Interactions at the interface between proteins and minerals

Amy Stegmann

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
James De Yoreo, Chair
Francois Baneyx
Shuai Zhang

Program Authorized to Offer Degree:
Molecular Engineering and Sciences

University of Washington

Abstract

Interactions at the interface between proteins and minerals

Amy Stegmann

Chair of the Supervisory Committee:
James De Yoreo
Department of Chemistry
Department of Materials Science and Engineering

Interactions at the interfaces between proteins and minerals drive both protein binding onto the surface of minerals and mineral formation on the surface of proteins. Understanding these interactions is fundamental to the rational design of self-assembling hierarchical structures. De novo designed proteins enable the precise placement of chemical functional groups in three-dimensional space and are thus ideal for investigating organic-mineral interactions. This research has been targeted to address both protein assembly and mineral formation. We investigated the role of charge patchiness in proteins that drive the formation of mineral crystals from precursors in solution. By designing charged template proteins, we gain insight into the assembly mechanisms of titanium oxides from titanium(IV) bis(ammonium lactate)dihydroxide (TiBALDH), where surface-displayed carboxyl and amine groups direct nucleation at room temperature, demonstrating sequence-driven control over titania nucleation phase and spatial organization using homo-oligomeric protein assemblies with D3 symmetry. We characterized the formation and growth of titanium dioxide as directed by the surface charge of various protein assemblies and developed the design principle that alternating positive and negative regions is optimal for forming titanium dioxide. Our research demonstrates that the precise control over functional group placement available to de novo designed proteins enables access to novel assemblies and can influence the crystallinity and morphology of mineral formation. We also employed machine learning and conventional analysis of high-speed atomic force microscopy data to better describe the assembly process of protein nanorods into a liquid crystal arrangement on the surface of mica and the interactions and solution conditions required for that liquid crystal to form. The symmetry of the mineral substrate and the resultant solution structure at the mineral surface influences the symmetry of the protein assembly. The designed interface of the protein allows access to an ordered assembly at the surface. These findings highlight the potential of de novo designed proteins to serve as precise scaffolds for controlling mineral nucleation and assembly, paving the way for the development of tailored bioinspired materials with tunable properties.

1. Introduction:

1.1 Nucleation

This section has been adapted from De Yoreo & Vekilov[1]. Particles in solution require a driving force to move from a solvated state to a solid, crystalline phase. Precipitation to a new phase is only thermodynamically favorable when the free energy of this new phase in the final solution is less than that of its components in the original solvated phase at equilibrium. This situation occurs when system is in a supersaturated state ($\sigma > 1$) which is a true when the activity product of the reactants (AP) surpasses the equilibrium activity product (Ksp). The chemical potential (μ) is the driving force for nucleation and is affected by the degree of supersaturation, as outlined in the equations below. The greater the chemical potential of the crystallizing species, the larger the driving force for crystallization. In this context, k_B represents Boltzmann's constant, and T indicates the absolute temperature.

Equation 1.1

$$\Delta\mu = k_B T ln\sigma$$

Equation 1.2

$$\sigma \equiv ln\{AP/K_{sp}\}$$

The above equations are only applicable to the bulk material because surface molecules are less bound to their neighbors than are those in the bulk. Thus, surface molecules contribute more to the free energy of the new phase than molecules in the bulk. The interfacial free energy (α) is the difference in free energies of the surface and the bulk. This term opposes nucleation and always carries a positive sign. The interfacial free energy can be defined as the energy required to form and stabilize a surface or boundary between two dissimilar phases. In the new phase, very small particles are dominated by interfacial free energy which makes them unstable and likely to dissolve, while larger particles are

governed by changes in chemical potential and tend to grow. The smaller the interfacial energy, the smaller the critical size and the more likely nucleation becomes for any given supersaturation. Consequently, by varying either the solution composition or the supersaturation, the probability of nucleation can be manipulated.

Water molecules play a significant role in the energetics of crystallization. Either releasing or trapping of water molecules significantly affects both enthalpy and entropy. The change in enthalpy varies from system to system because it is influenced by the strength of the bonds between water and solute molecules, as well as the bonds that form in the crystals. However, the change in entropy is consistent with the energy required to form ice from solution, with a positive value for release and a negative value for trapping. Furthermore, the structuring of water plays a crucial role in the free energy barrier for crystallization, and the kinetics of water restructuring around a solute molecule seem to affect the overall crystallization rate.

The free energy barrier of nucleation (Δg_n) can be expressed in terms of the interfacial energy and supersaturation.

Equation 1.3

$$\Delta g_n = (16/3)\pi\alpha^3(\Omega/k_B T\sigma)^2$$

$$\propto \alpha^3/\sigma^2$$

Combining all the factors other than interfacial energy and supersaturation into the coefficient B shows that the nucleation rate can also be expressed in terms of the interfacial energy and supersaturation, where A is a geometric term.

Equation 1.4

$$J_n = Aexp(-B\alpha^3/\sigma^2)$$

Salt screening, pH, redox potential, and ionic strength all contribute to the saturation limit. Organisms in nature can leverage the ability to regulate these conditions to direct nucleation of inorganic phases. Causing local gradients or fluctuations in solution conditions, proteins can manipulate the chemical potential to promote nucleation and can exert an influence over the size, shape, and polymorph of the formed mineral. In this way, proteins can serve as templates which reduce the energy barrier to nucleation. Protein templates provide a preferential nucleation site which is more energetically favorable for nucleation than in the bulk solution. Geometrically and electrostatically compatible surfaces to a target mineral surface will reduce the off rate of monomers into solution.

1.2 Organic templates for mineralization

Templates for nucleation leverage the interfacial energy between a crystal nucleus and a solid substrate being lower than the interfacial energy between crystal and the solution. This is because the molecules in the crystal can form stronger bonds with the molecules of the template than the bonds of solvation. Protein templates can thus reduce the interfacial free energy of the new mineral phase to promote nucleation and growth. Moreover, in principle *de novo* designed proteins can be far more effective than small molecules or unstructured peptides because they can present a surface area which is both large enough to meet or exceed the critical nucleus size and exhibit a regular pattern of binding sites.

Binding affinity has been shown to correlate with the promotion of nucleation[2]. Building on this principle, others have taken the approach of Lattice Matching where a protein is designed to specifically to match a targeted crystal facet of titania using a similar strategy as in the Mica18 protein design[3]. Alternatively, others screened protein libraries by using yeast or phage display[4] and experimentally identified proteins which bind, allowing further investigation into how protein binding influences mineralization.

In an alternative strategy to matching a crystal lattice as a nucleation template, regions of charge can be manually arrayed to investigate the role of electrostatics in nucleation. Wallace *et al* used patterns of carboxyl or amine decorated regions to create positively and negatively arrayed regions on a surface[5]. The charge involved was sufficient to drive the nucleation of silica on the surface, and AFM studies demonstrated that the silica preferentially formed at the interfaces where positive and negative charges met (Fig. 1.1).

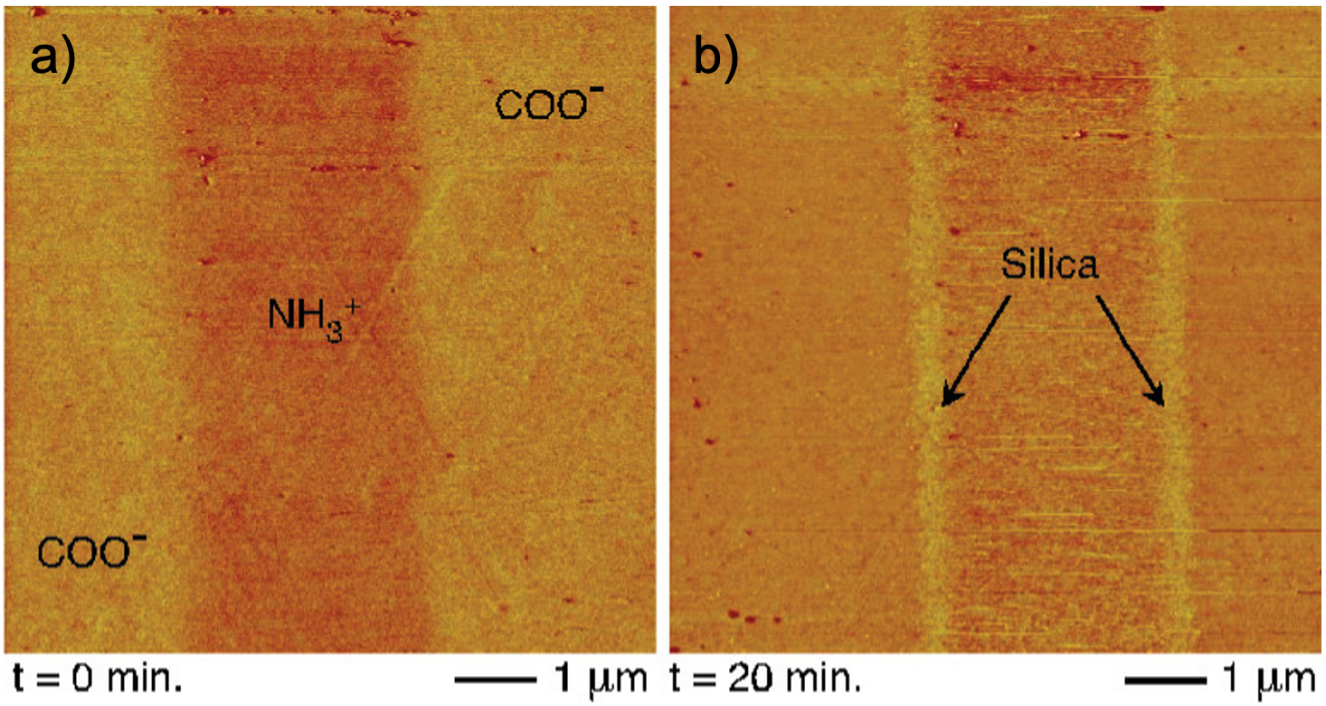

Figure 1.1: Example of *in vitro* silica mineralization on a charged surface:
AFM images showing a) positive and negatively charged regions and b) silica formed at the interface between positive and negatively Charged regions[6].

Silaffins are a naturally occurring type of protein that direct the formation of silica in diatoms. They template silica formation, which begins with silica nucleation on the protein surface at a site where negative and positive residues are juxtaposed. Purified silaffins have been shown to drive formation of silica, and even titania in solution (Fig. 1.2).

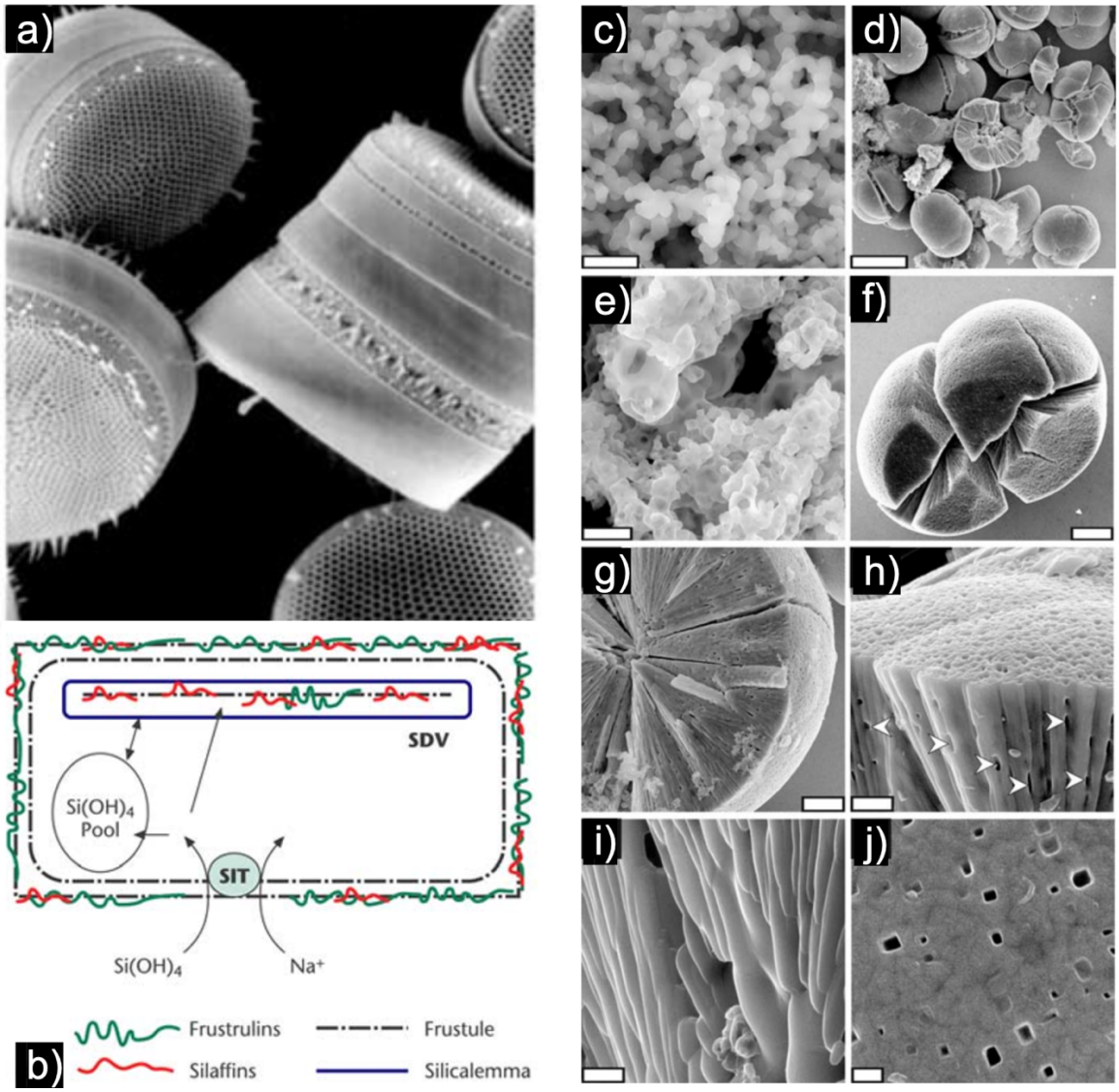

Figure 1.2: Examples of silaffin templated mineral growth:
a) Scanning electron micrograph (SEM) of silica frustules of the diatom *Thallassiosira eccentrica.* b) diagram of silica biogenesis in diatoms. Inside the cell, soluble silica is transported to the silica deposition vesicle (SDV), where it undergoes polycondensation under acidic conditions to form insoluble amorphous silica. Silaffin proteins regulate insoluble silica formation, while frustulins, polysaccharides, and lipids stabilize the mature silica wall. The completed valve is then exocytosed onto the cell surface. a) - b) taken from[7]. c) – j) SEM showing Rutile TiO^2 formed using silaffins[8]

The findings shown in Figure 1.1 and Figure 1.2 suggest that silica and titania form under similar conditions and that electrostatics play an important role in the formation mechanism. This study is the first to investigate charge patchiness on proteins to template nucleation in solution.

1.3 TiO_2 theory

Titanium bisammonium lactatodihydroxide (TiBALDH) is a widely used precursor for nucleating titania because it is water soluble and commercially available. Titanium ions require negatively charged ligands to be water soluble because without the electrostatic stability provided by the ligands Ti ions will rapidly hydrolyze into insoluble compounds[9]. Using biomolecules for nucleation generally requires water solubility, because biomolecule self-assembly requires the hydrophobic effect and thus an aqueous solution. Buffers, especially amine containing buffers like Tris or phosphate buffers can interact with TiBALDH and cause precipitation, additionally pH and concentration of TiBALDH can drive nucleation[9-13]. It is therefore advisable to use a pH below 9[10], and a concentration of TiBALDH around 10 mM[11-13].

Hydrolysis is the mechanism for growth of TiO_2 from TiBALDH, and there is a metalorganic cofactor $[Ti(Lactate)3]^{2-}$ which is also a product in the reaction. The chemical equation for the reaction of TiBALDH to TiO_2 is shown in Equation 1.5[14,15]:

Equation 1.5

$$3[Ti_4O_4(lactate)_8]^{8-} + 8NH_4^+ \rightleftarrows 8[Ti(lactate)_3]^{2-} + 4TiO_2 + 8NH_3 + 4H_2O$$

The metalorganic cofactor (Fig. 1.3) is rarely observed via TEM as excess sample solution is typically wicked away after an incubation period and the cofactor is usually removed in this step of sample preparation.

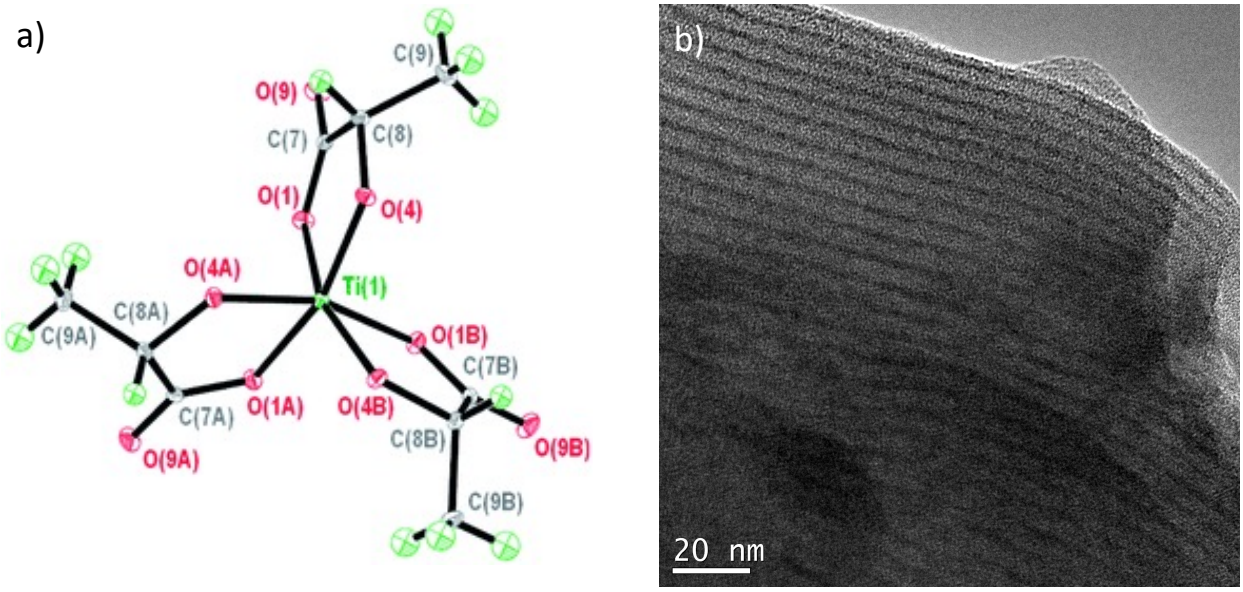

Figure 1.3: Example of metalorganic byproduct
a) molecular structure of the $[Ti(Lactate)3]^{2-}$ anion[15], b) TEM image obtained of the metalorganic byproduct of the reaction from TiBALDH.

1.4 Peptide interactions with TiBALDH

Peptides have been shown to direct nucleation of TiO_2[16-19]. Looking for similarities between titania nucleating peptides (Fig. 1.4), it is apparent that repeating positive charges are a common motif[16,19]. Slightly different sequences can influence the phase of titania, where XRD confirms a shift between percentage of the anatase and β polymorphs of titania[19].

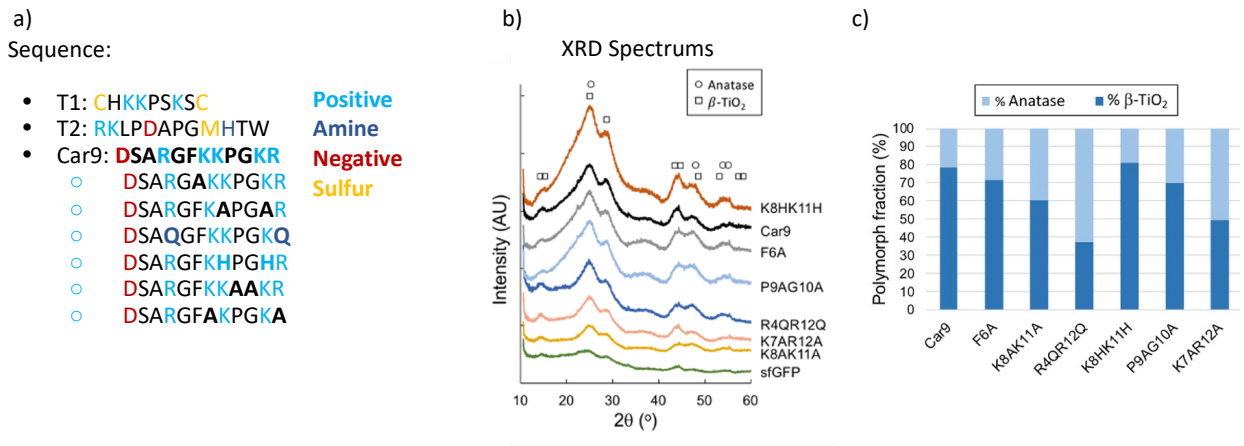

Figure 1.4: The effect of sequence on phase of titanium dioxide that peptides direct
a) sequences of T1, T2, and Car9 peptides as well as the point mutations studies by Helner *et al* showing a trend in patches of positive residues in the peptides that are reported to direct titania nucleation b) XRD spectrums showing the effect of point mutations of Car9 on different phases that nucleate, c) a breakdown of the percent of each phase that is nucleated by the point mutations to Car9. a)-c)[19].

The ability to direct different phases of a material during the self-assembly process is desirable because different phases have different materials properties like conductivity. Thus, studies into the mechanisms of nucleation and growth for materials like titania which can form multiple phases are particularly relevant to developing the fundamental understanding needed for directed self-assembly of complicated multi-phase systems.

1.5 Surface binding via lattice matching

It has been shown that a periodic arrangement of negatively charged groups on a protein rod will align along thermodynamically favorable directions that coincide with the sub-lattice of potassium ions on mica[20]. Mica N is a designed helical repeat protein, with the N representing the number of repeats, which has been expressed with up to 34 repeated alpha helices. The surface binding region of the protein was designed to contain glutamate residues which have carboxylate functional groups that match the crystal sub-lattice of potassium ions on the [001] surface of muscovite mica (Fig 1.5).

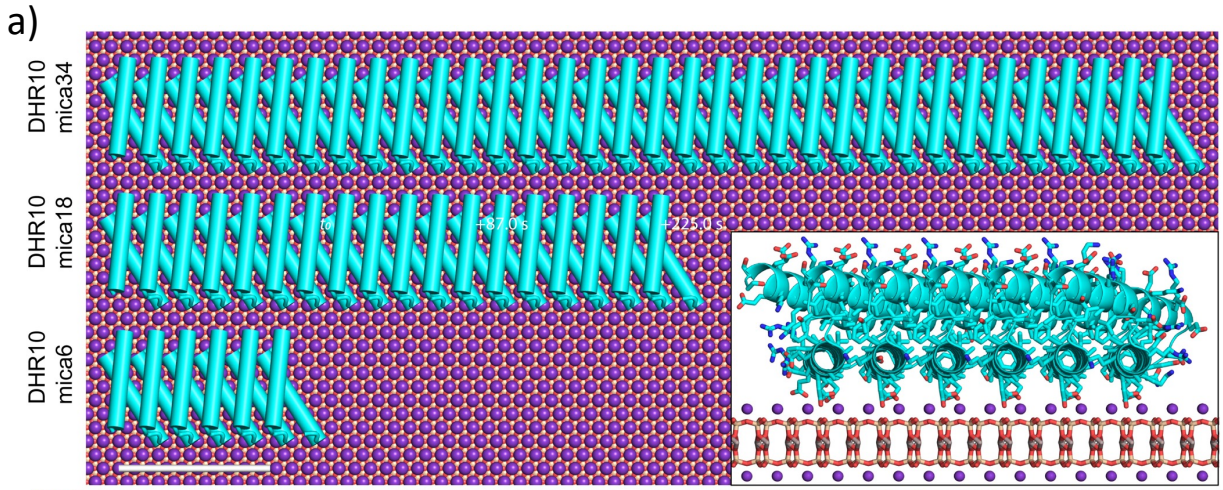

Figure 1.5: MicaN protein design principles
a) Model of Mica 6, Mica 18, and Mica 34 arrayed on a muscovite mica [001] surface. Inset shows a cross section of carboxylate groups that are matched to the repeating lattice of the mica surface. Model generated by Dr. Harley Pyles.

A change in the crystal structure of the substrate can result in very different arrangements on the surface of the substrate[21]. For example: the fluorophlogopite mica [001] surface has a hexagonal structure which results in three angles of alignments of rods on the surface with a roughly equal distribution of alignments, but the muscovite mica surface [001] has a slightly different bond structure in one direction such that the rods on the surface have one strongly preferred angle of alignment[20]. These bond structures give rise to different patterns of water at the mineral-solution interface (Fig. 1.6) which then leads to different assembly patterns of MicaN on the surface of mica.

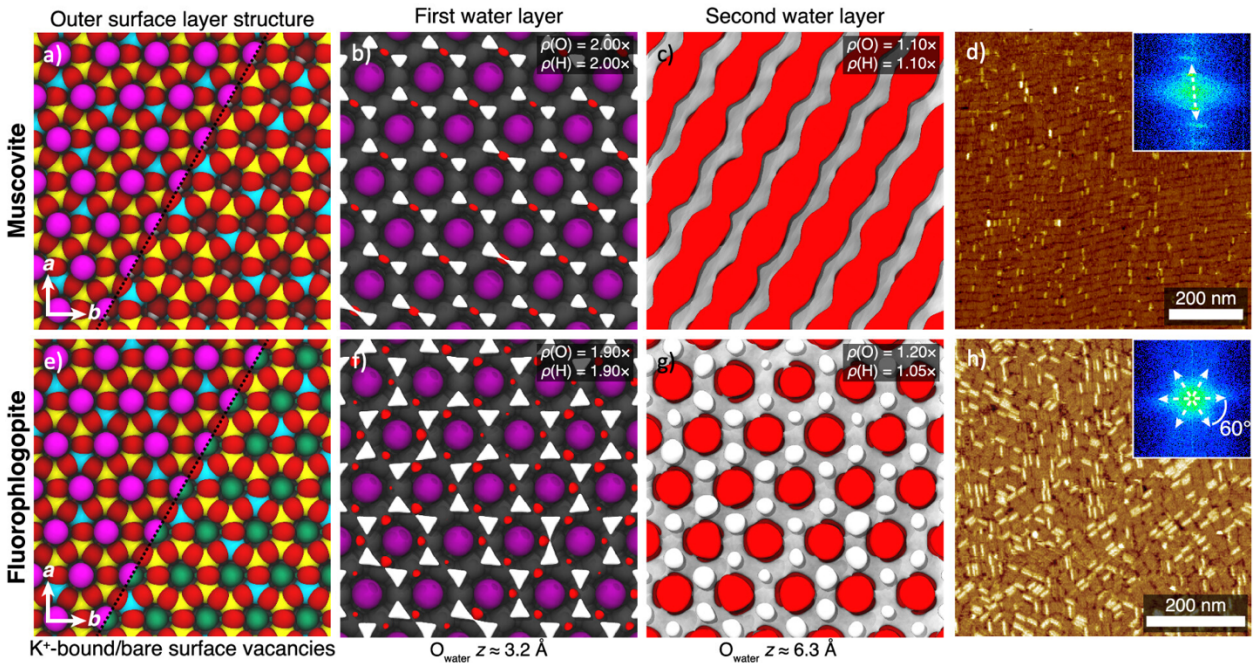

Figure 1.6: Different mica crystal phases direct different protein assemblies
a),e) computer renderings of muscovite and fluorophlogopite mica showing subtle differences between the crystal structures of the two phases, b),f) organized "Stern-layer" water structures where muscovite mica has a unidirectionally oriented first layer of solvent and fluorophlogopite has hexagonally ordered first-layer waters c),g) structure of the second water layer for muscovite and fluorophlogopite: muscovite mica exhibits ribbons of oxygen density in the second layer through hydrogen bonding and the hexagonal structure for fluorophlogopite produces circular oxygen densities directly above the K+ ions d),h) AFM showing the different alignments of protein rods on the surface of muscovite and fluorophlogopite mica with inset FFT showing one or three dominate angles of alignment a)-h) [21]

Mobility affects the order of a 2D smectic phase of rods[22]. Potassium chloride concentration in the solution affects the mobility of the rods on the surface of the substrate, as the rods are designed to interact with potassium ions on the surface of mica[20]. Video of the Mica18 rods aligning on the surface was obtained and analyzed for 100 mM KCl[23], where the mobility of rods on the surface is low, and the noise content of the collected data is also low. At higher concentrations of KCl the rods on the surface are far more mobile and noise from high salt buffers negatively affects the image quality. For this reason, videos of the protein rods assembling on mica in the presence of 3 M KCl have not previously been analyzed.

1.6 2D Smectic order

In solution, particles can assemble into a type of liquid crystalline array known as a smectic phase. Particles in a 3D smectic arrangement are arranged in layers that can slide against each other, but when reduced to a 2D arrangement of particles on a surface, if this phase exists, it is better described by ordered rows and columns (Fig. 1.7).

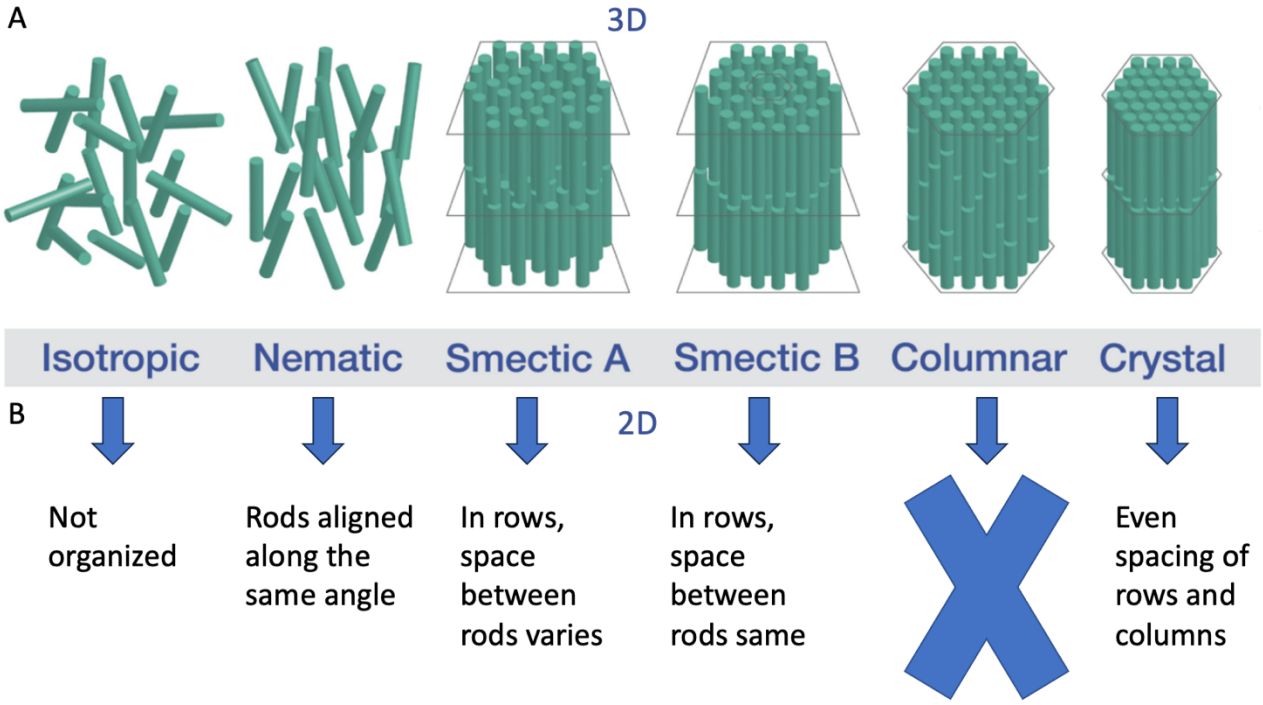

Figure 1.7: Descriptions of smectic phases:
a) smectic phases of hard rods and how they appear in 3D[24] b) descriptions of how the various phases reduce to 2D. Note that columnar is a special 3D smectic case where rods slide along columns instead of along layers and thus would be reduced to a nematic array in 2D.

The arrangement of the protein MicaN on the surface of M type mica can be described as a 2D smectic arrangement ordered into rows and columns. However, hard rods have previously only been shown to form a 2D smectic arrangement in the presence of repulsive interactions at the rod ends. Without this repulsive interaction a nematic phase is the most ordered phase rods can form in 2D. Yet the micaN surface both lacks repulsive forces at the ends of the rods and forms a 2D smectic array.

1.7 Co-assembly on a surface

Colloidal co-assembly of multiple species can result in interesting assemblies that are not possible with a single species of monodisperse colloids[24,25]. By introducing a second population of particle, we could gain access to additional formations and assemblies. Multiple sizes of particle are not always miscible with each other[25,26]. Hard rod assemblies are significantly affected by aspect ratio: changing the aspect ratio can affect the thermodynamically accessible phases of rod systems[27,28]. Different sizes and shapes of hard particles tend to phase separate, and the mechanism for forming clusters can vary based on the type of polydispersity[29].

1.8 References

1. de Yoreo, J. J. Principles of Crystal Nucleation and Growth. *Rev Mineral Geochem* **54**, 57–93 (2003).
2. Davies, P. L. Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. *Trends Biochem Sci* **39**, 548–555 (2014).
3. Davila-Hernandez, F. A. *et al.* Directing polymorph specific calcium carbonate formation with de novo protein templates. *Nature Communications 2023 14:1* **14**, 1–11 (2023).
4. Douglas, T. & Young, M. Viruses: Making friends with old foes. *Science (1979)* **312**, 873–875 (2006).
5. Lee, J. R. I. *et al.* Cooperative Reorganization of Mineral and Template during Directed Nucleation of Calcium Carbonate. *The Journal of Physical Chemistry C* **117**, 11076–11085 (2013).
6. Wallace, A. F., DeYoreo, J. J. & Dove, P. M. Kinetics of silica nucleation on carboxyl- and amine-terminated surfaces: Insights for biomineralization. *J Am Chem Soc* **131**, 5244–5250 (2009).
7. Koester, J., Brownlee, C. & Taylor, A. R. Algal Calcification and Silicification. *Encyclopedia of Life Sciences* 1–10 (2016) doi:10.1002/9780470015902.A0000313.PUB2.
8. Kröger, N. *et al.* Bioenabled Synthesis of Rutile (TiO2) at Ambient Temperature and Neutral pH. *Angewandte Chemie International Edition* **45**, 7239–7243 (2006).
9. Katsumata, K. I. *et al.* Preparation of TiO2 Thin Films Using Water-soluble Titanium Complexes and Their Photoinduced Properties. *Photochem Photobiol* **87**, 988–994 (2011).
10. Puddu, V., Slocik, J. M., Naik, R. R. & Perry, C. C. Titania binding peptides as templates in the biomimetic synthesis of stable titania nanosols: Insight into the role of buffers in peptide-mediated mineralization. *Langmuir* **29**, 9464–9472 (2013).
11. Forgács, A. *et al.* Kinetic Model for Hydrolytic Nucleation and Growth of TiO2 Nanoparticles. *Journal of Physical Chemistry C* **122**, 19161–19170 (2018).
12. Seisenbaeva, G. A., Daniel, G., Nedelec, J. M. & Kessler, V. G. Solution equilibrium behind the room-temperature synthesis of nanocrystalline titanium dioxide. *Nanoscale* **5**, 3330–3336 (2013).
13. Hernández-Gordillo, A., Hernández-Arana, A., Campero-Celis, A. & Vera-Robles, L. I. TiBALDH as a precursor for biomimetic TiO $_2$ synthesis: stability aspects in aqueous media. *RSC Adv* **9**, 34559–34566 (2019).
14. Hernández-Gordillo, A., Hernández-Arana, A., Campero-Celis, A. & Vera-Robles, L. I. TiBALDH as a precursor for biomimetic TiO2 synthesis: stability aspects in aqueous media. *RSC Adv* **9**, 34559–34566 (2019).
15. Seisenbaeva, G. A., Daniel, G., Nedelec, J. M. & Kessler, V. G. Solution equilibrium behind the room-temperature synthesis of nanocrystalline titanium dioxide. *Nanoscale* **5**, 3330–3336 (2013).
16. Bedwell, G. J. *et al.* Selective Biotemplated Synthesis of TiO 2 Inside a Protein Cage. *Biomacromolecules* **16**, 214–218 (2015).
17. Ma, J. *et al.* Controlling Mineralization with Protein-Functionalized Peptoid Nanotubes. *Advanced Materials* **35**, 2207543 (2023).
18. Pushpavanam, K., Hellner, B. & Baneyx, F. Interrogating biomineralization one amino acid at a time: amplification of mutational effects in protein-aided titania morphogenesis through reaction-diffusion control. *Chemical Communications* **57**, 4803–4806 (2021).
19. Hellner, B., Stegmann, A. E., Pushpavanam, K., Bailey, M. J. & Baneyx, F. Phase Control of Nanocrystalline Inclusions in Bioprecipitated Titania with a Panel of Mutant Silica-Binding Proteins. *Langmuir* **36**, 8503–8510 (2020).
20. Pyles, H., Zhang, S., De Yoreo, J. J. & Baker, D. Controlling protein assembly on inorganic crystals through designed protein interfaces. *Nature* vol. 571 251–256 Preprint at https://doi.org/10.1038/s41586-019-1361-6 (2019).

21. Alberstein, R. G. *et al.* Discrete Orientations of Interfacial Waters Direct Crystallization of Mica-Binding Proteins. *Journal of Physical Chemistry Letters* **14**, 80–87 (2023).
22. Tu, X. *et al.* Phase diagram of a bidispersed hard-rod lattice gas in two dimensions. *Europhys Lett* **112**, 66002 (2016).
23. Zhang, S. *et al.* Rotational dynamics and transition mechanisms of surface-adsorbed proteins. *Proc Natl Acad Sci U S A* **119**, e2020242119 (2022).
24. De Braaf, B., Oshima Menegon, M., Paquay, S. & Van Der Schoot, P. Self-organisation of semi-flexible rod-like particles. *J Chem Phys* **147**, (2017).
25. Xia, Y. *et al.* Self-assembly of self-limiting monodisperse supraparticles from polydisperse nanoparticles. *Nature Nanotechnology 2011 6:9* **6**, 580–587 (2011).
26. Singh, P. *et al.* Molecular-like hierarchical self-assembly of monolayers of mixtures of particles. *Scientific Reports 2014 4:1* **4**, 1–7 (2014).
27. Horsch, M. A., Zhang, Z. & Glotzer, S. C. Simulation studies of self-assembly of end-tethered nanorods in solution and role of rod aspect ratio and tether length. *J Chem Phys* **125**, 184903 (2006).
28. Bates, M. A. & Frenkel, D. Phase behavior of two-dimensional hard rod fluids. *J Chem Phys* **112**, 10034 (2000).
29. Schilling, T., Miller, M. A. & Van Der Schoot, P. Percolation in suspensions of hard nanoparticles: From spheres to needles. *EPL* **111**, 044901 (2015).

2. Using patches of charge to direct the nucleation of titanium dioxide in aqueous solution at STP

The material in this chapter is partially reproduced from a manuscript in preparation by Stegmann *et al.*

2.1 Abstract

De novo designed proteins enable the precise placement of functional groups in three-dimensional space, which can be leveraged to guide nucleation and growth of inorganic phases to form complex hierarchical nanostructures. Proteins have diverse functional groups which enable them to access an array of applications. We reasoned that leveraging a strict control of chemical moieties in building blocks by designing charged template proteins would enable insight into the mechanism of nucleation for titanium oxides from the precursor titanium(IV) bis(ammonium lactate)dihydroxide (TiBALDH). Surface display patterns of carboxylic acids and amines are shown to direct nucleation of TiBALDH in solution at room temperature. Understanding the effects that spatial presentation of different amino acid moieties has on the nucleated crystal phase and structure will facilitate the intentional design of self-assembled inorganic materials. Herein we demonstrate the influence that sequence can have on the phase of nucleated titania as well as the spatial control over nucleation that we can leverage using homo oligomeric *de novo* protein assemblies. *De novo* designed protein homo oligomers with D3 symmetry which present a favorable surface for nucleation are presented as templates in this work. We analyze the mechanism of growth and phase as it is influenced by *de novo* designed proteins.

2.2 Introduction

Fundamental understanding of the formation of oxides is important to be able to design materials with finely tuned materials properties. The main ways to influence a material's properties are by controlling the material's phase and/or morphology and by incorporating defects. Thus, guiding nucleation and growth at the atomic scale is ideal to form a desired material. Using biological materials as a template or catalyst for mineral formation is an emerging field in materials science to leverage the fine control over nucleation and growth that biomineralizing organisms exhibit[1].

Silica is a material that is biomineralized in nature[2,3], and biomolecules have been isolated and used to nucleate silica *in vitro*[4-7]. Phage display of peptides has also led to the discovery of peptides that are not naturally occurring but have been shown to nucleate silica[6-9], and these same isolated proteins and peptides have been shown to nucleate titania[10-14].

Titania is similar enough to silica that it can be grown on silica and the studies on nucleation of silica using biomolecules are relevant. Titania is an ideal model system because: it is an energetically relevant oxide, has multiple phases, and has the aqueous precursor TiBALDH. Although TiBALDH is a titanium salt, the titanium ions are bound to lactate groups and the compound has been crystalized and identified as $[Ti_4O_4(lactate)_8]^{8-}$ (Eq. 2.1) [15].

Equation 2.1

$$3[Ti_4O_4(lactate)_8]^{8-} + 8NH_4^+ \rightleftarrows 8[Ti(lactate)_3]^{2-} + 4TiO_2 + 8NH_3 + 4H_2O$$

De novo designed proteins can be designed to have arbitrary morphology and functional group placement[16-18]. Proteins have been used to template nucleation of various crystal systems[19-21].

We hypothesize that highly charged proteins promote a concentration gradient of charged species near the protein which generates a supersaturated solution from which oxide particles precipitate. Herein we employ *de novo* designed proteins as a substrate to direct nucleation of titania because of the ability

to arbitrarily place regions of charges in three dimensions.

2.3 Results

We investigated the role of 3D charged substrate regions in directing nucleation of titanium dioxide from TiBALDH using *de novo* designed proteins. Two previously designed homo-oligomers, ACC5 and A113, were designed to have highly charged surfaces[22]. The surface of A113 was modified so that the ends of the protein are more negative, giving A113-N. A positive band of lysine residues was added around the center of the protein complex to make A113-KM2N. Bovine serum albumen (BSA) is a native protein which was chosen as a control because it contains regions of positive and negative charge in proximity with each other (Fig. 2.1).

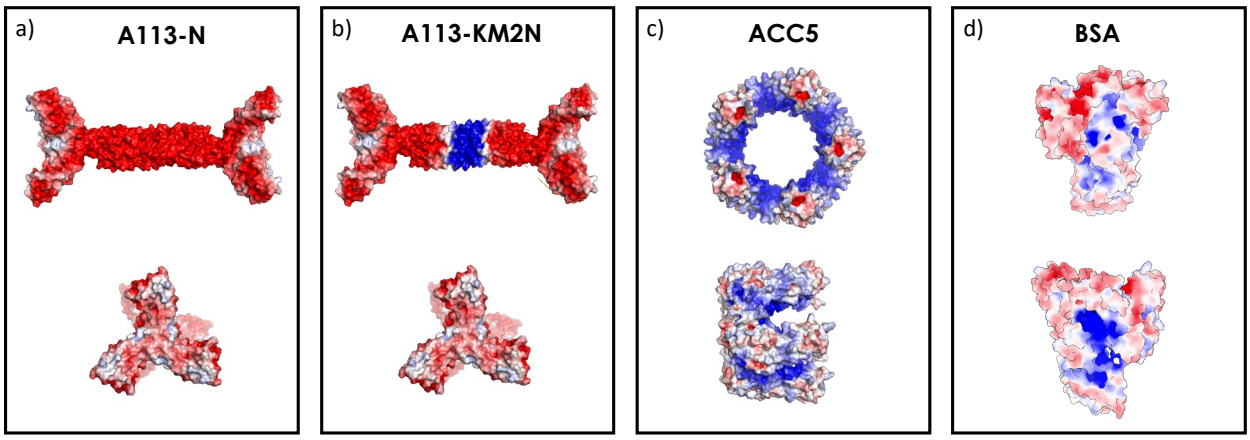

Figure 2.1: Schematic showing the charge distribution of the proteins used in this study.
a) side and end view of the negatively charged A113-N b) side and end view showing the positive band around the center of A113-KM2N c) top and side view of ACC5 d) front and back view of BSA. For all schematics red regions are negatively charged, white regions are neutral, and blue regions are positively charged.

A113-N and A113-KM2N are 10 nm in length and are homo oligomers made up of 6 subunits which assemble in D3 symmetry, and each has a mass of 21.46 kD. ACC5 is a cyclic homo oligomer that is 9 nm in diameter and made up of 5 subunits that have C5 symmetry and is 29.2 kD. BSA is a 69.3 kD monomeric protein. This study includes multiple types of symmetry and overall morphology to probe the effect different arrangements of charged regions have on the nucleated mineral.

20 mM TiBALDH in water was measured to have a pH of 6.5. The isoelectric point (pI), or mean of all pKa values within the molecule determines the net charge of a molecule at a given pH. At pH 6.5 the amine functional group on lysine residues (pkA of 10.5 ± 1.1) will be positively charged[23] and the carboxylate functional group on glutamate (pKa of 4.2 ± 0.9) will be negatively charged[23]. The pI's of the proteins in this study are: A113-N is 4.06, A113-KM2N is 4.49, ACC5 is 9.16 and BSA is 5.82. The net charge of the protein is relevant because proteins that are very positive as well as polylysine, which is also positively charged, have been shown to nucleate large particles of amorphous TiO2 from TiBALDH[24]. At pH 6.5, the net charges which were calculated for the proteins in this study in their folded state, are: A113-N is -2448.44, A113-KM2N is -1074.51, ACC5 is 3610.93, and BSA is -114.87. Our study only includes one positively charged protein, and the positive region is confined within the center of the ring rather than exposed to the outside of the oligomer.

A113-N and A113-KM2N were confirmed to have the expected kD via mass spectroscopy, and the oligomeric state was confirmed by size exclusion chromatography to be the expected with D3 symmetry. Protein mutants were then characterized using TEM, confirming that the modifications made to the surface chemistry did not alter the overall structure of the original proteins.

There are multiple phenomena observed in protein directed nucleation: small nanocrystals ranging from 2.5 to 5 nm in size, larger crystals of at least 20 nm, lamellar structures, and core shell structures. The small nanocrystals range in quality from poorly aligned planes of 10 or fewer atoms, which we refer to as nanoparticles, to well-ordered nanocrystals containing at least 25 atoms, which we refer to as small nanocrystals. Larger crystals are all larger than the protein substrate so have been tracked together, even though they range in size from ~20 n to the micron scale. Lamellar structures are long lines of TiO_2, we are describing them as lamellar because the aspect ratio is very long, resembling

stacks of titania. The results of reactions with 0.1 mg mL^{-1} and 0.5 mg mL^{-1} of protein with 10 mM TiBALDH in water were compared at 15 and 60 minutes.

We do not observe any crystals in the absence of protein when we dilute TiBALDH to 10 mM and wait for 1 hour before preparing a TEM grid (Fig. 2.2). This demonstrates that for these buffer conditions any observed phenomenon is the result of a reaction which only occurs in the presence of protein.

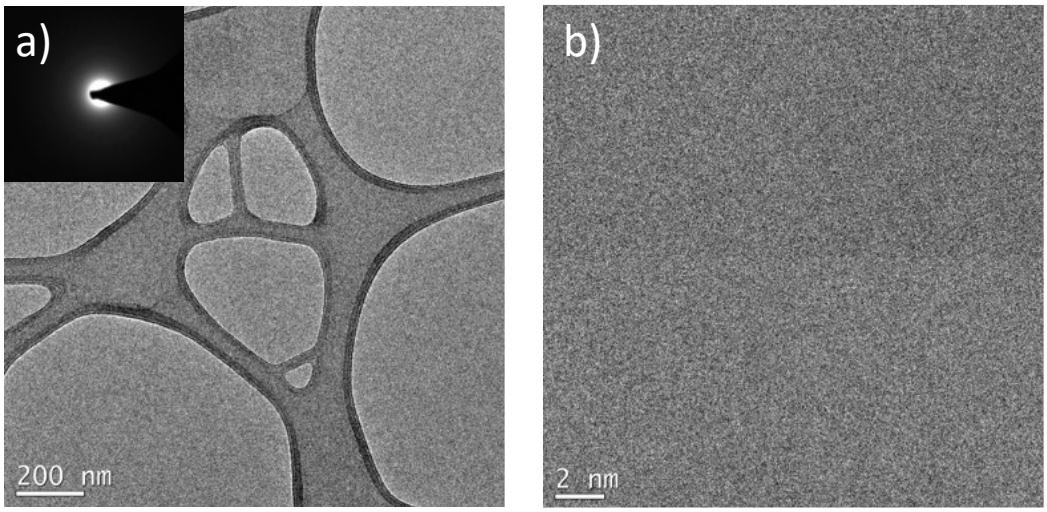

Figure 2.2: Control showing lack of nucleation in absence of protein
a) low magnification showing a lacey grid which had the TiBALDH solution deposited onto it with inset diffraction showing no crystalline signal b) high resolution of lacey grid substrate showing no observable sample

Figure 2.3 shows small particles and nanocrystals demonstrates that any protein in this study can direct nucleation of this phenomena, although the quality of crystal varies.

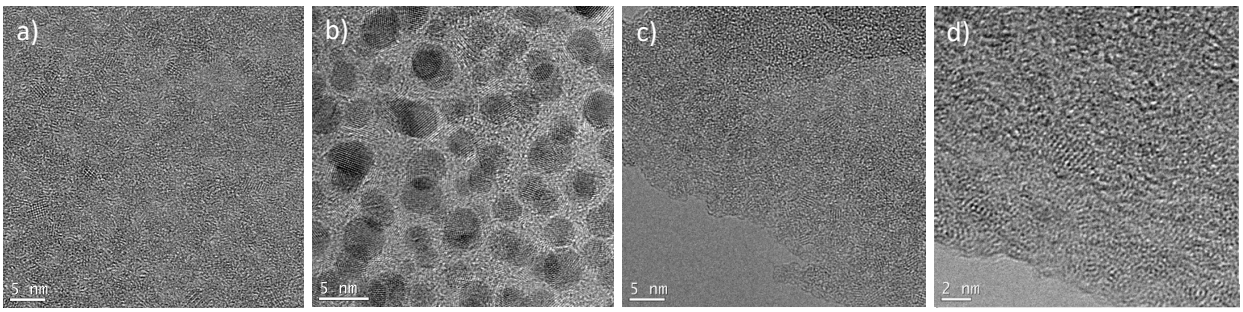

Figure 2.3*:* Small nanoparticles and nanocrystals
Brightfield ex situ TEM images of reactions with a) A113-N showing roughly 2 nm nanocrystals, b) A113-KM2N showing ~5 nm nanocrystals, c) ACC5 showing under ~1.5 nm nanocrystals, d) BSA showing thin ~2 nm long nanoparticles

In water, larger crystals only grew in reactions that involve de novo designed proteins with smooth exposed surfaces. Both A113 proteins direct the formation of large highly ordered crystals (Fig 2.4). Fast Fourier Transform (FFT) of the crystal nucleated in the presence of A113-KM2N is a close match to rutile, but angles between planes are 7 degrees off from the expected value.

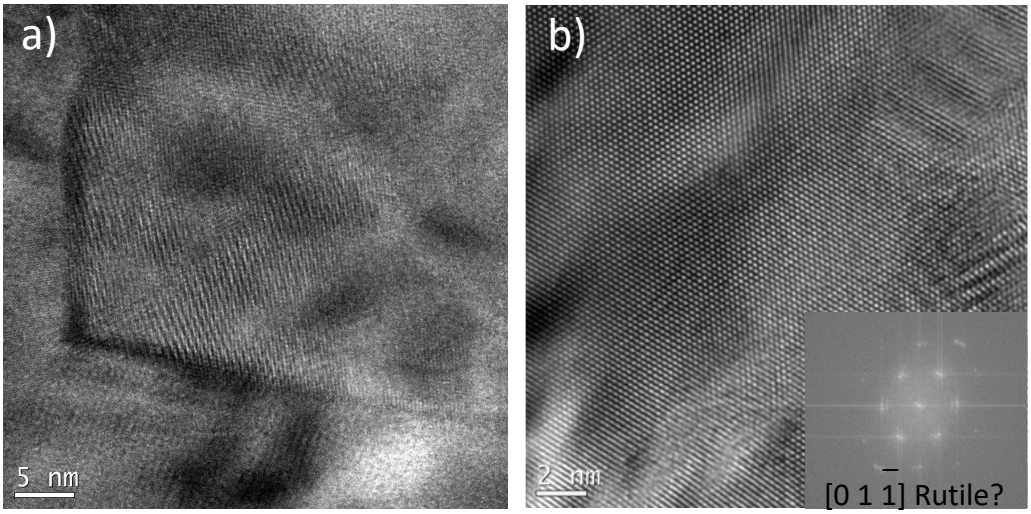

Figure 2.4: Large crystals
Brightfield TEM images of larger crystals which grew as part of a reaction with a) A113-N and b) A113-KM2N with inset FFT showing a potential match for the [01-1] zone axis, however the angles are off by 7 degrees, which indicates either very high strain or that this is not a match for rutile. For the reaction conditions in this research, larger crystals only grew in reactions that involve de novo designed proteins with smooth exposed surfaces.

Only the *de novo* designed proteins A113-KM2N and ACC5 direct the formation of core shell structures (Fig. 2.5). These two proteins are similar in scale, roughly 10 nm in diameter, and both contain alternating regions of positive and negative charge. The morphological difference in shape between the elongated and round structures is a qualitative indication of the influence exerted by the protein in templating the formation of these structures.

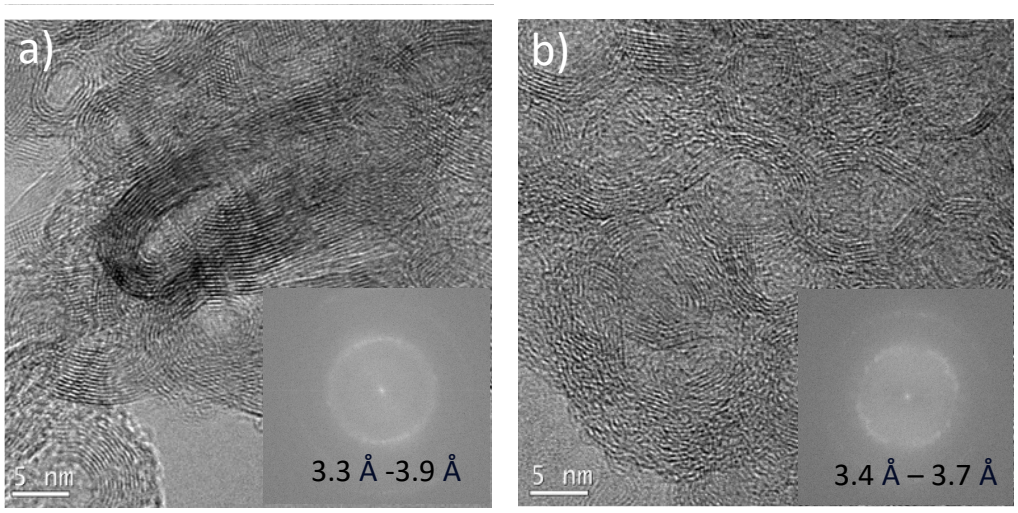

Figure 2.5: Core Shell structures
Brightfield TEM images of core shell structures which formed in a reaction with a) A113-KM2N and b) ACC5, with inset FFT showing the distance between planes of atoms. The morphological difference in shape between the elongated and round structures is a qualitative indication of the influence exerted by the protein in templating the formation of these structures.

Both BSA and A113-KM2N can direct the formation of lamellar structures (Fig. 2.6). These proteins both contain alternating regions of positive and negative charges which are solvent exposed. We hypothesize that the high curvature and protected positive charges of ACC5 contribute to the lack of any observable lamellar structures in this research.

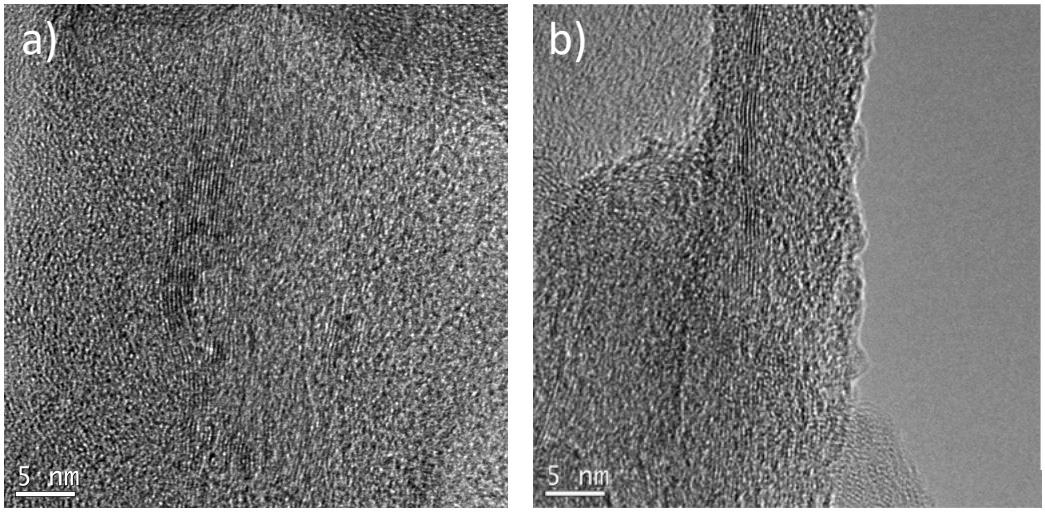

Figure 2.6: Lamellar structures
Brightfield TEM image showing the lamellar structures formed in a reaction with a) A113-KM2N and b) BSA. Lamellar structures have 3.6-3.8 Å spacing between planes of atoms.

We performed cryo-EM on a reaction between 0.5 mg ml-1 A113-KM2N and 10 mM TiBALDH as a control to ensure that the phenomenon observed are consistent between cryo-EM and *ex situ* brightfield TEM images of the same reaction. Figure 2.7 shows confirmation that small particles and core shell structures exist in cryo-EM, which indicates that these observed phenomena are not the result of a drying effect.

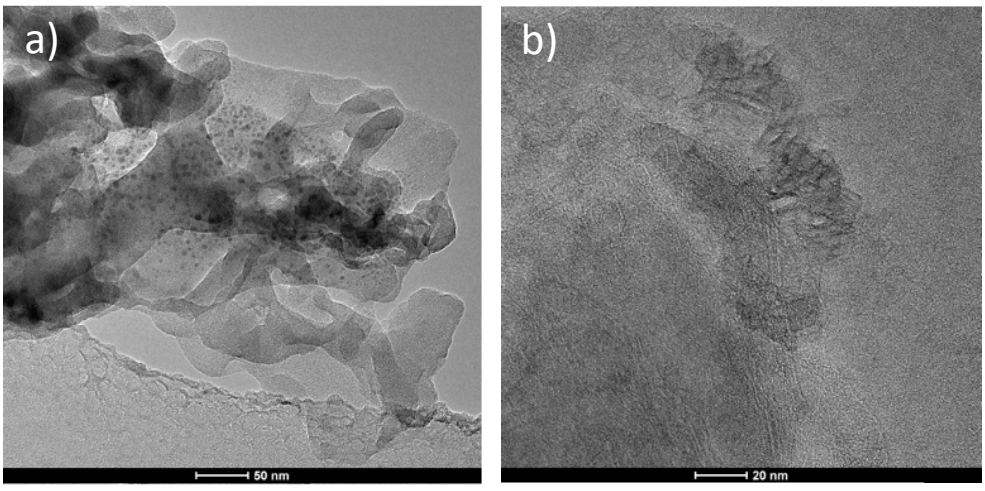

Figure 2.7: Cryo-EM of small nanoparticles and core shell structures
Cryo-EM images taken of the same reaction between A113-KM2N that show a) small nanoparticles and b) core shell structures

Table 1 summarizes the observed phenomenon for each reaction in unbuffered water.

Table 1:

	15 minutes		60 minutes	
A113-N 0.1 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Always present in abundance	Small nanocrystals	Always present in abundance
	Large crystals	Not observed	Large crystals	Not observed
	Lamellar structures	Not observed	Lamellar structures	Not observed
	Core Shell structures	Not observed	Core Shell structures	Not observed
A113-N 0.5 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Always present in abundance	Small nanocrystals	Always present in abundance
	Large crystals	Found on reaction 1, 2	Large crystals	Found on reaction 1, 2
	Lamellar structures	Not observed	Lamellar structures	Found on reaction 2
	Core Shell structures	Not observed	Core Shell structures	Not observed
A113-KM2N 0.1 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Found on reaction 1, 2, 3, and 4	Small nanocrystals	Found on reaction 1, 2, 3
	Large crystals	Found on reaction 1, 2, 3, and 4	Large crystals	Found on reaction 2 and 3
			Lamellar structures	Not observed

	Lamellar structures	Found on reaction 2 and 4		Core Shell structures	Found on reaction 2
	Core Shell structures	Not observed			
A113-KM2N 0.5 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Always present in abundance		Small nanocrystals	Always present in abundance
	Large crystals	Found on reaction 2, 4, 5		Large crystals	Found on reaction 3 and 4
	Lamellar structures	Found on reaction 3		Lamellar structures	Found on reaction 1
	Core Shell structures	Found on reaction 1 and 2		Core Shell structures	Found on reaction 1, 3, and 5
ACC5 0.1 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Not observed		Small nanocrystals	Sparse areas ~2.5 nm particles
	Large crystals	Not observed		Large crystals	Not observed
	Lamellar structures	Not observed		Lamellar structures	Not observed
	Core Shell structures	Not observed		Core Shell structures	Circular particles with non-mineral centers the size of the protein

ACC5 0.5 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Not observed		Small nanocrystals	Sparse areas ~3 nm nanocrystals
	Large crystals	Not observed		Large crystals	Not observed
	Lamellar structures	Not observed		Lamellar structures	Not observed
	Core Shell structures	Not observed		Core Shell structures	Not observed
BSA 0.5 mg mL^{-1} 10 mM TiBALDH	Small nanocrystals	Present near larger aggregates		Small nanocrystals	Present near larger aggregates
	Large crystals	Not observed		Large crystals	Not observed
	Lamellar structures	Always present in abundance		Lamellar structures	Always present in abundance
	Core Shell structures	Not observed		Core Shell structures	Not observed
No protein 10 mM TiBALDH	Small nanocrystals	Not observed		Small nanocrystals	Not observed
	Large crystals	Not observed		Large crystals	Not observed
	Lamellar structures	Not observed		Lamellar structures	Not observed
	Core Shell structures	Not observed		Core Shell structures	Not observed

To further test the role of charge in directing the formation of TiO_2 we used glow discharge on a TEM grid to negatively charge the surface of the grid, deposited 10 mM TiBALDH on the surface, and allowed the reaction to proceed for 1 hour. Figure 2.8 demonstrates that although there are regions of blank grid which do not appear to have TiO2 deposited on the surface, there are many small particles on the surface as well as rare crystals. The particles are approximately 2 nm in diameter and this is consistent with the 3.5 nm hydrodynamic radius that 10 mM TiBALDH has in dynamic light scattering (DLS) measurements. Thus, it is unlikely that the negatively charged surface alone is directing mineral formation.

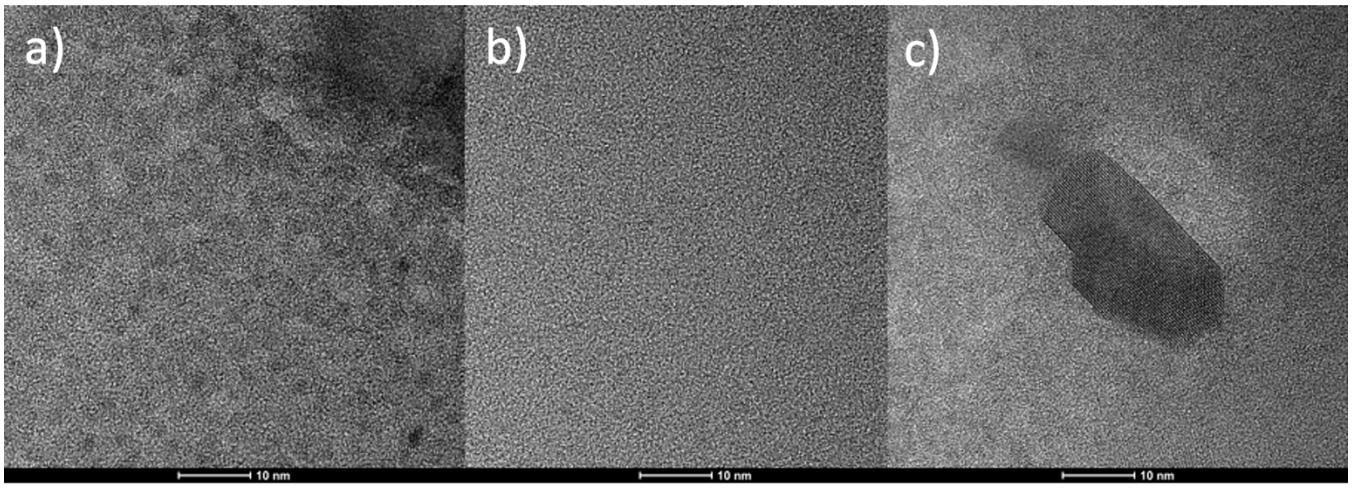

Figure 2.8: *TEM of a reaction between TiBALDH and a glow discharged TEM grid showing*
a) regions with 2 nm particles deposited on the surface b) blank regions of grid with no deposited material c) small crystals are rarely seen

However, when .5 mg/mL of the entirely negative construct is deposited onto a glow discharged grid along with 10 mM TiBALDH the result is drastically different than when the same reaction is deposited onto a surface that has not been glow discharged. Figure 2.9 shows that in the presence of a negatively charged substrate the formed TiO_2 takes on a morphology which resembles the size and shape of the proteins.

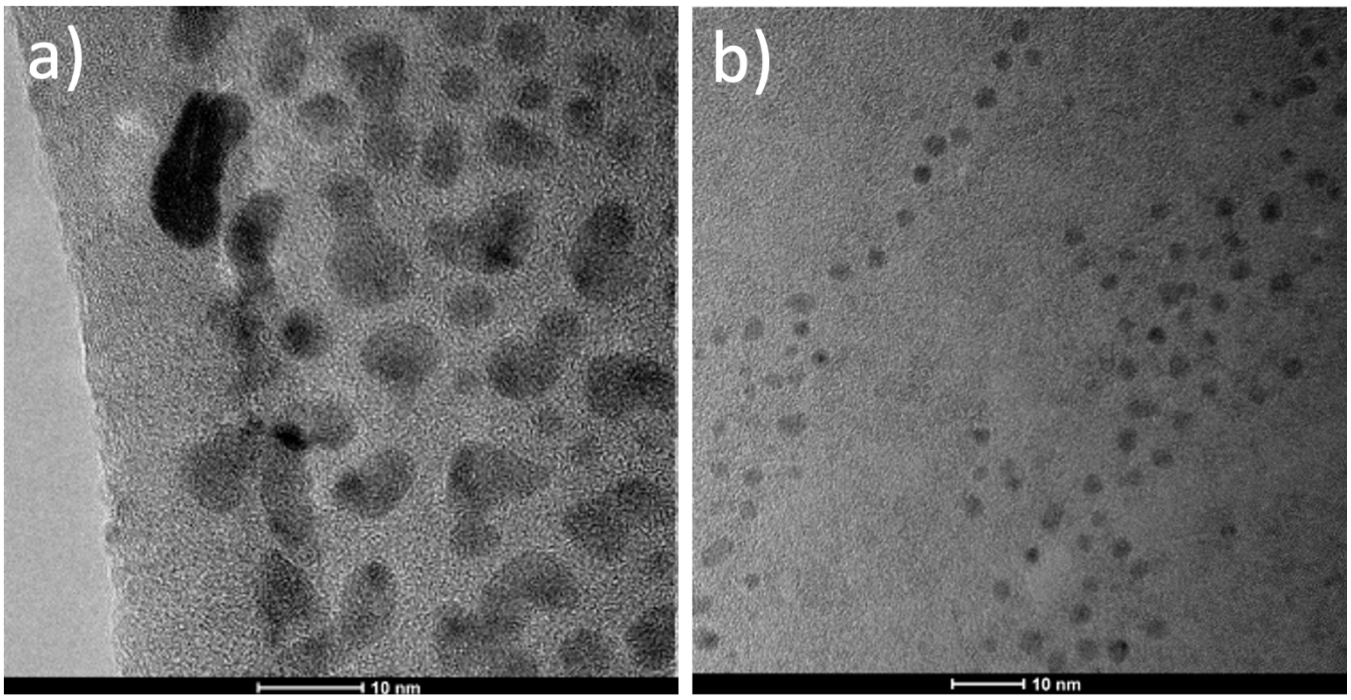

Figure 1.9: Effects of a charged substrate on particle growth
TEM of reactions with 0.5 mg/mL A113-N and 10 mM TiBALDH a) on a grid which had been plasma cleaned to possess a negative charge b) on a grid that has not been plasma cleaned

Understanding the mechanism of nucleation requires understanding the interaction between protein and precursor. To understood whether protein is incorporated into the mineral or if the surface acts as a catalyst where the mineral dissociates from the protein after nucleation, we leveraged HIS tag bioconjugation to immobilize proteins on a nickel substrate by applying a -2V bias to the substrate and incubating the protein solution on the biased substrate, adapting a dip pen lithography procedure[25]. AFM of the blank nickel substrate reveals spherical grains that are 4 nm high and 10 nm in diameter, which is consistent with the grain size of nickel[25]. *Ex situ* AFM of the substrate after the immobilization procedure using A113-KM2N contains features that are consistent in size and shape to A113-KM2N (Fig. 2.10).

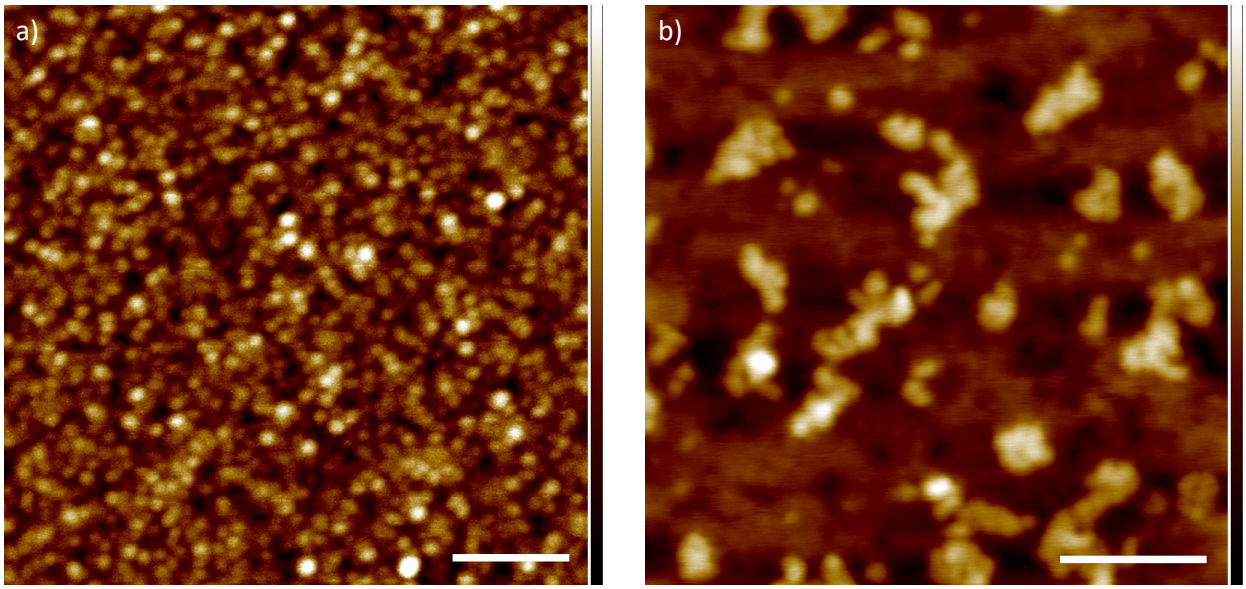

Figure 2.10: Immobilization of A113-KM2N on a nickel substrate
AFM images showing a) black nickel substrate and b) nickel substrate after the protein has been immobilized on the substrate. Scale bars are 100 nm and the z height for both images is 4 nm.

Protein coverage can be varied by adjusting the duration of voltage, with 10s leading to sparse coverage and 30s leading to dense coverage. Photo-induced force microscopy (PiFM), shown in Figure 2.11, of a substrate after 30s incubation demonstrates dense signals at the wavenumber 1657 nm^{-1} which are consistent with the N-H bond from the primary amine in lysine, which is densely populated in this protein construct.

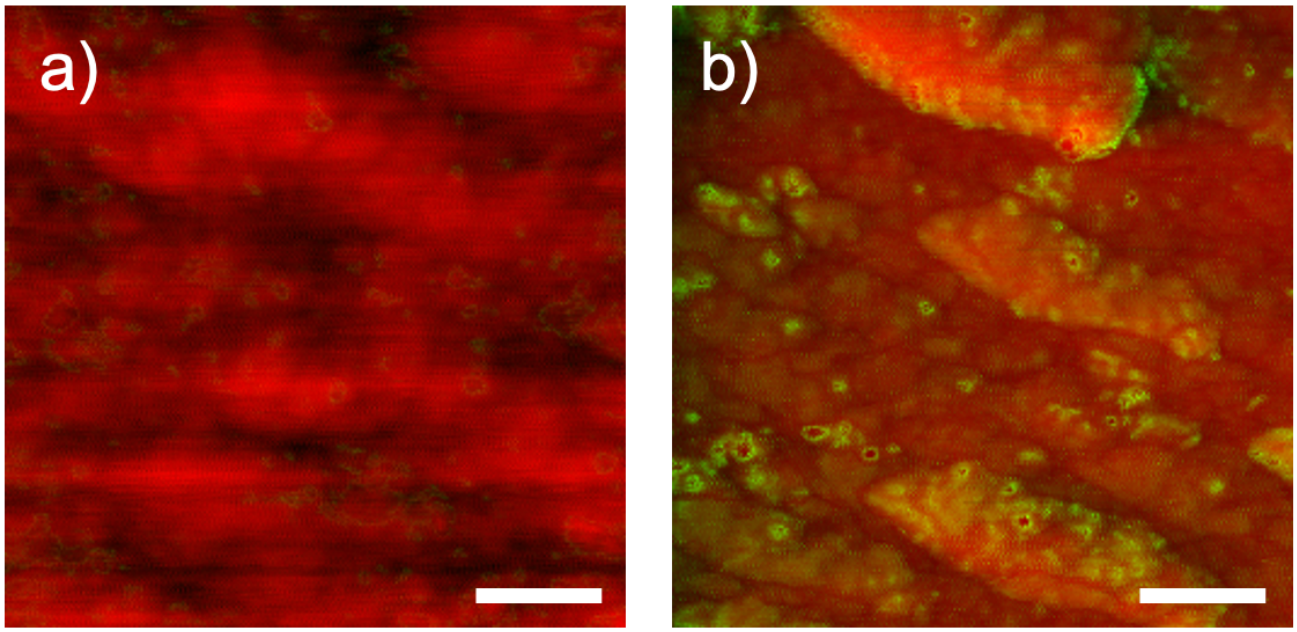

Figure 2.11: Evidence of protein coverage on Ni substrate
PiFM data showing topography in red with absorbance at 1657 nm^{-1} normalized to 2.5 mV of amplitude a) blank Ni substrate and b) substrate after protein immobilization. Scale bars 100 nm.

Flow cell experiments were performed on both blank and protein decorated substrates, where images were collected while 1 mM TiBALDH was flowed through the cell at a rate of 100 µL / min. In the samples with protein present, particle growth was observed for the first 20 mins of flow which agrees with the time point in the TEM data. Figure 2.12 shows the growth of particles on an A113-KM2N decorated surface compared to a control of the nickel substrate without proteins on the surface.

Three particles were tracked over the course of the experiment on the substrate with A113-KM2N. The particles grew at an average rate of 21.7 nm^2 per minute with a standard deviation of 8.2 nm^2 per minute. An area of background in a featureless region of the image was also tracked and no growth was observed. The area of the particles was tracked using the python package cv2 to detect the contours of the features throughout the experiment.

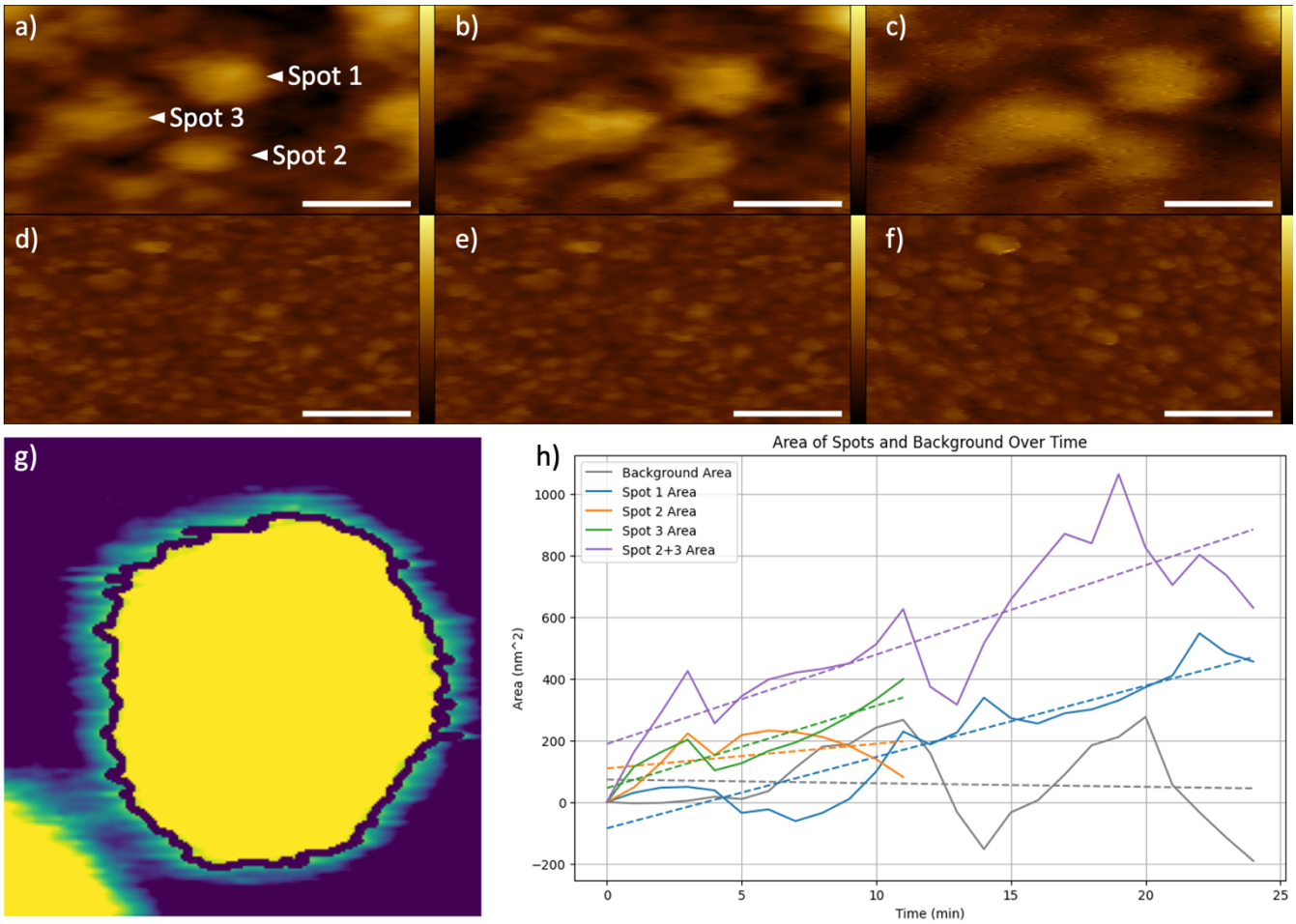

Figure 2.12: Flow Cell experiment of mineral growth
a)-c) A113-KM2N which has been incubated on the Ni surface with 10s of -2 V applied a) initial flow b) after 15 mins of reaction time c) after 30 mins of reaction time. d)-f) the surface of a blank Ni substrate after d) initial flow e) after 15 mins of reaction time f) after 30 mins of reaction time. g) zoomed in region of flow experiment showing the contours used to calculate the area of spot 1. h) plot showing the growth of particles over time- spots 2 and 3 eventually grew together so they have been plotted separately until joining and the combined growth has been plotted separately for the duration of the experiment. Scale bars 50 nm. Height of images from -10 nm to 20 nm.

2.4 Discussion:

The use of *De Novo* designed proteins in directing mineralization is a significant step forward in nucleation, growth, and self-assembly. Rather than using a SAM or unstructured peptide, we have introduced a structured morphology which can either be anchored to a substrate or free in solution (Fig. 2.13).

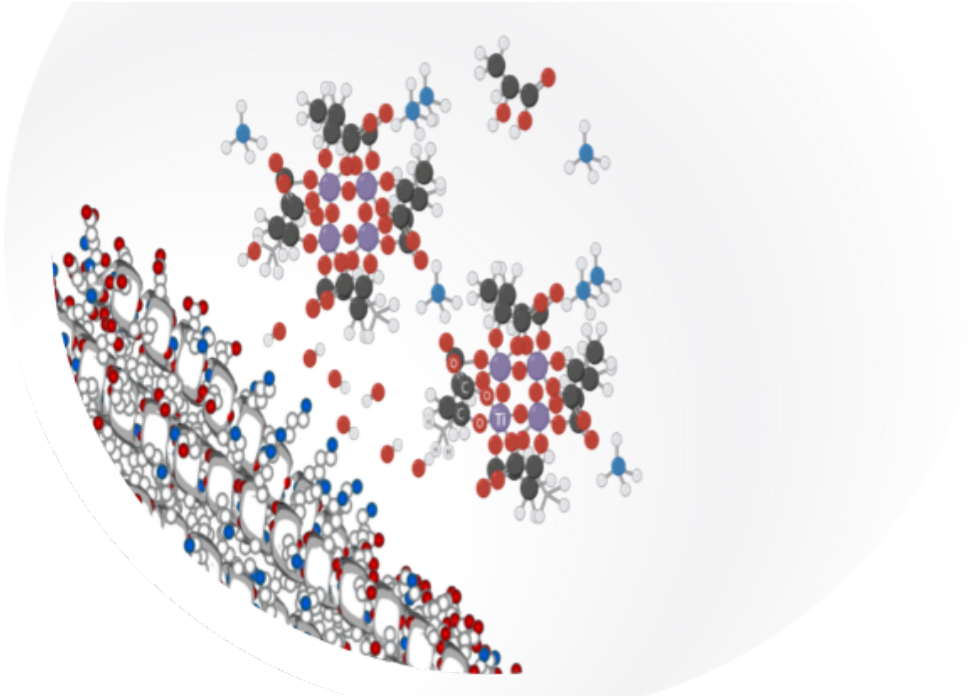

Figure 2.13: Proposed mechanism for protein directed nucleation from TiBALDH to titanium dioxide
Cartoon suggesting a mechanism for hydrolysis of TiBALDH into titanium dioxide, where the positive residues of the protein attract the precursor to the substrate and a local gradient of hydroxides can act as the nucleophile for hydrolysis with the titanium atom acting as the electrophile.

We propose a mechanism where regions of charge are instrumental to forming titania for the hydrolysis of TiBALDH to TiO_2. We expect to find a local gradient of HO^- molecules localized around the positively charged regions on the proteins and hypothesize that the negatively charged carboxylate functional groups can participate in ligand exchange to be on the surface of titania.

In hydrolysis reactions, the chemical potential governs whether the reaction will proceed spontaneously, be at equilibrium, or require external energy input. Biological systems exploit differences in chemical potential to regulate energy flow and drive essential biochemical processes like biomineralization. However, ex situ reactions do not have the ability to contain local regions of differing chemical potential. The local charge patchiness on the surface of the proteins will drive local concentration gradients of free radicals in solution, which should result in a uniform chemical potential by conventional theory. However, mineral formation is only observed in the presence of the proteins.

This presents a conundrum where conventional models are insufficient to explain the phenomenon we observe and leads to a need for a greater theoretical understanding of the underlying mechanism.

2.5 References

1. Dirkzwager, A. *et al.* Green Engineering of Silicon and Titanium Dioxide Architectures and Realizing Downstream Applications. *Adv Sustain Syst* **9**, 2400591 (2025).
2. Poulsen, N., Sumper, M. & Kröger, N. Biosilica formation in diatoms: Characterization of native silaffin-2 and its role in silica morphogenesis. *Proc Natl Acad Sci U S A* **100**, 12075–12080 (2003).
3. Koester, J., Brownlee, C. & Taylor, A. R. Algal Calcification and Silicification. *Encyclopedia of Life Sciences* 1–10 (2016) doi:10.1002/9780470015902.A0000313.PUB2.
4. Kröger, N., Deutzmann, R., Bergsdorf, C. & Sumper, M. Species-specific polyamines from diatoms control silica morphology. *Proceedings of the National Academy of Sciences* **97**, 14133–14138 (2000).
5. Kröger, N., Deutzmann, R. & Sumper, M. Silica-precipitating peptides from diatoms: The chemical structure of silaffin-1A from Cylindrotheca fusiformis. *Journal of Biological Chemistry* **276**, 26066–26070 (2001).
6. Naser, N. Y. *et al.* Biomimetic mineralization of positively charged silica nanoparticles templated by thermoresponsive protein micelles: applications to electrostatic assembly of hierarchical and composite superstructures. *Soft Matter* **21**, 166–178 (2025).
7. Pushpavanam, K., Ma, J., Cai, Y., Naser, N. Y. & Baneyx, F. Solid-Binding Proteins: Bridging Synthesis, Assembly, and Function in Hybrid and Hierarchical Materials Fabrication. *Annu Rev Chem Biomol Eng* **12**, 333–357 (2021).
8. Torkelson, K. *et al.* Rational Design of Novel Biomimetic Sequence-Defined Polymers for Mineralization Applications. *Chemistry of Materials* **36**, 786–794 (2024).
9. Wallace, A. F., DeYoreo, J. J. & Dove, P. M. Kinetics of silica nucleation on carboxyl- and amine-terminated surfaces: Insights for biomineralization. *J Am Chem Soc* **131**, 5244–5250 (2009).
10. Hellner, B., Stegmann, A. E., Pushpavanam, K., Bailey, M. J. & Baneyx, F. Phase Control of Nanocrystalline Inclusions in Bioprecipitated Titania with a Panel of Mutant Silica-Binding Proteins. *Langmuir* **36**, 8503–8510 (2020).
11. Sewell, S. L. & Wright, D. W. Biomimetic Synthesis of Titanium Dioxide Utilizing the R5 Peptide Derived from Cylindrotheca fusiformis. (2006) doi:10.1021/cm060342p.
12. Sumerel, J. L. *et al.* Biocatalytically Templated Synthesis of Titanium Dioxide. *Chemistry of Materials* **15**, 4804–4809 (2003).
13. Curnow, P. *et al.* Enzymatic synthesis of layered titanium phosphates at low temperature and neutral pH by cell-surface display of silicatein-α. *J Am Chem Soc* **127**, 15749–15755 (2005).
14. Kröger, N. *et al.* Bioenabled Synthesis of Rutile (TiO2) at Ambient Temperature and Neutral pH. *Angewandte Chemie International Edition* **45**, 7239–7243 (2006).
15. Seisenbaeva, G. A., Daniel, G., Nedelec, J. M. & Kessler, V. G. Solution equilibrium behind the room-temperature synthesis of nanocrystalline titanium dioxide. *Nanoscale* **5**, 3330–3336 (2013).
16. Doyle, L. *et al.* Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528**, 585–588 (2015).
17. Hsia, Y. *et al.* Design of multi-scale protein complexes by hierarchical building block fusion. *Nature Communications 2021 12:1* **12**, 1–10 (2021).
18. Divine, R. *et al.* Designed proteins assemble antibodies into modular nanocages. *Science* **372**, (2021).
19. Mann, S. & Meldrum, F. C. Controlled synthesis of inorganic materials using supramolecular assemblies. *Advanced Materials* **3**, 316–318 (1991).
20. Douglas, T. *et al.* Protein Engineering of a Viral Cage for Constrained Nanomaterials Synthesis**. doi:10.1002/1521-4095.

21. Ping, H. *et al.* Organized Arrangement of Calcium Carbonate Crystals, Directed by a Rationally Designed Protein. *Cryst Growth Des* **18**, 3576–3583 (2018).
22. Courbet, A. *et al.* Computational design of mechanically coupled axle-rotor protein assemblies. *Science (1979)* **376**, 383–390 (2022).
23. Grimsley, G. R., Scholtz, J. M. & Pace, C. N. A summary of the measured pK values of the ionizable groups in folded proteins. *Protein Sci* **18**, 247 (2009).
24. Kharlampieva, E. *et al.* Bioenabled surface-mediated growth of titania nanoparticles. *Advanced Materials* **20**, 3274–3279 (2008).
25. Agarwal, G., Naik, R. R. & Stone, M. O. Immobilization of histidine-tagged proteins on nickel by electrochemical dip pen nanolithography. *J Am Chem Soc* **125**, 7408–7412 (2003).

3. Machine learning-driven descriptions of protein dynamics at solid-liquid interfaces

This work was adapted from Machine learning driven descriptions of protein dynamics at solid liquid interfaces

Amy Stegmann,[1] Benjamin A. Legg,[2] James J. De Yoreo,[2,3] Shuai Zhang[1,2,3]

1 Molecular Engineering and Science Institute, University of Washington, Seattle WA 98105

2 Physical Sciences Division, Pacific Northwest National Laboratory, Richland WA, 99354

3 Materials Science and Engineering, University of Washington, Seattle WA, 98105

Key Words: protein dynamics, solid-liquid interfaces, atomic force microscopy, Machine Learning

3.1 Introduction

Proteins are more than simple nutrients. They are the building blocks of nature, thanks to their innate ability to assemble into complex hierarchal structures. For billions of years, living creatures have harnessed that ability to create a huge array of functional materials. More recently, scientists have begun to harness the self-assembly of these molecules to develop new biomimetic materials, with applications in health, energy, environment, and beyond, including [1-3] optics, [4] catalysis, [5, 6] medical diagnostics, [7] and medical therapeutics. [8, 9] To enable materials development, scientists are currently working to better understand and control the process of self-assembly. However, the pathways of assembly have historically been hidden from view. The proteins involved are complex, dynamic, and nanoscale in size. It is only recently that high resolution imaging techniques such as liquid-phase transmission electron microscopy [10] and high-speed atomic force microscopy (HS-AFM) have been introduced, which provide the capability to directly watch protein assembly. This chapter focuses on how machine learning (ML) techniques help to harness the full potential of these imaging tools, and thus provide new insights into protein assembly.

Protein assembly is a hierarchal process.[11] Individual protein strands fold into structured monomeric units, which in turn assemble into structures exhibiting order across multiple length scales. In many situations, these processes are coupled, since individual proteins can sometimes change their confirmation to facilitate the processes of assembling and disassembling [12-14]. The assembly process may directly produce functional materials [15, 16] or may produce templates that can support the synthesis of three-dimensional hybrid protein-inorganic materials. [17, 18] Because proteins can be designed with diverse shapes and precise placement of chemical functional groups in 3D space [8, 19], they are ideal building blocks for producing functional designer materials with tunable properties. [20-22] Likewise, hybrid protein-inorganic assemblies offer a rich platform for investigating self-assembled nanoparticles, opening the door to designing new functional materials with tunable material properties.

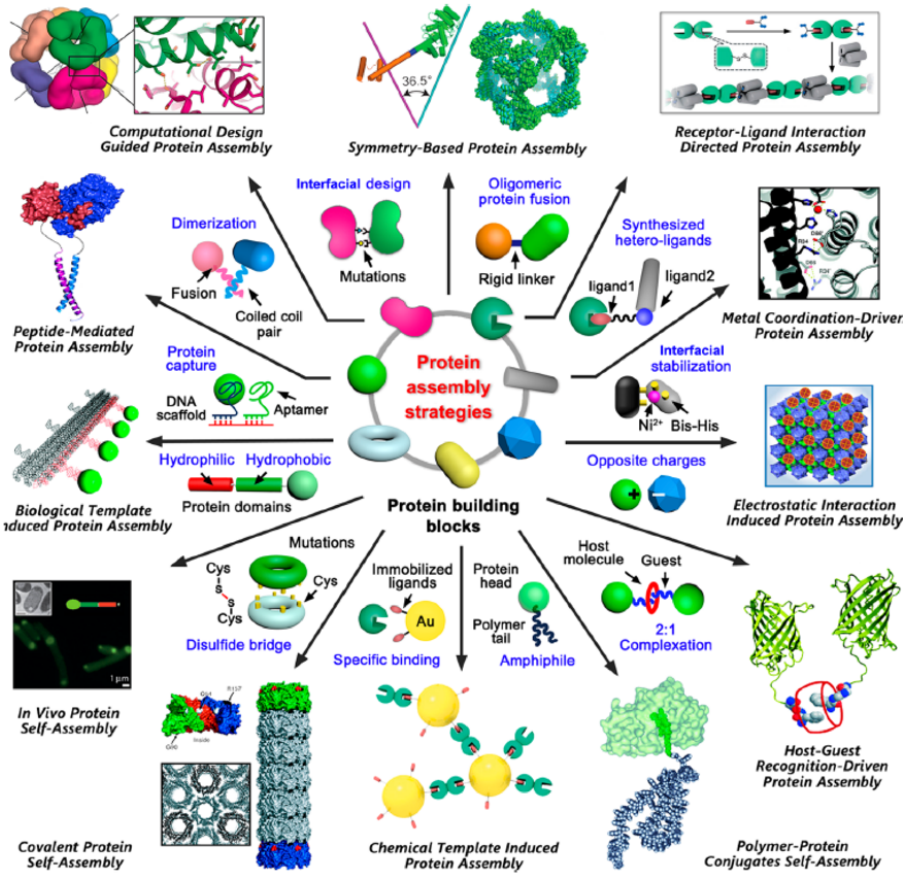

Figure 3.1: Examples of protein assemblies and protein-based hybrid materials.[11]

Frequently, protein assembly occurs preferentially at the interface between a liquid and a supporting substrate. The liquid is essential because it provides a solvent that facilitates protein motion and conformational changes, while the solid surface can template patterns of order and drive locally enhanced concentrations that facilitate assembly. Moreover, the driving forces for assembly can change at interfaces. The interfaces are typically a high-energy region when compared with the bulk of either phase. The solid phase has higher enthalpy than bulk, due to broken bonds at the surface, and the solvent phase often has reduced entropy than bulk due to solvent structuring. [23, 24] These local differences can, in some circumstances, provide a driving force that promotes assembly.[26]

Because AFM is uniquely capable of imaging surface structures with sub-nanometer resolution, it has been widely used to study protein assemblies for decades. [25-29] AFM can clearly reveal the kind of complex hierarchal structures, such as ribbons, fibrils, and lattices that protein assembly produces (Fig. 3.2). These studies have shown that the resulting structures are influenced by numerous parameters, including the type of substrate, solution pH, electrolyte concentration, temperature, light, etc. [30, 31]

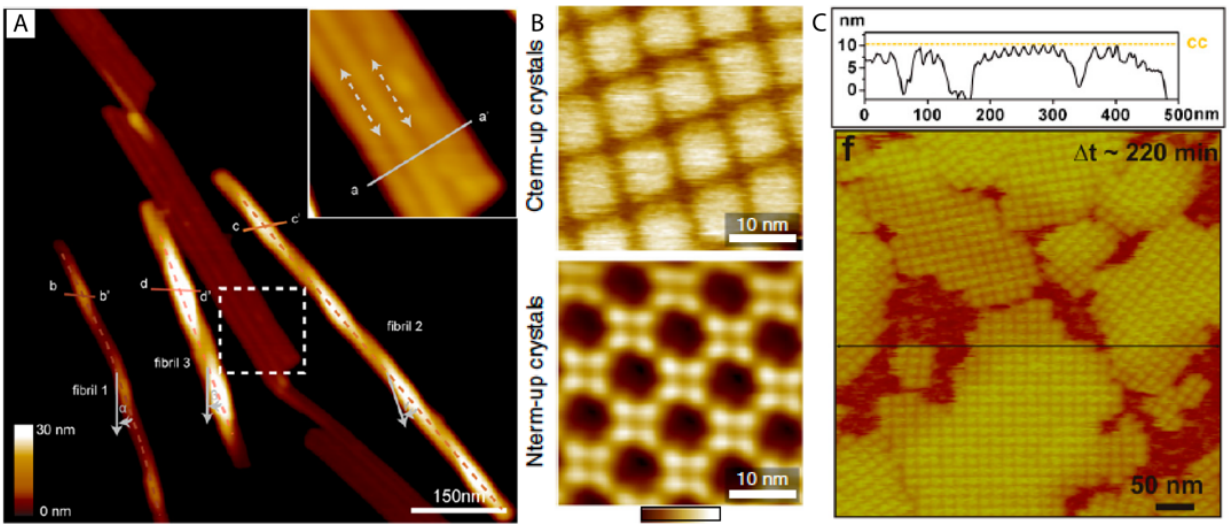

Figure 3.2: Examples of protein assembly resolved by AFM.
(A) Assembled ribbon and fibrils of an amyloid peptide. [32] (B) polymorphs of 2D protein crystals. [32] (C) 2D matrix of an S-layer protein. [27]

More recently, imaging tools have been turned toward directly imaging the *process* of assembly as it happens, to determine assembly pathways *in-situ* [33, 34], and understand the influence of environmental factors. To achieve this understanding, it is typically necessary to visualize dynamics at several second to sub-second temporal resolution, using techniques such as liquid-phase transmission electron microscopy [10] and high-speed AFM (HS-AFM). [26] HS-AFM provides a unique ability to visualize bio-macromolecular assembly with sub-nanometer spatial resolution and temporal resolution of seconds to milliseconds. [25] This has been demonstrated by a variety of recent studies. The Ando group directly observed the myosin V protein walking along an actin filament using HS-AFM, to shed new insights on proton motion. [35] Their group also directly visualized the movement of two glycoside hydrolase family 18 chitinases (ChiA and ChiB) from the chitinolytic bacterium Serratia marcescens on crystalline b-chitin. [36] (Fig. 3.3A) The Scheuring group employed HS-AFM to elucidate the protein channel formations on lipid bilayers and their response to external stimuli. [37, 38] (Fig. 3.3B) And the

De Yoreo and Huang groups used HS-AFM to resolve one-dimensional nucleation of peptides at MoS_2-water interfaces. [39] (Fig. 3.3C)

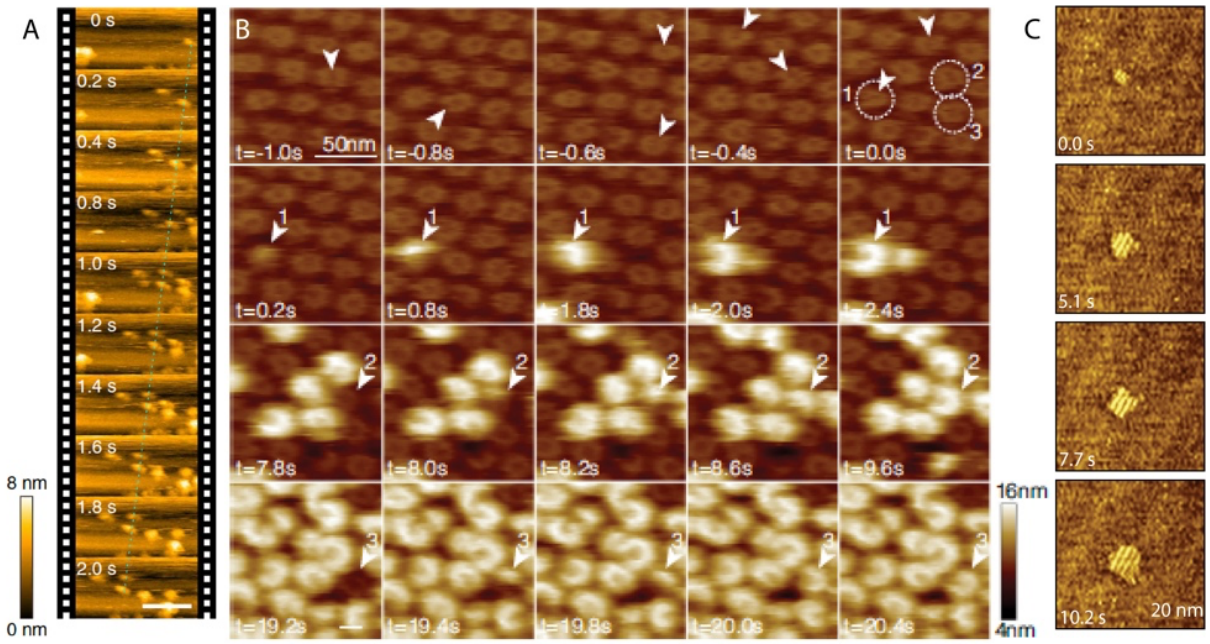

Figure 3.3: Examples of HS-AFM studies on protein assembly.
A) Real-time observation of crystalline chitin incubated with 1 mM ChiA by means of HS-AFM. [36] (B) HS-AFM imaging of protein PFN2 clockwise (CW) pre-pore to pore transition. [38] (C) HS-AFM data records the assembly of a peptide forming a 2D matrix via 1D nucleation on MoS^2. [39]

However, with the development of HS-AFM, simply dealing with the generated data becomes a barrier to describing protein dynamics. A single HS-AFM video might allow researchers to watch hundreds of proteins, as they assemble across thousands of consecutive frames. To understand the dynamics of this process statistically, one needs to quantitatively track and analyze these proteins over time. This can be a burdensome task if one is dependent on manually annotating the location and structure of each protein in each image. Traditional image analysis tools can help with automated data-analysis, but AFM datasets often include intrinsic artifacts (such as streaks, tip-convolution, or general noise) that impede direct analysis.

Fortunately, over the last few years, machine learning (ML) tools for automated feature recognition in experimental data have been developed using deep convolutional neural networks (DCNN). [40] These tools have become incredibly useful for analyzing these large datasets. Rather than manually annotating each image, researchers can now mark a few selected frames and ‘teach’ the neural network how to identify the features of interest. Once trained, the neural network can automatically process the remaining images. Deep convolutional neural networks are a proven tool for automating feature recognition during video analysis [41]. They are robust to noise, and once properly trained they can pick out features in a noisy image. [40, 42] When combined with traditional image analysis tools, ML can help to improve the recognition performance to a high enough standard for quantitative analysis. For example the Ginger group used ML to analyze microscopy data using a DCNN to identify protein fibers on a substrate and then refine the analysis by comparing the image data with spectroscopic data to get high-throughput ensemble analysis of the structure of gold nanorods adhered on the fiber surfaces [43]. ML was also used to determine the chirality of individual molecules in AFM data, as reported by the Wang group. [44]

This chapter will focus on the use of machine learning to interpret AFM images, via DCNN. A convolutional neural network describes a machine learning process or architecture that learns directly from data by extracting features and patterns. Segmentation is a term for classifying the pixels in an image. Semantic segmentation refers to classifying pixels into broad categories, such as 'protein' and 'background'. Instance segmentation is the further classification of the features into discrete objects. Training any segmentation machine learning model involves dividing data into two sets: a training set that is manually labeled to 'teach' a model, and the data set that the model is then used on to make predictions. The training set is further separated into two groups – ‘train’ and ‘test’. The training set is

used to demonstrate the desired outcome to the model being trained, and the test set is used to calculate how similar the predictions of the model are to the human labels.

To demonstrate how ML methods can be used to support HS-AFM studies of protein assembly, this chapter discusses several case studies of assembly for one protein: DHR10-micaN (micaN). This protein was specifically designed to assemble on the surface of the sheet-silicate mineral, muscovite (m) mica. Inspired by ice-binding proteins, the micaN protein contains an array of negatively charged glutamate amino acids, which are spaced appropriately to interact with the potassium sublattice on the m-mica (001) plane, which has hexagonal symmetry. [45] (Figure 3.4 A-C). The 'N' in micaN denotes the number of repeats of the same amino acid sequence. Thus, a higher number indicates a longer protein rod. This protein has been found to show a diversity of fascinating assembly behaviors. For example, small changes in electrolyte concentration have dramatic effects. In a Tris buffer solution with 10 mM KCl, mica18 attaches to the m-mica surface sparsely, with minimal protein-protein interactions. (Fig. 3.4D) But if the KCl concentration is increased to 100 mM, the mica18 assembles into numerous closely-packed clusters aligned in the three directions of the cation sublattice. (Fig. 3.4E) Furthermore, at 3 molar potassium chloride (3M KCl), mica18 assembles into 2D liquid crystals that extend across the m-mica surface for millimeters. (Fig. 3.4F) These dramatic changes in response to something as subtle as electrolyte concentration demonstrate both the challenges, and exciting opportunities, for controlling protein assembly.

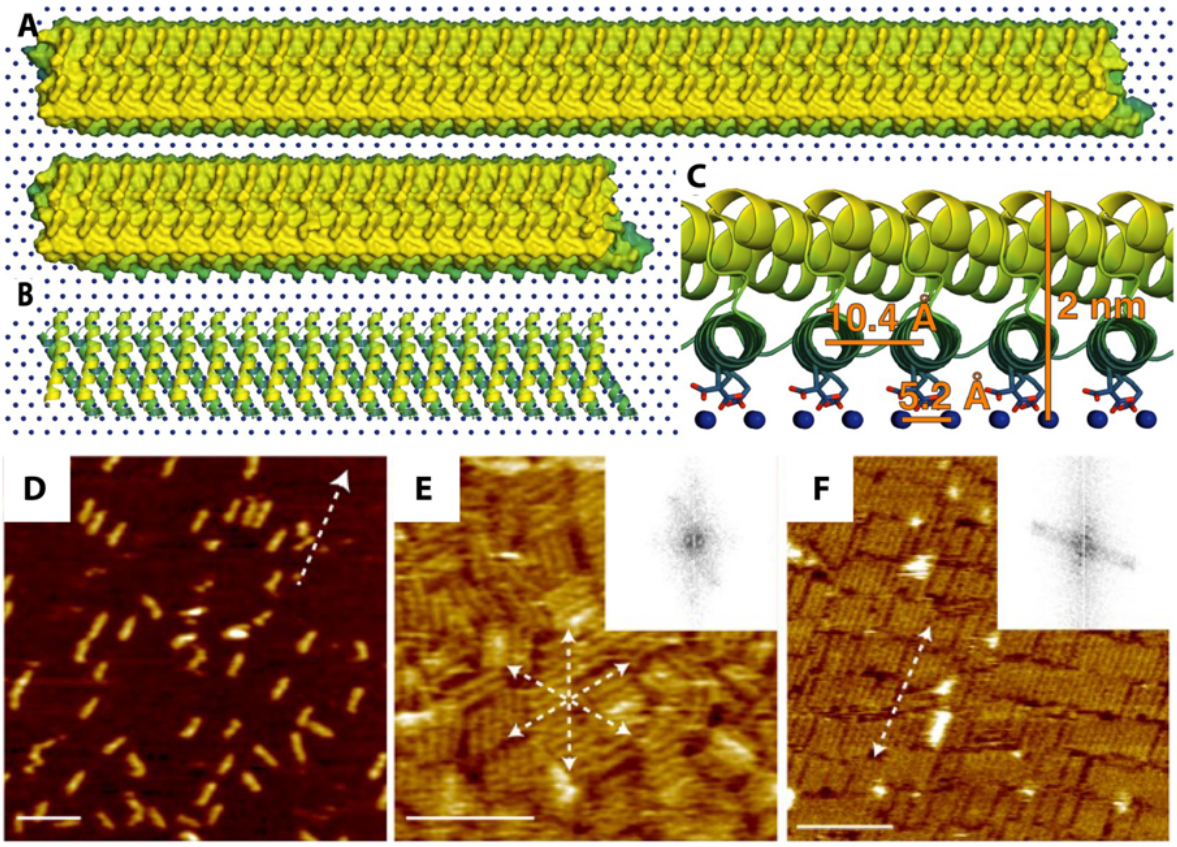

Figure 3.4: MicaN protein system on the surface of mica.
(A-C) Computer model of micaN: (A) The surfaces of mica34 (top) and mica18 (bottom). (B) Cartoon of the backbone showing the repeating helical segments of the protein, (C) the designed interface between the glutamate side chains on the protein surface and the potassium ions on the mica surface. (D) Mica18 on the surface of m-mica in the presence of 10 mM KCl. (E) Mica18 on the surface of m-mica in the presence of 100 mM KCl- the inset gives the fast Fourier Transform (FFT) showing three dominant orientations. (F) Mica18 on the surface of m-mica in the presence of 3M KCl exhibiting one dominant direction as shown in the FFT (inset). Scale bars are 50 nm. [45]

Each case study in this chapter describes an example where ML facilitated data analysis that would otherwise have been infeasible. In the first case, ML improved the automated recognition of micaN proteins in HS-AFM data, so that their motions across a mica surface could be tracked, with enough fidelity to identify anomalous rotational diffusion pathways. The second case details the implementation of a powerful open-source computational package involving ensembles of deep convolutional neural networks , which enabled ML image recognition of proteins in images that contained too much noise to employ the methods of the first example. The third case details the use of the tools developed in the second case to analyze incredibly noisy HS-AFM data and track the process of assembly of a 2D liquid

crystal, starting from a blank substrate. These cases exemplify ways that ML can help to answer increasingly difficult questions about the fundamental dynamics of bio-macromolecular assembly at solid-liquid interfaces.

3.2 Case 1: Rotational dynamics and transition mechanisms of mica18 protein on mica

This section discusses a recent study, in which HS-AFM was used to resolve the rotational dynamics of micaN nanorods on m-mica. The key objectives were (1) to understand how interactions between micaN proteins and the m-mica substrate produce an orientation-dependent energy landscape that biases the micaN protein toward specific orientations, and (2) to understand how proteins traverse that orientational energy landscape to explore different orientations. To achieve these objectives, several HS-AFM datasets were obtained, each one tracking hundreds of proteins across hundreds of frames, with approximately 1 second time resolution. [32] Although manual analysis can be used to track limited numbers of the micaN nanorods over selected HS-AFM frames, the large workload means that a major percentage of the data remains unused. By pairing the HS-AFM with ML analysis, it became possible to track each protein across each frame, to generate enough observations for statistical analysis. A key finding of this study, was that the rotations largely followed Brownian diffusion, but the proteins frequently made large jumps in orientation, to rapidly overcome barriers that usually inhibit rotation. The dependance of rotational dynamics on protein size and solution chemistry were also explored, providing tools for controlling biomolecular assembly at inorganic interfaces.

Thanks to the high aspect ratio of the micaN proteins, the first goal is relatively easy to achieve using traditional image segmentation approaches, as it simply required that the angle of each protein be measured and recorded in each image. After making tens of thousands of measurements, the equilibrium

angular distribution could be obtained and inverted using Boltzmann statistics to determine the orientational energy landscape.

However, the second goal is more challenging, because it requires tracking of the proteins from frame to frame to measure the angular 'jump-sizes' between frames. By quantifying the distribution of jump sizes, great insights into the rotational dynamics can be obtained, but the tolerance for error is now lower. A single error in the angle-determination will negatively impact *two* jump-size estimates, creating a signal that is nearly impossible to distinguish from random motion. Beyond that, the authors also sought to generate continuous trajectories, so that an individual protein could be tracked throughout the course of the experiment. This constitutes a special challenge, since failing to identify just one protein in one frame can cause an entire trajectory to be broken and split into two. Thus, the analysis of rotational dynamics via trajectories requires a very accurate classification.

Conventional recognition methods, such as pattern-matching to reference images via correlation algorithms, can provide the center of mass and angles of particles with high enough accuracy to get average angles of the protein nanorods and pick up trends in density on the surface, but not high enough accuracy to track individual protein nanorods from frame to frame. Thus, ML methods were employed to improve recognition of rods such that tracking frame to frame was possible.

One common issue with deep learning segmentation tasks is the challenge of getting a good separation between objects in cluttered microscopy images. The high level of noise makes learning boundaries difficult. Before training the ML model, preprocessing is required to reduce the multiple types of noise and artifacts that are inherent to AFM. These include background slopes due to tilted-substrates, streaks due to defective lines occurring in the AFM raster pattern, and parachuting and/or ringing, which arise from inefficiencies in the proportional-integral-derivative (PID) systems that control AFM image acquisition.[46] These can be dealt with using a variety of standard algorithms, such as a

line-flattening algorithm used to reduce background slope and reduce the impact of bad lines. In this study, two preprocessing techniques were used to reduce the amount of noise and improve the contrast of the input images (Fig. 3.5A-B). First, a bilateral filter was applied [47]. The bilateral filter is an edge-preserving and noise-reducing filter that averages pixels based on their spatial closeness and radiometric similarity. This allows one to remove high-frequency noise while still preserving the size, shape, and outline of large structures. Finally, a contrast-limited adaptive histogram equalization (CLAHE) [48] was applied. This method used histograms computed over different tile regions of the image so that local details were enhanced even in regions that were darker or lighter than most of the image. The CLAHE method helps to additionally flatten the background and ensure a consistent contrast-difference between the proteins and mica across the sample. After this preprocessing, the images were ready for the ML-based segmentation pipeline.

The next step in the pipeline involved segmenting the regions of the image containing nanorods into discrete objects using U-NET, which is a deep neural network architecture commonly used for microscopy and medical image segmentation. [48] U-NET has a set of convolution layers that compress and then expand, which enables accurate high-resolution predictions at an efficient cost to the CPU. The output of U-NET is the classification of each pixel in the image into either belonging to a protein or belonging to the background. U-NET Centroid Detection function (UNet-DET)was used to train a model to learn centroids of rods (Fig. 3.5C). The results from the centroid detection were then input into the U-NET segmentation routine (UNet-SEG)[49]. UNet-SEG in turn provided instance segmentation with bounding boxes from which the orientational distributions of micaN across the time span of the experiments were extracted. The total workflow is illustrated in Figure 3.5.

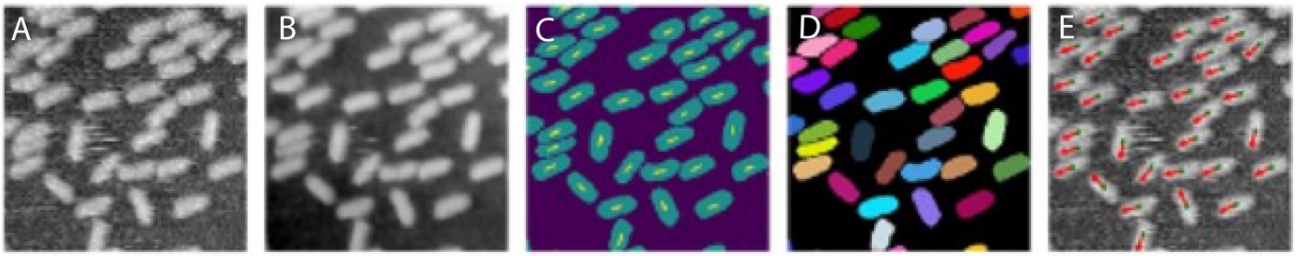

Figure 3.5: Selected steps in the workflow used to identify mica-N protein positions and orientations. (A) Line-flattened AFM image, with evidence for artifacts such as streaks still visible. (B) Image after bilateral filtering, showing a distinct reduction in noise. (C) A binary image, where particles (in cyan) have been identified and separated from the substrate (in purple) using UNet-DET. The particle centers are noted in yellow. (D) The centroids of the particles were then identified using a second U-NET algorithm, UNet-SEG which categorizes distinct particles using a watershed algorithm. (E) final product showing particle center-of-mass (green points) and orientations (red arrows) superimposed upon the original AFM image.

For each image-series, all protein-observations were combined to generate orientational histograms, which were used to calculate the orientation-dependent energy landscape according to the Boltzmann distribution, $G(\theta) = -kBT \ln(N(\theta)/N(\theta^*))$, where $N(\theta^*)$ is the number of proteins in the most favorable orientation. An advantage of using HS-AFM is that the distribution could be tracked over time, to confirm that it was indeed a stable, equilibrium distribution (Fig. 3.6A). In this case, several preferred orientations were found, which are observed as minima in the free energy landscape.

Rod tracking was performed by constructing a graph of association between the protein rods in sequential frames [50]. In the first frame, every identified protein marks the starting node of a new trajectory. Then, frame-by-frame, each protein in frame *i* was matched to a subsequent protein by identifying the protein in frame $i+1$ with the highest number of overlapping pixels. If no protein could be found in the subsequent frame that satisfied a minimum number of overlapping pixels, the trajectory would terminate. If a protein is found without any parent in the previous frame, then it is marked as the starting node for a new trajectory. In this way, each protein could be assigned and labeled a unique particle trajectory (Fig. 3.6B), and for each particle a trajectory for orientation vs time could be constructed (Fig. 3.6C)

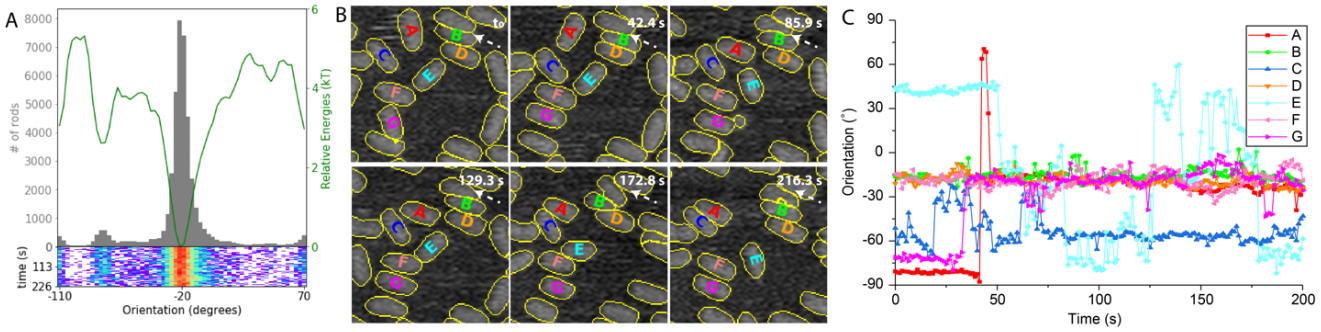

Fig 3.6: (A) Resulting angular populations of the micaN nanorods and the corresponding energy landscape of mica18 with 10 mM KCl on m-mica (001). (B) Selected frames tracking mica18 with 10 mM KCl on m-mica (001). The arrows in the upper right corners indicate the dominated orientations of micaN in each dataset. The times in the upper right corners indicate the relative capture time of each frame in the corresponding dataset. The frame speeds are 0.92 Hz. (C) Direction tracking of the selected Mica18 molecules in panel B.

The combination of equilibrium energy landscapes and time-resolved trajectories can provide a powerful tool for understanding particle dynamics, and allow new insights into the microscopic mechanisms of particle motion.[51] Simple models of Brownian motion can predict how particles with constant mobility would traverse a given energy landscape and overcome energy barriers to sample different orientations. Here, it was found that most of the nanorod motion followed Brownian expectations, but the rods occasionally made very large jumps in angle, with a frequency that would not be expected by simple Brownian models. The distribution was more consistent with a Levy-flight [52, 53], and could be explained if the particle occasionally enters into a spectrum of activated states with enhanced mobility. This appears to be an example of the "anomalous diffusion" that has recently been observed in several transmission electron microscopy and AFM studies of nanoparticle motion [53]. But here, the motion is confounded by the influence of the orientational energy landscape, and the anomalous diffusion would not have been detectable without the detailed trajectory analysis enabled by HS-AFM and ML.

3.3 Case 2: AtomAI framework for deep learning analysis of micaN in-plane dynamics

In the previous example, the micaN was observed under solution conditions that produce sparse coverages and relatively slow dynamics, that made it comparatively easy to obtain high-resolution AFM images that can be segmented with relative ease. However, subtle changes in solution chemistry can significantly increase the rates of micaN particle motion or produce much high surface coverages that are much harder to segment. This makes micaN an ideal test case for developing machine learning packages to analyze complicated or noisy microscopy datasets.

A recurring challenge for high-speed microscopy, is finding ways to segment complex structures in the presence of noise. In the previous section, we mentioned how tools like bilateral filters or line flattening can reduce noise to make simple segmentation algorithms more tractable, but some systems cannot be handed in this way. A complementary strategy for overcoming noise is to use ML semantic segmentation models for pixel classification. The U-NET architecture begins to fail at producing accurate results once the micrographs being analyzed are sufficiently dense. To overcome this issue, the U-NET architecture was adapted to have less information condensed through the successive layers of convolution. The new architecture, dilnet, was developed by the Kalinin group and also employs multiple networks trained in parallel to further increase the accuracy of predictions. [55]

AtomAI is an open-source Python-based software package that is optimized for this problem. It implements multiple toolkits for controlling instruments, while also providing microscopy-specific deep learning and simulation tools. [40] AtomAI has primarily been used for scanning transmission electron microscopy (STEM), but it can also be applied to other imaging techniques, including scanning tunneling microscopy [54] and AFM [55, 56]. AtomAI's ML tools include trainers that can be used to train a model with weights that can then be used to recognize features on experimental data by passing

the model to a predictor. The correct term for this feature recognition is a prediction – the model predicts the location of features. AtomAI incorporates recent advances in deep learning, such as stochastic weight averaging, on-the-fly data augmentation, and can train an ensemble of neural networks to collectively predict the data so that uncertainty values in the predictions can be generated. [40]

As discussed in Case 1, ML-based tools can be superior to simple segmentation tools (such as high thresholds), especially for apparently noisy systems. Often, features that appear to be noise are correlated with the real structure. A well-trained ML tool can thus utilize aspects of the local noise-structure to aid in the segmentation. The dilnet architecture of AtomAI improves upon previous DCNNs to enable prediction in very noisy and crowded micrographs. The semantic segmentation models trained by AtomAI are used to classify each pixel in an input image as belonging to a specific object or background. This can be used to find mica18 nanorods in AFM data.

The first step in implementing AtomAI for semantic segmentation is to select and prepare training data. (Fig. 3.7 A-C) Here, a subset of mica18 AFM images were chosen, and a corresponding set of marked images are manually prepared to denote the location of features. The marked images contain only 0 and 1 values, where 0 is a pixel that does not contain the feature of interest, and 1 is a pixel that does (protein rods in this case). Image editing programs with layers are most convenient for this stage. Note that how the images are marked will profoundly impact the outcome of the machine learning predictions. Marking a feature only when entirely sure a rod exists will result in the DCNN under-predicting locations, where aggressive marking can lead to noise incorrectly being interpreted as a feature. The training process of AtomAI only classifies the features of interest, the rest is background. Therefore training images must be marked in entirety with no ambiguous regions. The order of images and marked data must be the same so that the first element in each array matches each other.

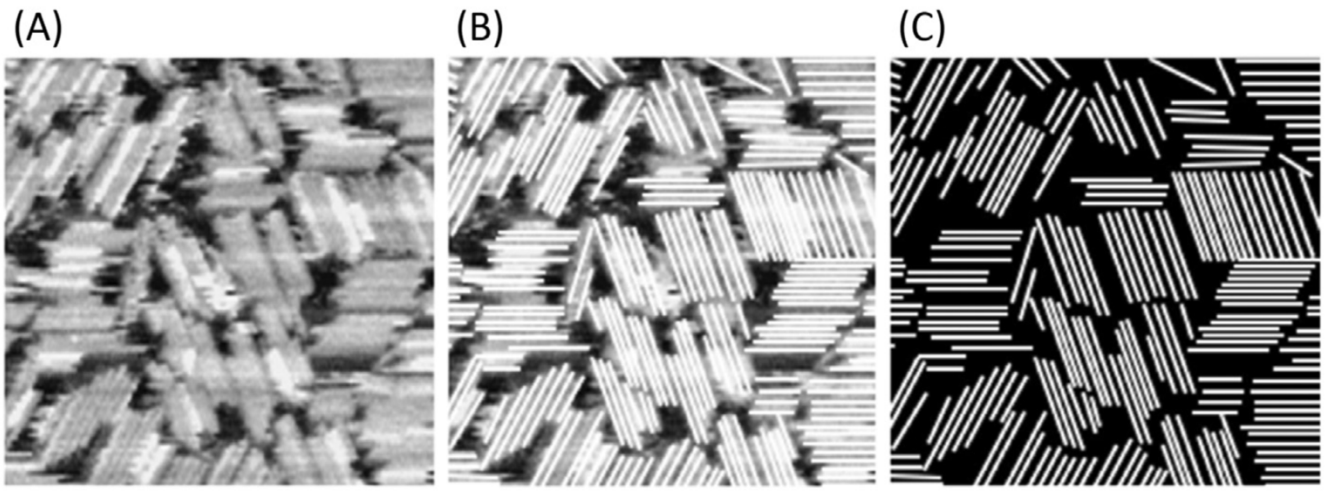

Figure 3.7: Example of the data marking process.
(A) AFM data of micaN rods. (B) An overlayed image, produced may manually drawing the location of rods, (C) the marked-image used at input to AtomAI.

AtomAI has utility functions for training data preparation, including data augmentation that can be performed before and/or during model training with both built-in functions (including blurring, Gaussian and Poisson noises, rotation, zooming, and resizing) and extracting smaller images from a larger marked image. This functionality is implemented with a feature that extracts patches from test and training data:

atomai.utils.extract_patches(*images*, *masks*, *patch_size*, *num_patches*, ***kwargs*)

Takes batches of images and batches of corresponding masks as an input and for each image-mask pair it extracts stack of subimages (patches) of the selected size.

Return type:

Tuple [**ndarray**]

The reason to extract patches is to inflate the training set. Only a small subset of images are marked. The more marked images, the better the DCNN will perform, but marking data is a massive bottleneck in the workflow. Therefore, it is desirable to mitigate the time spent manually marking images. Extracting patches artificially adds to the test train dataset by using many small slices from the original images and masks (Fig. 3.8). Further augmentation could be performed on the test/train dataset by using atomai.transforms.datatransform(). Noise can be added, images can be rotated or flipped, and zoom can be performed if desired.

To test that the patches of image and mask align properly, the n^{th} elements in the extracted patches image and marked arrays are plotted, as shown in Figure 3.8:

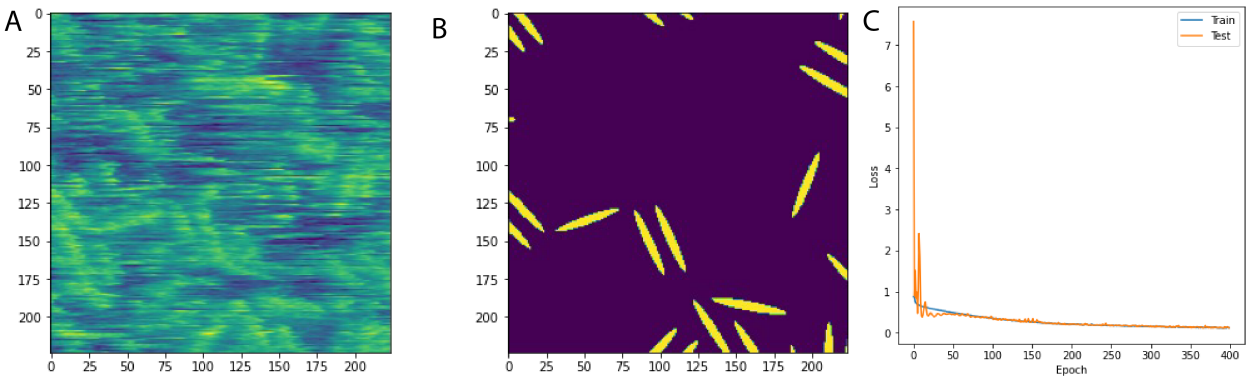

Figure 3.8: (A) Extracted patch from the image training set and (B) matching extracted patch from the marked training set. (C) This loss plot can be used to look for convergence between the test and the train. The more training cycles, the less slope the test and train prediction plots will have at the endpoints. The model shown in this figure has not yet converged and should be trained for more cycles.

After the test and training datasets are compiled, the next step is to train the DCNN. AtomAI can train ensembles of models, which can provide more accurate and reliable predictions compared to a single model. [54, 55] An ensemble is a user defined number of networks that are trained in parallel, but will be trained on slightly different groupings of training data and will end up with similar but slightly different predictions. A model or ensemble can be trained from scratch or from a baseline provided either as a previously trained ensemble of models or from a baseline created by first training a single

model for N epochs, or by sampling along a Gaussian subspace of a single training trajectory. The code below

```
etrainer = aoi.trainers.EnsembleTrainer("dilnet", nb_classes=1, with_dilation=False,
                                batch_norm=True, nb_filters=32, layers=[1, 2, 2, 1],use_dropouts=False,
                                loss='ce', upsampling="nearest")
etrainer.compile_ensemble_trainer(training_cycles=400,training_cycles_ensemble=100,
                                compute_accuracy=True, swa=True, memory_alloc=0.5)
smodel, ensemble = etrainer.train_ensemble_from_scratch(image_train, marked_train,
                                n_models=30)
```

The swa argument applies statistical weight analysis. To use swa and compute accuracy, a minimum of 100 training cycles is required. To obtain more accurate models, increase the training cycles. n_models is the number of models of DCNN within the ensemble. To decrease processing time, this number can be reduced. If swa is set to true, the trainer will plot the loss for every increment of the training_cycles_ensemble (Fig. 3.8C).

Now plot the predictions of each ensemble of models that was trained, shown in Figure 3.9. The code below demonstrates a plotting method:

```
for m in ensemble.values():
    smodel.load_state_dict(m)
    predictor = aoi.predictors.SegPredictor(smodel)
    out, _ = predictor.run(expdata[your_frame] * 255, norm=False)
```

```
plt.imshow(out.squeeze())

plt.show()
```

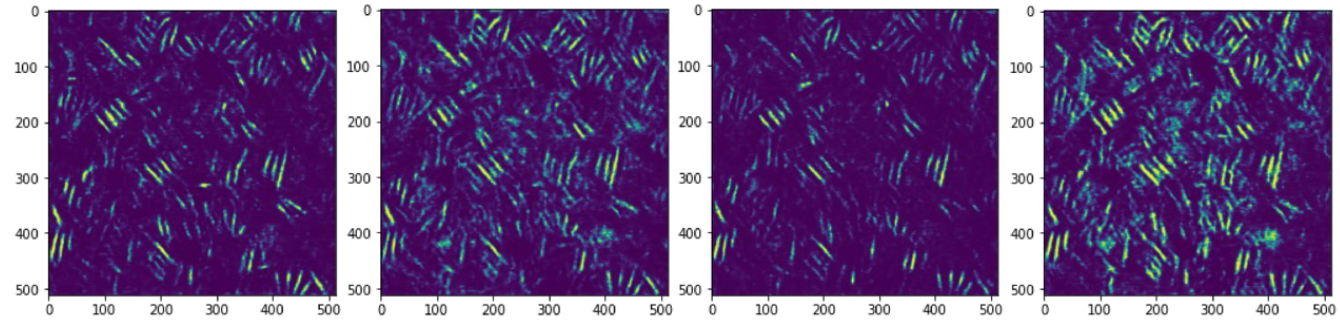

Fig. 3.9: A subset of the ensembles of DCNN- each group has slightly different weights and different predictions. The confidence is plotted in intensity – the brighter the pixel, the more of the models within the ensemble agree that the pixel contains a rod.

The micaN system was used as a test case in developing the tools to analyze noisy AFM data with classifiers. The challenge addressed here was to develop a successful approach to analyzing the data. The workflow enabled by AtomAI can now be used to answer physical questions about protein assembly using HS-AFM in high-noise environments.

3.4 Case 3: ML methods to analyze the assembly of mica18 liquid crystals

Using a combinatorial approach of Case 1 and Case 2, very noisy data was analyzed to understand the dynamics of mica18 liquid crystal formation. An ensemble of neural networks was necessary to gain predictions on the data that was accurate enough to track trends. Although the accuracy was still not high enough to track rods from frame to frame, using the approach described in this case it was possible to track the emergence of order by quantifying an order parameter over the course of a video.

HS-AFM was used to characterize micaN assembly from the bulk solution to two-dimensional liquid crystal phases on m-mica. The concentration of KCl has previously been shown to control the mobility of the mica18 proteins on the surface of mica. [45] Case 1 was performed using 10 millimolar

potassium chloride (mM KCl), resulting in slow motion, and video data with relatively low noise. However, when higher concentrations of 3 molar potassium chloride (3M KCl) concentration are used, the rods become more mobile, which interferes with high-resolution imaging.

Using the tools introduced in Case 2, it becomes possible to reduce noise and track the emergence of molecular order during mica18 assembly in 3M KCl. Finally, order parameters for crystallinity were extracted using each nanorod's angles and center of mass. Ten predictors with ensembles of 100 models each were trained to provide semantic segmentation predictions for every movie frame. The methods described in Case 2 were a starting point. Several changes were necessary to improve upon the predictions. The images were pre-processed to reduce the noise amplitude; essentially, the z height of the image was clipped to twice the height of one layer of micaN on the surface. Training cycles were optimized to achieve ~85% accuracy on the test data set. The ensemble of 10 predictors was then averaged to precisely predict each frame.

Because the mica18 nanorods are densely packed in 3M KCl, the instance segmentation used in previous works [55] was not a viable option. Instance segmentation was achieved using a convolution-based pattern matching process with a reference image matching algorithm. A reference rod image was compared to each feature on the screen, and the best match for angle and position was kept. In this way, the centers of mass and the angles of all the rods were aggregated into a database. For each frame, the center of mass for each rod was plotted. FFTs were generated from the center of mass plots, as a method to find periodicities in the structure. The intensity of the center peak is proportional to the number of recognized rods in each image, [57] and the intensity of all FFTs of the videos was normalized to the intensity of the brightest center peak from the videos. By carefully comparing the intensity of the second brightest peak with simulations of many differently ordered rod assemblies, it was found that the intensity of the second brightest peak is proportional to the degree of smectic order for the image. The

degree of 2D liquid crystallinity (2D smectic order parameter) was tracked by plotting the intensity of the second-highest peak in the FFT. (Fig. 10)

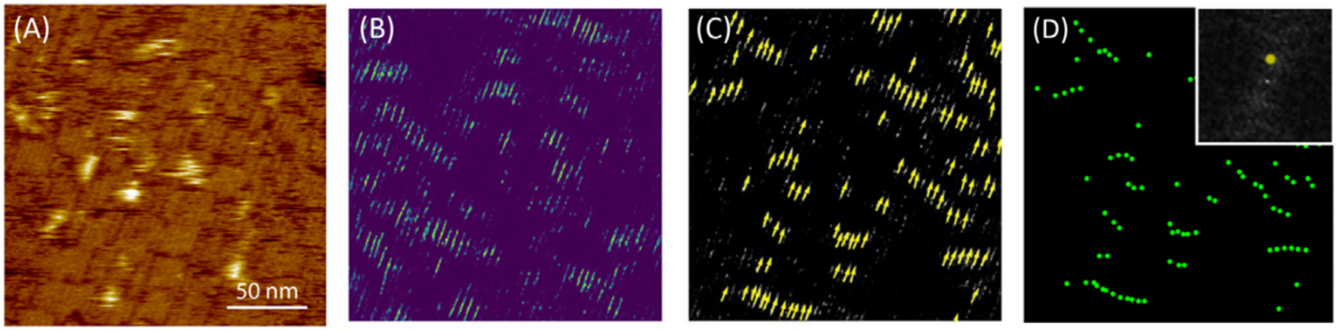

Figure 3.10: Pipeline for applying ML to HS-AFM data collected in 3mM KCl.
(A) AFM Data (B) DCNN semantic segmentation (C) Rod recognition (D) plot with the center of mass of each detected rod, inset FFT with the second-highest peak marked in yellow

As shown in Figure 3.10, even with the adapted tools for this application protein rods are lost to noise in each processing step of the pipeline. FFT is a tool that recognizes periodicity thus we can extract an order parameter to quantify the relative order of the structures that are recognized by the machine learning process herein.

Employing and optimizing DCNN technology offers a solution to analyze data that was previously of a purely qualitative nature as there was simply no way to discern enough of the data to make meaningful calculations or models for assembly, because the signal-to-noise ratio is too high while the rods are still organizing on the surface. Once the rods assemble, there is a stabilizing interaction between the rods, the mobility decreases, and the signal-to-noise increases. Thus, we now have a tool for analyzing protein assembly dynamics and assembly mechanisms with a ML-boosted HS-AFM protocol, even in very challenging imaging conditions.

3.5 Case 4: Adapting analysis methods to recognize multiple lengths of rods

The micaN protein itself is modular, in that it can be expressed with more or fewer repeats to make the protein longer or shorter. The number of repeats affects the K_d of the molecule based on the number of interactions that the molecule has with the surface. An assay which adds an additional degree of complexity to the assembly process is to study the assembly of multiple lengths of protein rods in the same solution. This is a step towards designing hierarchical multi component functional materials. End state analysis at the high mobility condition of 3M KCl produces densely packed domains which again need sophisticated analysis methods, as rods are packed together too closely for conventional recognition software. Herein we explore the physical properties and order parameters that can be learned through the end state analysis of Mica N co-assembly.

The analysis methods which were applied to the Case 3 studies were adapted and used to analyze the MicaN co-assembly data. Number of rods from each length are recognized and grouped into domains.

To achieve this multi-length recognition, new datasets were marked which included both lengths of rods. Semantic segmentation was again used to obtain pixel by pixel recognition of whether a given pixel contained a rod or did not. A reference rod image was generated for each length of micaN, and the previously used convolution-based pattern matching process with a reference image matching algorithm was adapted. Each rod was binned into either mica18 or mica34.

Figure 3.11 shows exemplar plots for co-assembly, from the AFM image through semantic segmentation, instance segmentation with maximal separation, and plotting physical phenomenon extracted from the AFM image.

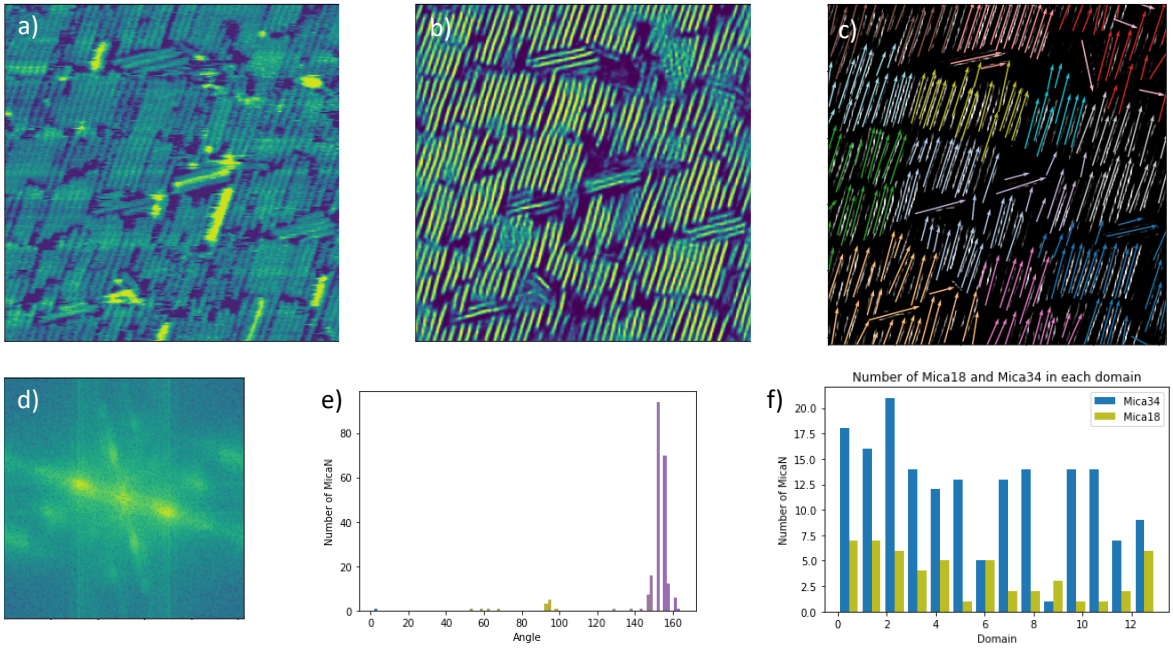

Figure 3.11: Co-Assembly analysis example
a) original AFM data obtained by Sakshi Yadav of a 200 nm square sample region, b) image with reduced noise after semantic segmentation, c) plot of recognized rods after rod recognition, maximal separation, and domain classification, d) FFT of the image after semantic segmentation, showing a dominant angle peak with a secondary angle peak for rod alignment, and additional structure peaks for the average distance between rods, e) histogram showing the angle of every rod that was recognized from the semantic segmentation image, f) histogram showing the number of Mica18 and Mica34 in each domain

This pipeline enabled analysis of the assembly of two species of molecules on the surface of an interface between liquid and substrate.

3.6 Summary

This chapter has described how ML has enabled quantitative analysis of HS-AFM data to discover the physical phenomena governing protein dynamics and ordering at solid-liquid interfaces. The research detailed in this chapter modeled the rotation models of protein nanorods, the discovery of which would otherwise not be possible. By tracking the trajectories of individual protein rods from frame to frame, it was possible to model Brownian type motion and behaviors and Levy-flight dynamics that had not previously been shown.

We also described the application of the Python package AtomAI, which has been developed specifically to analyze and extract physical phenomena, providing exemplar code for training an ensemble of deep neural networks to produce the semantic segmentation of AFM data and functions for encoding and decoding local environments.

We last described a combinatorial approach to analyze very noisy data with a densely covered substrate where the emergence of order for the protein liquid crystals could be elucidated. By combining the methods from Case 1 and 2, it was possible to obtain the center of mass and angle for each rod in the images and track the assembly of the rods over time into a 2D liquid crystal array on the surface of mica. Further refinement of the pipeline enabled analysis of the co-assembly of multiple aspect ratios of the micaN system.

3.7 References

1. Liu, X., et al., *Power generation from ambient humidity using protein nanowires.* Nature, 2020. **578**(7796): p. 550-554.
2. Peydayesh, M. and R. Mezzenga, *Protein nanofibrils for next generation sustainable water purification.* Nature Communications, 2021. **12**(1): p. 3248.
3. Dong, Z., et al., *Bridging Hydrometallurgy and Biochemistry: A Protein-Based Process for Recovery and Separation of Rare Earth Elements.* ACS Central Science, 2021. **7**(11): p. 1798-1808.
4. Yaman, M.Y., et al., *Learning and Predicting Photonic Responses of Plasmonic Nanoparticle Assemblies via Dual Variational Autoencoders.* Small, 2023. **19**(25): p. 2205893.
5. Oohora, K., A. Onoda, and T. Hayashi, *Hemoproteins Reconstituted with Artificial Metal Complexes as Biohybrid Catalysts.* Accounts of Chemical Research, 2019. **52**(4): p. 945-954.
6. Li, X., et al., *Highly active enzyme–metal nanohybrids synthesized in protein–polymer conjugates.* Nature Catalysis, 2019. **2**(8): p. 718-725.
7. Quijano-Rubio, A., et al., *De novo design of modular and tunable protein biosensors.* Nature, 2021. **591**(7850): p. 482-487.
8. Anishchenko, I., et al., *De novo protein design by deep network hallucination.* Nature, 2021. **600**(7889): p. 547-552.
9. Boyoglu-Barnum, S., et al., *Quadrivalent influenza nanoparticle vaccines induce broad protection.* Nature, 2021. **592**(7855): p. 623-628.
10. Li, D., et al., *Nanoparticle Assembly and Oriented Attachment: Correlating Controlling Factors to the Resulting Structures.* Chemical Reviews, 2023. **123**(6): p. 3127-3159.
11. Luo, Q., et al., *Protein Assembly: Versatile Approaches to Construct Highly Ordered Nanostructures.* Chemical Reviews, 2016. **116**(22): p. 13571-13632.
12. Lin, Y.-C., et al., *Force-induced conformational changes in PIEZO1.* Nature, 2019. **573**(7773): p. 230-234.
13. Ruan, Y., et al., *Direct visualization of glutamate transporter elevator mechanism by high-speed AFM.* Proceedings of the National Academy of Sciences, 2017. **114**(7): p. 1584-1588.
14. Pashley, R.M., *Hydration forces between mica surfaces in aqueous electrolyte solutions.* Journal of Colloid and Interface Science, 1981. **80**(1): p. 153-162.
15. Ido, S., et al., *Immunoactive two-dimensional self-assembly of monoclonal antibodies in aqueous solution revealed by atomic force microscopy.* Nature Materials, 2014. **13**(3): p. 264-270.
16. Uchida, M., et al., *Modular Self-Assembly of Protein Cage Lattices for Multistep Catalysis.* ACS Nano, 2018. **12**(2): p. 942-953.
17. Said, M.Y., et al., *Exploration of Structured Symmetric Cyclic Peptides as Ligands for Metal-Organic Frameworks.* Chemistry of Materials, 2022. **34**(21): p. 9736-9744.
18. Courbet, A., et al., *Computational design of mechanically coupled axle-rotor protein assemblies.* Science, 2022. **376**(6591): p. 383-390.
19. Brunette, T.J., et al., *Exploring the repeat protein universe through computational protein design.* Nature, 2015. **528**(7583): p. 580-584.
20. Divine, R., et al., *Designed proteins assemble antibodies into modular nanocages.* Science, 2021. **372**(6537): p. eabd9994.
21. Hellner, B., et al., *Phase Control of Nanocrystalline Inclusions in Bioprecipitated Titania with a Panel of Mutant Silica-Binding Proteins.* Langmuir, 2020. **36**(29): p. 8503-8510.

22. Coloma, A., A. Velty, and U. Díaz, *Hybrid organic–inorganic nanoparticles with associated functionality for catalytic transformation of biomass substrates.* RSC Advances, 2023. **13**(15): p. 10144-10156.
23. Berg, J.C., *An Introduction to Interfaces and Colloids*. An Introduction to Interfaces and Colloids.
24. Israelachvili, J.N., *10 - Unifying Concepts in Intermolecular and Interparticle Forces*, in *Intermolecular and Surface Forces (Third Edition)*, J.N. Israelachvili, Editor. 2011, Academic Press: San Diego. p. 191-204.
25. Ando, T., *High-speed AFM imaging.* Current Opinion in Structural Biology, 2014. **28**: p. 63-68.
26. De Yoreo, J.J., S. Chung, and R.W. Friddle, *In Situ Atomic Force Microscopy as a Tool for Investigating Interactions and Assembly Dynamics in Biomolecular and Biomineral Systems.* Advanced Functional Materials, 2013. **23**(20): p. 2525-2538.
27. Chung, S., et al., *Self-catalyzed growth of S layers via an amorphous-to-crystalline transition limited by folding kinetics.* Proceedings of the National Academy of Sciences, 2010. **107**(38): p. 16536-16541.
28. Zhang, S., et al., *Assembly of a patchy protein into variable 2D lattices via tunable multiscale interactions.* Nature Communications, 2020. **11**(1): p. 3770.
29. Müller, D.J. and Y.F. Dufrêne, *Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology.* Nature Nanotechnology, 2008. **3**(5): p. 261-269.
30. Morrow, B.H., G.F. Payne, and J. Shen, *pH-Responsive Self-Assembly of Polysaccharide through a Rugged Energy Landscape.* Journal of the American Chemical Society, 2015. **137**(40): p. 13024-13030.
31. Walther, A., *Viewpoint: From Responsive to Adaptive and Interactive Materials and Materials Systems: A Roadmap.* Advanced Materials, 2020. **32**(20): p. 1905111.
32. Zhang, S., et al., *Coexistence of ribbon and helical fibrils originating from hIAPP_{20–29} revealed by quantitative nanomechanical atomic force microscopy.* Proceedings of the National Academy of Sciences, 2013. **110**(8): p. 2798-2803.
33. De Yoreo, J.J. and P.G. Vekilov, *Principles of Crystal Nucleation and Growth.* Reviews in Mineralogy and Geochemistry, 2003. **54**(1): p. 57-93.
34. Zhang, S., et al., *Engineering Biomolecular Self-Assembly at Solid–Liquid Interfaces.* Advanced Materials, 2021. **33**(23): p. 1905784.
35. Kodera, N., et al., *Video imaging of walking myosin V by high-speed atomic force microscopy.* Nature, 2010. **468**(7320): p. 72-76.
36. Igarashi, K., et al., *Two-way traffic of glycoside hydrolase family 18 processive chitinases on crystalline chitin.* Nature Communications, 2014. **5**(1): p. 3975.
37. Heath, G.R. and S. Scheuring, *Advances in high-speed atomic force microscopy (HS-AFM) reveal dynamics of transmembrane channels and transporters.* Current Opinion in Structural Biology, 2019. **57**: p. 93-102.
38. Jiao, F., et al., *Perforin-2 clockwise hand-over-hand pre-pore to pore transition mechanism.* Nature Communications, 2022. **13**(1): p. 5039.
39. Chen, J., et al., *Building two-dimensional materials one row at a time: Avoiding the nucleation barrier.* Science, 2018. **362**(6419): p. 1135-1139.
40. Ziatdinov, M., et al., *AtomAI framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy.* Nature Machine Intelligence, 2022. **4**(12): p. 1101-1112.

41. Oxley, M.P., et al., *Probing atomic-scale symmetry breaking by rotationally invariant machine learning of multidimensional electron scattering.* npj Computational Materials, 2021. **7**(1): p. 65.
42. Ziatdinov, M. and S. Kalinin, *AtomAI: Open-source software for applications of deep learning to microscopy data.* Microscopy and Microanalysis, 2021. **27**(S1): p. 3000-3002.
43. Yaman, M.Y., et al., *Alignment of Au nanorods along de novo designed protein nanofibers studied with automated image analysis.* Soft Matter, 2021. **17**(25): p. 6109-6115.
44. Li, J., et al., *Machine Vision Automated Chiral Molecule Detection and Classification in Molecular Imaging.* Journal of the American Chemical Society, 2021. **143**(27): p. 10177-10188.
45. Pyles, H., et al., *Controlling protein assembly on inorganic crystals through designed protein interfaces.* Nature, 2019. **571**(7764): p. 251-256.
46. Giessibl, F.J., *Advances in atomic force microscopy.* Reviews of Modern Physics, 2003. **75**(3): p. 949-983.
47. Tomasi, C. and R. Manduchi. *Bilateral filtering for gray and color images*. in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*. 1998.
48. Zuiderveld, K., *VIII.5. Contrast Limited Adaptive Histogram Equalization*, in *Graphics Gems*. 1994. p. 474-485.
49. Akeret, J., et al., *Radio frequency interference mitigation using deep convolutional neural networks.* Astronomy and Computing, 2017. **18**: p. 35-39.
50. Yang, F., et al. *Cell Segmentation, Tracking, and Mitosis Detection Using Temporal Context*. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*. 2005. Berlin, Heidelberg: Springer Berlin Heidelberg.
51. Shin, S.-H., et al., *Direct observation of kinetic traps associated with structural transformations leading to multiple pathways of S-layer assembly.* Proceedings of the National Academy of Sciences, 2012. **109**(32): p. 12968-12973.
52. Zanette, D.H. and P.A. Alemany, *Thermodynamics of Anomalous Diffusion.* Physical Review Letters, 1995. **75**(3): p. 366-369.
53. Skaug, M.J., J. Mabry, and D.K. Schwartz, *Intermittent Molecular Hopping at the Solid-Liquid Interface.* Physical Review Letters, 2013. **110**(25): p. 256101.
54. Gordon, O., et al., *Scanning tunneling state recognition with multi-class neural network ensembles.* Review of Scientific Instruments, 2019. **90**(10).
55. Kalinin, S.V., et al., *Disentangling Rotational Dynamics and Ordering Transitions in a System of Self-Organizing Protein Nanorods via Rotationally Invariant Latent Representations.* ACS Nano, 2021. **15**(4): p. 6471-6480.
56. Ziatdinov, M., et al., *Quantifying the Dynamics of Protein Self-Organization Using Deep Learning Analysis of Atomic Force Microscopy Data.* Nano Letters, 2021. **21**(1): p. 158-165.
57. Taylor, F.J., *Signal Processing, Digital.* 2003.

4. Symmetry Breaking Drives Protein Assembly into a Forbidden Two-Dimensional Liquid-Crystal Phase

The material in this chapter is partially reproduced from a manuscript in preparation by Yadav Schmid, Helfrecht, Stegmann *et all.*

4.1 Abstract

The era of protein design has enabled the creation of hybrid protein-inorganic interfaces but resulting patterns of protein assembly often diverge from target designs, implying that essential interactions are unaccounted for in current design platforms. Here, we use high-speed AFM analyzed through machine learning to follow the assembly of protein nanorods in aqueous electrolytes on two types of mica exhibiting disparate symmetry elements, which are imprinted on the overlying hydration structure. Using Monte Carlo simulations, we reproduce the observed phases and show that an observed smectic phase, previously thought to be forbidden for non-interacting rods in two dimensions, emerges when crystal symmetry introduces a directional bias. The findings demonstrate the importance of incorporating colloidal forces and hydration structure inherent to interfacial systems into protein design platforms.

4.2 Introduction

Nature's ability to create hybrid biomolecular-inorganic materials with a hierarchy of structural motifs has inspired extensive research into the assembly of macromolecules at solid surfaces[1–3]. Efforts to design proteins, peptides and peptide mimetics to self-assemble at surfaces and exhibit order across multiple length scales have led to promising progress with potential applications in energy harvesting, catalysis, photonics, biomedicine, and structural materials[4–10].

Strategies to direct protein assembly have generally focused on manipulating chemical bonds between the proteins, electrostatic interactions between charged functional sidechains and surfaces, and hydrophobic interactions both between the proteins themselves and with the surface[11–13]. However, even when a set of interactions has intentionally been designed into the sequence, phases emerge that do not represent the design point. For example, when the enzyme RhuA was site-specifically modified to form two-dimensional (2D) lattices through in-plane chemical bonds, assembly in bulk solution produced the target structure, but assembly on mica surfaces led to three distinct and unexpected monolayer and bilayer phases [14]. In other research using *de novo* design to create protein nanorods exhibiting a precise match to the crystal lattice of mica, a rich set of ordered 2D structures was observed, but many were not design targets and could not have been anticipated from lattice matching alone[5].

The common occurrence of unexpected phases demonstrates that the designed interactions do not fully encompass the suite of forces that drive protein assembly at surfaces and cannot be used to predict the outcome. Thus, computational protein design platforms that have proven successful for predicting protein folding and designing quaternary protein architectures can not yet be exploited to design hybrid structures. Given that proteins, despite their atomic precision, are akin to patchy particles interacting with both the surface and the surrounding solution, the role of colloidal forces and the impact of the interfacial solution structure, both of which are modulated by solution electrolytes and the broken symmetry imposed by the substrate, [6, 8, 15, 16] must be considered but are not currently an element of these design platforms.

Here, we explore the impact of inherent interfacial forces on the off-target assembly of designed proteins. We use machine learning to analyze high-speed (HS-) atomic force microscopy (AFM) data and quantify the assembly dynamics and degree of order achieved by protein nanorods during assembly on surfaces to which they are lattice-matched. We additionally employ Monte Carlo simulations to

predict the dependence of order on chemical potential, mobility, and symmetry of the potential energy landscape. The results show that the consideration of these interfacial forces leads to accurate prediction of the observed 2D phases. In particular, the emergence of a 2D smectic phase, which is predicted to be unstable by long-standing theories of colloidal systems[17], is shown to arise naturally when symmetry breaking by the underlying surface creates a two-fold symmetric potential energy landscape. The results point to a way of integrating coarse grain simulations of colloidal systems into *de novo* protein design platforms to accurately design supramolecular protein architectures at inorganic interfaces.

4.3 Results and Discussion

To investigate the role of interfacial structure on protein assembly we utilized a rectangular rod-shaped *de novo* designed helical repeat (DHR) protein[5] — referred to as DHR10-mica18 — with an aspect ratio of 5.6. The protein consists of 18 repeating subunits, each with three glutamate residues positioned to exhibit a structural match to the (001) potassium sublattice of muscovite (m-) and fluorophlogopite (f-) mica (Fig. 4.1A, B). In agreement with previous research, we find that, in 100 mM K^+ aqueous solutions, these protein nanorods assemble into a disordered phase consisting of small domains of coaligned proteins oriented along the three principal axes of the underlying mica lattice, regardless of whether f- or m-mica is used (Fig. 4.1C, D). However, when the K^+ concentration is increased to 3M, DHR10-mica18 continues to form the three-fold disordered phase on f-mica (Fig. 4.1E), but it assembles into an ordered phase on m-mica in which all rods are coaligned along a single direction and arranged in parallel rows (Fig. 4.1F).

In the parlance of the liquid crystal literature, the observed disordered phase is known as a 2D high-density disordered (HDD) phase (Fig. 4.1G) and is predicted for a sufficiently high rod concentration in a 3-fold potential when the translational and rotational mobility are low[18, 19]. As the mobility increases, theoretical treatments predict that the rods will align due to purely entropic forces,

forming a 2D nematic phase[20] (Fig. 4.1H), which we do not observe experimentally. Instead, the ordered phase observed on m-mica at 3M K^+ has smectic order (Fig. 4.1I) which is not predicted for non-interacting rectangular rods in 2D(*20*), though the introduction of excluded volume interactions at the rod tips due to the addition of polymer tails[17] or charges that produce electrostatic repulsion[21] has been predicted to stabilze smectic order.

The above observations thus present a conundrum: if the rod-substrate potential is established by the interaction of the glutamate side chains with the K^+ ions of the mica lattice, then why should the nanorods assemble into two distinct phases on m- and f-mica, which possess identical K^+ sublattices, and, for the case of m-mica, why does smectic order emerge in a two-dimensional system? To answer these questions, we used high-speed AFM (HS-AFM) to observe protein adsorption and assembly at the water-mica interface[1, 20], follow the emergence of order, and quantify the degree of nematic and smectic order on both substrates.

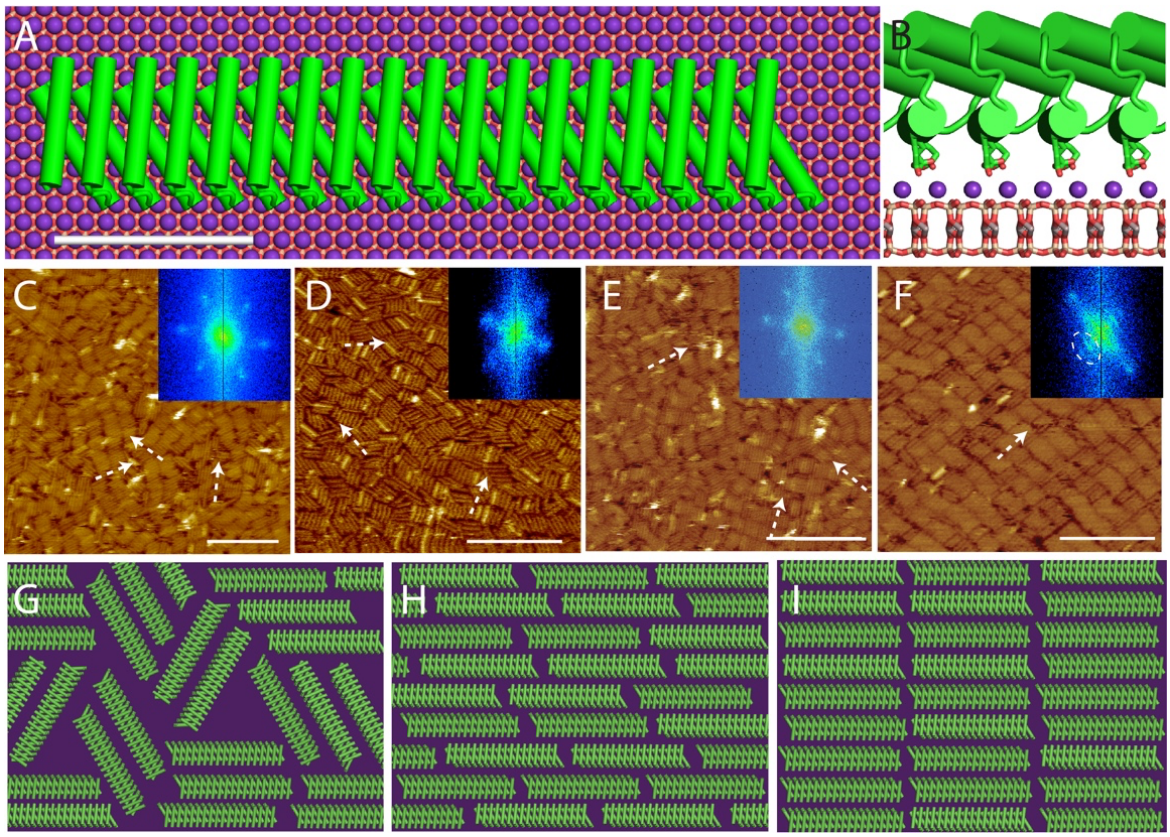

Figure 4.1: (A) The Rosetta model of a DHR10-mica18 protein nanorod with dimesions 3.6 nm x 20 nm adsorbed on mica surface. The protein consists of 18 tandem repeat units shown in green with alpha-helices rendered as cylinders. An aluminosilicate layer of the mica substrate is shown with the K+ sublattice (shown as purple spheres). Scale bar is 5nm. (B) Side view of the protein-mica interface showing negatively charged glutamate side chains (green and red sticks, respectively) extending from the protein with a periodicity that forms a 2-to-1 lattice match with the mica surface. (C-D) AFM images of the final assembly states of DHR-mica18 on f-mica and m-mica in 100 mM and (E-F) 3 M KCl, respectively, the fast Fourier transform is shown in the inset. Scale bars are 100 nm. G-I) 2D phases of hard rods possible at high concentrations in a three-fold potential: (G) 2D high-density disordered (HDD) phase, (H) nematic phase, and (I) smectic phase. Note that the smectic phase is not predicted when the rods are non-interacting(20).

In the case of m-mica in 3M K^+ (Fig. 4.2), individual proteins are rarely observable in the initial frames; rather DHR10-mica18 initially forms a 2D liquid phase in which the proteins have high in-plane mobility (Fig. 4.2A, t_0+18.92 s). Gradually, the proteins become visible as small domains of two to four coaligned nanorods that are short-lived, often appearing for only a single frame (Fig. 4.2A, t_0+75.68 s to t_0+277.03 s), but, with time, become larger and longer lived (Fig. 4.2A, t_0+277.03 s to t_0+337.84 s) until a stable smectic phase emerges (Fig. 4.2A, t_0+337.84 s to t_0+940.54 s).

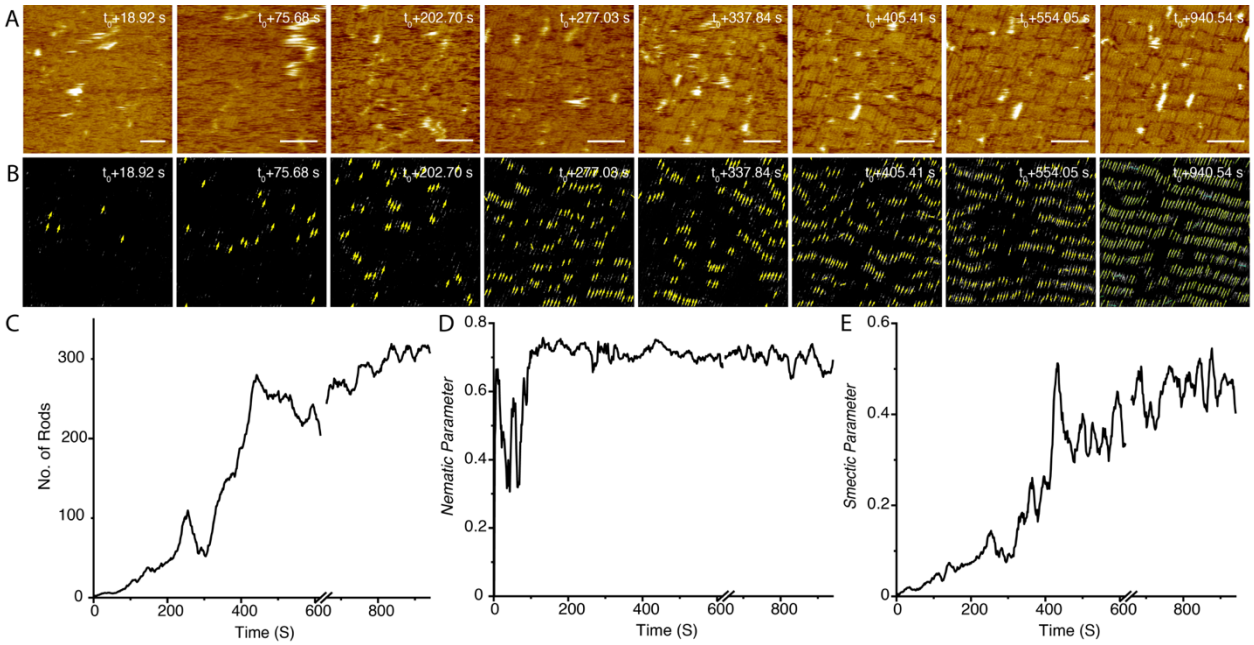

Figure 4.2: (A) HS-AFM images from supplementary movie M1 showing the translational motion of protein rods and their assembly on m-mica into a smectic phase. Scale bar equals 50 nm (B) Machine-learning-based workflow to recognize the protein rods in the HS-AFM images in (A). (C) Number of recognized protein rods as a function of time. (D – E) Nematic and Smectic order parameters as a function of time, respectively, for the observed rod assembly in (A).

To quantify the dynamics of assembly and the degree of order, we used a computational workflow where an ensemble of deep convolutional neural networks was employed to achieve semantic segmentation, which was then used as input to a conventional algorithm for rod recognition (Fig. 4.2B). This approach was used to establish the total coverage (Fig. 4.2C), as well as the orientation and center of mass of each protein, which were in turn used to obtain the nematic (Fig. 4.2D) and smectic (Fig. 4.2E) order parameters, respectively (see methods section for detailed calculation of order parameters.) As the analysis shows, virtually every nanorod visible in any frame is oriented along a single direction. Thus, the nematic order rises rapidly and saturates. In contrast, the smectic order increases gradually, growing slowly at first and then rapidly transitioning to higher values before saturating as the surface becomes densely packed (Fig. 4.2E). This behavior mirrors that of the surface coverage (Fig. 4.2C) and shows that, as the domains become closely spaced, a percolation threshold is reached at which the

probability of rod attachment increases rapidly and leads to high coverage, while the high degree of translational mobility enables the domains to align and reach high smectic order.

In the case of f-mica at 3M K^+ (Fig. 4.3), rods are already visible in the initial frames, both as individual rods and small domains, and unlike the case of m-mica, are oriented along all three K^+ sublattice directions with roughly equal probability (Fig. 4.3A). Individual domains fluctuate in size but, on average, grow as individual rods attach. Furthermore, unlike the m-mica case, no rapid transition in either coverage or order parameter is observed. Most importantly, the resulting HDD phase exhibits low values for the nematic and smectic order parameters, with the latter being nearly zero at all times.

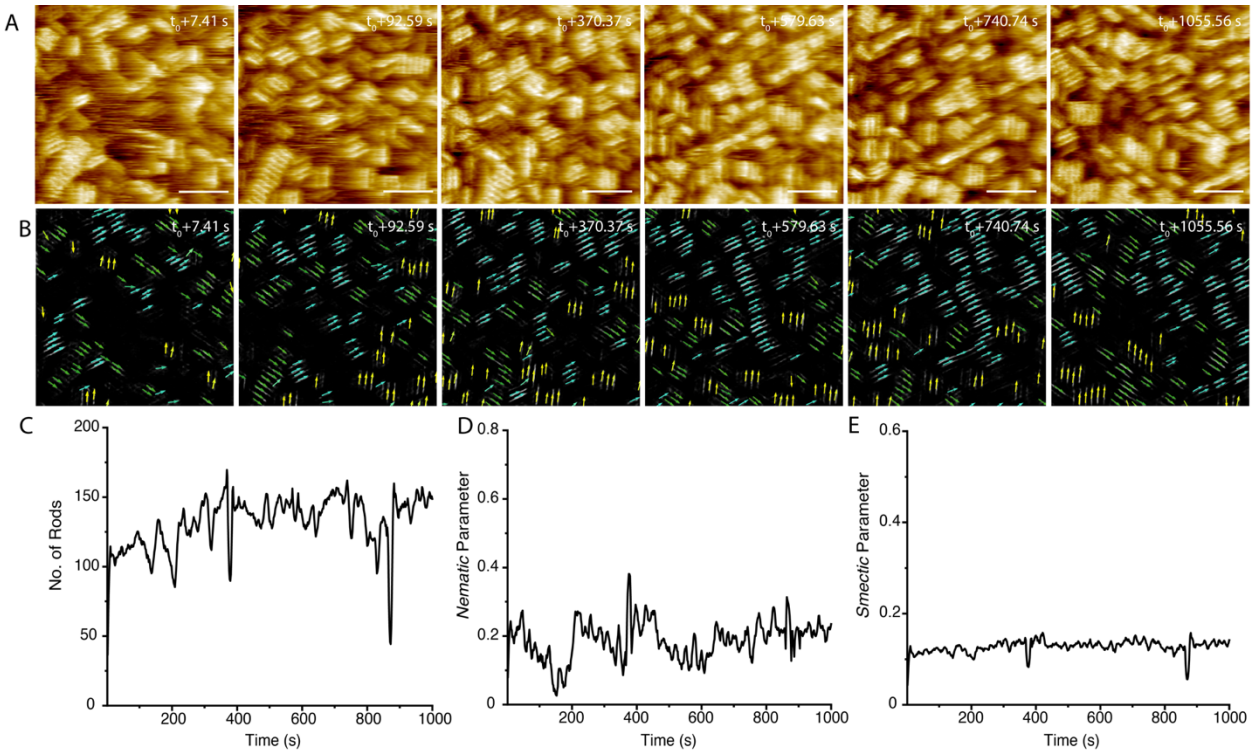

Figure 4.3: (A) HS-AFM images from supplementary movie M2 showing the translational and rotational motion of protein rods and their assembly on f-mica along the three K+ sublattice directions into a high-density disordered phase. Scale bar equals 50 nm. (B) Machine learning-based workflow to recognize the protein rods in the HS-AFM images in (A). (C) Number of recognized protein rods as a function of time. (D – E) Nematic and Smectic order parameters as a function of time, respectively, for the observed rod assembly in (A).

The differences in assembly on m- and f-mica are correlated with two differences in the structure of the interface, one associated with the mica lattice itself and the other with the overlying hydration layers. The structural difference lies in the inherent symmetry of the tetrahedral-octohedral-tetrhedral (TOT) layers that comprise each mica. When cleaved, mica splits between TOT layers to expose a framework of silicate tetrahedra and a nearly-hexagonal sublattice of K^+ ions that occupy the cavities of that framework. Neither f-mica nor m-mica TOT sheets have perfect hexagonal symmetry, but the silicate framework for f-mica is much closer to hexagonal symmetry than for m-mica. The structural-driver of m-mica's broken surface symmetry lies in the distribution of octahedral cations within the TOT layer. As a dioctoehdral mica, only 2/3 of the octahedral sites in m-mica are filled with trivalent cations, leaving one third of the octohedral sites empty for charge balance. These vacant sites are ordered and cause the overlying silicate tetrahedra to systematically tilt, an effect which is enhanced strongly in one direction due to couplings between silicate-oxygens and hydroxyl groups of the TOT layer. As a result, the cleaved m-mica surface displays distinct corrugations in one direction. In contrast, fluorphlogopite f-mica is trioctohedral, meaning that divalent cations fill every octahedral site to produce a symmetric core. The f-mica also lacks hydroxyls. Together, this leaves the tetrahedral silicate layer of f-mica much less distorted and closer to hexagonal symmetry.

The consequence of the underlying vacancies in the structure of the m-mica surface lattice is minor: only a sub-Å decrease in the height of a set of surface oxygens is produced, but that change in height has a dramatic effect on the overlying hydration structure. Recent molecular simulations predict two highly organized water layers above both mineral surfaces at high K^+ concentrations[8]. In the first layer, both minerals exhibit a hexagonal pattern of water oxygen density just above the K^+ ions. However, while the orientation of the water dipoles above f-mica point in three equivalent directions along the surface, those above m-mica are rotated into a single alignment due to the lower position of

the surface oxygens. The consequence is that the pattern of both water hydrogen and oxygen densities in the second layer remains hexagonal for f-mica but evolves into stripes on m-mica that are oriented along the unique axis of m-mica, which is the direction along which the nanorods align. These predictions were experimentally corroborated by using 3D atomic force microscopy (3D-AFM) to image the solution structure above the mica surface at high K^+ concentrations and indeed, a hexagonal pattern of water in the first layer above m-mica evolves into a striped pattern in the second layer[8].

To test whether a change in the potential energy landscape due to the altered symmetry of the water structure in going from f- to m-mica can lead to the observed behavior of DHR10-mica18 as a consequence of purely colloidal forces, we performed grand canonical Monte Carlo simulations of hard rods depositing on a surface in two scenarios (Fig. 4.4): 1) the three lattice vectors that define the rod orientations have equal probability of occupancy — i.e., the potential is three-fold symmetric — to represent the case of f-mica (Fig. 4.4B, G-J), and 2) one of the three orientations is twice as energetically favorable compared to the others — i.e., the potential is quasi-two-fold — to represent m-mica (Fig. 4.4A, C-F). For each, we systematically varied the chemical potential, which controls the surface concentration of rods, and rod mobility, which determines the maximum distance a rod is allowed to move during a Monte Carlo step.

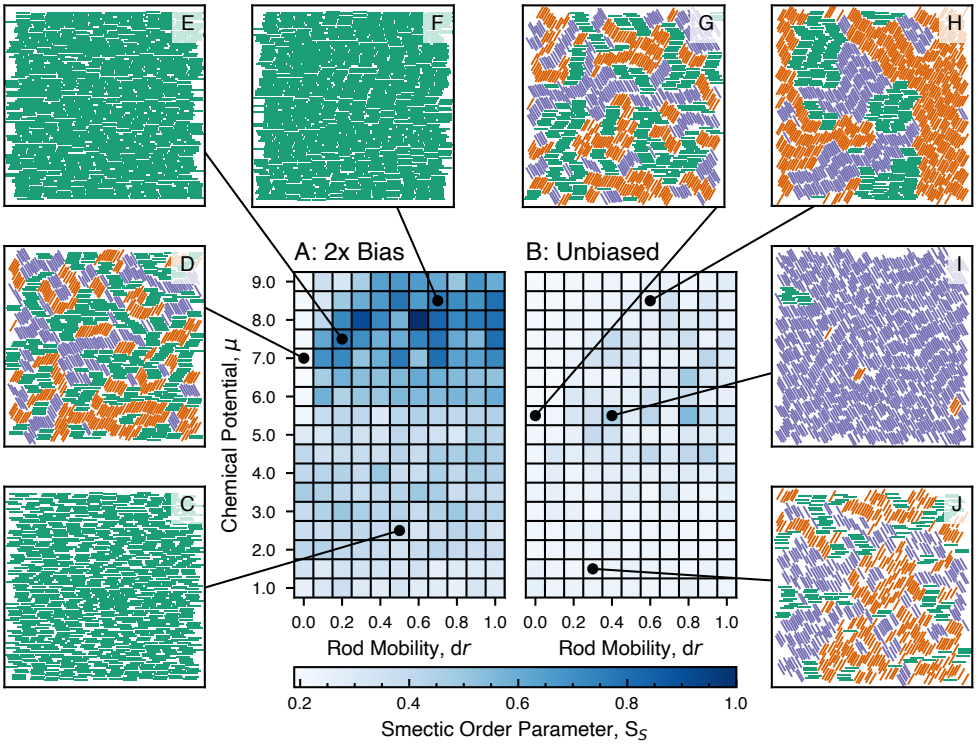

Figure 4.4: (A-B) The smectic order parameters for a collection of Monte Carlo simulations of hard rods with aspect ratio $\ell = 7$ are presented in grids, where each box in the grid represents a single simulation defined by its chemical potential and rod mobility; the coloring indicates the value of the smectic order parameter. Separate grids are plotted for simulation collections where (B) all three rod orientations are equally favorable ("Unbiased"), and (A) where the horizontal rods are twice as energetically favorable as the other orientations ("2x Bias"). (C-J) Several select simulations are annotated with a snapshot of the final rod configuration.

The results show that when the potential is three-fold symmetric (Fig. 4.4B), a rod mobility of zero leads to a three-fold disordered phase regardless of the value of the chemical potential, consistent with previous findings for low mobility rods on a triangular lattice[19]. As the chemical potential increases across simulations of rods with nonzero mobility, a transient nematic phase emerges (Fig. 4.4I) before being replaced, at high chemical potential, by a phase composed of large ordered domains. The smectic order parameter is nearly zero for all conditions where all three rod orientations are equally favorable. These results are consistent with the experimental observations for f-mica, on which the protein mobility is too low for even small domains to reorient into alignment with their neighbors.

In contrast, when one of the rod orientations is more energetically favorable than the other two (Fig. 4.4A), a significant degree of smectic order emerges for all conditions investigated, provided the

rods have adequate mobility and sufficiently high chemical potential, while a three-fold HDD phase is observed only at zero mobility and high chemical potential. The fact that we only observe the smectic phase when we apply an orientational bias to the model system suggests that the smectic phase observed for *non-interacting* rectangular rods is solely due to the emergence of a two-fold rod-surface potential on m-mica[8, 16]. This remarkable result demonstrates that the smectic order observed in these systems is mediated by the molecular details of solution-surface interactions rather than rod-rod interactions.

Overall, the simulation results are consistent with the experimental observations of protein nanorod assembly on m-mica and f-mica at low and high ion concentration, respectively (Fig. 4.4C-F), with one caveat: while the simulation model predicts a three-fold HDD phase at high chemical potential for m-mica (Fig. 4.4D), it is only observed when the rods have zero mobility, a condition that currently cannot be corroborated experimentally.

4.4 Conclusion

The above findings show that designed interactions between proteins and substrates do not solely govern the 2D assembly of proteins on surfaces. Rather, the entropic and colloidal forces that drive the assembly at the interface must be accounted for. However, the results also show that the colloidal forces at work are distinct from those in the bulk solution, because the symmetry of the substrate is imprinted onto the solvent structure near the interface. Most importantly, as shown by DHR10-micaN protein nanorods on m-mica at 3M K^+, the combination of designed interactions to create a preference for binding along the three equivalent directions of the K^+ mica sublattice and the two-fold bias introduced into the potential energy landscape by the interfacial solvent structure in response to the broken three-fold symmetry of the m-mica lattice enables the proteins to assemble into a smectic phase, which would otherwise not exist in 2D.

The results presented here also suggest an approach to utilizing computational platforms for *de novo* protein design, such as Rosetta, to improve the design of protein-inorganic interfaces. While those platforms have proven capable of designing specific patterns of molecular interactions both between proteins and substrates as well as between the proteins themselves, combining those predictions with coarse grain simulations that can account for colloidal forces due to both shape complementarity[23] and the structure of the interfacial solution can lead to a new class of design platforms for hybrid materials.

4.5 References

1. S. Y. Schmid, K. Lachowski, H. T. Chiang, L. Pozzo, J. De Yoreo, S. Zhang, Mechanisms of Biomolecular Self-Assembly Investigated Through In Situ Observations of Structures and Dynamics. *Angewandte Chemie - International Edition* **202309725** (2023).
2. B. Jin, F. Yan, X. Qi, B. Cai, J. Tao, X. Fu, S. Tan, P. Zhang, J. Pfaendtner, N. Y. Naser, F. Baneyx, X. Zhang, J. J. DeYoreo, C. Chen, Peptoid-Directed Formation of Five-Fold Twinned Au Nanostars through Particle Attachment and Facet Stabilization. *Angewandte Chemie International Edition* **61** (2022).
3. Q. Luo, C. Hou, Y. Bai, R. Wang, J. Liu, Protein Assembly: Versatile Approaches to Construct Highly Ordered Nanostructures. American Chemical Society [Preprint] (2016). https://doi.org/10.1021/acs.chemrev.6b00228.
4. X. Zhang, S. Kang, K. Adstedt, M. Kim, R. Xiong, J. Yu, X. Chen, X. Zhao, C. Ye, V. V. Tsukruk, Uniformly aligned flexible magnetic films from bacterial nanocelluloses for fast actuating optical materials. *Nat Commun* **13** (2022).
5. H. Pyles, S. Zhang, J. J. De Yoreo, D. Baker, Controlling protein assembly on inorganic crystals through designed protein interfaces. *Nature* **571**, 251–256 (2019).
6. S. Zhang, R. G. Alberstein, J. J. De Yoreo, F. A. Tezcan, Assembly of a patchy protein into variable 2D lattices via tunable multiscale interactions. *Nat Commun* **11**, 1–12 (2020).
7. J. Chen, E. Zhu, J. Liu, S. Zhang, Z. Lin, X. Duan, H. Heinz, Y. Huang, J. J. De Yoreo, Building two-dimensional materials one row at a time: Avoiding the nucleation barrier. *Science (1979)* **362**, 1135–1139 (2018).
8. R. G. Alberstein, J. L. Prelesnik, E. Nakouzi, S. Zhang, J. J. De Yoreo, J. Pfaendtner, F. A. Tezcan, C. J. Mundy, Discrete Orientations of Interfacial Waters Direct Crystallization of Mica-Binding Proteins. *Journal of Physical Chemistry Letters* **14**, 80–87 (2023).
9. S. Yadav Schmid, X. Ma, J. A. Hammons, S. T. Mergelsberg, B. S. Harris, T. Ferron, W. Yang, W. Zhou, R. Zheng, S. Zhang, B. A. Legg, A. Van Buuren, M. D. Baer, C. L. Chen, J. Tao, J. J. De Yoreo, Influence of Peptoid Sequence on the Mechanisms and Kinetics of 2D Assembly. *ACS Nano* **18**, 3497–3508 (2024).
10. M. A. Boles, M. Engel, D. V. Talapin, Self-assembly of colloidal nanocrystals: From intricate structures to functional materials. American Chemical Society [Preprint] (2016). https://doi.org/10.1021/acs.chemrev.6b00196.
11. I. W. Hamley, Protein Assemblies: Nature-Inspired and Designed Nanostructures. American Chemical Society [Preprint] (2019). https://doi.org/10.1021/acs.biomac.9b00228.
12. A. Quijano-Rubio, H. W. Yeh, J. Park, H. Lee, R. A. Langan, S. E. Boyken, M. J. Lajoie, L. Cao, C. M. Chow, M. C. Miranda, J. Wi, H. J. Hong, L. Stewart, B. H. Oh, D. Baker, De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).
13. A. Courbet, J. Hansen, Y. Hsia, N. Bethel, Y.-J. Park, C. Xu, A. Moyer, S. E. Boyken, G. Ueda, U. Nattermann, D. Nagarajan, D.-A. Silva, W. Sheffler, J. Quispe, A. Nord, N. King, P. Bradley, D. Veesler, J. Kollman, D. Baker, "Computational design of mechanically coupled axle-rotor protein assemblies;" https://www.science.org.
14. R. Alberstein, Y. Suzuki, F. Paesani, F. A. Tezcan, Engineering the entropy-driven free-energy landscape of a dynamic nanoporous protein assembly. *Nat Chem* **10**, 732–739 (2018).
15. J. L. Prelesnik, R. G. Alberstein, S. Zhang, H. Pyles, D. Baker, J. Pfaendtner, J. J. de Yoreo, F. A. Tezcan, R. C. Remsing, C. J. Mundy, Ion-dependent protein–surface interactions from intrinsic solvent response. *Proc Natl Acad Sci U S A* **118**, 1–9 (2021).

16. D. Martin-Jimenez, E. Chacon, P. Tarazona, R. Garcia, Atomically resolved three-dimensional structures of electrolyte aqueous solutions near a solid surface. *Nat Commun* **7** (2016).
17. D. A. King, R. D. Kamien, What promotes smectic order: Applying mean-field theory to the ends. *Phys Rev E* **107** (2023).
18. J. Kundu, J. F. Stilck, R. Rajesh, Phase diagram of a bidispersed hard-rod lattice gas in two dimensions. *Epl* **112**, 0–6 (2015).
19. J. Kundu, R. Rajesh, D. Dhar, J. F. Stilck, Nematic-disordered phase transition in systems of long rigid rods on two-dimensional lattices. *Phys Rev E Stat Nonlin Soft Matter Phys* **87**, 1–9 (2013).
20. M. A. Bates, D. Frenkel, Phase behavior of two-dimensional hard rod fluids. *Journal of Chemical Physics* **112**, 10034–10041 (2000).
21. D. A. Triplett, L. M. Quimby, B. D. Smith, D. Hernández Rodríguez, S. K. St. Angelo, P. González, C. D. Keating, K. A. Fichthorn, Assembly of gold nanowires by sedimentation from suspension: Experiments and simulation. *Journal of Physical Chemistry C* **114**, 7346–7355 (2010).
22. Z. Zhai, S. Y. Schmid, Z. Lin, S. Zhang, F. Jiao, Unveiling the nanoscale architectures and dynamics of protein assembly with in situ atomic force microscopy. *Aggregate*, doi: 10.1002/agt2.604 (2024).
23. F. Gao, J. Glaser, S. C. Glotzer, The role of complementary shape in protein dimerization. *Soft Matter* **17**, 7376–7383 (2021).

5. Methods and their background

5.1 Protein design

We used *de novo* designed proteins which were designed to have strong electrostatic patches[1] as a building block. Using point mutations which were modeled using the software PyMOL[2], we changed functional groups on the surface of the protein to generate regions of positive or negative charge. When making point mutations, considerations include helical propensity (the amino acid's tendency to form an alpha helix), surface hydrogen bond networks, and interactions between cations and aromatic functional groups. Using these considerations in our modifications, A113-KM2N was mutated to have very negative end areas, and a strip of positive charge around the center of the axle.

5.2 Isoelectric point

Isoelectric point was calculated using Isoelectric Point Calculator[3]

5.3 Plasmid design and synthesis

Each protein in this study was encoded into a pet29b+ plasmid and ordered from IDT. Pet29b+ was the chosen plasmid because it contains a kanamycin resistance as well as a HIS tag at the N terminus of the protein to aid in purification after expression and immobilization onto a substrate.

5.4 Protein expression

Proteins were expressed in limo E. coli cells from Novagen using Studiers autoinduction media (M2) with 0.5 L cultures in 2 L baffled flasks at 37 °C for 24 h. 50 μg ml^{-1} kanamycin was included in the media. Cells were pelleted at 4,000g for 20 min, resuspended in 30-ml lysis buffer (20 mM Tris HCl pH 8, 150 mM NaCl, 30 mM imidazole, 10 mM lysozyme, 1 mM DNase1 mM phenylmethylsulfonyl

fluoride), and lysed using sonification. Lysate was clarified at 17,000 × *g* and the soluble fraction batch bound to 1 mL Ni-NTA resin (Qiagen) via the HIS tag at the N terminus of the protein over night while being shaken at 4° C. Lysate and resin were transferred to a gravity column and washed with 20 mL wash buffer (25 mM Tris HCl pH 8, 30 mM NaCl, 30 mM imidazole) before eluting the target protein with 12 mL elution buffer (25 mM Tris HCl pH 8, 35 mM NaCl, 250 mM imidazole). The eluate was concentrated in a 3,000 MWCO centrifugal filter (Amicon Ultra-15). We purchased BSA (Sigma Aldrich) and resuspended in MilliQ water.

5.5 Measuring protein concentration

A Nanodrop 8000 spectrometer (Thermo Scientific) was used to measure the absorbance of 280-nm-wavelength of 2 µL of protein sample. The concentration was determined from the measured absorbance and the calculated extinction coefficient following the Beer–Lambert law.

Size-exclusion chromatography: The concentrated proteins (Fig. 5.1) were fractionated by size with an AKTA pure chromatography system on a Superdex 6 GL column. A TBS running buffer (30 mM NaCl and 25 mM Tris pH 8) buffer was used.

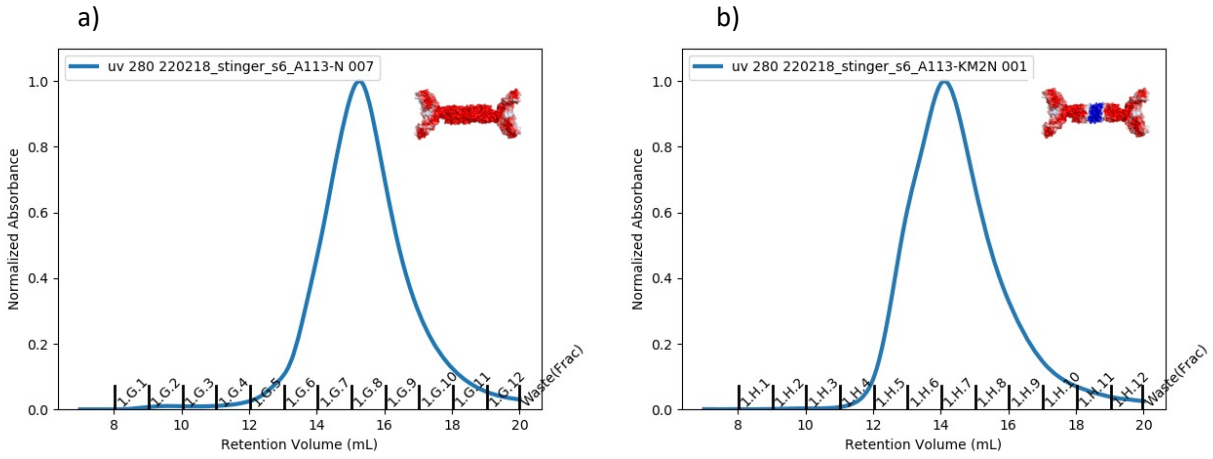

Figure 5.1: Size Exclusion Chromatography traces of modified proteins
a), b) size exclusion chromatography showing one clean peak at the expected retention volumes for a D3 oligomeric state.

5.6 Protein Dialysis

Protein was dialyzed in 3,500 molecular weight cut-off (MWCO) dialysis cassettes (Thermo) into 2 liters of MilliQ water for two hours, changing to fresh water after one hour.

5.7 TiBALDH reaction

Dialyzed Protein was then concentrated to 1 mg ml^{-1}. TiBALDH solution (20 mM) was prepared by diluting long titanium(IV) bis(ammonium lactate)dihydroxide (TiBALDH) (Sigma) with MilliQ water. The protein solution was quickly added to the TiBALDH solution to ensure that the TiBALDH is always in excess of the protein. Solution was then further mixed using a pipette and left to sit at room temperature. Aliquots were taken at the 15-minute time point and the one-hour time point and deposited onto a TEM grid.

5.8 TEM Grid Preparation

Lacey TEM Grids (Ted Pella) were used to analyze the reaction. 6 μL of reaction was incubated for one minute on the TEM grid and excess was wicked using filter paper. Grids were left to dry in a desiccator for one day prior to characterization by TEM.

5.9 TEM

Images were acquired on a Tecnai G2 F20 Supertwin TEM, in brightfield mode using 200kv acceleration voltage.

5.10 Ni Substrate prep

A 50 nm coating of nickel was deposited onto freshly cleaved muscovite mica (Ted Pella) using vapor deposition (performed at UW by a staff scientist at the WNF).

Attachment of proteins to nickel substrate: 50 µL of Dialyzed protein was deposited onto the nickel substrate. The nickel substrate was held at -2V potential using a Keithley 237 High Voltage Source-Measure Unit (SMU) and a grounding wire was placed near but not touching the surface so that the nickel atoms on the surface would be ionized and have a high affinity for the HIS tags on the protein (Fig. 5.2). After 1 minute of the applied voltage, the (SMU) went through a grounding cycle and the solution was exchanged from the protein containing solution to 10 mM TiBALDH. The substrate was inverted, and the reaction was allowed to continue for one hour. After one hour the substrate was washed three times with 1 mL of milliQ water which was pipetted onto the surface. The sample was then allowed to dry in a desiccator.

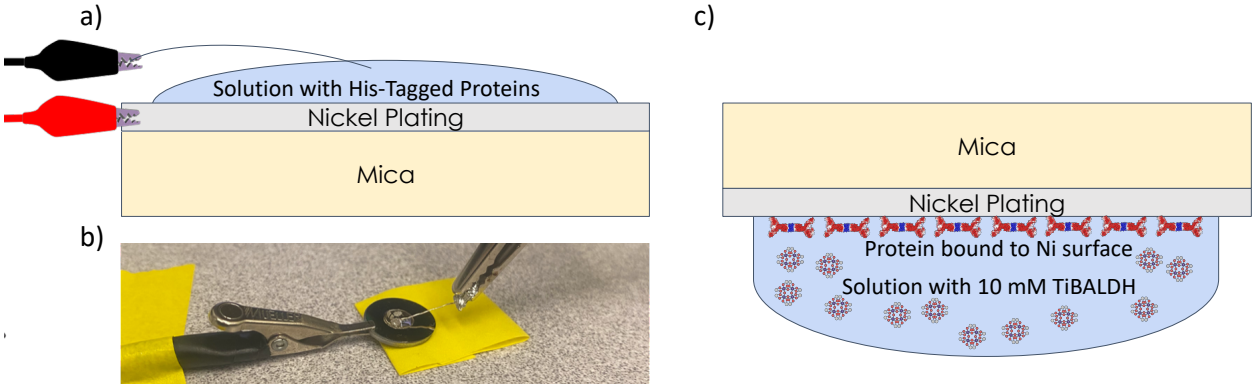

Figure 5.2: Experimental method for immobilizing a protein on a nickel substrate
a, b) cartoon and photo of the experimental set up where the nickel substrate is ionized and the proteins are attached via HIS tag affinity, c) cartoon showing the reaction set up where the substrate is inverted and the reaction proceeds in a manner that discourages precipitates to settle on the surface and bind non-specifically

5.11 *ex situ* AFM

The images were captured in air by a Bruker Icon atomic force microscope (Asylum Research) using tapping mode and RFESPA-75 (Bruker) probes using a setpoint of 148.2 with a target amplitude of 250 mV.

5.12 AFM Flow Cell

After protein was immobilized on the Ni substrate and excess was rinsed, flow experiments were performed using a Cypher AFM (Asylum) with liquid flow cell. A SLN C cantilever from Bruker was used to obtain data in the flow cell. Flow rate was 0.1 mL / min. Concentration of TiBALDH was 1 mM.

5.13 PiFM

PiFM data was collected on a Vista One instrument from Molecular Vista which combines AFM with IR spectroscopy. NCH-PtIr PiFM cantilevers from Molecular Vista were used. The wavenumber 1657 nm^{-1} is representative of the N-H bond in Lysine. AFM topography was overlaid with the absorbance signal at 1657 nm^{-1}.

5.14 Machine Learning

Employing a deep convolutional neural network to identify features is a proven way to automate feature recognition for video analysis[4]. Machine learning is robust to noise and once trained, a model can pick out features in a noisy environment[4-8]. AtomAI is a GitHub repository that was developed for analysis of scientific imaging[8]. The machine learning process was applied to 100 mM KCl data[4], but the focus of this research has been to apply machine learning to reduce noise in the 3 M KCl condition, adapt the rod recognition algorithms to function on videos of data, and provide analysis of the physical phenomenon of the rods organizing on the mica surface.

5.15 ML methods for segmentation

The paragraphs below describe the methods used to obtain a pixel by pixel classification into two categories: rods or background. This process reduces noise and enables computation analysis that would otherwise not be possible.

5.16 Preparing experimental files for analysis

Experimental data is exported from the AFM software as .ibw files. Using python, these files are converted into numpy[9] image arrays using the igor library[10] for python. The thresholds for these numpy arrays-were set to roughly 1 nm positive and negative on the z axis using the numpy clip function[9]. The images are then resized such that they cover a 200 nm physical space and are 256 pixels by 256 pixels.

5.17 Creating a training set

A selection of images from the dataset are set aside to use for training an ensemble of deep convolutional neural networks to recognize the features of interest. These training images are manually marked in Adobe Photoshop to delineate the location of each rod (Fig. 5.3). The marked data is then converted into an array containing only 0s and 1s, where each pixel is either part of a rod (1) or not part of a rod (0). Each training experimental data image is then paired with the corresponding marked image. Using AtomAI[8], a python package developed by ORNL, pairs of square patches are taken from the training set. These patches can then be augmented with additional noise or rotated if desired. It's best for rotation on AFM data to only flip horizontally or vertically to keep the horizontal aberrations which are intrinsic to AFM data in alignment.

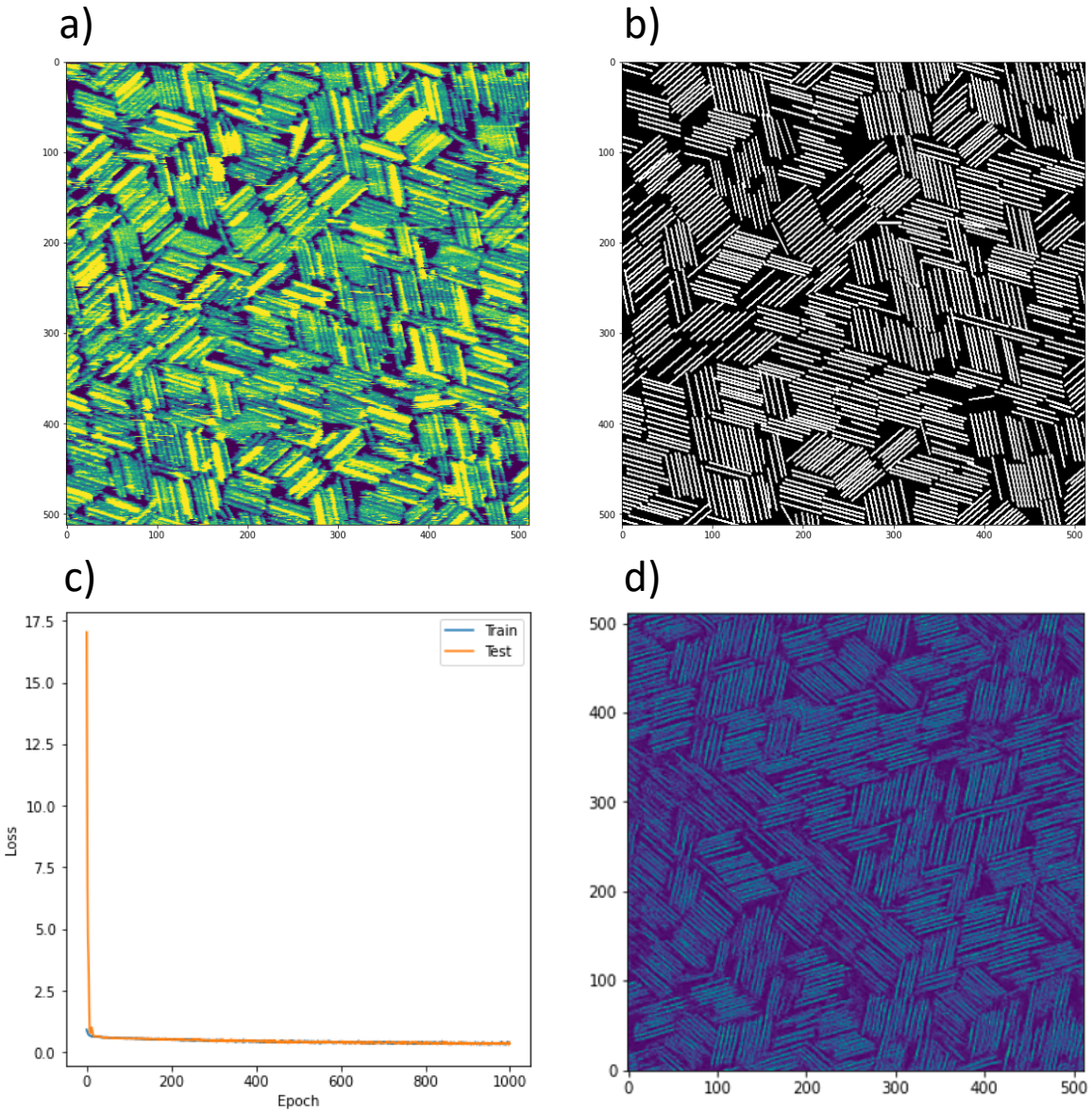

Figure 5.3: Pipeline for training an ensemble of deep convolutional neural networks to predict feature locations a) experimental AFM image acquired by Sakshi Yadav for analysis, b) ground truth of the single image to be used in training the ensemble of neural networks, c) the Loss for one set of neural networks, after 1000 training cycles which shows a plateau around 900 cycles and the agreement between the train and test performance indicates that the models are accurately trained to recognize the desired features , d) an average of all ensembles of the neural network output which shows very high accuracy and very low noise

5.18 Training the AI

After extracting and augmenting files from the training set, thousands of training images have been generated. These images are further split into a train and a test data set, with roughly 90 percent of

the images in the training set. The training set is used as an input into the ATOMAI ensemble trainer[5,6], which trains an ensemble of neural networks to recognize the features of interest. Each ensemble is set to contain 100 neural networks, and we train 10 ensembles. Using the marked data, the neural networks score the percentage of pixels that get correctly categorized into either features or background. A training cycle consists of the network using a set of weights for feature recognition, scoring the percent of correctly identified feature pixels, and then adjusting the weights for the next round of training. The loss and accuracy of each ensemble is reported every 100 training cycles for both the training and testing dataset. The testing dataset is only used for reporting and is never used to set weights for feature recognition. After 1000 training cycles, the accuracy of the neural networks no longer improves, and the ensemble of models is saved to use for feature recognition in the dataset. The loss for an exemplar training cycle is shown in Figure 4 and compares the test and train datasets- it is important that both test and train plateau because this indicates that the AI has learned how to recognize real features rather than simply memorizing the position of the ground truths from the training set. The loss of each ensemble after a stochastic weight averaging was an average of 7.864 percent with a standard deviation of 0.34 and the accuracy was 88.67 percent with a standard deviation of 0.48.

5.19 Feature Recognition

The experimental data which has been prepared as described above is then input to the ensemble of neural networks. The output is an image where noise is greatly reduced, shown in Figure 4d is a heatmap of the confidence of the ensemble that any given pixel is a part of a feature, or part of the background. This semantic segmentation has greatly improved the clarity of the image such that we can proceed to identifying individual features for analysis.

5.20 Algorithm for instance segmentation

The paragraphs below describe the process of obtaining the center of mass and angle for each rod, which classifies the features on the image into discrete rods.

5.21 Rod recognition

To identify the rods, we use a modified algorithm developed by Ben Legg at PNNL[11]. Providing an input of a reference rod, feature length, and the data image, we obtain the location of rod centers, and the angle of each rod segment. Feature lengths are measured manually and input as an integer. Using Adobe Photoshop, multiple features are averaged together to create an image of an ideal feature. This reference image contains both the feature, and the background space around the feature, with a transparent background. This reference rod is then applied as a sliding window across the entire experimental image, rotated, and then applied again to the entire image. This process is done for a full arc spanning 180°. To save in computational time, the method for applying a sliding window is FFT convolution, where the reference feature FFT is multiplied by the data image FFT. The output of the convolution is a heatmap where the peaks represent all possible placements for the rods. A score function based on the quality of alignment between the reference image and the data image stores the score of each rod placement. At this step there are roughly 3000 potential rod placements in the 200nm data image.

5.22 Data Management

A pandas dataframe[12] was used to store values throughout the analysis.

5.23 Maximal Separation

To narrow down rod placements to only contain the same rods that we recognize, we first sort all rod placements by the score of how well the rod reference image match. We then eliminate rods based on the following criteria: if two rods are on top of each other the better scoring rod is kept, if two rods intersect with each other the better scoring rod is kept, if two rods are next to each other and parallel but too close to represent a real feature the better scoring rod is kept. After this process of elimination, the only remaining identified rods are those we expect to see when looking at the data image with a human eye.

5.24 Clustering

To obtain locations of domains, a clustering algorithm is applied. We use sklearn's agglomerative clustering with the single linkage parameter[13]. This type of clustering was chosen because it does not require a number of domains to be input into the function but will group the rods based on the distance between rods. We input the x and y coordinates into the function along with the angle. The angle is plotted as a third dimension and imposes an extra distance between rods if they are not coaligned with each other. Each rod is assigned into a domain.

5.25 Plots over time

Number of rods over time, number of domains over time, and the average size of domains over time were taken by plotting the number of rods recognized, the count of individual domains, and the average of the count of rods within each unique domain, respectively.

Scatterplots, histograms, and heatmaps were plotted using Matplotlib[14].

5.26 Co-Assembly Rod Recognition

The rod recognition code was modified to include a second class of rods. Two reference images were generated using Adobe Photoshop containing the average of five rods for each length. The restoring window method was applied to the data image twice – once for each rod. All resultant rod matches were combined and sorted by score.

5.27 Co-Assembly Non-Maximal Separation

Previous iterations of the code found it sufficient to check if an arc extending from the end rod A intersected with rod B to rule out rods that are too close to each other. Once a smaller rod was introduced, it was possible for a smaller rod to nest in between the arcs and avoid detection by the algorithm. The code was modified to include a check if arcs extending from rod B overlap with rod A.

5.28 References

1. Courbet, A. *et al.* Computational design of mechanically coupled axle-rotor protein assemblies. *Science (1979)* **376**, 383–390 (2022).
2. The PyMOL Molecular Graphics System.
3. Kozlowski, L. P. IPC - Isoelectric Point Calculator. *Biol Direct* **11**, 1–16 (2016).
4. Oxley, M. P. *et al.* Probing atomic-scale symmetry breaking by rotationally invariant machine learning of multidimensional electron scattering. *npj Computational Materials 2021 7:1* **7**, 1–6 (2021).
5. Ziatdinov, M., Ghosh, A., Wong, C. Y. & Kalinin, S. v. AtomAI framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy. *Nature Machine Intelligence 2022 4:12* **4**, 1101–1112 (2022).
6. Ziatdinov, M. & Kalinin, S. AtomAI: Open-source software for applications of deep learning to microscopy data. *Microscopy and Microanalysis* **27**, 3000–3002 (2021).
7. Ziatdinov, M. *et al.* Quantifying the Dynamics of Protein Self-Organization Using Deep Learning Analysis of Atomic Force Microscopy Data. *Nano Lett* **21**, 158–165 (2021).
8. Ziatdinov, M. & Kalinin, S. AtomAI: Open-source software for applications of deep learning to microscopy data. *Microscopy and Microanalysis* **27**, 3000–3002 (2021).
9. Harris, C. R. *et al.* Array programming with NumPy. *Nature 2020 585:7825* **585**, 357–362 (2020).
10. igor · PyPI. https://pypi.org/project/igor/.
11. Zhang, S. *et al.* Rotational dynamics and transition mechanisms of surface-adsorbed proteins. *Proc Natl Acad Sci U S A* **119**, e2020242119 (2022).
12. Science, W. M.-P. of the 9th P. in & 2010, undefined. Data structures for statistical computing in python. *conference.scipy.org* (2010).
13. Hierarchical clustering: structured vs unstructured ward — scikit-learn 1.2.1 documentation. https://scikit-learn.org/stable/auto_examples/cluster/plot_ward_structured_vs_unstructured.html#sphx-glr-auto-examples-cluster-plot-ward-structured-vs-unstructured-py.
14. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**, 90–95 (2007).

6. Future Work

Future work in this field could explore several promising avenues, each offering unique opportunities for advancing our understanding and control of inorganic-mineral formation processes. One potential direction is to investigate the role of electrostatics in directing mineral formation on a broader scale. This could involve using in situ atomic force microscopy (AFM) to observe mineral growth on charged regions of a substrate in real-time. Such studies could provide valuable insights into the mechanisms of mineral nucleation and growth, as well as the influence of surface charge distribution on these processes. By examining the interplay between electrostatic forces and mineral formation at the nanoscale, researchers may uncover new principles for guiding the assembly of complex mineral structures.

Another area of interest is the development of more complex templating systems. This could include larger protein constructs, DNA-based templates, hybrid DNA-protein materials, or other patchy colloids. These more sophisticated templates could potentially allow for greater control over mineral formation and structure, enabling the creation of materials with specified properties and functionalities. For instance, DNA origami structures could be used to create intricate 3D templates for mineral growth or used as a scaffold to present proteins in specific arrangements. Protein-based templates could be engineered to incorporate specific binding sites or catalytic domains that influence mineral formation. Additionally, research could focus on sequentially arranging proteins onto a surface before initiating material formation, to organize nucleation sites and then directly grow crystals with the desired placements. This approach could involve developing new techniques for protein patterning and immobilization, as well as exploring the effects of protein orientation and spacing on mineral nucleation and growth. By carefully orchestrating the spatial arrangement of proteins on a surface, researchers may be able to guide the formation of complex mineral architectures with tailored material properties.

Further investigation into lattice matching could also prove fruitful, exploring questions such as the importance of directly matching the mineral surface in various solution conditions, the allowable margin of error in lattice matching, and the potential application of these principles to other material systems beyond those currently studied. This research could involve computational modeling to predict optimal lattice matches between proteins and target minerals, as well as experimental studies to validate and refine these predictions. By deepening our understanding of lattice matching principles, researchers may be able to design more generic protein templates for a wider range of mineral systems or be able to repurpose binding proteins from one system to another with minimal redesign.

Moreover, future work could explore the integration of multiple strategies, combining electrostatic interactions and lattice matching to achieve higher levels of control over mineral formation. This could be achieved by using proteins which have been designed to lattice match one mineral one side but have electrostatic patchiness on the other side. Another protein-oriented approach would be to design larger constructs which use lattice matching proteins in conjunction with other designed proteins that have a more three dimensional surface.

Another promising avenue for research is the investigation of dynamic templating systems that can adapt or respond to environmental stimuli during the mineral formation process. This could involve the use of stimuli-responsive polymers or switchable surface chemistries that allow for real-time adjustment of template properties. Such dynamic systems could enable the creation of minerals with gradient structures or composition-dependent properties, opening up new possibilities for material design.

While protein-inorganic interfaces hold great promise for material design, discrepancies between target and realized protein assemblies suggest that critical interactions remain unaccounted for in current

design frameworks. Therefore, another important future research area is to accurately incorporate these interactions into computational models and design software.

7. Acknowledgements

Using patches of charge to direct the nucleation of titanium dioxide in aqueous solution at STP: This work was produced in collaboration with Mingyi Zhang under the direction of David Baker, Shuai Zhang, and Jim De Yoreo

Machine learning-driven descriptions of protein dynamics at solid-liquid interfaces: This work was produced in collaboration with Shuai Zhang, and Jim De Yoreo

Symmetry Breaking Drives Protein Assembly into a Forbidden Two-Dimensional Liquid-Crystal Phase: This work was produced in collaboration with Sakshi Yadav, and Benjamin Helfrecht, under the direction of Christopher Mundy, Maxim Ziatdinov, Benjamin Legg, Shuai Zhang, and Jim De Yoreo

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (DGE-1762114 and DGE-2140004). This material is based in part upon work supported by the state of Washington through the University of Washington Clean Energy Institute under a fellowship award.

This work was supported by the Department of Energy (DOE), Office of Basic Energy Sciences, Energy Frontier Research Center- The Center for the Science of Synthesis Across Scales located at the University of Washington (UW), award number DE-SC0019288, and through Pacific Northwest National Laboratory (PNNL), FWP 72448. The design and synthesis of DHR10-micaX-H and its variants were performed at UW and supported by the US DOE BES Biomolecular Materials Program, award number DE-SC0018940. PNNL is a multi-program national laboratory operated for DOE by

Battelle under Contract No. DE-AC05-76RL01830. The MC simulation modal was originally developed at Molecular Foundry, Lawrence Berkeley National Laboratory, through a user proposal.
Work on analysis of protein rotation dynamics was supported by the DOE BES Division of Materials Science and Engineering at PNNL under award FWP-77246. PNNL is a multiprogram national laboratory operated for DOE by Battelle under contract number DE-AC05–76RL01830. Part of this work was conducted at Molecular Analysis Facility, a National Nanotechnology Coordinated Infrastructure (NNCI) site at the UW, which is supported in part by funds from the National Science Foundation (awards NNCI-2025489, NNCI-1542101), the Molecular Engineering & Sciences Institute, and the Clean Energy Institute.

I'd like to thank the Molecular Analysis Facility and staff scientists for support and training in characterization methods. In particular I'd like to thank Ellen Lavoi for her patience and mental support with everything from temperamental days at the TEM to frustrations with grad school and life as a whole.

Moreover, I would like to thank the whole CSSAS center for providing such a wonderful environment for conducting this research, with a wealth of scientific discussions and friendship, especially Fátima Dávila Hernández, Nada Naser, Helen Larson, and Abdul Moeez.

It's been an honor and privilege to be a part of the De Yoreo research group both because Jim has been fantastic in his support as an advisor but also because of the people who make up the group. I'd like to thank Ben Legg and Shuai Zhang for their mentorship in science and I feel lucky to have had the chance to work with them. Working together with Sakshi Yadav Schmid was one of the most enjoyable experiences from my PhD. I'd also like to thank Lili Liu, Yuna Bae, Alex Bard, Emily Saccuzzo, Chenyang Shi, Mingyi Zhang, and Biao Jin for research discussions and friendship. There are several

fellow students in the lab, Ying Xia, Jayesh Dua, Kyle Kluherz, Madison Monahan, Brenda Kessenich and Guomin Zhu who's commiseration and camaraderie have been invaluable.

I'm also grateful to my grad school friends Erich Pederson, Laura Carlucci, Molly Molica, Casey Kiyohara, Marike Reimer, Gokce Altin, and Yuhuan Meng. Meeting on campus or playing board games together has been invaluable to me.

Finally, my family has provided me with love, encouragement, and tolerance. My mom Sandi Brehm has been available to help whenever an experiment ran long and I needed someone to pick up my son Erich. My brother David Brehm is always available for last minute proof reading. My husband Ulrich Stegmann has been the best partner throughout that I could wish for. I'd like to posthumously thank my dad George Brehm for his belief in me.