Resolving non-equilibrium shape variations amongst millions of gold nanoparticles

Zhou Shen,[†,‡,¶] Salah Awel,[§] Anton Barty,[§] Richard Bean,[‖] Johan Bielecki,[‖] Martin Bergemann,[‖] Benedikt J. Daurer,[⊥,#] Tomas Ekeberg,[@] Armando D. Estillore,[§] Hans Fangohr,[‖] Klaus Giewekemeyer,[‖] Mark S. Hunter,[△] Mikhail Karnevskiy,[‖] Richard A. Kirian,[∇] Henry Kirkwood,[‖] Yoonhee Kim,[‖] Jayanath Koliyadu,[‖] Holger Lange,[††,‡‡] Romain Letrun,[‖] Jannik Lübke,[††,§,¶¶] Abhishek Mall,[‡,¶] Thomas Michelat,[‖] Andrew J. Morgan,[§§] Nils Roth,[§,¶¶] Amit K. Samanta,[††,§] Tokushi Sato,[‖] Marcin Sikorski,[‖] Florian Schulz,[‡‡] Patrik Vagovic,[§,‖] Tamme Wollweber,[‡,¶,††] Lena Worbs,[§,¶¶] Paul Lourdu Xavier,[‡,††,§] Filipe R. N. C. Maia,[@,‖‖] Daniel A. Horke,[††,§,⊥⊥] Jochen Küpper,[††,§,¶¶,##] Adrian P. Mancuso,[‖,@@,#] Henry N. Chapman,[††,§,¶¶] Kartik Ayyer,[‡,¶,††] and N. Duane Loh[*,⊥,†]

†*Department of Physics, National University of Singapore, Singapore 117551, Singapore*

‡*Max Planck Institute for the Structure and Dynamics of Matter, 22761 Hamburg, Germany*

¶*Center for Free-Electron Laser Science, 22761 Hamburg, Germany*

§*Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron DESY, 22607 Hamburg, Germany*

‖*European XFEL, 22869 Schenefeld, Germany*

⊥*Center for BioImaging Sciences, National University of Singapore, Singapore 117557, Singapore*

#*Diamond Light Source, Harwell Campus, Didcot, OX11 0DE, UK*

@*Department of Cell and Molecular Biology, Uppsala University, 75124 Uppsala, Sweden*

△*Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA*

Contents

Abstract

Nanoparticles, exhibiting functionally relevant structural heterogeneity, are at the forefront of cutting-edge research. Now, high-throughput single-particle imaging (SPI) with x-ray free-electron lasers (XFELs) creates unprecedented opportunities for recovering the shape distributions of millions of particles that exhibit functionally relevant structural heterogeneity. To realize this potential, three challenges have to be overcome: (1) simultaneous parametrization of structural variability in real and reciprocal spaces; (2) efficiently inferring the latent parameters of each SPI measurement; (3) scaling up comparisons between 10^5 structural models and 10^6 XFEL-SPI measurements. Here, we describe how we overcame these three challenges to resolve the non-equilibrium shape distributions within millions of gold nanoparticles imaged at the European XFEL. These shape distributions allowed us to quantify the degree of asymmetry in these particles, discover a relatively stable 'shape envelope' amongst nanoparticles, discern finite-size effects related to shape-controlling surfactants, and extrapolate nanoparticles' shapes to their idealized thermodynamic limit. Ultimately, these demonstrations show that XFEL SPI can help transform nanoparticle shape characterization from anecdotally interesting to statistically meaningful.

Keywords

XFEL, Gold Nanoparticle, Monte Carlo, Structural heterogeneity, High-throughput single-particle imaging

Colloidal, solid-state nanoparticles have properties defined by their size and shape, making them attractive for applications ranging from the broad field of photonics and electronics to catalysis.[1–3] In the case of catalysis, for example, the nanoparticle's catalytic activity strongly depends on its size and its exposed facets, which have a strong correlation with the shape.[4,5] Hence, understanding and controlling nanoparticles' structural variations is an important aspect of synthesis.[6] Commonly used post-synthesis characterization techniques (like UV-VIS,[7] small angle x-ray scattering (SAXS)[8]), however, mostly measure the mean of and standard deviation of the size of nanoparticles.

To directly resolve shape variations among nanoparticles, however, requires imaging many nanoparticles individually, for example, using scanning or transmission electron microscopy (SEM or TEM).[9,10] Tomography is sometimes used, but is time-consuming and hence limited to a few nanoparticles.[11] Nevertheless, electron microscopy-based characterization typically numbers in the hundreds (e.g., 300-500 particles[10]). Furthermore, for larger nanoparticles, multiple scattering limits such three-dimensional (3D) shape characterization. As such, shape characterization by SEM remains largely two-dimensional (2D). Furthermore, characterization by SEM and TEM suffer from orientation bias[12] since the nanoparticles are arrested on substrates for imaging.

In contrast, high-throughput single-particle imaging with intense, ultrafast, x-ray free electron lasers (XFELs)[13] can fundamentally transform how we characterize nanoparticles. Single particle imaging (SPI) at the European XFEL can interrogate millions of nanoparticles in a few hours.[14] Compared to electron microscopy, XFEL SPI is less limited by multiple scattering. Hence, XFEL diffraction patterns of single nanoparticles closely correspond to Ewald sphere sections of the particles' Fourier volume, which in turn allows the matching of 3D structure to single-particle 2D diffraction patterns.[15] Furthermore, nanoparticles are injected at different random orientations into the XFEL interaction region, thus avoiding the orientation bias when imaging substrate-bound nanoparticles.

Resolving the 3D shape variations among a million nanoparticles can uncover statistically meaningful insights about nanoparticles' non-equilibrium synthesis pathways. Lurking within this opportunity, however, is a formidable statistical learning challenge: to infer the hidden parameters of the measurement of each nanoparticle, such as orientation, incident photon fluence, structural class, complex phases missing from the diffraction intensities. This problem is typically tackled by two types of approaches.

The first approach induces a family of statistically likely 3D structures *d*e novo from large numbers of SPI patterns. Each measurement's hidden parameters are iteratively co-refined together with these induced 3D structures. This approach uses only prior knowledge from basic scattering physics (e.g., weak phase approximation, shot-noise limited images, etc). Some examples in this class extend the expand-maximize-compress algorithm (EMC),[16–19] to multiple structural models.[14,20] Notably, this approach recovers an over-sampled 3D diffraction volume of each 3D structure from which its corresponding real-space electron density map is recovered using computational phase retrieval.[21] However, the number of candidate 3D structures recoverable is limited ($\lesssim$ 100) by the computational memory needed to store them.[22]

The second approach to learning each pattern's hidden parameters uses diffraction template matching, which draws heavily on structural prior knowledge about the samples. Template diffraction patterns, typically created from a pool of idealized models, are used to match and classify experimentally measured SPI patterns. This approach does not generally require phase retrieval because each template is associated with a particular real-space model. Template-matching approaches were used to study variations among XFEL pulses[23,24] and recover the histogram of sizes in $> 10,000$ organelles by assuming their protein shells are spheroids.[25] Atsushi Tokuhisa et al. proposed a template matching method for biomolecules[15] using diffraction templates generated from 3D structures in molecular dynamics simulations. However, just like the first approach, the space of possible conformations if non-parametric is again limited by memory and compute requirements.

Here, we show how particles' shape variations (beyond the mere radius of gyration) can be simultaneously and efficiently parameterized in both real and reciprocal spaces. This simultaneous parametrization allows us to efficiently infer the latent parameters (including complex phases) of the individual SPI patterns given a pool of 3D structures. More importantly, this parametrization allows a principled and efficient approach to proposing and evaluating upwards of 10^5 candidate 3D structures *de novo*.

The recovered distribution of shapes (and sizes) of the millions of gold nanoparticles (two ensembles with edge lengths of approximately 30 nm and 40 nm, respectively[14]) is telling. We quantified the degree of asymmetry in each nanoparticle from the distribution of their (111) and (100) facet areas. We also discovered a relatively stable 'shape envelope' in two different ensembles of nanoparticles. Since both ensembles were extracted at different times in a common crystal growth trajectory, we could extrapolate their particle shapes to large crystals in the thermodynamic limit. Furthermore, we found hints of finite-size effects related to the surfactant used to control the nanoparticles' shape. These studies demonstrate the potential of XFEL for studying nonequilibrium systems that are difficult to image directly by conventional means or too heavy for molecular dynamics simulation.

Results and Discussion

Synthesis and measurements of nanoparticle ensembles.

Two ensembles of truncated octahedral gold nanoparticles were synthesized using the protocols described elsewhere.[5,26] We used a solution comprising $HAuCl_4$ as the precursor and poly (diallyldimethylammonium) chloride (PDDA) as a surfactant. This mixture was introduced into a round-bottom flask containing 1,5-pentanediol (PD) solution and refluxed in an oil bath at a temperature maintained at up to 225 °C. Two ensembles of nanoparticles were created by quenching this mixture in a room-temperature water bath after approximately 4 min (for sample oct30) or 7.5 min (for sample oct40) of reaction time. Following quenching, the resulting crude nanoparticle mixture underwent thorough purification to eliminate excess ligands. This purification involved sequential centrifugation steps in acetone and water. The resulting nanoparticle pellet was re-suspended in ultra-pure water, which is used for subsequent XFEL imaging. Using scanning electron microscopy (SEM) (fig. 4) on small batches of the nanoparticles from these two ensembles, we determined their nominal average widths as 30 nm (oct30) and 40 nm (oct40).

The samples were injected by an electrospray injector and focussed with an aerodynamic lens stack into the stream of XFEL pulses at the European XFEL (EuXFEL), as described in Ayyer *et al.*[14] Due to the high pulse repetition rates of the EuXFEL, approximately 105 and 65 diffractions of single particles were accumulated per second to comprise the oct30 and oct40 datasets respectively. Millions of such XFEL measurements were used to reconstruct two average 3D structures, each representing either the oct30 or oct40 ensembles.[14] The widths of these two average 3D structures (for oct30 and oct40 respectively) were 35 nm and 40 nm; the longest edge lengths of their (111) facets were 20 nm and 27 nm.

Two-dimensional (2D) *in-silico* classification[14] filtered out empty shots, multiple-particle shots, and diffraction patterns likely belonging to non-octahedral nanoparticles in both oct30 and oct40 ensembles. After post filtration, 1 287 570 and 823 202 patterns remained in the oct30 and oct40 datasets respectively.

Parametrizing structural variations

Earlier analyses of the large oct30 and oct40 datasets showed[14] noticeable structural variations amongst truncated octahedral nanoparticles, which led to the averaged 3-dimensional models showing 'rounded' (100) facets. Our goal here is to characterize these variations in a statistically robust and meaningful way.

The space of nanoparticle structural variations, which is resolvable by our experiment, lives in a 10^5-dimensional space (see Methods Section Degrees of freedom in nanoparticles' structure). However, we seek only the *posterior distribution of their first-order distortions from the average truncated octahedron.* Such distortions can be efficiently parameterized with a simpler 42-dimensional *free-facet truncated octahedron (FFTO) model*, which consists of the vertex positions of 14 facets of a truncated octahedron (fig. 2(b)).

We used a weighted Monte Carlo importance sampling scheme to sample the oct30 and oct40 ensembles' posterior distribution in the 42-dimensional FFTO space. We then parameterized the dominant structural variations within these posterior distributions, which allowed us to infer the nanoparticles' synthesis conditions directly.

Posterior estimation using Monte Carlo importance sampling

Exhaustively resolving the posterior probability of a nanoparticle's structure in a 42-dimensional FFTO is computationally prohibitive. In a naive approach, this would involve comparing each diffraction pattern against a large number of possible 3-dimensional models that densely cover this 42-dimensional space. Each comparison, in turn, requires checking the most likely orientation in which each pattern could arise within each model. Instead, we know these structures stay close to a truncated octahedron,[14] making the vast majority of these models in the FFTO space unlikely or, equivalently, *unimportant* in this analysis. Hence, a much smaller and non-uniformly spaced pool of FFTO models can capture the most important nanoparticle structures.

To estimate the posterior distribution of likely structures, we seek a pool of FFTO models, $\mathbb{M} = \{\rho_1, \rho_2, \dots\}$, that efficiently sample the posterior space (see fig. 1(a)). To paraphrase, this pool should encompass the set of FFTO models that are most likely to produce the experimentally measured oct30 and oct40 diffraction patterns, $\mathbb{K} = \{K_1, K_2, \dots\}$.

We use a weighted Monte Carlo (MC) importance sampling scheme to efficiently accumulate this pool of models (fig. 1(b)). This MC model pool starts with an initial model that is randomly perturbed from the average 3D structure, which is approximated from the single-model reconstruction result of the whole dataset. New models in the FFTO space are iteratively added to this pool in three steps: (1) select a weighted random model from the existing model pool; (2) perturb this selected model; (3) add the perturbed model to the pool and update all the models' weights. For this MC scheme to sufficiently sample the FFTO space, it needs to explore the space of less likely models. We do this by penalizing excessive selections of the most likely models in the pool. Hence, the weights used to select random models in the first step depend on the ratio between the following two quantities: the percentage of diffraction patterns that are likely due to each model in the pool as defined by eq. (5) and shown as numerators in fig. 1(b); and the number of times each model was selected for perturbation in step 2, which starts at 1 for each added model, and shown as denominators in fig. 1(b). The numerator ensures that we explore the neighbourhood around likely models, while the denominator favors selecting less frequently visited models.

With this pool of models, we can evaluate the posterior probability of various nanoparticle features ν (e.g., length, shape, volume, asymmetry, etc) given the diffraction measurements of the oct30 and oct40 ensembles ($\mathbb{K}$). This probability is similar to a weighted voting scheme: each model in the pool (ρ_i) casts a vote for a particular feature, and this vote is weighted by that model's posterior probability given all measurements. This leads to the posterior estimates in eqs. (1) and (4), which we derive in the Methods section.

We demonstrate this framework on an artificial ensemble of flexible particles. Each particle consists of four identical balls that are sequentially attached (fig. 1(c)). Each particle's structure is described by three angles defined in their body axes (bond angles α, and β; dihedral angle γ). All possible particle structures are confined to a ground truth linear trajectory (black line in fig. 1(c)). Since the four balls in each particle are identical, swapping the first and last balls, which also swaps α and β, yields identical diffraction patterns. This leads to a duplicate of the ground truth trajectory in (α, β, and γ) space. From 100 000 diffraction patterns of randomly rotated particles with random structures along this ground truth trajectory, we correctly reconstructed the posterior distribution of structures shown in red in fig. 1(d). Details about this artificial ensemble are discussed in the Methods section.

Validating our estimated posterior distribution is important, especially when the raw data is sparse and incomplete. Since we did not have the ground truth posterior for the oct30 and oct40 datasets for validation, we checked that our estimated posterior has converged and is self-consistent. Briefly, we used the reconstructed model pools as a 'proxy' to the ground truth to generate random test diffraction patterns. These generated test patterns were then used to accumulate a second pool of models, which were compared to the ground truth 'proxy' for repeatability. Details are described in Methods.

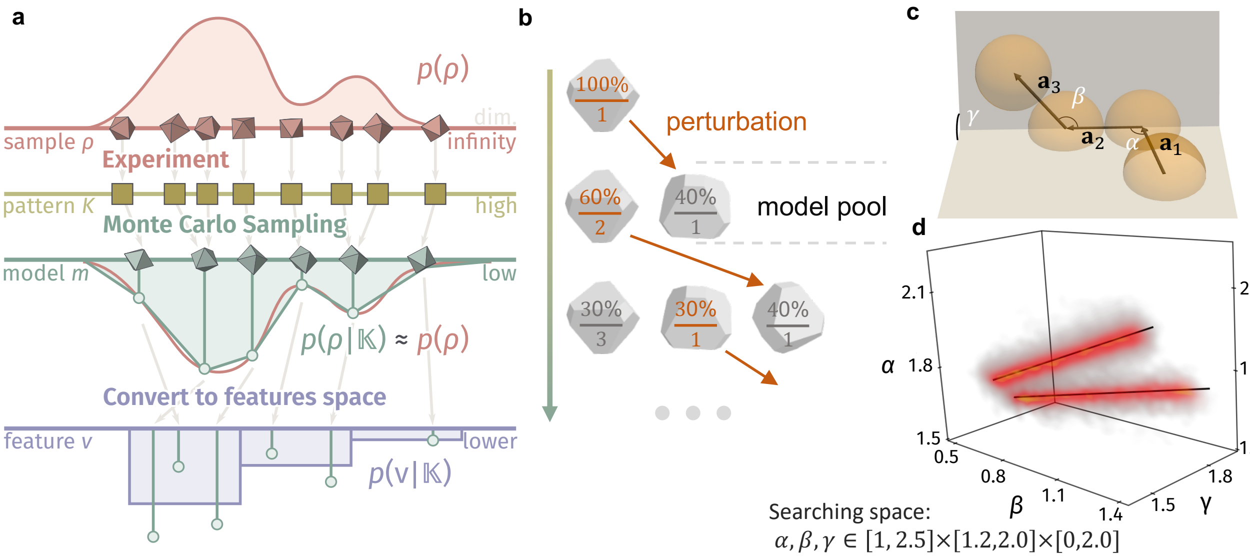

Figure 1: (a) Framework to estimate the structural posterior distribution of particles from their experimental measurements (i.e., diffraction patterns). (b) The weighted Monte Carlo importance sampling scheme, which includes model selection (red numbers), perturbation (red arrows), and weights updating. (c) Rotational degrees of freedom (α, β, γ) in our artificial ensemble of nanoparticles, each as a 4-ball chain. (d) Posterior distribution of ensemble in (c) using our Monte Carlo scheme in (a). The ground truth structural trajectory is shown with the twin black lines from which diffraction patterns are randomly generated. The pool of Monte Carlo models is rendered in a semi-transparent point cloud, where higher red intensity indicates models with higher data likelihood given the diffraction patterns.

Dominant structural modes in nanoparticle ensembles

Our weighted Monte Carlo importance sampling of the nanoparticles' posterior distribution yielded a pool of FFTO models representing the most probable nanoparticle structures in our oct30 and oct40 diffraction datasets. However, each FFTO model is described by a 42-element vector, which *still* has far too many dimensions for us to visualize.

Fortunately, these 42 numbers are not mutually independent, as they can describe the same nanoparticle structure but at a different orientation and/or translation. Hence, dimensionality reduction should be possible. To accomplish this, we mapped each FFTO model into a *facet-area representation*, which consists of an ordered list of the areas of each model's 14 facets. In this representation, the areas of the (100) direction facets are indexed from 0 to 5, and those of the (111) direction facets are indexed from 6 to 13 (fig. 2(a)). This set of facet-areas is not only invariant under rotations and translations but is also conveniently related to each model's surface free energy.[6]

To simplify our analysis of the estimated posterior, we 'hard-assigned' each diffraction pattern K only to its most probable model in the Monte Carlo FFTO model pool. Each time an FFTO model is deemed most likely for a pattern, we projected this FFTO model into its 14-dimensional facet-area feature space and then appended this facet-area model to a growing list. Since an FFTO model might be deemed most likely by multiple diffraction patterns, this model's facet-area features might appear multiple times within this list. For brevity, we will refer to this list of features as the *facet-area point cloud*, or equivalently, the $|\mathbb{K}| \times 14$ matrix X.

Due to the octahedral symmetry of our models, the order of these 14 numbers in a facet-area feature can be changed by applying any rotation operation within this symmetry group. This redundancy is eliminated (details in Methods section) since we are not interested in orientational differences amongst nanoparticles.

The primary structural variations manifest in this facet-area point cloud (X_{oct30} or X_{oct40}), which has been reduced in symmetry, are examined using principal components

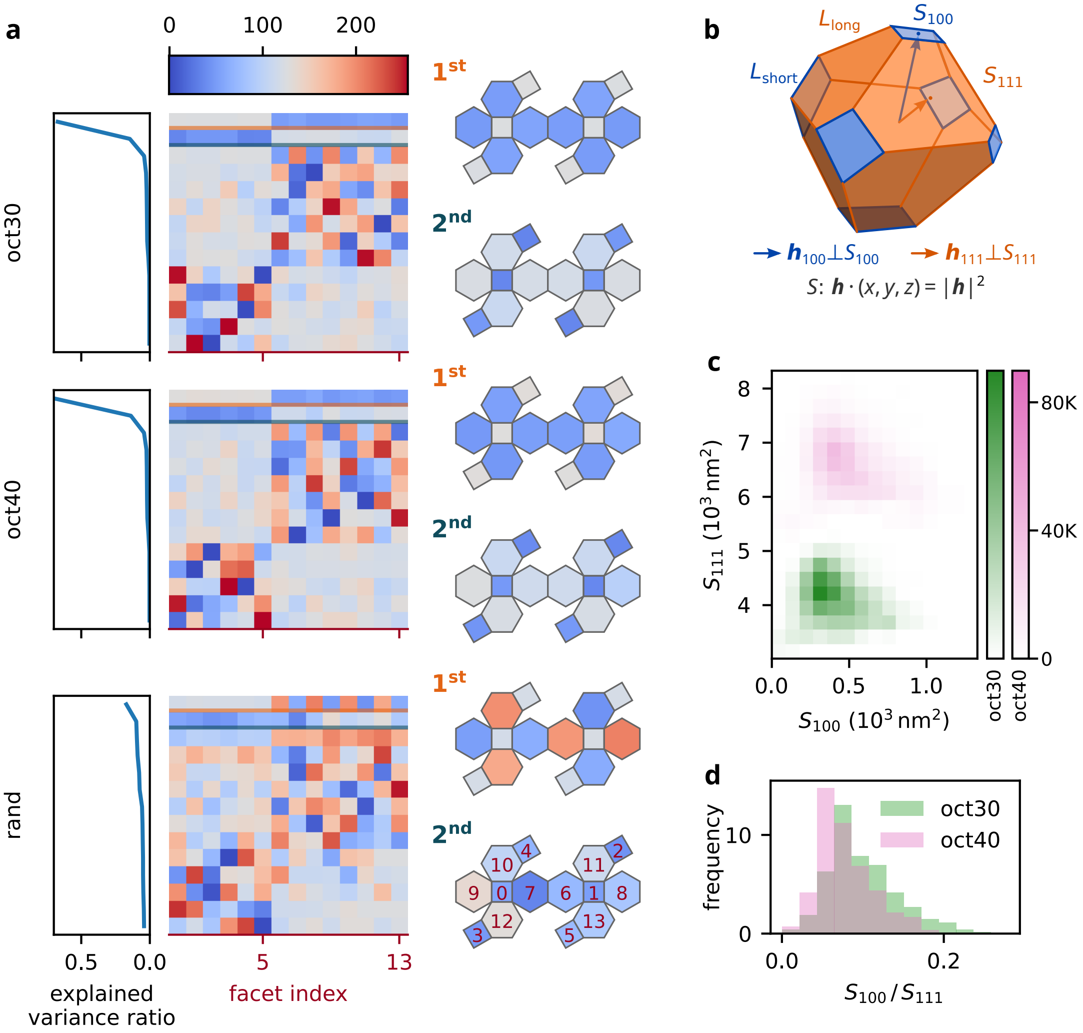

Figure 2: Quantifying the dominant modes of variation in nanoparticle ensembles (a) Principal Component Analysis (PCA) was applied to the *facet-area* features of three nanoparticle ensembles: oct30, oct40, and a randomly generated synthetic one. The PCA modes of these features are shown as rows of the matrices (middle block, linear color scale from blues to reds as negative to positive); these modes (i.e., rows) are sorted by their explained variance ratio (left block). In the right block, we see the corresponding net plots of the two most dominant PCA modes of each ensemble, where the facets are colored according to the modal variations. The facet indices of the PCA mode columns are laid out on the bottom net plot. (b) The coordinate system for our FFTO models, where the total areas of the (111) and (100) facets, denoted as S_{111} and S_{100}, are shown in orange and blue respectively. (c) The facet-area features of oct30 and oct40 are projected onto the S_{111}-S_{100} subspace. (d) The distribution of the S_{111}/S_{100} area ratio for oct30 and oct40 demonstrates that the former experiences a more significant truncation along the (100) directions.

analysis (PCA) (fig. 2(a)). To do this, we decomposed X into 14 modes, sorted by their explained variances. Modes with higher explained variance describe more frequent structural variations. We color the modes of these facet-area variations in fig. 2(a) for both X_{oct30} and X_{oct40}. For comparison, we include an ensemble of randomly perturbed truncated octahedra X_{rand} with 10 000 points. Each point was perturbed from an average canonical FFTO model whose facets are aligned perfectly along the (111) or (100) directions. The first two PCA modes of X_{oct30}, X_{oct40}, and X_{rand} are colored in the same manner in their accompanying polyhedral net plot.

These dominant facet-area PCA variations reveal that the surface energy densities of the nanoparticles' (111) and (100) facets are distinct. More than 80% of the facet-area variations of the *millions of nanoparticles* in X_{oct30} and X_{oct40} can be explained by their respective first two PCA modes. The most dominant PCA mode shows that the (111) facet-areas tend to be correlated, while the next mode shows similar correlations amongst the (100) facet-areas. This correlation is notably absent in X_{rand}, where no constraints were imposed on the ratios amongst the surface energy densities of different facets. The correlations within the first two modes can be explained by the fact that the free energy of a nanoparticle includes the terms $\gamma_{111}S_{111}$ and $\gamma_{100}S_{100}$, where γ and S denote the surface energy densities and total areas of the subscripted facets. Our observed correlations are hence consistent with expectation that γ_{111} and γ_{100} are different for these octahedral nanoparticles.[5,26]

Relatedly, the variations in the (111) facet areas are approximately four times higher than those of (100) facets. This indicates that much of the changes in the surface area of oct30 and oct40 nanoparticles still lie on their (111) facets.

The third-ranked dominant PCA modes of X_{oct30} and X_{oct40} in fig. 2(a) are similar, but likely due to random fluctuations since they resemble the top-ranked mode for X_{rand}. The alternating signature of this mode is largely due to eliminating symmetries in these features, which was also performed on X_{rand}.

These observations quantify the degree to which each nanoparticle's structural variations are highly correlated to the areas of their (111) and (100) facets. Furthermore, since the areas of the (111) and (100) facets are separately correlated, these variations can be further reduced to just the two-dimensional space of S_{111} vs S_{100} (the sum of the (111) and (100) facet-areas respectively). We project the posterior distributions for X_{oct30} and X_{oct40} into the S_{111}-S_{100} subspace in fig. 2(c).

Finally, the S_{100}/S_{111} ratio of each FFTO model is proportional to the extent of truncation along the (100) octahedral facet, where smaller ratios indicate less truncation. By projecting the posterior distribution into the S_{100}/S_{111} subspace in fig. 2(d), we see that the oct40 is less truncated than the oct30 ensemble. In the size range of our experiment (30 nm to 50 nm), smaller particles exhibit a tendency towards being more spherical. A similar behavior was observed in decahedral multiply twinned gold NPs.[27]

Evidence of non-equilibrium growth from posterior distributions

The PCA of the posterior distributions in fig. 2 show that the first-order structural variations in either oct30 or oct40 can be further reduced to features associated with either each nanoparticle's (111) facets or those with their (100) facets. Here are two possible feature pairs that can be physically interpreted. The first pair we chose is (h_{100}, h_{111}): the average distances of its (100) and (111) facets from each nanoparticle's origin respectively. These distances are key parameters in the Wulff construction used to describe the equilibrium shapes of crystals. The second pair of features is $(L_{\mathrm{short}}, L_{\mathrm{long}})$, which are the average lengths of two types of edges: twenty-four shorter edges of (100) facets (blue edges in fig. 2(b)), and the remaining twelve longer edges (orange edges in fig. 2(b)) respectively.

We can gain valuable insights into the overall growth trajectory of both nanoparticle ensembles by extrapolating from and interpolating between the (h_{100}, h_{111}) features of the oct30 and oct40 ensembles (fig. 3(b)). According to the Gibbs-Wulff theorem, when a constant volume crystal attains its equilibrium shape, the ratio $R = h_{100}/h_{111}$ equals the ratio between the surface tensions of its (100) and (111) facets, denoted as $\gamma_{100}/\gamma_{111}$. In our specific case,

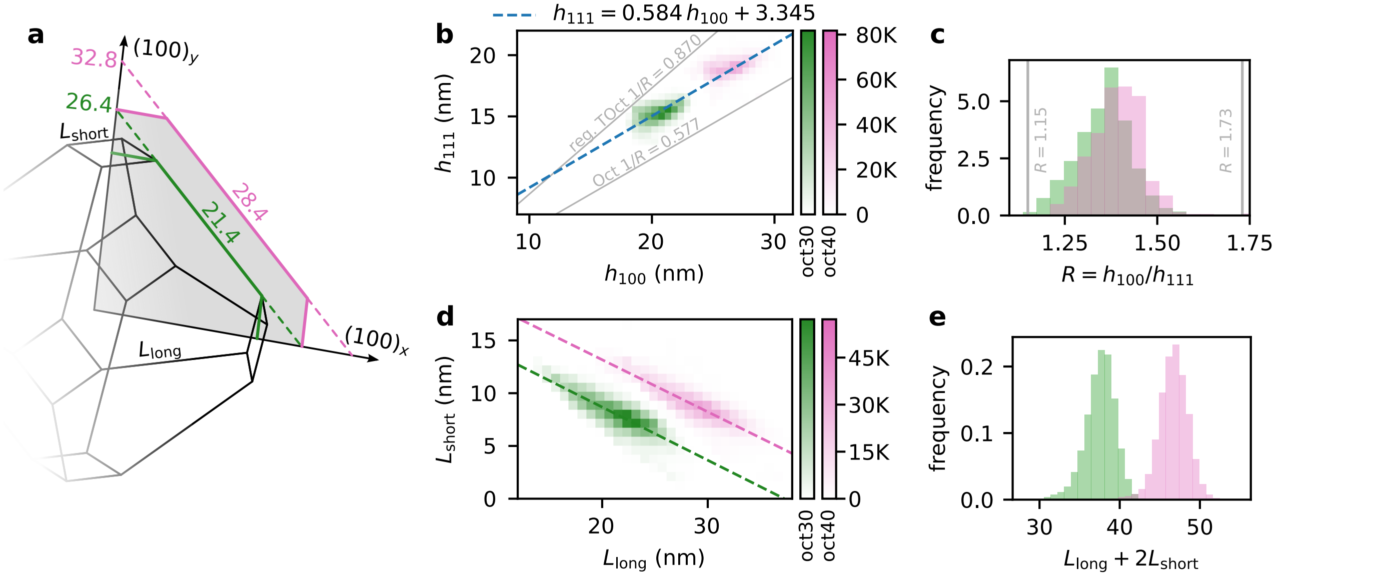

Figure 3: Signs of non-equilibrium growth in truncated octahedra. (a) A truncated octahedron (black edges) shown with the oct30 and oct40 octahedral envelopes (green and pink dashed lines). The distances of the h_{100} vertices in the oct30 and oct40 envelopes are 26.4 nm and 32.8 nm respectively. The average longest edge lengths of the oct30 and oct40 (111) facets are 21.4 nm and 28.4 nm respectively. (b) The posterior distributions of h_{111} vs h_{100} for oct30 and oct40 denote the average distances between the origin and the (111) and (100) facets respectively. Gray lines show the (h_{111}, h_{100}) relationship for an equilibrium regular octahedron and a regular truncated octahedron. Blue dotted line interpolates between the oct30 and oct40 posterior distributions (fit function as plot title). (c) Frequency histogram of $R = h_{111}/h_{100}$ projected from (b), annotated with the ratios of a regular truncated octahedron (0.577) and an octahedron (0.870). (d) The distributions of the average lengths of the shorter and longer edges of the truncated octahedra in oct30 and oct40 (L_{long} vs L_{short}), each fitted to a line. Both lines fit for -0.5 slope, indicating most nanoparticles are constrained to an octahedral envelope of edge length (i.e., $L_{\mathrm{long}} + 2L_{\mathrm{short}}$) of 26.4 nm (oct30), or 32.8 nm (oct40). (e) The frequency-distribution of the envelopes' edge length, $L_{\mathrm{long}} + 2L_{\mathrm{short}}$, show $\leq$3 nm FWHM deviation in both oct30 and oct40.

density functional theory[28] predicts that $R_0 = 1.27$. Additionally, the ideal (untruncated) octahedron and regular truncated octahedron exhibit R values of $\sqrt{3}$ and $\sqrt{4/3}$, respectively.

In fig. 3(b), the posterior distributions of oct30 and oct40, when projected to the (h_{100}, h_{111}) subspace, fit the dashed blue line given by $h_{111} = 0.584h_{100} + 3.345$. Since the only difference between the synthesis of the oct30 and oct40 ensembles is their reaction times,[14,26] we assume that these two ensembles are two "snapshots" of the same crystal growth trajectory connected by this fitted blue line.

If we extrapolate this fitted growth trajectory forward in time, assuming the nanocrystals could grow towards their thermodynamic limit (i.e., $h_{100} \to \infty$)[30] it would approach the facet displacement ratio of $R = h_{100}/h_{111} \to 1.71$. This ratio is smaller than $R = \sqrt{3} \approx 1.73$ of a regular (untruncated) octahedron. This trend suggests that larger crystals beyond those in oct40 will always exhibit some truncation on their (100) facets. This is consistent with the larger octahedra synthesized by Lu *et al.* that show 'rounded' (100) facets.[26] Extrapolating this growth trajectory backward in time, it intersects the regular truncated octahedra ratio of $R = \sqrt{\frac{4}{3}} \approx 1.15$ when h_{100} is around 11 nm, where the particles are most spherically symmetric.

Both nanoparticle ensembles in fig. 3(c) deviate significantly from the reference R_0. The oct40 ensemble ($= 1.42$) deviates more prominently than the oct30 ensemble ($= 1.34$). This deviation can be attributed to the synthesis process of these nanocrystals, wherein the cationic surfactant PDDA was employed to inhibit the growth of (111) facets. The presence of adsorbed PDDA molecules impedes the contact between the crystal facets and free gold atoms in the solution. It preferentially attaches to the (111) facets rather than the (100) facets, this results in an increased value of the sample's $R = h_{100}/h_{111}$ compared to the R_0.

In the Methods section, we propose a phenomenological model that shows how the adsorption efficiency of PDDA might change as the nanocrystals grow due to finite-area effects. Consider how each PDDA polymer (about 400 kDa to 500 kDa) is a linear molecule that spans several nanometers, which is comparable to the sizes of small nanocrystals. We assume that at our elevated nanocrystal growth temperatures, each PDDA polymer must be securely adsorbed onto a crystal facet via a minimum number of van der Waals contacts, $N_{\min}$. Hence, the attachment of an incoming PDDA polymer to a crystal facet will be frustrated by adsorbed PDDA that occupy possible attachment sites (fig. 8(a)). When the area of a facet that is covered by randomly adsorbed PDDA polymers reaches a critical fraction, the average number of contiguous attachment sites available to an incoming PDDA molecule falls below $N_{\min}$. Consequently, the attachment rate of new PDDA molecules slows dramatically due to frustrated attachment. Our simple model shows that this critical area fraction is reached sooner for smaller facet areas due to the size of randomly adsorbed PDDA. Conversely, this finite-area effect will become unimportant when the crystals are much larger than the average size of the PDDA molecule.

The finite-area effect mentioned in the previous paragraph suppresses both γ_{100} and γ_{111}. Recall that the Gibbs-Wulff theorem states that in an ideal crystal of constant volume at equilibrium, $\gamma \propto h$. Since $R = h_{100}/h_{111} = \gamma_{100}/\gamma_{111}$, one might expect R to be constant for such idealized nanocrystals. However, as shown in fig. 2(c), we observe that the area of the (111) facets expands relative to the (100) facets as the crystals grow from oct30 to oct40. Therefore, the finite-area effect suppresses γ_{111} less than γ_{100} in oct40 compared to oct30. This leads to R increasing in fig. 3(c) from oct30 to oct40.

The posterior distributions of oct30 and oct40 in the $(L_{\text{short}}, L_{\text{long}})$ subspace of fig. 3(d) show a linear trend with a $-\frac{1}{2}$ slope, which means $L_{\text{long}} + 2L_{\text{short}}$ is close to a constant. From fig. 3(a) we see that $L_{\text{long}} + 2L_{\text{short}}$ is the edge length of the enveloping octahedron. Hence, the distribution in fig. 3(d) for the millions of nanoparticles reveals two notable insights: a separate octahedral envelope that encompasses the structural variations within each ensemble; and the relative truncation of the (100) facets within this envelope decreases from oct30 to oct40.

The edge lengths of the enveloping octahedra for oct30 and oct40 are calculated from linear fits to their projected posterior distributions in fig. 3(d). Consistent with the explained variance ratios of the PCA in fig. 2(a), the nanoparticles' shape variations within oct30 or oct40 are largely due to different extents of (100) facet truncations within each ensemble's octahedral envelope. The $h_{(100)}$ distances of oct30 and oct40 octahedral envelopes increased from 26.4 nm to 32.8 nm (fig. 3(a)) respectively, while their corresponding edge lengths increased from 37.3 nm to 46.4 nm (fig. 3(e)). Notably, the average longest edge lengths of the (111) facets for oct30 and oct40 in fig. 3(a) are 21.3(28) nm and 28.4(33) nm respectively. These lengths overlap with those from the average 3-dimensional models reconstructed in Ayyer *et al.*[14] (see Section on Synthesis and measurements of nanoparticle ensembles).

Conclusions

Megahertz XFEL sources offer tremendous potential for inferring properties of particle ensembles numbering in the millions. The posterior distributions of these large ensembles of nanoparticles detail their structural dynamics and interactions. However, estimating these distributions is a computationally expensive and data-intensive endeavour.

This paper describes a scalable Monte Carlo importance-sampling framework to robustly estimate the posterior distribution of structural variations amongst very large numbers of single nanoparticles. By explicitly parameterizing these structures in the free-facet truncated octahedra (FFTO) space, we were able to avoid ambiguous features that often arise in prior-free induced 'manifolds' on XFEL datasets. Additionally, we also propose methods to validate the consistency of the recovered posterior distributions that circumvent the issues with Fourier Shell Correlation which is typically used in XFEL single particle imaging.[31]

Our manuscript also details practical implementation strategies to accelerate this importance sampling for millions of noisy and incomplete single-particle XFEL diffraction patterns. This includes an analytical approach to directly compute the diffraction pattern from polyhedra that can be efficiently implemented on GPGPUs (Supplementary). These strategies allow us to infer structural heterogeneity from datasets that are at least two orders of magnitude larger than what was previously attempted for single-particle XFEL imaging.

We interpreted such uncommonly high-dimensional posterior distributions using PCA, which showed that the structural variations within our truncated octahedra ensembles can be described by two independent degrees of freedom. By picking different projections of these two degrees of freedom, we inferred key signatures of non-equilibrium growth dynamics of nanocrystal growth, which led us to hypothesize a finite-area effect that might drive these dynamics away from equilibrium.

Our work shows a scalable statistical learning path to posterior estimation on massive datasets in high-throughput XFEL facilities worldwide. More broadly, it illuminates a similar path for data-driven heterogeneity mapping in single particle imaging, including cryo-electron microscopy. The four-ball model example in our manuscript shows that our framework also works for flexible particle chains (e.g., polymers, polypetides, etc). Here, efficiently parameterizing an object's structure is critical. Since the free-energy landscape of biomolecules can be embedded in a low-dimensional surface,[32] a low-dimensional parameterization of their structures might be possible. Ultimately, we have both the datasets and statistical learning tools for an unprecedented window into the hidden and chaotic world of nanoparticle dynamics.

Acknowledgements

The authors acknowledge the support of John C.H. Spence, who continues to inspire many of us. N.D.L. would like to thank the support of the Early Career Grant by the National University of Singapore, and the Singapore Ministry of Education AcRF Tier 1 grant. Z.S. thanks the Ph.D. scholarship from the Physics Department at the National University of Singapore. Both N.D.L. and Z.S. are grateful for the support by Bai Chang from the Centre for Bio-imaging Sciences at the National University of Singapore for IT infrastructure support. Paul Lourdu Xavier acknowledges a fellowship from the Joachim Herz Stiftung, Germany.

Paul Lourdu Xavier and Henry N. Chapman acknowledge support from the Human Frontiers Science Program, France (grant No. RGP0010/2017). Richard A. Kirian would like to thank the support of the National Science Foundation, BioXFEL Science and Technology Center (award #1231306) and the National Science Foundation, Directorate for Biological Sciences (award #1943448).

Methods

Degrees of freedom in nanoparticles' structure

Upon inspecting the 2D class averages of oct40 nanoparticles (fig. 5), it is observed that most diffraction patterns are characterized by approximately 12-15 radial resolution elements, as defined by Loh and Elser.[16] Consequently, the electron density maps of each nanoparticle can be represented by a 3D grid containing approximately 10^5 resolution elements, calculated from $\sim (2 \times 15 + 1)^3$. Although it is possible to determine the modal structures[20] of our nanoparticle ensemble in this 10^5-dimensional space, efficiency can be significantly enhanced with prior knowledge about these variations.

Data likelihood model

Our aim is to infer the posterior distribution $p(\rho \,|\, \mathbb{K})$, representing the structural conformations (ρs) within an ensemble, using collected diffraction patterns ($\mathbb{K}$). However, directly applying Bayes' theorem, $p(\rho \,|\, \mathbb{K})p(\mathbb{K}) = p(\mathbb{K} \,|\, \rho)p(\rho)$, to estimate $p(\rho \,|\, \mathbb{K})$ is not feasible due to the imprecision in defining both terms on the right-hand side. Conformation ρ is conceptualized as a function that assigns electron densities to points in real space, indicating that ρ's domain is infinitely dimensional. Defining $p(\rho)$ on such domain is a significant challenge. Furthermore, since observed patterns are derived from different instances of ρ, for any specific pair of K and ρ, $p(K \,|\, \rho)$ is likely to be zero. This leads to $p(\mathbb{K} \,|\, \rho)$ frequently approaching zero, causing the formula to be ill-defined and computationally unstable.

Instead of studying the full dataset, we should focus on a single pattern. In the context of a specific pattern pattern K and a particular feature ν, the posterior probability, $p(\nu \,|\, K)$, essentially quantifies how much K is distributed or "voted" to a ν. Then the averaged $p(\nu \,|\, K)$ over $K \in \mathbb{K}$, $p(\nu \,|\, \mathbb{K}) \equiv \langle p(\nu \,|\, K) \rangle_{K \in \mathbb{K}}$, gives us an overall posterior estimation over a whole ensemble.

According to Bayes' theorem:

$$p(\nu \,|\, K) p(K) = p(K \,|\, \nu) p(\nu) \ . \tag{1}$$

To make progress here, we will need an uninformative-prior assumption about the feature space: $p(\nu)$ is a constant. In addition, the value of $p(K \,|\, \nu)$ is approximated by $p(K \,|\, \nu; \mathbb{M}) \equiv \sum_{\rho \in \mathbb{M}} p(K \,|\, \rho) p(\rho \,|\, \nu)$. Assuming the uninformative prior, $p(\rho \,|\, \nu) = 1/N_{\nu,\mathbb{M}}$, where $N_{\nu,\mathbb{M}}$ is the number models in $\mathbb{M}$ having feature ν. The definition of $p(K \,|\, \rho)$ will be discussed later in eq. (5). To summarize the discussion above with formulae, we have

$$\begin{aligned} p(\nu \,|\, K) &\approx p(\nu \,|\, K; \mathbb{M}) \\ &\propto p(K \,|\, \nu; \mathbb{M}) \\ &= \sum_{\rho \in \mathbb{M}} p(K \,|\, \rho) p(\rho \,|\, \nu) \ , \end{aligned} \tag{2}$$

and

$$p(\nu \,|\, \mathbb{K}; \mathbb{M}) \equiv \frac{1}{|\mathbb{K}|} D(\nu \,|\, \mathbb{K}; \mathbb{M}) \ , \tag{3}$$

where

$$D(\nu \,|\, \mathbb{K}; \mathbb{M}) \equiv \sum_{K \in \mathbb{K}} p(\nu \,|\, K) \ . \tag{4}$$

It is worth to notice that $\sum_{\nu \in V} D(\nu \,|\, \mathbb{K}; \mathbb{M}) = |\mathbb{K}|$.

In XFEL-SPI, numerous far-field diffraction patterns ($\mathbb{K}$) are captured, each originating from a distinct particle in the ensemble illuminated by a single x-ray pulse. Occasionally, multiple particles may diffract from a single pulse, but this is predominantly filtered out *in* *silico* (as explained below). Disregarding background and inelastic scattering, these patterns represent the far-field diffraction resulting from the phase shift induced on the x-ray pulses by a particle's two-dimensional (2D) projected scattering potential. The orientation of each particle is unmeasured and has to be inferred.[16] Due to the photon limitation, these patterns essentially represent the Poisson-sampled Ewald sphere tomograms of the target particle's three-dimensional (3D) diffraction intensity W. This diffraction intensity varies linearly with the unmeasured local fluence of the XFEL pulse that illuminated each particle. Taken together, the likelihood[19] of measuring a particular pattern K given a tomogram W_Q of the particle presented at orientation Q is

$$p(K \,|\, Q, W, \phi) = \prod_{t \in \text{detector}} \frac{\mathrm{e}^{-\phi_K W_{Qt}} \left(\phi_K W_{Qt}\right)^{K_t}}{K_t!}, \tag{5}$$

where t indexes the detector's pixels, and ϕ_K is the local fluence rescaling factor for K.[17]

For a weakly scattering particle ρ, its diffraction intensities W are the squared modulus of the Fourier transform of the particle's real-space electron density distribution $\rho(\mathbf{r})$, which is represented as $W(\mathbf{q}) = \left|\mathcal{F}_{\mathbf{r}\to\mathbf{q}}[\rho(\mathbf{r})]\right|^2$. Thus, the likelihood of measuring a pattern K given an electron density ρ is

$$p(K \,|\, \rho) \equiv p(K \,|\, W) = \int d\phi_K \sum_{Q \in \mathbb{Q}} p(K \,|\, Q, W, \phi_K)\; p(Q) p(\phi_K)\;, \tag{6}$$

where $\mathbb{Q}$ is the set of orientations in $SO(3)$ space considered for particle ρ. The likelihood $p(K \,|\, \rho)$ here estimates how well each pattern K is matched to our Monte Carlo model ρ.

To simplify eq. (6), we once more apply the uninformative prior but this time on orientations: that the aerosolized particles do not have any orientation bias when injected into the path of the x-ray pulses (i.e. $p(Q)$ is a constant). Following this, we need to determine the most probable fluence rescaling factor for each pattern, ϕ_K, as it has been demonstrated to be vital for accurate multiple model reconstruction.[33] For this purpose, we conducted a single model EMC reconstruction[16] on each dataset K to ascertain the most probable rescaling factor $\widetilde{\phi}_K$ for each pattern.[19] Subsequently, we made the assumption that $p(\phi_K) = \delta(\phi_K - \widetilde{\phi}_K)$. The two assumptions outlined in this paragraph result in a streamlined version of the likelihood function in eq. (6), which is employed to assign weight to model importance in our Monte Carlo scheme:

$$p(K \mid \rho) \propto \sum_{Q \in \mathbb{Q}} p(K \mid Q, W, \widetilde{\phi}_K) \ . \tag{7}$$

Four-ball artificial model

As shown in fig. 1(c), the artificial model consists of four identical balls with centers at $\mathbf{0}$, $\mathbf{a}_1$, $\mathbf{a}_1 + \mathbf{a}_2$, and $\mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3$. The diameters of these balls are of unit length. In other words, $|\mathbf{a}_1| = |\mathbf{a}_2| = |\mathbf{a}_3| = 1$. We are only concerned with the model's structure, which is described by three degrees of freedom: α, β and γ. These are chosen as follows:

$$\begin{aligned} \alpha &= \langle -\mathbf{a}_1, \mathbf{a}_2 \rangle, \\ \beta &= \langle -\mathbf{a}_2, \mathbf{a}_3 \rangle, \\ \gamma &= \langle \mathbf{a}_1 \times \mathbf{a}_2, \mathbf{a}_2 \times \mathbf{a}_3 \rangle, \end{aligned}$$

where $\langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ represents the angle between vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. To generate new models, we perturb our canonical four-ball model, $(\alpha, \beta, \gamma) \to (\alpha + \delta_1, \beta + \delta_2, \gamma + \delta_3)$, where $\delta_{i=1,2,3} \in [-0.02, 0.02]$ are three independent uniform random numbers. We impose an extra constraint on the perturbed models that their individual α, β, γ cannot exceed the range $[1, 2.5] \times [1.2, 2.0] \times [0, 2.0]$. Perturbations that violate this constraint are discarded.

To generate the 100 000 diffraction patterns (i.e., upper black line in fig. 1(d), we first generated an ensemble of 100 perturbed models within the constrained angular ranges in the previous paragraph. Then 1000 diffraction patterns were generated from each perturbed model at random 3D orientations.

In the Monte Carlo search, to average out the effect from the choice of first sampled model, we sampled the dataset with 12 different random initial models, resulting 12 trajectories. The each trajectory has a length of 5000.

Strategies to accelerate Markov Chain Monte Carlo

The importance weight of each candidate model ρ is linked to the model's data likelihood $p(K \mid \rho)$ (eq. (5)). We accelerated this calculation with the following four strategies.

First, we partitioned the Monte Carlo model searches into smaller searches performed in parallel. Each dataset of diffraction patterns, $\mathbb{K}_{\text{oct30}}$ or $\mathbb{K}_{\text{oct40}}$, was randomly split into five similarly-sized, non-overlapping partitions. We accumulated eight different pools of Monte Carlo models for each partition, each containing 5000 models. Each pool was started from a randomly perturbed version of the same average model. Eventually, we accumulated 400 000 models: 200 000 for $\mathbb{K}_{\text{oct30}}$, and 200 000 for $\mathbb{K}_{\text{oct40}}$.

Second, the determination of every single diffraction pattern's orientation with respect to each 3D model is performed only once – when the 3D model is first added to the model pool. When additional models are added to this pool, we only need to rescale existing models' weights without comparing the latter against the diffraction data again. Overall, the number of orientations to be determined scaled like the product of the number of models and number of diffraction patterns.

Third, we accelerated the calculation of the likelihood $p(K \mid \rho)$, which as defined in eq. (6), compares each pattern K against tomograms of all possible orientations of each 3D FFTO model ρ in the model pool. However, in practice, only the likelihoods of a few orientations within each model were significant.[31] Put differently, $p(K \mid \rho)$ is sparse. Hence, we used coarse orientation sampling to first identify rotational neighbourhoods near these significant orientations. Then we increased the orientation sampling around these neighbourhoods for each data-model pair (K and ρ).

Fourth, we employed a memory-efficient approach to compute the two-dimensional Ewald sphere intensity section of each model ρ. A direct way to perform this job is to voxelize the real-space electron density of ρ, and then apply a fast Fourier transform on this. Instead, since ρ is a polyhedron with uniform density, we can compute its Fourier transform more accurately using a finite-element approach (see Methods). Briefly, each polyhedron is partitioned into non-overlapping tetrahedra, whose separately complex-valued Fourier transforms can be analytically computed and then coherently added together to give the Ewald sphere section of the original polyhedron.

These four time- and memory-saving approaches were implemented across 20 parallel-running NVIDIA GTX 1080 Ti GPUs. The 400 000 models for the $\mathbb{K}_{\mathrm{oct30}}$ and $\mathbb{K}_{\mathrm{oct40}}$ datasets were accumulated in approximately 240 hours.

In silico filtration with 2D EMC

As SEM images in fig. 4 show, our synthesized particles contained shapes that did not resemble truncated octahedra. Similar to previous analyses of this EuXFEL dataset[14] , we filtered out (*in silico*) some of the undesirable data heterogeneities using 2D classification via EMC method.[14] This method classifies diffraction patterns into multiple 2D models up to an overall in-plane rotation. This effectively helps us to identify significant patterns unlikely to arise from single truncated octahedra without having to reconstruct or compare them against 3D models. In fig. 5, these non-conforming patterns (dark red) clusters include: multiple-particle shots or triangular particles (cluster 1, 2 and 8), spherical patterns (cluster 2, 3, 4, 6, 11 and 27; absence of prominent streaks), and patterns with feature-less stripe (cluster 4). To increase the concentration of truncated octahedra, 2D EMC was applied in three rounds on the oct30 and oct40 datasets separately. After each round, non-conforming clusters were manually identified (like these dark red clusters in fig. 5) and discarded before the next round. Only patterns that survived all three times of filtration were used for this paper: 1282k out of 1608k for the oct30 dataset, and 823k of 1032k for the oct40 dataset.

For reference, we show typical 2D intensity slices of an ideal truncated octahedron from different orientations in the supplementary (fig. 9).

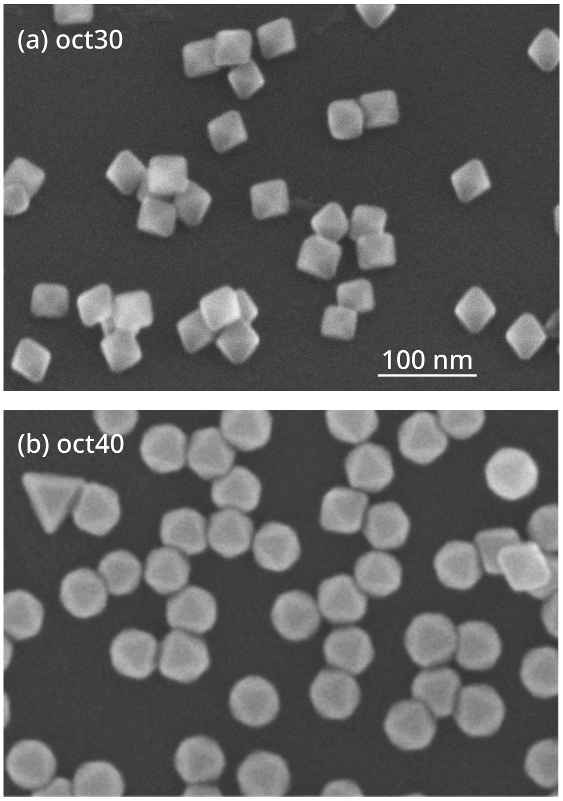

Figure 4: (a) and (b) represent SEM images of the oct30 and oct40 nanoparticle samples, respectively, each exhibiting nominal average widths of 30 nm and 40 nm. It is important to acknowledge that within the original sample, not all particles adopt an octahedral shape. Nevertheless, these non-octahedral variants can be effectively distinguished and filtered out through the application of the 2D EMC classification process.

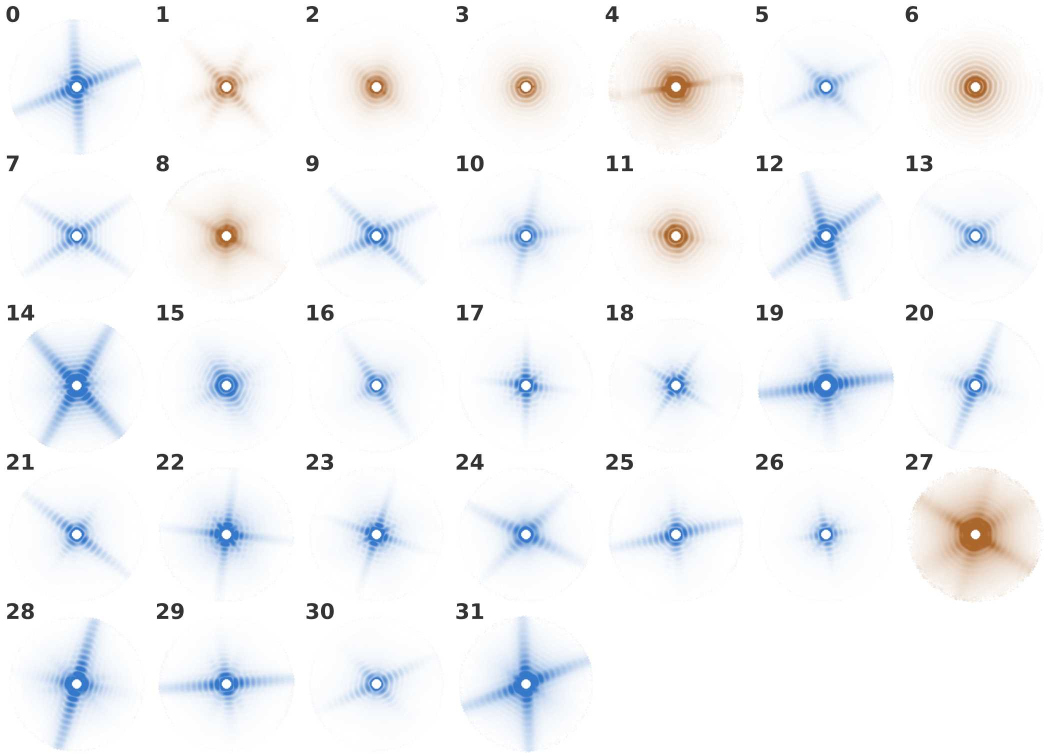

Figure 5: This is the 2D EMC classification of the raw oct40 dataset. Clusters colored with dark red are unlikely generated from an octahedral sample, hence filtered out for study in this paper.

Finite-element Fourier transform

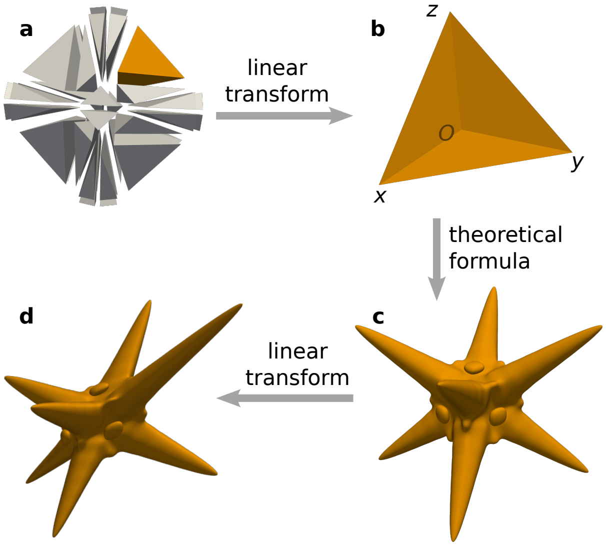

Figure 6: (a) A polyhedral volume is divided into several smaller non-overlapping tetrahedra. Each tetrahedron includes the coordinate origin $(0, 0, 0)$ and the three vertices of its triangulated meshes. (b) Each tetrahedra is linearly transformed to a standard trirectangular tetrahedron. (c) We can compute the complex-valued Fourier transform of each suitably "rectangularized" tetrahedron in (b). Here we show contours of such Fourier intensities. (d) The linear transformation from (a) to (b) is be reversed to obtain the Fourier transform of the highlighted (in orange) tetrahedron in (a).

The finite-element Fourier transform method explicitly calculates the Fourier transform of a uniform density polyhedral volume parameterized by its surface vertices without initially 'voxelizing' the volume onto a grid. A 'voxelized' electron density is conventionally needed to compute its 3D discrete Fourier transform (DFT), which can be readily compared with its XFEL diffraction patterns. Here, the inter-particle variations of L_{long} within the oct30 ensemble measures only 2-voxels in a 3D electron density array with size 251^3 according to the Nyquist–Shannon sampling theorem.[19] To minimize significant truncation errors when describing small but measurable size/shape variations among the nanoparticles using the voxelization approach, these volumes are typically padded with extra zeros (equivalent to oversampling their Fourier volumes by a multiplicative factor of α).

As shown in fig. 6, due to the linearity of the Fourier transform, the Fourier transform of a 3D volume is the sum of the Fourier transforms of its constituent non-overlapping tetrahedra.

We derive the analytical formula for computing the Fourier transform of arbitrary tetrahedra in the Supplementary material.

The complexity of computing Fourier transforms using this tetrahedralization method is $O(NM)$, where N is the number of point samples in the 3D Fourier volume and M is the number of tetrahedra. For the truncated octahedra in this work, $M = 48$. Comparatively, the computational complexity of 'voxelizing' then performing Fast Fourier Transform (FFT) on each unique polyhedra scales like $O(N\alpha \log_2(\alpha N))$, where α is the typical oversampling parameter needed to overcome the truncation issue with voxelizing polyhedra (discussed above).

Hence our tetrahedralization method will generally perform faster when $M < \alpha \log_2(\alpha N)$. This is true when we coarse-grain our candidate nanoparticles to polyhedra with relatively few faces (small M) while our diffraction patterns are highly oversampled (large αN). There are further time savings for the tetrahedralization method when we only need to compute a fraction of the full Fourier intensities (e.g., only along a handful of Ewald sphere slices).

Free Facet Truncated Octahedron (FFTO)

The FFTO model consists of 14 facets (fig. 2(b)): six for the (100) directions, and eight for the (111) directions. Each facet is described by a 3D vector $\mathbf{h} = (A, B, C)$, where (A, B, C) is a point on the facet where the vector (A, B, C) is also normal to the facet. The plane equation for such a facet is $\mathbf{h} \cdot (x, y, z) = |\mathbf{h}|^2$ or

$$Ax + By + Cz = A^2 + B^2 + C^2. \tag{8}$$

In total, $42 = 14 \times 3$ parameters are needed for each FFTO model. For the special case of an ideal FFTO model with perfect octahedral symmetry, their facets are described by the six cyclic permutations of $(\pm a, 0, 0)$ that describe the (100) facets, plus the eight combinations of $(\pm b, \pm b, \pm b)$ for the (111) facets.

To perturb an FFTO model, each facet, (A, B, C) is mapped to a new facet $(A, B, C) + \mathbf{v}$, where $\mathbf{v}$ is 3D uniform random vector within a 0.84 nm-radius ball. Each perturbation also needs to satisfy two constraints. The first constraint is that a model has to be a convex volume. The second constraint ensures each FFTO model stays reasonably close to the ideal truncated octahedron. To enforce these two constraints, the closest ideal truncated octahedron model (described in the previous paragraph) is found first for a given candidate FFTO model. This closeness is defined as the Euclidean distance between the 24 corresponding pairs of vertices between the two models. Hence, the closest ideal truncated octahedron to a perturbed FFTO model minimizes this total distance between the two models. If the distance between any two paired vertices between these two models is larger than 1.68 nm, then the perturbed FFTO candidate is rejected. For the Monte Carlo importance sampling, we will continue to perturb each FFTO model until these two constraints are satisfied.

Eliminate symmetry redundancy

Here we explain how we checked if two FFTO models are similar up to a particular permutation of their facet indices. This check is used to re-order the facet indices of our pool of models in fig. 2 to then distill model-model differences that are not due to trivial permutations of each model's facet indices. Each FFTO model is uniquely represented by an 14-element vector that detailed the areas of each FFTO model's 14 facets. Before any new facet-index permutation is attempted on model ρ, its 14-element area vector, $\mathbf{A}_\rho$, is normalized to $\mathbf{A}_\rho / V^{2/3}$ where V is the model's volume.

Rather than checking and permuting all possible pairs of models in our pool $\mathbb{M}$, we aim to to re-order each model's facet index to have the smallest distortion from the pool's average area vector $\overline{\mathbf{A}} = \frac{1}{|\mathbb{M}|} \sum_{\rho \in \mathbb{M}} \mathbf{A}_\rho$. This re-ordering is performed iteratively with two alternating steps: in the first step we compute $\overline{\mathbf{A}}$ given each model's current index order; then in the second step we re-order each model's indices to minimize the model's area vector from the mean vector using $\arg\min_{\hat{P}} |\hat{P}(\mathbf{A}_\rho) - \overline{\mathbf{A}}|^2$ where $\hat{P}$ refers to the facet-index permutation over the symmetry orbit of the ideal truncated octahedra. This iterative procedure is repeated until all re-ordering ceases and the mean vector stops changing.

Convergence of posterior estimation

We need to determine if our Markov Chain Monte Carlo (MCMC) scheme has accumulated a sufficiently large pool of models $\mathbb{M}$ that adequately samples the posterior distribution over all possible models $\{\rho\}$. Our demonstration of convergence comprises two steps. First, for our posterior estimation to have converged, it is *necessary* that the distribution differences between two iterations should be sufficiently small after convergence, or

$$\sum_{\nu} \left\| p(\nu \,|\, \mathbb{K}; \mathbb{M}^{(n)}) \cdot \nu - p(\nu \,|\, \mathbb{K}; \mathbb{M}^{(m)}) \cdot \nu \right\| \,, \tag{9}$$

is a sufficiently small value when m, n are sufficiently large, where m and n are iteration numbers, $\mathbb{M}^{(n)}$ is the sampled model pool at n^{th} iteration, and some normalization, $\| \cdot \|$, is used here for multiple-dimensional ν. Second, we further corroborate this convergence if $\mathbb{M}$ is a *self-consistent generative model.* We demonstrate this self-consistency by assuming a subset model pool $\mathbb{M}' \in \mathbb{M}$ as the synthetic ground truth from which a number of diffraction data are generated $\mathbb{K}'$; we then repeated our MCMC posterior estimation on $\mathbb{K}'$ to obtain a third model pool, $\mathbb{M}''$. For our posterior estimation $\mathbb{M}$ to have converged, it is necessary that the posterior predictive $p_T(\nu|\mathbb{K})$ marginalized over $\mathbb{M}$, $\mathbb{M}'$, and $\mathbb{M}''$ are sufficiently similar.

Figure 7(a) shows the convergence of our posterior predictive distribution $p(S_{111}, S_{100} \,|\, \mathbb{K})$, where the feature pair $\nu = \{S_{100}, S_{111}\}$ are the total areas of each FFTO model's (100) and (111) facets respectively. In practice, as most patterns are only in favor of one model, to speed up the calculation, we count only the best matched model for each pattern instead of strictly following the definitions in eqs. (3) and (9). We denote the area difference between two models, ρ_a and ρ_b, as $d(\rho_a, \rho_b) = d_{111}(\rho_a, \rho_b) + d_{100}(\rho_a, \rho_b) = \left|S_{111}(\rho_a) - S_{111}(\rho_b)\right| + \left|S_{100}(\rho_a) - S_{100}(\rho_b)\right|$. Then fig. 7(a) summarizes the change in area between the n^{th} and

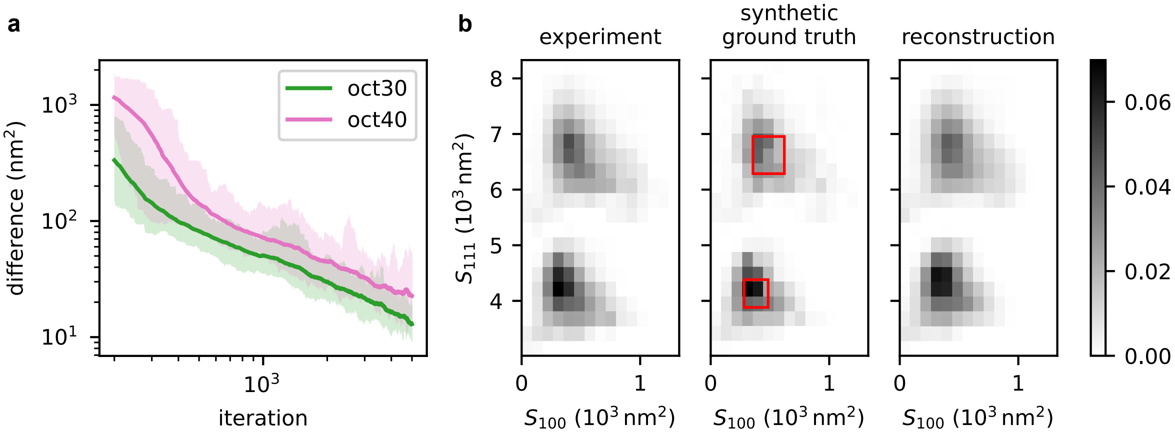

Figure 7: Validating the convergence and self-consistency of our reconstructed posterior predictive distribution $p(S_{111}, S_{100} \,|\, \mathbb{K})$. (a) *Convergence* in the difference areas between the most likely models in our reconstructed model pools from the n-th and $(n+200)$-th MCMC iteration. The maximum and minimum changes in areal differences averaged over 40 trajectories shown as faint color fills around the mean change (darker line). (b) *Self-consistency* in our reconstructed $p(S_{111}, S_{100} \,|\, \mathbb{K})$. Left panel (experiment): computed from Bayes model averaging over the pool of FFTO models $\mathbb{M}$ given the diffraction data $\mathbb{K}$. Middle panel (synthetic ground truth): $p(S_{111}, S_{100} \,|\, \mathbb{K})$ of a synthetic ground truth model ensemble generated from a random subset of $\mathbb{M}$. Right panel (reconstructed): $p(S_{111}, S_{100} \,|\, \mathbb{K})$ of a new pool of models $\mathbb{M}'$ reconstructed from random patterns generated by the synthetic ground truth.

$(n+200)^{\text{th}}$ MCMC iteration as $\big\langle d\big(\rho_n^{(K)}, \rho_{n+200}^{(K)}\big)\big\rangle_{K\in\mathbb{K}}$, where the $\rho_n^{(K)}$ stands for the best matched model for a pattern K in the model pool $\big\{\rho_1, \rho_2, \ldots, \rho_n\big\}$. The colored fills in fig. 7(a) span the largest and smallest values among all 40 trajectories (eight trajectories for all five non-overlapping partitions of the full dataset) at each iteration. By iteration $n = 5000$, the magnitude of this areal differences is about $10\,\text{nm}^2$, which is less than 1% compared to the total area of a particle.

In the second step of our validation, we tested for self-consistency of the MCMC model pool $\mathbb{M}$ that was reconstructed from diffraction data $\mathbb{K}$. From $\mathbb{M}$ we picked the 2000 best-matched models of 2000 randomly selected patterns in $\mathbb{K}$. These 2000 models forms a *synthetic ground truth* pool of models $\mathbb{M}'$. We then generated 1000 diffraction patterns from each model in $\mathbb{M}'$, denoting these patterns as $\mathbb{K}'$. Each of these patterns are randomly oriented, and rescaled from the distribution of factors recovered in the earlier single-model EMC reconstruction of $\mathbb{K}$ that initialized the reconstruction of $\mathbb{M}$.[19] Thereafter, we used the same MCMC procedure used to recover a third model pool $\mathbb{M}''$ from $\mathbb{K}'$. Figure 7(b) shows three posterior predictive distributions marginalized over the model pool $\mathbb{M}$ reconstructed from $\mathbb{K}$, the synthetic ground truth $\mathbb{M}'$, and $\mathbb{M}''$ reconstructed from $\mathbb{K}'$. Since we know the ground truth models $\mathbb{M}'$ for every pattern in the synthetic dataset $\mathbb{K}'$, we can evaluate the area differences, d_{111} and d_{100}, between the ground truth models and reconstructed best-matched models in $\mathbb{M}''$. The two red rectangles in fig. 7(b) mark the average difference in d_{111} and d_{100} for the oct30 and oct40 datasets.

PDDA coverage

A simple model is proposed to show the finite size effect on PDDA coverage on crystal facet growth at elevated temperatures (fig. 8(b)). The synthesis protocol creates different shaped Au nanoparticles by adding PDDA polymer chains to the growth solution.[5,26] Each PDDA polymeric molecule has probability of attaching to the crystal facets only if there is sufficient areal contact between them. In this model, we use an $N \times N$ square lattice (the gray lattice

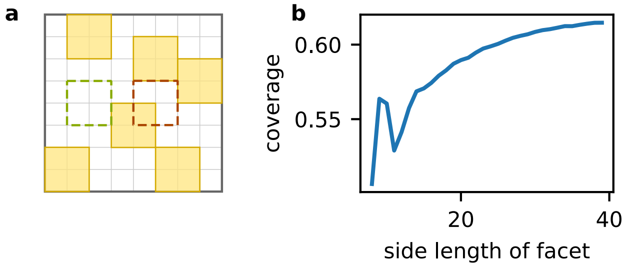

Figure 8: Fractional coverage of non-overlapping PDDA molecules on a crystal facet depends on the facet's area. (a) The grey lattice represents a crystal facet covered by PDDA molecules (yellow squares). The PDDA molecules are not allowed to overlap (i.e., a next PDDA can occupy the green dashed square but not the red one). (b) The PDDA coverage vs. the side length of a facet.

in fig. 8(a)) to simulate a crystal facet. Thus N could be regarded as the side length of a crystal facet whose typical size is few tens nanometers. The PDDA molecules that attach to the facet are abstracted as an $L \times L$ square (yellow square in fig. 8(a)). We attempt to place PDDA molecules randomly over this facet such that no two PDDA molecules overlap. This mutual exclusion requirement expresses

Then the coverage is nL^2/N^2, where n is the number of PDDAs placed. Since the size of a PDDA is few nanometers, we choose $L = 4$ in the simulation. For each N, simulations were run 20 000 times. As shown in fig. 8(b), the average coverage is increasing with the side length of a facet, which causes effectively smaller surface tension.

Supplementary

Match 2D EMC clusters with ideal 2D intensity slices

In the fig. 9, we manually match several typical 2D EMC clusters with 2D intensity slices of an ideal octahedra from different orientations. The shape of this octahedra is given by the average model we reconstructed.

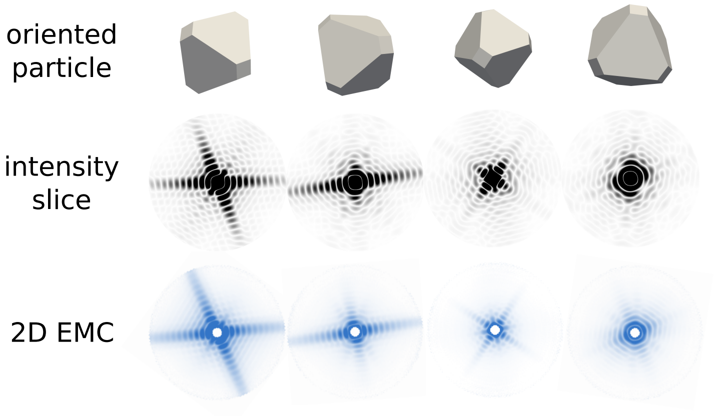

Figure 9: Here, we select four typical 2D EMC clusters (blue). For each one, a 2D intensity (black) slice is manually matched, and the corresponding oriented particle is showed in grey.

Finite volume Fourier Transformation

Comparing to biomolecules, nanoparticles are also a common kind of sample in XFEL SPI experiment but has a much simpler structure. Usually, we could assume its density is uniform and has a polyhedron shape. In this subsection, we propose a numerical Fourier transformation scheme for any uniform polyhedron S. This method could avoid the voxelization of S. As the difficulty of the Fourier transformation of S , $\iiint_S \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x})\ \mathrm{d}\mathbf{x}$ comes from the complexity of the shape of S, we convert this one big integral into several smaller integrals over tetrahedrons. The surface (boundary) of S, ∂S, can be triangulated into a list of n triangular faces

$$\left\{[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}]\,\middle|\, i = 1, 2, \ldots, n\right\},$$

where $v^{(i)}$s are the vertices of triangular i, $[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}]$. Suppose S is convex and the origin v_O is inside S. Then $S = \sum_i [v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_O]$ where $[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_O]$ is the tetrahedron with base $[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}]$ and apex v_O. Immediately, we get

$$\iiint_S \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x})\ \mathrm{d}\mathbf{x} = \iiint\limits_{\sum_i [v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_O]} \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x})\ \mathrm{d}\mathbf{x}. \tag{10}$$

Equation (10) retains its validity even when these two assumptions are relaxed. A straightforward argument in support of this is that both sides of eq. (10) exhibit continuity across all vertices, and the integral remains independent of the choice of origin. The critical factor here is ensuring the correct orientation of a surface, which is determined by the sequential selection of three vertices, $\{v_1^{(i)}, v_2^{(i)}, v_3^{(i)}\}$, on the surface following the right-hand rule. It is imperative that the orientation, represented by the "thumb" direction, points outward from the S. A more rigorous mathematical statement pertaining to this concept, "orientability", can be found in most algebraic topology textbooks.

The region covered by $[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_O]$ and $[v_x, v_y, v_z, v_O]$ can be converted into each other by a linear transform, A,

$$A^{-\mathrm{T}}[\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_O] = \begin{bmatrix} 1 & & & 0 \\ & 1 & & 0 \\ & & 1 & 0 \end{bmatrix} = [\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z, \mathbf{v}_O] \tag{11}$$

where

$$A = \begin{bmatrix} \mathbf{v}_1^{\mathrm{T}} \\ \mathbf{v}_2^{\mathrm{T}} \\ \mathbf{v}_3^{\mathrm{T}} \end{bmatrix}, \tag{12}$$

v_x, v_y, and v_z are three unit points of x, y, and z-axis, and the $\mathbf{v}$ emphasizes that it is a column vector of vertex v. The integrals over $[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_O]$ in eq. (10) can be converted to integrals over the same trirectangular triangular pyramid $[v_x, v_y, v_z, v_O]$.

$$\begin{aligned} \iiint_S \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x})\ \mathrm{d}\mathbf{x} &= \sum_i \iiint\limits_{[v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_O]} \exp(\mathrm{i}\mathbf{k}\cdot\mathbf{x})\ \mathrm{d}\mathbf{x} \\ &= \sum_i \det A^{(i)} \iiint\limits_{[v_x, v_y, v_z, v_O]} \exp(\mathrm{i}A^{(i)}\mathbf{k}\cdot\mathbf{x})\ \mathrm{d}\mathbf{x} \\ &= \sum_i \det A^{(i)} \cdot F^\star(A^{(i)}\mathbf{k}) \end{aligned}$$

where $F^\star$ is the Fourier transformation of $[v_x, v_y, v_z, v_O]$. The closed-form expression for $F^\star$

can be found by

$$\begin{aligned} F^\star(\mathbf{k}) &= \int_0^1 \int_0^{1-x} \int_0^{1-x-y} \exp\left[\mathrm{i}(k_x x + k_y y + k_z z)\right] \,\mathrm{d}z\,\mathrm{d}y\,\mathrm{d}x \\ &= \frac{\mathrm{i}\sum_{xyz}\left[\exp(\mathrm{i}k_x) - 1\right] k_y k_z (k_z - k_y)}{\prod_{xyz} k_x \prod_{xyz}(k_x - k_y)}, \end{aligned} \tag{13}$$

where $\sum_{xyz}$ and $\prod_{xyz}$ are the index-rolling summation and production, for instance, $\sum_{xyz} k_x = k_x + k_y + k_z$, $\prod_{xyz}(k_y - k_z) = (k_y - k_z)(k_z - k_x)(k_x - k_y)$.

In the numerical calculation of eq. (13), zero denominator in eq. (13) causes zero division error. Those cases are supposed to be handled separately by calculating the limits of eq. (13). It should be pointed out that as the right-triangular pyramid has $\mathrm{C_{3v}}$ symmetry, all permutations $(\sigma(x)\ \sigma(y)\ \sigma(z))$, like $(y\ x\ z)$, to the index list $(x\ y\ z)$ in $F^\star(k_x, k_y, k_z)$ give the same values, which also could be verified directly from eq. (13). Therefore, all zero-division cases are divided into six classes, and the condition label for each class is a representative of its index-permutated class.

1. $k_x = k_y = k_z = 0$

$$F^\star(0,0,0) = \frac{1}{6};$$

2. $k_x = k_y = k_z \neq 0$

$$F^\star(k_x, k_x, k_x) = \frac{-2\mathrm{i} + \exp(\mathrm{i}k_x)(2\mathrm{i} + 2k_x - \mathrm{i}k_x^2)}{2k_x^3};$$

3. $k_x \neq 0$, $k_y = k_z = 0$

$$F^\star(k_x, 0, 0) = \frac{1}{k_x^2} + \frac{\mathrm{i}\left[-2 + 2\exp(\mathrm{i}k_x) + k_x^2\right]}{2k_x^3};$$

4. $k_z = k_x \neq 0$, $k_y = 0$

$$F^\star(k_x, 0, k_x) = -\frac{-2\mathrm{i} + k_x + \exp(\mathrm{i}k_x)(2\mathrm{i} + k_x)}{k_x^3};$$

5. $k_x = 0,\ k_y \neq 0,\ k_z \neq 0,\ k_y \neq k_z$

$$F^\star(0, k_y, k_z) = \frac{\mathrm{i}\big[\exp(\mathrm{i}k_y) - 1\big]}{k_y^2(k_y - k_z)} + \frac{\mathrm{i}\big[\exp(\mathrm{i}k_z) - 1\big]}{k_z^2(k_z - k_y)} - \frac{1}{k_y k_z};$$

6. $k_x = k_y \neq 0,\ k_z \neq 0$

$$\begin{aligned} F^\star(k_x, k_x, k_z) = \frac{1}{k_y^2(k_y - k_z)^2 k_z} \Big\{ & \mathrm{i}\big[\exp(\mathrm{i}k_y) - 1\big] k_z^2 + \\ & \mathrm{i}k_y^2\big[\exp(\mathrm{i}k_z) - 1 + \mathrm{i}\exp(\mathrm{i}k_y) k_z\big] + \\ & k_y k_z \big[2\mathrm{i} + \exp(\mathrm{i}k_y)(-2\mathrm{i} + k_z)\big] \Big\}. \end{aligned}$$

References

1. Li, X.-B.; Tung, C.-H.; Wu, L.-Z. Semiconducting Quantum Dots for Artificial Photosynthesis. *Nature Reviews Chemistry* **2018**, *2*, 160–173.

2. Liu, M.; Yazdani, N.; Yarema, M.; Jansen, M.; Wood, V.; Sargent, E. H. Colloidal Quantum Dot Electronics. *Nature Electronics* **2021**, *4*, 548–558.

3. Shi, Y.; Lyu, Z.; Zhao, M.; Chen, R.; Nguyen, Q. N.; Xia, Y. Noble-Metal Nanocrystals with Controlled Shapes for Catalytic and Electrocatalytic Applications. *Chemical Reviews* **2021**, *121*, 649–735.

4. Qin, R.; Liu, K.; Wu, Q.; Zheng, N. Surface Coordination Chemistry of Atomically Dispersed Metal Catalysts. *Chemical Reviews* **2020**, *120*, 11810–11899.

5. Li, C.; Shuford, K. L.; Chen, M.; Lee, E. J.; Cho, S. O. A Facile Polyol Route to Uniform Gold Octahedra with Tailorable Size and Their Optical Properties. *ACS Nano* **2008**, *2*, 1760–1769.

6. Modena, M. M.; Rühle, B.; Burg, T. P.; Wuttke, S. Nanoparticle Characterization: What to Measure? *Advanced Materials* **2019**, 1901556.

7. Haiss, W.; Thanh, N. T. K.; Aveyard, J.; Fernig, D. G. Determination of Size and Concentration of Gold Nanoparticles from UV-Vis Spectra. *Analytical Chemistry* **2007**, *79*, 4215–4221.

8. Nakamura, K.; Kawabata, T.; Mori, Y. Size Distribution Analysis of Colloidal Gold by Small Angle X-ray Scattering and Light Absorbance. *Powder Technology* **2003**, *131*, 120–128.

9. Shields, S. P.; Richards, V. N.; Buhro, W. E. Nucleation Control of Size and Dispersity in Aggregative Nanoparticle Growth. A Study of the Coarsening Kinetics of Thiolate-Capped Gold Nanocrystals. *Chemistry of Materials* **2010**, *22*, 3212–3225.

10. Woehl, T. J.; Park, C.; Evans, J. E.; Arslan, I.; Ristenpart, W. D.; Browning, N. D. Direct Observation of Aggregative Nanoparticle Growth: Kinetic Modeling of the Size Distribution and Growth Rate. *Nano Letters* **2014**, *14*, 373–378.

11. Florea, I.; Feral-Martin, C.; Majimel, J.; Ihiawakrim, D.; Hirlimann, C.; Ersen, O. Three-Dimensional Tomographic Analyses of CeO $_2$ Nanoparticles. *Crystal Growth & Design* **2013**, *13*, 1110–1121.

12. Tan, Y. Z.; Baldwin, P. R.; Davis, J. H.; Williamson, J. R.; Potter, C. S.; Carragher, B.; Lyumkis, D. Addressing Preferred Specimen Orientation in Single-Particle Cryo-EM through Tilting. *Nature Methods* **2017**, *14*, 793–796.

13. Chapman, H. N. X-Ray Free-Electron Lasers for the Structure and Dynamics of Macromolecules. *Annual Review of Biochemistry* **2019**, *88*, 35–58.

14. Ayyer, K.; Xavier, P. L.; Bielecki, J.; Shen, Z.; Daurer, B. J.; Samanta, A. K.; Awel, S.; Bean, R.; Barty, A.; Bergemann, M.; Ekeberg, T.; Estillore, A. D.; Fangohr, H.; Gieweke-

meyer, K.; Hunter, M. S.; Karnevskiy, M.; Kirian, R. A.; Kirkwood, H.; Kim, Y.; Koliyadu, J. *et al.* 3D Diffractive Imaging of Nanoparticle Ensembles Using an X-Ray Laser. *Optica* **2021**, *8*, 15–23.

15. Tokuhisa, A.; Kanada, R.; Chiba, S.; Terayama, K.; Isaka, Y.; Ma, B.; Kamiya, N.; Okuno, Y. Coarse-Grained Diffraction Template Matching Model to Retrieve Multiconformational Models for Biomolecule Structures from Noisy Diffraction Patterns. *Journal of Chemical Information and Modeling* **2020**, *60*, 2803–2818.

16. Loh, N.-T. D.; Elser, V. Reconstruction Algorithm for Single-Particle Diffraction Imaging Experiments. *Physical Review E* **2009**, *80*, 026705.

17. Loh, N. D.; Bogan, M. J.; Elser, V.; Barty, A.; Boutet, S.; Bajt, S.; Hajdu, J.; Ekeberg, T.; Maia, F. R. N. C.; Schulz, J.; Seibert, M. M.; Iwan, B.; Timneanu, N.; Marchesini, S.; Schlichting, I.; Shoeman, R. L.; Lomb, L.; Frank, M.; Liang, M.; Chapman, H. N. Cryptotomography: Reconstructing 3D Fourier Intensities from Randomly Oriented Single-Shot Diffraction Patterns. *Physical Review Letters* **2010**, *104*.

18. Ekeberg, T.; Svenda, M.; Abergel, C.; Maia, F. R. N. C.; Seltzer, V.; Claverie, J.-M.; Hantke, M.; Jönsson, O.; Nettelblad, C.; van der Schot, G.; Liang, M.; DePonte, D. P.; Barty, A.; Seibert, M. M.; Iwan, B.; Andersson, I.; Loh, N. D.; Martin, A. V.; Chapman, H.; Bostedt, C. *et al.* Three-Dimensional Reconstruction of the Giant Mimivirus Particle with an X-Ray Free-Electron Laser. *Physical Review Letters* **2015**, *114*, 098102.

19. Ayyer, K.; Lan, T.-Y.; Elser, V.; Loh, N. D. *Dragonfly* : An Implementation of the Expand–Maximize–Compress Algorithm for Single-Particle Imaging. *Journal of Applied Crystallography* **2016**, *49*, 1320–1335.

20. Cho, D. H.; Shen, Z.; Ihm, Y.; Wi, D. H.; Jung, C.; Nam, D.; Kim, S.; Park, S.-Y.; Kim, K. S.; Sung, D.; Lee, H.; Shin, J.-Y.; Hwang, J.; Lee, S. Y.; Lee, S. Y.; Han, S. W.;

Noh, D. Y.; Loh, N. D.; Song, C. High-Throughput 3D Ensemble Characterization of Individual Core–Shell Nanoparticles with X-ray Free Electron Laser Single-Particle Imaging. *ACS Nano* **2021**,

21. Elser, V. Phase Retrieval by Iterated Projections. *JOSA A* **2003**, *20*, 40–55.

22. Shen, Z. Data Heterogeneity in Single Particle Imaging Experiment with X-ray Free Electron Laser. Ph.D. thesis, National University of Singapore (Singapore), 2021.

23. Daurer, B. J.; Okamoto, K.; Bielecki, J.; Maia, F. R. N. C.; Mühlig, K.; Seibert, M. M.; Hantke, M. F.; Nettelblad, C.; Benner, W. H.; Svenda, M.; Tîmneanu, N.; Ekeberg, T.; Loh, N. D.; Pietrini, A.; Zani, A.; Rath, A. D.; Westphal, D.; Kirian, R. A.; Awel, S.; Wiedorn, M. O. *et al.* Experimental Strategies for Imaging Bioparticles with Femtosecond Hard X-ray Pulses. *IUCrJ* **2017**, *4*, 251–262.

24. Loh, N. D.; Starodub, D.; Lomb, L.; Hampton, C. Y.; Martin, A. V.; Sierra, R. G.; Barty, A.; Aquila, A.; Schulz, J.; Steinbrener, J.; Shoeman, R. L.; Kassemeyer, S.; Bostedt, C.; Bozek, J.; Epp, S. W.; Erk, B.; Hartmann, R.; Rolles, D.; Rudenko, A.; Rudek, B. *et al.* Sensing the Wavefront of X-Ray Free-Electron Lasers Using Aerosol Spheres. *Optics Express* **2013**, *21*, 12385.

25. Hantke, M. F.; Hasse, D.; Maia, F. R. N. C.; Ekeberg, T.; John, K.; Svenda, M.; Loh, N. D.; Martin, A. V.; Timneanu, N.; Larsson, D. S. D.; van der Schot, G.; Carlsson, G. H.; Ingelman, M.; Andreasson, J.; Westphal, D.; Liang, M.; Stellato, F.; DePonte, D. P.; Hartmann, R.; Kimmel, N. *et al.* High-Throughput Imaging of Heterogeneous Cell Organelles with an X-ray Laser. *Nature Photonics* **2014**, *8*, 943–949.

26. Lu, Y.; Zhang, H.; Wu, F.; Liu, H.; Fang, J. Size-Tunable Uniform Gold Octahedra: Fast Synthesis, Characterization, and Plasmonic Properties. *RSC Advances* **2017**, *7*, 18601–18608.

27. Alpay, D.; Peng, L.; Marks, L. D. Are Nanoparticle Corners Round? *The Journal of Physical Chemistry C* **2015**, *119*, 21018–21023.

28. Vitos, L.; Ruban, A.; Skriver, H.; Kollár, J. The Surface Energy of Metals. *Surface Science* **1998**, *411*, 186–202.

29. Vanselow, R., Howe, R. F., Eds. *Chemistry and Physics of Solid Surfaces VII*; Springer Series in Surface Sciences; Springer-Verlag: Berlin Heidelberg, 1988.

30. Ref. 29, Chapter 13.

31. Shen, Z.; Teo, C. Z. W.; Ayyer, K.; Loh, N. D. An Encryption–Decryption Framework to Validating Single-Particle Imaging. *Scientific Reports* **2021**, *11*, 971.

32. Neupane, K.; Manuel, A. P.; Woodside, M. T. Protein Folding Trajectories Can Be Described Quantitatively by One-Dimensional Diffusion over Measured Energy Landscapes. *Nature Physics* **2016**, *12*, 700–703.

33. Daurer, B. J.; Sala, S.; Hantke, M. F.; Reddy, H. K. N.; Bielecki, J.; Shen, Z.; Nettelblad, C.; Svenda, M.; Ekeberg, T.; Carini, G. A.; Hart, P.; Osipov, T.; Aquila, A.; Loh, N. D.; Maia, F. R. N. C.; Thibault, P. Ptychographic Wavefront Characterization for Single-Particle Imaging at x-Ray Lasers. *Optica* **2021**, *8*, 551–562.