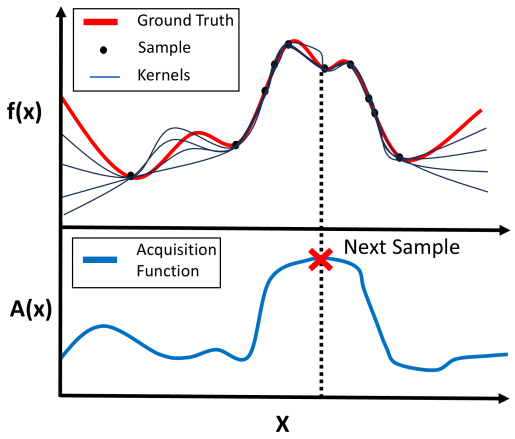



**Figure 2.2:** Bayesian Optimization works by using a gaussian process as a surrogate model which is trained on sampled point. The acquisition function is used to decide where to sample next.

### 2.1.3 Genetic Algorithm

Genetic Algorithms are useful to optimize objective functions that are inexpensive to evaluate. This is done by mimicking the process of evolution to optimize a desired feature. The algorithm consists of four steps: evaluation, selection, crossover, and mutation. The evaluation step starts with candidate solutions to the objective function and the score of each solution according to the objective function. This step is simply ordering the inputs of each candidate according to its score or fitness. In the selection step, the samples

from the evaluation step are selected according to their fitness value (i.e., samples with higher fitness have a higher chance of being selected). The next step is the crossover step where candidate solutions that were selected are randomly mixed with one another. The purpose of this is to combine the characteristics of good solutions to create a better one. Finally, the mutation step introduces a random change in one of the new candidate solutions. The probability of a mutation occurring is controlled by a hyperparameter, which controls the amount of exploration and exploitation [31]. An advantage of a GA is that the calculations are computationally inexpensive which makes it suitable for simulated experiments such as model-fitting algorithms [32]. It is not well suited for experiments where evaluating the objective function is expensive such as when physical experiments are performed. Bayesian optimization would be a better choice for that task. The code for the genetic algorithm which was used in this thesis can be found online (`https://github.com/huatc/GA`).

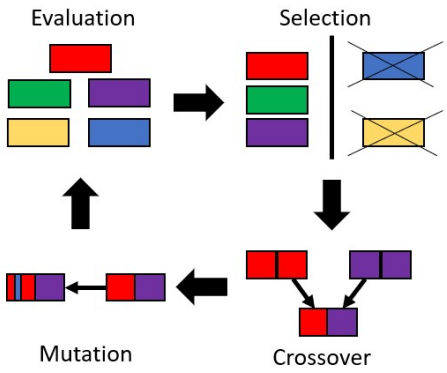



**Figure 2.3:** Procedures of a genetic algorithm.

## 2.2 Robotic Synthesis

### 2.2.1 Opentrons OT2

The Opentrons OT2 liquid handling robot was used for many of the high-throughput experiments in this thesis. This low-cost robot (<$10,000) transfers liquids with pipettes of volumes of 1 to 1000 microliters with an accuracy of 1 microliter. In addition, it can be controlled by a Python-based API, which allows for highly customizable protocols. In this thesis, the OT2 robot was used for the synthesis of inorganic nanoparticles,

where different volumes of stock solutions were mixed in a well plate to synthesize nanoparticles of different morphologies and sizes. Python code to perform this task for a large number of samples was developed in the Pozzo research group (`https://github.com/pozzo-research-group/OT2-DOE`).

Nanoparticle synthesis was the main application of the OT2 in this thesis, leading to several useful lessons and tips, many of which may seem trivial. The first lesson is to always watch the OT2 when performing a protocol for the first time. There are various points of failure that will result in unsuccessful experiments. The second is that proper calibration is extremely important for the accurate transfer of liquids. Before any experiment, it is useful to test a protocol without any liquids to ensure that the pipette properly picks up pipette tips and aspirates/dispenses in the center of the specified labware. If the pipette is not centered, offsets can be applied to correct for this misalignment using the calibration tool in the OT2 application. The third lesson is carefully set the pipetting options depending on the liquid being transferred. This is because the ability of the pipette to dispense small volumes of fluid greatly depends on the fluid being transferred. For example, when dispensing an aqueous solution of surfactant, air bubbles will often form at the end of the dispensing. To solve this issue, a solution could be to aspirate more liquid than the amount to be dispensed. Another scenario involves the transfer of viscous fluids such as an aqueous solution of a bulky polymer. In this case, liquid bubbles can remain at the end of the pipette tip after dispensing the liquid, which would result in less liquid being transferred to the sample. One possible solution is to implement a mixing protocol where the pipette uses the same tip to mix the sample solution, ensuring that any residual liquid, such as a bubble, is properly incorporated into the sample. However, this would require the pipette tip to be washed or discarded before the next step. To effectively wash the pipette tip, it was determined that sequentially aspirating and dispensing a large volume of water across three separate deionized water reservoirs was sufficient to eliminate sample contamination. Finally, liquid evaporation is a significant issue especially for long protocols. To mitigate this, adding flexible covers to labware or using a low temperature are some possibilities.

Apart from the advantages of improved efficiency and reproducibility that the liquid handling robot offers, it also opens up the possibility of performing experiments with timed interventions. A protocol can be created where the time delays and order in which the robot performs a pipetting step are variables. This is applicable to colloidal syntheses and assemblies that are kinetically driven, where the time and order in

which reagents are added have an effect on the final result. One advantage of using a liquid-handling robot for experiments with timed interventions over microfluidic devices is that batches of samples can be created in parallel. This means that optimizing the schedule in which the samples are created (the transfer of stock solutions to the sample well) to reduce dead time, can drastically reduce the total experiment time. This is essential for high-throughput experiments where the time delays are long and where numerous samples are synthesized. Code to perform scheduled optimization was created (`https://github.com/pozzo-research-group/otto`) and used in some projects [33]. In short, the inputs of the code are the volumes of each reagent to add to each sample, the time delay of the addition of each reagent, and the order of addition. Using these inputs, an algorithm schedules the pipetting to minimize the total experiment time by minimizing the dead time between adding reagents.

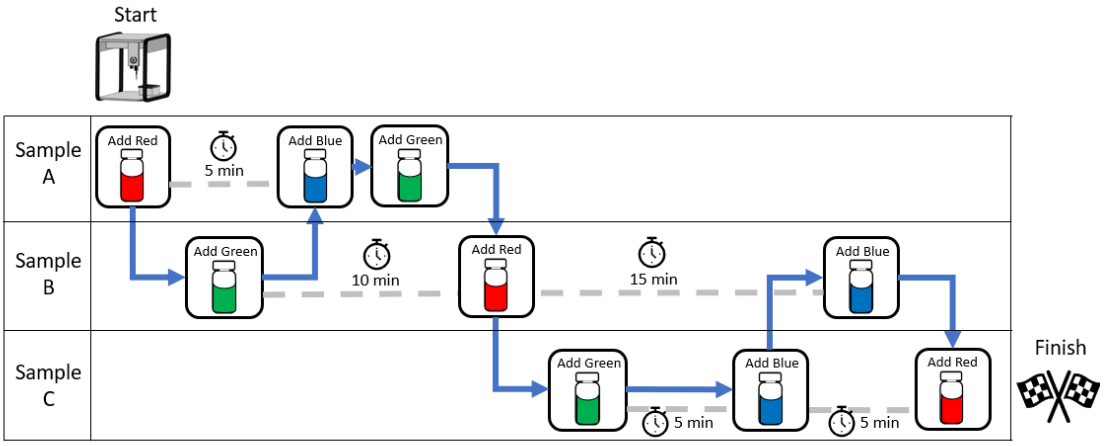



**Figure 2.4:** A protocol for the OT2 liquid handling robot was created to vary the volume, time, and order of mixing stock solutions. An algorithm optimizes the experiment scheduling to minimize the total experiment time by minimizing the dead time between the addition of reagents.

## 2.3 Characterization Methods

### 2.3.1 UV-Vis Spectroscopy

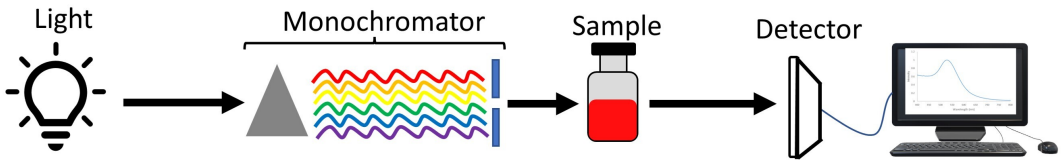



**Figure 2.5:** Experimental setup for UV-Vis spectroscopy. A monochromator is used to select a wavelength of light to expose the sample to. The detector then measures the transmittance of the sample.

UV-Vis spectroscopy obtains information on how matter interacts with different wavelengths of UV to visible light. In the experiment, a monochromator selects a specific wavelength of light that is then exposed to the sample. A detector directly behind the sample then detects how much light of that specific wavelength gets transmitted through the sample. Normally, a graph of the absorbance of each wavelength of light is obtained using Beer's law.

$$A = log_{10}(T) = \epsilon l c \tag{2.1}$$

Where $A$ is the absorbance, $T$ is the transmittance, $\epsilon$ is the molar absorptivity coefficient, $l$ is the path length of light, and $c$ is the concentration of the sample. In the Beer-Lambert Law, the absorbance is equal to the product of the molar absorptivity coefficient, the path length of light, and the concentration of the sample. The molar absorptivity coefficient is the only term that is a function of wavelength, which means that it accounts for the shape of the spectrum. The path length and the concentration are constants that account for the intensity position of the spectrum but do not change the shape. Because of this, UV-Vis spectroscopy can also be used to measure the concentration of samples, by measuring the change in absorbance. This can be used for turbidity tests where the absorbance of a specific wavelength is determined, which is useful for quantifying aggregation in some colloidal systems.

In the special case of plasmonic nanoparticles, the structure of the nanoparticle can be indirectly inferred from the shape of the UV-Vis spectrum due to localized surface plasmon resonance. Generally, the

peak position of the spectrum is related to the size of the nanoparticle and the number of peaks is related to the shape. In addition, the UV-Vis spectrum can serve as an identification tool by comparing it to computationally simulated spectra in order to identify a plasmonic nanoparticle's structure. However, this must be done cautiously since many different structures can have similar spectroscopic signatures. Finally, a major advantage of UV-Vis spectroscopy is that it can be a relatively fast and inexpensive technique using plate readers, which makes it ideal for screening plasmonic nanoparticles in high-throughput.

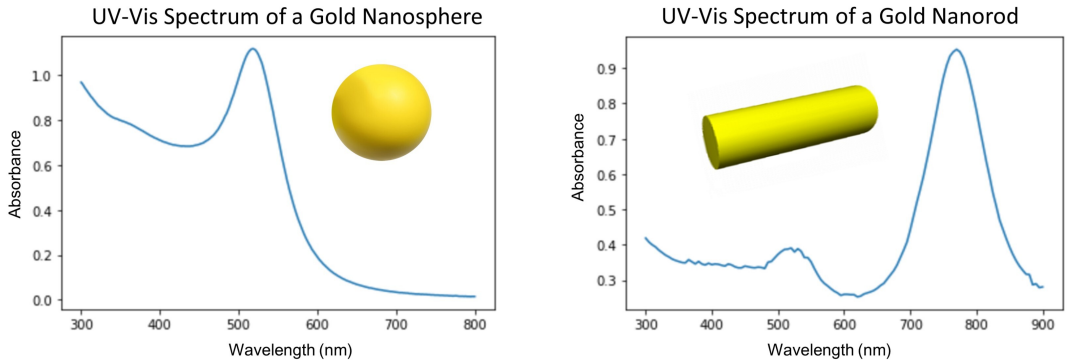



**Figure 2.6:** Differences in the UV-Vis spectrum of a gold nanorod and a gold nanosphere. The spectra are normalized to have a maximum value of 1 and a minimum of 0. The absorbance is reported in arbitrary units.

### 2.3.2 Dynamic Light Scattering

Dynamic Light Scattering (DLS) can be used to determine the size and dispersity of colloidal particles of nanometers to tens of microns. The theory behind DLS is that the movement of colloidal particles can be described with Brownian motion because thermal fluctuations are the dominant force acting on the particles. In addition, colloidal particles scatter light due to the differences in refractive index between the solvent and solute. DLS measures the intensity of the light, as a function of time, that is scattered by each particle from an incident laser beam. The intensity of the scattered light depends on the position and size of the particles and results from the interference (constructive or destructive) between the scattered rays originating from each particle [20]. A schematic of the experimental setup of a DLS experiment is shown in Figure 2.7.

The theory behind DLS is that larger particles have lower scattering fluctuations over time than smaller particles, because of longer diffusion times. The extent of the fluctuations can be quantified by the auto-

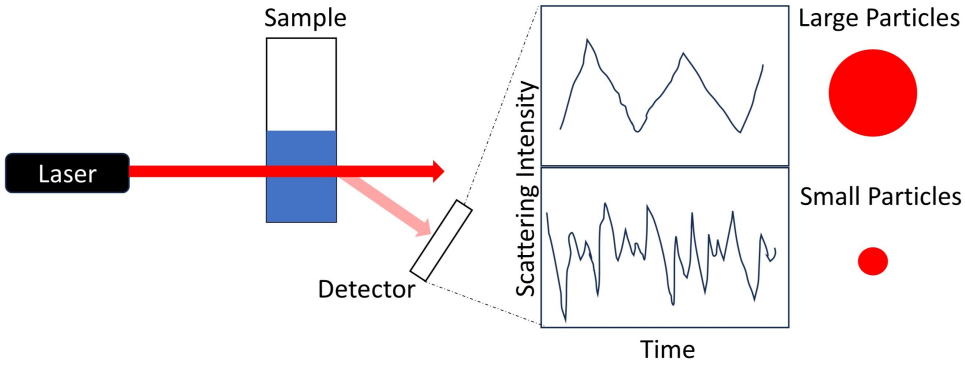



**Figure 2.7:** Experimental setup of a DLS experiment. A laser is scattered off a sample, which is then measured by a detector. The intensity of the scattered light is measured as a function of time. Large particles have lower scattering fluctuations over time than smaller particles due to longer diffusion times.

correlation function, which measures the similarity between a scattering intensity signal and a copy of itself taken at short delay time interval $\tau$, from which the autocorrelation function is derived (2.2). At $\tau \to 0$, the low elapsed time for the particles to undergo significant displacement relative to their prior positions results in two similar scattering intensity patterns and a high correlation coefficient. At $\tau \to \infty$, there is enough time for the particles to be displaced relative to their prior positions which results in a low correlation coefficient. This results in the typical "S" curve in Figure 2.8, where there is initially a high correlation coefficient which then drastically lowers to 0. The shape of the curve is used to determine the particle size and dispersity.

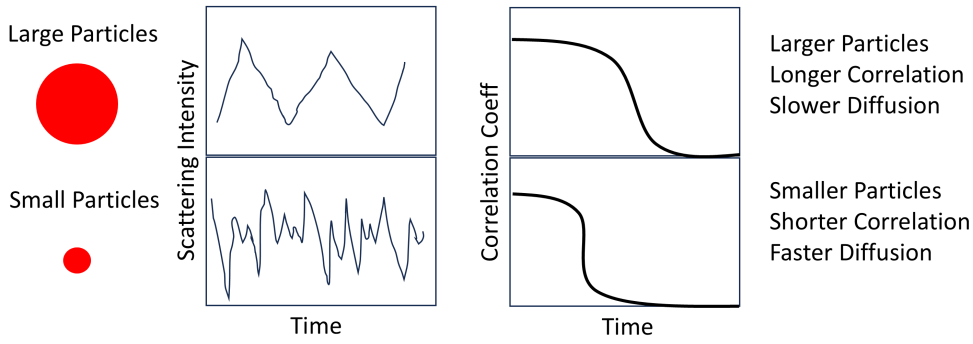



**Figure 2.8:** The differences between the scattering intensities of a large particle and a small one. The intensity autocorrelation function quantifies the difference in fluctuations between the particles of different sizes.

The intensity autocorrelation function (second-order autocorrelation) is derived from the scattering intensity and is defined as:

$$g^{(2)}(\tau) = \frac{\langle I(t)I(t+\tau)\rangle}{\langle I(t)\rangle^2}$$

Where $g^{(2)}(\tau)$ is the intensity autocorrelation function, $I(t)$ is the scattered light intensity at time $t$, $\tau$ is the delay time, and $\langle\cdot\rangle$ is a time average. The Siegert relation connects the intensity autocorrelation function to the field autocorrelation function $g^{(1)}(\tau)$ as follows:

$$g^{(2)}(\tau) = 1 + \beta \left| g^{(1)}(\tau) \right|^2$$

Where $g^{(1)}(\tau)$ is the electric field autocorrelation function (first-order), and $\beta$ is the coherence factor (depends on experimental setup). For monodisperse particles, the following equation can be used to obtain the diffusion constant $D$.

$$g^{(1)}(\tau) = 1 + e^{-2Dq^2\tau} \tag{2.2}$$

Where $q$ is the scattering vector. The diffusion constant $D$ can be calculated from the fit and then used to calculate the hydrodynamic radius $R_H$ from the Stokes-Einstein equation [34].

$$D = \frac{k_B T}{6\pi\mu R_H} \tag{2.3}$$

Where $k_B$ is the Boltzmann constant, $T$ is temperature, $\mu$ is the viscosity, and $R_H$ is the hydrodynamic radius. The hydrodynamic radius is defined as the radius of a sphere that diffuses at the same rate as the particle being measured [35]. Therefore, DLS is sensitive to adsorbed polymers, the electrical double layer, or other large molecules on the surface of particles [20].

### 2.3.3 Zeta Potential

Zeta potential is defined by the electrostatic potential difference between the slipping plane and the dispersing medium. It is an important property that determines the stability of a colloid since its value is related

to the electric double-layer repulsion between particles. Zeta potentials with high magnitudes indicate a strong electric potential while magnitudes close to zero indicate a weak one. A useful application of the zeta potential in colloid science is to determine if molecules are adsorbed onto a particle's surface. For example, a negatively charged particle will have a drastically different zeta potential when cationic molecules are adsorbed onto its surface. To measure zeta potential, the electrophoretic mobility must first be calculated, which is the particle velocity divided by the electric field strength.

$$u_E = \frac{V_p}{E_x} \tag{2.4}$$

Depending on the size of the double layer, determined by the product of the particle size $a$ and the inverse Debye length $\kappa$, different equations should be used to calculate the zeta potential. Equation (2.5) is the Helmholtz-Smoluchowski limit and (2.6) is the Hückel limit. Both of these limits are also summarized in Henry's formula which calculates the zeta potential for both of these limits.

$$u_E = \frac{\epsilon\epsilon_0\zeta}{\mu} \quad \text{for } \kappa a > 200 \tag{2.5}$$

$$u_E = \frac{2}{3}\frac{\epsilon\epsilon_0\zeta}{\mu} \quad \text{for } \kappa a < 0.1 \tag{2.6}$$

The variables used in the equations are as follows: $u_E$ represents the electrophoretic mobility, defined as the particle velocity $V_p$ divided by the electric field strength $E_x$. The electric field strength is denoted as $E_x$, and $V_p$ is the electrophoretic velocity of the particle. The dielectric constant of the medium is $\epsilon$, while $\epsilon_0$ refers to the permittivity of free space, which has a value of $8.854 \times 10^{-12}\,\mathrm{C/V \cdot m}$. The zeta potential is denoted by $\zeta$. The viscosity of the medium is represented by $\mu$. The Debye length, $\kappa^{-1}$, is a measure of the thickness of the electrical double layer [20].

The most common way to calculate electrophoretic mobility is through laser-Doppler electrophoresis. In short, a laser beam of known wavelength is split into two and then reflected to intersect on the stationary plane of the electrophoretic cell. This causes interference fringes of a known distance, which can calculated by Equation (2.7). In this equation $\lambda$ denotes the wavelength of the incident light, $n$ is the refractive index of the medium, and $\theta$ is the scattering angle. The fringe spacing in this method is represented as $\Delta$.

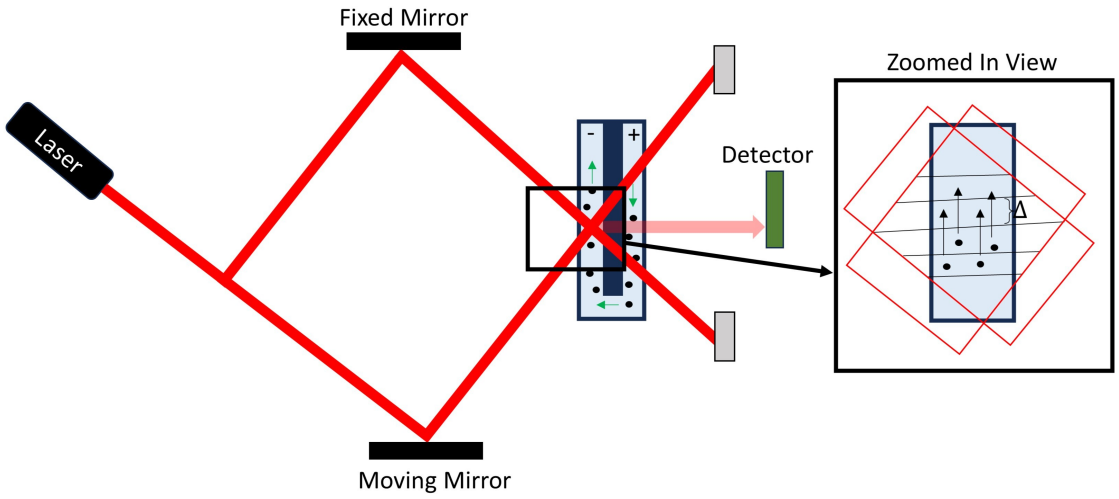



**Figure 2.9:** The experimental setup to determine the zeta potential by laser-Doppler electrophoresis.

$$\Delta = \frac{\lambda}{2n\sin(\theta/2)} \tag{2.7}$$

After the fringe spacing is calculated, the time a particle takes to traverse one fringe spacing is calculated. This is done by determining the flickering frequency, which is measured from the autocorrelation function (discussed in the previous DLS section). The autocorrelation function is created from the scattering of light from the particles as they traverse the distance of one fringe spacing. Finally, the velocity of the particles is calculated using Equation (2.8) [20].

$$V_P = \tau_\Delta \Delta \tag{2.8}$$

### 2.3.4 Small Angle Scattering

Small-angle scattering (SAS) is a powerful tool used to characterize the structure of colloidal nanoparticles ranging from 1 to 1000's nm in the dispersion state, non-invasively and, in some cases, in high throughput. The use of X-rays or neutrons, which have wavelengths comparable to the sizes of the nanostructures, results in the high sensitivity of structural features of these sizes and is advantageous over visible light-based characterization techniques such as dynamic light scattering. In addition, the ability to characterize a sample in its native state or while performing an external manipulation differentiates it from many other techniques.

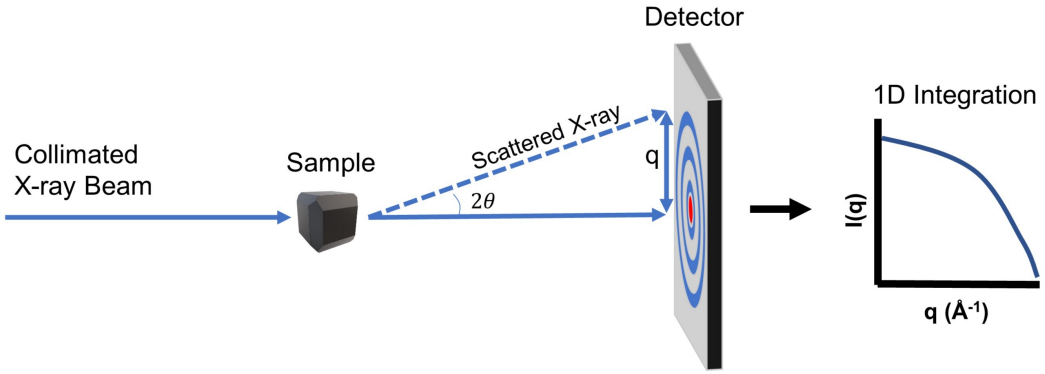



**Figure 2.10:** Experimental setup of a SAXS experiment. A collimated beam of X-rays is scattered off a sample, which is then measured by a detector. The 2D image is then integrated to create a 1D scattering curve.

While the data interpretation of SAXS curves can be difficult (inverse problem), the scattering curve of any scattering object can be mathematically calculated and compared with the experimental scattering curve to identify it (forward problem). In theory, small-angle X-ray scattering (SAXS) can be used to study any material. However, the contrast, or the difference in electron density between materials, often determines how feasible and time-consuming a scattering experiment will be. Scattering experiments on materials with high contrast such as metals, crystals, and others are often easy to perform with quick measurement times. On the other hand, biomolecular samples such as proteins, peptides, lipids, and others often require longer measurement times and sample environments with low background noise. SAXS is also accessible since it can be performed in laboratory-based sources, which are becoming popular due to advancements in detector technology and X-ray brightness. All these characteristics make SAXS an important tool to study colloidal synthesis or self-assembly.

Another characterization technique that is similar to SAXS is small-angle neutron scattering (SANS), which is performed at large scale facilities. This technique is like SAXS except that the beam is composed of neutrons instead of X-rays. While the X-ray scattering length of elements monotonically increases with atomic number, the neutron scattering length is related to the strong neutron force, making it unrelated to the atomic number and sensitive to isotopes. The main advantage of this is the ability to perform contrast variation experiments due to differences in scattering lengths of hydrogen and deuterium. By changing the composition of the solvent ($D_2O/H_2O$), contrast variation allows for the ability to separate the scattering

contributions of individual subunits in complex systems. This is especially useful for studying hybrid systems that undergo conformational rearrangement upon the formation of a weakly associated complex, such as studying how the structure of a protein changes when attached to a ligand. Another common application of SANS is studying the structure of a membrane protein in a lipid bicelle. In this experiment, scattering contributions of the lipid bicelle are matched out using the correct solvent composition, resulting in the scattering of only the membrane protein. In this thesis, SANS is used to study the encapsulation of silica nanoparticles in lipid bilayers. At high concentrations of $D_2O$, there is high scattering contrast from both the silica and lipids, enabling the extraction of structural information from both materials from the scattering curve. Another unique concept of SANS is the strong incoherent scattering of hydrogen, which contributes to a high flat background signal in scattering curves from samples containing aqueous solvents. This elevated background is undesirable, as it requires the material of interest to scatter neutrons more intensely than the background to achieve a high-quality scattering curve. Because of this, when choosing a solvent composition ($D_2O$/$H_2O$), it is often advised to use a solvent with as much $D_2O$ as possible while still maintaining the highest amount of contrast between the material in the sample and the solvent.

**Theory**

In SAS, a collimated beam of X-rays or neutrons is scattered off a sample of interest resulting in a change in the momentum of the X-ray or neutron. The X-rays or neutrons are assumed to scatter elastically and scattered only once. The interactions between the scattered X-rays or neutrons result in a scattering pattern that is measured by a detector located behind the sample. This 2D pattern is then azimuthally integrated to obtain the intensity of the scattered X-rays or neutrons ($I(q)$) as a function of the momentum transfer vector ($q$).

$$q = \frac{4\pi}{\lambda} \sin(\frac{\theta}{2}) \tag{2.9}$$

Equation (2.9) shows the equation of the momentum transfer vector where $\theta$ is the angle of the scattered beam and $\lambda$ is the wavelength of the X-ray or neutron. The scattering equation for many two-phased systems that do not have any anisotropic ordering can then be simplified and written as:

$$I(q) = \frac{N}{V}(\Delta\rho V_p)^2 P(q) S(q) + background \tag{2.10}$$

In eq. (2.10), $\frac{N}{V}$ is the volume density of the scattering particles, $V_p$ is the volume of the particles and $\Delta\rho$ is the contrast or the difference between scattering length densities of the particles and the background. $P(q)$ is the form factor which is the scattering due to the size and shape (e.g., spheres, disks, rods, cubes) of the particle. $S(q)$ is the structure factor which is the scattering due to the position of the particles in relationship to other particles. Finally, the $background$ refers to the scattering of any material in the path of the beam other than the material of interest.

**Data Analysis**

Data from SAS is often ambiguous and hard to interpret, which is why additional information from other characterization techniques or from prior knowledge is needed to interpret SAS data. In the case where the shape of the material can be approximated with a geometric object (e.g., sphere, cylinder, cube, and others), model fitting of the analytical solution of the geometric object to the data can be performed. The analytical solutions of simple geometric objects are straightforward to implement and easily solvable. For example, the equation for the scattering of a sphere is:

$$I(q) = \left( \frac{3V\Delta\rho(\sin(qr) - qr\cos(qr))}{(qr)^3} \right)^2 \tag{2.11}$$

In (2.11), $r$ is the radius of the sphere, which can be tuned in the model fitting algorithm to fit the experimental data. Similar expressions exist for the scattering of different geometric structures.

Using the analytical solutions of geometric structures to fit data is valid only if the sample is known to be completely monodisperse in size and shape, which is uncommon in colloidal systems. Polydispersity in size and shape can be accounted for by calculating the scattering curves of structures with different sizes and taking the weighted average of the curves. Depending on the polydispersity distribution, this calculation can be computationally expensive. Several software packages have been developed for this purpose such as Sasmodels (`www.github.com/SasView/sasmodels`) or McSAS [36].

Many biomolecular systems, such as proteins, have irregular shapes, which prevents the use of analytical

solutions of geometric shapes to fit the data. In these cases, the scattering curve of the protein can be calculated and compared to the experimental curve. A common method to perform this calculation is through the Debye equation. The Debye equation calculates the scattering intensity from the atomic coordinates (3D-cartesian coordinates) of any structure and also accounts for changes in the scattering length density of atoms. This makes it an extremely useful tool in modeling small-angle scattering data. The derivation of the Debye equation [37] starts with the amplitude of a scattered wave, which is the result of the interference pattern of all the scattered waves. The assumption is that every wave is scattered once and only elastic scattering occurs.

$$\Psi(q) = \sum_j f_j \exp(-iq \cdot r_j) \tag{2.12}$$

In (2.12), $f_j$ refers to the scattering length density of the $j^{th}$ atom and $r_j$ is the position of the $j^{th}$ atom. The intensity of the scattered wave can then be calculated using the amplitude.

$$I(q) = |\Psi(q))|^2 = \sum_k \sum_j f_k f_j \exp(-iq \cdot (r_k - r_j)) \tag{2.13}$$

The Debye equation is then the spherically averaged intensity of the intensity of the scattered wave (2.13). By spherically averaging the exponential part of (2.13) the Debye equation is derived.

$$\langle \exp(-iq \cdot (r_k - r_j)) \rangle = \frac{\sin(qr_{jk})}{qr_{jk}} \tag{2.14}$$

In equation (2.14), $r_{jk}$ refers to the euclidean distance between points $r_j$ and $r_k$.

$$I(q) = \sum_k \sum_j f_k f_j \frac{\sin(qr_{jk})}{qr_{jk}} \tag{2.15}$$

## 2.4 Experiments at Large Scale Facilities

Large scale facilities are government funded laboratories with beamlines that are used for experiments. Each beamline consists of a beam of X-ray or neutrons and equipment such as collimators, lenses, sample environment, and detectors for the desired experiment. In this thesis, ultra small angle x-ray scattering

(USAXS) and SANS experiments were performed at the beamlines. USAXS was used to characterize large nanoparticle assemblies of hundreds of nanometers in size. The large size of the assemblies results in an ultra small scattering angle. To measure these small angles, the USAXS instrument uses a Bonse-Hart device which allows for the measurement of ultra small angles, with the drawback of a loss of X-ray flux. However, synchrotrons are able to produce extremely bright X-rays, which allows for quick and high quality USAXS measurements. Currently, experiments with neutrons such as SANS, can only be performed at large scale facilities due to the complexity and high cost of producing neutrons that are intense enough to use in experiments. In this thesis, SANS was used to investigate the encapsulation of silica nanoparticles in lipid bilayers.

The first step to perform experiments at large scale facilities is to write a proposal. For most experiments, beamtime is scarce and highly competitive, so the proposed experiment must be carefully planned. Before writing a proposal it is advised to contact the beamline scientist, who is responsible for the instrument. They can help determine if their beamline is necessary for the user's experiment and what kinds of modifications or sample holders are available for the experiment. They can also help determine a rough estimate of the exposure time needed for each of your samples. After this, a proposal should be written to obtain beamtime. In the proposal, the user should provide a brief summary of their experiment and justify why the specific beamline is needed to meet their objectives. For highly competitive beamlines, it is recommended to justify why other beamlines at other facilities are not suitable for the proposed experiment. In addition to this, it is beneficial to provide complementary data from other characterization techniques showing that the experiment being proposed is feasible. Finally, while proposing the experiment, a concise experiment plan should be developed to demonstrate how the user will make use of their time. Details such as what samples will be exposed to the beam and the measurement time of each sample need to be stated in the proposal. Beamline allocation committees determine which proposals are funded.

If the proposal is accepted and beamtime is allocated, the next step is to plan the experiment. This is the most important step and preparation is essential for a successful experiment. First, a method to track sample details should be developed, such as what samples are made, when they are made, how much volume is made, and the order that they should be measured in the beam. Once this is done, the user also should decide what materials and equipment to ship to the facility. This should be done carefully, because

unpredictable events are common in beamline experiments, so it is often advised to bring extra sample materials or lab supplies in case they are needed. Synchrotron beamlines can experience outages, which will cut short the experiment, so it is important to prioritize the order in which samples are measured. It is also recommended to perform all possible preparations at the user's home institution, such as printing labels or labeling vials, to minimize the amount of unnecessary work done at the beamline. When shipping materials, the user should consider the hazards of the materials and the sample's stability due to temperature fluctuations during shipping, since they can experience extremely hot or cold temperatures depending on the time of the year. Additionally, it is recommended not to perform beamline experiments alone. They can be highly stressful experiences and support will be needed from colleagues for several tasks, such creating samples, driving, getting food, or emotional support. Nevertheless, please treat your colleague respectfully during the trip and be grateful for their support. Failure to do so may result in a future beamline trip where your experiments are conducted alone. For some large scale facilities (Argonne and Oak Ridge), it is highly recommended for someone to have a driver license, since these facilities are located in remote areas. Finally, beamline experiments also provide a valuable opportunity to network and gain firsthand insight into working at a national laboratory. Throughout the experiment, users will engage with various staff members who may share perspectives on their career paths.

## 2.5 Self Driving Labs

As discussed in this thesis, controlling the synthesis or assembly of colloidal systems involves navigating a vast experimental design space. One way to effectively navigate a large design space is using a self-driving lab (SDL), a strategy for efficiently optimizing desired outcomes within an experimental design space. A SDL is composed of robotic tools to assist in sample formulation, high-throughput scattering techniques for sample characterization, and machine learning methods to guide decision-making. In an SDL, the first step is to identify the specific structure or property to be optimized, followed by selecting the appropriate characterization method(s) capable of analyzing that structure or property. In this step, a signal (e.g., scattering curve) from the characterization method corresponding to the desired structure or property is simulated, serving as the target for the SDL. The objective of the SDL is to produce a sample with a signal that most closely aligns with the simulated signal from the nanomaterial possessing the targeted

structure or property. In nanoparticle synthesis, this approach is particularly valuable because the structure of nanomaterials often determines their function. As a result, applying nanomaterials to real-world problems typically requires first synthesizing them into specific structures. Numerous examples of SDLs applied to the synthesis of plasmonic nanoparticles can be found in the literature [38] [29].

While the concept of a SDL is simple, implementing it requires careful planning and execution. The first step is to select an experimental system that can be automated with the tools that are available. The chemical synthesis of plasmonic nanoparticles is a common choice for liquid handling robots because it can be performed by mixing different reagents and precursors to obtain nanoparticles. After determining the system, one must determine the design space within the experimental system to be explored. In the context of the chemical synthesis of plasmonic nanoparticles, this involves defining the concentration range and resolution of each reagent that the robotic system can prepare. This selection is crucial, as the design space must be sufficiently expansive to enable thorough exploration of different regions and optimize synthesis outcomes effectively. However, the design space cannot be too expansive which would require a extremely large number of sampling for the SDL to converge to a solution. After selecting the design space, an experimental protocol must be developed and executed using a robotic platform, such as the OT-2 liquid handling robot. Key experimental design considerations including the choice of sample holder, pipetting protocol, and pipette volume range, play a crucial role in the success of the SDL. Once these design considerations are finalized, the reproducibility of the robotic platform should be tested with samples from several different regions in the design space to ensure consistency in sample preparation before implementation with the SDL.

After the creating the samples, the next step is to characterize them. Several of the scattering-based characterization techniques discussed in this chapter (e.g., DLS, SAXS) are suitable for this task because they can be performed in high-throughput with autonomous sample loaders, require minimal sample preparation, measure the samples in their native state, and because they can characterize the structure of an ensemble of nanoparticles. For the special case of plasmonic nanoparticles, UV-Vis Spectroscopy is a powerful characterization method because of its ability to indirectly determine the structure of an ensemble of nanoparticles through surface plasmon resonance effects. The raw data from these scattering-based techniques is obtained in the form of a 1-dimensional function or curve, and structural information can then be inferred from fea-

tures of this curve such as peak positions, peak widths, or slopes. This information is then compared to the initial target which informs the AI-agent on its decisions for the next samples. As shown in the next chapter, the choice of how to compute this "distance" between sample and target is nontrivial and greatly affects the performance of the SDL [39]. Finally, while many examples of SDLs in the literature rely on a single characterization method, incorporating multiple characterization techniques can be beneficial for obtaining higher-quality information about the nanoparticle's structure or property. This is particularly important because data interpretation from scattering-based techniques is inherently ambiguous, as different nanoparticle structures can produce identical scattering curves. By introducing a second independent characterization method, the likelihood of two distinct structures generating identical data signals across both methods is significantly reduced.

The next component of a SDL is the AI-agent, which is responsible for decision making and ultimately creating a sample with the targeted structure or function. One common choice is Bayesian optimization which uses a Gaussian process and acquisition function to guide the optimization [39]. As data is collected, the Gaussian process gets trained, which allows the acquisition function to make more informed decisions on the regions in the design space to sample next. Hyperparameters are critical for the success of the algorithm and can control behaviors such as the exploration or exploitation rate. One limitation of Bayesian optimization is that it is generally restricted to low-dimensional spaces due to the high computational cost associated with training the Gaussian process and sampling the acquisition function [30]. However, in the case of SDLs, this is typically not a concern, as the number of dimensions and the size of the generated dataset are usually small. Another widely used optimization method is the genetic algorithm [40], which operates based on a set of predefined rules (evaluation, selection, crossover, and mutation) to guide optimization. These rules are applied only to the current iteration of data, making the approach computationally efficient. However, unlike Bayesian optimization, genetic algorithms do not use information from previous iterations, which may result in a need for more data to achieve convergence. Finally, a stopping criterion must be established for any SDL. Common approaches include stopping the process after a predefined number of iterations or setting a distance threshold relative to the target structure to be achieved before stopping.

In summary, a SLD is a powerful tool to efficiently optimize desired outcomes within an experimental design space. While it is mostly autonomous, its implementation requires careful planning and execution

in order to be successful. This section contains some design considerations that should be accounted for a successful implementation of a SDL.

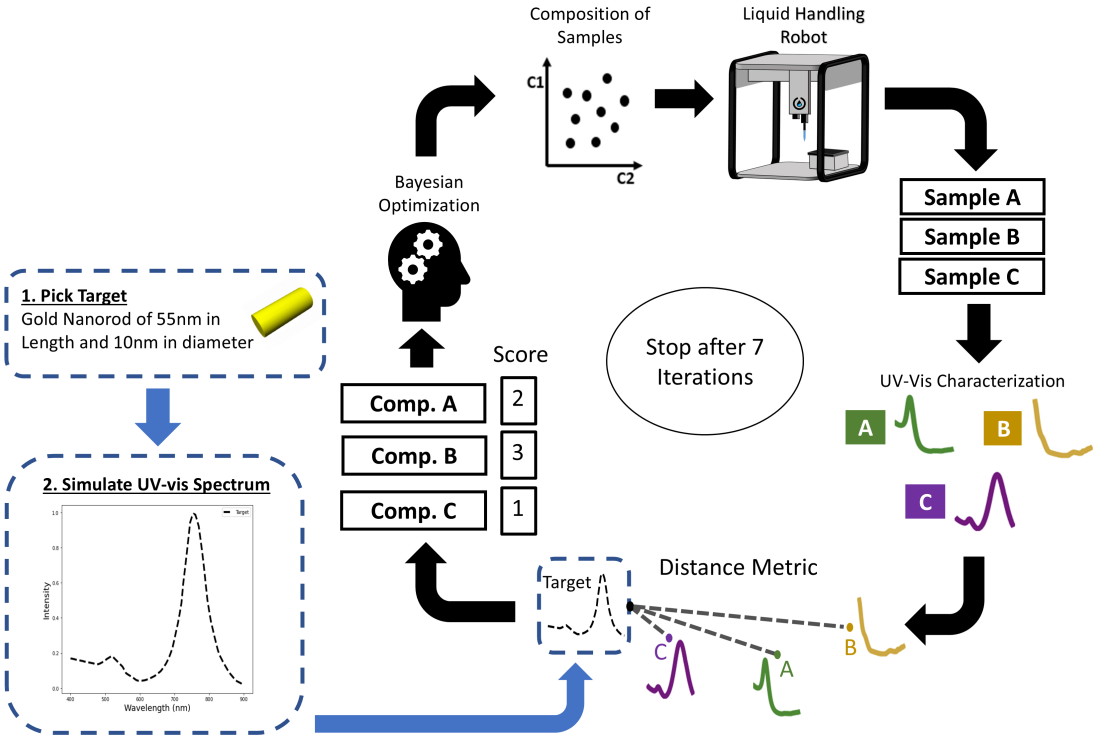



**Figure 2.11:** Detailed Setup of a self driving lab for the synthesis of gold nanoparticles. First, a targeted structure is chosen and its UV-Visible spectrum is simulated. The objective of the self driving lab is to synthesize a sample that most closely matches the targeted spectrum.

# Chapter 3

# Autonomous retrosynthesis of gold nanoparticles via spectral shape matching

This work presented in this chapter is from:

- Kiran Vaddi, Huat Thart Chiang, Lilo D Pozzo, Autonomous retrosynthesis of gold nanoparticles via spectral shape matching, *Digital Discovery*, 2022, 1, 502-510. [39]

## 3.1 Introduction

A critical part of closed-loop retrosynthesis systems is the need to determine a reward function to provide feedback to the AI agent on the outcomes of its choices. In the case of data defined by scalar values (i.e., nanoparticle size, conductivity, viscosity) the analysis is simple as these kinds of datasets are easily comparable to one another. However, data analysis becomes much more challenging in the case of functional datasets where the shape of the curve is related to the property of interest. Functional data is ubiquitous in many high-throughput material characterization techniques such as spectroscopy and small-angle scattering. Expert knowledge can be used to extract information from these curves by identifying critical features, such as peak positions in the case of UV-Vis spectroscopy or the slopes of the curves in the case of small-angle scattering. Another method is to define a score function that results in a scalar similarity between two curves such as the Euclidean or the Cosine distance. These score functions, however, only measure differences in

the intensity scale (differences in the y-axis), which in some cases can be an inaccurate representation of the shape of the curves. To solve this problem, the Amplitude-phase distance metric is introduced, which is able to compare the shape of two curves based on variations in the x and y axes and assign a scalar similarity score between them.

## 3.2 Methods

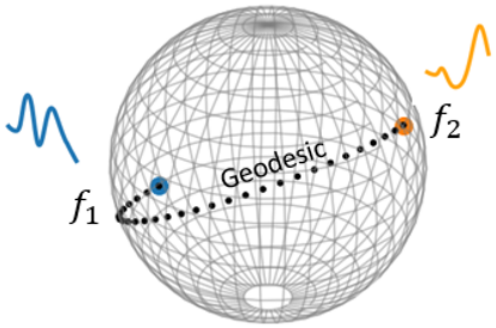
The Amplitude-phase metric was created to take into account the shape of a function or curve while calculating a distance between two functions $f_1$ and $f_2$. By using function spaces, it is possible to analyze data with differential geometry methods. The Amplitude-phase metric is a two-part measurement method. First, the phase component (3.1) uses a warping function to decouple the x and y variations of the function, which allows the computation of variance in both axes independently. It then calculates the variation in the x-axis. The amplitude component (3.2) then calculates variations along the y-axis using the square root slope function transformation. The total Amplitude-phase distance can then be calculated by adding the amplitude and phase components. By adding weights for each component, the contributions of each component to the total score can be adjusted.

$$d_p(f_1, f_2) = \cos^{-1}\left(\int_0^1 \sqrt{\dot{\gamma}(t)}dt\right) \tag{3.1}$$

$$d_a(f_1, f_2) = d([q_1, q_2]) = \inf ||q_1 - q_2^{\circ}\gamma||_{L^2} \tag{3.2}$$

## 3.3 Results and Discussion

In this section, case studies of the effect of the distance metric in retrosynthesis campaigns were performed. The performance of retrosynthesis campaigns with the Amplitude-phase metric and with the Euclidean distance were compared in the synthesis of a gold nanorod target. In Case Study 1, the design space was limited to 2 variables and used a target that was known to be in the design space. In Case Study 2, the design space consisted of 8 variables and used an arbitrarily chosen simulated nanorod target. All other variables in




**Amplitude (Variations in the y-axis)**

$$d_a(f_1, f_2) = d([q_1], [q_2]) = \inf_{\gamma} \|q_1 - q_2 \circ \gamma\|_{\mathbb{L}^2}$$

**Phase (Variations in the x-axis)**

$$d_p(f_1, f_2) = \cos^{-1}\left(\int_0^1 \sqrt{\dot{\gamma}}(t)\, dt\right)$$

**Amplitude Phase**

$$d_{ap}(f_1, f_2) = d_a(f_1, f_2) + d_p(f_1, f_2)$$

**Figure 3.1:** The formulas used to calculate the Amplitude and Phase distances.

the retrosynthesis campaigns were kept constant so differences in performance can be attributed exclusively to the distance metric.

### 3.3.1 Case Study 1: 2-Dimensional Optimization

To showcase the Amplitude-phase distance, a high-throughput experimental retrosynthesis campaign of a gold nanorod structure was performed starting with a targeted UV-Vis spectra. Retrosynthesis campaigns were performed in parallel, with a different similarity metric in a two-dimensional reaction space. The results for two retrosynthesis campaigns are described and discussed: (a) using the Euclidean distance between the raw spectra, (b) using the Amplitude-phase distance. Following the seed mediated synthesis procedure described in Nikoobakht and El-Sayed [8], gold nanorods were synthesized with five chemicals: gold(III) chloride trihydrate, hexadecyltrimethylammonium bromide (CTAB), ascorbic acid (AA), silver nitrate ($AgNO_3$), and gold seeds. An arbitrary nanorod was synthesized by pipetting a pre-specified volume of the five solutions and its UV-Vis spectrum was used as the target for the optimization. Each optimization had a batch size of 4 samples and the iterative process continued until a total of 7 iterations had been synthesized. The concentrations of CTAB, gold(III) chloride trihydrate, and gold seeds were kept constant and equal to those that were used to synthesize the target sample. The search space for the autonomous retrosynthesis is then defined as the two-dimensional reaction space of the concentrations of silver nitrate ([$AgNO_3$]) and ascorbic acid ([AA]). An OT2 liquid handling robot was used to autonomously synthesize the samples, and a Biotek plate reader was used to characterize the samples using UV-Vis spectroscopy with wavelengths of 400–900 nm in increments of 5 nm. The samples were made in 96-well polystyrene

microplates, which were heated to around 30°C during the synthesis using a hot plate. After the synthesis, the samples were kept at the same temperature for 50 minutes, so that the nanoparticles could fully grow before being characterized by UV-Vis spectroscopy. All the retrosynthesis campaigns had identical initial conditions (i.e., the first iteration had the same concentrations and measured spectra).

**Table 3.1:** Table I. Concentrations of the Chemical Design Space and the Arbitrary Nanorod Target

| Reagent | Stock Solution Concentration (M) | Target Concentration (M) | Concentration Range (M) |
|---|---|---|---|
| CTAB | $2.0 \times 10^{-1}$ | $6.40 \times 10^{-2}$ | $6.40 \times 10^{-2}$ |
| Gold (III) Chloride Trihydrate | $1.0 \times 10^{-3}$ | $1.96 \times 10^{-4}$ | $1.96 \times 10^{-4}$ |
| Silver Nitrate | $6.4 \times 10^{-4}$ | $6.20 \times 10^{-5}$ | $0 - 7.38 \times 10^{-5}$ |
| Ascorbic Acid | $6.3 \times 10^{-3}$ | $3.60 \times 10^{-4}$ | $0 - 7.27 \times 10^{-4}$ |
| Gold Seeds | $1.8 \times 10^{-5}$ | $1.44 \times 10^{-6}$ | $1.44 \times 10^{-6}$ |

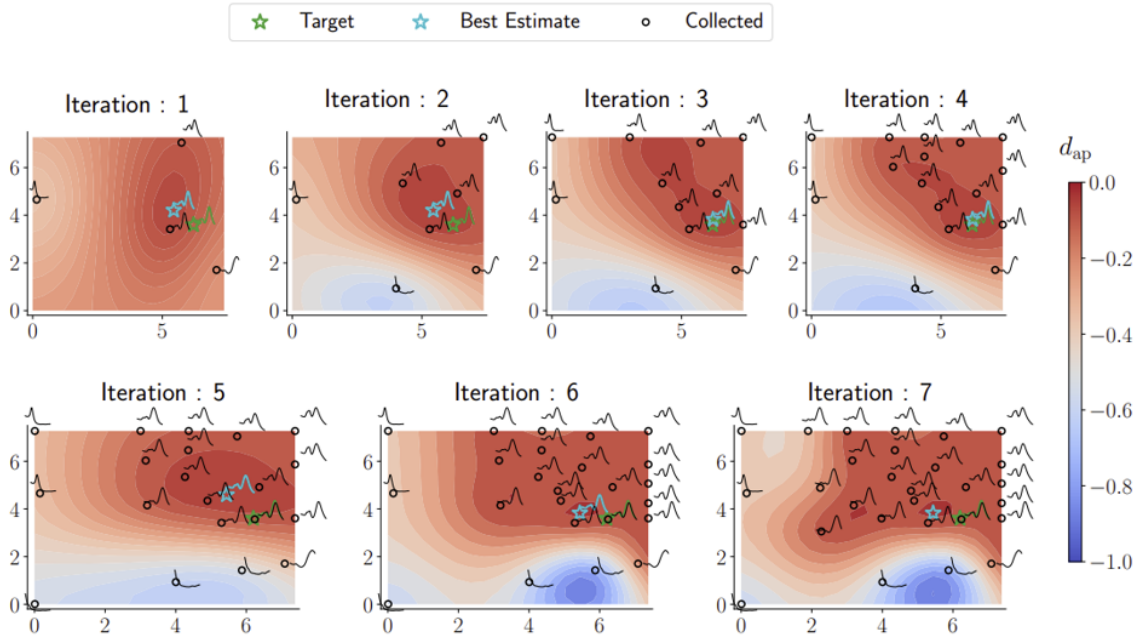



**Figure 3.2:** Optimization trace for a gold nanorod target using the amplitude–phase distance. Each panel shows the surrogate model as a contour plot, data points collected/queried from the experiment in circles, the current best estimate using an aqua-colored star, and the retrosynthesis target using a green-colored star. The x-axis of each plot represents the concentration of silver nitrate $(M \times 10^{-5})$ and the y-axis represents the concentration of ascorbic acid $(M \times 10^{-4})$. All the compositions are annotated with the respective spectra obtained from the experiment. We observe gradual changes to the surrogate approximation with an increase in data collected and the optimization mainly focuses on improving the region with a lot of "target-like" spectra.

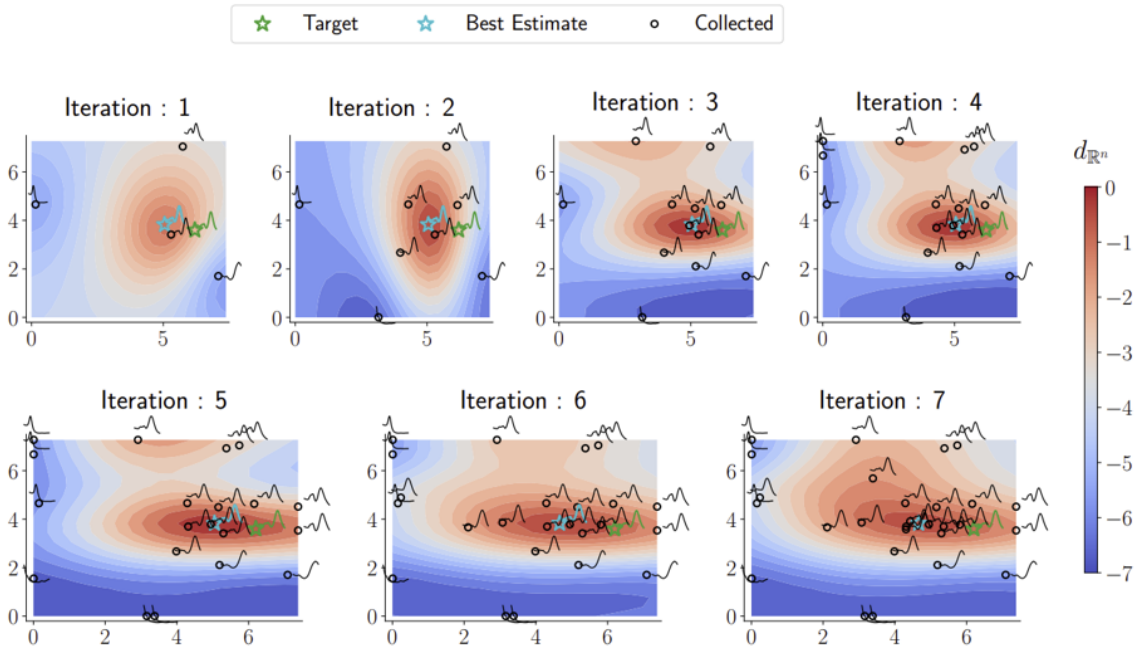



**Figure 3.3:** Optimization trace for a gold nanorod target using a Euclidean distance similar to Figure 3.2

From the phasemaps in Figure 3.2 and Figure 3.3, there are clear differences in the way that the AI agent sampled the design space, as seen in the contours. In Figure 3.2, the AI agent using the Amplitude-phase distance sampled more dispersedly compared to the one using the Euclidean distance metric in Figure 3.3. It also seems that the AI agent with the Amplitude-phase metric quickly learned how to synthesize nanorods, and explored the design space where nanorods are most likely to be found. In contrast to this, the AI agent with the Euclidean distance also quickly learned how to create samples close to the targeted one, but then mostly sampled close to this target, favoring exploitation over exploration. The final contours in iteration 7 of both retrosynthesis campaigns differ significantly. The one using the Amplitude-phase metric resembles the "ground truth" shown in Figure 3.4, where there are clear boundaries between the regions where nanorods, nanospheres, and no structures are located. In addition, the AI agent was able to distinguish between nanospheres and samples with no structures, as seen by the different distance scores assigned to them. Conversely, the AI agent using the Euclidean distance had a significantly different contour plot in iteration 7. It seems to assign a higher distance as the samples are further away from the target, and because of this, it was not able to distinguish the different structures being formed in the design space. In summary, both optimization campaigns were able to synthesize a sample with the correct concentrations as the targeted

sample, so the distance metric had minimal effect on the final result of the optimization campaign. The main difference lies in the construction of the surrogate model of the AI agent, which affects how the AI agent samples the design space.

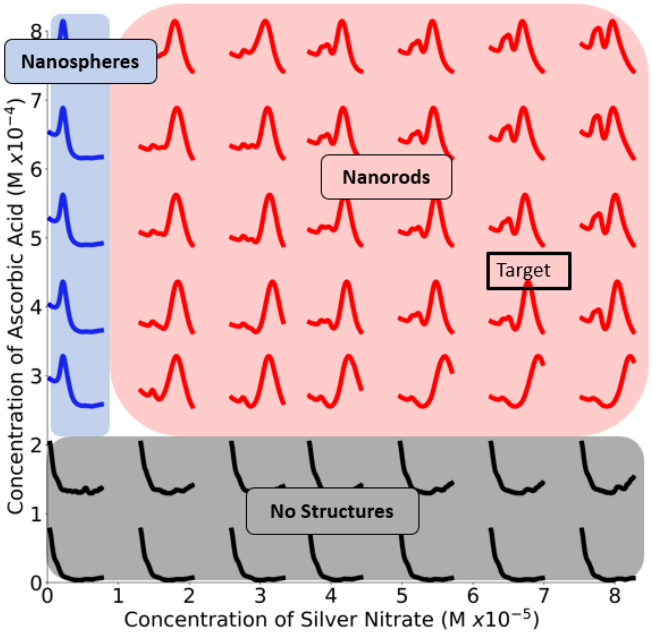



**Figure 3.4:** Spectra obtained from a coarse grid sampling of the two-dimensional design space. Observe that the space is continuous in terms of nano-structural geometries with three broad classes: no nano-structures (black), nanospheres (blue), and nanorods (red). The retrosynthesis target spectrum is labeled.

### 3.3.2 Case Study 2: 8-Dimensional Optimization

In Case Study 1, a 2-Dimensional optimization campaign was performed with a target chosen from the design space, and the campaigns using the Amplitude-phase and Euclidean distance were able to achieve the goal of synthesizing the targeted sample. In Case Study 2, the design space was expanded to include 8 dimensions to test for differences between the two optimization campaigns. Hydrochloric acid and sodium hydroxide were added to control the pH of the synthesis and sodium chloride was added to control the presence of counter ions. It was hypothesized that the addition of these reagents would allow for greater control over the morphology of the nanoparticles [5].

Table 3.2 shows the stock solution concentrations and the concentration range of each sample. The optimization algorithm was free to choose any concentration in this range without any constraints. The

**Table 3.2:** Table II. Concentrations of the Chemical Design Space and the Nanorod Target

| Reagent | Stock Solution Concentration (M) | Concentration Range (M) |
|---|---|---|
| CTAB | $2.0 \times 10^{-1}$ | $0 - 7.50 \times 10^{-2}$ |
| Gold (III) Chloride Trihydrate | $1.0 \times 10^{-3}$ | $0 - 1.50 \times 10^{-4}$ |
| Silver Nitrate | $6.4 \times 10^{-4}$ | $0 - 6.00 \times 10^{-5}$ |
| Ascorbic Acid | $6.3 \times 10^{-3}$ | $0 - 6.40 \times 10^{-4}$ |
| Gold Seeds | $2.0 \times 10^{-5}$ | $0 - 6.00 \times 10^{-5}$ |
| Hydrochloric Acid | $2.0 \times 10^{-1}$ | $0 - 1.40 \times 10^{-2}$ |
| Sodium Hydroxide | $1.0 \times 10^{-1}$ | $0 - 7.20 \times 10^{-3}$ |
| Sodium Chloride | $2.0 \times 10^{-1}$ | $0 - 1.40 \times 10^{-2}$ |

target of the optimization campaign was arbitrarily chosen to be gold nanorods of 55 nm in length and 10 nm in diameter. The extinction spectrum of these gold nanorods was numerically simulated and used as the target for the experiments.

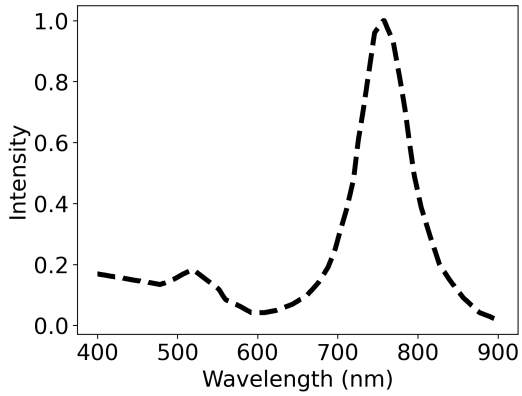



**Figure 3.5:** The simulated extinction spectrum of a gold nanorod of 55 nm in length and 10 nm in diameter. This was used as the target spectrum of the 8-dimensional optimization campaigns.

Due to the high dimensional space, it was not possible to visualize the changes in the surrogate model over the iterations like in Case Study 1. To visualize the differences in performance of the campaigns that used the Euclidean distance metric compared to the Amplitude-phase, the spectra of all the samples in iterations 0, 1, 2, and 6 were plotted and shown in Figure 3.6. From that figure, it is clear that there are major differences in the samples over the iterations. In iteration 1, of the campaign, using the Amplitude Phase metric, all of the spectra have two peaks that resemble the spectrum of a nanorod. This indicates that

with only one interaction, the AI agent already identified how to synthesize nanorods. Nanorod spectra are mostly observed in iterations 2 and 6 which shows that the AI agent continued to synthesize these nanorods.

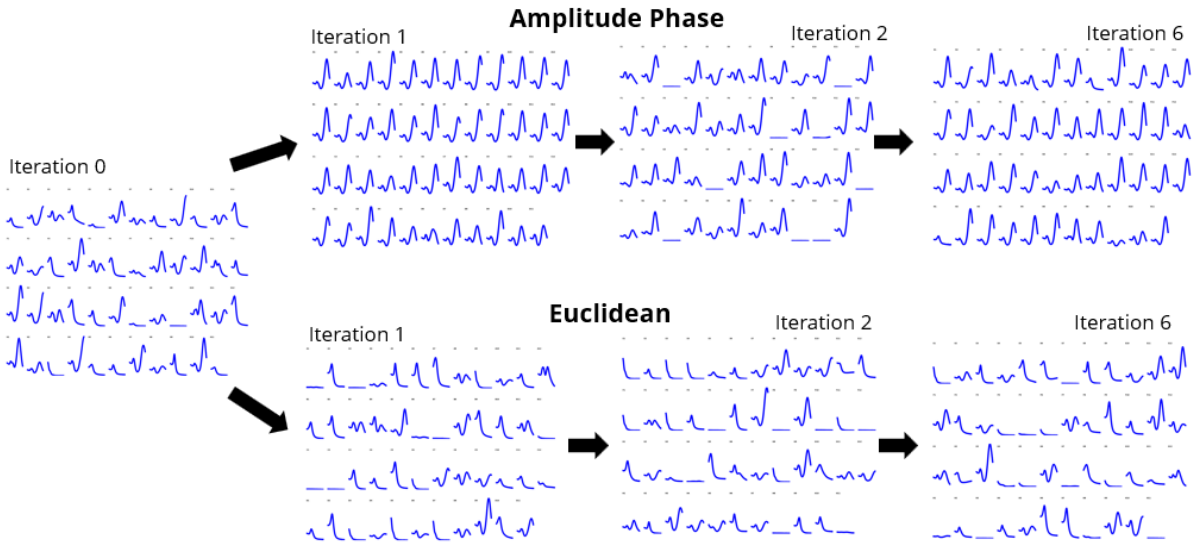



**Figure 3.6:** The results of the 8-dimensional optimization campaigns. The spectra of iterations 0, 1, 2, and 6 are shown for the campaigns using the Amplitude Phase metric (top) and the Euclidean metric (bottom). The two campaigns start with the same samples which is why they share the samples from iteration 0.

In contrast, the AI agent using the Euclidean distance metric failed to synthesize a large number of nanorods. In iteration 1, many spectra with a single peak that resemble nanospheres or flat spectra that suggest no reaction had occurred, are common. This trend continued until iteration 6, where there was still no evidence that the algorithm learned how to synthesize nanorods. To further compare the performance of the optimization campaigns, the spectra of each campaign that most closely matched the targeted nanorod spectra were plotted in Figure 3.7. In this figure, it is clear that the campaign that used the Amplitude Phase metric synthesized a sample that more closely resembled the targeted spectrum. While the campaign that used the Euclidean distance metric synthesized a nanorod, it had some inconsistencies with the chosen target.

In Case Study 2, the 8-dimensional optimization successfully synthesized a sample whose UV-Vis spectrum showed good agreement with the targeted spectrum, as seen in the Amplitude Phase Best Sample whose spectrum is shown in Figure 3.7. To verify if the best sample actually consisted of nanoparticles with the targeted morphology of gold nanorods of 55 nm in length and 10 nm in diameter, SAXS and TEM were

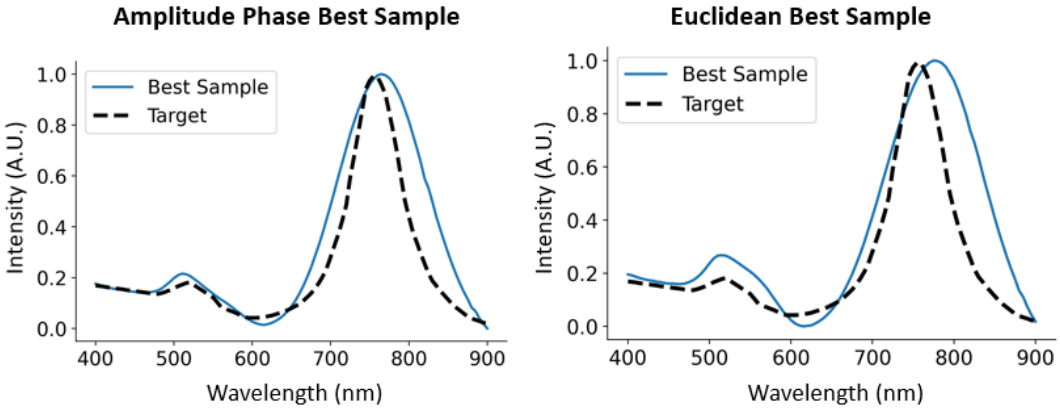



**Figure 3.7:** The normalized spectra that most closely matched the targeted nanorod spectra of each optimization campaign.

performed on the sample.

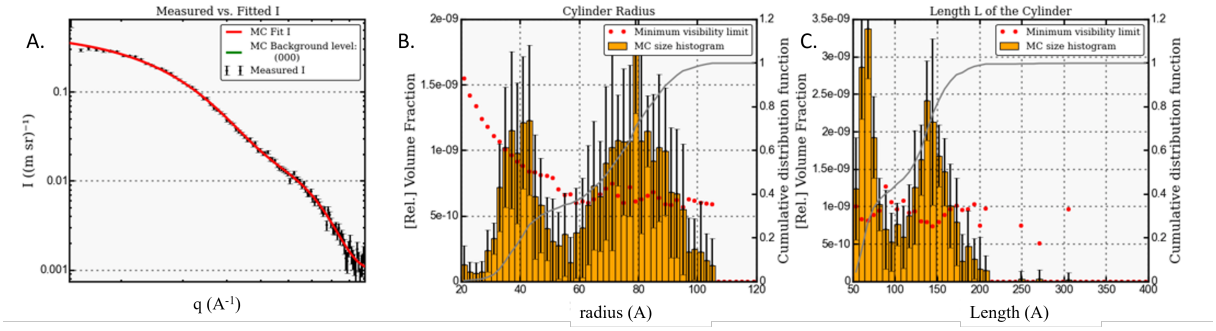



**Figure 3.8:** The SAXS scattering curve of the Amplitude Phase Best Sample whose UV-Vis spectrum is shown in Figure 3.7. A. shows the experimental scattering curve together with the cylinder model fit. B. shows the distribution of cylinder radius calculated from the cylinder model. C. shows the distribution in cylinder length from the cylinder model.

The SAXS scattering curve was fit using a cylindrical model, and the distributions of the cylinder radius and length were obtained. The distribution of cylinder radius shown in Figure 3.8 B. seems to be bimodal with large populations of radii of 4 and 8 nanometers. The distribution of cylinder length shown in Figure 3.8 C., is also bimodal with large populations that contain lengths of 7 and 15 nanometers. The information from SAXS suggests that the Amplitude Phase Best Sample is polydisperse and does not contain the targeted structure of monodisperse nanorods with 55 nm in length and 10 in radius. To further validate the information from SAXS, TEM was taken from the same sample.

The TEM images show that the Amplitude Phase Best Sample contains both nanorods and nanospheres,

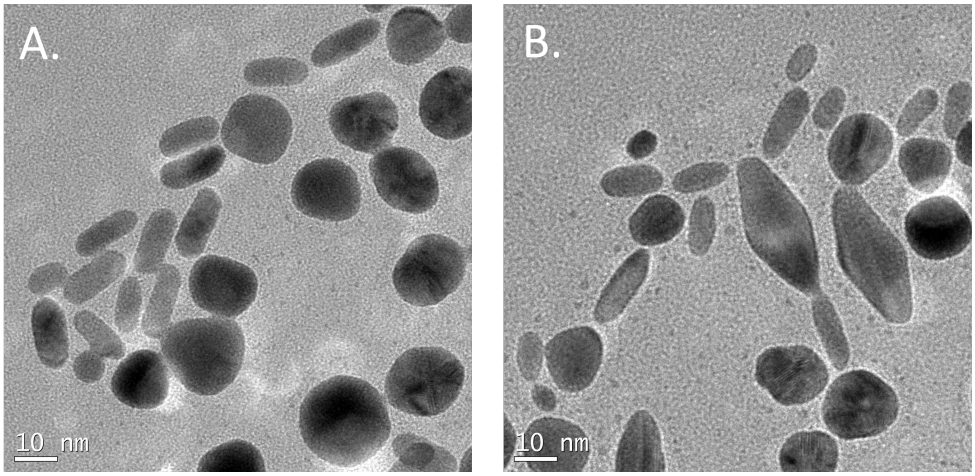



**Figure 3.9:** TEM images of the Amplitude Phase Best Sample whose UV-Vis spectrum is shown in Figure 3.7 and SAXS scattering curve is shown in Figure 3.8 A.

which already is enough evidence that the Amplitude Phase Best Sample does not contain the targeted structure of monodisperse nanorods. In addition, the nanorods in the TEM images in fig. 3.9 seem to be much smaller than the target ones. Evidence from both SAXS and TEM suggests that although the Amplitude Phase Best Sample had a good UV-Vis Spectra match with the targeted spectrum, the actual structure of the sample did not match the targeted one. One hypothesis for this result could be that the exclusive use of UV-Vis spectroscopy as a proxy for the nanoparticle's structure resulted in the failure of our optimization campaign to achieve the desired structure of the nanoparticle. To successfully achieve the desired nanoparticle structure, it is could be advantageous to use a combination of characterization methods instead of relying on a single method. For instance, the use of UV-Vis spectroscopy and SAXS for nanoparticle characterization could yield better results. This is because both techniques are degenerate, meaning that multiple structures can have the same spectrum or scattering curve, but it is highly unlikely for a structure to have the same spectrum and scattering curve at the same time. We hypothesize that future optimization campaigns would benefit from multimodal characterization methods.

### 3.3.3 Discussion on Closed Loop Optimization Campaigns

While performing the closed-loop optimization campaign in case study 2, several limitations were observed. The first limitation, which has already been discussed in the previous section, is the degeneracy of UV-Vis

spectroscopy leading to the failure of the campaign to achieve the desired structure. To mitigate this problem, future campaigns should include multimodal characterization methods to obtain more detailed information on the material's structure.

Another limitation is the lack of knowledge gained from the optimization campaign, such as the effect of each reagent on the material's structure. Because the AI agent completely controls the campaign, no knowledge is extracted from the experiment. This can be problematic when the optimization campaign fails to achieve a good match with the targeted UV-Vis spectrum, which was a common occurrence experienced firsthand during the development of this closed-loop optimization campaign. By understanding how each reagent affects the final material structure, it becomes easier to troubleshoot and fix errors, such as by expanding or contracting the design space to achieve a higher chance of success.

The third limitation is the use of extensive prior knowledge to determine the design space. In the optimization campaign that was performed in this chapter, and in many other ones from literature [29] [40], the synthesis of gold/silver nanoparticles was used as a model system. Due to extensive prior research and ease of execution, the synthesis of gold and silver nanoparticles is a popular choice. Because of this, it is common for the design space, defined by the stock solution concentrations and the concentration range of each sample, to be chosen based on values extracted from the literature, which is in direct conflict with the ultimate goal of applying closed-loop optimization on novel material systems. An ideal optimization campaign would not need to rely on extensive prior knowledge for it to properly explore the material system of interest.

## 3.4 Conclusion

In this chapter, the performance of optimization algorithms using the Amplitude Phase and Euclidean distance metric were compared. In Case Study 1, a 2-dimensional design space was chosen with a target that was known to be in the design space. The results show that the AI agent using the Amplitude Phase distance was able to identify the different regions in the design space where different structures were formed. In Case Study 2, the effect of the distance metric on an 8-dimensional optimization with a simulated nanorod target was performed. The results show that the AI agent using the Amplitude Phase metic quickly learned how to synthesize nanorods, continued to synthesize them throughout the iterations, and eventually had a

candidate spectrum that closely matched the targeted one. Despite this success, three limitations of closed-loop optimization campaigns were discussed. The first was the need for the implementation of multiple characterization methods to overcome the degeneracy of UV-Vis Spectroscopy. The second was the need for an interpretable optimization campaign to extract knowledge from the material system. Finally, the third was the need to reduce the reliance on the literature to create the design space. In the next chapter, a novel data-driven exploration method will be introduced that addresses all the previously identified challenges.

# Chapter 4

# Data-Driven Exploration of Silver Nanoplate Formation in Multidimensional Chemical Design Spaces

This work presented in this chapter is from:

- Huat Thart Chiang, Kiran Vaddi, Lilo D Pozzo, Data-driven exploration of silver nanoplate formation in multidimensional chemical design spaces, *Digital Discovery*, 2024, 3, 2252-2264. [7]

## 4.1 Abstract

We present an autonomous data-driven framework that iteratively explores the experimental design space of silver nanoparticle synthesis to obtain control over the formation of a desired morphology and size. The objective of the method is to identify design rules such as the effects of the design variables on the structure of the nanoparticle. The framework balances multimodal characterization methods (i.e. UV-Vis spectroscopy, SAXS, TEM), taking into account the cost of performing a measurement and the quality of information gained. By integrating with an AI agent, we identify important design variables in the synthesis of small colloidally stable plate-like silver particles and outline how each variable affects plate thickness, radius, polydispersity, and relative concentration. Our findings are consistent with the literature, demonstrating that

the framework could be further applied to new systems that have not been well characterized and understood. The framework is generalizable and allows tangible knowledge extraction from the high-throughput experimental runs while still considering inherent stochasticity.

## 4.2 Introduction

Silver nanoparticles have shown to be extremely useful for purposes such as catalysis[41], therapeutics[42], drug delivery[43], and surface-enhanced Raman spectroscopy (SERS)[44]. It is also well known that the optical properties of silver nanoparticles depend on their shape and size[45], thus the ability to synthesize particles of a specific structure is highly desirable. The literature on the synthesis of silver nanoparticles is vast including physical, photochemical, and chemical methods[46]. Still, it is often difficult to obtain control over their shape and size due to a limited understanding of the processes that affect the final structure. For example, nucleation, growth, aggregation, and Ostwald ripening, are often affected by both thermodynamic and kinetic parameters associated with reaction conditions[47]. Because of this, the experimental design space used to synthesize these nanoparticles is often large and complex[48]. In the chemical synthesis of silver nanoparticles, factors such as light, temperature, age of stock solutions, and duration of the reaction affect the nanoparticle's structure. Finally, due to the limited understanding of the processes that affect the final structure, the relationship between the experimental design parameters (e.g, the concentration of reagents, temperature) and the final structure is often determined by trial and error, which is time-consuming and laborious[29].

Large and multidimensional experimental design spaces are common in material synthesis. Recently, the combination of artificial intelligence, automation, and a characterization method has emerged as a powerful method to achieve control over the results of nanoparticle synthesis [39]. Automation can be used to synthesize nanoparticles in high-throughput and then screen for a desired structure or property using a characterization method, which facilitates the collection of large datasets. While it is often difficult for a human to interpret the generated large datasets, artificial intelligence algorithms have been successfully used to make decisions for experimental design[39][49], build predictive models[50], and extract knowledge via model interpretation[51]. One common application of artificial intelligence and automation is the concept of a closed-loop design or "retrosynthesis", where optimization algorithms (e.g., bayesian

optimization[39] or genetic algorithms[40]) work together with robots (e.g., liquid handling robots[39] or microfluidic devices[52]) and one or more characterization methods (e.g., UV-Vis Spectroscopy) to iteratively discover the reagent compositions and conditions that yield a desired material or structure.

Many attempts to demonstrate this concept with inorganic nanoparticle synthesis have been successful [39][38][53]. Despite these successes, there are several limitations to closed-loop systems. For example, in most of the above-mentioned frameworks, the experimental design space, which is usually specified by variations in reagent concentrations, is chosen based on values extracted from the literature. This is in direct conflict with the ultimate goal of the closed-loop approach which is to accelerate the discovery of novel materials. Thus, it is unlikely that existing literature would contain information on the design parameters for the completely new materials we aim to discover. Another limitation is the lack of interpretability of the data that is generated from the experiment. Since an optimization algorithm drives the experiment and causes the data to be biased to samples that are close to the target, almost no information on the effect of the experimental design parameters on the structure or property of the material (i.e. outcomes) is obtained that is interpretable by humans. Understanding these relationships is important to obtain a better understanding of the reaction mechanisms that occur during the experiment, which can be useful for modifying the design space, optimizing material structure, or extracting higher-level knowledge that can then be applied to other problems.

In addition to closed-loop systems, another method to study large and complex design spaces is to combine design of experiments (DOE) concepts and high throughput experimentation to perform systematic studies of large design spaces. Some sampling methods include full factorial designs and Latin hypercube sampling, which can increase interpretability when combined with data science and artificial intelligence methods[54]. However, an important disadvantage of such systematic studies is that they do not scale well in high-dimensional experimental design spaces which is defined in terms of the number of tunable design variables. As the number of variables or dimensions gets larger, the number of experiments that need to be performed to sufficiently explore the design space increases drastically[55]. In addition, just like in closed-loop systems, the design space used in systematic studies needs to be chosen very carefully to maximize the amount of information gained from the experiment, which is again difficult to implement in completely new and unknown material systems.

The objective of this work is to demonstrate a hierarchical data-driven framework that can efficiently explore large multidimensional experimental design spaces of material synthesis to converge onto targeted outcomes. This framework combines the advantages of a fully automated closed-loop system, by sampling iteratively, and systematic studies (similar to traditional DOE), by being interpretable. To demonstrate our framework, we chose the synthesis of silver nanoparticles, where we aim to identify regions where nanoparticles of a specific shape (i.e., small plate-like particles) can be formed with greater accuracy and control. In addition, after identifying particles of the desired shape, we seek to extract information on the relationship between the design parameters and structural features of the nanoparticles such as feature size and polydispersity. In this work, we leverage the power of a liquid-handling robot to perform the synthesis of silver nanoparticles in high-throughput sampling campaigns. For the characterization of samples, we use a hierarchical analysis campaign starting with UV-Vis spectroscopy as a fast and inexpensive proxy for the nanoparticle's structure, small-angle x-ray scattering as a more expensive but direct characterization method, and transmission electron microscopy (TEM) as the most expensive but most information-rich characterization method. Due to the compromise between costs and structural detail that is gained from each of these complementary methods, we also seek to apply a hierarchical experimental design to maximize the value of the information that is obtained while minimizing the total costs. Our data-driven exploration starts with UV-Vis spectroscopy, as the fastest and least expensive technique, to infer the shape of nanoparticles based on their plasmonic resonance. We analyze vast amounts of spectra collected over a very large design space and use a distance metric to determine which samples are small, colloidally stable, monodisperse, plate-like particles. The data is then used to train a Gaussian process classifier which is used to iteratively explore regions of the design space that allow us to synthesize nanoparticles of the targeted morphology (i.e. silver nanoplates). Once we constraint the design space to primarily form particles with the target shape, we perform small angle x-ray scattering (SAXS) on these samples to obtain quantitative information on the size of features (i.e. radius and thickness) as well as the polydispersity and relative concentration of particles. We then use transmission electron microscopy, which is the most expensive and time-intensive technique, to verify models and to help validate the SAXS data. Finally, interpretable design rules are extracted from the aggregate data to identify the effect of design parameters on the structural features of 2D silver nanoparticles.

## 4.3 Materials and Methods

### 4.3.1 Materials

Silver nanoparticles were synthesized using polyvinylpyrrolidone (PVP) 40 kDa, tannic acid, ascorbic acid ($\geq 99\%$), silver nitrate ($\geq 99\%$), sodium borohydride ($\geq 98\%$), and methyl cellulose (4000 cP). All chemicals were purchased from Sigma Aldrich (St. Louis, MO, USA) and used as received. Deionized water was used in all syntheses from a Direct-Q 3 UV water purification system with a resistivity of 18.2 M $\Omega$ (Millipore Corporation, Bedford, MA, USA)

### 4.3.2 Silver Nanoparticle Synthesis

We used a procedure similar to that described in Samanta et al.[56] to synthesize nanoparticles. The first step was to synthesize silver seeds. To do this, 0.50 mL of a 10 mM silver nitrate solution and 4.5 mL of water were added to 15 mL of 9.35 mM ice-cold methyl cellulose solution. Under rigorous stirring, 0.050 mL of 10 mM sodium borohydride solution was added and the color of the solution immediately turned dark yellow. The seeds were left for 2 hours under stirring before use. All stock solutions were created in 20 mL scintillation vials. The synthesis of silver nanoparticles was performed in clear 96-well polystyrene microplates (Corning, NY, USA) with a maximum well volume of 350 uL at approximately 22 degrees Celsius. Water was added to each sample so a total volume of 325 uL could be achieved. The order in which the reagents were added is as follows: PVP, water, tannic acid, ascorbic acid, silver nitrate, and silver seeds. Samples were made simultaneously in batches of 48 samples, meaning that a given reagent was added to every well, in the specified volumes, before changing the pipette tip and performing the same step with the next reagent. After the addition of tannic acid, ascorbic acid, silver nitrate, and silver seeds, a mixing step was performed in each well by repeatedly aspirating and dispensing 100 uL of sample three times. The pipette tip was then washed by performing this same mixing step in three different reservoirs of deionized water to avoid cross-contamination of the stock solutions and of the samples. All pipetting was performed by an Opentrons (Brooklyn, NY, USA) OT2 liquid handling robot. The OT2 control code can be found at (`https://github.com/pozzo-research-group/papers/tree/main/Silver%20Nanoplates`) to perform the pipetting commands.

### 4.3.3 UV-Vis Spectroscopy

Samples were characterized using an Epoch 2 microplate spectrophotometer (BioTek, Winooski, VT, USA) from 350 nm to 800 nm in increments of 5 nm. Background subtraction was performed by subtracting the optical extinction of a water sample with the same volume (325 uL). UV-vis spectroscopy was performed 24 hours after the synthesis of the silver nanoparticles to allow enough time for the complete growth of the particles, which continued to evolve over several hours after the reagents were added.

### 4.3.4 Small Angle X-ray Scattering

SAXS was performed on a Xenocs Xeuss 3.0 (Grenoble, France) instrument with an x-ray energy of 8.04 keV (wavelength 1.54 Å) using a copper K-$\alpha$ microfocus source. Data was collected in three configurations: low-q (0.003 - 0.007 $Å^{-1}$) for 7 minutes, mid-q (0.007 - 0.020 $Å^{-1}$) for 3 minutes, and high-q (0.020 - 0.100 $Å^{-1}$) for 2 minutes. Samples were autonomously loaded into a 2 mm diameter quartz capillary flow cell using the "Biocube" sample environment and the robotic loading capabilities of the Xeuss instrument. After each measurement, the capillary was autonomously flushed with water for 15 seconds and dried for 50 seconds with compressed air using the robotic arm. It was discovered that this cleaning protocol prevented fouling in the capillary after changing samples. Background reduction was performed with the XSCAT software by subtracting the scattering of water in the same capillary flow cell and the data was merged automatically with code from Python. The models used to fit the data were implemented with Sasmodels (`www.github.com/SasView/sasmodels`). Code to autonomously merge data and fit SAXS data can be found in the online repository (`https://github.com/pozzo-research-group/papers/tree/main/Silver%20Nanoplates`).

### 4.3.5 Transmission Electron Microscopy

Transmission electron microscopy (TEM) samples were prepared by casting 10 uL over a carbon-coated film, 200 mesh, copper grids which were purchased from Electron Microscopy Sciences (Hatfield, PA, USA). No centrifugation or other sample preparation was performed. The grids were then imaged on an FEI Tecnai F20 at 200 kV. Image analysis was performed by individually measuring all the plate diameters in the TEM images with the ImageJ software [57].

### 4.3.6 Scanning Electron Microscopy

Scanning electron microscopy (SEM) was performed using an Apreo-S instrument (Thermo Fisher, Waltham, MA, USA). Samples were first diluted 100 times and then 5 $\mu$L aliquots were deposited on clean silicon wafers.

### 4.3.7 Machine Learning and Data Analysis

Machine learning algorithms were used for several purposes. A Gaussian process classifier was used to explore the nanoparticle morphologies in the experimental design space. A Gaussian process regressor was used to create contour plots. Both algorithms were implemented in Python with the sci-kit learn library. A radial basis function kernel was selected. Code to reproduce the algorithms can be found in the online repository (`https://github.com/pozzo-research-group/papers/tree/main/Silver%20Nanoplates`).

## 4.4 Results and Discussion

For the data-driven exploration, we used three different characterization techniques that have different costs, throughput, and information content. UV-Vis spectroscopy was used in the Fast Spectroscopic Exploration section to determine the experimental design parameters that give the highest probability of forming small, colloidally stable, monodisperse, plate-like particles. SAXS was then used in the SAXS Structural Exploration section to determine how design parameters affect the thickness, radius, polydispersity, and scale, a parameter proportional to the concentration of the plate-like particles. Finally, transmission electron microscopy (TEM), which is the most expensive and intensive of the characterization tools, was used on select samples to validate observations via direct imaging. In each section, we provide a description of how each characterization method is used followed by a discussion of the results.

### 4.4.1 Fast Spectroscopic Exploration

Our first objective is to narrow down the design space to find regions where small, colloidally stable, monodisperse, plate-like structures are formed in high purity. This first step is advantageous because it

enables starting experiments from an arbitrarily large design space, as intended for the application of these systems in the discovery of novel materials. The viable design space, which is bounded or limited by the maximum concentrations of stock solutions and the maximum/minimum volumes that can be transferred by the robotic tool, is shown in Table 4.1.

**Table 4.1:** Experimental design parameters used to synthesize silver nanoparticles. The volumes of each regent can be independently varied from 0-60 $\mu$L in increments of 1 $\mu$L. Water was added to each sample to obtain a volume of 325 $\mu$L. The lowest concentration achieved is 0.09 $\mu$M for the silver seeds and 0.006 mM for the other reagents. Reagent concentrations are reported assuming a total volume of 325 $\mu$L.

| Reagent | Stock Concentration (mM) | Volume Range (uL) | Concentration Range (mM) |
|---|---|---|---|
| PVP | 2.00 | 0-60 | 0.000-0.370 |
| Tannic Acid | 2.00 | 0-60 | 0.000-0.370 |
| Ascorbic Acid | 2.00 | 0-60 | 0.000-0.370 |
| Silver Nitrate | 2.00 | 0-60 | 0.000-0.370 |
| Silver Seeds | 0.03 | 0-60 | 0.000-0.001 |

In order to maximize the chances of obtaining a diverse set of morphologies and sizes, several reducing agents and stabilizers were selected based on common silver nanoparticle synthesis reagents. Instead of relying on the literature for stock solution concentrations, a relatively high, but reasonable, concentration of 2 mM was chosen for all reagents except for silver seeds because the method used for the seed synthesis could not achieve such a high concentration. A high stock solution concentration ensures that a large design space can be searched by the algorithm.

An experiment was also performed to determine the effect of the order and time of reagent addition on the final nanoparticle's structure (Supporting Information S20). In summary, it was found that the order of addition had a major effect on the final structure, but the time of addition only affected certain sample compositions when the order of addition was: silver nitrate, seeds, ascorbic acid, tannic acid, PVP. Experiments evaluating the order and/or time of reagent addition can be easily evaluated and explored using robotic systems but this is otherwise very challenging to explore due to the manual labor involved. Data-driven workflows driven by robotic tools enable effective analysis of colloidal systems where timed-intervention affects the final structure[33]. Despite this, we decided to constrain our experimental design space to exclusively explore the concentrations of the reagents.

UV-Vis spectroscopy was chosen as the primary screening step because of its low cost and high-throughput capabilities to characterize plasmonic samples. It is well known that this method can be used to indirectly determine a metallic nanoparticle's structure by measuring its localized surface plasmon resonance (LSPR), which results in the local enhancement of an oscillating electric field due to nanoscale confinements. In a UV-vis spectrum, this is usually the wavelength where the peak extinction occurs. Using LSPR to determine the size of a nanoparticle must be done cautiously since the LSPR is also affected by other factors such as particle assembly/aggregation, the solvent or medium, the orientation of the electric field, impurities, etc [45], which makes it difficult to attribute variations of the LSPR exclusively to particle size. The shape of a UV-Vis spectrum can also be used to determine the nanoparticle's shape. While suffering from similar limitations (e.g. sensitivity to aggregation), the shape of a UV-Vis spectrum has significant variation when used to differentiate between nanoparticle shapes (i.e. spheres, rods or plates). For example, the spectrum of a nanorod has transverse and longitudinal peaks, a sphere has a single peak, and an aggregate of particles has a very broad peak[58]. In addition to particle shape, UV-Vis spectroscopy can identify conditions that do not lead to any reaction or nanoparticle formation. For example, the spectrum of a sample where no nanoparticles are formed will have low extinction, similar to that of water. It can be challenging to determine which experimental parameters are most effective in producing nanoparticles when working with a large design space. Therefore, the shape of the UV-Vis spectra was used as a fast and effective screening method to identify regions leading to nanoparticle morphologies of interest. This approach proved helpful in identifying the most promising candidates for further study.

The analysis of functional data such as UV-Vis spectroscopy curves is challenging due to the high-dimensional space. Scalar-value features such as peak position and peak width are frequently used to characterize nanoparticle shapes; however, in an experimental spectrum of real samples, the shape of the spectrum represents an ensemble average of all individual nanoparticle signals. Thus, polydispersity, which is a measure of particle size distributions and variations in shape, can change the width of the primary peak in a spectrum. To account for this information and provide a generic method that is agnostic to prior knowledge about observing all the different features of the synthesized morphology, we use information in the form of shape rather than selecting a few scalar features that narrowly represent the full shape of the spectrum.

In our experiment, we expected that we would synthesize hundreds of samples that would later have to be autonomously classified as small, colloidally stable, monodisperse, plate-like particles or not. To carry out this task, a distance metric was established that used the UV-Vis spectroscopy curve of each sample. Before creating this distance metric, the targeted UV-Vis spectroscopy curve of small, colloidally stable, monodisperse, plate-like particles was simulated using the discrete dipole approximation for the scattering (nanoDDSCAT)[59] with a refractive index of 1.33 corresponding to water. Because we were concerned with initially classifying particles based on shape, we simulated the spectra of several small plates of different radii and used them as targets. In addition to nanoplates, the formation of spherical silver nanoparticles is also likely using the experimental design parameters[42]. Therefore, the extinction spectrum of silver nanospheres of different sizes was also simulated to identify the spectrum of these undesired structures. Using data from simulations, a distance metric that rewards spectra that are close to that of plates and penalizes those that are close to spheres was created. A threshold distance was carefully chosen so that spectra with distances lower than this value would be classified as "Below Threshold" and samples with distances greater than the threshold would be classified as "Above Threshold". Since the selection of the distance threshold could vary, we accept that there will be samples that have some of the small, colloidally stable, monodisperse, and/or plate-like characteristics that are cut off by the distance threshold. This is because we expect the targeted characteristics to change gradually within an experimental design space, unlike a binary design space where there is a sudden change between regions with particles that have all the targeted characteristics and regions that do not.

From the simulations shown in Figure 4.1, the simulated spectra of both spheres and plates have a similar shape (i.e. single peak). The main difference is the wavelength where the peak intensity is located. The position of the peaks of spheres of 20 to 40 nm in diameter varies from about 390 nm to 410 nm, while that of the plates of the same size varies from 480 nm to 590 nm. These observations are consistent with the explanation that anisotropic morphologies have longer plasmon lengths which in turn shifts its spectrum's peak position to higher wavelengths[60]. Using these observations, a distance metric was created to classify UV-Vis spectra. This distance metric is composed of four terms shown in equation (4.1). Each distance term was scaled with a coefficient so that all the terms had values that ranged within the same order of magnitude. This was to ensure that the contributions of each term to the final distance metric were comparable. The

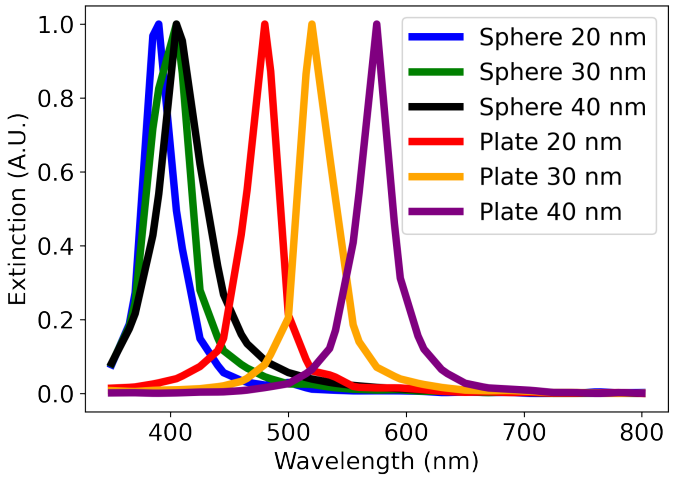



**Figure 4.1:** The simulated spectra of spheres and plates using nanoDDSCAT. The legend refers to the diameter of simulated the plates and spheres. All simulated plates have a thickness of 7 nm.

effect of each coefficient on the final distance metric was also investigated and this is demonstrated in the Supporting Information (S4).

$$d = (7 \times d_{AP}) + (2 \times d_{peak}) + (85 \times d_{area}) + (1 \times d_{intesnity}) \tag{4.1}$$

- $d_{AP}$ refers to the Amplitude Phase distance as described in Vaddi et al.[39]. This distance primarily accounts for the shape of the spectra by decomposing the required distance into x-component (phase) and y-component (amplitude). The amplitude and phase components are then computed by an iterative two-step optimization procedure to compute the amount of x and y-component distance between the query and target spectra. This approach represents spectra as points in a high-dimensional function space. The amplitude distance is defined using a function norm between the query and target after separating the phase component. The phase distance is calculated by measuring the distance between the identity phase component (which means no change along the x-axis of the query spectrum) and the phase component of the query spectrum that is needed to derive it from the target spectrum. We refer interested readers to the supporting information or the original paper and references therein for more details.

- $d_{peak}$ refers to the wavelength of the peak intensity. Since the simulated spectra of spheres and plates in Figure 4.1 have similar shapes, this distance term penalizes spectra that have a peak position lower

than 450nm, which are most likely to belong to spheres. This term adds a penalty of 100 to the total distance score if the primary peak's wavelength is lower than 450nm.

- $d_{area}$ refers to the area under the curve, which encourages monodispersity. This facilitates the interpretation of data when determining how the reagents affect the size of the plates. This term is determined by integrating the UV-Vis spectrum and adding this value to the distance score.

- $d_{intensity}$ refers to the intensity below 450nm. Since the simulated spectra of plates have low intensity at these wavelengths, this term was added to help in the classification of plates. This term is determined by finding the maximum intensity of the UV-Vis spectrum at wavelengths less than 450nm and adding this value to the distance score.

The distance metric was used as the primary reward function for the workflow presented in Figure 4.2. In the initial iteration, the compositions for 48 samples were randomly selected from the predetermined design space and synthesized using an Opentrons OT2 robot, which took around 3 hours. After one day, UV-Vis spectroscopy characterization was performed, and the distance metric was used to classify each sample. The information on the composition and classification of each sample was then used to train a Gaussian process classifier. Due to the probabilistic nature of a Gaussian process, we could identify regions in the design space where the probability of forming small, colloidally stable, monodisperse plate-like structures was greater than a certain limit, which was chosen to be 90%. It is important to clarify that this 90% limit is a hyperparameter independent from the distance threshold, but both can be used to bias how strongly the algorithm forms the targeted structure. Next, we randomly selected 48 new samples from these high-probability regions to obtain the next iteration of compositions to synthesize. In this closed-loop system, the Gaussian process gets more accurate as data is obtained, which allows it to improve its prediction of where samples with high concentrations of the desired plates are most likely to be formed. It was observed that convergence was achieved after performing a total of six iterations.

**Results of Fast Spectroscopic Exploration**

The primary goal for the interpretation of data in this section was to determine the design rules that the Gaussian process classifier learned to synthesize plate-like particles. Understanding these design rules can

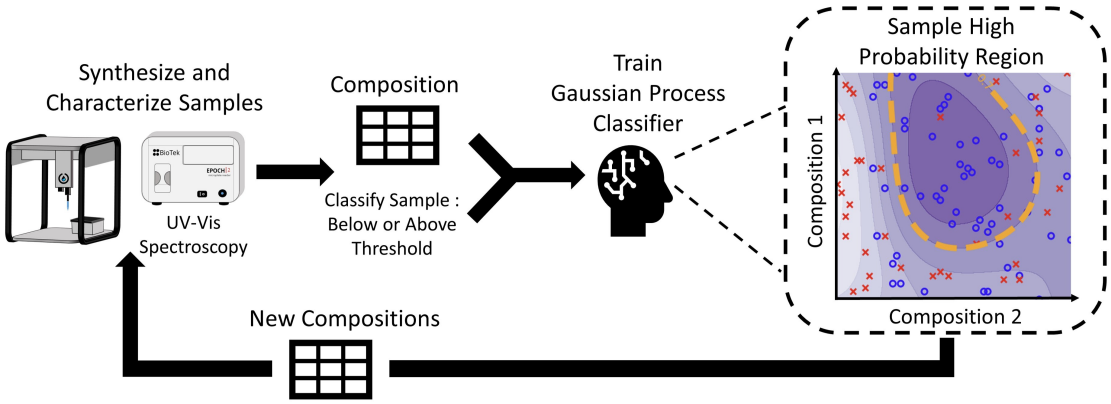



**Figure 4.2:** The workflow for Fast Spectroscopic Exploration. Samples were synthesized with an Opentrons OT-2 liquid handling robot and then characterized using UV-Vis spectroscopy. Each UV-Vis spectrum that was obtained was classified, using the distance metric, into whether it had characteristics of small, colloidally stable, monodisperse, plate-like particles. This information was then used to train a Gaussian process classifier. Using this classifier, the region where desired plates are most likely to be formed was identified, and samples for the next iteration were randomly chosen from this region. This method can be applied in high-dimensional design spaces, but a two-dimensional design space is shown in the figure for visualization purposes.

be helpful when performing future experiments that attempt to synthesize plate-like particles. Figure 4.3 shows the changes in the samples' composition over the iterations and gives insight into how the Gaussian process classifier updated its predictions throughout the experiment. Since we used random sampling, each iteration consists of samples that, collectively, are an unbiased representation of the region in the design space where plates are formed. The ability to randomly sample in different regions of the design space gives rise to the interpretability of our data-driven exploration and distinguishes it from other closed-loop retrosynthesis systems.

In Figure 4.3, we attempt to visualize the multivariate effect of the reagents on the shape of the nanoparticle by representing the volume fractions of the 5 reagents using multidimensional scaling plots, adapted from Li et al.[61]. Iteration 0 in Figure 4.3 was the first iteration, meaning that it was sampled randomly without any input from the Gaussian process classifier. Because of this, the samples were dispersed, and a high number of samples that had distances above the threshold were synthesized. In iteration 1, the Gaussian process classifier influenced the sampling by sampling in the region with low-volume fractions of ascorbic acid. However, there were still many samples that had distances above the threshold. In iteration 2, the sampling was biased towards low-volume fractions of ascorbic acid and tannic acid, moderate-volume fractions

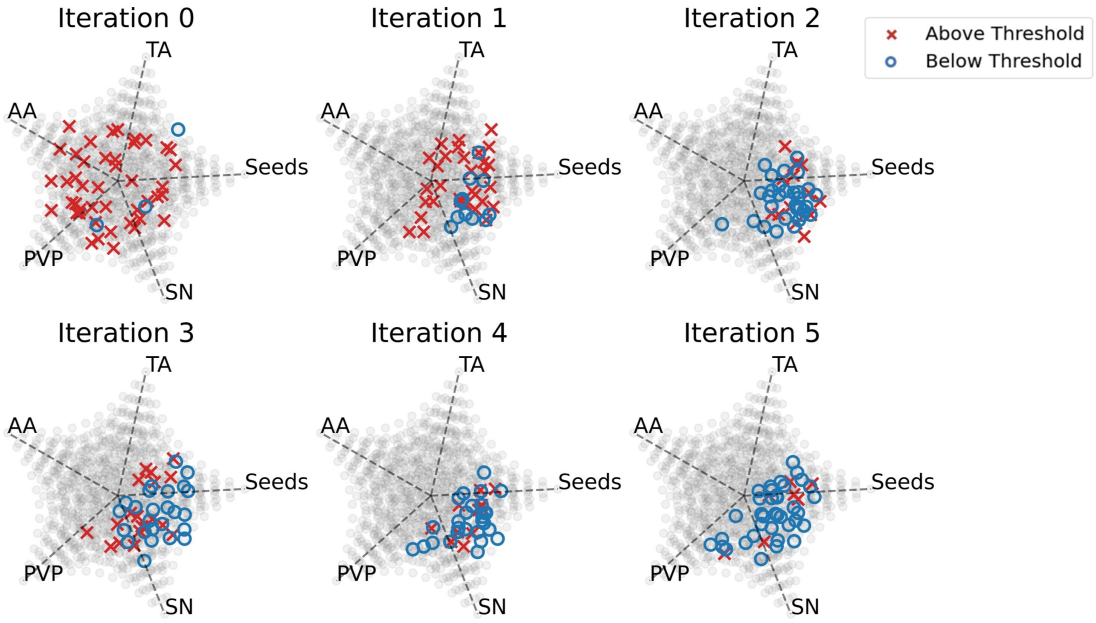



**Figure 4.3:** This figure shows the volume fractions of each sample that was synthesized in each iteration. Each iteration contains 48 samples. The plot of each iteration is a 2-dimensional representation of a 5-dimensional space. Each corner of the pentagon represents a reagent: ascorbic acid (AA), tannic acid (TA), silver seeds (Seeds), silver nitrate (SN), and polyvinylpyrrolidone (PVP). Samples that are located closer to a corner indicate a higher volume fraction of the reagent labeled in the respective corner. Samples located near the center of the plot, indicated by the intersection of the dotted black lines, suggest equal volume fractions of all reagents. The grey dots are visual aids to show the shape of the plot. A red "x" represents a sample that was classified as "Above Threshold" using the distance metric. A blue "o" represents a sample that was classified as "Below Threshold". There were 3/48 samples labeled "Below Threshold" in iteration 0, 11/48 in iteration 1, 34/48 in iteration 2, 26/48 in iteration 3, 36/48 in iteration 4, and 41/48 in iteration 5.

of PVP, and high-volume fractions of silver nitrate and silver seeds. In this iteration, the majority of the samples were plate-like and had distances below the threshold. The design rules learned from iteration 2 were maintained through iteration 5, which indicates convergence to a set of rules.

There are several explanations for why these design rules led to the synthesis of small, colloidally stable, monodisperse, plate-like particles. According to Yang et al., slow growth kinetics of the nanoparticles is essential for the formation of plate-like particles, which means that weaker reducing agents are favorable over stronger ones[62]. In our experiment, ascorbic acid was the strongest reducing agent followed by tannic acid and then PVP. This could explain why moderate-volume fractions of PVP and low-volume fractions of tannic acid and ascorbic acid were selected for the formation of plates. In addition to being a weak reducing

agent, PVP can also bind to the Ag(111) facet of the plate (i.e., the axial direction) via van der Waals interactions and induce growth in the radial direction[63]. The data also shows that high concentrations of silver seeds and silver nitrate were needed to form our targeted particles. In our distance metric, we simulated small plates (20-40 nm), and we included a term to bias towards monodispersity. High-volume fractions of silver seeds result in a large number of nucleation sites for the nanoparticles to grow and high-volume fractions of silver nitrate are necessary for the production of the silver atoms that grow on these sites[64]. This combination of reagents would thus result in a high concentration of small monodisperse plates.

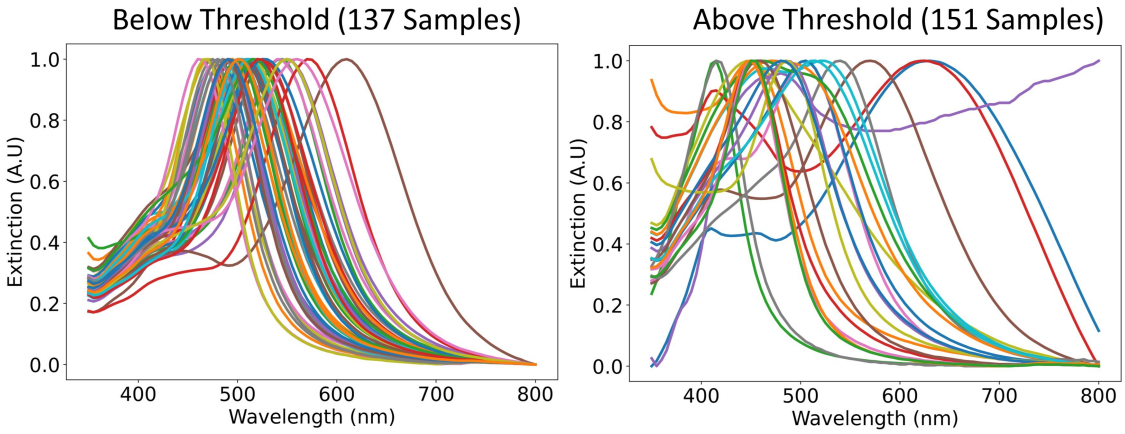



**Figure 4.4:** Representative UV-vis spectra of the samples that were classified as "Below Threshold" and the ones that were classified as "Above Threshold".

The classified UV-vis spectra of all the samples were plotted and shown in Figure 4.4. The spectra that were classified as "Below Threshold" have similar shapes (i.e., single narrow peak) and peak intensity positions (i.e., greater than 450nm) to the simulated plates in Figure 4.1. In addition, the variance in the peak intensity positions indicates that plates of different sizes are being synthesized, which means that the effect of the reagents on the plate size can be explored in the next section. The spectra of the samples that were classified as "Above Threshold" are also shown in Figure 4.4. These spectra have vastly different shapes such as ones with broad peaks, ones with high extinction at high wavelengths, and also ones that are similar to the simulated sphere spectra in Figure 4.1. The successful classification of UV-Vis spectra shown in Figure 4.4 supports the implementation of our chosen distance metric. Geometric metrics, such as the aspect ratio of plates or spheres, could also be great metric descriptors of morphology. However,

determining metrics such as the aspect ratio of a particle usually requires microscopy, which is typically a low throughput and costly technique. Instead, we rely on using bulk classification metrics (e.g. UV-Vis, SAXS) that can be used as proxies for the ensemble geometry of particles. As seen by the results of our classification, UV-Vis spectroscopy was able to differentiate samples that formed small, monodisperse, (triangular or circular) plate-like particles from other kinds of morphologies and sizes. Therefore, UV-Vis spectroscopy is an effective and cost-effective technique for this classification. We would also like to clarify that, for this work, we focused on demonstrating the hierarchical analysis method, which could be applied to any particle of interest. We chose to focus on making small monodisperse plate-like particles. However, the methodology could be used for any other particle shape of interest with modification of the simulations and the definition of the distance metric.

### 4.4.2 SAXS Structural Exploration

From Fast Spectroscopic Exploration, we selected 137 samples that were classified as "Below Threshold" based on their UV-Vis spectroscopy curve. To determine how the concentration of the reagents affects the size of the plates, we took SAXS measurements on these samples. Unlike UV-Vis spectroscopy, SAXS is a direct method of characterizing nanoparticle structure in the dispersion state due to its sensitivity to structural parameters in this size range. The scattering can be mathematically modeled as the Fourier transform of the continuous electron density in all sample orientations[65], and because of this, geometric models can be used to fit experimental data to obtain information on size, shape, shape fraction, polydispersity, or concentration. Like other scattering techniques, SAXS data has the limitation of being degenerate, which means that many structures can have the same SAXS curve. This is due to the phase problem where the detector can only measure the intensity of the scattered x-rays and not the phase, which contains the majority of the structural information[65]. Because of this, it is well known that the choice of geometric model used to fit the data must be supported with data from other characterization methods or other pieces of information[66]. In our method we only perform SAXS on samples that we hypothesize are plates based on the shape of their UV-Vis spectroscopy curve. After collecting the SAXS data, we can then perform electron microscopy on a few samples to confirm this hypothesis.

Supported by evidence primarily from UV-Vis spectroscopy, the geometric model used to fit the SAXS

data was a dilute cylinder model because a cylinder with a radius much larger than its length is representative of a plate. Although the shape of particles is most frequently rounded, in certain samples we also observe nanoplates with triangular shapes. In SAXS analysis these were approximated with a circular plate model because of the low difference between the scattering patterns for somewhat polydisperse samples. However, their presence could have a significant effect on the spectroscopic properties. The fundamental equation for small angle x-ray scattering is shown in equation (4.2). The variable $n$ refers to the number density of the particles, $\Delta\rho$ refers to the contrast, $V$ refers to the volume of a single particle, $S(q)$ refers to the structure factor, and $P(q)$ refers to the form factor[67].

$$I(q) = n\Delta\rho^2 V^2 P(q) S(q) + background \tag{4.2}$$

This equation was simplified with the following expression:

$$A = n\Delta\rho^2 V^2 \tag{4.3}$$

Since $\Delta\rho$ and $V$ are constant:

$$A \propto n \tag{4.4}$$

By substituting (4.3) back in equation (4.2), assuming that the nanoparticles are dilute($S(q)$ = 1), assuming a background of 0, and using the form factor of a plate, the scattering equation for a plate was created. The subscript "p" refers to a plate.

$$I(q) = A_p P_p(R, T, PD, q) \tag{4.5}$$

Using sasmodels (`www.github.com/SasView/sasmodels`), the parameters in equation (4.5) were obtained by fitting the model to the experimental data. From the form factor $P_p(R, T, PD, q)$, the radius $R$, the lognormal polydispersity of the radius $PD$, and the thickness $T$ of the plates were obtained. The scale parameter $A$ was also obtained, which is a parameter proportional to the concentration of the particles in the solution.

After fitting all the data with a plate model, it was discovered that 23 out of the 137 fits were sub-optimal

by the residuals or errors of the fits. To fit these datasets, a combined model of a plate and sphere (equation (4.6)) was used only on the data from the 23 samples. In equation (4.6), the subscript "p" refers to a plate and "s" refers to a sphere.

$$I(q) = A_p P_p(PD, R, T, q) + A_s P_s(R, q) \tag{4.6}$$

This method of hierarchical fitting ensures that the most simple model (i.e., polydisperse plates) is used first to fit the data. This means the data that agreed with this simple model most likely come from samples that contain polydisperse plates. As we update the model by adding spheres, we risk overfitting the data and are less confident in the structural parameters obtained from these fits. While the plate and sphere model improved the fits of the 23 samples that did not fit the polydisperse plate model, they were not used in further data analysis, since the objective of this work is to study the structural features of nanoplates.

After fitting the SAXS data with the models, we randomly chose a sample that was fit using the polydisperse plate model. Electron microscopy was performed to validate the model choice that was used. The SAXS curve of the sample that was fit using a plate model and an image of the sample is shown in Figure 4.5 B below.

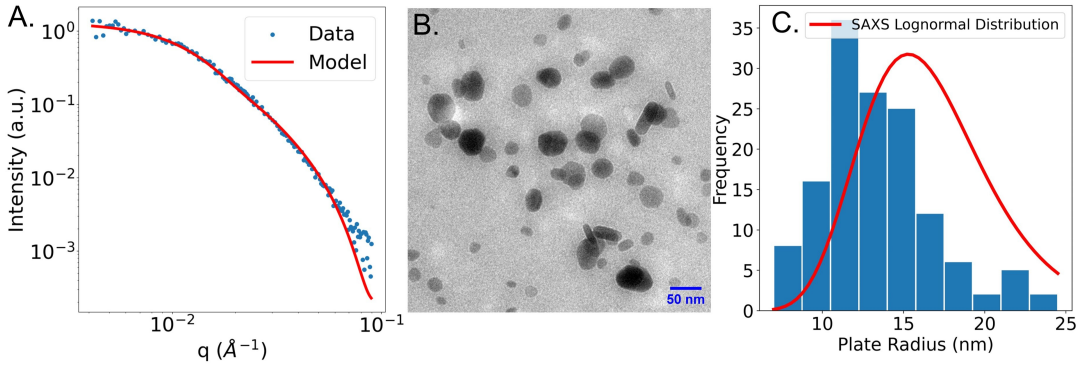



**Figure 4.5:** This figure shows the data from the randomly chosen sample that was fit using a plate model. (A) The SAXS curve and the plate model that was used to fit the data. (B) An electron microscopy image from the set of images that were taken of the sample. (C) A histogram of the plate radii from all the images that were taken. The lognormal distribution with the plate radius and polydispersity obtained from SAXS is plotted in red.

Figure 4.5 A shows the experimental SAXS scattering curve of a plate-like sample and the plate model used to fit the experimental data. From the fit, an average plate radius of 16.2 nm, a lognormal plate

radius polydispersity of 0.24, and a plate thickness of 7.2 nm was obtained. Figure 4.5 B shows a TEM image taken of the same sample. From that image, circular plates of different radii are observed. It is likely that some plates are stacked on top of each other, which causes some plates to have darker contours in the image. It is also possible that spheres are stacked on top of plates. However, we believe, based on evidence from UV-Vis spectroscopy and SAXS, that the sample is primarily composed of plates. In addition, some plates stand on their edges which could explain the rod-like shape in TEM. This analysis is consistent with that of similar silver nanoplates in Gestraud et al [68]. The image in Figure 4.5 B is a representative image from a much larger set. By measuring the plate radii using all images (included in the Supporting Information S11-S18), a histogram of the plate radius was created and compared to the lognormal distribution obtained from SAXS. The comparison between the radius distribution derived from TEM and SAXS is shown in Figure 4.5 C. From the TEM images, the radius of 140 individual plates was measured. The mean radius was 13.3 nm and the lognormal radius polydispersity was 0.26. The lognormal radius polydispersity obtained from microscopy and SAXS have a good agreement, but the plate radius obtained from SAXS is slightly larger than that from microscopy. There are several reasons for this discrepancy. The first reason for the disagreement between SAXS and TEM could be that data obtained from SAXS is often biased towards larger particles because of their stronger scattering signal[69]. Another reason is the difficulty in distinguishing between thickness and radius in the TEM images since the particles can be imaged in all possible orientations.

**Results of SAXS Structural Exploration**

The interpretation of structural features from samples obtained via SAXS was performed in several steps. The first step relates the composition of the synthesis to the plate thickness, scale parameter, radius, and polydispersity. We first use histograms to visualize the different types of plates that were formed in our experiment. We then attempt to discover relationships between the concentrations of the reagents and the structural features using contour plots. This information can be used to obtain a better understanding of the role of each reagent in the synthesis of silver nanoplates. Understanding the roles of each reagent could be useful to modify the experimental design space in case a target structure cannot be synthesized.

### Plate Thickness, Plate Radius, Plate Radius Polydispersity, and Scale

The histograms in Figure 4.6 show the plate thicknesses, plate radius, lognormal radius polydispersity, and scale parameter of the 114 samples that were obtained from SAXS data that were fit using the polydisperse plate model. Since compositions were randomly sampled, the histograms are representative of the samples in the targeted design space of plate-like particles learned by the Gaussian process classifier after the Fast Spectroscopic Exploration step. The plate thicknesses ranged from 2.4 nm to 11.4 nm with a mean of 5.7 nm and a standard deviation of 1.3 nm. Most of the thicknesses (about 90 %) are less than 7 nm. The plate radius ranges from 5.0 nm to 34.0 nm with a mean of 15.8 nm and a standard deviation of 5.1 nm. The plate radii seem to be somewhat normally distributed. The lognormal polydispersity of the plate radii ranges from 0 to 1 with a mean of 0.45 and a standard deviation of 0.28. Most of the samples seem to have a polydispersity of less than 0.5, but there are some with a polydispersity of 1.0. A polydispersity of greater than 0.5 can be considered excessive and likely caused by another factor such as aggregation or low concentrations that limit the quality and reliability of the SAXS data. Despite this, the samples with high polydispersity were kept for the analysis, since they were a minority of all the samples (31 out of 114 samples). Finally, the concentration of the particles was determined by the scale parameter in (4.4), which is proportional to the concentration of the particles in the solution. It ranges from 0.07 to 0.38 and seems to be normally distributed with a mean of 0.18 and a standard deviation of 0.06. All this information can inform us of the sizes of the plates that are most commonly synthesized by the algorithm used in the Fast Spectroscopic Exploration section and their size limits.

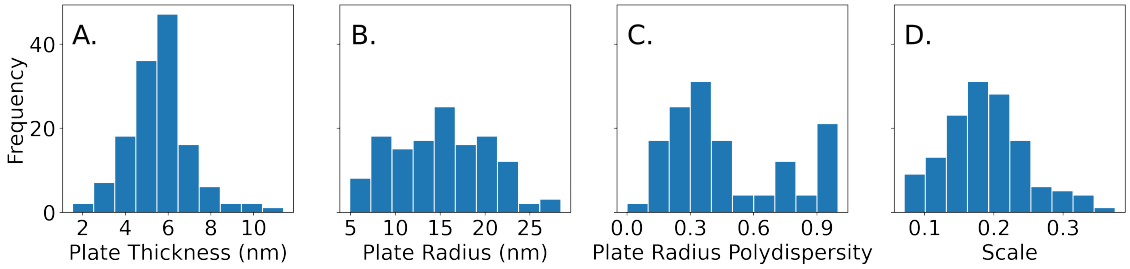



**Figure 4.6:** A histogram of the plate thickness (A), plate radius (B), lognormal plate radius polydispersity (C), and scale (D) that were obtained from the 114 samples that were fitted with the polydisperse plate model.

To determine the effects of the experimental design parameters on the structural features of the samples,

the 5-dimensional design space was simplified into 2 dimensions. SHAP, a method to explain machine learning models[70], was first used to identify the top two most important reagents that contributed to each of the measured structural features. This ignores data on the concentrations of the other three reagents but increases the interpretability since it is much easier to visualize data in two dimensions. From SHAP, it was determined that PVP and silver nitrate were the strongest contributors to variations in plate thickness, while tannic acid and PVP contributed the most to variations in plate radius. A Gaussian process regressor was trained with the top two contributing factors (i.e. reagents) and the structural parameter to obtain a contour plot that shows how the regions in the design space affect each structural parameter.

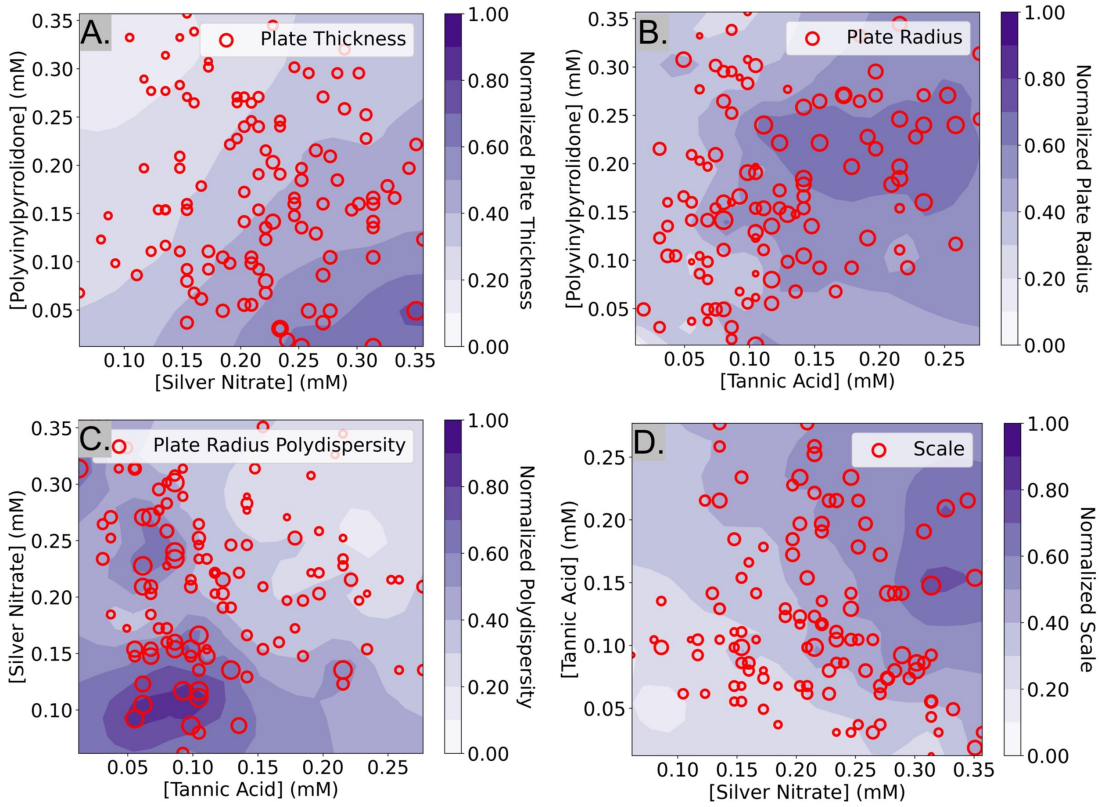



**Figure 4.7:** Contour plots of the top two most influential reagents on plate thickness (A), plate radius (B), lognormal plate radius polydispersity (C), and the scale parameter of the particles (D), which is proportional to the concentration of the particles. The structural information was obtained by fitting a polydisperse plate model to SAXS data. The marker size is directly correlated to the value of the structural feature, and the contours represent the design space learned by the Gaussian process regressor.

In the contour plot of plate thickness (Figure 4.7 A) there is a noticeable pattern where low concentrations of PVP and high concentrations of silver nitrate resulted in thicker plates. A reason for this could be

that higher concentrations of silver nitrate lead to higher amounts of silver atoms available in the solution. Low concentrations of PVP result in larger plates because PVP can act as a stabilizer to prevent large particles from forming. As discussed before, it has been reported that PVP preferentially binds to the Ag(111) facet (i.e., the axial direction) via van der Waals interactions due to the larger surface area. This explains why lower concentrations of PVP result in thicker plates. In Figure 4.7 B., tannic acid and PVP were identified as the top two most influential reagents on plate radius. From Figure 4.7 B., it seems that the region where samples with the largest plate radii are obtained is in the darker purple region, between 0.10 - 0.25 mM tannic acid and 0.15 - 0.25 mM PVP. This indicates that relatively high concentrations of both tannic acid and PVP are essential for the formation of large silver nanoplates, but excessive amounts result in a reduced size. As discussed previously, tannic acid and PVP can act as both reducing agents and stabilizers that control the size and shape of the nanoplates[71][72][73]. Low concentrations of reducing agents could result in smaller nanoparticles by limiting the number of silver atoms available in the solution. However, we hypothesize that high concentrations of these reagents could hinder the growth of the nanoplates by inducing steric hindrance effects on the surface of the particles[63]. Another reason why high concentrations of both PVP and tannic acid are detrimental to the plate radius could be due to the cross-linking of the reagents. During the synthesis, it was observed that when high concentrations of tannic acid were mixed with PVP, the solution became turbid. The cross-linking of tannic acid and PVP is likely caused by the pyrogallol compounds in tannic acid interacting with PVP via hydrogen bonding[74] or $\pi$-$\pi$ stacking[75]. Therefore, it is hypothesized that high concentrations of tannic acid could contribute to a lower plate radius by inducing cross-linking and reducing the number of PVP and tannic acid molecules available for the reaction. Finally, the largest plates of the samples labeled “Below Threshold” are around 30-34 nm in radius and are located in the central region of Figure 4.7 B, supporting the hypothesis that moderate concentrations of tannic acid and PVP lead to large plates.

In Figure 4.7 C, a contour plot of the effect of the top two reagents on the lognormal polydispersity of the plate radius is shown. It was determined that silver nitrate and tannic acid had the greatest effect on this feature and that high concentrations of both reagents increase the probability of synthesizing monodisperse plates. Due to the lack of studies involving the effect of silver nitrate and tannic on the polydispersity of plate radius, no comparisons to the literature were made. In equation Figure 4.7 D, the effect of tannic

acid and silver nitrate on the scale parameter of the particles is shown. The contour plot shows that high amounts of silver nitrate and tannic acid increase the scale parameter of the particles. This is consistent with the explanation that silver nitrate is the precursor of the silver ions in the reaction and that tannic acid acts as a reducing agent. In summary, we have described many testable hypotheses that could explain the observations in our experiment. While these hypotheses have not been experimentally verified in this manuscript, we aim to tackle that in a follow-up work.

### UV-Vis Spectroscopy as a Proxy for Particle Size

In the Fast Spectroscopic Exploration section of this paper, it was claimed that UV-vis spectroscopy was limited in determining the size of a nanoparticle. Since both a UV-vis spectrum and a SAXS scattering curve were obtained for each sample, that claim could be evaluated. By assuming that the plate radius determined by SAXS was the "true" value, the radius determined by SAXS and the peak wavelength of the sample's UV-vis spectra were plotted. Only the 114 samples that were determined to be polydisperse plates are shown in the plot. In addition to this, the peak wavelength positions of the simulated UV-vis spectra using nanoDDSCAT[59] were added to the plot to compare how the peak wavelength position changes as the plates get bigger in simulations compared to experiments.

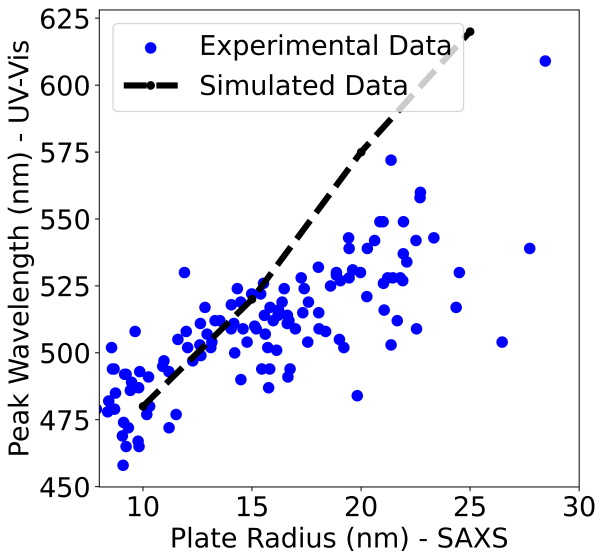



**Figure 4.8:** The comparison between the plate radius determined by the peak wavelength position from UV-vis Spectroscopy and the radius determined by SAXS. The data from the simulated plates are also plotted assuming that the radius determined from SAXS is the "true" radius.

From Figure 4.8, it is evident that the peak wavelength position increases as the plate radius increases. This is seen in both the experimental data and the simulations and confirms that the peak wavelength position of the spectra is positively correlated with the size of the nanoparticle. However, it seems that the nanoDDSCAT simulation overestimates the peak wavelength position compared to the experimental data. An explanation for this could be the presence of organic molecules in the solution such as PVP or tannic acid which could account for this disagreement.

In summary, Figure 4.8 shows that the peak wavelength positions from UV-Vis spectra are directly correlated with the size of the nanoparticle. However, caution must be exercised when using the peak wavelength position to determine the absolute size of a nanoparticle especially when analyzing samples with impurities or with anisotropically shaped nanoparticles. This underscores the importance of SAXS for determining the size of silver nanoparticles in dispersion since X-rays are only sensitive to the size and morphology of these electron-dense nanoparticles. Dynamic light scattering (DLS) is another technique that is sensitive to the size and shape of nanoparticles and is often more accessible than SAXS. However, DLS relies on size measurements based on the diffusion of particles, which has a much lower structural resolution and is model-dependent. Thus, the interpretation of DLS data requires an assumption of either spherical particles to convert a diffusion coefficient into a diameter or an assumption of a shape (e.g., plates) and an aspect ratio to quantitatively interpret the size from the diffusion coefficient. For these reasons, SAXS is more powerful for the analysis of silver nanoparticles as we report in this study.

## 4.5 Conclusion

We performed a data-driven exploration of the synthesis of silver nanoplates, which combines the iterative sampling capabilities of a closed-loop system and the straightforward data interpretation and visualization of a systematic study. Our closed-loop system goes beyond solving an optimization problem and focuses on the generation of scientific knowledge. To do this, we take advantage of multimodal and complementary characterization methods. Our data-driven method balances the use of characterization techniques based on cost and information gained. UV-vis spectroscopy was used as a fast proxy to determine how to synthesize silver nanoplates, while SAXS, a more expensive but direct characterization method, was used for information on the structural features of these nanoplates. In summary, using a Gaussian process classifier

and UV-Vis spectroscopy, we started from a relatively large design space and narrowed it down to a region where plate-like particles are most likely to be formed. In this region of interest, random sampling and SAXS were used to obtain simplified contour plots containing information on the effect of the top two most contributing chemical reagents on the structural features of nanoplates. Our findings on the compositions of the reagents that are necessary to form plate-like nanoparticles as well as the effect of these reagents on the nanoparticle's structural features are consistent with the literature. Because of this, we envision that all the methods that were used for the synthesis, data science, and characterization can be applied to other, more complex, systems to accelerate the discovery of novel materials.

# Chapter 5

# Efficient Analysis of Small-Angle Scattering Curves for Large Biomolecular Assemblies Using Monte Carlo Methods

This work presented in this chapter is based from:

- Zhiyin Zhang, Huat T Chiang, Ying Xia, Nicole Avakyan, Ravi R Sonani, Fengbin Wang, Edward H Egelman, James J De Yoreo, Lilo D Pozzo, F Akif Tezcan, Design of light-and chemically responsive protein assemblies through host-guest interactions, *Chem*, 2025. [76]
- Efficient Analysis of Small-Angle Scattering Curves for Large Biomolecular Assemblies Using Monte Carlo Methods, Huat Thart Chiang, Zhiyin Zhang, Kiran Vaddi, Akif Tezcan, Lilo Pozzo, Efficient Analysis of Small-Angle Scattering Curves for Large Biomolecular Assemblies Using Monte Carlo Methods, *Journal of Applied Crystallography*, 2025. [77]

## 5.1 Abstract

Structure elicitation from small angle scattering (i.e. SAXS, SANS) curves of large biomolecular assemblies is notoriously challenging. This is because the simulation of high-resolution features in the structure of large macromolecular assemblies, such as de-novo protein assemblies, is computationally demanding when

it needs to cover a broad range of length scales. Conventional methods such as the numerical approximation to the Debye equation or the use of spherical harmonics do not scale well as the size of the assembly increases, which limits their application to small structures (e.g. individual proteins). This work explores the effectiveness of a Monte Carlo method to simulate and fit scattering curves for large biomolecular assemblies spanning over ranges covering atomic and molecular detail (e.g., spacing and orientation of proteins in an assembly), as well as large scale (100's of nm) features. Due to its excellent speed and scalability, it can be combined with a fitting algorithm to extract structural features from experimental small-angle scattering curves in biomolecular assemblies that are otherwise intractable for interpretation. This work first demonstrates the effectiveness of the tool using experimental small angle X-ray scattering (SAXS) data from tile-like proteins that assemble into 1D tube-like macromolecular structures. The diameter distribution of tubes is extracted from SAXS fits, and this quantitatively compared with distributions from electron microscopy. SAXS data is also obtained from 2D sheet-like protein assemblies and the proposed method is used to quantify structural features such as the separation distance between protein building blocks and the flexing of the sheet. An open-source implementation of the methodology is provided for use in a broad range of biological systems involving multi-scale scattering analysis.

## 5.2 Introduction

The study of biomolecular assemblies and hierarchical systems is important because many of these complex structures are responsible for the physiological processes that make life possible (e.g. microtubules, fibrin, actin filaments). Several characterization methods can be used to study these structures. Some of the most powerful characterization methods include electron microscopy [78] [79], atomic force microscopy [80], and small angle scattering (SAS), each of which has its strengths and weaknesses. In this paper, we focus on SAS, which is particularly useful for studying biomolecular self-assembly because of its ability to perform *in situ* measurements with high spatial (angstrom to micron) and temporal (up to milliseconds) resolution [49]. These characteristics make SAS a preferred characterization method for the study of biomolecular assembly. Interpreting SAS data for unknown structures is challenging since it records signals from an ensemble of particles, which can be affected by shape and size polydispersity, and because it requires a solution to an ill-posed "inverse problem" without a guaranteed unique solution. During a scattering experiment,

the isotropic intensity of scattered X-rays or neutrons observed on a detector is typically converted to a 1D curve of intensity as a function of the momentum transfer vector $q$ using an azimuthal average [81]. This scattering curve is mathematically defined as the Fourier transform of the scattering-length density, or electron density for the case of X-rays, of the 3D scattering object [82]. Because of this, a good match between an experimental scattering curve and a mathematically calculated one (model) is interpreted as suggesting that they originate from the same 3D object. Quantitative shape and dimensional parameters are then obtained from the model.

The process of mathematically calculating the experimental scattering curve can be computationally expensive depending on the object of interest. The scattering curves of geometric objects such as spheres, ellipsoids, and cylinders can be calculated analytically from established mathematical models [83] [84] [85] [48]. In contrast, biomolecular objects with complex shapes, such as proteins, do not usually have an exact analytical solution and a numerical method must be used. Common methods used to calculate the scattering curve of biomolecules and their assemblies are based on approximating the Debye equation and often use methods such as spherical harmonics to accelerate the computation [86].

The enumerated Debye equation and the spherical harmonics methods work well on small biomolecular structures [87] [86] but do not scale well for large biomolecular assemblies, such as those involving many protein sub-units. The computational time of the Debye equation scales as $O(N^2)$ where $N$ is the number of atomic coordinates in the model of the structure and for the spherical harmonics as $O(N)$ [86]. Because of this, calculating the scattering curve of large biomolecular assemblies, which can be composed of many millions of atoms, often involves either coarse-graining the structure with geometric objects, using a small representative sample of all the atoms in the structure, or an expensive calculation that requires significant computational resources. The Debye equation is defined in Equation (5.1), where $f_k$ and $f_j$ are the differences in the scattering length density between the atoms and the solvent background (water), $r_{jk}$ is the interatomic distance between coordinates $j$ and $k$, and $q$ is the momentum transfer vector [88].

$$I(q) = \sum_k \sum_j f_k f_j \frac{\sin(q r_{jk})}{q r_{jk}} \tag{5.1}$$

An alternative method to approximate the scattering intensity is the Monte Carlo Distribution Function Method (MC-DFM) [89] [90] [91], which is a Monte Carlo approximation to the fully enumerated pairwise

summation of the Debye equation. This method first calculates the pairwise distribution function from the atomic coordinates and then transforms it into the scattering intensity using Fourier inversion. Since the pairwise distribution function is a probability distribution of the distances between two randomly selected points, it can be created using a simple algorithm that randomly selects two atomic coordinates and calculates the pairwise distance between them. The pairwise distribution is then defined as a histogram of these distances. With enough sampling, the distance distribution approximates the distribution of scatterers and can thus be used to calculate the intensity of the scattering using Equation (5.2). The MC-DFM is computationally advantageous because it scales well with an increase in the number of atomic coordinates $N$. The computationally expensive step of the calculation is the creation of the pairwise distribution function, which involves random sampling of atomic coordinates. This means that, as the number of atomic coordinates increases, the number of sampled coordinates used to create the pairwise distribution function should also increase. However, we show that a large sampling of pairwise distances ($n$ = 10 million) can be used to accurately and rapidly calculate the scattering curve of various biomolecular assembly structures up to micrometers in size. The MC-DFM balances the need to retain the small-scale detail of the non-spherical biological building blocks (e.g. proteins) while still enabling the computation of scattering from biomolecular assemblies that are much larger than any individual subunit. There are several examples where it has been used to simulate the scattering of complex biomolecular structures [92], biomolecular structures with high symmetry [93] [94], and simulations based on information from electron microscopy [95].

The computation time of the MC-DFM methodology is efficient and comparatively short, about a few seconds [89], making it compatible with optimization algorithms to fit experimental data and extract unknown structural parameters [96]. In this work, the MC-DFM is used with a genetic algorithm to numerically fit small angle scattering curves to experimental data to extract structural parameters of large hierarchical biomolecular assemblies. The MC-DFM resolves structural details ranging across multiple length scales and does not require coarse-graining approximations. A novel aspect of this work is the implementation of the MC-DFM using matrix operations, which allows for the efficient calculation of scattering curves in timescales of seconds. In addition, modeling of assemblies from a single building block using rigid-body transformations allows for rapid structural adjustments without explicitly recalculating the positions of every atom.

## 5.3 Methods

### 5.3.1 Monte Carlo Distribution Function Method (MC-DFM)

The mathematical equation of the MC-DFM is defined in Equation (5.2), where $p(r)$ is the pairwise distribution function, $r$ is the interatomic distance, $N_{bins}$ is the number of bins in the pairwise distribution function, and $q$ is the momentum transfer vector. Note that the MC-DFM is similar to the Debye equation since both involve a sum of the $sinc$ function. The scattering length density difference between each atom and the solvent background is accounted for in the pairwise distribution function, Equation (5.3), where each pairwise distance, $r_{jk}$, is weighted according to the product of the scattering length density differences, $f_j f_k$, of each pair of atoms. Pairwise distances of zero are eliminated from the pairwise distribution to satisfy the constraint of $j \neq k$. The delta Dirac function $\delta$ is used to represent the value at each bin in the pairwise distribution function and the weighted sum of all the delta functions makes up the pairwise distribution [89]. The pairwise distribution $p(r)$ is computed to an arbitrary number of distances, $n$, that determine the resolution. Sampling all distances within a structure is unnecessary, as the pairwise distribution is expected to converge after a sufficient number of samples.

$$I(q) = \sum_{i}^{N_{bins}} p(r) \frac{\sin(qr)}{qr} \tag{5.2}$$

$$p(r) = \sum_{j \neq k}^{n} f_j f_k \delta(r - r_{jk}) \tag{5.3}$$

Notice that the MC-DFM in Equation (5.2) contains a single summation over the number of bins in the pairwise distribution compared to the double sum in the Debye Equation (5.1) over the number of atomic coordinates. Because of this, the algorithmic complexity of the MC-DFM should not depend on the number of atomic coordinates in the structure, which should allow it to be applied to larger structures. A comparison of the procedures used to calculate the scattering intensity from a PDB file of a protein structure is shown in Figure 5.1.

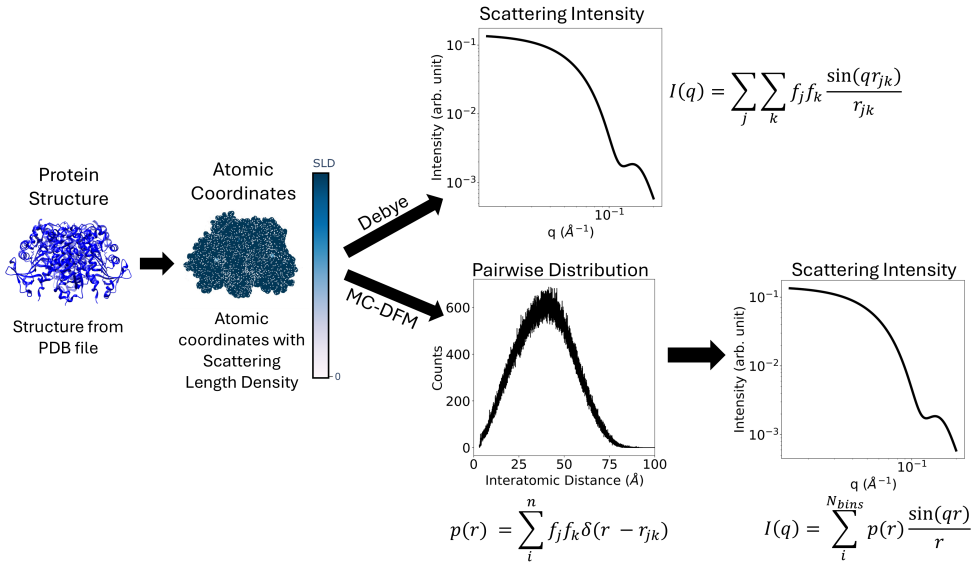



**Figure 5.1:** The process of calculating the scattering curves of a structure from a PDB file. First, atomic coordinates are extracted and scattering length density differences between each atom and the solvent background (water) are assigned to each atom. The Debye method calculates the scattering intensity from the coordinates, while the MC-DFM first calculate the pairwise distribution and then converts that to the scattering intensity using Fourier Inversion.

### 5.3.2 Convolution of the Building Block and Lattice

A challenge in calculating the scattering curve of a large biomolecular assembly is the need to obtain the coordinates of all the atoms in the structure. This can be computationally expensive because biomolecules, such as proteins, are each composed of thousands of atoms, and many relevant constructs are created by assembling thousands or millions of these sub-units. Therefore, creating an accurate computational representation can be prohibitively expensive. In addition, these structures can have complex shapes, which is a critical limitation for using simple geometric approximations (e.g. sphere, ellipsoid, cube, cylinder) to model the full assembly.

Using MC-DFM, the scattering of large hierarchically organized biomolecular assemblies can be calculated by taking advantage of the periodicity inherent in many structures. A large biomolecular assembly can be modeled by utilizing the lattice coordinates (i.e. relative positions of repeating subunits), the relative orientations of these sub-units within the assembly, and the internal atomic coordinates of each building block or repeating sub-unit (e.g., a biomolecule or a group of biomolecules), without explicitly defining the

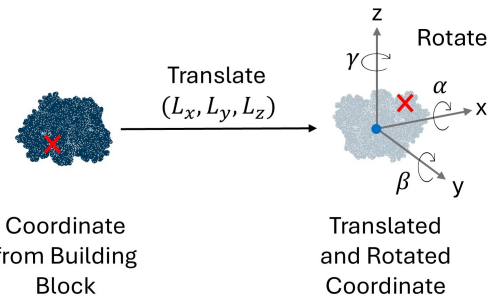



**Figure 5.2:** Translational and rotational transformations can be applied to randomly sampled coordinates of the building block.

coordinates of every atom in the overall assembly. This is achieved by applying rigid-body transformations (i.e. vectorial translation and relative rotation) to the randomly selected coordinates in each building block used to create the pairwise distribution to fill the assembly structure rather than explicitly defining all atomic coordinates.

The relevant rigid body transformations can be broken down into translations and rotations. The translation matrix representation shown in Equation (5.4) is used to translate the coordinates of the rigid body building block to the lattice coordinate $(L_x, L_y, L_z)$.

$$T = \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} L_x \\ L_y \\ L_z \end{bmatrix} \tag{5.4}$$

The rotation matrix representation shown in Equations (5.5), (5.6), and (5.7) is used to rotate the coordinates of the rigid body building block by $\alpha, \beta, \gamma$ degrees around the $x, y, z$ axes, respectively.

$$R_x(\alpha) = \begin{bmatrix} x_{rx} \\ y_{rx} \\ z_{rx} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5.5}$$

$$R_y(\beta) = \begin{bmatrix} x_{ry} \\ y_{ry} \\ z_{ry} \end{bmatrix} = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5.6}$$

$$R_z(\gamma) = \begin{bmatrix} x_{rz} \\ y_{rz} \\ z_{rz} \end{bmatrix} = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{5.7}$$

This approach of defining a hierarchical structure through the coordinates of a building block, along with the rigid-body translational and rotational transformations at each lattice point, is effective because the MC-DFM uses random sampling to first create the pairwise distribution. Consequently, the necessary transformations only need to be applied to randomly chosen building block and lattice coordinates, which is especially beneficial for handling large structures. Additionally, this method allows for easy structural modifications by adjusting the lattice coordinates, making it compatible with a fitting algorithm where the scattering curves of assemblies of various sizes or shapes can be tested. Finally, implementing the transformations through matrix operations enhances computational speed and efficiency.

Consider modeling a helical tube-like structure from an assembly of a cuboid-shaped protein called L-Rhamnulose-1-phosphate aldolase (RhuA) [97], as shown in Figure 5.3. This protein will be used as a model system in this work and further details about it can be found in the results section. The first step is to define the building block which will be duplicated, translated to each of the lattice coordinates, and rotated to form the assembly. In this example, the building block is a single RhuA protein, but it could be a multimer in other cases for other structures. The next step is to determine the lattice coordinates, which can be calculated using the parametric equations for a helix. This is shown in Equations (5.8) and (5.9), where $R$ is the radius of the helix and $f$ is a parameter that controls the pitch of the helix. The $z$ values range from 0 to the length of the tube, are equally spaced apart, and contain a number of points equal to the number of RhuA building blocks in the tube assembly.

$$x = R\cos(fz) \tag{5.8}$$

$$y = R\sin(fz) \tag{5.9}$$

Next, the rotation applied to each lattice coordinate is calculated by converting the coordinates of the helix to polar coordinates to obtain values for $\alpha, \beta, \gamma$, which are the rotation, in degrees, applied to each

protein about the $x, y, z$ axes, respectively, so that it is oriented relative to the center of the helix. For this helical tube-like structure, only rotation about the $z$ axis is necessary, and it was calculated using Equation (5.10). However, if necessary, rotations about the $x, y$ axes can also be implemented for other structures, such as icosahedrons.

$$\gamma = -\frac{180}{\pi} \arctan\left(\frac{y}{x}\right) \tag{5.10}$$

For this specific structure, it is also expected that adjacent proteins are flipped by 180 degrees relative to the center of the helix, which results in an alternating checker-board relative orientation of the proteins in the final assembly. This orientation was previously observed using AFM and TEM, on an assembly of RhuA [98]. This structural effect can easily be implemented by adding 180 degrees to $\gamma$ for every other lattice coordinate. After this, the array containing the specified translation and rotation for each of the lattice coordinates is complete. The first three columns of the array specify the magnitude of translation required $(L_x, L_y, L_z)$, and the last three columns specify the amount of rotation required $(\alpha, \beta, \gamma)$ for a coordinate from a building block to be placed correctly at each successive lattice coordinate.

Using the MC-DFM, coordinates from the building block are randomly sampled and appropriate translational and rotational transformations are applied to the building block coordinates according to another independent randomly sampled lattice coordinate and rotation. This is repeated with another building block and lattice coordinate and the distance between the two points is calculated. Distances are repeatedly calculated between two randomly sampled coordinates until the pairwise distribution is created, which is later converted into the scattering curve. These procedures are summarized in Figure 5.3. To clarify, the method of independently sampling the building block and lattice can be used for structures with periodicity but this is not required. The advantage of this convolution is that it bypasses the expensive step of explicitly modeling the full biomolecular assembly, which is advantageous (i.e. fast computation and low memory requirements) when using MC-DFM together with a fitting algorithm or for modeling very large assemblies.

### 5.3.3 Implementation

A Python implementation of MC-DFM can be found at (`https://github.com/pozzo-research-group/MC-DFM/tree/main`). For all the case studies presented in this work, a personal laptop com-

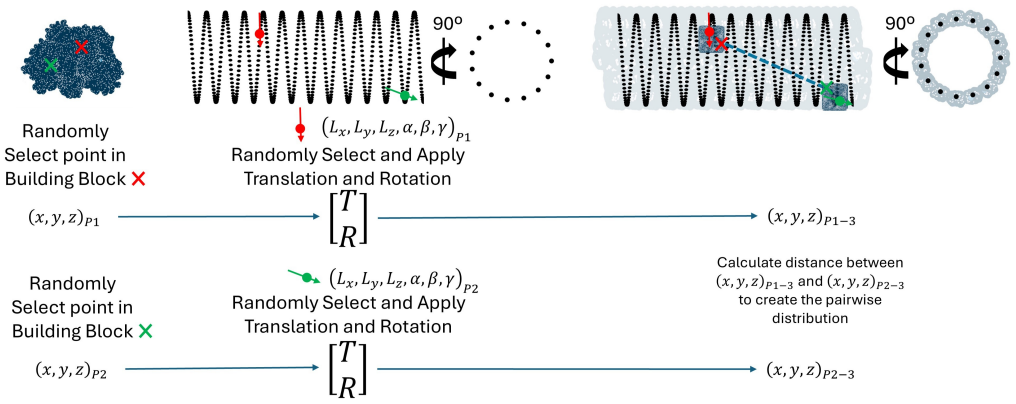



**Figure 5.3:** A convolution can be performed with the MC-DFM to avoid calculating the coordinates of the whole structure by randomly sampling the building block and the lattice coordinates independently. The randomly sampled building block coordinate is translated and rotated according to the randomly sampled lattice coordinate. This is repeated with another coordinate and the distance between the two points is calculated. This process is then repeated to create the pairwise distribution.

puter with 8GB RAM and a 12th Gen Intel Core i7, 2500 Mhz processor was used to simulate and fit the SAS data. Since the time and memory-intensive step of the MC-DFM is the random sampling of atomic coordinates and the calculation of the pairwise distances between them, our implementation of the algorithm performs these steps entirely using matrix operations which allows the calculation to be performed rapidly and efficiently. To achieve even more efficiency, the calculations can be integrated with a GPU using CUDA, leveraging its parallel processing capabilities, but this was not done in this work due to limitations in hardware availability. The input for the Python program in the simplest case is the 3D cartesian coordinates of the atoms of the basic protein sub-structure and each atom's scattering length density. Alternatively, the 3D cartesian coordinates of the building block, the lattice coordinates, which includes the amount of translation and rotation applied at each repeating subunit, and the scattering length density of each coordinate can also be used. The output is the scattering intensity as a function of $q$.

Our implementation of the MC-DFM is compared to CRYSOL [87] and D+ [99], which are software used to simulate the scattering curve of large biomolecular assemblies. The results are shown in the supplemental information, and it is clear that the MC-DFM outperforms the other methods in terms of computational time. Some disadvantages of our implementation of the MC-DFM is that it currently cannot model the scattering of the hydration layer surrounding the protein, cannot model polydispersity, and can only be run in python without a GUI.

### 5.3.4 Experimental methods

#### Small Angle X-ray Scattering

Experimental SAXS data was obtained on a Xenocs Xeuss 3.0 (Grenoble, France) instrument with an x-ray energy of 8.04 keV (wavelength 1.54 Å) using a copper K-$\alpha$ microfocus source. Data was collected in three configurations: low-q (0.004 - 0.007 $Å^{-1}$) for 3600s, mid-q (0.007 - 0.020 $Å^{-1}$) for 3600s, and high-q (0.020 - 0.200 $Å^{-1}$) for 2700s. Samples of 20 $\mu$L were loaded into a quartz capillary flow cell of 1.5 mm in diameter using the “Biocube" configuration of the instrument. Data reduction and merging were performed with the XSCAT software (Xenocs, Grenoble, France). Buffer subtraction was performed by subtracting the scattering of water from a previous measurement in the same capillary and configurations.

#### TEM Imaging

For the preparation of negative-stain TEM samples, 3-3.5$\mu$L aliquots of protein solutions were applied to negatively glow-discharged copper grids with Formvar/Carbon-coating (Ted Pella Inc.). The grids were washed with 50$\mu$L of filtered milliQ water and stained with 3.5$\mu$L of 2% filtered uranyl acetate solution at $4^{\circ}$C. TEM micrographs were collected using JEM-1400 plus transmission electron microscope (JEOL Ltd.) operating at 80 keV, equipped with a tungsten filament and bottom-mounted Gatan OneView camera (4k x 4k).

#### Cryo-EM Imaging

Cryo-EM sample preparation and image acquisition preparation of cryoEM grids was performed on an FEI Vitrobot Mark IV (ThermoFisher Scientific) at 95% humidity and $4^{\circ}$C. 3.5$\mu$L of 100$\mu$M protein solution was applied directly onto glow-discharged QuantiFoil Cu 2/1 300 grids (Electron Microscopy Sciences) without further dilution. The grids were blotted with a filter paper for 4-5 s before flash-freezing in liquid ethane. The grids were then clipped under liquid nitrogen and stored in liquid nitrogen until data collection. The grids were imaged at the S2C2 Stanford-SLAC CryoEM Center on TEM Beta (Titan Krios G3i (Thermo Fisher Scientific) equipped with a Gatan K3 direct electron detector) operating at 300 keV. Images were collected at 0.86 Å/pixel with a fluence exposure of $\sim$50 electrons/$Å^2$ at a dose rate of $\sim$1.5 electrons/$Å^2$ per frame. Defocus range of the objective lens was set between -1.5 to -2.1 $\mu$m.

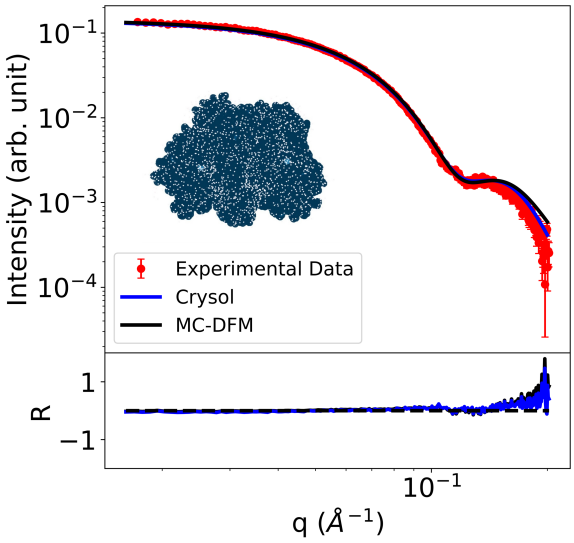



**Figure 5.4:** Experimental data from the RhuA monomer protein and the scattering curve simulated from its PDB file with the MC-DFM and Crysol. The scattering curves from both Crysol and the MC-DFM were simulated without the hydration layer and an aqueous background was assumed.

## 5.4 Results

### 5.4.1 RhuA Protein

Our MC-DFM implementation was first tested on the scattering of the RhuA protein (PDB: 1GT7) [97]. This protein is a C4-symmetric homotetramer of about 7nm x 7nm x 5nm in size, which we will use as a model building block protein in this work. By performing mutations on the four corners of the protein, intermolecular forces (e.g., host-guest interactions) can be implemented that allow the protein to assemble into 2D lattices of different shapes such as tubes or sheets [100]. To test the validity of the MC-DFM, we can simulate SAXS curves using the MC-DFM and compare them to experimentally obtained SAXS curves of the RhuA assemblies. In this section, we start with the unassembled RhuA building block. For simplicity, we will refer to the unassembled RhuA building block from now on as "RhuA monomer". First, small angle x-ray scattering data was collected (see Methods section) from an aqueous sample of 50 $\mu$M RhuA in 20 mM Tris. Then, the simulated scattering curve was obtained from the PDB file of the protein using both the MC-DFM from this work and Crysol [87] from the ATSAS software [101]. All 8,476 atoms from the protein were used for the calculation. As shown in Figure 5.4, there is very good agreement between the scattering of the models and the experimental data.

### 5.4.2 Polydisperse RhuA Tube-like Assemblies

The MC-DFM, combined with a traditional SAS fitting algorithm, can be used to obtain structural features of biomolecular assemblies. In this example, the MC-DFM and a genetic algorithm were used sequentially to obtain the size distribution of polydisperse tube-like assemblies. The biomolecular assembly was physically created by assembling RhuA monomers via host-guest interactions. Host-guest interactions refer to the noncovalent bonding between two distinct molecules, leading to the formation of unique complexes. In these complexes, guest molecules are typically encapsulated or incorporated within the host molecules, linking them together. This type of interaction is frequently used to create supramolecular systems because of its responsiveness to external stimuli like light, pH, temperature, etc. [26]. The host-guest complex used to assemble the RhuA protein was chosen to be the light-responsive azobenzene (guest) and $\beta$-cyclodextrin (host) complex. In short, $\beta$-cyclodextrin has a high affinity to (E)-azobenzene, but not to (Z)-azobenzene. The (E)-isomer is thermodynamically more stable than the (Z)-isomer, so the formation of the complex is the most stable form. Exposure to UV light triggers a conformational change from (E)-azobenzene to (Z)-azobenzene, leading to the separation of the complex, while visible light reverses this process [27]. In this work, the RhuA protein was first mutated to contain cysteine residues on each of its four corners. The protein was then separated into two identical batches. In the first batch, an azobenzene molecule was attached to each of the four corners of the RhuA protein by treating the solution with 4-phenylazomaleinanil dissolved in DMF, which takes advantage of the thiol-malemide cross-linking reaction. This resulted in a RhuA protein functionalized with an azobenzene group ($^{azo}$RhuA). In the second batch, a $\beta$-cyclodextrin molecule was attached to the four corners of the RhuA protein using a similar cross-linking strategy. This resulted in a RhuA protein functionalized with a $\beta$-cyclodextrin molecule ($^{\beta CD}$RhuA). The resulting proteins, $^{\beta CD}$RhuA and $^{azo}$RhuA, were then mixed in equal molar ratios to create 50 $\mu$M of RhuA in a 20 mM Tris and 100 mM KCl solution. Tube-like assemblies of different diameters and lengths were observed from cryo-EM and ns-TEM [76].

From Figure 5.5, the outer diameters of the tubes range from 40-120 nm, while the lengths are significantly longer than the maximum size that can typically be measured using SAXS ($\leq$ 300 nm). A histogram of tube diameters was created from the microscopy images and a normal distribution was fitted to the data. Given the range of tube diameters observed in TEM, several computational RhuA tube assembly models

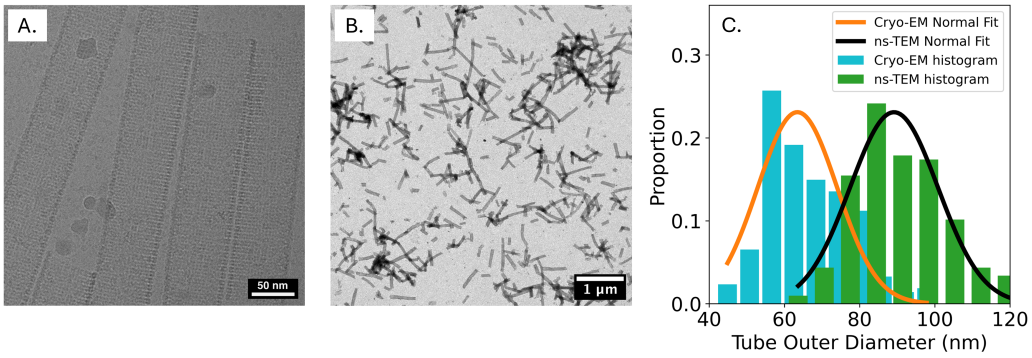



**Figure 5.5:** (A) Cryo-EM image showing a zoomed-in view of the tube-like assemblies of RhuA proteins. (B) ns-TEM image showing a zoomed-out view of the tube-like assemblies. (C) A histogram of the outer diameters of the tubes in the images shown in this figure, as well as from additional images that are not shown. The tubes are expected to be flattened due to drying effects for ns-TEM, which would make them appear larger than if they were imaged in their native state. The histogram data was fitted with a normal distribution. The mean of the distribution obtained from cryo-EM is 63.6 nm with a standard deviation of 10.6 nm. The mean of the distribution obtained from ns-TEM is 88.9 nm with a standard deviation of 11.7 nm.

of different diameters were created by placing the coordinates of the RhuA monomer building block from the PDB file in a helical tube-like structure. Monodisperse tube-like models with varying diameters ranging from 40-120 nm were created with a constant tube length of 240 nm using the process described in Figure 5.3. The spacing between adjacent proteins was always kept constant, and the diameter was increased by increasing the number of proteins in each type of tubular assembly. To increase the diameter of the tube, two additional proteins were added to each loop of the helix, which increased the outer diameter by approximately 6 nm. Starting with a tube of 15 proteins in a loop and 44 nm in diameter, this process was repeated until tubes of 35 proteins in a loop and 104 nm in diameter were created. Each RhuA monomer contains 8,476 atoms, and the smallest tube model of 44 nm in diameter contains around 450 monomers which results in around $N = 3.8$ million atomic coordinates for the tube model of the smallest diameter. The SAXS instrument used in this work has a measurable q-range of about 0.004 to 0.2 $^{-1}$, which is a range that is primarily sensitive to structural features of roughly less than about 150 nm. Therefore, the small-angle scattering curve is primarily sensitive to changes in the inter-protein organization and the cross-sectional size of the tubes, but it is not very sensitive to fluctuations in the tube length. In addition to tubes, unassembled monomers were also present in the images in Figure 5.5. Therefore, the SAXS models also accounted for free unassembled monomer proteins present in the samples. The MC-DFM was then used to simulate the

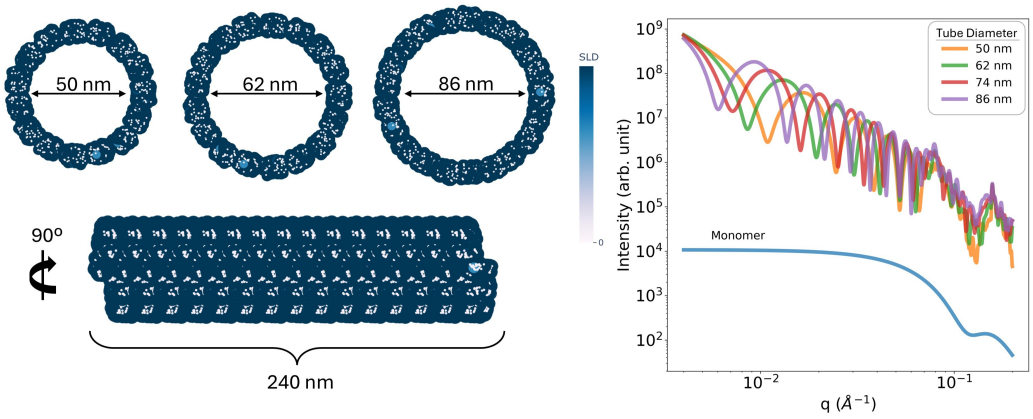



**Figure 5.6:** Models of tube-like assemblies of different outer diameters and the unassembled RhuA monomer were created and their scattering curves were simulated using the MC-DFM. Tube-like models with outer diameters of 44, 50, 56, 62, 68, 74, 80, 86, 92, 98, and 104 nm were simulated, but not all are shown in the figure.

weighted scattering curves of each tube and unassembled monomer as shown in Figure 5.6.

To quantitatively determine the distribution of tube diameters in the sample, the simulated scattering curves were first weighted so that their invariant would be linearly proportional to the number of RhuA monomers in each tube assembly, which is necessary to correctly scale the intensities of the simulated scattering curves with respect to the total protein fraction in any given model. This was done by dividing each simulated scattering curve by its calculated invariant. By definition, the invariant is a constant that is directly proportional to the number of scattering objects (i.e., electrons for SAXS) in the path of the X-ray beam and is independent of size, shape, or interactions between the scattering objects [81]. The invariant, $Q^*$, is defined in Equation (5.11).

$$Q^* = \int_0^\infty q^2 I(q)\, dq \tag{5.11}$$

This step of scaling each curve according to its invariant was crucial to obtain the correct population distribution of each structure in the sample. After this, the weighted average of all the curves was calculated as shown in Equation (5.12). The optimal weights were determined by the genetic algorithm that minimized the difference between the weighted average of the simulated curves and the experimental scattering curve, Equation (5.13). The optimal weights were then plotted to determine the size distribution of tube diameters in the sample. The distribution was then compared to the one found by Cryo-TEM and ns-TEM as shown in

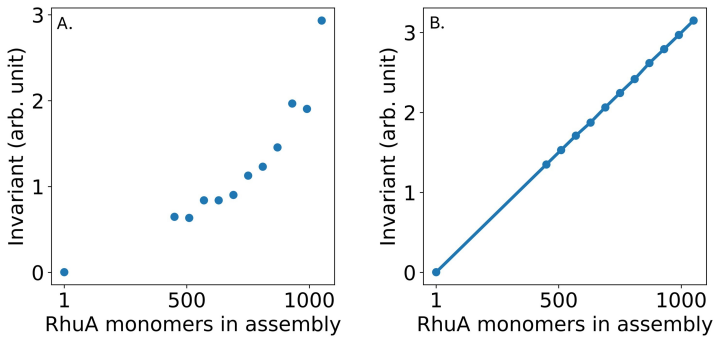



**Figure 5.7:** The invariant of the simulated SAXS curves as a function of the number of RhuA monomers in the tube-like assemblies. Each point represents the invariant of a single simulated SAXS curve of the tube-like assembly and was calculated with a Guinier extrapolation at low q and a fourth power law extrapolation at high q. (A) shows the invariant before scaling and (B) shows the scaled, linearly proportional, invariant of the simulated SAXS curves after diving them by their invariant calculated in (A).

Figure 5.5.

$$I(q)_{avg} = W_{monomer}I(q)_{monomer} + W_{44nm}I(q)_{44nm} + ... + W_{86nm}I(q)_{86nm} \tag{5.12}$$

$$Score = \sum_{q_{min}}^{q_{max}} |\log(I(q)_{avg}) - \log(I(q)_{exp})| \tag{5.13}$$

As shown in Figure 5.8, the weighted average of the simulated SAXS curves has a good agreement with the experimental scattering data of the RhuA tubes. As expected, the MC-DFM was able to simulate the high-resolution features of the structure shown by the inter-monomer correlation peaks at high-q. A histogram was then created from the optimal weights of the weighted average of the SAXS curves and was fit with a normal distribution for comparison with the ns-TEM data. The distribution obtained from SAXS data had a mean of 56.8 nm and a standard deviation of 6.3 nm. In comparison, the fit for the data obtained from cryo-EM had a mean of 63.6 nm with a standard deviation of 10.6 nm, which is close to the findings from SAXS. The fit from ns-TEM data had a mean of 88.9 nm and a standard deviation of 11.7 nm. From the data, it is clear that the mean of the normal distribution fit from ns-TEM data is significantly greater than that from SAXS and cryo-EM. A reason for this could be the flattening of the tubes that occurs when the sample is dried for ns-TEM characterization. This would cause the tubes to appear larger than they were in their true native hydrated state. However, the effect of flattening can be calculated, where the flattened diameter

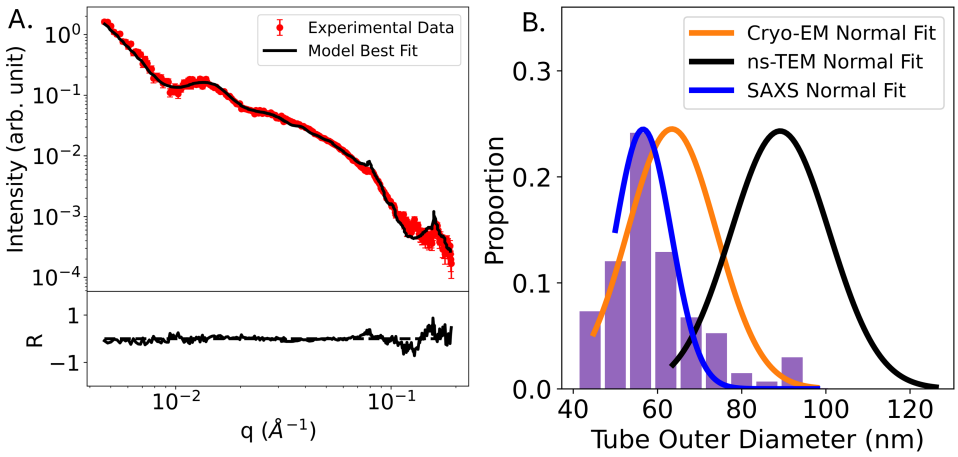



**Figure 5.8:** (A) SAXS fit of the RhuA tube-like assemblies and a histogram showing the distribution of tube diameters in the sample. (B) A normal distribution was used to fit the data and compared to the fits from cryo-EM and ns-TEM. The fit of the data obtained from SAXS had a mean of 56.8 nm and a standard deviation of 6.4 nm. The proportion of unassembled RhuA monomers is around 0.50.

should be half the circumference of the tube in its native hydrated state, which means that the diameters should differ by a factor of $2/\pi$. By applying this correction, the unflattened diameter from ns-TEM is 56.6 nm, extremely close to the diameter obtained from SAXS. The flattening effect should not affect the standard deviation of the diameter size distribution since all tubes get flattened equally. The standard deviations of the normal distribution fits from the microscopy characterization methods are in agreement, while the one found from SAXS is slightly lower. In this example, the MC-DFM was successfully used with a genetic algorithm to determine the size distribution of polydisperse biomolecular assemblies. We envision that this method can be used for other similar systems such as determining the size distributions of tobacco mosaic virus aggreagates [102], microtubules [103], or other 1D helical protein assemblies [104].

### 5.4.3 RhuA Sheets

The MC-DFM combined with a genetic algorithm can also be used in a closed-loop algorithm to determine the structural features of biomolecular assemblies, similar to a typical least-squares analytical model fitting approach. This was demonstrated by extracting structural features of 2D sheet-like assemblies made from RhuA monomers. This assembly was experimentally created using the same host-guest interactions used to create the tubes, but a salt with a higher valence was used. To create the sample, equimolar ra-

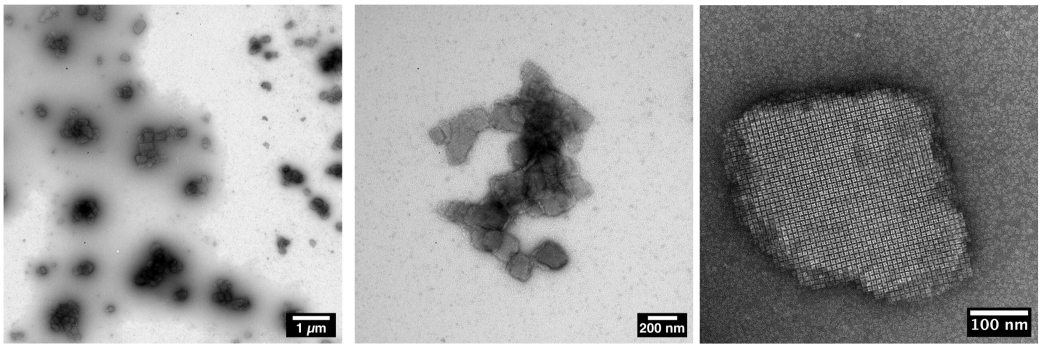



**Figure 5.9:** ns-TEM images of the RhuA sheet-like assemblies. The sheets seem to have rectangular shapes, and some of the sheets stack on top of each other, indicating that the height of the sheets is much smaller than the length and width.

tios of $^{\beta CD}$RhuA and $^{azo}$RhuA were mixed to create a solution of 50 $\mu$M RhuA in 20 mM Tris and 100 mM $CaCl_2$. As seen in Figure 5.9, sheet-like assemblies were observed in ns-TEM instead of tube-like assemblies because of the higher ionic strength of the solution [76].

In a recent publication by Zhang et al., it was discovered that the RhuA protein is negatively charged and forms tube-like assemblies at high pH and in the presence of monvalent salts. At low pH or in the presence of divalent salts, it assembles into sheets. This is because the screening of electrostatic forces allow the proteins to stack onto each other in the sheet formation. All atom molecular dynamics simulations indicated that the RhuA surface has high affinity to $Ca^{2+}$ ions, which would reduce the electrostatic forces on the surface of the protein, leading to the sheet assembly [76]. Furthermore, we do not believe the images of the sheets in Figure 5.9 are of collapsed tubes because they are not elongated like the tubes and do not have straight borders which are expected on the edges of the tubes. Instead, they are shaped more like squares with rough borders which is more likely to come from a stack of sheets.

From the observations made in Figure 5.9, computational models of the sheets were created by assembling the RhuA monomers in a 2D sheet-like structure. After simulating the scattering curves of models of flat sheets, it was noticed that the peak intensities, which corresponds to the spacing between the proteins in the sheet, were extremely sharp and not consistent with the experimental scattering data which has broad peaks. Because of this, we hypothesized that there could be some distortion in the spacing between the proteins in the sheets, which would broaden the peaks. Due to the large aspect ratio (i.e. length over thickness) of the sheets and the relatively weak physical bonds between the proteins, it was hypothesized that they

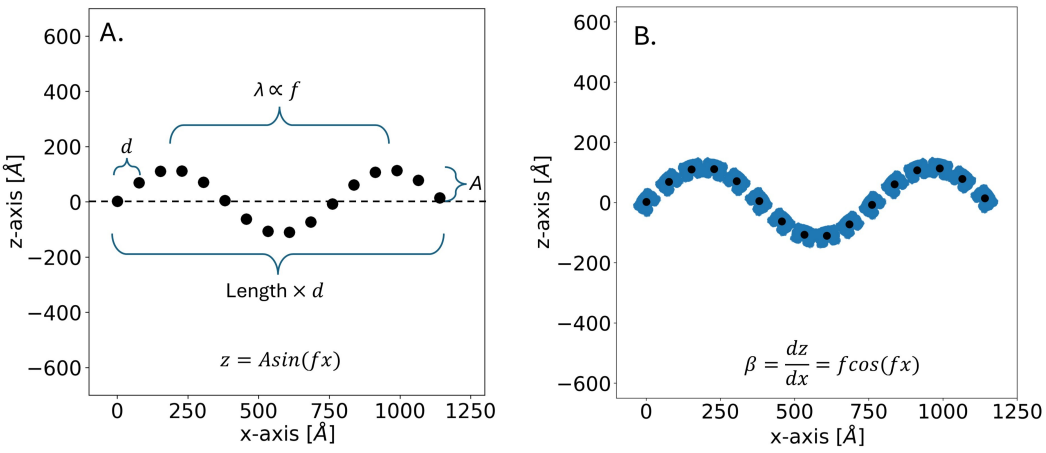



**Figure 5.10:** The steps to create the sheet model with the undulating wave-like structural effect. (A) First, the coordinates of the proteins are determined using Equation (5.14). (B) The relative rotation applied to each protein is determined using Equation (5.15). The sampled coordinates of the proteins can be translated and rotated to each position, creating the final assembly.

could experience undulating wave-like distortions in solution, so this structural effect was also included in the model. To create this model, Equation (5.14) was used, where $A$ is the wave's amplitude and $f$ is the wave's frequency. The $x$ values range from 0 to the length of the sheet and are equally spaced apart by $d$, the spacing between the proteins.

$$z = A\sin(fx) \tag{5.14}$$

Each protein is then rotated along the y-axis so that they are oriented side-by-side as shown in Figure 5.10 (B). Equation (5.15) is used to find the precise rotation needed for each protein.

$$\beta = f\cos(fx) \tag{5.15}$$

To create a protein sheet-like assembly of arbitrary length, width, and height, the lattice coordinates (translational and rotational transformations) created in Figure 5.10 (A) can be duplicated and translated in any direction as shown in Figure 5.11. Finally, adjacent proteins in the length and width direction are rotated by 180 degrees to recreate the alternating checker-board pattern, previously observed in AFM and TEM, in the final assembly [98].

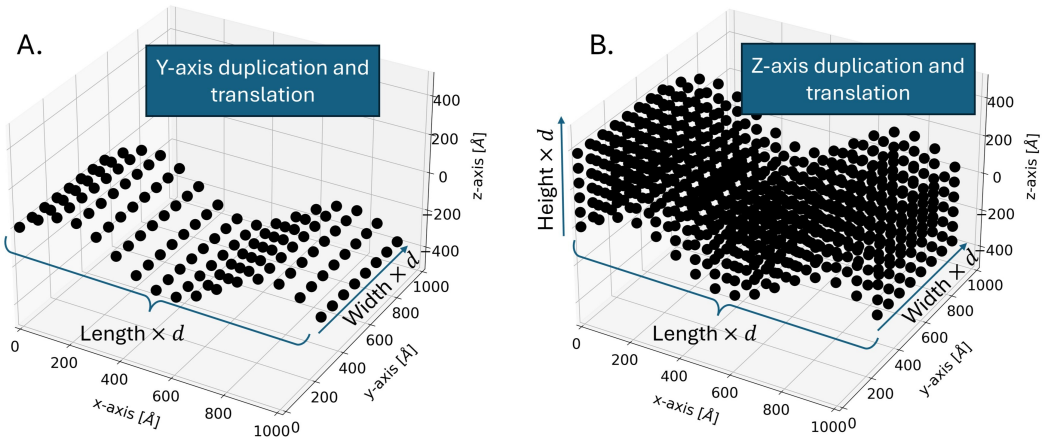



**Figure 5.11:** The lattice coordinates can be duplicated and translated in the width (y-axis) and height (z-axis) directions to create protein sheet-like assemblies of any length, width, or height (defined by the number of proteins in each size direction).

The values of the structural parameters used to create the sheet-like model were determined by fitting the simulated scattering data of the model to the experimental one. The fit parameters were the separation distance between the proteins in the width and length direction $d_{wl}$ (ranging from 7.5-8.5 nm), the separation distance between the proteins in the height direction $d_h$ (ranging from 5.5-6.5 nm), the number of proteins making up the sheet's length (ranging from 5-30), the sheet's width (ranging from 5-30), and the sheet's height (ranging from 1-6). The sheet's length, width, and height parameters were controlled by changing the number of monomers in the respective size dimension. Parameters that describe the undulative wave-like structural effect were also determined from the fit. The frequency of the wave $f$ ranged from 0 to 0.01, corresponding to a completely flat structure to a full wave cycle occurring every 60 nm. The amplitude $A$ ranged from 0 to 60 nm.

The presence of unassembled RhuA monomers was accounted for by the RhuA monomer binary volume fraction parameter, $\sigma$, (ranging from 0-1). The scattering curve of the sample which contains unassembled RhuA monomers and assembled RhuA sheets was found by adding the weighted scattering curves of the two structures. To calculate this, the RhuA monomer volume fraction parameter ($\sigma$) was multiplied by the scattering curve of the monomer, shown in Figure 5.4, and then added to the scattering curve of the sheets multiplied by the volume fraction of the sheets ($1 - \sigma$). The scattering intensities of the RhuA monomer and the sheets were correctly scaled by dividing each curve by its invariant. Finally, the scale parameter

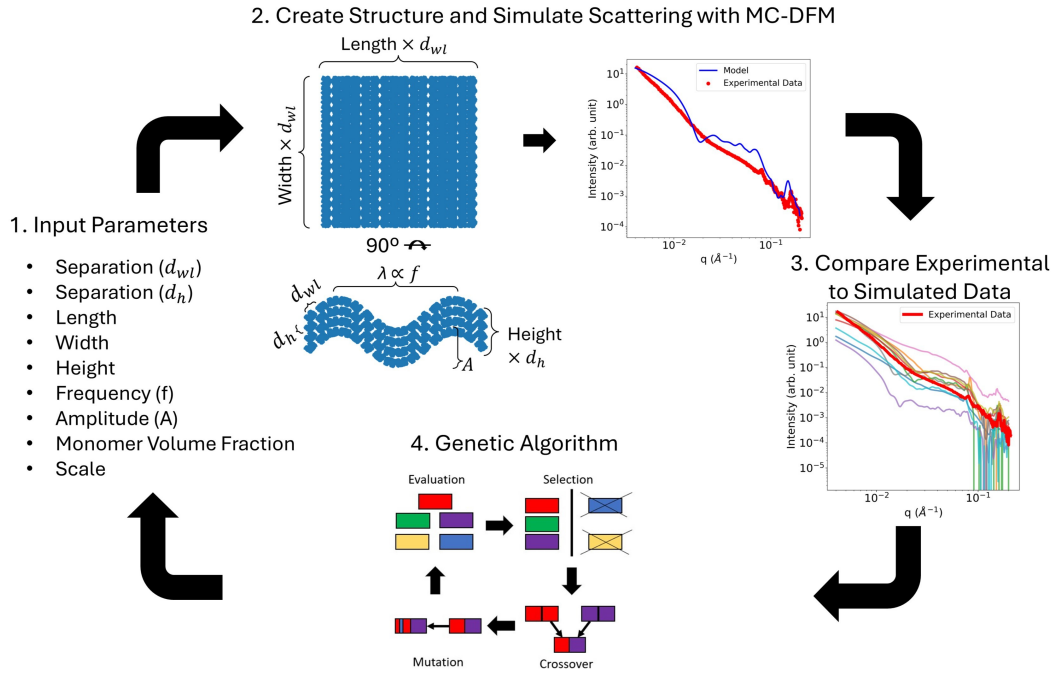



**Figure 5.12:** The closed-loop optimization workflow used to determine the structural features of the sheet model. Input parameters were used to create a model of the sheet-like assembly. The scattering curve of the assembly was then calculated with the MC-DFM and compared to the experimental scattering curve obtaining a distance score. Finally, the distance score and the input parameters were sent to a genetic algorithm to determine the next set of input parameters to test. The genetic algorithm ran for 20 iterations with a batch size of 30 samples, and took around 27 minutes to complete.

(ranging from $10^{-5}$ to $10^{-9}$) was used to determine the concentration of sheets in the sample. Since the concentration of sheets in the sample is unknown because of sedimentation and because the data was not collected on an absolute scale, the scale parameter was only used to align the intensities of the experimental and simulated scattering curves. The equation that combines the scattering of the monomer and the sheet is shown in (5.16).

$$I(q)_{sample} = scale\left[\frac{(\sigma)I(q)_{monomer}}{invariant(I(q)_{monomer})} + \frac{(1-\sigma)I(q)_{sheets}}{invariant(I(q)_{sheets})}\right] \tag{5.16}$$

The optimization workflow illustrated in Figure 5.12 was used to identify the optimal structural parameters to create the sheet model whose scattering curve best matched the experimentally obtained scattering curve. This was done using the genetic algorithm to minimize the difference between experimental and simulated scattering curves.

As shown in Figure 5.13, our closed-loop workflow was able to obtain a reasonable fit to the experimental data. The fit at high-q (0.03-0.2 $^{-1}$) is adequate since the correlation peaks in the simulated curve are almost identical to the ones in the experimental curve, in terms of position and intensity. The discrepancy between the peak intensities could be attributed to instrumental smearing and to thermal fluctuations that are unaccounted for in the model. From the fit at high-q, the separation distance of the proteins was determined to be 7.6 nm in the width and length direction and 6.0 nm in the height direction. This is consistent with the known size of the RhuA protein of 7 nm in width and length and 5 nm in height [100]. The parameters that describe the undulating wave-like effect were an amplitude of 39 nm and a complete wavelength every 240 nm. These values are reasonable, but hard to verify with microscopy. The RhuA monomer volume fraction of 0.44 is also hard to verify with microscopy, but is consistent with the volume fraction found in the previous section when fitting the RhuA tube-assembly data. In the low-q region (0.004 - 0.03 $^{-1}$), there is a disagreement between the slopes of the experimental and simulated curves. The low-q region in SAXS is sensitive to the sheet size, highlighting a discrepancy between the length and width values obtained from the fit compared to the ones suggested by the SAXS data. An explanation for this could be polydispersity in sheet size as observed in the microscopy images in Figure 5.9, and possibly to defects in the sheet. The microscopy images show a wide distribution of sheet sizes, many of which are large ($\geq$ 200nm). Our model assumed monodisperse sheets, and therefore, it is unlikely that the best-fit parameters for a single sheet of specific length and width are a good representation of the sizes found in an ensemble of sheets. This also explains why quantitative sizes obtained from SAXS fits are not consistent with sheet sizes observed from microscopy in Figure 5.9. Implementing polydispersity would further improve fits and lead to a more realistic sheet size distribution, which will be attempted in future work.

This example demonstrates the power of the MC-DFM combined with a fitting algorithm to extract structural features of large biomolecular assemblies. This was only possible due to the fast and scalable characteristics of the MC-DFM. We envision that this method can be integrated with other optimization routines and can also be applied to other systems such as 2D protein nanosheets [105], peptide assemblies in honeycomb lattices [106], or 3D protein crystals [107] [108].

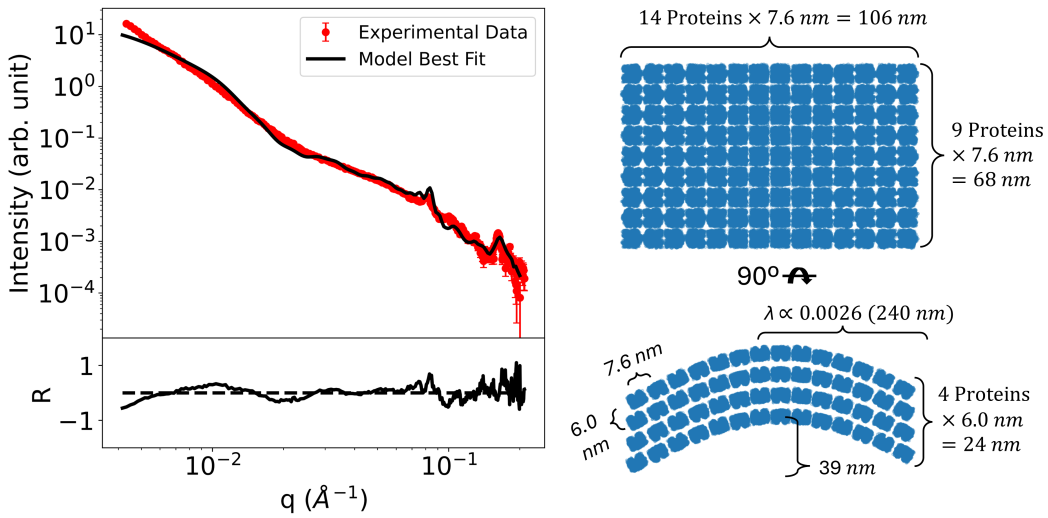



**Figure 5.13:** The results of the optimization show the model curve that best fits the experimental data. The structural parameters of the model curve were a protein separation distance of 7.6 nm in the length and width direction, a protein separation distance of 6.0 nm in the height direction, a sheet length of 14 proteins (106 nm), a sheet width of 9 proteins (68 nm), a sheet height of 4 proteins (24 nm), an undulating wave-like amplitude of 39 nm, an undulating wave-like frequency of one wavelength every 240 nm, and an unassembled RhuA monomer volume fraction of 0.44.

### 5.4.4 SAS Simulations from HOOMD-blue

Another potential application for the MC-DFM is in simulating SAS curves from molecular simulations. This is especially useful to enable direct comparisons with experimental data by matching simulated and experimental scattering profiles, which could quantify how accurate a simulation is compared to reality. To demonstrate this ability, we use HOOMD-blue [109], a powerful simulation toolkit, to simulate the assembly of spheres and cubes. Because the objective of this section is to simulate scattering profiles, the molecular simulations we used were designed to be convenient, and may not realistically reflect the true physics of colloidal assembly.

For the sphere assembly, molecular dynamics simulations were conducted using HOOMD-blue to study the self-assembly of 5000 monodisperse spherical particles with diameter $\sigma = 1.0$ angstrom at a number density of 0.3. Particles were initialized in non-overlapping random positions within a cubic simulation box of edge length $L = (N/\rho)^{1/3}$. Interactions were modeled using a truncated Lennard-Jones potential with $\epsilon = 3.0$, $\sigma = 1.0$, and a cutoff of $2.5\sigma$. The system was evolved in the NVT ensemble using a velocity-Verlet integrator with timestep $\Delta t = 0.002$, and a Bussi thermostat maintained temperature at $k_B T = 1.0$.

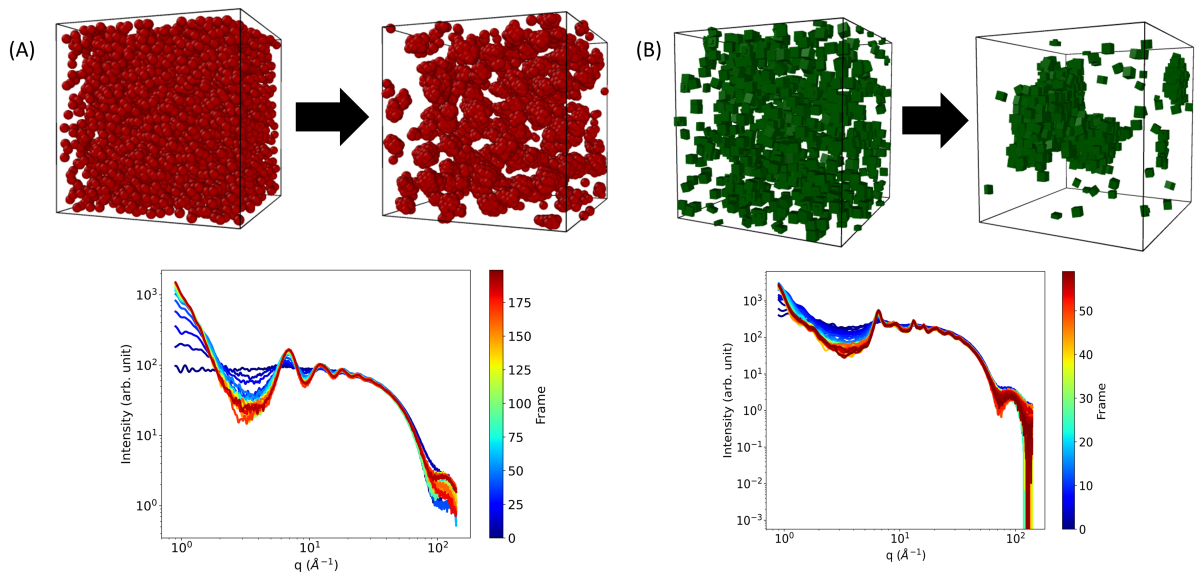



**Figure 5.14:** (A) the simulation of the assembly of spheres and the SAXS curve from the simulation showing the change in structure as the assembly occurs. The spheres were assumed to be core shell spheres with a core diameter of 0.1 angstrom and a shell thickness of 0.4 angstrom. (B) the simulation of the cube assembly as well as the SAXS curves from the simulation. The cubes were assumed to be core shell cubes with an inner cube edge length of 0.1 angstrom and an outer cube length of 1 angstrom.

Particle momenta were initialized thermally, and trajectories were recorded every 200 timesteps over a total of 20,000 steps.

The rigid cube assembly was also performed using HOOMD-blue. Each cube was modeled as a rigid body composed of eight Lennard-Jones (LJ) spheres (Particle "`B`") arranged in a $2 \times 2 \times 2$ configuration, centered on a parent particle (Particle "`A`"). The constituent particles had a diameter of 0.3 units and were placed at $\pm 0.25$ units along each axis, yielding a cube edge length of 0.5 angstrom. A total of 800 cubes were initialized at random non-overlapping positions and orientations in a periodic box sized to achieve a number density of 0.7. Rigid bodies were defined using HOOMD's `md.constrain.Rigid` constraint, with parent particles of type "`A`" and LJ interactions applied only between constituent particles of type "`B`". The LJ potential used parameters $\varepsilon = 3.0$, $\sigma = 0.45$, and a cutoff of $r_{\text{cut}} = 3.0\sigma$. Langevin dynamics with $\Delta t = 0.002$ was used to integrate the motion of rigid bodies with rotational degrees of freedom enabled. Trajectories were recorded every 60 timesteps over a total of 20,000 steps.

Figure 5.14 shows molecular dynamics simulation snapshots of the assembly of the spheres and cubes as well as the corresponding scattering profiles at each timestep. The simulated scattering curves seem reasonable since they both contain an increase in slope at low-q which is consistent with the formation of large assemblies. Meanwhile, correlation peaks emerge which suggests the presence of a constant interparticle spacing between the particles in the assembly. The advantage of using the MC-DFM for this simulation that it calculates the intensity, which is the product of the form factor and structure factor. This is advantageous because it means that the scattering curve is directly comparable to data obtained from experiments. Another advantage is that it can be used for objects of any shape or size, which is advantageous when dealing with objects that have form factors that are too complex to derive analytically, such as biomolecules like proteins.

## 5.5 Conclusion

The MC-DFM is a well-established method used to simulate the scattering curves of large protein assemblies and structures. In this work, we investigated its effectiveness in simulating the scattering curves of large biomolecular assemblies with periodic structures, including tubes and sheets composed of several hundreds of protein sub-units. Due to the accelerated speed and scalability, we successfully combined it with a genetic algorithm to extract structural features from experimental scattering curves of large host-guest protein assemblies. We first demonstrated this method on a 1D tube-like assembly where we determined the size distribution of tube diameters in the sample. Next, we used the MC-DFM together with a genetic algorithm in a closed loop to fit data from another sample of 2D sheet-like assemblies. With our method, we quantitatively determined several structural features such as the protein separation distance and the height of the sheet-like assemblies. As expected, the structural features obtained from SAXS were consistent with observations from microscopy images. The implementation of the MC-DFM within a fitting algorithm was feasible primarily because the MC-DFM was efficiently coded using matrix operations, making it highly suitable for handling large structures. Additionally, our strategy for modeling large biomolecular assemblies involves using a single building block and applying a series of rigid-body transformations to generate the complete structure. This method offers a significant advantage, as adjustments to the assembly's structure (e.g., altering tube radii) can be rapidly achieved by modifying only the transformations. Consequently,

explicit calculation of all atoms in the large structure is avoided, greatly reducing computational time. We imagine that the method of using the MC-DFM and a genetic algorithm can be used to extract structural features from the scattering curves of other large biomolecular assemblies.

# Chapter 6

# Assembly of Silica Nanoparticles using Physically Tethered DNA Bonds

This work presented in this chapter has contributions from:

- Huat Thart Chiang, Naomi Kern, Zachery R. Wylie, Abdul Moeez, Haoqing Zhang, Daniel McKeen, Nicholas Herringer, Oleg Gang, Andrew Ferguson, Zachary Sherman, Lilo D. Pozzo

## 6.1 Abstract

Single-stranded DNA molecules modified with cholesterol functional groups are physically tethered to silica nanoparticles (Diameter 25 nm) engulfed in a lipid bilayer, which increases DNA mobility over the surface of the nanoparticle, and enables generalized bonding without depending on a specific surface chemistry (e.g. silane or thiol ligands). Nanoparticles are first encapsulated in lipid bilayers using ultrasound treatment of dispersions in the presence of lipids. The formation of the desired structure is confirmed with dynamic light scattering (DLS), small angle x-ray scattering (SAXS), and small angle neutron scattering (SANS) measurements. To induce assembly of nanoparticles coated with lipid bilayers, single-stranded DNA modified with cholesterol functional groups are first added followed by NaCl to reduce electrostatic repulsion, allowing for a higher grafting density of DNA on the surface of the nanoparticles. Double-stranded DNA 'bridge' molecules are then added with complementary nucleotides to the DNA 'anchor' molecules that are physi-

cally grafted to the lipids on the surface of the particles. Assembly is observed to occur at room temperature and without the need for temperature annealing. Using automated liquid handling tools, assemblies are created in high throughput and rapidly characterized using SAXS to screen the impact of design variables. It is determined that the relative concentration of DNA-to-silica and the ionic strength of the solution are important parameters outlining the resulting assembly. Analysis of SAXS is performed using coarse-grained particle dynamics simulations as a function of the interparticle potential. The results support the spontaneous formation of semi-crystalline particle assemblies by particle condensation, where the interparticle distance is tuned by the sequence of the DNA 'bridge' used to link the particles. Crystallinity analysis performed on the resulting simulations, optimized to match SAXS observations, suggest that particle clusters display increased crystallinity in the center of the clusters, but their maximum size remains relatively small (100's of nm) before settling occurs, which limits the extent of crystallization.

## 6.2 Introduction

DNA-mediated assembly has emerged as a powerful strategy for organizing colloidal particles into ordered structures, owing to the high degree of programmability and tunability inherent in DNA molecules [49]. This programmability arises from the vast number of possible nucleotide base pair combinations, which can be precisely engineered to control key parameters such as strand length and melting temperature. In the literature, the most established and common approach utilizes single-stranded DNA molecules functionalized with thiol groups to chemically bind to gold nanoparticles. These DNA functionalized nanoparticles can then be programmed with 'sticky ends' which are short, single-stranded DNA molecules designed to hybridize selectively with complementary sequences on neighboring particles. After thermal annealing, which involves heating to the DNA melting temperature followed by controlled slow cooling, these sticky ends repeatedly hybridize and dehybridize, leading to the reorganization of nanoparticles into highly ordered crystal superlattices, as demonstrated in several foundational studies [110] [111].

While gold-thiol chemistry is a well-established method for anchoring DNA to nanoparticle surfaces, the resulting covalent bonds render the DNA immobile. This rigidity can hinder structural rearrangements during the assembly process. In contrast, anchoring DNA-cholesterol conjugates within a fluid lipid bilayer via hydrophobic interactions offers an alternative strategy, which has also been well documented in

the literature but for colloids of larger sizes [112] [113]. The bilayer's inherent fluidity allows DNA strands to diffuse laterally across the nanoparticle surface, enabling real-time reconfiguration of interparticle connections and potentially enhancing the formation of ordered superlattices. Given this distinct difference in surface dynamics, it is important to assess the functional impact of fluid bilayers in the context of DNA-mediated nanoparticle assembly. In this work, we investigate how this mode of DNA attachment compares to conventional gold-thiol systems in assembling three-dimensional colloidal superlattices [114].

There are several possible advantages of DNA surface mobility in forming colloidal superlattices. For example, mechanical flexibility at the nanoscale has been increasingly recognized as a crucial factor in promoting crystal formation. In our system, the expected increase in lateral mobility of DNA on fluid-supported nanoparticles introduces an added degree of freedom that enables the particles to adjust their positions relative to one another. This dynamic behavior could facilitate assembly by helping the system escape kinetic traps and resolve local mismatches that would otherwise stall the assembly process [112]. Furthermore, particle dynamics simulations have demonstrated that bond mobility can give rise to a rich variety of assembled structures such as aperiodic arrangements, shape-diverse tilings, and open porous frameworks that are typically inaccessible in systems with fixed DNA linkers [115]. Despite these insights, experimental investigations into mobile DNA bonds on nanoparticles remain scarce. This work begins to address this gap by exploring how physically-bound DNA with increased surface mobility influences the assembly dynamics and structural outcomes of DNA-programmed colloidal assembly at the nanoscale.

## 6.3 Materials and Methods

### 6.3.1 Chemicals

LUDOX TM-50 Colloidal Silica (50% weight) and Sodium Chloride ($\geq$ 99%) were purchased from Sigma Aldrich (St. Louis, MO, USA) and used as received. 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) ($\geq$ 99%) was purchased from Avanti Polar Lipids (Alabaster, Alabama) and used as recieved. Deionized water was obtained from a Direct-Q 3 UV water purification system with a resistivity of 18.2 M$\Omega$ (Millipore Corporation, Bedford, MA, USA).

### 6.3.2 DNA

All DNA was ordered from Integrated DNA Technologies (IDT) (Coralville, IA, USA). The sequences are defined in the (5' $\rightarrow$ 3') direction. Strands modified with cholesterol were ordered with HPLC purification and all other stands were ordered with standard desalting purification. All DNA was ordered in the LabReady ($100\mu M$ in IDTE, pH 8.0) formulation and then diluted to $60\mu M$ for the DNA-Chol and 50 $\mu M$ for the Bridge in sterile DNase free water. The nucleotides in red indicate the "sticky ends".

- DNA-Chol (10 PolyT, 18 HBP): TAT GAA GTG ATG GAT GAT TTT TTT TTT T/3CholTEG/
- DNA-Brigde (1) (20 Bridge, 18 HBP): ATC ATC CAT CAC TTC ATA ACT GAG CAG CAC TGA CAG CA
- DNA-Brigde (2) (20 Bridge, 18 HBP): ATC ATC CAT CAC TTC ATA TGC TGT CAG TGC TGC TCA GT
- DNA-Bridge (1) (40 Bridge, 18 HBP): ATC ATC CAT CAC TTC ATA TCA ACC TGA GTA TAA TTG TTA CTG AGC AGC ACT GAC AGC A
- DNA-Bridge (2) (40 Bridge, 18 HBP): ATC ATC CAT CAC TTC ATA TGC TGT CAG TGC TGC TCA GTA ACA ATT ATA CTC AGG TTG A
- DNA-Bridge (1) (80 Bridge, 18 HBP): ATC ATC CAT CAC TTC ATA GTA TTC CAT AGG TCG CCT CGC TTT GCA GAT AAG CTA ACT ATC AAC CTG AGT ATA ATT GTT ACT GAG CAG CAC TGA CAG CA
- DNA-Bridge (2) (80 Bridge, 18 HBP): ATC ATC CAT CAC TTC ATA TGC TGT CAG TGC TGC TCA GTA ACA ATT ATA CTC AGG TTG ATA GTT AGC TTA TCT GCA AAG CGA GGC GAC CTA TGG AAT AC

### 6.3.3 Dynamic Light Scattering

DLS measurements were taken on a Malvern PANalytical (Malvern, United Kingdom) Zetasizer Nano ZS in a polystyrene cuvette. The wavelength of the laser was 633 nm. The refractive index of Silica and lipids

was assumed to be 1.44 [116] [117] and that of the water solvent was 1.33. To analyze the data, the General Purpose algorithm provided by the Zetasizer software was used.

### 6.3.4 Small Angle X-ray Scattering

Experimental SAXS data was obtained on a Xenocs Xeuss 3.0 (Grenoble, France) instrument with an x-ray energy of 8.04 keV (wavelength 1.54 Å) using a copper K-$\alpha$ microfocus source. Data was collected in two configurations: low-q (0.003 - 0.007 $Å^{-1}$) using the ESAXS standard setting for 1200s and mid-q (0.007 - 0.10 $Å^{-1}$) for 600s using the SAXS high intensity setting. The samples had a volume of 200 $\mu$L and were created in a specially designed wellplate with 48 wells. The wellplate was designed so that it would collect any sedimentation from samples in a small pocket at the bottom of the wells. The X-ray beam can then be positioned on this small pocket to obtain the scattering curve of the sediment. Design files to create the wellplate can be found at (`https://github.com/pozzo-research-group/Automation-Hardware/tree/master/Cartridge%20Sample%20Holder%20for%20SAS%20Experiments/SAXS-USAXS%20Liquids%2048%20well%20plate%20holder`). Background reduction was performed with the XSCAT software by subtracting the scattering of water in the same wellplate and the data was merged automatically using a Python code.

### 6.3.5 Small Angle Neutron Scattering

Experimental SANS data was obtained on the Bio-SANS instrument at the High Flux Isotope Reactor at Oak Ridge National Laboratory. Samples were placed in a 2 mm thick Hellma (Banjo) cell with a volume of 600 $\mu L$, and a sample exposure time of 1 hour was used. A dual detector system composed of a curved wing detector at 1.13 m and a flat main detector at 15.5 m provided a continuous range of momentum transfer values q from (0.003 - 0.8 $Å^{-1}$). The neutron wavelength ($\lambda$) was 6 Å with a distribution ($\delta\lambda/\lambda$) of 13.2%. Data reduction was performed with the DRTSANS reduction software provided by the facility.

### 6.3.6 Electron Microscopy

Scanning electron microscopy and energy dispersive spectroscopy (SEM-EDS) were performed on samples using a Thermo Fischer Phenom Pharos G2 SEM-STEM. Samples were prepared by the drop casting of

diluted nanoparticles onto carbon film, 200 mesh nickel grids (Electron Microscopy Sciences; Hatfield, PA). Grids were placed onto a circular piece of filter paper, carbon side up, and 10 $\mu$L of the dilute solution was dropped onto the grid and allowed to dry.

### 6.3.7 Liquid Handling

The pipetting to mix samples in the wellplate for SAXS characterization was performed by an Opentrons (Brooklyn, NY, USA) OT2 liquid handling robot. The OT2 control code can be found at (`https://github.com/pozzo-research-group/OT2-DOE`).

### 6.3.8 Sonication

Sonication was performed with a Branson 450 digital sonifier using a probe of 3/4" diameter. Samples were sonicated in 20 mL scintillation vials for 1 minute at a power of 10% with a pulse interval of 10s (5s on/off). Immediately after this first sonication, they were resonicated for 5 minutes at a power of 30% with a pulse interval of 10s (5s on/off).

### 6.3.9 Lipid Encapsulation with DOPC

To encapsulate the silica nanoparticles in a lipid bilayer, the total surface area of the nanoparticles was first calculated. Based on this calculation, the required amount of DOPC to completely coat their surface was determined. The first step was to create a silica nanoparticle stock solution of 60 mg/mL silica nanoparticles in water using the Ludox TM-50 product, which has a silica weight percent of 50. A DOPC stock solution was also created containing 50 mg/mL DOPC in chloroform. The sample of silica nanoparticles encapsulated in DOPC was created in a 20 mL scintillation vial. First, 142 $\mu L$ of the 50 mg/mL DOPC stock solution (7 mg of DOPC) was added to the vial. The uncapped vial was placed in a vacuum oven under vacuum at room temperature for an hour in order for the chloroform to evaporate. After this, 333 $\mu L$ of the 60 mg/mL silica stock solution (20 mg of silica) was added as well as 9666 $\mu L$ of deionized water. The final step was to sonicate the sample using the sonicator horn with the settings described in the sonication section.

### 6.3.10 Particle Simulations

Particle dynamics simulations were performed using HOOMD-blue (v2.9.4) [118] to model the self-assembly of DNA-functionalized nanoparticles under Langevin dynamics. The simulation setup closely follows the implicit interaction model described by Mao et al. [119], in which DNA-mediated attractions are represented as short-range isotropic pair potentials between spherical particles. In their work, the pair potenitals are validated using the potential of mean force calculated from a more realistic coarse grained simulation of the DNA-mediated assembly of nanoparticles. Our simulations consisted of $N = 5000$ spherical particles of diameter $\sigma = 1.0$, randomly distributed in a cubic simulation box with periodic boundary conditions. The box length $L$ was computed based on the target number density $\rho = N/V$, with the initial configuration generated by randomly placing particles with a minimum center-to-center distance of $1.1\sigma$ to prevent overlaps. The interparticle interactions were modeled using a modified Lennard-Jones potential of the form:

$$U(r) = \frac{U_0}{n-m}\left[m\left(\frac{r_0}{r}\right)^n - n\left(\frac{r_0}{r}\right)^m\right] \tag{6.1}$$

Where $U_0$ is the interaction strength in units of kT, $r$ is the interaction distance in units of $\sigma$, $r_0$ is the distance where the interaction is minimum in units of $\sigma$, and $n$ and $m$ define the 'steepness' of the repulsive and attractive components, respectively. The potential was tabulated between $r_{\text{min}} = 0.75r_0$ and $r_{\text{max}} = 5.0$ using 1000 linearly spaced points and implemented via HOOMD's `pair.table` module. All potential values were pre-validated to avoid non-finite values that could disrupt force evaluations. The system was simulated under Langevin dynamics using a time step $\Delta t = 0.001$ (reduced units) and thermal energy $k_BT = 1.0$. Simulations were run for $1.5 \times 10^7$ time steps. System snapshots were recorded every 1000 steps using GSD trajectory output. In addition, the potential energy of the system was visualized and saved to confirm that the system reached equilibrium.

Simulated scattering curves were calculated from particle configurations of the last six saved snapshots of the simulation evenly spaced apart by 50,000 timesteps using a Monte Carlo sampling of the Debye equation [77], and averaged to obtain a single simulated scattering curve of the system at equilibrium. The simulated curve was converted into a structure factor and compared to the experimental structure factor using the Amplitude-Phase similarity metric [39] to obtain a similartiy score. The score was then used to

inform a Bayesian optimization algorithm (botorch v0.10.0) [120] on simulation parameters for the next iteration. A total of 100 iterations were run for each optimization.

## 6.4 Results and Discussion

### 6.4.1 Overview of Assembly Strategy

The method presented in this work to assemble silica nanoparticles involves first encapsulating them in a lipid bilayer composed of DOPC. Next, single-stranded DNA modified with a cholesterol molecule on the 3' end (DNA-Chol) is added to the system. The cholesterol embeds itself into the bilayer due to hydrophobic interactions, while the DNA forms a corona on the lipid bilayer. Because the lipids used are unsaturated, the bilayer is expected to be fluid at room temperature, allowing for DNA mobility on the surface of the bilayer. To induce assembly, a double-stranded DNA molecule (DNA-Bridge) is added. The double-stranded DNA has single-stranded sticky ends on each side which are complementary to the single-stranded DNA-Chol molecule previously embedded into the bilayer. More detailed discussions on the choice of nanoparticles, lipids, and DNA sequences are presented in the following sections.

### 6.4.2 Choice of Nanoparticle

Using the method described in this work, it should be possible to assemble many different types of nanoparticles of variable shape, size, and composition, provided that they can be encapsulated within the lipid bilayer structure. To demonstrate this concept, we chose to assemble monodisperse silica nanoparticles of 25 nm in diameter. This material was chosen for several reasons. The first reason is that there is high electron density contrast between silica and water, which facilitates SAXS measurements on the assembled structure, and there is also good availability at high particle concentrations, variable particle sizes, and with good monodispersity. This contrasts to traditionally used gold nanoparticles, where typical particle concentrations are many orders of magnitude lower. A second reason is that monodisperse silica nanoparticles can be inexpensively synthesized using the Stöber method [121]. Good monodispersty is essential for the formation of crystals since disorder can propagate and result in a loss of periodicity. Ludox-TM50 meets these requirements and is selected for this experiment.

### 6.4.3 Choice of Lipid

Several criteria must be met to select the appropriate lipid for assembling the silica nanoparticles. The first criterion is that the lipid must be able to encapsulate the nanoparticles with a bilayer. This means that the lipid heads must have a strong adsorption affinity to negatively charged silica nanoparticles. The second criterion is that the lipids must not be cationic which would cause the adsorption of negatively charged DNA molecules on the surface of the bilayer, preventing them from being used to induce assembly. Finally, to achieve assembly without thermal annealing, the lipid must have a low melting point to allow for a fluid bilayer to form. A lipid that satisfies all criteria is DOPC, which is chosen as the primary lipid in this work. DOPC has a critical packing parameter (CPP) of about 0.66 which results in lamellar self-assembled structures [122]. Despite being zwitterionic, it also has a strong affinity towards silica nanoparticles, which is well reported in the literature [123] [124] [125], and hypothesized to originate from hydrogen bonding, ion-dipole, dipole-dipole, and van der Waals interactions [126]. It is also unsaturated, with a gel phase melting point of -20°C [127]. Thus, a fluid bilayer should form at room temperature, allowing for increased mobility of the 'anchored' DNA on the surface.

### 6.4.4 DNA Sequence

The DNA sequence is also crucial for forming the desired type of crystal. Factors such as the 'sticky end' binding energy and the total length of nucleotides are common design variables. In this work, we first functionalize the lipid-encapsulated silica nanoparticles with single-stranded DNA which is modified with a cholesterol molecule on the 3' end (referred to as DNA-Chol). Separately, two single-stranded DNA molecules are designed and mixed to hybridize forming a DNA 'bridge' with double-stranded DNA in the center with unhybridized single-stranded ends (referred to as DNA-Bridge). These single-stranded ends are complementary to the 'sticky ends' from the DNA-Chol, which is used to functionalize the nanoparticles. Assembly is then induced by adding DNA-Bridge to the nanoparticles functionalized with DNA-Chol. One advantage of this strategy is that assembly can be triggered simply by adding the DNA-Bridge. This allows for well-controlled experiments to compare the structures of the nanoparticle assemblies with and without the addition of the DNA-Bridge. Such comparisons would not be feasible in a strategy where nanoparticles are grafted with self-complementary DNA-Chol. The second advantage is that all DNA-Chol grafted

nanoparticles are identical, making the formation of the crystal statistically more likely than if nanoparticles were individually grafted with two different complementary DNA-Chol strands and then mixed. The last advantage relates to the cost of DNA. DNA-Chol is considerably more expensive than DNA-Bridge due to the cholesterol modification. With this approach, we can test various bridge lengths and 'sticky end' base pairs by adjusting the sequence of the DNA-Bridge, which is more cost-effective and simpler than altering the sequence of two DNA-Chol molecules.

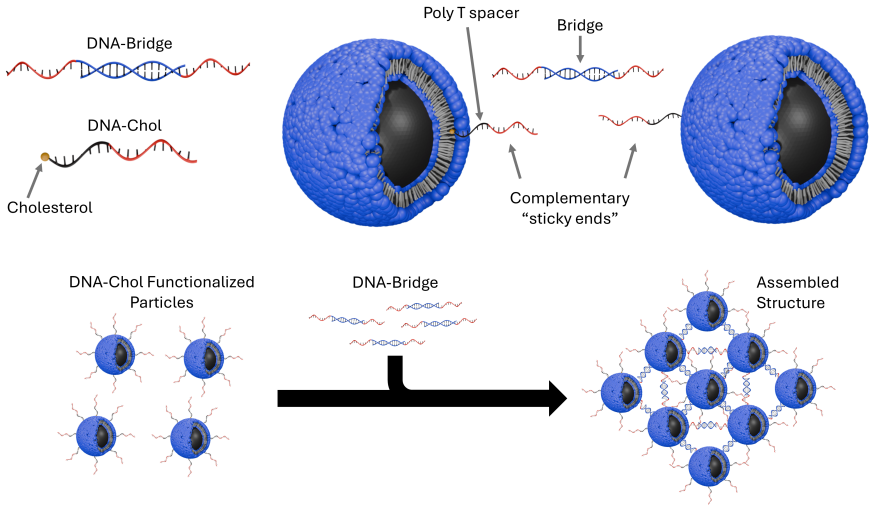



**Figure 6.1:** The DNA design used to assemble the silica particles (figure not drawn to scale). DNA-Chol is first added to functionalize the particles followed by a premixed DNA-Bridge that links two DNA-Chol strands together. DNA-Bridge can be inexpensively tuned to test different lengths or binding energies on the assembly.

### 6.4.5 Lipid Encapsulation of Silica Nanoparticles

The initial step to assemble nanoparticles with lipids and DNA involved encapsulating the nanoparticles within a lipid bilayer, as detailed in the Methods section. To test if this structure was actually formed, a control sample consisting of silica nanoparticles of the same concentration (2 mg/mL) without any DOPC was also created. The two samples were characterized using Dynamic Light Scattering (DLS), Small-Angle Neutron Scattering (SANS), and Small-Angle X-ray Scattering (SAXS).

Small angle scattering techniques are useful to determine if the silica nanoparticles are successfully encapsulated in lipid bilayers because they are sensitive to structures over broad length scales and contrast can be manipulated to highlight different molecular components in composite systems. For example, informa-

tion on the diameter of a lipid vesicle, as well as the thickness of the bilayer, can be obtained from small angle scattering. The main difference between SAXS and SANS is the contrast of silica and DOPC with the suspending solvent. SAXS is more sensitive to silica than to lipids due to the higher X-ray contrast (i.e. electron density) between silica and water versus lipids and water. Because of this, the scattering curves of bare silica nanoparticles and those of encapsulated particles are expected to be reminiscent of monodisperse spherical particles. On the other hand, using contrast variation neutron scattering, high neutron contrasts of both lipids and silica can be obtained, meaning that the scattering curve would have contributions from both materials. SAXS and SANS characterization were performed on the two silica samples, one with and one without DOPC and the scattering curves are shown in Figure 6.2.

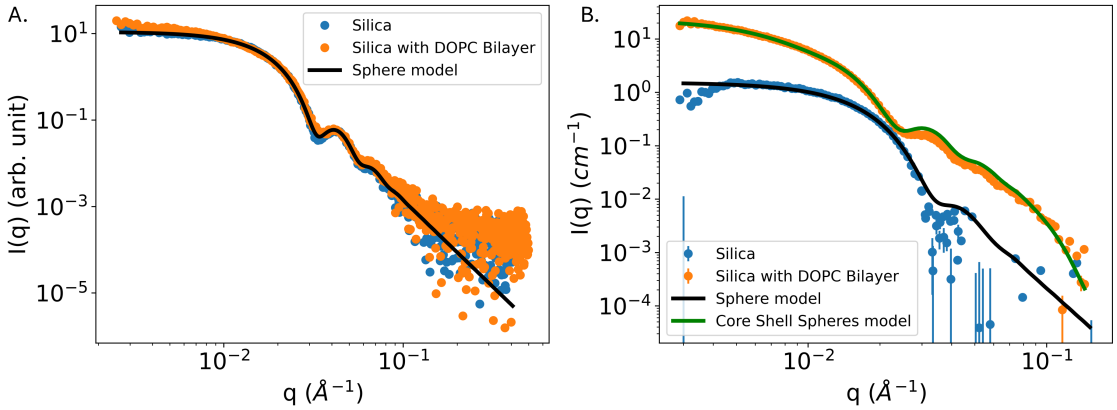
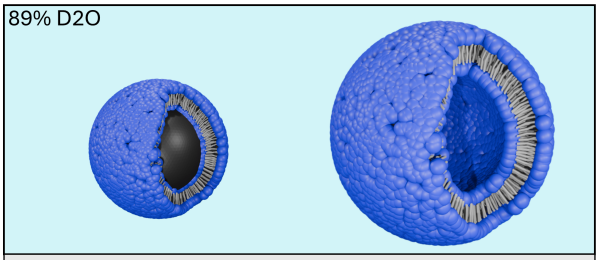



**Figure 6.2:** Two samples were characterized with SAXS and SANS. The first sample contains a dispersion of (2mg/mL) silica nanoparticles. The second sample contains a dispersion of (2mg/mL) silica nanoparticles encapsulated with DOPC, which is prepared using the techniques discussed in the Methods section. (A) shows the SAXS curves of the two samples along with a sphere model fit. Both samples were prepared in water. (B) shows the SANS curves of the two samples along with a sphere model fit and a core shell sphere model fit. Both samples were prepared in 89% $D_2O$.

From Figure 6.2 (A), the SAXS curves of the two silica samples with and without DOPC are largely comparable. The sphere model used to fit the data of the sample containing bare silica nanoparticles has good agreement to the data. From the fit, a sphere diameter of 25 nm with a polydispersity of 0.11 was obtained. The SAXS profile of the sample containing silica and DOPC is also plotted and closely resembles the data of the sample without any lipid. This can be explained because lipids do not significantly alter the scattering signature because of the low X-ray contrast with respect to water. The slight increase in the power-law slope at low q may be attributed to minor nanoparticle aggregation or to the presence of some

larger DOPC vesicles in solution. Figure 6.2 (B) shows the SANS data of the two samples along with the models used to fit the data. As expected, there are major differences between the two SANS curves due to the high neutron contrast of the lipids with the $D_2O$ enriched solvent. To model the data of bare silica, a sphere model was used and a diameter of 24.6 nm with a polydispersity index of 0.15 was obtained which is very similar to the values independently obtained from SAXS. To model the data of the sample containing silica nanoparticles with DOPC, a model composed of the sum of two core shell spheres was used. The first core shell sphere represents the silica nanoparticles encapsulated in a lipid bilayer and the second core shell sphere represents a vesicle. This was done by assigning the appropriate scattering length densities to each material as shown in Figure 6.2. The parameters describing the size of the silica nanoparticle inside the lipid bilayer were also fixed since they were determined from the sphere fit. From the fit of the sample containing silica and DOPC, additional structural parameters were obtained and are shown in Figure 6.3.




| Model: Core Shell Sphere + Core Shell Sphere | | | |
|---|---|---|---|
| **Fit Parameters** | **Value** | **Fit Parameters** | **Value** |
| Bilayer Thickness | 3.6 nm | Bilayer Thickness | 4.5 nm |
| Vol. Frac Structure | 0.60 | Vol. Frac Structure | 0.40 |
| **Fixed Parameters** | | Vesicle Diameter | 41.8 nm |
| Silica Diameter | 24.6 nm | Vesicle Diam. PD | 0.27 |
| Silica Diameter PD | 0.15 | **Fixed Parameters** | |
| SLD Core | 3.48 | SLD Core | 5.58 |
| SLD Shell | 0.09 | SLD Shell | 0.09 |
| SLD Solvent | 5.58 | SLD Solvent | 5.58 |

**Figure 6.3:** The hypothesized structure of the materials in the sample consisting of silica and DOPC. The table in the figure shows the parameters obtained from fitting the SANS data of the sample containing silica and DOPC. The values for the fixed parameters of the silica diameter and silica diameter polydispersity were obtained from fitting a sphere model to the sample that only contained silica. The values from the fit parameters were obtained from fitting the SANS data of the sample with silica and DOPC using a model consisting of the sum of two core shell sphere models. The units of the scattering length densities are ($\times 10^{-6}$ Å$^{-2}$). References for the scattering length densities can be found in the supporting information.

The values from the fit show that the sample is consistent with silica nanoparticles encapsulated in a

lipid bilayer of around 3.6 nm in thickness, which is close to the thickness of a lipid bilayer of POPC, a lipid with size and structure similar to DOPC, found in the literature [128]. This structure makes up the majority of the sample with a volume fraction of about 0.60. In addition to this structure, there are also vesicles in the solution that have a diameter of 41.8 nm and occupy a volume fraction of about 0.40. The thickness of the 'free vesicle' bilayer is 4.5 nm which slightly increased compared to when it was supported by the silica nanoparticle. This increase is also consistent with findings from the literature [128]. Overall, the scattering data from both SANS and SAXS is consistent with silica nanoparticles encapsulated with a DOPC bilayer.

In addition to small angle scattering, Dynamic Light Scattering was also performed on the two samples. The hydrodynamic diameter of the sample with bare silica is around 32 nm and when the DOPC is added, it increases to about 60 nm. This increase in size is significantly more than the size increase that is expected due to a bilayer forming on the particles (7.2 nm). An explanation for this could be the presence of some larger vesicles in the sample determined by SANS, which were found to be around 42 nm in diameter with some polydispersity. It is well known that larger structures scatter more light, which could dominate the scattering detected by DLS, resulting in the larger average size. Moreover, DLS can also overestimate size when there is significant hydration of the interface that can reduce the expected diffusion coefficient of the particles.

### 6.4.6 High-Throughput Assembly Screening

To rapidly explore the large experimental design space, automated robotic synthesis and SAXS characterization were used. Samples were made with the silica nanoparticles encapsulated in a DOPC bilayer created in the previous section, DNA-Chol, NaCl and DNA-Bridge. Water was added to each sample to obtain a sample volume of 200 $\mu L$. The high-throughput experiment was designed to test different concentrations of DNA-Chol (20, 60, 100, 140 DNA Molecules / Particle) and NaCl (0, 10, 20, 30, 40 mM). The concentration of the DNA-Bridge was set so that there would always be one DNA-Bridge molecule for every two DNA-Chol molecules in the sample. This is important since excess addition of the 'bridge' DNA construct could cap the particles and prevent them from binding. Controlled experiments were also performed where all samples were recreated with the same formulation, except for the DNA-Bridge. This would corroborate that the assembly is induced by the addition of the DNA-Bridge and does not occur in its absense. The

concentration of silica nanoparticles encapsulated in DOPC bilayers was kept constant at 1 mg/mL for all samples. Samples were prepared in a custom well plate designed to hold 48 samples, with each well featuring a small chamber at the bottom to collect the assemblies as they sediment due to gravity. This design enables precise positioning of the X-ray beam on the sediment for high-throughput SAXS characterization of the resulting assemblies. While the sample temperature can be adjusted by placing the well plate in an oven, SAXS measurements were limited to room temperature when using this well plate.

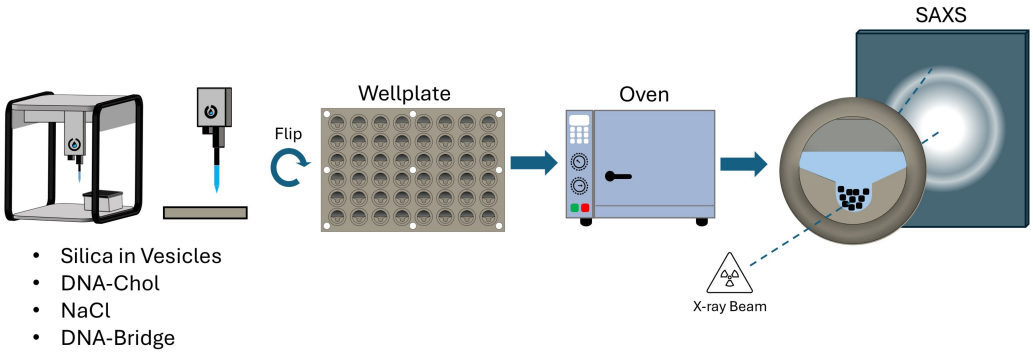



**Figure 6.4:** The process used to screen the effect of NaCl and DNA-Chol concentrations on the assembly of DOPC encapsulated silica nanoparticles in high throughput. Samples were first mixed using a liquid handling robot in a custom wellplate. After sealing and flipping the wellplate in the upright position, any sediment from the sample accumulates in a small pocket in the bottom of the well, where the X-ray beam for SAXS can be positioned. To also assess the effects of thermal annealing, SAXS was measured both before heating the wellplate in an oven (not shown in the figure) and also after heating the wellplate to the melting point of the DNA in the oven and then cooling to room temperature. All SAXS data was taken at room temperature.

The results of the high-throughput screening experiment presented in Figure 6.4 are summarized in Figure 6.5, which shows the resulting SAXS data for the samples assembled at room temperature without any thermal annealing. The SAXS data for samples that were heated to $50°C$ and slowly cooled to room temperature (i.e. thermally annealed) is almost identical and can be found in the supporting information. It was concluded that thermal annealing did not significantly affect the state of assembly under these conditions. Based on the SAXS profiles, three distinct structural regimes were identified and are highlighted in green, purple, and light blue. In the green region, the SAXS curves are characteristic of dispersed spherical particles. Notably, the similarity between the red and blue curves suggests that the presence of DNA-Bridge does not alter the sample structure in this regime. Since no NaCl is present in this region, these observations imply that some electrostatic screening is beneficial for particle assembly. One possible explanation is that addition

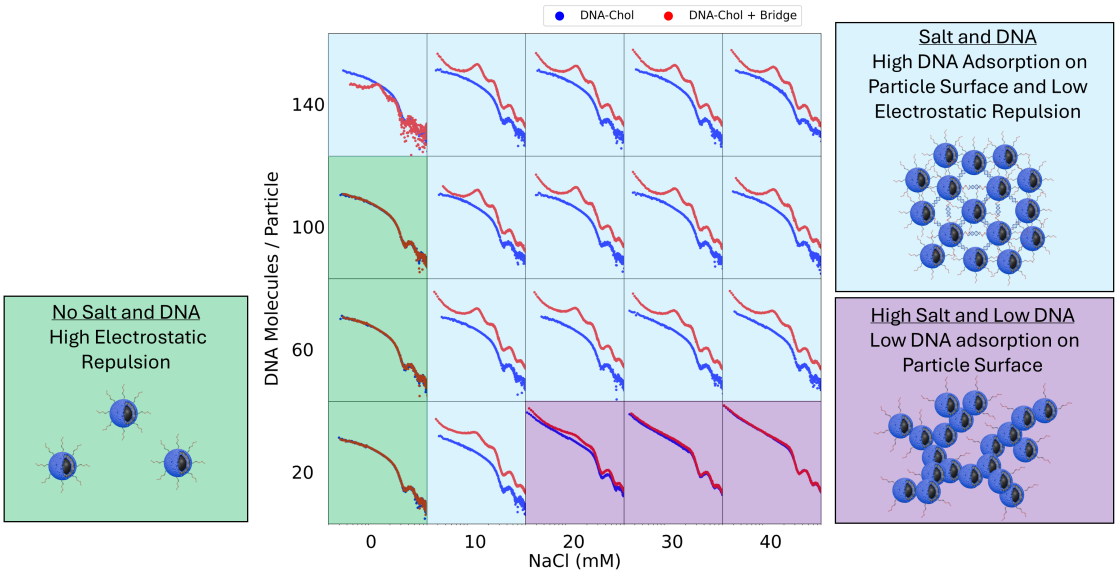



**Figure 6.5:** The SAXS curves from the high-throughput assembly screening without any heating. The curves in blue represent the samples that do not contain DNA-Bridge, while the curves in red contain DNA-Bridge. Regions in the experimental design space that have similar SAXS curves were manually identified and highlighted with the colors green, purple, and light blue. The DNA-Bridge molecule has a length of 20 double stranded base pairs plus two single stranded 'sticky ends' of 18 base pairs on each end of the molecule.

of NaCl reduces electrostatic repulsion between anionic DNA-Chol strands adsorbed onto the lipid bilayer, or between DNA-Chol and the particle surface, thereby facilitating a higher density of DNA molecules on the surface [129]. It is noteworthy that in the sample prepared with 140 DNA-Chol molecules per particle at 0 mM NaCl, DNA mediated assembly is observed in the SAXS profile upon addition of DNA-Bridge. This indicates that some DNA-Chol adsorption onto the particle surface occurs even in the absence of added salt. These results suggest that the grafting density of DNA-Chol on the particle surface is directly proportional to its concentration in solution, with the proportionality influenced by the solution's ionic strength. Therefore, while NaCl is not strictly required for DNA-Chol adsorption, it enhances the efficiency of grafting, likely by mitigating electrostatic repulsion between the negatively charged DNA strands or between the strands and the particle surface.

In the purple region of Figure 6.5, the SAXS curves exhibit a steep power-law decay, indicative of large, aggregated structures. Similar to the green region, the addition of DNA-Bridge does not lead to significant structural changes. This regime corresponds to conditions of low DNA-Chol density per particle combined

with high NaCl concentration. We hypothesize that insufficient surface coverage by DNA-Chol prevents effective steric stabilization, resulting in dense aggregation prior to the introduction of DNA-Bridge. In contrast, the light blue region features a pronounced scattering peak near 0.0136 $Å^{-2}$, consistent with an interparticle spacing that arises due to the length of DNA linking the particles. In this case, the SAXS curves differ notably upon addition of DNA-Bridge, supporting the conclusion that DNA-Bridge actively drives the formation of linked particle assemblies. Despite this, it is clear that the assembly is not completely crystalline because of the relatively broad peak and the absence of additional peaks, which are characterisitic of crystal structures.

If the DNA-Bridge is indeed responsible for inducing assembly, then varying the length of DNA-Bridge should result in assemblies of different interparticle spacings, which would result in SAXS curves with peaks in different locations. This hypothesis was tested by using modified DNA-Bridge molecules that have the same sequence on its complementary 'sticky ends', but a higher number of base pairs in the section of the bridge that does not interact with DNA-Chol. The results of this expeirment are shown in Figure 6.6, where the increase in DNA-Bridge length clearly shows the peak positions shift to a lower $q$, indicating a longer interparticle spacing.

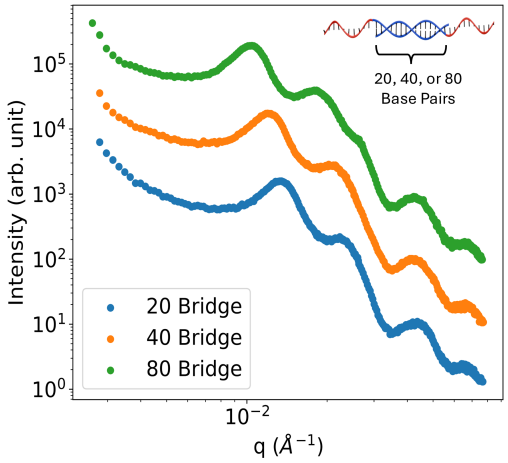



**Figure 6.6:** SAXS data showing the effect of different DNA-Bridge lengths on the assembly without any thermal annealing protocol. All DNA-Bridge molecules have the same base pairs on the ends of the molecule, but differ in the amount of base pairs in the center. These samples were made with 10 mM NaCl and 100 DNA molecules/particle but with DNA-Bridge molecules of varying lengths.

### 6.4.7 SAXS Analysis

To analyze the experimental SAXS data shown in Figure 6.5, a fitting algorithm that combines Bayesian optimization with molecular simulations was used. This approach is inspired by inverse design strategies, where interaction potentials are discovered by matching simulated structures to a desired target [130]. In these simulations, an ultra-coarse-grained model in HOOMD was used, where each silica nanoparticle of diameter $\sigma = 1$ is represented by a single point, and interparticle interactions are implemented using the modified Lennard-Jones potential (6.1) described in Mao et al. [119]. In that work, the potential was validated by a more realistic simulation of the assembly of DNA coated particles. The modifed Lennard-Jones potential is defined by four parameters which were optimized using the algorithm, $U_0$ (optimziation range of 0-150), which controls the depth of the potential, $r_0$, which is the distance where the interparticle interaction is minimum, and $n$ (1-30) and $m$ (1-30), which control the shape and sharpness of the potential. The range of $r_0$ was set to approximately match the calculated interparticle spacing of the particles based on the length of DNA-Bridge. In addition, the number density of particles (0.0005-0.05) in units of particles/$\sigma^3$ inside the simulation box was also added to control the aggregation kinetics and the resulting structure of the final assembly. The Bayesian optimization algorithm was free to suggest combinations of these parameters within the specifed ranges, which were then used to perform the molecular simulation. To capture the system at equilibrium, SAXS curves were simulated for six frames that were spaced apart by 50,000 timesteps near the end of the simulation using an efficient Monte Carlo method [77], and averaged to generate a representative scattering profile. The potential energy of the system was plotted for each simulation to verify that the simulation time was enough for the system to achieve equilibrium. The structure factor was then calculated from the scattering profile and then compared to the corresponding experimental structure factor using a geometry-based similarity metric [39]. The resulting similarity score was then fed back into the optimization algorithm, which iteratively refined the parameters over multiple rounds. It is important to acknowledge the limitations of using simulations to fit experimental scattering data. One limitation is that the optimal parameters are not unique, meaning that several combinations of simulation paramters could give similar results. Another is that the simulation parameters (e.g., $U_0$, $m$, and $n$) are not directly translatable to physical experimental variables such as DNA concentration or the number of hybridizable sticky ends. Solving these limitations would require more detailed and expensive simulations that consider the effect of

DNA sequence and concentration on the interparticle potential. Nevertheless, we find this approach to be essential to capture the complexity of the dispersed semi-crystalline assembled structures that are observed in SAXS, which cannot be properly represented by analytical scattering models.

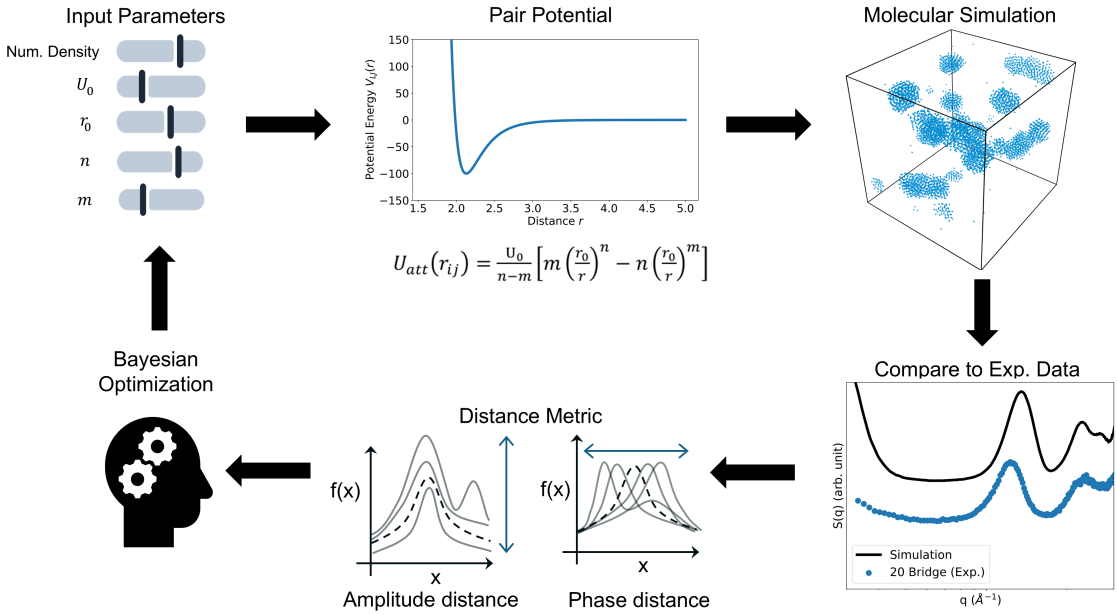



**Figure 6.7:** The optimization protocol used to match the simulated SAXS curve from a molecular simulation to the experimental curve. The optimal input parameters that define the pair potential as well as the molecular simulation are determined by Bayesian optimization.

The optimization results are presented in Figure 6.8 (A) with reasonable agreement between the simulated and experimental structure factors. Simulation snapshots of the structure of the particles are shown in Figure 6.9. The good match between the structure factors at low-q ($\leq 10^{-2}$Å$^{-2}$) for the samples with DNA-Bridge suggests that the cluster size from the simulations match the experiments. Using the cluster analysis tool from the OVITO software [131] with a cut off distance of 3.5, the average spherical cluster diameter from the molecular simulations were determined to be 824 ($\pm$) 324 nm for 20 Bridge, 1392 ($\pm$) 646 nm for 40 Bridge, and 2282 ($\pm$) 648 nm for 80 Bridge. It is difficult to experimentally verify the cluster size, due to sedimentation effects. However, in Figure 6.10 (F), the average size of the cluster of the sample with 20 Bridge determined by DLS is around 600 nm which is close to the size of 824 ($\pm$) 324 nm determined from the simulation. At mid-q ($\geq 10^{-2}$Å$^{-2}$), the agreement between the correlation peaks suggests that the structures from the molecular simulations are also consistent with the interparticle spacing within the

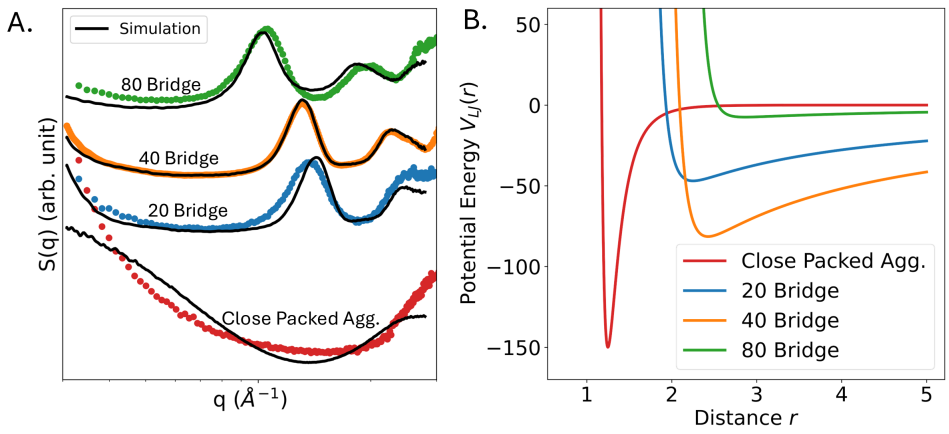



**Figure 6.8:** (A) The results of four independent optimization campaigns are shown as simulated structure factors plotted with the targeted experimental ones. The curves are manually shifted in intensity for visualization purposes. The major tickmark on the x-axis, roughly in the center of the x-axis, represents a q-value of $10^{-2}$ Å$^{-2}$. (B) shows the optimized interparticle potentials used to run the simulations. The unit of the Potential Energy is in kT.

clusters, as well as the relative arrangement of the particles in the experiments. The optimized simulation parameters, including the interparticle distance, are shown in Table 6.1. The values for the interparticle distance can be compared to calculated values, since the length of the silica nanoparticle, lipid bilayer, and DNA are known. Assuming a DNA length of 0.34 nm/base pair [132], a lipid bilayer thickness of about 3.6 nm from SANS data shown in Figure 6.3, and a silica particle diameter of 25 nm from SAXS data, the interparticle distance of 20 Bridge should be 58.2 nm, 40 Bridge should be 65.2 nm and 80 Bridge should be 79.2 nm. These values are close to the interparticle distances in Table 6.1, suggesting a reasonable fit. There is a small disagreement between the interparticle spacing as the length of the bridge gets longer. This could be explained by the increased flexibility of longer DNA bridge strands, which may cause the actual distance to be shorter than the expected one. Although the 'best' fit for the example Close-Packed Aggregate sample was not ideal, the interparticle spacing of 32 nm from the simulations is close to the expected spacing between silica nanoparticles encapsulated in a liposome, which was calculated to be 32.2 nm.

The final equilibrated structures from the molecular simulations can be further analyzed to determine the degree of crystallinity in the structures. The polyhedral template method [1] in OVITO was used to do this. This method classifies particles into categories according to the topology of the local environment, and is a robust method for structures that are not completely crystalline. The results of the analysis are

| Sample | Num. Density | $U_0$ | $r_0$ | $n$ | $m$ |
|---|---|---|---|---|---|
| Close Packed Agg. | 0.0005 | 150 | 1.24 (32 nm) | 8.41 | 28.92 |
| 20 Bridge | 0.0050 | 46.95 | 2.25 (59 nm) | 1.00 | 20.89 |
| 40 Bridge | 0.0050 | 81.42 | 2.42 (63 nm) | 1.00 | 21.78 |
| 80 Bridge | 0.0050 | 7.46 | 2.86 (74 nm) | 1.00 | 30.00 |

**Table 6.1:** Optimal simulation parameters found by fitting structure factors derived from molecular simulations to experimental ones. $U_0$ has units of kT. The number density has units of particles/$\sigma^3$, where $\sigma$ is the size of one particle. $r_0$, the interparticle distance, is reported in units of $\sigma$ as well as the actual length in nm.

shown in Figure 6.9, where the percent of particles categorized in each structure are shown in a bar plot. From the figure, it is clear that DNA induces the assembly of the particles into semi-crystalline composite structures of HCP and FCC ($\sim$30%) and amorphous arrangements of primary particles. For all samples, most primary particles ($\geq$ 60%) are found in amorphous regions, which is consistent with the relatively broad peak observed in the structure factors and the absence of higher-order reflections. By visualizing a single cluster, shown in Figure 6.9 (E), it was observed that the crystalline regions are located in the center of the cluster, while the amorphous regions are located on the surface of the clusters. Therefore, we hypothesize that one way to increase the crystallinity of the sample is to increase the cluster size, which would maximize the crystalline regions.

According to these results, crystalline structures can be achieved without a thermal annealing protocol by using physically tethered DNA bonds on the nanoparticle surface. This could suggest that the mobility of the tethers enables particles to reconfigure and rearrange into crystalline arrangements. This behavior stands in contrast to the more common approach of using immobile gold–thiol bonds to attach DNA molecules to nanoparticle surfaces, where crystallinity cannot be obtained without a thermal annealing step [110] [133]. Although we have not yet achieved the well-defined crystal superlattices reported in many other systems, we believe that increasing the cluster sizes in our system could make this possible.

The SAXS analysis demonstrated in this section also demonstrates that efficient particle dynamics simulations can be used to directly fit SAXS data by automatically tuning simulation parameters, such as the interparticle potential and particle number density, within a closed-loop optimization framework. With this method, a simulation file containing the positions of each particle at equilibrium can be obtained and further

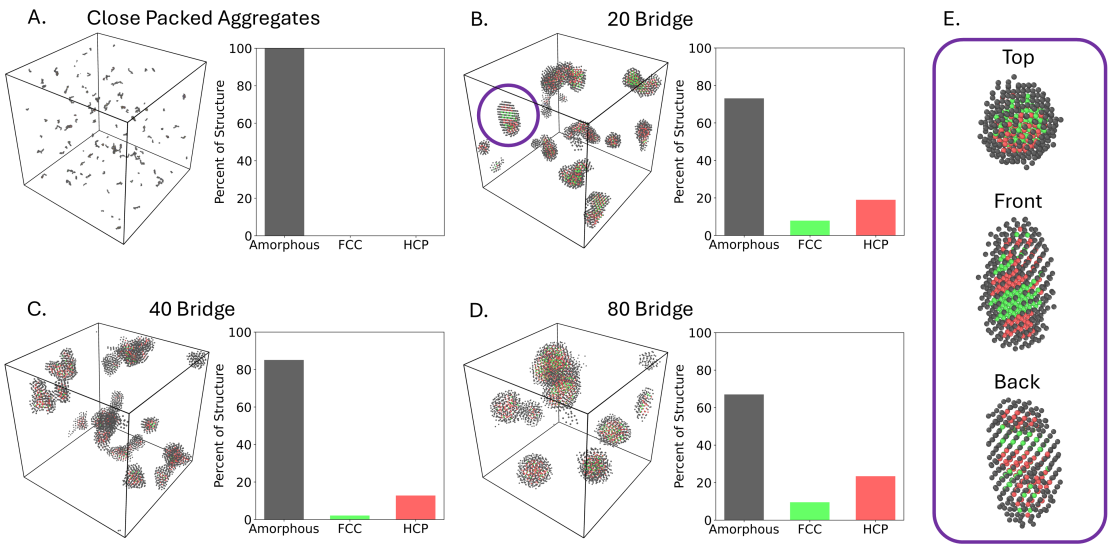



**Figure 6.9:** (A-D) Crystallinity analysis on the final equilibrium snapshots of the simulation for each of the samples. The polyhedral template method [1] categorizes each particle into known structures (Amorphous, FCC, HCP) according to the topology of the local environment. (E) A zoomed in view of the top, front, and back of the cluster circled in purple.

analyzed for cluster size and crystallinity. This approach is valuable because it allows for the influence of various experimental design parameters (e.g., particle shape, particle size, DNA length) on the final assembly to be evaluated *in silico* before conducting the experiment. For example, simulations can aid in designing large crystalline assemblies with targeted symmetries (e.g., FCC, BCC) by first identifying, through parameter sweeps, the trends that promote their formation. This information can then be applied to the experimental setup.

### 6.4.8 Assembly Kinetics and Mechanism

The mechanism at which the particles assemble was investigated using time resolved SAXS and DLS. The sample containing 1 mg/mL silica encapsuled in DOPC, 10 mM NaCl, and 100 DNA Molecules/Particle and the DNA-Bridge with 20 base pairs was recreated and characterized at approximately every 10 minutes after inducing assembly. The results of the experiment are shown in Figure 6.10. From Figure 6.10 (A) the peak corresponding to the interparticle spacing at around 0.0136 $Å^{-2}$ is observed to form and dramatically increase in peak prominence in the first 100 minutes. Figure 6.10 (B) shows the SAXS curves from minute 100 to 360. These curves mostly have the same shape, but only increase in intensity, which is a sign of

sedimentation. The invariant of all the SAXS curves is plotted in Figure 6.10 (C). The invariant is calculated from the area under the curve and is proportional to the number of scattering objects in the beam path. From the data, the invariant stays relatively constant until 180 minutes when it starts to increase, which is evidence of the formation of assemblies that are large enough to sediment due to gravity. Figure 6.10 (D) shows the peak prominence of the SAXS curves as a function of time. The prominence is calculated by subtracting the intensity at the peak from the intensity at the trough. Therefore, it is insensitve to sedimentation and should represent the amount of material in the particle assemblies that have an ordered spacing. The data shows that after around 30 minutes, the peak prominence dramatically increases until about 200 minutes when it finally converges to a single value. This indicates that the assembly process is complete by this time and that the assemblies do not increase in size. In addition to SAXS, DLS was also used to characterize the sample. Figure 6.10 (E) shows the intensity distribution of the sample which is related to the size of objects in it. Figure 6.10 (F) was created by extracting the peak positions of the intensity distribution to track the size of the assembly as a function of time. The data shows that the assembly process takes around 10 minutes to start, but after this, the size of the assembly rapidly grows until it reaches a large size.

Finally, electron microscopy was peformed on the sample at 100 minutes after DNA-Bridge was added. The sample was diluted by a factor of 10 and then drop casted on a microscopy grid. The images in Figure 6.11 show several large particle clusters of silica nanoparticles of about 1-2 microns in diameter. This size is slightly larger than the size from DLS. However, the DLS data increases rapidly at around 100 minutes and drying effects may result in cluster coalesence, which would increase the cluster size. Despite this, the images show a presence of large clusters of silica nanoparticles which are consistent with the structures shown in the simulations.

Together, the data from SAXS, DLS, and microscopy suggest that the assembly follows a classical nucleation and growth mechanism. Initially, small clusters form, and once they reach a critical size, further growth becomes thermodynamically favorable [134]. This is reflected in the SAXS data, where a correlation peak appears early during assembly and gradually increases in prominence, indicating the development of ordered structures. DLS data further support this mechanism, showing a progressive increase in the size of the assemblies over time. Eventually, we hypothesize that the clusters coalesce as shown in the microscopy images, making them large enough to undergo sedimentation. This mechanism is consistent

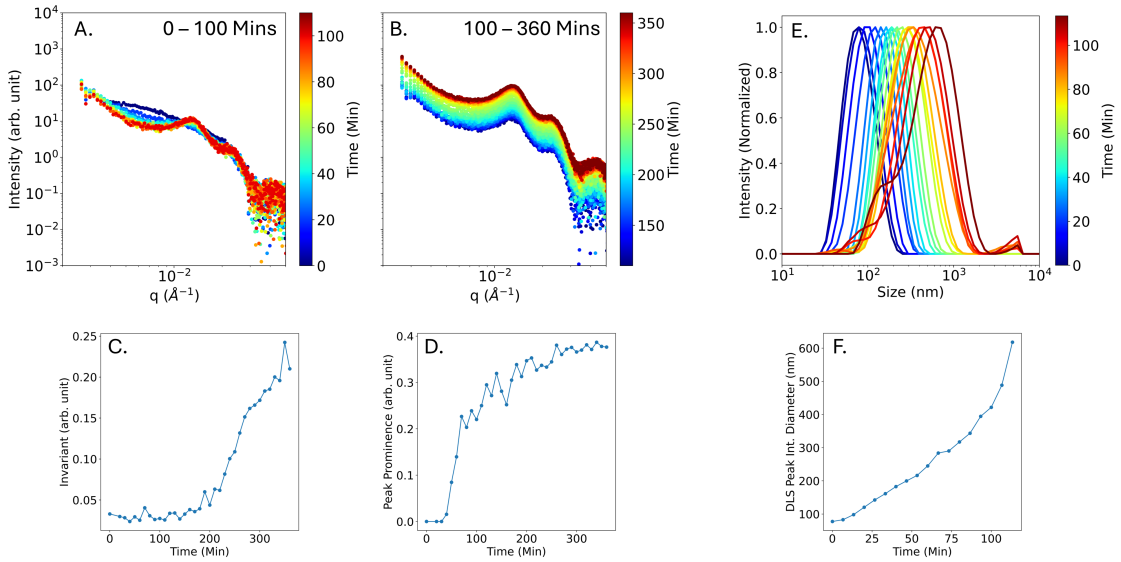



**Figure 6.10:** Time resolved SAXS and DLS measurments of the sample 20 Bridge. The initial time (t=0) corresponds to the time at which the DNA-Bridge was added to the sample. (A) shows the SAXS curves of the first 100 minutes. (B) shows the SAXS curves of minute 100 to 360. (C) shows the invariant of all the SAXS curves as a function of time. (D) shows the peak prominence of the SAXS curves, which was determined by the height of the peak minus the trough. (E) shows the results from DLS of the first 100 minutes of assembly. (F) is derived from the DLS data and shows the location of the peak position as a function of time.

with the structures in the particle simulations as shown in Figure 6.12.

## 6.5 Conclusion

This study demonstrates the behavior of DNA-mediated assembly of lipid encapsulated silica nanoparticles. After verifying that the silica nanoparticles were properly encapsulated using a combination of scattering techniques, a high-throuhput platform was used to test the effect of NaCl and DNA concentration on the assembly. The results show that while NaCl is not strictly required for DNA-Chol adsorption on the surface of the particle, it enhances the efficiency of grafting, likely by mitigating electrostatic repulsion between the negatively charged DNA-Chol strands and the particle surface. DNA-Chol strands anchored to the particle surface provide steric stabilization, and allows for controlled assembly of the nanoparticles after the addition of DNA-Bridge. The resulting structures exhibit controlled interparticle spacing that is determined by the length of the DNA-Bridge. Molecular simulations combined with Bayesian optimization are used to fit

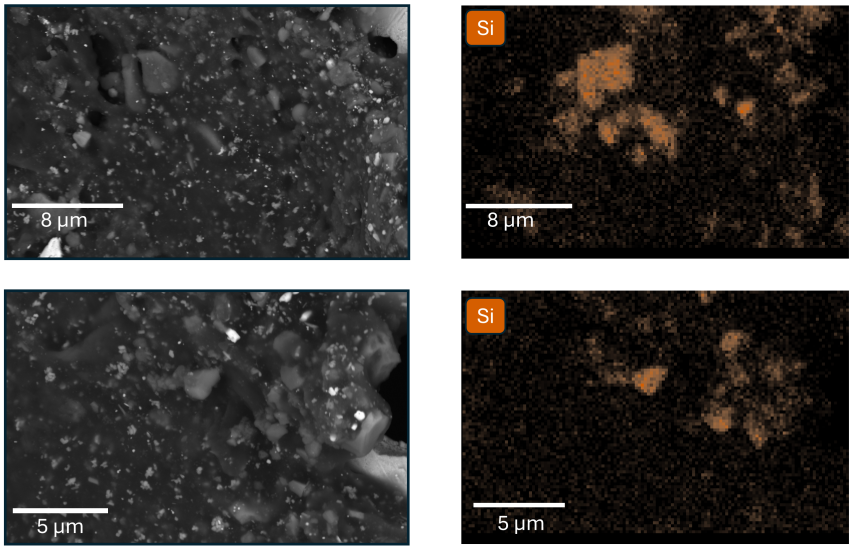



**Figure 6.11:** Electron microscopy with energy dissipative spectroscopy was performed on the 20 Bridge sample. At 100 minutes after the DNA-Bridge was addded, the sample was diluted by a factor of 10 and drop casted on a microscopy grid.

the experimental structrue factors and optimal simulation parameters are obtained. The resulting particle configurations reveal structural arrangements consistent with the expected interparticle distances derived from the length of DNA linking the particles. Crystallinity analyis reveals that around 30% of all particles in the sample are in a crystalline arrangement. Furthermore, the crystalline particles are located in the center of the cluster of particles while the surface of the cluster is amorphous. The ability to form crystalline structures without any thermal annealing could be due to the lateral mobility of DNA on the surface of the nanoparticles, but further experiments are needed to validate this. Finally, time resolved SAXS and DLS measurements are used to monitor the assembly process over time. Together, the data suggests that the assembly follows a classical nucleation and growth mechanism, where small clusters form and slowly increase in size, until they coalesce and undergo sedimentation.

## 6.6 Future Experiment Plan

Several future experiments can be designed to advance the DNA-mediated assembly of lipid-encapsulated silica nanoparticles. One key experiment should focus on understanding and optimizing the encapsulation

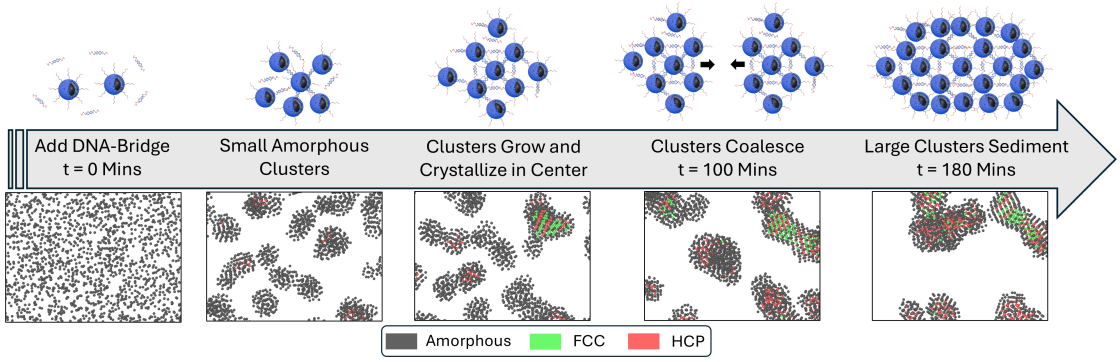



**Figure 6.12:** The proposed mechanism of the assembly of the DNA coated DOPC encapsulated silica nanoparticles. The timesteps in the figure are based on the interpretation of experimental SAXS, DLS, and microscopy data. The mechanism is consistent with simulation snapshots of the sample with 20 Bridge.

process of silica nanoparticles within lipid bilayers. SANS data indicated that samples containing lipid-encapsulated silica nanoparticles also included a significant population of larger, free vesicles. This is undesirable, as DNA-Chol can adsorb onto both the encapsulated nanoparticles and the free vesicles, potentially disrupting the assembly process and preventing the formation of a well-ordered, close-packed crystal. To address this issue, the primary objective should be to determine the optimal DOPC to silica ratio that maximizes nanoparticle encapsulation while minimizing the presence of vesicles in solution. To do this, several conditions of lipid to silica ratios can be measured and the data can be fitted with a core-shell-sphere + core-shell-sphere model. The first core-shell-sphere represents the silica encapsulated in lipid bilayers and the second represents free vesicles. This can be done by manipulating the scattering length densities of the core and shell. From the fit, the proportion of each structure can be obtained, and the optimal ratio of lipid to silica should be the one where the proportion of vesicles is the minimum. These samples should be performed in the highest proportion of $D_2O/H_2O$ to minimize incoherent scattering from hydrogen. Both silica and lipids have high scattering length densities at high $D_2O$ so the scattering of both materials should contribute to the scattering curve. A second experiment is also proposed where the same samples are measured at 60.12% $D_2O$, which is the point of minimum intensity (PMI) of silica determined from a previous beamline experiment. At this solvent composition, the scattering length density of silica matches the one of the solvent background, resulting in scattering contributions from only the lipids. By fitting the data with the appropriate core-shell-sphere model, the structure of the lipid bilayers should be consistent with the results from the same samples measured at 99% $D_2O$.

**Test Lipid Encapsulation of Nanoparticles (6 samples)**

- Sample: 0.4, 0.6, 0.8, 1.0, 1.2, 1.4 DOPC_SA:Silica_SA in 99% $D_2O$

**PMI Silica (6 samples)**

- Sample: 0.4, 0.6, 0.8, 1.0, 1.2, 1.4 DOPC_SA:Silica_SA 60.12% $D_2O$

The second objective is to measure the amount of cholesterol modified DNA molecules adsorbed onto the surface of the lipid encapsulated particles. This is important to quantify the amount of DNA-Chol being adsorbed onto the surface of the particle, which is responsible for assembly. From previous SANS data, it was hypothesized that the attractive force between the cholesterol in the DNA molecule to the lipid bilayer could be insufficient to achieve a high grafting density. To test this hypothesis, an additional DNA cholesterol molecule could be added which hybridizes with the first DNA-Chol molecule. This results in a pair of cholesterol molecules anchoring the DNA strands on to the surface of the nanoparticle. SANS can be used to test if this modification results in an increased DNA grafting density. The proposed experiments include adding different concentrations of DNA-Chol to the silica nanoparticles encapsulated in lipid bilayers. Sodium chloride can be added to reduce electrostatic repulsion which leads to a higher DNA grafting density. These samples should also be created in the highest amount of $D_2O$ as possible to minimize incoherent scattering. This can be done by ordering DNA-Chol in a dried state and resuspending it in $D_2O$. To fit this data, a multi-core-shell model could be used to model the silica, lipids, and DNA.

**Test the functionalization of particles with DNA (4 samples)**

- Sample: 1.0 DOPC_SA:Silica_SA in 99% $D_2O$ with 20, 50, 100, 150 DNA/Particle

**Test the functionalization of particles with DNA with NaCl (4 samples)**

- Sample: 1.0 DOPC_SA:Silica_SA in 99% $D_2O$ with 20, 50, 100, 150 DNA/Particle in 10 mM NaCl

**Test the functionalization of particles with DNA and the additional DNA-Chol anchor strand (4 samples)**

- Sample: 1.0 DOPC_SA:Silica_SA in 99% $D_2O$ with 20, 50, 100, 150 DNA/Particle and the other anchoring DNA strand

**Test the functionalization of particles with DNA with NaCl and the additional DNA-Chol anchor strand (4 samples)**

- Sample: 1.0 DOPC_SA:Silica_SA in 99% $D_2O$ with 20, 50, 100, 150 DNA/Particle, 10 mM NaCl, and the other anchoring DNA strand

Additional experiments are also proposed such as measuring the kinetics of assembly. In this experiment, the DNA-Bridge is added to induce assembly, and SANS is collected to determine structural changes over time. Another experiment is to encapsulate larger silica nanoparticles with a lipid bilayer and to characterize their structure.

**Kinetics of Assembly (1 Sample, least priority)**

- Add the DNA-Bridge and measure sample as a function of time. One measurement approximately every 20 mins for 2–3 hours of total measurement time

**Larger Silica Sample**

- Use DOPC to encapsulate silica particles of a larger size (approximately 5–6 samples)

To perform this experiment, several DNA sequences should be ordered. The following sequences can be ordered on Integrated DNA technologies (IDT). Based on previous experiments, the following product and quantities should be enough to perform the proposed experiment. DNA that is ordered without any services will arrive in a dried state and can be resuspended to the desired concentration in $D_2O$. Details on the procedure can be found on the IDT website.

## DNA Sequences

### DNA-Chol_15_B_24_HBP

- Product/Scale: 250 nmole DNA Oligo
- Purification: HPLC
- Services: None
- Sequence: ACT CTG TAT GAA GTG ATG GAT GAT TTT TTT TTT TTT TTT /3CholTEG/
- Quantity: 3

### DNA-Chol_15_B_Anchor

- Product/Scale: 250 nmole DNA Oligo
- Purification: HPLC
- Services: None
- Sequence: /5Chol-TEG/AA AAA AAA AAA AAA A
- Quantity: 2

### DNA-Brigde_A_20B_24_HBP

- Product/Scale: 250 nmole DNA Oligo
- Purification: Standard Desalting
- Services: LabReady (Normalized to 100µM in IDTE pH 8.0)
- Sequence: ATC ATC CAT CAC TTC ATA CAG AGT ACT GAG CAG CAC TGA CAG CA
- Quantity: 1

**DNA-Brigde_B_20B_24_HBP**

- Product/Scale: 250 nmole DNA Oligo
- Purification: Standard Desalting
- Services: LabReady (Normalized to 100μM in IDTE pH 8.0)
- Sequence: ATC ATC CAT CAC TTC ATA CAG AGT TGC TGT CAG TGC TGC TCA GT
- Quantity: 1

### Sample Preparation

The sample containing silica and DOPC should be prepared in the highest possible amount of $D_2O$ to minimize incoherent scattering. It was found in a previous beamline experiment that a silica concentration of 1 mg/mL, a banjo cell of 2 mm thickness (600 μL sample volume) and a measurement time of 60 minutes was enough to obtain high quality data. Therefore, all the samples in this proposed experiment can be prepared at that concentration and volume. It is suggested to prepare at least 1 mL of each sample, so that DLS and SAXS can also be used to characterize them.

### Code Resource

The code used to calculate how much silica, DOPC, and DNA is needed to create a sample can be found at:

`https://github.com/pozzo-research-group/silica_lipid_DNA`

## 6.7 Acknowledgments

This work was supported by the DOE Energy Frontiers Research Center (EFRC) the Center for the Science of Synthesis Across Scales (DE-SC0019288). The authors acknowledge the use of facilities and instrumentation supported by the U.S. National Science Foundation through the Major Research Instrumentation (MRI) program (DMR-2116265) and the UW Molecular Engineering Materials Center (MEM-C), a Materials Research Science and Engineering Center (DMR-2308979). This research used resources at the High

Flux Isotope Reactor, a DOE Office of Science User Facility operated by the Oak Ridge National Laboratory. The beam time was allocated to CG-3 Bio-SANS on proposal number IPTS-33819.1. We thank Dr. Wellington Leite for assisting with the SANS experiments. This work benefited from the use of the SasView application, originally developed under NSF award DMR-0520547. SasView contains code developed with funding from the European Union's Horizon 2020 research and innovation program under the SINE2020 project, grant agreement No 654000. This work was also facilitated by the advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and the Department of Chemical Engineering at the University of Washington.

# Chapter 7

# Conclusion

## 7.1 Summary and Perspectives

Colloidal nanomaterials owe their diverse functionalities and applications to their structural characteristics (see Chapter 1 for examples). As a result, precise control over their synthesis and self-assembly is crucial for leveraging these properties. However, this is challenging beacuse the experimental design space for colloidal synthesis and assembly is vast and highly complex. One way to navigate this design space is the concept of a self-driving lab, introduced in Chapter 2, which combines machine learning with high-throughput experimentation and characterization. In this approach, the AI agent selects the experiments to be conducted and sends them to autonomous robotic systems. Once the experiments are completed, the results are characterized and the resulting data is fed back to the AI agent to inform subsequent decisions forming a closed loop which continues until a stopping criteria is met. Despite being an effective method to navigate experimental design spaces, as discussed in Chapter 3, there remain several opportunities for enhancing these data-driven systems. One key area for improvement lies in the distance metric used to provide feedback to the AI agent. Specifically, when functional data is employed as a proxy for nanoparticle structure, refining this metric could lead to better guidance and outcomes. In Chapter 3, the Amplitude-phase distance metric is introduced to solve this problem. This metric primarily takes into account the shape of a function or curve while calculating a distance between two functions. By using function spaces, it is possible to analyze data with differential geometry methods. The results show that an AI agent using this distance

metric is able to distinguish between the shapes of nanoparticles being formed (e.g., spheres and rods), which results in a more dispersed sampling of the design space compared to when it used the Euclidean distance metric. In the same chapter, we also identify three other limitations of a self-driving lab which are the use of a single characterization method, the use of extensive knowledge from literature to set the experimental design space, and the lack of knowledge extraction from the experiments. In Chapter 4, we develop a data-driven system and test it on silver nanoplate synthesis. This system had several improvements over the one demonstrated in Chapter 3 that were designed to overcome the limitations previously mentioned. First of all, it employed several characterization methods to navigate the design space, using a combination of UV-Vis spectroscopy, SAXS, and TEM while considering the quality of information gained and the cost of performing the measurement. It also used an interpretable AI-agent to navigate the design space, where knowledge was able to extracted. The design rules gained from the experiment agreed with the ones found in the literature indicating that this data-driven method could be used to extract knowledge from more complex material systems. Finally, the design space of the system was arbitrarily chosen and did not rely much on information from the literature. The ability to start from a large design space is essential for data-driven systems to be applied on materials that have not been well-studied.

In Chapter 5, we transition from inorganic nanoparticle synthesis to protein assembly. In this chapter, a model protein called RhuA is modified by covalently attaching either a $\beta$-cyclodextrin (host) or azobenzene (guest) molecule to it. This results in chemically and light responsive protein assemblies when mixed together, where assembly is promoted in visible light and disassembly in UV-light. To characterize the structure of the protein assembly, small angle x-ray scattering is used and a Monte Carlo method combined with a fitting algorithm is developed to efficiently analyze the scattering curves. The Monte Carlo Distribution Function Method (MC-DFM) uses random sampling to calculate pairwise distances of atoms in the large protein assembly. It is also efficiently coded using matrix operations, making it highly suitable for handling large structures. When combined with a fitting algorithm, the modeling of large biomolecular assemblies involves using a single building block (e.g. RhuA) and applying a series of rigid-body transformations to generate the complete structure such as tubes or sheets. This method offers a significant advantage, as adjustments to the assembly's structure (e.g., altering tube radii) can be rapidly achieved by modifying only the transformations. Consequently, explicit calculation of all atoms in the large structure is avoided, greatly

reducing computational time. The SAXS analysis shows that the RhuA protein assembles into tube or sheet like structures, depending on the ionic strength of the solution, which is consistent with data obtained from microscopy. We envision that our method of analyzing SAXS curves can be used on data of other large biomolecular assemblies, so we publish it as open source software. In addition, there are several opportunities to expand on the work of chemical and light responsive RhuA assemblies. For example, the use of a combination of UV and Visible light at the correct sequence and intensity could be used to achieve a metastable assembly other than the thermodynamically stable tubes or sheet configuration. As discussed throughout this thesis, data-driven methods could be used together with *in situ* SAXS measurements to determine the correct light sequence and intensity to achieve metastable structures. Finally, in Chapter 6, the DNA-mediated assembly of lipid encapsulated nanoparticles in investigated. Due to the grafting of DNA to the lipid bilayer, this system is advantageous because it can be applied on nanoparticles of any material, and because it allows for DNA mobility on the particle surface. We use high-throughput experimentation to investigate the effect of the solution's ionic strength and the DNA concentration on the assembly. The results show that while NaCl is not strictly required for DNA adsorption on the surface of the particle, it enhances the efficiency of grafting, likely by mitigating electrostatic repulsion between the negatively charged DNA strands and the particle surface. DNA strands anchored to the particle surface provide steric stabilization, and allows for controlled assembly of the nanoparticles after the addition of the linking DNA strand. SAXS analysis shows that the assembly structures exhibit controlled interparticle spacing that is determined by the length of the DNA-Bridge. Despite this, the assemblies do not form close-packed crystalline arrangements, which has been demonstrated in the literature with DNA-mediated nanoparticle assembly. The future step in this work could be to determine how to form closed-packed crystalline structures. Since the high-throughput framework has already been developed, it could be used together with an AI-agent to explore the vast design space like in Chapter 4 and Chapter 5.

# Bibliography


[1] Peter Mahler Larsen, Søren Schmidt, and Jakob Schiøtz. Robust structural identification via polyhedral template matching. *Modelling and Simulation in Materials Science and Engineering*, 24(5):055007, may 2016.

[2] Reema Narayan, Usha Y. Nayak, Ashok M. Raichur, and Sanjay Garg. Mesoporous silica nanoparticles: A comprehensive review on synthesis and recent advances. *Pharmaceutics*, 10(3), 2018.

[3] F. Pelayo García de Arquer, Dmitri V. Talapin, Victor I. Klimov, Yasuhiko Arakawa, Manfred Bayer, and Edward H. Sargent. Semiconductor quantum dots: Technological progress and future challenges. *Science*, 373(6555):eaaz8541, 2021.

[4] Nadeem Baig, Irshad Kammakakam, and Wail Falath. Nanomaterials: a review of synthesis methods, properties, recent progress, and challenges. *Mater. Adv.*, 2:1821–1871, 2021.

[5] Leonardo Scarabelli, Ana Sánchez-Iglesias, Jorge Pérez-Juste, and Luis M. Liz-Marzán. A "tips and tricks" practical guide to the synthesis of gold nanorods. *The Journal of Physical Chemistry Letters*, 6(21):4270–4279, 2015. PMID: 26538043.

[6] Gonzalo Villaverde-Cantizano, Marco Laurenti, Jorge Rubio-Retama, and Rafael Contreras-Cáceres. Reducing agents in colloidal nanoparticle synthesis – an introduction. In *Reducing Agents in Colloidal Nanoparticle Synthesis*. The Royal Society of Chemistry, 05 2021.

[7] Huat Thart Chiang, Kiran Vaddi, and Lilo Pozzo. Data-driven exploration of silver nanoplate formation in multidimensional chemical design spaces. *Digital Discovery*, 3:2252–2264, 2024.

[8] Babak Nikoobakht and Mostafa A. El-Sayed. Preparation and growth mechanism of gold nanorods (nrs) using seed-mediated growth method. *Chemistry of Materials*, 15(10):1957–1962, 2003.

[9] Dessie Belay Emrie. Sol–gel synthesis of nanostructured mesoporous silica powder and thin films. *Journal of Nanomaterials*, 2024(1):6109770, 2024.

[10] Zhiwei Li, Qingsong Fan, and Yadong Yin. Colloidal self-assembly approaches to smart nanostructured materials. *Chemical Reviews*, 122(5):4976–5067, 2022. PMID: 34747588.

[11] Guang Yang, Jagjit Nanda, Boya Wang, Gang Chen, and Daniel T. Jr. Hallinan. Self-assembly of large gold nanoparticles for surface-enhanced raman spectroscopy. *ACS Applied Materials & Interfaces*, 9(15):13457–13470, 2017. PMID: 28328194.

[12] Jia-Yu Lin, Fang-Chi Hsu, Yu-Chieh Chao, Guan-Zhang Lu, Mujahid Mustaqeem, and Yang-Fang Chen. Self-assembled monolayer for low-power-consumption, long-term-stability, and high-efficiency quantum dot light-emitting diodes. *ACS Applied Materials & Interfaces*, 15(21):25744–25751, 2023. PMID: 37199533.

[13] Hyo-Jun Lim, Thi Huong Thao Dang, Nayoon Lee, Sunwoo Jin, Van-Khoe Vo, Joon-Hyung Lee, Won Sik Shin, Byoung-Seong Jeong, and Young-Woo Heo. Self-assembled monolayer-functionalized nio hole injection layer for improved charge injection in quantum dot light-emitting diodes. *ACS Applied Materials & Interfaces*, 17(1):1533–1541, 2025. PMID: 39780384.

[14] Nitu L. Wankhede, Mayur B. Kale, Aman B. Upaganlawar, Brijesh G. Taksande, Milind J. Umekar, Tapan Behl, Ahmed A.H. Abdellatif, Prasanna Mohana Bhaskaran, Sudarshan Reddy Dachani, Aayush Sehgal, Sukhbir Singh, Neelam Sharma, Hafiz A. Makeen, Mohammed Albratty, Hamed Ghaleb Dailah, Saurabh Bhatia, Ahmed Al-Harrasi, and Simona Bungau. Involvement of molecular chaperone in protein-misfolding brain diseases. *Biomedicine Pharmacotherapy*, 147:112647, 2022.

[15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav

Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[16] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, August 2023.

[17] Cara W. Chao, Kaitlin R. Sprouse, Marcos C. Miranda, Nicholas J. Catanzaro, Miranda L. Hubbard, Amin Addetia, Cameron Stewart, Jack T. Brown, Annie Dosey, Adian Valdez, Rashmi Ravichandran, Grace G. Hendricks, Maggie Ahlrichs, Craig Dobbins, Alexis Hand, Jackson McGowan, Boston Simmons, Catherine Treichel, Isabelle Willoughby, Alexandra C. Walls, Andrew T. McGuire, Elizabeth M. Leaf, Ralph S. Baric, Alexandra Schäfer, David Veesler, and Neil P. King. Protein nanoparticle vaccines induce potent neutralizing antibody responses against MERS-CoV. *Cell Reports*, 43(12):115036, December 2024.

[18] Susana Vázquez Torres, Melisa Benard Valle, Stephen P. Mackessy, Stefanie K. Menzies, Nicholas R. Casewell, Shirin Ahmadi, Nick J. Burlet, Edin Muratspahić, Isaac Sappington, Max D. Overath, Esperanza Rivera-de Torre, Jann Ledergerber, Andreas H. Laustsen, Kim Boddum, Asim K. Bera, Alex Kang, Evans Brackenbrough, Iara A. Cardoso, Edouard P. Crittenden, Rebecca J. Edge, Justin Decarreau, Robert J. Ragotte, Arvind S. Pillai, Mohamad Abedi, Hannah L. Han, Stacey R. Gerben, Analisa Murray, Rebecca Skotheim, Lynda Stuart, Lance Stewart, Thomas J. A. Fryer, Timothy P. Jenkins, and David Baker. De novo designed proteins neutralize lethal snake venom toxins. *Nature*, January 2025.

[19] Qin Li, Ulrich Jonas, X. S. Zhao, and Michael Kappl. The forces at work in colloidal self-assembly: a review on fundamental interactions between colloidal particles. *Asia-Pacific Journal of Chemical Engineering*, 3(3):255–268.

[20] John C. Berg. *An Introduction of Interfaces  Colloids: The Bridge to Nanoscience*. World Scientific, Singapore, 2010.

[21] Kjersta Larson-Smith and Danilo C. Pozzo. Scalable synthesis of self-assembling nanoparticle clusters based on controlled steric interactions. *Soft Matter*, 7:5339–5347, 2011.

[22] Dymtro Nykypanchuk, Matthew Maye, Daniel van der Lelie, and Oleg Gang. Dna-guided crystallization of colloidal nanoparticles. *Nature*, 451:549–552, 2008.

[23] Melissa Rinaldin, Ruben W. Verweij, Indrani Chakraborty, and Daniela J. Kraft. Colloid supported lipid bilayers for self-assembly. *Soft Matter*, 15:1345–1360, 2019.

[24] RobertJ. Macfarlane, MatthewR. Jones, AndrewJ. Senesi, KaylieL. Young, Byeongdu Lee, Jinsong Wu, and ChadA. Mirkin. Establishing the design rules for dna-mediated programmable colloidal crystallization. *Angewandte Chemie International Edition*, 49(27):4589–4592, 2010.

[25] Robert J. Macfarlane, Ryan V. Thaner, Keith A. Brown, Jian Zhang, Byeongdu Lee, SonBinh T. Nguyen, and Chad A. Mirkin. Importance of the dna "bond" in programmable nanoparticle crystallization. *Proceedings of the National Academy of Sciences*, 111(42):14995–15000, 2014.

[26] Mhejabeen Sayed and Haridas Pal. An overview from simple host–guest systems to progressively complex supramolecular assemblies. *Phys. Chem. Chem. Phys.*, 23:26085–26107, 2021.

[27] Gonzalo Rivero-Barbarroja, Carlos Fernández-Clavero, Cristina García-Iriepa, Gema Marcelo, M. Carmen Padilla-Pérez, Tania Neva, Juan M. Benito, Stéphane Maisonneuve, Carmen Ortiz Mellet, Juan Xie, José M. García Fernández, and Francisco Mendicuti. Reversible light-induced dimerization of secondary face azobenzene-functionalized -cyclodextrin derivatives. *The Journal of Organic Chemistry*, 88(13):8674–8689, 2023. PMID: 37341522.

[28] Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. *Chemical Reviews*, 121(16):10073–10141, 2021.

[29] Huachen Tao, Tianyi Wu, Sina Kheiri, Matteo Aldeghi, Alán Aspuru-Guzik, and Eugenia Kumacheva. Self-driving platform for metal nanoparticle synthesis: Combining microfluidics and machine learning. *Advanced Functional Materials*, 31(51), 2021.

[30] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. 25, 2012.

[31] Tarak K. Patra, Venkatesh Meenakshisundaram, Jui-Hsiang Hung, and David S. Simmons. Neural-network-biased genetic algorithms for materials design: Evolutionary algorithms that learn. *ACS Combinatorial Science*, 19(2):96–107, 2017.

[32] Ziyu Ye, Zijie Wu, and Arthi Jayaraman. Computational reverse engineering analysis for scattering experiments (crease) on vesicles assembled from amphiphilic macromolecular solutions. *JACS Au*, 1(11):1925–1936, 2021.

[33] Kacper J. Lachowski, Huat Thart Chiang, Kaylyn Torkelson, Wenhao Zhou, Shuai Zhang, Jim Pfaendtner, and Lilo D. Pozzo. Anisotropic gold nanomaterial synthesis using peptide facet specificity and timed intervention. *Langmuir*, 39(45):15878–15888, 2023. PMID: 37910774.

[34] Sourav Bhattacharjee. Dls and zeta potential – what they are and what they are not? *Journal of Controlled Release*, 235:337–351, 2016.

[35] Jorg Stetefeld, Sean A. McKenna, and Trushar R. Patel. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophysical Reviews*, 8(4):409 – 427, 2016.

[36] I. Bressler, B. R. Pauw, and A. F. Thünemann. *McSAS*: software for the retrieval of model parameter distributions from scattering patterns. *Journal of Applied Crystallography*, 48(3):962–969, Jun 2015.

[37] Christopher L. Farrow and Simon J. L. Billinge. Relationship between the atomic pair distribution

function and small-angle scattering: implications for modeling of nanoparticles. *Acta Crystallographica Section A*, 65(3):232–239, May 2009.

[38] Flore Mekki-Berrada, Zekun Ren, Tan Huang, Wai Kuan Wong, Fang Zheng, Jiaxun Xie, Isaac Parker Siyu Tian, Senthilnath Jayavelu, Zackaria Mahfoud, Daniil Bash, et al. Two-step machine learning enables optimized nanoparticle synthesis. *npj Computational Materials*, 7(1):1–10, 2021.

[39] Vaddi Kiran, Chiang Huat Thart, and Pozzo Lilo D. Autonomous retrosynthesis of gold nanoparticles via spectral shape matching. *Digital Discovery*, 1(502-510), 2022.

[40] Daniel Salley, Graham Keenan, Jonathan Grizou, Abhishek Sharma, Sergio Martín, and Leroy Cronin. A nanomaterials discovery robot for the darwinian evolution of shape programmable gold nanoparticles. *Nature Communications*, 11, 2020.

[41] Zhong-Jie Jiang, Chun-Yan Liu, and Lu-Wei Sun. Catalytic properties of silver nanoparticles supported on silica spheres. *Journal of Physical Chemistry B*, 109:1730–1735, 2005.

[42] Xi-Feng Zhang, Zhi-Guo Liu, Wei Shen, and Sangiliyandi Gurunathan. Silver nanoparticles: Synthesis, characterization, properties, applications, and therapeutic approaches. *International Journal of Molecular Sciences*, 17(9), 2016.

[43] Parteek Prasher, Mousmee Sharma, Harish Mudila, Gaurav Gupta, Abhishek Kumar Sharma, Deepak Kumar, Hamid A. Bakshi, Poonam Negi, Deepak N. Kapoor, Dinesh Kumar Chellappan, Murtaza M. Tambuwala, and Kamal Dua. Emerging trends in clinical implications of bio-conjugated silver nanoparticles in drug delivery. *Colloid and Interface Science Communications*, 35:100244, 2020.

[44] Ana Isabel Pérez-Jiménez, Danya Lyu, Zhixuan Lu, Guokun Liu, and Bin Ren. Surface-enhanced raman spectroscopy: benefits, trade-offs and future developments. *Chem. Sci.*, 11:4563–4577, 2020.

[45] Luis M. Liz-Marzán. Tailoring surface plasmons through the morphology and assembly of metal nanoparticles. *Langmuir*, 22(1):32–41, 2006.

[46] S. Iravani, H. Korbekandi, S.V. Mirmohammadi, and B. Zolfaghari. Synthesis of silver nanoparticles: chemical, physical and biological methods. *Research in Pharmaceutical Sciences*, 9(6):385–406, 2014.

[47] Andrea R. Tao, Susan Habas, and Peidong Yang. Shape control of colloidal metal nanocrystals. *Small*, 4(3):310–325, 2008.

[48] Kacper J. Lachowski, Huat Thart Chiang, Kaylyn Torkelson, Wenhao Zhou, Shuai Zhang, Jim Pfaendtner, and Lilo D. Pozzo. Anisotropic gold nanomaterial synthesis using peptide facet specificity and timed intervention. *Langmuir*, 39(45):15878–15888, 2023. PMID: 37910774.

[49] Sakshi Yadav Schmid, Kacper Lachowski, Huat Thart Chiang, Lilo Pozzo, Jim De Yoreo, and Shuai Zhang. Mechanisms of biomolecular self-assembly investigated through in situ observations of structures and dynamics. *Angewandte Chemie International Edition*, 62(48):e202309725, 2023.

[50] Mehrad Ansari and Andrew D. White. Serverless prediction of peptide properties with recurrent neural networks. *Journal of Chemical Information and Modeling*, 63(8):2546–2553, 2023.

[51] Yongtao Liu, Maxim Ziatdinov, and Sergei V. Kalinin. Exploring causal physical mechanisms via non-gaussian linear models and deep kernel learning: Applications for ferroelectric domain structures. *ACS Nano*, 16(1):1250–1259, 2022.

[52] Amanda A. Volk, Robert W. Epps, Daniel T. Yonemoto, Benjamin S. Masters, Felix N. Castellano, Kristofer G Reyes, and Milad Abolhasani. Alphaflow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nature Communications*, 14, 2023.

[53] Yibin Jiang, Daniel Salley, Abhishek Sharma, Graham Keenan, Margaret Mullin, and Leroy Cronin. An artificial intelligence enabled chemical synthesis robot for exploration and optimization of nanomaterials. *Science Advances*, 8(40):eabo2626, 2022.

[54] Kacper J. Lachowski, Kiran Vaddi, Nada Y. Naser, François Baneyx, and Lilo D. Pozzo. Multivariate analysis of peptide-driven nucleation and growth of au nanoparticles. *Digital Discovery*, 1:427–439, 2022.

[55] Maria Politi, Fabio Baum, Kiran Vaddi, Edwin Antonio, Joshua Vasquez, Brittany P. Bishop, Nadya Peek, Vincent C. Holmberg, and Lilo D. Pozzo. A high-throughput workflow for the synthesis of cdse nanocrystals using a sonochemical materials acceleration platform. *Digital Discovery*, pages –, 2023.

[56] Sadhan Samanta, Priyanka Sarkar, Santanu Pyne, Gobinda Prasad Sahoo, and Ajay Misra. Synthesis of silver nanodiscs and triangular nanoplates in pvp matrix: Photophysical study and simulation of uv–vis extinction spectra using dda method. *Journal of Molecular Liquids*, 165:21–26, 2012.

[57] Schneider Caroline, Rasband Wayne, and Eliceiri Kevin. Nih image to imagej: 25 years of image analysis, 2012.

[58] Allison Siehr, Bin Xu, Ronald A. Siegel, and Wei Shen. Colloidal stability versus self-assembly of nanoparticles controlled by coiled-coil protein interactions. *Soft Matter*, 15:7122–7126, 2019.

[59] Prashant K Jain, Nahil Sobh, Jeremy Smith, AbderRahman N Sobh, Sarah White, Jacob Faucheaux, and John Feser. nanoddscat, Jan 2014.

[60] Christina Boukouvala and Emilie Ringe. Wulff-based approach to modeling the plasmonic response of single crystal, twinned, and core–shell nanoparticles. *The Journal of Physical Chemistry C*, 123(41):25501–25508, 2019. PMID: 31681455.

[61] Y. J. Li, A. Savan, A. Kostka, H. S. Stein, and A. Ludwig. Accelerated atomic-scale exploration of phase evolution in compositionally complex materials. *Mater. Horiz.*, 5:86–92, 2018.

[62] Yun Yang, Wenfang Wang, Xingliang Li, Wei Chen, Nini Fan, Chao Zou, Xian Chen, Xiangju Xu, Lijie Zhang, and Shaoming Huang. Controlled growth of ag/au bimetallic nanorods through kinetics control. *Chemistry of Materials*, 25(1):34–41, 2013.

[63] Rok Mravljak and Aleš Podgornik. Simple and tailorable synthesis of silver nanoplates in gram quantities. *ACS Omega*, 8(2):2760–2772, 2023.

[64] Zao YI, Jian bo ZHANG, Hua HE, Xi bin XU, Bing chi LUO, Xi bo LI, Kai LI, Gao NIU, Xiu lan TAN, Jiang shan LUO, Yong jian TANG, Wei dong WU, and You gen YI. Convenient synthesis

of silver nanoplates with adjustable size through seed mediated growth approach. *Transactions of Nonferrous Metals Society of China*, 22(4):865–872, 2012.

[65] Cedric J. Gommes, Sebastian Jaksch, and Henrich Frielinghaus. Small-angle scattering for beginners. *Journal of Applied Crystallography*, 54(6):1832–1843, 2021.

[66] Brian Richard Pauw. Everything saxs: small-angle scattering pattern collection and correction. *Journal of Physics: Condensed Matter*, 25(38):383201, aug 2013.

[67] Cy M. Jeffries, Jan Ilavsky, Anne Martel, Stephan Hinrichs, Andreas Meyer, Jan Skov Pedersen, Anna V. Sokolova, and Dmitri I. Svergun. Small-angle X-ray and neutron scattering. *Nature Reviews Methods Primers*, 1(1):70, October 2021.

[68] Cécilia Gestraud, Pierre Roblin, Jeffrey F. Morris, Martine Meireles, and Yannick Hallez. Injection time controls the final morphology of nanocrystals during in situ-seeding synthesis of silver nanodisks. *CrystEngComm*, 22:1769–1778, 2020.

[69] Siyu Wu, Xiaobing Zuo, and Yugang Sun. Size-refocusing fitting of small-angle X-ray scattering from polydisperse nanoparticles for shape determination. *Journal of Applied Crystallography*, 56(6):1739–1750, Dec 2023.

[70] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[71] Kallum M. Koczkur, Stefanos Mourdikoudis, Lakshminarayana Polavarapu, and Sara E. Skrabalak. Polyvinylpyrrolidone (pvp) in nanoparticle synthesis. *Dalton Trans.*, 44:17883–17905, 2015.

[72] I. Washio, Y. Xiong, Y. Yin, and Y. Xia. Reduction by the end groups of poly(vinyl pyrrolidone): A new and versatile route to the kinetically controlled synthesis of ag triangular nanoplates. *Advanced Materials*, 18(13):1745–1749, 2006.

[73] Tufail Ahmad. Reviewing the tannic acid mediated synthesis of metal nanoparticles. *Journal of Nanotechnology*, 2014.

[74] Hailong Fan, Le Wang, Xunda Feng, Yazhong Bu, Decheng Wu, and Zhaoxia Jin. Supramolecular hydrogel formation based on tannic acid. *Macromolecules*, 50(2):666–676, 2017.

[75] Chen Chen, Hao Yang, Xiao Yang, and Qinghai Ma. Tannic acid: a crosslinker leading to versatile functional polymeric networks: a review. *RSC Adv.*, 12:7689–7711, 2022.

[76] Zhiyin Zhang, Huat T. Chiang, Ying Xia, Nicole Avakyan, Ravi R. Sonani, Fengbin Wang, Edward H. Egelman, James J. De Yoreo, Lilo D. Pozzo, and F. Akif Tezcan. Design of light- and chemically responsive protein assemblies through host-guest interactions. page 102407, 2025.

[77] Huat Thart Chiang, Zhiyin Zhang, Kiran Vaddi, F. Akif Tezcan, and Lilo D. Pozzo. Efficient analysis of small-angle scattering curves for large biomolecular assemblies using Monte Carlo methods. *Journal of Applied Crystallography*, 58(3):963–975, Jun 2025.

[78] Marta Carroni and Helen R. Saibil. Cryo electron microscopy to determine the structure of macromolecular complexes. *Methods*, 95:78–85, 2016. Integrated Structural Biology.

[79] Kaung Su Khin Zaw, Chen Ma, Zhongzhen Wang, Meisha L. Shofner, and Sankar Nair. Effects of graphene oxide membrane thickness reduction on microstructure and crossflow separation performance in kraft black liquor dewatering. *Chemical Engineering Science*, 281:119194, 2023.

[80] Ingrid Tessmer. *AFM Studies of Biomolecules*, pages 15–21. Springer New York, New York, NY, 2014.

[81] O. Glatter and O. Kratky. *Small Angle X-Ray Scattering*. Academic Press, New York, 1982.

[82] Cedric J. Gommes, Sebastian Jaksch, and Henrich Frielinghaus. Small-angle scattering for beginners. *Journal of Applied Crystallography*, 54(6):1832–1843, Dec 2021.

[83] Jan Skov Pedersen. Analysis of small-angle scattering data from colloids and polymer solutions: modeling and least-squares fitting. *Advances in Colloid and Interface Science*, 70:171–210, 1997.

[84] Jan Skov Pedersen. Form factors of block copolymer micelles with spherical, ellipsoidal and cylindrical cores. *Journal of Applied Crystallography*, 33(3-1):637–640, 2000.

[85] Huat Thart Chiang, Kiran Vaddi, and Lilo Pozzo. Data-driven exploration of silver nanoplate formation in multidimensional chemical design spaces. *Digital Discovery*, 3:2252–2264, 2024.

[86] Daniel K. Putnam, Edward W. Lowe, and Jens Meiler. Reconstruction of saxs profiles from protein structures. *Computational and Structural Biotechnology Journal*, 8(11):e201308006, 2013.

[87] D. Svergun, C. Barberato, and M. H. J. Koch. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography*, 28(6):768–773, Dec 1995.

[88] Jérôme Deumer, Brian R. Pauw, Sylvie Marguet, Dieter Skroblin, Olivier Taché, Michael Krumrey, and Christian Gollwitzer. Small-angle x-ray scattering: Characterization of cubic au nanoparticles using debye's scattering formula, 2022.

[89] Daniel P. Olds and Phillip M. Duxbury. Efficient algorithms for calculating small-angle scattering from large model structures. *Journal of Applied Crystallography*, 47(3):1077–1086, Jun 2014.

[90] S. Hansen. Calculation of small-angle scattering profiles using Monte Carlo simulation. *Journal of Applied Crystallography*, 23(4):344–346, Aug 1990.

[91] O Glatter. X-ray small angle scattering of molecules composed of subunits. *Acta Physica Austriaca*, 36(307-315), 1972.

[92] Cristiano Luis Pinto Oliveira, Manja Annette Behrens, Jesper Søndergaard Pedersen, Kurt Erlacher, Daniel Otzen, and Jan Skov Pedersen. A SAXS study of glucagon fibrillation. 387(1):147–161, 2009.

[93] Cassio Alves, Jan Skov Pedersen, and Cristiano Luis Pinto Oliveira. Modelling of high-symmetry nanoscale particles by small-angle scattering. 47(1):84–94, 2014.

[94] Cassio Alves, Jan Skov Pedersen, and Cristiano L. P. Oliveira. Calculation of two-dimensional scattering patterns for oriented systems. *Journal of Applied Crystallography*, 50(3):840–850, Jun 2017.

[95] E. Pantos and J. Bordas. Supercomputer simulation of small angle x-ray scattering, electron micrographs and x-ray diffraction patterns of macromolecular structures. *Pure and Applied Chemistry*, 66(1):77–82, 1994.

[96] Jan Skov Pedersen, Cristiano L.P. Oliveira, Henriette Baun Hübschmann, Lise Arleth, Søren Manniche, Nicolai Kirkby, and Hanne Mørck Nielsen. Structure of immune stimulating complex matrices and immune stimulating complexes in suspension determined by small-angle x-ray scattering. 102(10):2372–2380, 2012.

[97] Markus Kroemer, Iris Merkel, and Georg E. Schulz. Structure and catalytic mechanism of l-rhamnulose-1-phosphate aldolase,. *Biochemistry*, 42(36):10560–10568, 2003. PMID: 12962479.

[98] Shuai Zhang, Robert G. Alberstein, James J. De Yoreo, and F. Akif Tezcan. Assembly of a patchy protein into variable 2D lattices via tunable multiscale interactions. *Nature Communications*, 11(1), July 2020. Publisher: Springer Science and Business Media LLC.

[99] Avi Ginsburg, Tal Ben-Nun, Roi Asor, Asaf Shemesh, Lea Fink, Roee Tekoah, Yehonatan Levartovsky, Daniel Khaykelson, Raviv Dharan, Amos Fellig, and Uri Raviv. *D+* : software for high-resolution hierarchical modeling of solution x-ray scattering from complex structures. 52(1):219–242, 2019.

[100] Yuta Suzuki, Giovanni Cardone, David Restrepo, PabloD. Zavattieri, TimothyS. Baker, and F.Akif Tezcan. Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals. *Nature*, 533(7603):369–373, May 2016.

[101] Karen Manalastas-Cantos, Petr V. Konarev, Nelly R. Hajizadeh, Alexey G. Kikhney, Maxim V. Petoukhov, Dmitry S. Molodenskiy, Alejandro Panjkovich, Haydyn D. T. Mertens, Andrey Gruzinov, Clemente Borges, Cy M. Jeffries, Dmitri I. Svergun, and Daniel Franke. *ATSAS 3.0*: expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography*, 54(1):343–355, Feb 2021.

[102] A.L. Ksenofontov, M.V. Petoukhov, A.N. Prusov, N.V. Fedorova, and E.V. Shtykova. Characterization of tobacco mosaic virus virions and repolymerized coat protein aggregates in solution by smallangle xray scattering. *Biochemistry (Moscow)*, 85(3):310–317, 2020.

[103] Guang Yang, Xiang Zhang, Zdravko Kochovski, Yufei Zhang, Bin Dai, Fuji Sakai, Lin Jiang, Yan Lu, Matthias Ballauff, Xueming Li, Cong Liu, Guosong Chen, and Ming Jiang. Precise and reversible

protein-microtubule-like structure with helicity driven by dual supramolecular interactions. *Journal of the American Chemical Society*, 138(6):1932–1937, 2016. PMID: 26799414.

[104] Jeffrey D. Brodin, Sarah J. Smith, Jessica R. Carr, and F. Akif Tezcan. Designed, helical protein nanotubes with variable diameters from a single building block. *Journal of the American Chemical Society*, 137(33):10468–10471, 2015. PMID: 26256820.

[105] Linlu Zhao, Haoyang Zou, Hao Zhang, Hongcheng Sun, Tingting Wang, Tiezheng Pan, Xiumei Li, Yushi Bai, Shanpeng Qiao, Quan Luo, Jiayun Xu, Chunxi Hou, and Junqiu Liu. Enzyme-triggered defined protein nanoarrays: Efficient light-harvesting systems to mimic chloroplasts. *ACS Nano*, 11(1):938–945, 2017. PMID: 28051843.

[106] Jordan M. Fletcher, Robert L. Harniman, Frederick R. H. Barnes, Aimee L. Boyle, Andrew Collins, Judith Mantell, Thomas H. Sharp, Massimo Antognozzi, Paula J. Booth, Noah Linden, Mervyn J. Miles, Richard B. Sessions, Paul Verkade, and Derek N. Woolfson. Self-assembling cages from coiled-coil peptide modules. *Science*, 340(6132):595–599, 2013.

[107] Joona Mikkilä, Eduardo Anaya-Plaza, Ville Liljeström, Jose R. Caston, Tomas Torres, Andrés de la Escosura, and Mauri A. Kostiainen. Hierarchical organization of organic dyes and protein cages into photoactive crystals. *ACS Nano*, 10(1):1565–1571, 2016. PMID: 26691783.

[108] Shunzhi Wang, Andrew Favor, Ryan Daniel Kibler, Joshua Morris Lubner, Andrew J Borst, Nicolas Coudray, Rachel Redler, Huat Thart Chiang, William Sheffler, Yang Hsia, Zhe Li, Damian Charles Ekiert, Gira Bhabha, Lilo D Pozzo, and David Baker. Bond-centric modular design of protein assemblies. *bioRxiv*, 2024.

[109] Joshua A. Anderson, Jens Glaser, and Sharon C. Glotzer. Hoomd-blue: A python package for high-performance molecular dynamics and hard particle monte carlo simulations. *Computational Materials Science*, 173:109363, 2020.

[110] Dmytro Nykypanchuk, Mathew M. Maye, Daniel van der Lelie, and Oleg Gang. DNA-guided crystallization of colloidal nanoparticles. *Nature*, 451(7178):549–552, January 2008.

[111] Robert J. Macfarlane, Ryan V. Thaner, Keith A. Brown, Jian Zhang, Byeongdu Lee, SonBinh T. Nguyen, and Chad A. Mirkin. Importance of the DNA "bond" in programmable nanoparticle crystallization. *Proceedings of the National Academy of Sciences*, 111(42):14995–15000, October 2014.

[112] Melissa Rinaldin, Ruben W. Verweij, Indrani Chakraborty, and Daniela J. Kraft. Colloid supported lipid bilayers for self-assembly. *Soft Matter*, 15(6):1345–1360, 2019.

[113] Indrani Chakraborty, Daniel J. G. Pearce, Ruben W. Verweij, Sabine C. Matysik, Luca Giomi, and Daniela J. Kraft. Self-Assembly Dynamics of Reconfigurable Colloidal Molecules. *ACS Nano*, 16(2):2471–2480, February 2022.

[114] Matthew R. Jones, Robert J. Macfarlane, Byeongdu Lee, Jian Zhang, Kaylie L. Young, Andrew J. Senesi, and Chad A. Mirkin. DNA-nanoparticle superlattices formed from anisotropic building blocks. *Nature Materials*, 9(11):913–917, November 2010.

[115] Daniel Ortiz, Kevin L. Kohlstedt, Trung Dac Nguyen, and Sharon C. Glotzer. Self-assembly of reconfigurable colloidal molecules. *Soft Matter*, 10:3541–3552, 2014.

[116] Boris N. Khlebtsov, Vitaly A. Khanadeev, and Nikolai G. Khlebtsov. Determination of the size, concentration, and refractive index of silica nanoparticles from turbidity spectra. *Langmuir*, 24(16):8964–8970, 2008. PMID: 18590302.

[117] A. Gadomski, N. Kruszewska, and J.M. Rubi. Derivation of the refractive index of lipid monolayers at an air-water interface. *Optical Materials*, 93:1–5, 2019.

[118] Joshua A. Anderson, Jens Glaser, and Sharon C. Glotzer. Hoomd-blue: A python package for high-performance molecular dynamics and hard particle monte carlo simulations. *Computational Materials Science*, 173:109363, 2020.

[119] Runfang Mao, Brian Minevich, Daniel McKeen, Qizan Chen, Fang Lu, Oleg Gang, and Jeetain Mittal. Regulating phase behavior of nanoparticle assemblies through engineering of dna-mediated isotropic interactions. *Proceedings of the National Academy of Sciences*, 120(52):e2302037120, 2023.

[120] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

[121] Yandong Han, Ziyang Lu, Zhaogang Teng, Jinglun Liang, Zilong Guo, Dayang Wang, Ming-Yong Han, and Wensheng Yang. Unraveling the growth mechanism of silica particles in the stöber method: In situ seeded growth model. *Langmuir*, 33(23):5879–5890, 2017. PMID: 28514596.

[122] Jan Kobierski, Anita Wnętrzak, Anna Chachaj-Brekiesz, and Patrycja Dynarowicz-Latka. Predicting the packing parameter for lipids in monolayers with the use of molecular dynamics. *Colloids and Surfaces B: Biointerfaces*, 211:112298, 2022.

[123] Juewen Liu. Interfacing Zwitterionic Liposomes with Inorganic Nanomaterials: Surface Forces, Membrane Integrity, and Applications. *Langmuir*, 32(18):4393–4404, May 2016.

[124] Hairong Wang, Jelena Drazenovic, Zhenyu Luo, Jiangyue Zhang, Hongwen Zhou, and Stephanie L. Wunder. Mechanism of supported bilayer formation of zwitterionic lipids on SiO2 nanoparticles and structure of the stable colloids. *RSC Advances*, 2(30):11336, 2012.

[125] Selver Ahmed and Stephanie L. Wunder. Effect of High Surface Curvature on the Main Phase Transition of Supported Phospholipid Bilayers on $SiO_2$ Nanoparticles. *Langmuir*, 25(6):3682–3691, March 2009.

[126] Hend I. Alkhammash, Nan Li, Rémy Berthier, and Maurits R. R. De Planque. Native silica nanoparticles are powerful membrane disruptors. *Physical Chemistry Chemical Physics*, 17(24):15547–15560, 2015.

[127] Mary Elizabeth Beattie, Sarah L. Veatch, Benjamin L. Stottrup, and Sarah L. Keller. Sterol structure determines miscibility versus melting transitions in lipid vesicles. *Biophysical Journal*, 89(3):1760–1768, 2005.

[128] Nicolò Paracini, Philipp Gutfreund, Rebecca Welbourn, Juan Francisco Gonzalez-Martinez, Kexin Zhu, Yansong Miao, Nageshwar Yepuri, Tamim A. Darwish, Christopher Garvey, Sarah Waldie, Johan Larsson, Max Wolff, and Marité Cárdenas. Structural characterization of nanoparticle-supported

lipid bilayer arrays by grazing incidence x-ray and neutron scattering. *ACS Applied Materials & Interfaces*, 15(3):3772–3780, 2023. PMID: 36625710.

[129] Haley D. Hill, Jill E. Millstone, Matthew J. Banholzer, and Chad A. Mirkin. The Role Radius of Curvature Plays in Thiolated Oligonucleotide Loading on Gold Nanoparticles. *ACS Nano*, 3(2):418–424, February 2009.

[130] Beth A. Lindquist. Inverse design of equilibrium cluster fluids applied to a physically informed model. *The Journal of Chemical Physics*, 154(17):174907, 05 2021.

[131] Alexander Stukowski. Visualization and analysis of atomistic simulation data with OVITO-the Open Visualization Tool. *MODELLING AND SIMULATION IN MATERIALS SCIENCE AND ENGINEERING*, 18(1), JAN 2010.

[132] John D. Le, Yariv Pinto, Nadrian C. Seeman, Karin Musier-Forsyth, T. Andrew Taton, and Richard A. Kiehl. Dna-templated self-assembly of metallic nanocomponent arrays on a surface. *Nano Letters*, 4(12):2343–2347, 2004.

[133] RobertJ. Macfarlane, MatthewR. Jones, AndrewJ. Senesi, KaylieL. Young, Byeongdu Lee, Jinsong Wu, and ChadA. Mirkin. Establishing the design rules for dna-mediated programmable colloidal crystallization. *Angewandte Chemie International Edition*, 49(27):4589–4592, 2010.

[134] S. Karthika, T. K. Radhakrishnan, and P. Kalaichelvi. A review of classical and nonclassical nucleation theories. *Crystal Growth & Design*, 16(11):6663–6681, 2016.