Supplemental material: Correlations in scattered x-ray laser pulses reveal nanoscale structural features of viruses

Ruslan P. Kurta,[1] Jeffrey J. Donatelli,[2,3] Chun Hong Yoon,[4] Peter Berntsen,[5] Johan Bielecki,[6,1] Benedikt J. Daurer,[6] Hasan DeMirci,[7,8] Petra Fromme,[9] Max Felix Hantke,[6] Filipe R. N. C. Maia,[6,10] Anna Munke,[6] Carl Nettelblad,[11,6] Kanupriya Pande,[12,3] Hemanth K. N. Reddy,[6] Jonas A. Sellberg,[13,6] Raymond G. Sierra,[4] Martin Svenda,[6] Gijs van der Schot,[6] Ivan A. Vartanyants,[14,15] Garth J. Williams,[16] P. Lourdu Xavier,[17,18] Andrew Aquila,[4] Peter H. Zwart,[12,3] and Adrian P. Mancuso[1]

[1]*European XFEL GmbH, Holzkoppel 4, D-22869 Schenefeld, Germany*

[2]*Mathematics Department, Lawrence Berkeley National Laboratory,*
1 Cyclotron Road, Berkeley, CA 94720, USA

[3]*Center for Advanced Mathematics for Energy Research Applications,*
1 Cyclotron Road, Berkeley, CA 94720, USA

[4]*Linac Coherent Light Source, SLAC National Accelerator Laboratory,*
2575 Sand Hill Road, Menlo Park, CA 94025, USA

[5]*Australian Research Council Centre of Excellence in Advanced Molecular Imaging,*
La Trobe Institute for Molecular Science,
La Trobe University, Melbourne 3086, Australia

[6]*Laboratory of Molecular Biophysics,*
Department of Cell and Molecular Biology, Uppsala University, Sweden

[7]*Biosciences Division, SLAC National Accelerator Laboratory,*
2575 Sand Hill Road, Menlo Park, CA 94025, USA

[8]*Stanford PULSE Institute, SLAC National Accelerator Laboratory,*
2575 Sand Hill Road, Menlo Park, CA 94025, USA

[9]*Biodesign Center for Applied Structural Discovery and School of Molecular Sciences,*
Arizona State University, Tempe, AZ 85287-1604, USA

[10]*NERSC, Lawrence Berkeley National Laboratory, Berkeley, California, USA*

[11]*Division of Scientific Computing, Science for Life Laboratory,*
Department of Information Technology, Uppsala University, Sweden

[12]*Molecular Biophysics and Integrated Bioimaging,*

Lawrence Berkeley National Laboratory,
1 Cyclotron Road, Berkeley, CA 94720, USA
[13]*Biomedical and X-Ray Physics, Department of Applied Physics,*
AlbaNova University Center, KTH Royal Institute
of Technology, Stockholm SE-106 91, Sweden
[14]*Deutsches Elektronen-Synchrotron DESY,*
Notkestraße 85, D-22607 Hamburg, Germany
[15]*National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),*
Kashirskoe shosse 31, 115409 Moscow, Russia
[16]*NSLS-II, Brookhaven National Laboratory,*
PO Box 5000, Upton, NY 11973, USA
[17]*Center for Free-Electron Laser Science,*
Deutsches Elektronen-Synchrotron DESY,
Notkestraße 85, 22607 Hamburg, Germany
[18]*Max-Planck Institute for the Structure and Dynamics of Matter, 22607 Hamburg, Germany*

(Dated: September 7, 2017)

EXPERIMENTAL DATA PREPROCESSING AND CLASSIFICATION

A large set consisting of more than two million images, containing single hits, as well as hits from multiple particles and aggregates [see Fig. S1], was classified by the diffusion map embedding method to extract single-particle hits, as described in ref. [1]. Additional manual filtering was performed to filter out the remaining alien images corresponding to diffraction from multiple viruses [see Fig. S1(d)], aggregates [Fig. S1(e)], and spherical droplets [Fig. S1(f)], as well as the images with visible detector failure, leading to a rejection of 8.6% and 8% of the images classified as single hits for PR772 and RDV, correspondingly. An intensity threshold was then applied to the extracted single-hits to reject all images with an average intensity smaller than 4500 ADUs/pixel (analog-to-digital units per pixel), leading to the final datasets consisting of 1400 "high-intensity" single PR772 hits and 760 "high-intensity" single RDV hits, where the scattered signal in most of the images was recorded up to the edges of the detector. For the last preprocessing step, the centers of the diffraction patterns were refined individually for each image in each of the datasets, as required for x-ray cross-correlation analysis (XCCA).

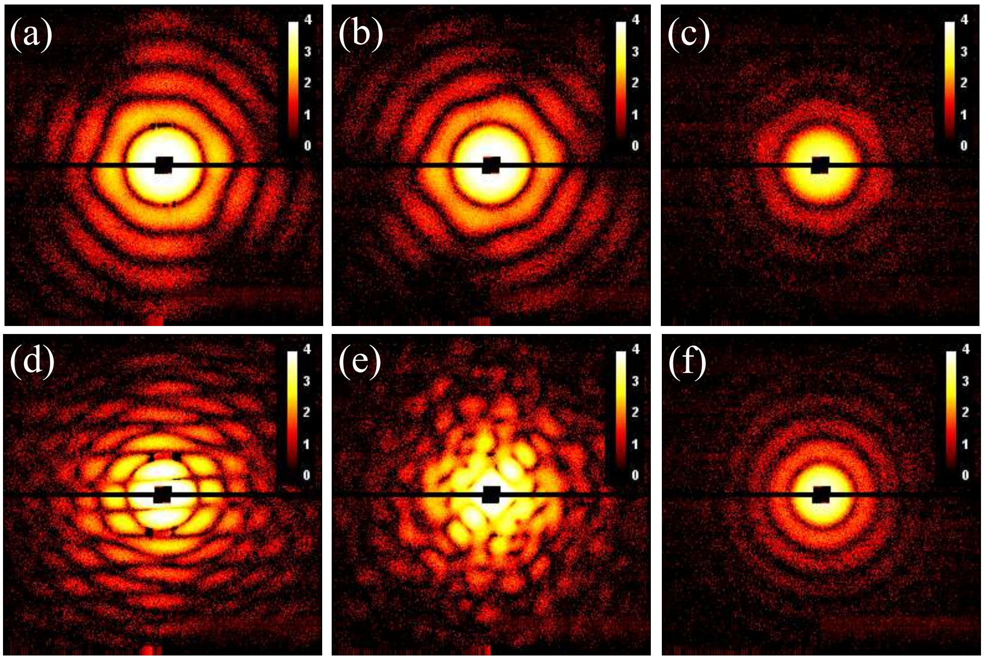

FIG. S1. Scattering data measured from RDV samples (logarithmic scale). High-intensity (a),(b) and low-intensity (c) diffraction patterns from single RDV particles. Diffraction patterns from multiple viruses (d), aggregates (e), and spherical droplets of buffer (f).

PARTICLE SIZE DETERMINATION

By inspecting individual radial intensity profiles $\langle I(q,\varphi)\rangle_\varphi$, one can find that the two datasets are characterized by a certain size distribution of RDV and PR772 particles, which was used to further classify the diffraction patterns according to their corresponding particle sizes. Here we explicitly define the size of an icosahedral particle to be equal to $d_{\rm max}$, the maximum pair distance in the particle. Note that $d_{\rm max} = 2R_{\rm cirscr}$, where $R_{\rm cirscr}$ is the radius of a circumscribed sphere [see inset in Fig. S2(a)]. Image classification was performed based on two types of fits of $\langle I(q,\varphi)\rangle_\varphi$ as a function of q for individual diffraction patterns: a Guinier-type fit [2] $\langle I(q,\varphi)\rangle_\varphi = I(0)\exp(-q^2R_{\rm g}^2/3)$ in the low-q range and a fit with a form factor of a spherical particle, $\langle I(q,\varphi)\rangle_\varphi = A[(\sin(qR_{\rm s}) - qR_{\rm s}\cos(qR_{\rm s}))/q^3]^2$, in the neighborhood of the first minimum of $\langle I(q,\varphi)\rangle_\varphi$ [see Fig. S2(a)]. Here $I(0)$ and A are scaling parameters, $R_{\rm g}$ is the radius of gyration [2], and $R_{\rm s}$ is the radius of the corresponding spherical particle. Fitting with the form factor of a sphere was restricted to the range of $q = (0.08, 0.17)$ nm^{-1} for RDV and $q = (0.11, 0.17)$ nm^{-1} for PR772 viruses, and the Guinier fitting was performed in the range $q = (0.054, 0.077)$ nm^{-1} for both types of particles.

Simulations show that, for particles with small shape anisotropy (like icosahedral-shaped RDV or PR772), the position of the first minimum is almost independent of particle orientation, and therefore, it can be used to characterize particle size, while positions of higher-order minima become dependent on particle orientation. Our results of simulations for solid icosahedral particles of different sizes show that particle sizes determined from the Guinier-type fits ($2R_{\rm g}$) and spherical form-factor fits ($2R_{\rm s}$) give systematically lower values as compared to the real particle size ($2R_{\rm cirscr}$) [see Fig. S2(b)]. Inaccuracies in the Guinier analysis appear because we perform fitting $\langle I(q,\varphi)\rangle_\varphi$ outside of the proper Guinier $q-$range, defined as $q < 1/R_{\rm g}$ (or $q < 1.3/R_{\rm g}$ in the worst case) [2], and also because we actually do not fit a proper SAXS intensity, but rather $\langle I(q,\varphi)\rangle_\varphi$ corresponding to a certain particle orientation (this is why we call it a “Guinier-type” fit). The parameter $R_{\rm s}$ determined from the spherical form-factor fit is equal to the radius of the volume-equivalent spherical particle. In principle, the size of an ideal icosahedral particle $d_{\rm max}$ defined above can be expressed just in terms of $R_{\rm s}$ as, $d_{\rm max} = R_{\rm s}2^{5/6}(5+\sqrt{5})^{1/2}(\pi/[5+(3+\sqrt{5})])^{1/3} \approx 2.364R_{\rm s}$. Alternatively, we empirically found that the value of the diameter of the inscribed sphere $2R_{\rm inscr}$ [see Fig. S2(a)] can be also quite accurately approximated by a combination of the size parameters determined

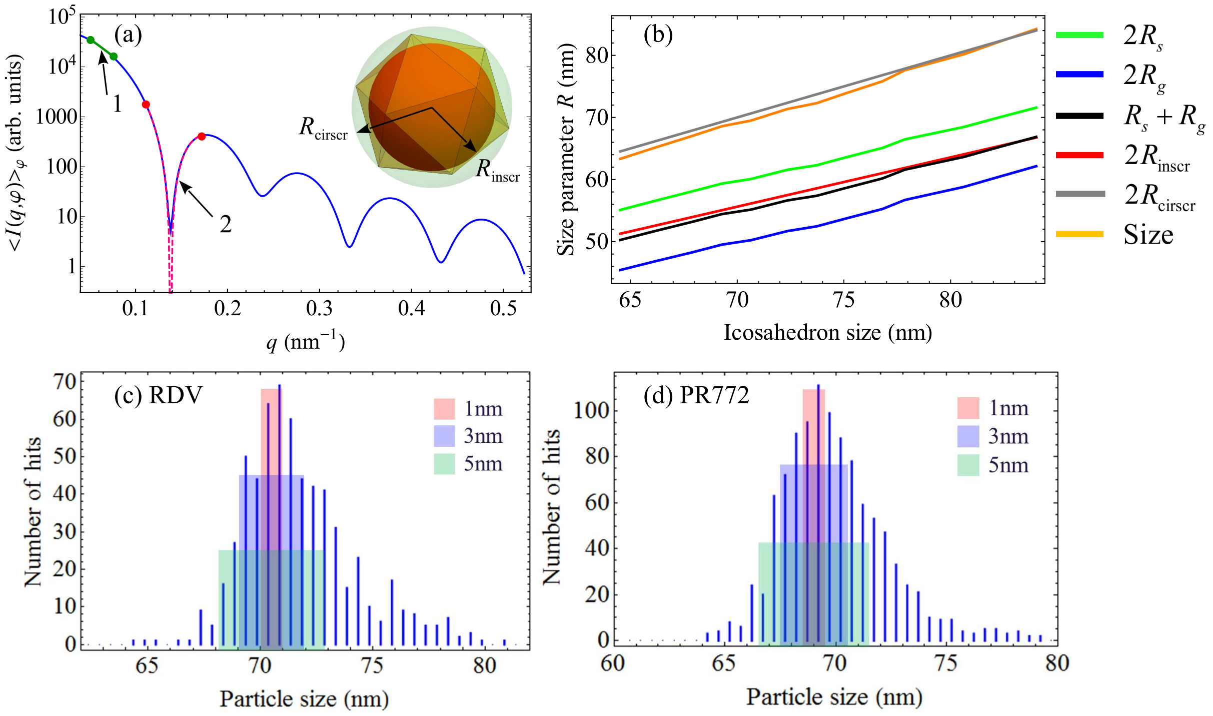

FIG. S2. Size determination of the virus particles. (a) Example of a radial intensity profile $\langle I(q,\varphi)\rangle_\varphi$ of an icosahedral particle approximated by the Guinier-type fit (1) in the low-q range, and fitted with a form factor of a spherical particle (2) around the position of the first minimum of $\langle I(q,\varphi)\rangle_\varphi$. The inset in (a) schematically shows an icosahedron with the specified radii of inscribed R_{inscr} and circumscribed R_{cirscr} spheres. (b) Dependence of the fitting parameters R_{s} and R_{g}, as well as other size parameters on the size of icosahedron, used in the size-determination procedure (see text). (c),(d) Size distribution histograms determined for the experimental RDV (c) and PR772 (d) diffraction patterns. Data portions corresponding to polydispersity PD = 1 nm, 3 nm and 5 nm are indicated by different shaded areas in (c) and (d).

from the two fits as $2R_{\text{inscr}} = R_{\text{g}} + R_{\text{s}}$ [Fig. S2(b)]. By using the exact geometric relation $R_{\text{cirscr}} = (15 - 6\sqrt{5})^{0.5} R_{\text{inscr}}$, the size of an icosahedral particle can be quite accurately determined from the fitted parameters as, size $\approx 1.26(R_{\text{g}} + R_{\text{s}})$. As one can see from Fig. S2(b), the particle size that is determined from this approach is in very good agreement with the exact value of $2R_{\text{cirscr}}$.

The size distribution histograms determined as size $\approx 1.26(R_{\text{g}} + R_{\text{s}})$, with the maximum allowed root-mean-square (RMS) errors for the spherical form-factor fits $\text{RMS}^{\text{s}}_{\text{RDV}} = 0.2$ and $\text{RMS}^{\text{s}}_{\text{PR772}} = 0.12$, and for the Guinier fits $\text{RMS}^{\text{g}}_{\text{RDV}} = 0.02$ and $\text{RMS}^{\text{g}}_{\text{PR772}} = 0.01$, are shown in Figs. S2(c) and S2(d) for RDV and PR772 viruses, respectively. After rejecting additional patterns using this RMS filtering of the fits, the histograms show size distributions for the remaining 619 RDV hits and 1058 PR772 hits. As one can see, the size distributions for both viruses have a Gaussian-like shape, with a bit steeper left wing due to missing "low-intensity" hits from smaller particles, which were rejected on the initial preprocessing stage.

X-RAY CROSS-CORRELATION ANALYSIS

The scattered intensity distribution from a single particle in an arbitrary orientation can be expressed as

$$I(\mathbf{q}) = \left| \int \rho(\mathbf{r}) \exp(\mathrm{i}\mathbf{q} \cdot \mathbf{r}) d\mathbf{r} \right|^2 . \tag{1}$$

where $\mathbf{q}$ is the scattering vector, $\mathbf{r}$ is the real-space vector, ρ is the 3D electron density distribution of the particle. A diffraction pattern measured on a 2D detector samples the 3D intensity distribution in Eq. (1) along the Ewald sphere, and can be expanded into the angular Fourier series

$$I(q, \varphi) = \sum_{n=-\infty}^{\infty} I^n(q) \exp(\mathrm{i}n\varphi), \tag{2}$$

where the angular Fourier transform is defined in the polar coordinate system of the detector (q, φ), and $I^n(q)$ are the Fourier components (FCs) of $I(q, \varphi)$.

The basic element of x-ray cross-correlation analysis (XCCA) is the two-point cross-correlation function (CCF), which can be defined at two momentum transfer values q_1 and q_2 as [3–5],

$$C_{ij}(q_1, q_2, \Delta) = \langle I_i(q_1, \varphi) I_j(q_2, \varphi + \Delta) \rangle_{\varphi} , \tag{3}$$

where Δ is the angular separation, $\langle \ldots \rangle_{\varphi}$ defines averaging over the angular coordinate φ, and the subscripts i and j indicate that intensities are correlated between the i-th and j-th diffraction patterns. It is customary to operate with the FCs $C_{ij}^n(q_1, q_2)$ of the CCF (3), with angular Fourier series of $C_{ij}(q_1, q_2, \Delta)$ written as

$$C_{ij}(q_1, q_2, \Delta) = \sum_{n=-\infty}^{\infty} C_{ij}^n(q_1, q_2) \exp(\mathrm{i}n\Delta). \tag{4}$$

It has been shown that the following relation holds between the FCs of intensity and of the CCF $C_{ij}^n(q_1,q_2)$ [3–5],

$$C_{ij}^n(q_1,q_2) = I_i^{n*}(q_1) I_j^n(q_2). \tag{5}$$

As one can see, the FCs $C_{ij}^n(q_1,q_2)$ are directly related to the FCs of intensity $I_i^n(q_1)$ and $I_j^n(q_2)$.

Eqs. (3) and (5) reduce to a commonly used single-diffraction-pattern CCF $C_{ii}(q_1,q_2,\Delta)$ and its FCs $C_{ii}^n(q_1,q_2)$ when intensities are correlated on the same diffraction pattern i. Also, the CCF can be calculated for a single momentum transfer $q_1 = q_2 = q$, with Eqs. (3) and (5) further reducing to $C_{ii}(q,\Delta) = \langle I_i(q,\varphi) I_i(q,\varphi+\Delta)\rangle_\varphi$ and $C_{ii}^n(q) = |I_i^n(q)|^2$.

The CCF and its FCs can be averaged over a set of M diffraction patterns to obtain the orientationally averaged result,

$$\langle C_{ii}(q_1,q_2,\Delta)\rangle_i = \frac{1}{M}\sum_{i=1}^{M} C_{ii}(q_1,q_2,\Delta), \tag{6}$$

$$\langle C_{ij}(q_1,q_2,\Delta)\rangle_{i\neq j} = \frac{1}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} C_{ij}(q_1,q_2,\Delta), \tag{7}$$

$$\begin{aligned}\langle C_{ii}^n(q_1,q_2)\rangle_i &= \frac{1}{M}\sum_{i=1}^{M} C_{ii}^n(q_1,q_2)\\ &= \frac{1}{M}\sum_{i=1}^{M} I_i^{n*}(q_1) I_i^n(q_2),\end{aligned} \tag{8}$$

$$\begin{aligned}\langle C_{ij}^n(q_1,q_2)\rangle_{i\neq j} &= \frac{1}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} C_{ij}^n(q_1,q_2)\\ &= \frac{1}{M(M-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{M} I_i^{n*}(q_1) I_j^n(q_2),\end{aligned} \tag{9}$$

where $\langle\ldots\rangle_i$ and $\langle\ldots\rangle_{i\neq j}$ denote statistical averages over M patterns and $M(M-1)$ pairs of diffraction patterns, respectively. Notice, that due to symmetry properties of $C_{ij}(q_1,q_2,\Delta)$, for a set of M measured diffraction patterns there are $M(M-1)$ nonequivalent pairs of patterns for $q_1 \neq q_2$, and $M(M-1)/2$ pairs for $q_1 = q_2 = q$.

For practical application, it is useful to define the difference CCF

$$\widetilde{C}(q_1,q_2,\Delta) = \langle C_{ii}(q_1,q_2,\Delta)\rangle_i - \langle C_{ij}(q_1,q_2,\Delta)\rangle_{i\neq j} \tag{10}$$

and its Fourier series,

$$\widetilde{C}(q_1, q_2, \Delta) = \sum_{n=-\infty}^{\infty} \widetilde{C}^n(q_1, q_2) \exp(\mathrm{i}n\Delta). \tag{11}$$

Due to the linear properties of the Fourier transform, the average difference Fourier spectrum can be also determined as [5],

$$\widetilde{C}^n(q_1, q_2) = \langle C^n_{ii}(q_1, q_2)\rangle_i - \langle C^n_{ij}(q_1, q_2)\rangle_{i\neq j}. \tag{12}$$

The key property of the ensemble averaged $\langle C_{ii}(q_1, q_2, \Delta)\rangle_i$ and $\langle C^n_{ii}(q_1, q_2)\rangle_i$ is that they preserve higher-order structural information of the 3D structure of a particle. In contrast, such information cannot be accessed in conventional SAXS analysis, where only orientationally averaged intensity $\langle I_i(q, \varphi)\rangle_{\varphi,i}$ is measured. In the absence of any background and uniform distribution of particle orientations the term $\langle C^n_{ij}(q_1, q_2)\rangle_{i\neq j}$ vanishes for $n > 0$, therefore $\widetilde{C}^n(q_1, q_2)$ should contain undistorted information about a particle, apart from the $n = 0$ term which can be recovered from the SAXS pattern. In the presence of nonuniformities in the measured data (for instance, structured background, nonuniform response of detector tiles, etc.), the difference FCs $\widetilde{C}^n(q_1, q_2)$ help to reduce the effect of various undesirable experimental factors that can contaminate $\langle C^n_{ii}(q_1, q_2)\rangle_i$ [5].

The ability of the difference FCs to filter out undesirable experimental factors is illustrated in Fig. S3, where the Fourier components $\langle C^n_{ii}(q_1, q_2)\rangle_i$ and $\widetilde{C}^n(q_1, q_2)$ are plotted for the case $q_1 = q_2 = q$, both for PR772 and RDV experimental data. It is clearly visible that the difference Fourier components $\widetilde{C}^n(q_1, q_2)$ look much cleaner than $\langle C^n_{ii}(q_1, q_2)\rangle_i$, and the FCs of even orders n can be clearly distinguished from the odd orders, which have vanishing values in the small-angle scattering geometry of our experiment [1].

SIMILARITY METRIC FOR CORRELATION DATA

While 2D maps of correlation coefficients $\widetilde{C}^n(q_1, q_2)$ provide convenient visual means for comparison of different experimental data and simulations, from a practical point of view, it is important to have a quantitative measure of similarity for correlation data. Here we adopted a similarity metric for correlation datasets based on the idea of Fourier ring correlation (FRC), which is commonly used as a resolution-dependent metric for comparison of 2D images in electron microscopy and x-ray imaging [6–8]. To compare 2D maps of the

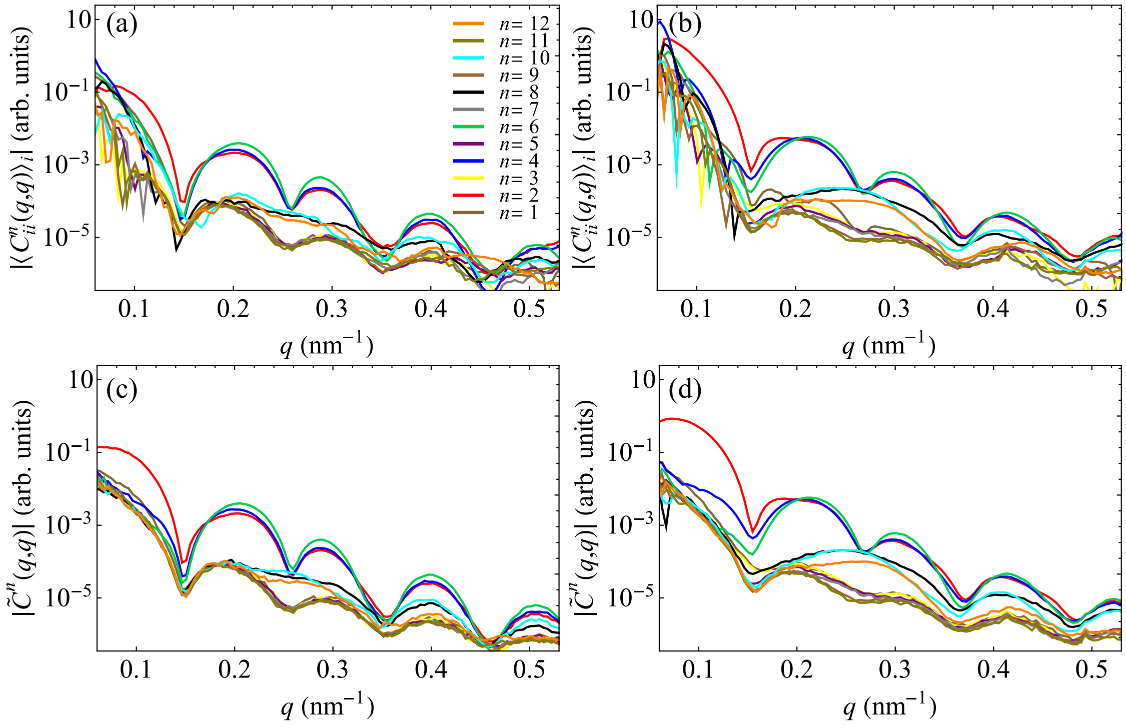

FIG. S3. Amplitudes (log scale) of the FCs (a),(b) $\langle C^n_{ii}(q_1, q_2)\rangle_i$ [see Eq. (8)] and (c),(d) $\widetilde{C}^n(q_1, q_2)$ [Eq. (12)] for $n = 1, \ldots, 12$ determined at $q_1 = q_2 = q$ for (a),(c) RDV and (b),(d) PR772 viruses, respectively.

Fourier components $\widetilde{C}^n(q_1, q_2)$ [Eq. 12] of different orders n, we propose to use the Fourier quadrant correlation (FQC) between two maps defined as

$$\mathrm{FQC}^n(q) = \frac{|CC^n_{1,2}(q)|}{\sqrt{CC^n_{1,1}(q) CC^n_{2,2}(q)}}, \tag{13}$$

where

$$CC^n_{v,w}(q) = \sum_{q_1 \leq q} \widetilde{C}^n_v(q_1, q) \cdot \widetilde{C}^n_w(q_1, q)^* + \sum_{q_2 < q} \widetilde{C}^n_v(q, q_2) \cdot \widetilde{C}^n_w(q, q_2)^*, \tag{14}$$

and $\widetilde{C}^n_v(q_1, q_2)$ [Eq. 12] defines the FC of the n-th order corresponding to the v-th map ($v, w = 1, 2$). The two summations in Eq. (14) are performed for each q over two orthogonal sections of the 2D maps that form edges of a quadrant (see Fig. S4), hence defining the name "FQC" of the similarity metric. Such a choice of the similarity metric is justified by the rather rectangular symmetry of the 2D correlation maps (compared to the rather circular symmetry of diffraction patterns) and also by its independence of experimental geometry (similar to classical FRC used for diffraction patterns).

Note that due to imperfections in the experimental data (e.g, limited statistics) the FCs $\widetilde{C}^n(q_1, q_2)$ are not precisely real-valued. This is reflected in Fig. S5, where the phases of $\widetilde{C}^n(q_1, q_2)$ are shown, determined for the experimental RDB and PR772 datasets for particle polydispersity PD $= 3\,\mathrm{nm}$ (see next section for polydispersity effects). For this reason, Eq. (14) is in general defined for complex values of $\widetilde{C}^n(q_1, q_2)$.

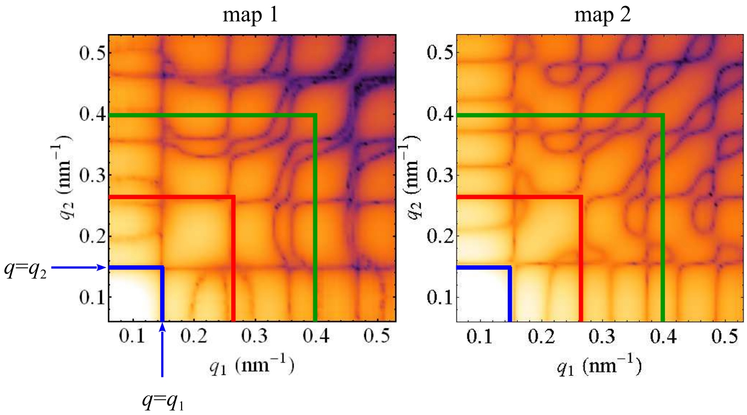

FIG. S4. Quadrants used for the summations in the definition of the FQC, as given by Eqs. (13) and (14). The two summations in Eq. (14) are performed for each q over two orthogonal sections, $q_1 = q$, $q_2 \leq q$ and $q_1 < q$, $q_2 = q$, forming edges of a quadrant. Edges of three quadrants corresponding to different values of q in Eq. (14) are shown in different color.

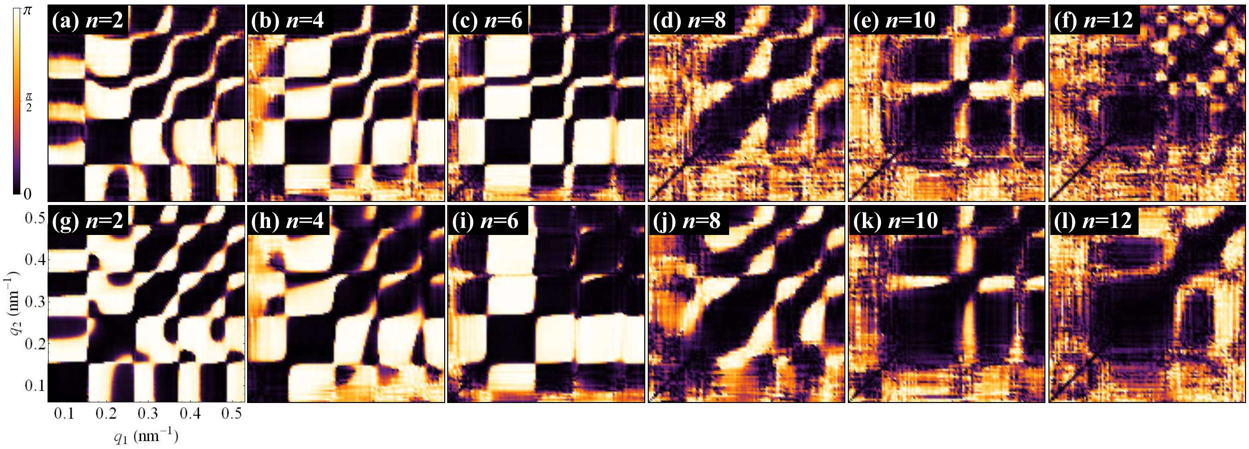

FIG. S5. Experimental 2D maps of the absolute values of phases $|\arg[\widetilde{C}^n(q_1, q_2)]|$ for $n = 2, 4, 6, 8, 10$ and 12, determined for (a)-(f) RDV and (g-l) PR772 viruses.

While the similarity metric defined in Eq. (13) can be used to compare FCs of different orders n separately, it cannot be used to characterize the correlation dataset as a whole because FCs of different orders n can have different orders of magnitude. For single-particle structure recovery by the MTIP algorithm, we use the whole correlation dataset including SAXS intensities. Therefore, we also defined a cumulative correlation metric $CC(q_1, q_2)$ in terms of the difference CCF $\widetilde{C}(q_1, q_2, \Delta)$ [Eq. (10)] and SAXS intensities $\langle I_i(q, \varphi)\rangle_{\varphi,i}$,

$$CC(q_1, q_2) = \frac{CC_{1,2}(q_1, q_2)}{\sqrt{CC_{1,1}(q_1, q_2)CC_{2,2}(q_1, q_2)}}, \tag{15}$$

where

$$\begin{aligned} CC_{v,w}(q_1, q_2) =& \langle \widetilde{C}_v(q_1, q_2, \Delta)\widetilde{C}_w(q_1, q_2, \Delta)\rangle_\Delta \\ &+\langle I_i^v(q_1, \varphi)\rangle_{\varphi,i}\langle I_i^v(q_2, \varphi)\rangle_{\varphi,i}\langle I_i^w(q_1, \varphi)\rangle_{\varphi,i}\langle I_i^w(q_2, \varphi)\rangle_{\varphi,i}, \end{aligned} \tag{16}$$

and the CCF $\widetilde{C}_v(q_1, q_2, \Delta)$ and SAXS intensity $\langle I_i^v(q, \varphi)\rangle_{\varphi,i}$ are specified for the v-th dataset $(v, w = 1, 2)$. The two-dimensional metric $CC(q_1, q_2)$ defined in Eq. (15) can be averaged over the quadrants to produce a 1D FQC as a function of q,

$$\mathrm{FQC}(q) = \frac{1}{N(q)}\left(\sum_{q_1 \le q} CC(q_1, q) + \sum_{q_2 < q} CC(q, q_2)\right), \tag{17}$$

where $N(q) = \sum_{q_1 \le q} 1 + \sum_{q_2 < q} 1$ is the number of sampled (q_1, q_2) pairs in the quadrant associated to q, which is described above [see Fig. S4]. The FQC defined in Eq. (17) can be used to compare entire correlation datasets, including SAXS intensities.

We should note, that there is no direct correspondence between the classical FRC, which is a linear function of the momentum transfer q, and the FQC introduced here, which has a nonlinear dependence on q. At each q both $\mathrm{FQC}^n(q)$ [Eq. (13)] and $\mathrm{FQC}(q)$ [Eq. (17)] are determined as an average over FCs or CCFs which are functions of two arguments, q_1 and q_2. However, our results show that the FQC-type metric can be used in a similar manner (to classical FRC for diffraction data) to estimate the data quality and similarity between different sets of correlation data. As an example, Fig. S6 shows the results of application of Eqs. (13), (15) and (17) to the RDV and PR772 datasets determined for particle polydispersity PD = 3 nm (the data used for structure analysis in the main text). As one can see, the FQC metric shows that the structures are indeed substantially different even at low resolution, in agreement with the visual observations of the 2D correlation maps in Fig. 2 of the main text.

It should also be stressed that when this metric is used to compare random-half datasets, it only measures the influence of random noise on the data, and doesn't gauge the impact that systematic effects, such as inadequate masking of bad pixels or incorrect background subtraction, have on the quality of the dataset.

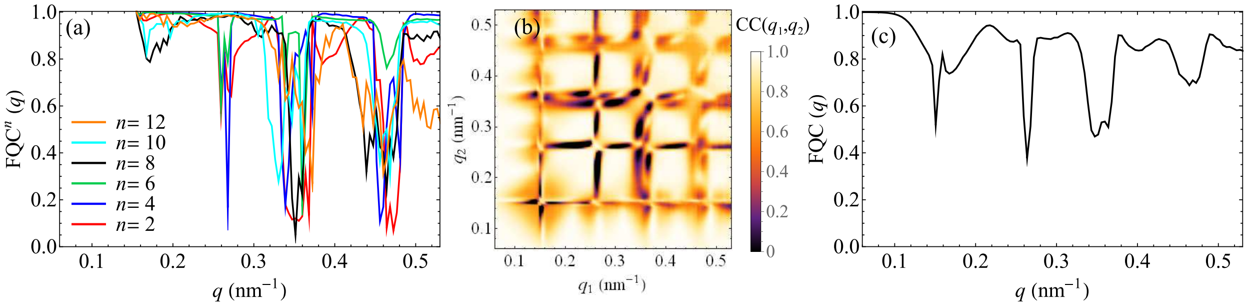

FIG. S6. Similarity analysis for the RDV and PR772 datasets at PD = 3 nm. (a) $\mathrm{FQC}^n(q)$ [Eq. (13)] for $n = 2, 4, 6, 8, 10$ an 12, (b) $CC(q_1, q_2)$ [Eq. (15)] and (c) $\mathrm{FQC}(q)$ [Eq. (17)]. All three similarity metrics reach values $\ll 1$ indicating substantial differences between the RDV and PR772 structures even at low resolution.

POLYDISPERSITY EFFECTS

To analyze polydispersity (PD) effects, for each experiment we selected three subsets of diffraction patterns corresponding to PD = 1 nm, 3 nm and 5 nm, as shown in Figs. S2(c) and S2(d) by shaded areas of different colors. The correlation data were averaged over a different number M of diffraction patters for different PD, i.e, $M = 132$ (PD = 1 nm), $M = 332$ (PD = 3 nm), and $M = 459$ (PD = 5 nm) in the case of RDV, and $M = 217$ (PD = 1 nm), $M = 566$ (PD = 3 nm), and $M = 796$ (PD = 5 nm) in the case of PR772. The experimental correlation maps corresponding to different PD are shown in Figs. S7 and S8 for RDV and PR772, respectively. One can clearly see that correlation maps for each of the viruses look very similar at different degrees of polydispersity. For instance, the features attributed to a 3% distortion (which corresponds to about 2 nm) of the RDV particle are still perfectly preserved on the maps even for PD = 5 nm [Fig. S7]. Even at high degree of polydispersity the correlation maps still contain features characteristic of each of the virus particles.

We also performed similarity analysis using the proposed FQC metric to quantify the differences between the results shown in Figs. S7 and S8. The results for the datasets corresponding to different PD are shown in Figs. S9 and S10 for RDV and PR772, respectively.

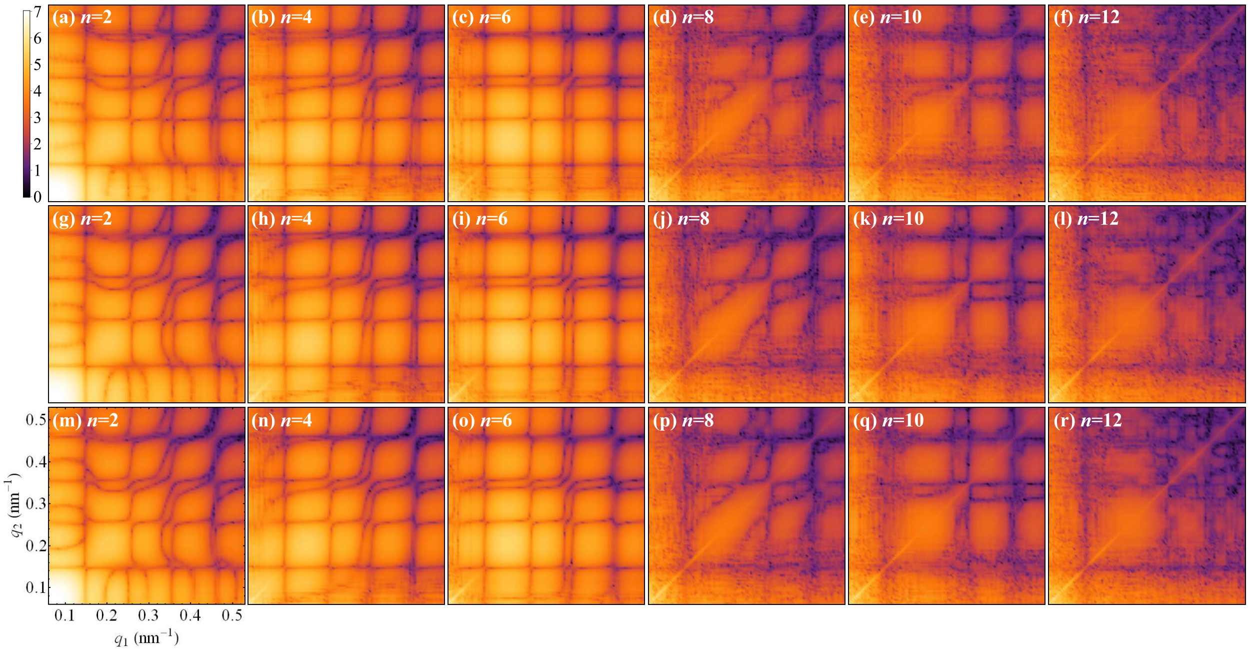

FIG. S7. Experimental amplitudes (log scale) of the FCs $|\widetilde{C}^n(q_1, q_2)|$ [Eq. (12)] for $n = 1, \ldots, 12$ determined for RDV with polydispersity of (a)-(f) PD = 1 nm, (g)-(l) PD = 3 nm, and (m)-(r) PD = 5 nm, respectively. Portions of the data corresponding to different PD are indicated in Fig. S2(c).

In Figs. S9(g)-(i) and S10(g)-(i) we show the similarity of two split subsets (containing equal number of patterns) of data for PD = 3 nm for RDV and PR772, respectively. We would like to note that in all calculations of FQC$^n(q)$ [Eq. (13)] we excluded from the analysis the low-q region ($q_1, q_2 < 0.16$ nm^{-1}) on each 2D map, to improve the visibility of the high-q data. In contrast, the low-q region is included in the calculations of $CC(q_1, q_2)$ [Eq. (15)] and FQC(q) [Eq. (17)]. Even if the low-q range is masked for the correlation data (but not for the SAXS terms) in the calculation of $CC(q_1, q_2)$ and FQC(q), the results look almost identical to those without masking. This indicates that SAXS terms make dominant contribution to $CC(q_1, q_2)$and FQC(q) at low q.

Inspection of the results shown in Figs. S9 and S10 allows us to make the following conclusions. In general, major differences between the correlation data for different PD are caused by the FCs $\widetilde{C}^n(q_1, q_2)$ of higher orders $n = 8, 10$ and 12, as evident from Figs. S9(a),(d),(g) and S10(a),(d),(g). Relatively small differences between the lower-order FCs ($n = 2, 4$ and 6) for different PD are observed mostly in the locations where the corresponding 2D

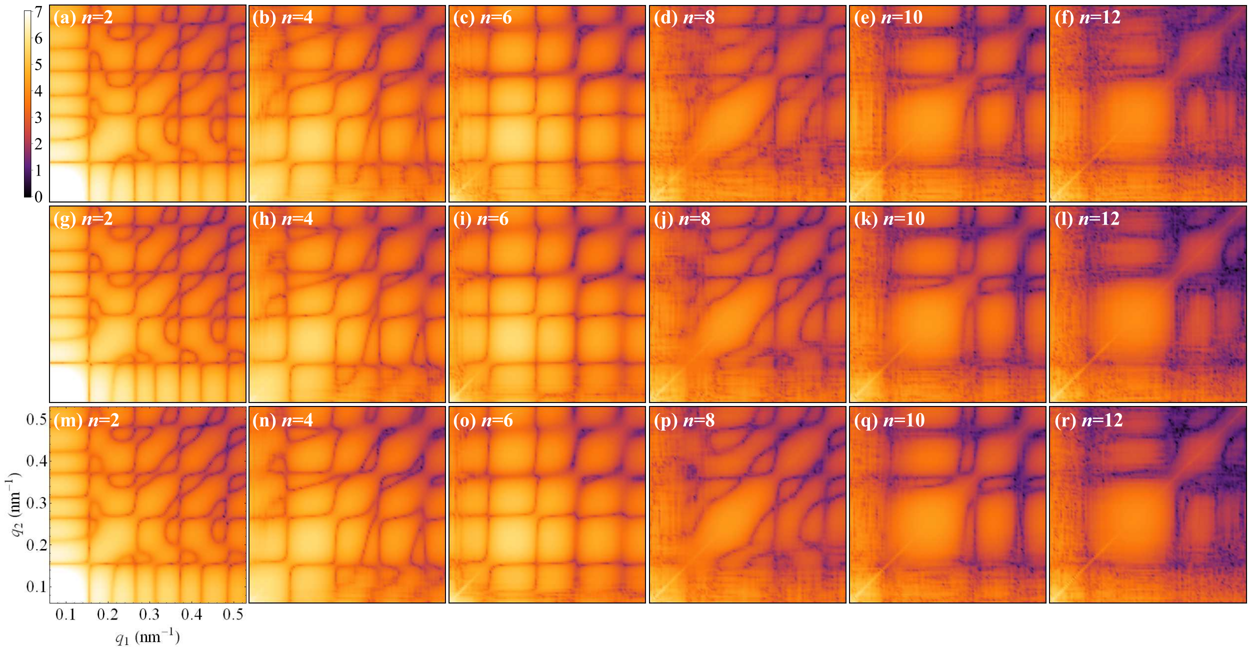

FIG. S8. Experimental amplitudes (log scale) of the FCs $|\widetilde{C}^n(q_1, q_2)|$ [Eq. (12)] for $n = 1, \ldots, 12$ determined for PR772 with polydispersity of (a)-(f) PD = 1 nm, (g)-(l) PD = 3 nm, and (m)-(r) PD = 5 nm, respectively. Portions of the data corresponding to different PD are indicated in Fig. S2(d).

maps of $\widetilde{C}^n(q_1, q_2)$ have minima. Such behavior is also typical for the metric $CC(q_1, q_2)$ [Figs. S9(b),(e),(h) and S10(b),(e),(h)] attributed to the whole correlation dataset, where the minima of the metric are observed at the same positions as the minima of the correlation maps [compare with Figs. S7 and S8]. Clearly, such behavior can be explained by limited statistics of the present experiment. However, noisier higher-order FCs have a comparably smaller contribution to $CC(q_1, q_2)$ and FQC(q), which leads to a substantial similarity of the whole correlation datasets at different PD [see Figs. S9(c),(f),(i) and S10(c),(f),(i)]. One may note that, according to the FQC(q), there is more similarity between the datasets for PD = 1 nm and PD = 3 nm [Figs. S9(f) and S10(f)], than between PD = 1 nm and PD = 5 nm [Figs. S9(c) and S10(c)]. This can be attributed to a different number of patterns in the datasets with different PD, and to the effect of polydispersity itself, which is hard to distinguish due to the limited statistics of our experiment. General comparison of the results for RDV [Fig. S9] and PR772 [Fig. S10] suggests that the major difference between results for different PD has statistical origin, since all metrics have systematically higher values for

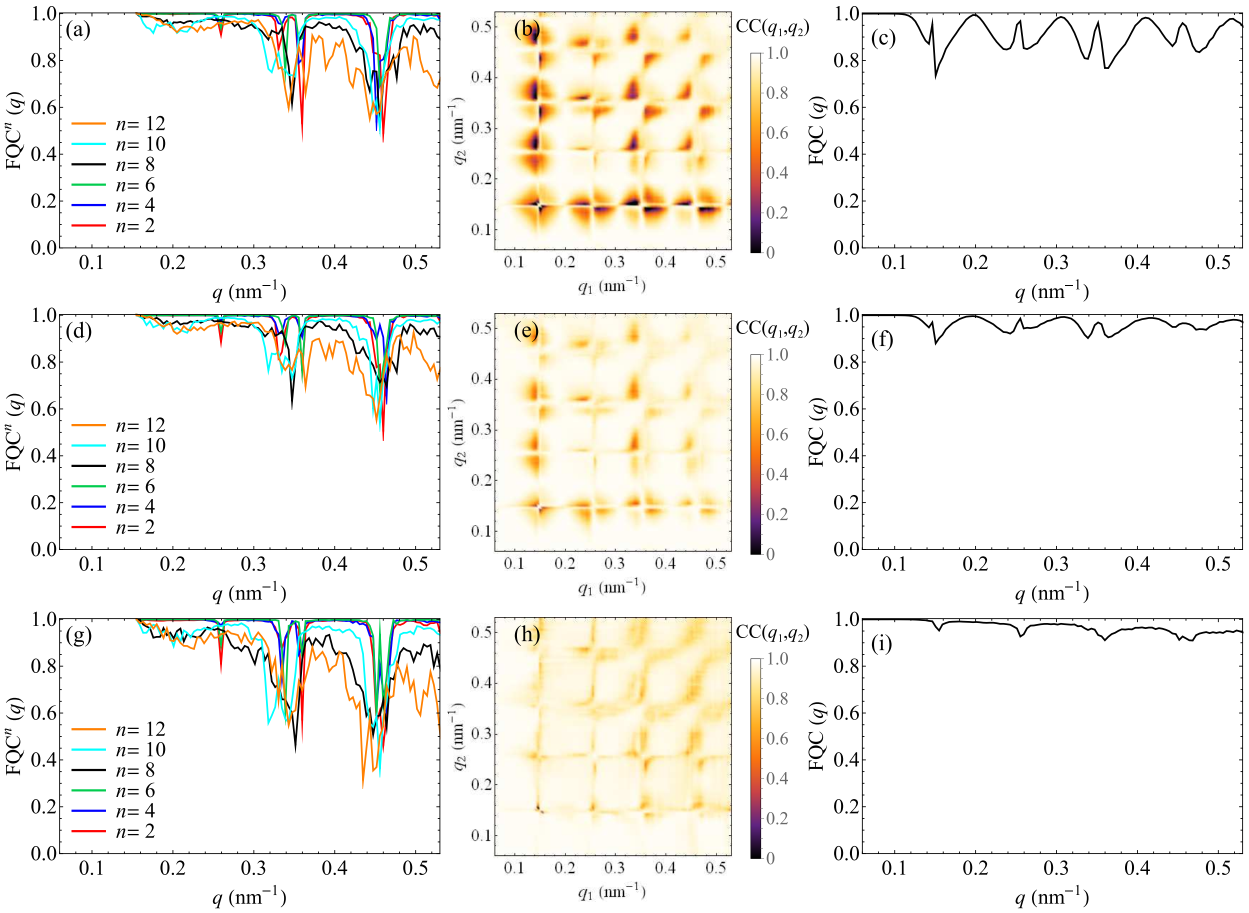

FIG. S9. Results of similarity analysis for RDV, showing (left column) FQC$^n(q)$ [Eq. (13)] for $n = 2, 4, 6, 8, 10$ an 12, (middle column) $CC(q_1, q_2)$ [Eq. (15)] and (right column) FQC(q) [Eq. (17)]. A pairwise comparison is done for the correlation data corresponding to (a)-(c) PD = 1 nm and PD = 5 nm, (d)-(f) PD = 1 nm and PD = 3 nm, and (g)-(i) two subsets of PD = 3 nm.

PR772, for which more diffraction patterns were available. For the major analysis of the present work we have chosen the datasets corresponding to PD = 3 nm, as a compromise between possible polydispersity effects and statistical issues.

The results of this section suggest that cross-correlation data can still be used to analyze particle structure by scattering from a system of N reproducible particles with a limited degree of polydispersity. This would enable one to exploit the full potential of the FXS approach and to go beyond the single-particle imaging scheme.

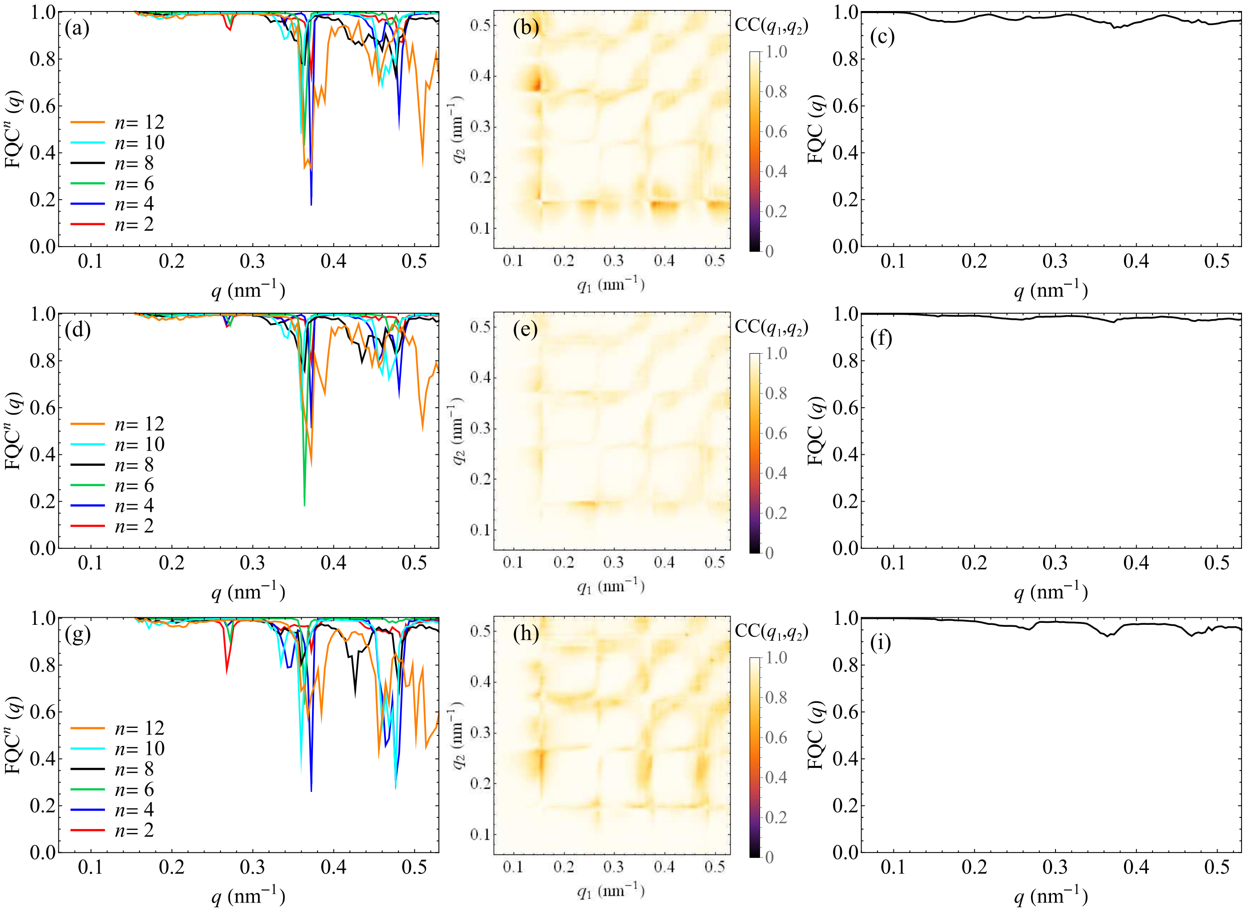

FIG. S10. Results of similarity analysis for PR772, showing (left column) FQC$^n(q)$ [Eq. (13)] for $n = 2, 4, 6, 8, 10$ an 12, (middle column) $CC(q_1, q_2)$ [Eq. (15)] and (right column) FQC(q) [Eq. (17)]. A pairwise comparison is done for the correlation data corresponding to (a)-(c) PD = 1 nm and PD = 5 nm, (d)-(f) PD = 1 nm and PD = 3 nm, and (g)-(i) two subsets of PD = 3 nm.

MODEL COMPARISON

To understand the correlation maps determined from the experimental data for RDV and PR772 (see Fig. 2 in the main text) we performed simulations using bead models of various structures, with a bead diameter of 1 nm and average electron density of 0.325 electrons/$\AA$. We also did simulations with the empty RDV capsid atomic structure determined at 3.5 $\AA$ resolution by x-ray crystallography [Protein Data Bank (PDB) entry 1UF2] [9], and reduced to a resolution matching the resolution of the single particle data of our experiment. Our simulations of x-ray diffraction were performed with parameters similar to those of the experiment [1]. The simulated 2D maps of the amplitudes of the FCs $|\widetilde{C}^n(q_1, q_2)|$ for several model particles possessing icosahedral symmetry are shown in Fig. S11. Direct comparison of these maps with the experimental results (see Fig. 2 in main text) shows that none of these model particles can adequately reproduce the experimental data for RDV or PR772. We then applied simple types of distortions to the ideal empty RDV capsid and bead model of a solid icosahedral particle, with the results shown in Fig. S12. One may notice that the correlation maps for the FCs of the 2-nd order shown in Figs. S12(g) and S12(m) are in very good agreement with the experimental maps shown in Figs. 2(a) and 2(g) of the main text for RDV and PR772, correspondingly. Our simulations show that such similarity observed for FCs ($n = 2$) can be explained by the deviations of the particle shape from an exact icosahedron, as well as by particle "caking" induced during buffer evaporation (see Fig. S13). For example, the 2D maps for $n = 2$ can be quite closely reproduced in the case of ellipsoidal caking of ideal icosahedral particle, with a longer axis of the ellipsoid coinciding with one of the five-fold symmetry axis of an icosahedron [compare Figs. S13(a) and S13(g) with the experimental results in Fig. 2(a) and 2(g) in main text, for RDV and PR772 respectively]. Also, observed similarity can be a result of a combined effect of both article distortion and caking [compare Figs. S13(a), S13(s) and S12(g)]. At the same time we were not able to reproduce the observed characteristic features for FCs of the 2-nd order with more symmetric model of spherical caking [compare Figs. S13(m) with S13(a) and S13(s)].

All other FCs of higher orders ($n > 2$) cannot be accurately reproduced by simple distortions, and require a more advanced modeling approach which goes beyond a uniform density approximation applied in our bead modeling. Instead of doing such sophisticated modeling, we perform *ab initio* reconstructions of the virus structures by applying the MTIP algorithm to the experimental correlation data (see next section). However, we would like to note that the correlation maps can still be used for model based comparison as illustrated in this section, to get an idea about the possible particle structure. The 2D correlation maps can be especially useful for following fast dynamical changes in the structure, for instance, as a response to external stimulus, which is a key component of structural studies at XFELs [see supplementary gif-animations showing evolution of the FCs $|\widetilde{C}^n(q_1, q_2)|$ of orders $n = 2, 4, 6, 8, 10$ and 12 during uni-axial distortion of an ideal icosahedral particle (icosahedron_distortion.gif) and caking of the ideal icosahedral particle during solution evaporation (icosahedron_caking.gif)].

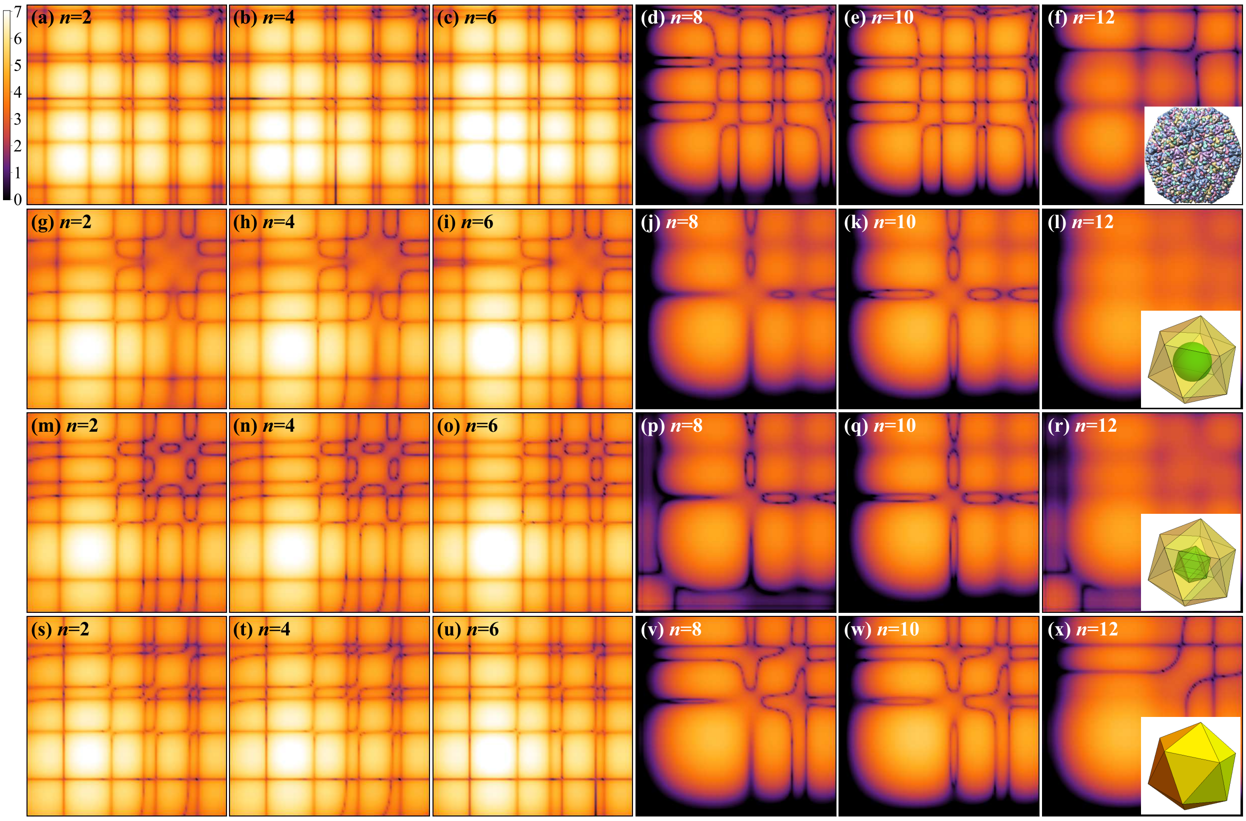

FIG. S11. Simulated 2D maps (log scale, arb. units) of the amplitudes of the FCs $|\widetilde{C}^n(q_1, q_2)|$ for $n = 2, 4, 6, 8, 10$ and 12 for ideal icosahedral particles. The results are shown for the atomistic model of empty RDV capsid (a)-(f), as well as for bead models of a hollow icosahedral particle of 71 nm in size with a spherical void of a diameter $d = 30$ nm (g)-(l), with an icosahedral void of size $d = 30$ nm (m)-(r), and a solid icosahedral particle of 71 nm in size (s)-(x). The corresponding particles are schematically shown in (f),(l),(r) and (x).

STRUCTURE RECOVERY BY THE MTIP ALGORITHM

In addition to the modeling discussed above, we obtain *ab initio* reconstructions using the multi-tiered iterative phasing (MTIP) algorithm introduced in ref. [10]. This approach to structure determination from correlation data is based on the theory developed by Kam in ref. [11], where it is shown that the angular correlations can be directly related to the spherical harmonic expansion of the 3D intensity function, given by

$$I(q, \theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} I_{lm}(q) Y_l^m(\theta, \phi). \tag{18}$$

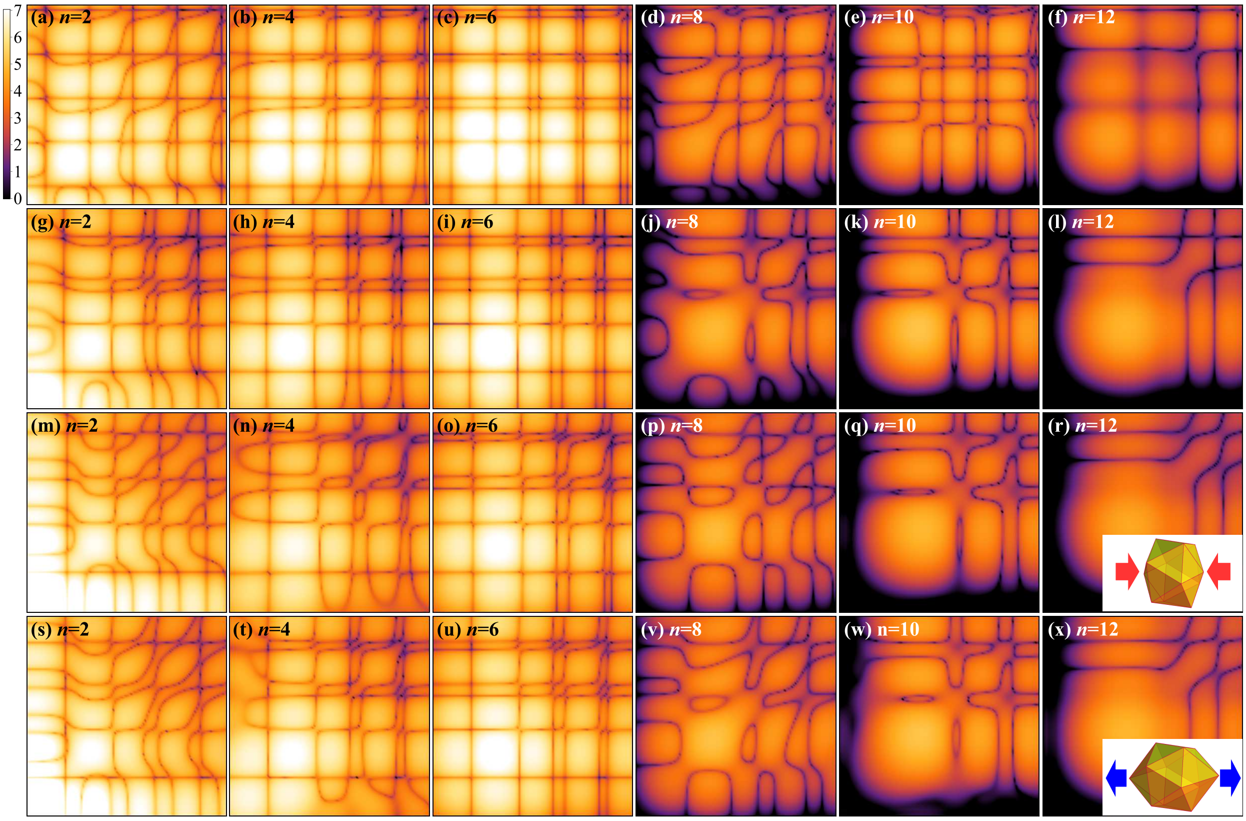

FIG. S12. Simulated 2D maps (log scale, arb. units) of the amplitudes of the FCs $|\widetilde{C}^n(q_1, q_2)|$ for $n = 2, 4, 6, 8, 10$ and 12 for distorted icosahedral particles. The results are shown for the atomistic model of empty RDV capsid compressed by 3 %, as well as for bead models of a solid icosahedral particle of 71 nm in size, compressed by 3 % (g)-(l), compressed by 7 % (m)-(r), and extended by 7 % (s)-(x) relative to the initial size of an undistorted particle. The applied compressive and extensive distortions are schematically shown in (r) and (x).

The average correlation function can be shown to be related to the $I_{lm}(q)$ expansion coefficients via the Legendre decomposition [in order to simplify the following presentation, hereafter we denote the theoretical orientationally averaged cross-correlation function simply as $C(q_1, q_2, \Delta)$]

$$C(q_1, q_2, \Delta) = \sum_{l=0}^{\infty} P_l(\cos\theta(q_1)\cos\theta(q_2) + \sin\theta(q_1)\sin\theta(q_2)\cos\Delta) B_l(q_1, q_2), \tag{19}$$

where

$$B_l(q_1, q_2) = \sum_{m=-l}^{l} I_{lm}(q_1) I^*_{lm}(q_2), \tag{20}$$

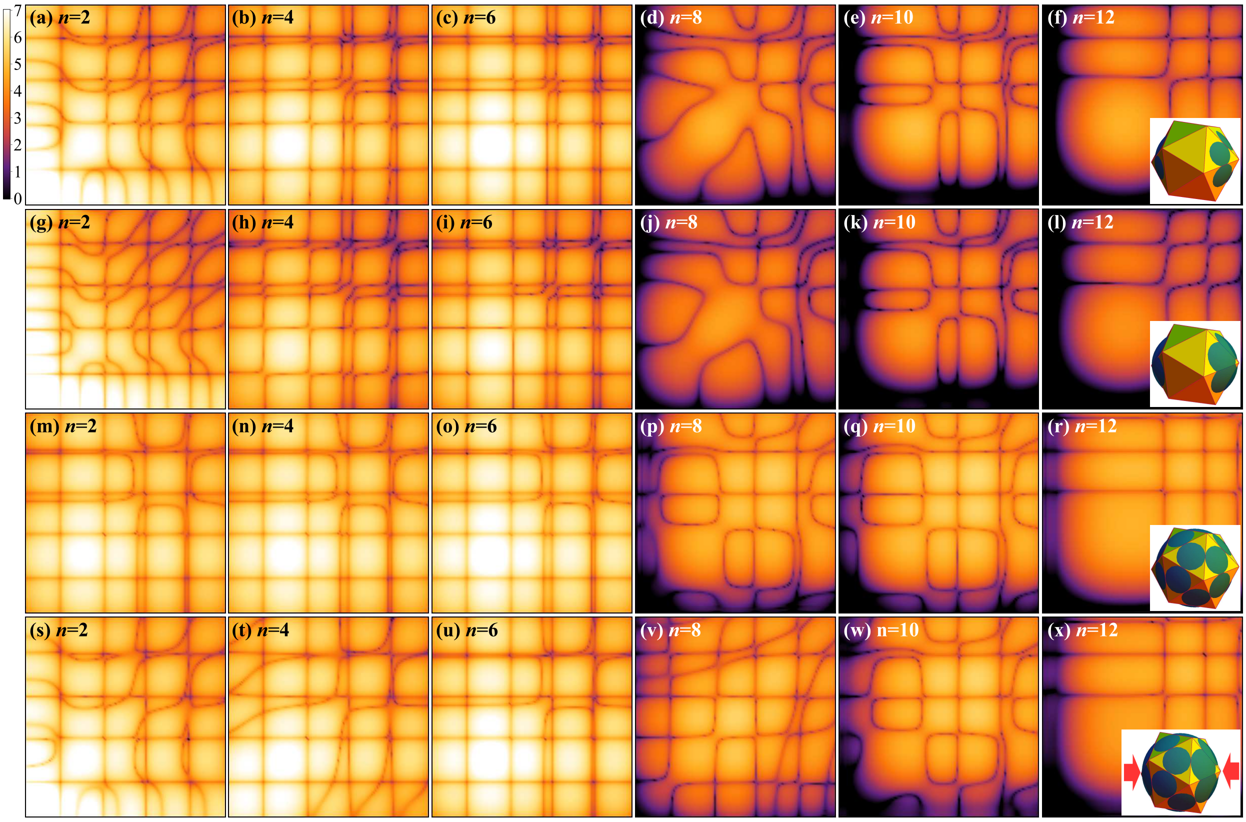

FIG. S13. Simulated 2D maps (log scale, arb. units) of the amplitudes of the FCs $|\widetilde{C}^n(q_1, q_2)|$ for $n = 2, 4, 6, 8, 10$ and 12 for icosahedral particles with different types of caking. The results are shown for bead models of an ideal solid icosahedral particle of 71 nm in size with ellipsoidal caking with ellipsoid semiaxes $a = b = 29.5$ nm, $c = 31.5$ nm (a)-(f) and ellipsoid semiaxes $a = b = 29.5$ nm, $c = 32.5$ nm (g)-(l), with spherical caking of radius $r = 31.5$ nm (m)-(r), and for icosahedral particle compressed by 7.5 % with spherical caking of radius $r = 31.5$ nm (s)-(x). In the case of ellipsoidal caking the longest c-axis coincides with one of the 5-fold symmetry axis. Particle models illustrating caking (green) are schematically shown in (f),(l),(r) and (e), where caking size is exaggerated for visibility purpose.

P_l is the l-th order Legendre polynomials, and $\theta(q) = \arccos(\frac{q\lambda}{4\pi})$. Eq. (20) can also be written in matrix notation, with indices q_1 and q_2, as

$$B_l = I_l I_l^*, \tag{21}$$

where I_l is the $N \times 2l + 1$ matrix of spherical harmonic coefficients, with rows indexed by q and columns indexed by m. Additionally, each B_l matrix can be viewed as a rank $2l + 1$

Gram matrix, and thus has the compact eigenvalue decomposition

$$B_l = V_l \Lambda_l V_l^*, \tag{22}$$

where V_l in an $N \times 2l+1$ unitary matrix, Λ_l is a $2l+1 \times 2l+1$ diagonal matrix of nonnegative eigenvalues, and N is the number of sampled q points. A standard linear algebra theorem allows us to relate the decompositions in Eqs. (21) and (22) via

$$I_l = V_l \sqrt{\Lambda_l} U_l, \tag{23}$$

where U_l is an unknown $2l + 1$-dimensional unitary matrix.

The relation in Eq. (23) is essentially a hyperphase generalization of the classical phase problem, where $V_l\sqrt{\Lambda_l}$ can be thought of as the known "amplitude matrix" and U_l can be thought of as the unknown "phase matrix", which needs to be determined in order to reconstruct the 3D intensity function. Therefore, in order to determine the 3D electron density ρ of the imaged structure, one must solve the hyperphase problem [i.e. determine the U_l matrices in Eq. (23)] in order to reconstruct the 3D intensity function I, in addition to the classical phase problem, in order to reconstruct the electron density ρ from I.

The MTIP scheme reconstructs an electron density from the correlation data by simultaneously solving the phase and hyperphase problems via a generalization of classical iterative phasing schemes that are typically used to solve the standard phase problem. In particular, this is accomplished by applying a series of projection operators, each of which seeks to find the minimum-norm perturbation of a model which is consistent with a given constraint, several times in an iterative scheme. These projections include the cross-correlation projector to project a 3D intensity model to be consistent with the B_l data, a magnitude projector to project a 3D electron density model to be consistent with a model intensity, and a support projector to project a density model to be 0 outside of a specified support region, which can be determined during the reconstruction. To reconstruct an electron density from the correlation data, these projection operators are applied in a combination of the error-reducing (ER) [12] and hybrid input-output (HIO) [13] iterative schemes and the shrinkwrap technique [14], which periodically updates an estimate of the support. The details of this procedure can be found in ref. [10].

NOISE MODELING IN MTIP

Here, we add one additional step to the original MTIP scheme in order to model noise in the correlations calculated from experimental data, based on the concept of the noise projection operator introduced in ref. [15]. Instead of first extracting the B_l coefficients from the correlations and then directly fitting to them, we instead update the B_l data during each iteration of MTIP. In particular, during each iteration, we generate a correlation function from the current intensity model, via equations Eqs. (19) and (20), and then use a noise projector P_N to project the correlation function C^{mod} of the current model so that the weighted second order moment of the data about the projected correlation function is less than the weighted sum of an estimated set of variances, i.e. the projected quantity $C^{\mathrm{proj}} = P_N C^{\mathrm{mod}}$ is given by the solution to

$$\min_{C^{\mathrm{proj}}} \sum_{q_1,q_2,\Delta} \left(C^{\mathrm{proj}}(q_1,q_2,\Delta) - C^{\mathrm{mod}}(q_1,q_2,\Delta)\right)^2 w(q_1,q_2),$$
$$\text{subject to} \tag{24}$$
$$\sum_{q_1,q_2,\Delta} \left(C^{\mathrm{proj}}(q_1,q_2,\Delta) - C^{\mathrm{data}}(q_1,q_2,\Delta)\right)^2 w(q_1,q_2) \le \sum_{q_1,q_2,\Delta} \sigma^2_{q_1,q_2} w(q_1,q_2),$$

where σ_{q_1,q_2} is an estimate of the standard deviation for $C^{\mathrm{data}}(q_1,q_2,\Delta)$, and $w(q_1,q_2)$ is a weighting function that can be used to alter the contribution of certain parts of the correlation function, depending on their relevance or noise levels.

The advantage of using the above weighting scheme, where the same weight is used in both the objective and constraint, is that the theory of Lagrange multipliers can be used to give a simple analytic expression for the solution to Eq. (24) as

$$C^{\mathrm{proj}}(q_1,q_2,\Delta) = \begin{cases} C^{\mathrm{mod}}(q_1,q_2,\Delta), & \text{if } \lambda \ge 0 \\ \frac{C^{\mathrm{mod}}(q_1,q_2,\Delta) - \lambda C^{\mathrm{data}}(q_1,q_2,\Delta)}{1-\lambda}, & \text{if } \lambda < 0, \end{cases} \tag{25}$$

where the Lagrange multiplier is given by

$$\lambda = 1 - \sqrt{\frac{\sum_{q_1,q_2,\Delta} \left(C^{\mathrm{mod}}(q_1,q_2,\Delta) - C^{\mathrm{data}}(q_1,q_2,\Delta)\right)^2 w(q_1,q_2)}{\sum_{q_1,q_2,\Delta} \sigma^2_{q_1,q_2} w(q_1,q_2)}}. \tag{26}$$

Any unmeasured or masked quantities are not included in the above optimization and are allowed to float during the reconstruction.

Once we have an updated correlation function C^{proj}, we would like to extract the B_l coefficients from its Legendre decomposition. Note that, for a curved Ewald sphere, even though the Legendre polynomials form an orthogonal basis on the interval $[-1,1]$, the B_l coefficients cannot be computed via an inner product with the correlation function in Eq. (19) since $\cos\theta(q_1)\cos\theta(q_2)+\sin\theta(q_1)\sin\theta(q_2)\cos\Delta$ does not span the entire interval. Alternatively, we can bypass this orthogonality issue by calculating an approximation to the curvature-corrected correlation function $C^{\mathrm{proj}}_{\mathrm{cc}}$, which estimates the correlation function that would be obtained if one had a flat Ewald sphere. However, since the argument of Legendre polynomials for the curved Ewald sphere case does not span $[-1,1]$, there are regions of $C^{\mathrm{proj}}_{\mathrm{cc}}$ that are not sampled by the data, and so in these regions we allow the the values of $C^{\mathrm{proj}}_{\mathrm{cc}}$ to float. More specifically, we calculate the curvature-corrected projected correlation for $0\le\Delta\le\pi$ as

$$C^{\mathrm{proj}}_{\mathrm{cc}}(q_1,q_2,\Delta)=\begin{cases}C^{\mathrm{proj}}(q_1,q_2,cc(q_1,q_2,\Delta)), & \text{if } -1\le\frac{\cos(\Delta)-\cos\theta(q_1)\cos\theta(q_2)}{\sin\theta(q_1)\sin\theta(q_2)}\le 1\\ \sum_{l=0}^{\infty}P_l(\cos\Delta)B_l^{\mathrm{mod}}(q_1,q_2) & \text{otherwise,}\end{cases}\tag{27}$$

where the curvature correction function is

$$cc(q_1,q_2,\Delta)=\arccos\left(\frac{\cos(\Delta)-\cos\theta(q_1)\cos\theta(q_2)}{\sin\theta(q_1)\sin\theta(q_2)}\right),\tag{28}$$

and where the B_l^{mod} are calculated from the most recent intensity model via Eq. (20).

Once $C^{\mathrm{proj}}_{\mathrm{cc}}$ is computed, we use it to update the B_l coefficients by integrating the Legendre polynomials against the curvature-corrected correlation function via

$$B_l(q_1,q_2)=\frac{2l+1}{2}\int_{-1}^{1}C^{\mathrm{proj}}_{\mathrm{cc}}(q_1,q_2,\arccos(x))P_l(x)dx.\tag{29}$$

The above operations are performed during each iteration of MTIP to update the set of B_l data in which MTIP fits a model density to. More specifically, during each iteration of the reconstruction, we perform the operations above to update an estimate of the B_l data from the correlation data and then perform one step of the MTIP procedure outlined in ref. [10] to fit to that B_l data, and repeat until convergence.

MTIP RECONSTRUCTION PARAMETERS

Here we discuss the parameters used in the MTIP reconstructions, each defined in detail in ref. [10]. We use $\beta=0.5$ in the MTIP HIO scheme, and for shrinkwrap we set ϵ to 5%

of the maximum density and σ to the size of a pixel on the computational grid. We apply 15 cycles of the MTIP algorithm, each consisting of 60 MTIP HIO iterations, followed by 40 MTIP ER iterations, and then shrinkwrap. The result is then refined by applying 200 MTIP ER iterations. During each iteration of MTIP, we extracted and fit to B_l values for even l values in the range $0 \leq l \leq 20$ because odd orders vanish in the presence of Friedel symmetry. Computations were done using a spherical-polar grid for both real and Fourier space with 64 radial nodes, up to 207 inclination angles, and 413 azimuthal angles.

The reconstructions were performed using the difference CCFs defined in Eq. (10) to extract B_l data for $l \geq 2$. However, the difference CCFs do not contain a DC component and, thus, the B_0 coefficients were obtained from the SAXS curves via $B_0(q_1, q_2) = \langle I_i(q_1, \varphi)\rangle_{\phi,i} \langle I_i(q_2, \varphi)\rangle_{\phi,i}$. Prior to analysis, the correlations were symmetry averaged about $\Delta = \pi$ in order to reduce noise levels. Due to noise, the autocorrelation curves, i.e. $C(q, q, \Delta)$, contain a large noise peak around $\Delta = 0$, which was masked out of the analysis. Further masking was performed on the correlation data for very low q, which appeared to suffer from large systematic issues; see the section "Generalized Guinier analysis" for details. In order to compute the average reconstructions and the reconstruction statistics (see next section) for each virus, we ran 48 independent MTIP reconstructions from different random starting conditions and aligned the reconstructions.

The standard deviations used in the noise projector were calculated by computing the ℓ^2 difference between Friedel-symmetric components of the correlation curves. More specifically, in the presence of Friedel symmetry, the Legendre decomposition in Eq. (19) only contains even orders of l, which are symmetric about 0, allowing us to estimate the variance of the data via

$$\sigma^2_{q_1,q_2} = \frac{1}{N_\Delta} \sum_{\Delta_{\min} \leq \Delta \leq \Delta_{\max}} \left(C^{\text{data}}_{\text{cc}}(q_1, q_2, \Delta) - C^{\text{data}}_{\text{cc}}(q_1, q_2, \pi - \Delta) \right)^2, \tag{30}$$

where $\Delta_{\min}$ is the smallest sampled value of Δ, $\Delta_{\max}$ is the largest sampled value of $\Delta \leq \pi$, and N_Δ is the number of measured values of Δ in the range $[\Delta_{\min}, \Delta_{\max}]$. For the noise projector, we used the weight $w(q_1, q_2) = (q_1 q_2)^3/\sigma^2_{q_1,q_2}$, which balances between compensating for the decay in signal as a function of q and weighting down the contribution from noisier components of the correlation function.

RECONSTRUCTION STATISTICS AND DISCUSSION

In order to the assess the quality of the reconstructions, we compute both a phase retrieval transfer function (PRTF) [16–18] from the full dataset, as well as the Fourier shell correlations (FSC) [19] between the average structure determined from two randomly generated halves of the correlation data. The PRTF quantifies the uniqueness of the reconstructions obtainable from a single dataset, whereas the FSC on the average structures quantifies the reproducibility of the features observed on the average structures, which are the main objects of interest.

However, one key difference in how MTIP works versus classical single-particle imaging (SPI) techniques, is that MTIP generates a different intensity function for each reconstruction, whereas classical SPI techniques generally reconstruct only one intensity function and then compute the PRTF by solving the phase problem multiple times from the same intensity function. In order to capture the possible variance in the intensity functions recovered from MTIP, we use a modification of the standard PRTF [16]

$$\mathrm{PRTF}(\mathbf{q}) = \frac{|\,\langle \hat{\rho}_k(\mathbf{q})\rangle_k\,|}{\sqrt{\langle I_k(\mathbf{q})\rangle_k}}, \tag{31}$$

where $\hat{\rho}_k$ is the Fourier transform of the k-th reconstructed electron density and I_k is the k-th reconstructed intensity function.

In addition to using the PRTF and FSC to measure consistency of the reconstructed structures, we also compute the average FQC, defined earlier, to assess how well the reconstructions fit the data. The PRTF, FSC, and FQC plots are shown in Fig. S14. Using the established cutoff values, $1/e$ for the PRTF and 0.5 for the FSC, we arrive at resolution estimates of 17.7 nm for RDV and 16.9 nm for PR772 using the PRTF, and 13.5 nm for RDV and 12.6 nm for PR772 using the FSC. The FQC lies above 0.84 for RDV and above 0.91 for PR772 over the entire resolution range, indicating that the reconstructed structures have an excellent amount of agreement with the data.

In order to visualize the distortions in the capsid of the reconstructed viruses, we compare the reconstructed viruses with their icosahedral projections [10] in Fig. S15. It can clearly be seen that the reconstructed RDV capsid is much less distorted than the PR772 capsid, which is extended along one direction. These features are consistent with the results described in the “Model comparison” section above and the “Generalized Guinier analysis” section below.

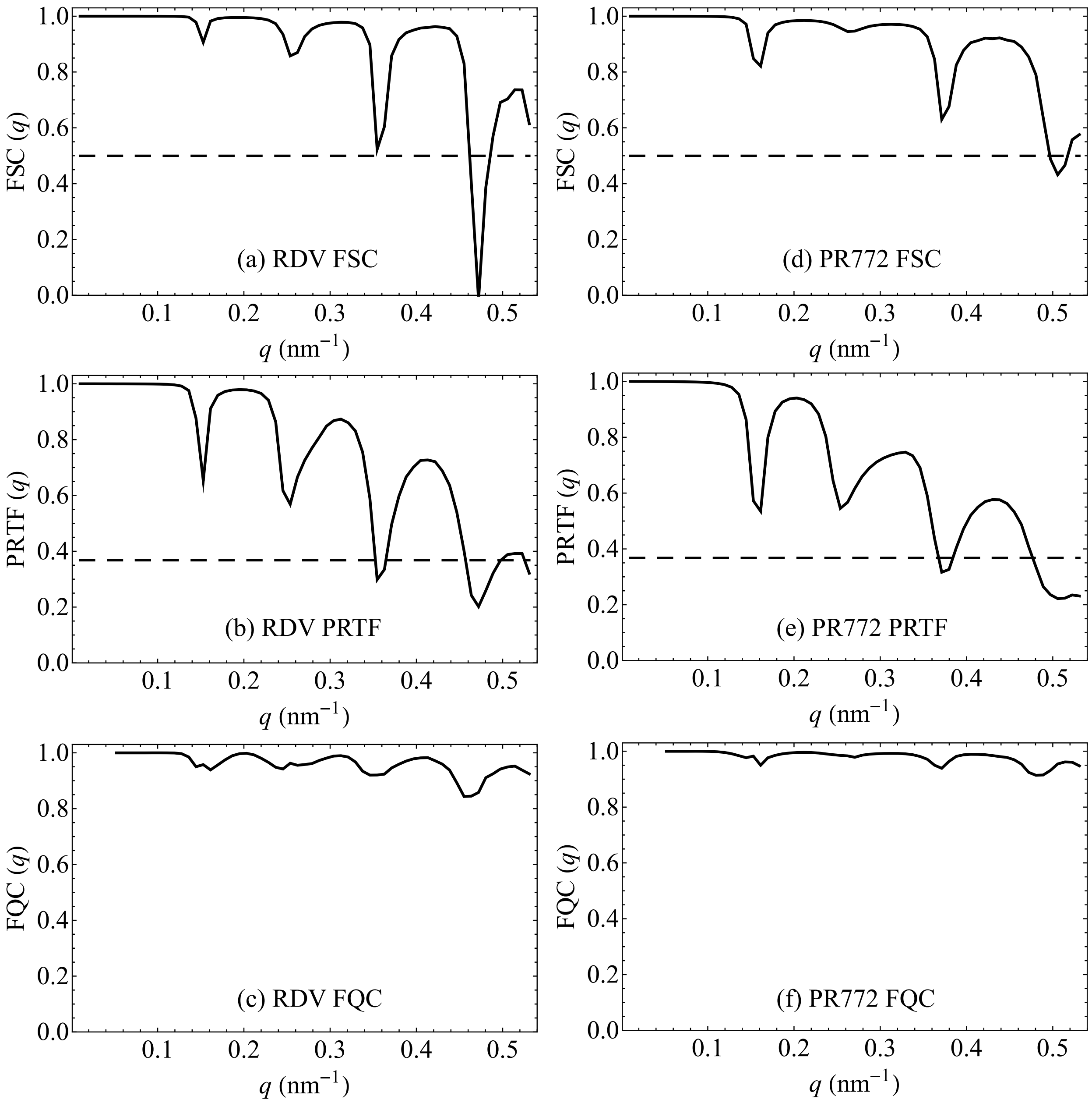

FIG. S14. PRTF, FSC, and FQC plots for the MTIP reconstructions of RDV (a)-(c) and PR772 (d)-(f). The dashed lines represent the cutoff values for the PRTF and FSC, given by $1/e$ and 0.5, respectively.

GENERALIZED GUINIER ANALYSIS

A generalized Guinier analysis on the Legendre-decomposed correlations, as outlined in ref. [20], was performed on the RDV and PR772 data. This type of analysis allows a very rapid, model-free determination of whether the shape of an object is prolate or oblate. The

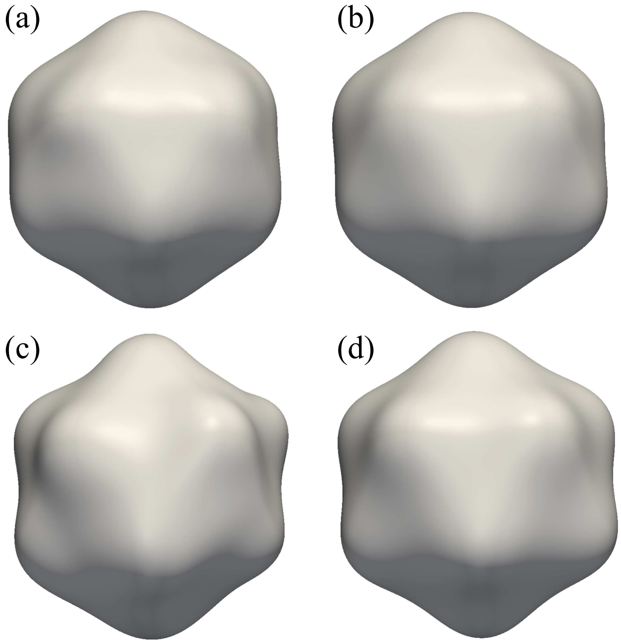

FIG. S15. Comparison of the averaged reconstructions of RDV (a) and PR772 (c) to their respective icosahedral projections (b) and (d).

generalized Guinier equation describing the low-resolution behavior of the B_l data is given by

$$\log B_l - 2l \log q = \log B_l^* - \frac{2q^2 R_l^2}{2l+3}, \tag{32}$$

where B_l^* and R_l are sample-dependent quantities that are related to the multipole moments of the sample's autocorrelation function and can be determined from a simple least squares fit. For $l = 0$, this can be related to the radius of gyration via $R_{\mathrm{g}} = \sqrt{2}R_0$. In ref. [20], it was empirically determined that R_2/R_{g} serves as an indicator of whether the underlying shape is oblate or prolate. Furthermore, it was shown that the first local maximum $\hat{q}_l$ of $B_l(q)$ can be approximated by

$$\hat{q}_l = \sqrt{\frac{l(2l+3)}{2}}\frac{1}{R_l}. \tag{33}$$

We performed the generalized Guinier analysis described above on the experimental correlation data and found that R_{g} for PR772 and RDV were 26.2 and 26.4 nm, respectively. A least squares analysis found R_2 to be equal to 36.2 nm for PR772 and 36.4 nm for RDV, corresponding to the first local maximum of $B_2(q)$ occurring around 0.073 nm^{-1}, which is consistent with the observed $B_2(q)$ curves extracted from both data sets [Fig. S16].

The ratio R_2/R_g is approximately 1.38, which indicates that the particles have an oblate character, as also suggested by the model-based analyses outlined in the main text.

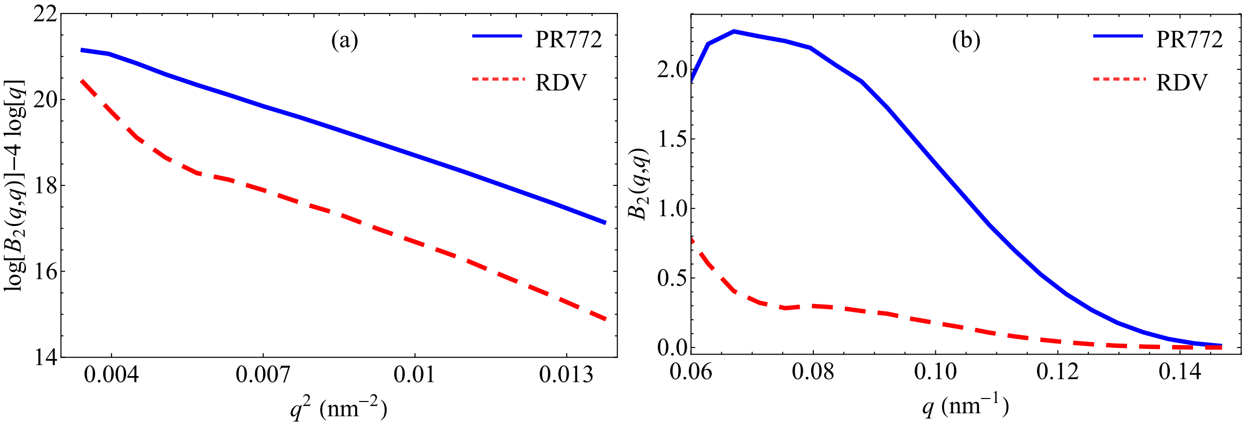

FIG. S16. (a) A generalized Guinier plot of $B_2(q,q)$ shows the expected linear dependence between q^2 and $\log[B_2(q,q)] - 4\log[q]$ for PR772. A strong departure from linearity is observed for RDV, indicating poor quality of the very low-resolution part of the data. The least-squares analyses of the generalized Guinier plot estimates the first maximum of $B_2(q,q)$ at 0.073 nm^{-1}, consistent with a plot of $B_2(q,q)$ vs q (b). Due to poor quality of the very low resolution RDV data, the generalized Guinier analyses on RDV data was unsuccessful. It is worth noting that the magnitude of $B_2(q,q)$ is substantially lower for RDV than for PR772, indicating that the latter has a large departure from icosahedral symmetry than the former.

A visual inspection of the RDV data [see Fig. S16] indicates that the very low resolution B_l data has poorer generalized Guinier properties as compared to the same resolution range in PR772. As is the case in the analyses of standard SAXS data, a deviation from the expected low resolution behavior is indicative of sample or experimental problems, and is typically corrected *post hoc* by omitting this region from any further analyses, as was done in the *ab initio* MTIP reconstructions.

[1] H. K. N. Reddy, C. H. Yoon, A. Aquila, S. Awel, K. Ayyer, A. Barty, P. Berntsen, J. Bielecki, S. Bobkov, M. Bucher, G. A. Carini, S. Carron, H. Chapman, B. Daurer, H. DeMirci, T. Ekeberg, P. Fromme, J. Hajdu, M. F. Hanke, P. Hart, B. G. Hogue, A. Hosseinizadeh, Y. Kim, R. A. Kirian, R. P. Kurta, D. S. D. Larsson, N. D. Loh, F. R. N. C. Maia, A. P. Mancuso,

K. Mühlig, A. Munke, D. Nam, C. Nettelblad, A. Ourmazd, M. Rose, P. Schwander, M. Seibert, J. A. Sellberg, C. Song, J. C. H. Spence, M. Svenda, G. van der Schot, I. A. Vartanyants, G. J. Williams, and P. L. Xavier, Scientific Data **4**, 170079 (2017).

[2] H. D. T. Mertens and D. I. Svergun, J. Struct. Biol. **172**, 128 (2010).

[3] M. Altarelli, R. P. Kurta, and I. A. Vartanyants, Phys. Rev. B **82**, 104207 (2010); Erratum: **86**, 179904(E) (2012).

[4] R. P. Kurta, M. Altarelli, E. Weckert, and I. A. Vartanyants, Phys. Rev. B **85**, 184204 (2012).

[5] R. P. Kurta, L. Grodd, E. Mikayelyan, O. Y. Gorobtsov, I. Fratoddi, I. Venditti, M. Sprung, S. Grigorian, and I. A. Vartanyants, J.Phys: Conf. Series **499**, 012021 (2014).

[6] W. O. Saxton and W. Baumeister, Journal of Microscopy **127**, 127 (1982).

[7] M. van Heel, W. Keegstra, W. Schutter, and E. F. J. van Bruggen, in *"The Structure and Function of Invertebrate Respiratory Proteins.", EMBO workshop 1982, Life Chemistry Reports Suppl. 1*, edited by E. J. Wood (1982) pp. 69–73.

[8] M. van Heel and M. Schatz, J. Struct. Biol. **151**, 250 (2005).

[9] A. Nakagawa, N. Miyazaki, J. Taka, H. Naitow, A. Ogawa, Z. Fujimoto, H. Mizuno, T. Higashi, Y. Watanabe, T. Omura, R. Cheng, and T. Tsukihara, Structure **11**, 1227 (2003).

[10] J. J. Donatelli, P. H. Zwart, and J. A. Sethian, Proc. Nat. Acad. Sci. **112**, 10286 (2015).

[11] Z. Kam, Macromolecules **10**, 927 (1977).

[12] R. W. Gerchberg and W. O. Saxton, Optik **35**, 237 (1972).

[13] J. R. Fienup, Opt. Lett. **3**, 27 (1978).

[14] S. Marchesini, H. He, H. N. Chapman, S. P. Hau-Riege, A. Noy, M. R. Howells, U. Weierstall, and J. C. H. Spence, Phys. Rev. B **68**, 140101(R) (2003).

[15] J. J. Donatelli, J. A. Sethian, and P. H. Zwart, Proc. Nat. Acad. Sci. **114**, 7222 (2017).

[16] S. Marchesini, H. N. Chapman, A. Barty, C. Cui, M. R. Howells, J. C. H. Spence, U. Weierstall, and A. M. Minor, arXiv:physics/0510033 (2005).

[17] D. Shapiro, P. Thibault, T. Beetz, V. Elser, M. Howells, C. Jacobsen, J. Kirz, E. Lima, H. Miao, A. M. Neiman, and D. Sayre, Proc. Nat. Acad. Sci. **102**, 15343 (2005).

[18] H. N. Chapman, A. Barty, S. Marchesini, A. Noy, S. P. Hau-Riege, C. Cui, M. R. Howells, R. Rosen, H. He, J. C. H. Spence, U. Weierstall, T. Beetz, C. Jacobsen, and D. Shapiro, J. Opt. Soc. Am. A **23**, 1179 (2006).

[19] S. H. W. Scheres and S. Chen, Nature Methods **9**, 853 (2012).

[20] E. Malmerberg, C. A. Kerfeld, and P. H. Zwart, IUCrJ **2**, 309 (2015).