The symmetries of image formation by scattering. II. Applications

Peter Schwander,[1] Dimitrios Giannakis,[2] Chun Hong Yoon[1] and Abbas Ourmazd[1,*]

[1]*Department of Physics, University of Wisconsin Milwaukee, 1900 E Kenwood Blvd, Milwaukee, WI 53211, USA*

[2]*Courant Institute of Mathematical Sciences, New York University, 251 Mercer St, New York, NY 10012, USA*

[*]*ourmazd@uwm.edu*

Abstract: We show that the symmetries of image formation by scattering enable graph-theoretic manifold-embedding techniques to extract structural and timing information from simulated and experimental snapshots at extremely low signal. The approach constitutes a physically-based, computationally efficient, and noise-robust route to analyzing the large and varied datasets generated by existing and emerging methods for studying structure and dynamics by scattering. We demonstrate three-dimensional structure recovery from X-ray diffraction and cryo-electron microscope image snapshots of unknown orientation, the latter at 12 times lower dose than currently in use. We also show that ultra-low-signal, random sightings of dynamically evolving systems can be sequenced into high quality movies to reveal their evolution. Our approach offers a route to recovering timing information in time-resolved experiments, and extracting 3D movies from two-dimensional random sightings of dynamic systems.

OCIS codes: (290.5825) Scattering theory; (290.5840) Scattering, molecules; (290.3200) Inverse scattering; (140.2600) Free-electron lasers (FELs); (180.6900) Three-dimensional microscopy.

References and links

1. D. Giannakis, P. Schwander, and A. Ourmazd, "The symmetries of image formation by scattering. I. Theoretical framework," Opt. Express (2012). Submitted.
2. V. L. Shneerson, A. Ourmazd, and D. K. Saldin, "Crystallography without crystals. i. the common-line method for assembling a 3D diffraction volume from single-particle scattering," Acta Cryst. A **64**, 303–315 (2008).
3. R. Fung, V. Shneerson, D. K. Saldin, and A. Ourmazd, "Structure from fleeting illumination of faint spinning objects in flight," Nat. Phys. **5**, 64–67 (2008).
4. P. Schwander, R. Fung, G. N. Phillips, and A. Ourmazd, "Mapping the conformations of biological assemblies," New J. Phys. **12**, 1–15 (2010).
5. http://www.youtube.com/watch?v=vqQfARtnsWw.
6. http://www.youtube.com/watch?v=hbFiNaire4o.
7. http://www.youtube.com/watch?v=unnHKBCT8XQ.
8. http://www.youtube.com/watch?v=9Y2DF-X5LSA.
9. J. Frank, "Single-particle imaging of macromolecules by cryo-electron microscopy," Annu. Rev. Biophys. Biomolec. Struct. **31**, 303–319 (2002).
10. S. H. W. Scheres, H. Gao, M. Valle, G. T. Herman, P. P. B. Eggermont, J. Frank, and J.-M. Carazo, "Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization," Nature Methods **4**, 27–29 (2007).
11. N. T. D. Loh and V. Elser, "Reconstruction algorithm for single-particle diffraction imaging experiments," Phys. Rev. E **80**, 026705 (2009).

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

12. N. Fischer, A. L. Konevega, W. Wintermeyer, M. V. Rodnina, and H. Stark, "Ribosome dynamics and tRNA movement by time-resolved electron cryomicroscopy." Nature **466**, 329–333 (2010).
13. J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science **290**, 2319–2323 (2000).
14. S. T. Roweis and S. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science **290**, 2323–2326 (2000).
15. M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Comput. **13**, 1373–1396 (2003).
16. D. L. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," Proc. Natl. Acad. Sci. **100**, 5591–5596 (2003).
17. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition on data," Proc. Natl. Acad. Sci. **102**, 7426–7431 (2005).
18. R. R. Coifman and S. Lafon, "Diffusion maps," Appl. Comput. Harmon. Anal. **21**, 5–30 (2006).
19. R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Reference free structure determination through eigenvectors of center of mass operators," Appl. Comput. Harmon. Anal. **28**, 296–312 (2010).
20. C. M. Bishop, M. Svensen, and C. K. I. Williams, "GTM: The generative topographic mapping," Neural Computation **463**, 379–383 (1998).
21. R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph Laplacian tomography from unknown random projections," IEEE Trans. Image Process. **17**, 1891–1899 (2008).
22. A. Singer, R. R. Coifman, F. J. Sigworth, D. W. Chester, and Y. Shkolnisky, "Detecting consistent common lines in cryo-EM by voting," J. Struct. Biol. **169**, 312–322 (2010).
23. J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State* (Oxford University Press, 2006).
24. A. H. Taub, "Empty space-times admitting a three parameter group of motions," Ann. Math. **53**, 472–490 (1951).
25. B. L. Hu, "Scalar waves in the mixmaster universe. I. The Helmholtz equation in a fixed background," Phys. Rev. D **8**, 1048–1060 (1973).
26. L. C. Biedenharn and J. D. Louck, *Angular Momentum in Quantum Physics* (Addison Wesley, Reading, 1981).
27. Y. LeCun, J. S. Denker, S. Solla, R. E. Howard, and L. D. Jackel, "Optimal brain damage," in *Advances in Neural Information Processing Systems (NIPS 1989)*, **2**, D. Touretzky, ed. (Morgan Kaufman, Denver, CO, 1990), pp. 598–605.
28. G. W. Stewart, "Error and perturbation bounds for subspaces associated with certain eigenvalue problems," SIAM Rev. **15**, 727–764 (1973).
29. L. Lovisolo and E. A. B. da Silva, "Uniform distribution of points on a hyper-sphere with applications to vector bit-plane encoding," IEEE Proc., Vis. Image Signal Process. **148**, 187–193 (2001).
30. D. T. Cromer and J. B. Mann, "Atomic scattering factors computed from numerical Hartree-Fock wavefunctions," Acta Cryst. A **24**, 321–324 (1968).
31. D. E. Knuth, *The Art of Computer Programming: Seminumerical Algorithms*, 3rd ed., (Addison-Wesley, 1997), Vol. 2.
32. P. Schwander, "Efficient interpolation of scattering data to an arbitrary grid," (in preparation, 2012).
33. L. Palatinus and G. Chapuis, "SUPERFLIP – a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions," J. Appl. Cryst. **40**, 786–790 (2007).
34. B. Moths and A. Ourmazd, "Bayesian algorithms for recovering structure from single-particle diffraction snapshots of unknown orientation: a comparison," Acta. Cryst. **A67**, 481–486 (2011).
35. H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. N. C. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Hömke, C. Reich, D. Pietschner, L. Strüder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K. Kühnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jönsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C. Schröter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. M. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, and J. C. H. Spence, "Femtosecond X-ray protein nanocrystallography," Nature **470**, 73–77 (2011).
36. J. Zhang, M. L. Baker, G. F. Schröder, N. R. Douglas, S. Reissmann, J. Jakana, M. Dougherty, C. J. Fu, M. Levitt, S. J. Ludtke, J. Frydman, and W. Chiu, "Mechanism of folding chamber closure in a group II chaperonin," Nature **463**, 379–383 (2010).
37. S. J. Ludtke, P. R. Baldwin, and W. Chiu, "EMAN: Semiautomated software for high-resolution single-particle reconstructions," J. Struct. Biol. **128**, 82–97 (1999).
38. J. Frank and L. Al-Ali, "Signal-to-noise ratio of electron micrographs obtained by cross-correlation," Nature **256**, 376–379 (1975).

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

39. J. F. M. Svensen, "GTM: The generative topographic mapping," Ph.D. thesis, Aston University (1998).
40. J. Frank, M. Radermacher, P. Penczek, J. Zhu, Y. Li, M. Ladjadj, and A. Leith, "SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields," J. Struct. Biol. **116**, 190 (1996).
41. J M. Glownia, J. Cryan, J. Andreasson, A. Belkacem, N. Berrah, C. I. Blaga, C. Bostedt, J. Bozek, L. F. DiMauro, L. Fang, J. Frisch, O. Gessner, M. Gühr, J. Hajdu, M. P. Hertlein, M. Hoener, G. Huang, O. Kornilov, J. P. Marangos, A. M. March, B. K. McFarland, H. Merdji, V. S. Petrovic, C. Raman, D. Ray, D. A. Reis, M. Trigo, J. L. White, W. White, R. Wilcox, L. Young, R. N. Coffee, and P. H. Bucksbaum, "Time-resolved pump-probe experiments at the LCLS," Opt. Express **18**, 17620–17630 (2010).
42. J. P. Cryan, J. M. Glownia, J. Andreasson, A. Belkacem, N. Berrah, C. I.Blaga, C. Bostedt, J. Bozek, C. Buth, L. F. DiMauro, L. Fang, O. Gessner, M. Guehr, J. Hajdu, M. P. Hertlein, M. Hoener, O. Kornilov, J. P. Marangos, A. M. March, B. K. McFarland, H. Merdji, V. S. Petrović, C. Raman, D. Ray, D. Reis, F. Tarantelli, M. Trigo, J. L. White, W. White, L. Young, P. H. Bucksbaum, and R. N. Coffee, "Auger electron angular distribution of double core-hole states in the molecular reference frame," Phys. Rev. Lett. **105**, 083004 (2010).
43. M. Balasubramanian and E. L. Schwartz, "The Isomap algorithm and topological stability," Science **295**, 5552 (2002).
44. B. Zhang, M. J. Fadili, J. L. Starck, and J. C. Olivo-Marin, "Multiscale variance-stabilizing transform for mixed-Poisson-Gaussian processes and its applications in bioimaging," in *Proceedings of IEEE International Conference on Image Processing*, **6** (Institute of Electrical and Electronics Engineers, New York), 233–236.
45. B. Zhang, J. Fadili, and J. Starck, "Wavelets, ridgelets, and curvelets for Poisson noise removal," IEEE Trans. Image Process. **17**, 1093–1108 (2008).
46. Y. Guan, "Variance stabilizing transformations of Poisson, binomial and negative binomial distributions," Stat. Probabil. Lett. **14**, 1621–1629 (2009).
47. L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in neural information processing systems*, **17** (2004), 1601–1608.
48. J. Chen and I. Safro, "Algebraic distance on graphs," SIAM J. Sci. Comput. (2010). Submitted.
49. UCSF CHIMERA package, Resource for Biocomputing, Visualization, and Informatics, University of California, San Francisco (supported by NIH P41 RR-01081).

1. Introduction

In an earlier paper [1], hereafter referred to as Paper I, we presented a theoretical framework for analyzing snapshots formed by scattering. In this paper, we demonstrate the power of this approach to reconstruct three-dimensional (3D) models and time-series from random sightings at extremely low signal, with no orientational or timing information. The theoretical framework in Paper I represents the information content of an ensemble of snapshots as a Riemannian manifold, and shows that the properties of operations in space give rise to object-independent symmetries. Purposeful navigation on this manifold is tantamount to reconstructing a 3D model of the sighted system and/or its evolution, in the sense that given any snapshot, any other can be produced on demand. The symmetries of the manifold reveal its natural eigenfunctions, thus allowing physically-based interpretation of graph-theoretic analysis, and enhanced noise discrimination. Simple algorithms then suffice to reach exceptionally low signal-to-noise levels unmatched by other approaches in terms of computational cost, noise robustness, or both. As examples, we demonstrate structure recovery from radiation-sensitive objects at doses at least an order of magnitude below current levels (signal-to-noise ratio (SNR): -16 dB), and reconstruction of time-series at SNR values as low as -21 dB. The versatility of the approach is demonstrated in the context of simulated and experimental data from X-ray diffraction, cryo-electron microscopy (cryo-EM), and optical snapshots using a variety of graph-theoretic techniques. These applications demonstrate the generality of the symmetry-based approach, elucidating at the same time the measures needed to deal successfully with experimental data, a key benchmark of the practical utility of any theoretical framework.

This paper is organized as follows. Without claim to be comprehensive, Sec. 2 briefly summarizes previous work in the field to provide a context for the applications discussed in this paper. For the convenience of non-mathematical readers, Sec. 3 provides a conceptual outline of the theoretical framework developed in Paper I. Sec. 4.1 describes 3D reconstruction from simulated diffraction snapshots of single biomolecules at the signal level expected in upcom-

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

ing experiments utilizing the new generation of X-ray Free Electron Lasers (XFELs) [2–4]. Section 4.2 establishes, in principle, the applicability of our approach to crystalline samples. Section 4.3 addresses structure recovery from simulated and experimental cryo-EM snapshots of single molecules. In this case, essential experimental issues such as defocus variation must be faced and incorporated into the theoretical formalism. Section 4.4 demonstrates reconstruction of time-series (movies) from random sequences of ultralow-signal optical snapshots. The paper concludes in Sec. 5 with a summary of our key findings and their implications. Detailed points of a technical nature are elucidated in appendices, and movies provided as online material [5–8].

2. Previous work

As described in Paper I, we are concerned with constructing a model from random sightings of a system viewed in some projection, i.e., by accessing a limited number of variables describing the state of the system. A 3D model of an object and its evolution, for example, can be constructed from an ensemble of low-signal 2D snapshots without orientational information [3, 4, 9–12]. Modern graph-theoretic algorithms can now be used to discover low-dimensional manifolds representing the information content of datasets in some high-dimensional space determined by the measurement apparatus [13–20]. The power of these methods stems from their generality, in the sense that few assumptions are made as to the nature of the data or their internal correlations. This brings with it four major challenges: (1) Interpretation of the analysis results ("what physical variables do the manifold dimensions represent?"); (2) Computational cost and scaling behavior on moving from simulated ("toy") datasets to experimental measurements; (3) Robustness against noise, particularly of non-additive, non-Gaussian types; and (4) Incorporation of inevitable and/or desirable experimental factors ("utility of the theoretical framework in practice".)

These issues can be brought to focus in the context of the much-discussed problem of recovering the orientation of cryo-EM snapshots of faint biological objects. Direct graph-theoretic attempts to determine the orientation of snapshots from a synthetic object were abandoned at a signal-to-noise ratio (SNR) of ~ 2 dB, even though only additive Gaussian noise was included [21]. Noting that graph-theoretic analyses often "fail to solve the cryo-EM problem, because the reduced coordinate system that each of them obtains does not agree with the projection directions" [19], properties specific to cryo-EM images were used to extract information from the snapshots. Graphs were then constructed using this information in order to assign physical meaning to the outcome of the analysis. Orienting low-signal cryo-EM snapshots by utilizing so-called (straight) common-lines identified primarily in simulated data with additive Gaussian noise has reached remarkably low SNR values [22]. However, such assumptions, while justified under some circumstances, are not generally valid. Common-lines, for example, are present only when elastic single-scattering dominates, are straight only when the wavelength of the incident radiation is so short that the Ewald sphere can be replaced by a plane, and are compromised by defocus variations essential for reliable structure recovery by cryo-EM [23].

Symmetry-based assignment of physical meaning to the outcome of graph-theoretic analysis of scattering data and its favorable computational consequences were addressed in Paper I. Here, we are concerned with ability of this theoretical framework to deal with noise and other important factors encountered in experimental datasets. This determines the practical utility of an approach as much as theoretical elegance and computational efficiency. Below, we demonstrate the utility of our symmetry-based approach by applying a number of manifold-embedding techniques to a variety of simulated and experimental datasets (see Table 1).

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

3. Conceptual summary of theoretical framework

A snapshot formed on a 2D detector by scattered radiation from an object can be represented by a vector, with the intensity values recorded at the n detector pixels as components (Paper I, Fig. 1). Object motion and/or evolution (dynamics) change the pixel intensities, causing "the vector tips" representing the ensemble of snapshots to trace out a surface — a manifold — in the n-dimensional data space. The number of degrees of freedom available to the object determines the dimensionality of the manifold traced out. Rotations of a rigid object in 3D, for example, result in a 3D manifold.

The data manifold represents the totality of information about the object gathered by the detector in the course of an observation. Learning is tantamount to understanding the properties of this manifold sufficiently to "navigate" on it. Learning the manifold generated by object rotations, for example, is equivalent to constructing a 3D model of the object, because, starting from any 2D projection (point on the manifold) any other 2D projection can be found by navigation, with the shortest route corresponding to a geodesic.

As we are initially concerned with constructing 3D models from 2D snapshots, we consider a formulation of scattering by a single object in Fourier space so as to concentrate on the effect of rotations. This circumvents issues such as rigid shifts, which would otherwise have to be corrected or incorporated as additional manifold dimensions. Rotation operations do not commute. One must therefore consider the order in which they are performed. This leads to a distinction between so-called left and right "translations," where a rotation operator T is placed to the left or right of another rotation operator R, i.e., TR vs. RT. A left translation can be thought of as an active rotation in 3D space of the incident beam-detector arrangement (frame rotation) after R. Similarly, a right translation corresponds to an active rotation of the object (object rotation) before R. As $\mathsf{TR} \neq \mathsf{RT}$, left and right translations must be considered separately. Each forms an SO(3) group, and the total set of possible operations to be considered corresponds to $\mathrm{SO}(3) \times \mathrm{SO}(3)$.

A key question is this: Which, if any, of these operations leave the distances on the manifold unchanged, i.e., which operations are "invisible" to an ant crawling on the manifold? These operations would represent symmetries (more precisely, isometries) of the manifold. For a detector with circular symmetry, the distances on the manifold are invariant under beam-detector rotations about the beam axis. This is obvious; a frame rotation about the beam axis rotates all the snapshots by the same amount about that axis without changing them. This leaves the distances on the manifold unchanged. The process of image formation on a circularly-symmetric detector at right angles to the illuminating beam thus has SO(2) isometry, i.e., of all possible SO(3) frame operations, the SO(2) subset of rotations about the beam direction leave the distances on the manifold unchanged. This is related to the projection of a 3D object on the 2D detector, which is equivalent to a "central slice" through the diffraction volume in reciprocal space.

Consider next the SO(3) set of operations corresponding to object rotations. It turns out that the metric measuring distance on the manifold can be decomposed into a homogeneous part, which varies uniformly with object rotation, plus a residual term, which acts as a fingerprint of the object (see Paper I, Sec. 4.4). Considering the homogeneous part only, the total set of symmetries is then $\mathrm{SO}(2) \times \mathrm{SO}(3)$. The same set of symmetries appears in certain models of the universe in general relativity [24, 25], and is associated with well-known eigenfunctions familiar in the context of spinning tops in classical and quantum mechanics [26].

The key point here is that the knowledge of the manifold symmetries, which stem from the nature of operations in space, allows one to determine the leading-order properties of the manifold under a very general set of scattering scenarios, including its natural eigenfunctions. Projection of noisy datasets on these eigenfunctions is tantamount to noise discrimination. The

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

components of a data point representing a snapshot can then be directly related to its orientation (see Paper I, Sec. 4.4).

4. Applications

It has long been known that the use of problem-specific constraints can substantially increase computational efficiency [27]. By combining wide applicability with class specificity, symmetries represent a particularly powerful example of such constraints. In Paper I, we used the object-independent symmetries of image formation to recover 3D structure from a large ensemble of simulated, noise-free diffraction snapshots with a computational complexity $10^4\times$ higher than the state of the art. Here we demonstrate the noise robustness stemming from exploiting the symmetries of image formation. Examples include orientation recovery, 3D reconstruction, and movie extraction from ultra-low-signal diffraction or image snapshots of periodic and non-periodic objects and dynamical systems. Each example was selected to highlight an important application area. As shown in Table 1, a variety of manifold-embedding techniques can be used.

Table 1. Summary of applications. For Diffusion Map see Refs. [17, 18], Isomap Ref. [13], GTM Refs. [3, 20].

Data type	Observed system	Snapshot type	Reconstruction	Manifold-embedding technique
Simulated	Adenylate kinase molecule[a]	Diffraction	3D structure	Diffusion Map
	Chignolin molecule	Diffraction	3D structure	Diffusion Map
Experimental	Superoxide dismutase-1 crystal	Diffraction	Orientation recovery	Isomap
	Chaperonin molecule	Cryo-EM images	3D structure	GTM
	Pirouette	Unsorted image frames	Time series	Diffusion Map
	Pas de deux	Unsorted image frames	Time series	Isomap

[a] see Paper I, Sec. 5.1.

4.1. Structure recovery from simulated diffraction snapshots of non-periodic objects at ultra-low signal

First, we demonstrate 3D structure recovery from a collection of two million simulated diffraction snapshots of the synthetic protein chignolin (Protein Data Bank (PDB) descriptor: 1UAO, model 1) at 4×10^{-2} scattered photons per Shannon pixel at 1.8 Å. (A Shannon pixel is of the size needed for appropriate sampling of the intensity distribution as prescribed by the Shannon-Nyquist theorem.) This scattered intensity is expected from a 500 kD protein exposed to a single pulse from an XFEL [2, 3]. At this signal level, Poisson (shot) noise dominates. The ability to deal with such levels of non-additive noise was previously demonstrated only by Bayesian algorithms [3, 11] with extremely unfavorable scaling behavior [1, 3, 4, 11], restricting the size of amenable objects to eight times the spatial resolution.

Here, we use the symmetry-based approach described in Paper I after modest denoising. The denoising scheme consists of two steps: (1) Convolve the snapshot pixels with a 2D Gaussian filter with a width approximately equal to that of a Shannon pixel; (2) Replace each snapshot vector by an average over its local neighbors. Depending on the SNR, a number iterations of step (2) may be needed, with a stopping criterion based on a least-squares residual determined through the first nine Laplacian eigenfunctions of the dataset (ranked in order of increasing eigenvalue). These eigenfunctions are employed in our scheme to assign an orientation to each snapshot. (For details see Appendix A and Paper I.) As shown in Appendix A below, the effect of this denoising scheme is benign, in the sense that the noisy snapshots tend smoothly toward

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

their noise-free analogues. Of course, the denoising procedure cannot be repeated indefinitely. The stopping criterion described above is a simple and practical means for terminating the denoising process.

To estimate the accuracy of orientation recovery, we use the following measure for root-mean-square (RMS) distance between the deduced and true orientations:

$$\varepsilon = \left[\frac{1}{s(s-1)} \sum_{i \neq j} (\tilde{D}_{ij} - D_{ij})^2 \right]^{1/2}, \tag{1}$$

where $D_{ij} = 2\arccos(|\tau_i \cdot \tau_j|)$ and $\tilde{D}_{ij} = 2\arccos(|\tilde{\tau}_i \cdot \tilde{\tau}_j|)$ are the true and estimated internal distances between orientations i and j, respectively, and $\cdot$ is the inner product between quaternions. Moreover, to assess the influence of local averaging on the eigenfunctions employed for orientation recovery, we compute the distance γ of the invariant subspace $\tilde{V}$ spanned by the leading nine eigenfunctions of the diffusion matrix P_ε in Table 5 from the corresponding invariant subspace V of the noise-free diffusion matrix. Note that V and $\tilde{V}$ consist of all linear combinations of the form $\sum_{k=1}^{9} c_k \underline{\psi}_k$, where $\underline{\psi}_k$ are the first nine eigenvectors of P_ε ranked in order of increasing eigenvalue. Moreover, P_ε has size $s \times s$, where s is the number of snapshots in the data set; i.e., $\tilde{V}$ and V are subspaces of $\mathbb{R}^s$.

Here, we employ a standard distance measure from matrix perturbation theory [28] (also used in Paper I), viz.

$$\gamma = \|\tilde{\Pi} - \Pi\|_2, \tag{2}$$

where $\tilde{\Pi}$ and Π are orthogonal projectors from $\mathbb{R}^s$ to $\tilde{V}$ and V, respectively, and $\|\cdot\|_2$ denotes the spectral norm of matrices. With this definition, γ lies in the interval $[0,1]$, and may be interpreted as the sine of an angle characterizing the deviation of $\tilde{V}$ from V. For our purposes, Eq. (2) is more appropriate than an error measure based on the difference between the noisy and noise-free diffusion matrices (or their generators), since the latter depends on higher eigenfunctions which are not used in our scheme.

Diffraction snapshots were simulated in 2×10^6 different orientations to a spatial resolution of 1.8 Å using 1 Å photons. The orientations were sampled over SO(3), as described in Ref. [29]. Cromer-Mann atomic scattering factors [30] were used for the 77 non-hydrogen atoms, and the hydrogen atoms neglected. The detector pixel was the appropriate Shannon pixel [3], which oversamples the scattered amplitudes by a factor of two, resulting in $40 \times 40 = 1600$ Shannon detector pixels. To model shot noise, diffracted intensities were scaled so that the mean photon count (MPC) per Shannon pixel was 0.04 at 1.8 Å resolution. The quantized photon count at each pixel was obtained from a Poisson distribution by the algorithm described in Ref. [31].

With no other information, the noisy diffraction patterns were provided to the algorithm in Appendix A (width of Gaussian filter for image smoothing $\sigma = 0.7$; number of nearest neighbors in the sparse distance matrix $d = 220$; number of nearest neighbors for local averaging $l = 20$; number of datapoints for least-squares fitting $r = 8 \times 10^4$; number of nearest neighbors for autotuning $n = 30$.) As illustrated in Fig. 1, the least-squares residual $\mathscr{G}^*$, the subspace distance γ, and the RMS orientation recovery error ε all decrease monotonically for the first five iterations of local averaging, but exhibit a mild increase at iteration six. At that point the algorithm was terminated in accordance with the stopping criterion described above and in Appendix A. The minimum orientation recovery error ε attained with this choice of parameters at iteration 5 is ~ 1.1 Shannon angles. We measured comparable levels of orientation-recovery accuracy for various combinations of l and n parameters in the range 10–50. In all cases, we observed that small values of $\mathscr{G}^*$ correlate strongly with small values of ε, indicating that the least-squares residual provides an effective guideline for setting the parameters of the algorithm.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

This is particularly important, because $\mathscr{G}^*$ depends solely on the Laplacian eigenfunctions (see Appendix A), and, unlike ε, can be evaluated in an experimental environment where the correct orientations are not known.

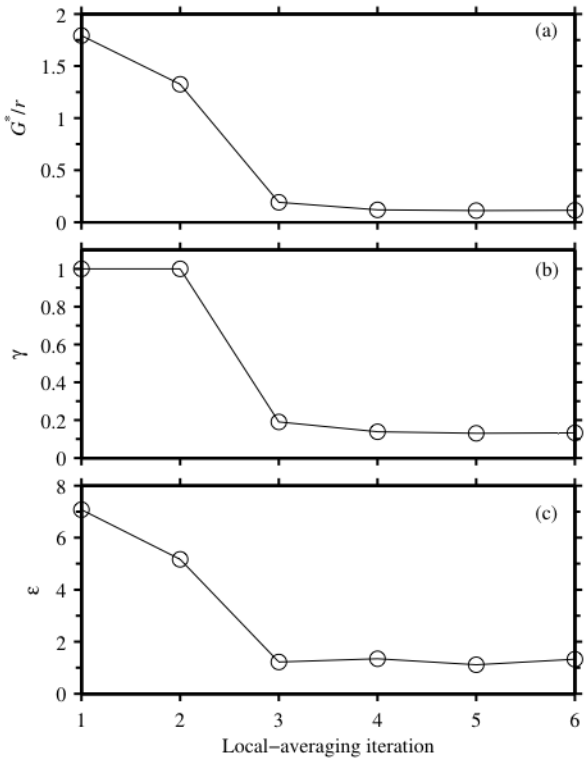

Fig. 1. (a) Least-squares residual $\mathscr{G}^*$, (b) invariant-subspace distance γ, (c) RMS internal distance error ε, shown as a function of the local-averaging iteration count. In Panel (a), $\mathscr{G}^*$ has been normalized by the number of samples $r = 8 \times 10^4$ used for least-squares fitting.

The quality of orientation recovery was further tested by inverting the reconstructed 3D diffraction volume compiled on a uniform Cartesian grid by an interpolation scheme consistent with the geometry of diffraction [32]. The *R*-factor between the gridded scattering amplitudes $\tilde{F}_i$ and those obtained from the Fourier transform of the recovered electron density from phasing, F_i, was defined as

$$R = \frac{\sum_i (|\tilde{F}_i| - |F_i|)^2}{\sum_i |F_i|^2}. \tag{3}$$

The 3D electron density obtained by iterative phasing with the SUPERFLIP algorithm [33] ($R = 0.20$) is shown in Fig. 2. The close agreement with the known structure of chignolin clearly demonstrates sufficient alignment accuracy for reconstruction to 1.8 Å resolution. This is on par with the computationally much more expensive Bayesian approaches [3, 11, 34].

4.2. Orienting diffraction patterns of crystals

So far we have shown that our symmetry-based approach can be used to orient diffraction patterns from single molecules to high accuracy. We now demonstrate that this approach can also orient diffraction snapshots from crystals. This is important, because recent XFEL-based

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

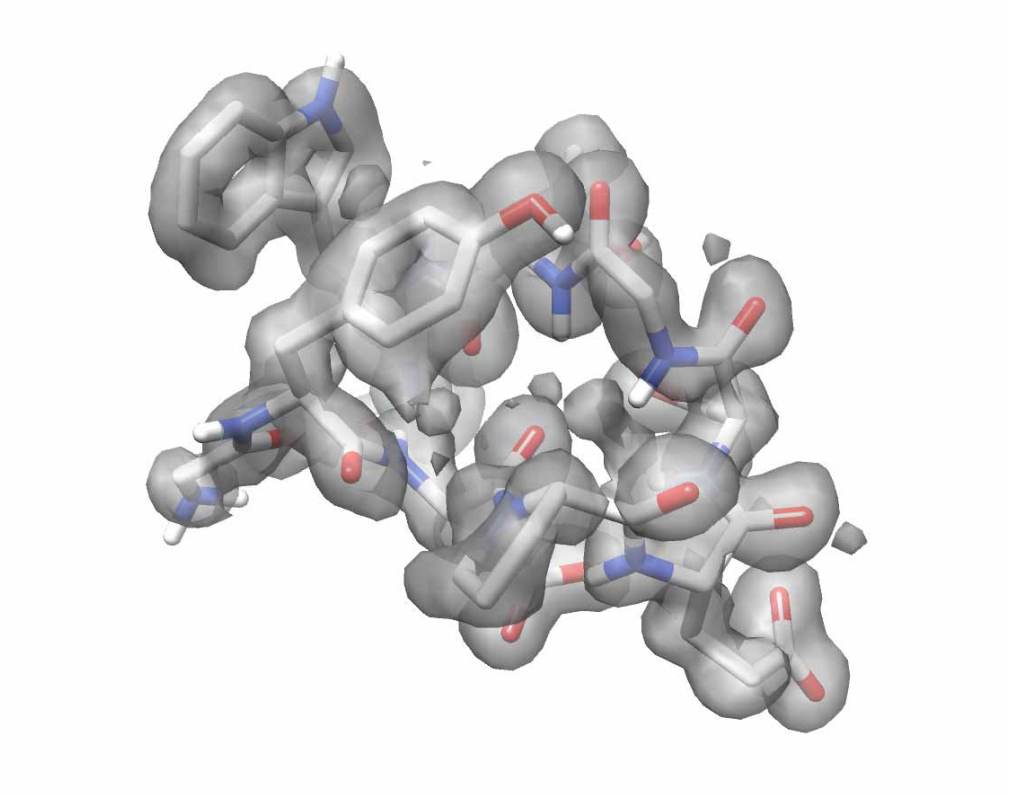

Fig. 2. Three-dimensional electron density of the synthetic protein chignolin, recovered from 2×10^6 noisy diffraction patterns of unknown orientation at a mean photon count of 0.04 per pixel at 1.8 Å resolution. The ball-and-stick model represents the actual structure.

"diffract-and-destroy" approaches, which use femtosecond X-ray pulses to "outrun radiation damage", produce diffraction snapshots of nanocrystals of unknown orientation [35]. As a representative example, we consider a biological crystal of the enzyme superoxide dismutase-1 (SOD1, PDB designation: 1AR4) with $\sim 3 \times 10^3$ atoms per unit cell, and thus highly complex diffraction patterns. The key issue is whether manifolds produced by diffraction snapshots of crystals are sufficiently homogeneous (possess sufficiently homogeneous metrics) for snapshot orientations to be recovered in a straightforward manner. To demonstrate this point, we intentionally utilized snapshots spanning an orientation range of 90° so as to produce an open 1D manifold, and analyzed the dataset with the Isomap manifold-embedding method [13]. In contrast to Diffusion Map, whose eigenfunctions are insensitive at the boundaries, Isomap maps a 1D open manifold to a straight line segment, and is sensitive to the snapshot orientation over the entire range.

Experimental diffraction patterns of a single crystal of superoxide dismutase-1 with a mosaicity of 0.8° were obtained at the Advanced Photon Source ($\lambda = 0.98$ Å). The crystal was rotated about an arbitrary axis with a step size of 0.5°, and 1800 diffraction patterns recorded over a range of 90°. To compensate for spurious beam intensity fluctuations, the diffraction pattern intensities were normalized. Isomap was used to embed diffracted amplitudes (square-roots of intensities), using two nearest neighbors for calculation of geodesic distances (integrals of the metric). As shown in Fig. 3, a one-dimensional and uniformly populated manifold results,

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

with the projection on the first eigenvector linearly proportional to the snapshot orientation to within 1°, compared with the crystal mosaicity of 0.8°. The homogeneity of this manifold (metric) establishes that, in principle, our symmetry-based approach can be used to treat crystalline objects in the same way as non-periodic single particles, provided, of course, object symmetry is appropriately incorporated.

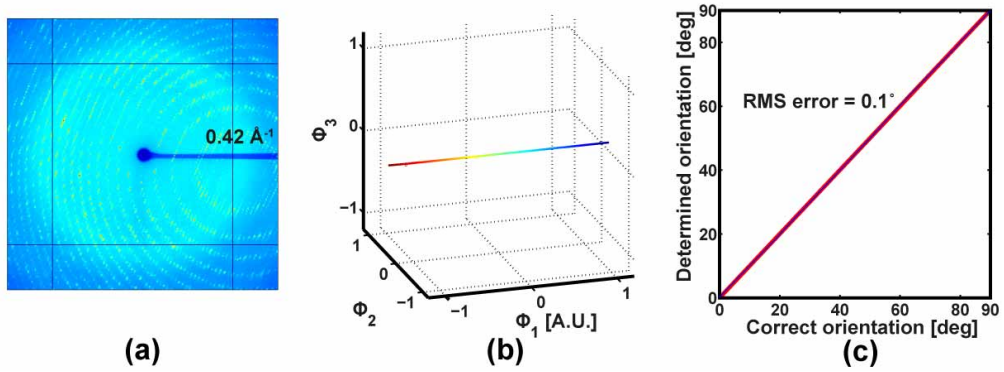

Fig. 3. (a) Typical experimental diffraction pattern of superoxide dismutase-1. (b) The embedding of the geodesic distances in the space of the first three eigenvectors, with each point representing a diffraction pattern. (c) Relation between the correct and the determined orientations.

4.3. Structure recovery from experimental cryo-electron micrographs

A well-studied application of graph-theoretic techniques concerns the 3D reconstruction of faint biological objects by single-particle cryo-EM without orientational information. In cryo-EM, the resolution is strongly degraded by radiation damage. As such, the lowest acceptable exposure to electrons and thus SNR must be employed. As described in Sec. 2, this has proved a fertile ground for testing new algorithms. By recourse to specific properties of cryo-EM images, impressive results have been obtained, primarily with simulated data. Beyond noise, however, reconstruction by cryo-EM must contend with a range of important issues, chief among them the loss of information due to zero-crossings in the transfer function of the microscope and thus partial loss of information in any single snapshot. The exact position of the zero-crossings depends sensitively on microscope defocus. This offers a means to recoup some of the lost information by insuring that the dataset includes micrographs obtained over a range of defocus values, each with a different set of zero-crossings in the transfer function. The key point is this: the object structure cannot be recovered in full detail from a single defocus, even if the imaging parameters were known exactly. Thus, for a reconstruction algorithm to be of practical use, it must deal with the effect of defocus variations — a test rarely passed by new algorithms. Here, we demonstrate structure recovery from experimental cryo-EM images of the biological molecule chaperonin. Specifically, we incorporate the effect of defocus, use the symmetry-based homogeneity of the manifold metric to deduce orientations, and thence recover the 3D object structure. This demonstration is mitigated by two factors: (1) in order to expedite the calculations, snapshots with only one orientational degree of freedom were selected from a set presorted by a standard orientation algorithm; and (2) to demonstrate structure recovery at ultra-low signal — far below what is normally used — experimental snapshots were preprocessed to simulate such low signal levels.

Randomly oriented single-particle cryo-EM images images of the wild-type group II chaperonin in *methanococcus maripaludis* (Mm-cpn), obtained with a mean incident electron count

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

(MEC) of $20/\text{Å}^2$ (equivalent to 135 electrons per 2.6 Å-square snapshot pixel) were kindly provided by Chiu and collaborators [36]. Each snapshot consisted of 96×96 pixels. A set of 413 side-view snapshots was selected from 5000 images, whose orientations had been previously determined by the EMAN program [37], resulting in a dataset with a single orientational degree of freedom about the object symmetry axis.

To investigate the performance of our method at lower dose, a second data set was produced by applying an additional Poisson process to the raw experimental images. The method is based on an approximation valid for low-contrast images with Poisson noise and sufficiently large MEC. The substitution $I \mapsto I' = \text{Pois}(I^{1/2})$ transforms a signal I to a signal I', with mean $\text{MEC}' = \text{MEC}^{1/2}$ and variance $\text{var}(I') = \text{var}(I)/4$. Simulations verified the accurate validity of this approach at MEC = 100, compared with an MEC per snapshot pixel of 135 for our experimental images. Twenty noisy versions of each image were thus generated to form a data set of 8260 images with an effective MEC of 1.7 per Å^2 .

Since neither the noise-free signal nor the noise variance was known for our experimental cryo-EM images, a method developed by Frank [23] was used to estimate the SNR directly from the experimental data. This determines the SNR from the cross-correlation coefficient C_{ij} between two images in the same orientational class using the definition:

$$\text{SNR} = 10 \log_{10} \text{mean}(C_{ij}/(1 - C_{ij})), \tag{4}$$

where the mean is taken over all classes and all images within each class. Provided two images represent different realizations of noise from an identical object in the same orientation, the above estimate for SNR agrees with the standard definition $\text{SNR} = 10 \log_{10} \frac{\text{var}(signal)}{\text{var}(noise)}$ [38]. With the classification obtained from EMAN [37], and the assumption that class members differ only in noise, we estimate a SNR of -6 dB for the raw experimental snapshots (MEC: $20/\text{Å}^2$) and a SNR of -16 dB for the preprocessed experimental snapshots (MEC: $1.7/\text{Å}^2$).

As described above, the defocus value and hence transfer function of the electron microscope vary from snapshot to snapshot. To analyze such cryo-EM data, we implemented a modified version of the manifold-embedding algorithm Generative Topographic Mapping (GTM) [20,39] to explicitly incorporate the effect of the microscope transfer function. GTM defines a manifold in data space by partitioning the noisy dataset into a number of Gaussians each centered around a point (node) on the manifold. The partitioning is based on a nonlinear mapping of a latent space, in this case the space of rotations. GTM is thus, in essence, a manifold-embedding technique, with the symmetries of scattering manifested in the homogeneity of the data manifold, as described in Paper I. However, the generative capability of GTM allows one to construct an image (in essence a model snapshot) at each node on the data manifold. In our approach, this model image extracted from the data corresponds to the aberration-free projected potential of the object. In order to assign an experimental snapshot to a model image, its distance from the model is calculated after convolving the model with the transfer function of the microscope at the defocus corresponding to that of the experimental snapshot. This convolution proceeds efficiently as multiplication in Fourier space, and is not computationally expensive. A similar approach based on more efficient manifold-embedding techniques will be published elsewhere.

The GTM-based approached was first validated with simulated cryo-EM images of chaperonin over a typical experimental defocus range of 10,000 Å to 30,000 Å (underfocus). The orientations were successfully recovered to within 1°. To reconstruct 3D density maps from experimental snapshots, model aberration-free projected potentials were generated (lifted) at 16 equally-spaced nodes of the data manifold produced by experimental images replicated according to the 8-fold object symmetry. 3D density maps were then reconstructed tomographically using the back-projection algorithm BG CG of the SPIDER software package [40]. For comparison, a simulated density map was obtained from 2D snapshots using the known chaperonin

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

atomic coordinates (PDB identifier: 3LOS) under the following imaging conditions: spherical aberration $C_s = 4.1$ mm; defocus $\Delta f = 24{,}000$ Å (underfocus); electron energy $E = 300$ keV; damping envelope parameter $B = 50$ Å^2; images phase-flipped. The resulting 3D density map was passed through a 5 Å Gaussian filter.

Figure 4(a) shows a typical experimental snapshot, Fig. 4(b) the average of the micrographs assigned to an orientation class by the cryo-EM reconstruction software package EMAN [37], and Fig. 4(c) the snapshots oriented by manifold embedding and reconstructed (lifted) from the manifold. (For a movie of the reconstructed tilt series see [5]). Note that the manifold is able to generate missing images by interpolation. The improved quality of the manifold-generated snapshots compared to the class averages offers the possibility to reconstruct at significantly reduced dose. Figure 4(d) is an experimental snapshot preprocessed to approximate snapshots expected from a single chaperonin molecule at a dose $12\times$ lower than commonly used [36] (SNR ~ -16 dB, i.e., 10 dB below a typical dose). Figure 4(e) is the snapshot lifted from the manifold after orienting an ensemble of 8000 different raw snapshots by manifold embedding. It is clear from this image, the corresponding tilt series [6], and the 3D reconstructions of Fig. 4(f), Fig. 4(g), and Fig. 4(h) that snapshots can be successfully oriented by manifold embedding to produce 3D models, even at $12\times$ lower signal than in use today. Note that images at this dose could not be oriented by standard cryo-EM approaches [40], even when accurately centered prior to analysis, as was performed here. In contrast, our orientation recovery results were similar to those obtained at an MEC of $20/$Å^2, indicating that the effect of lower signal levels can be compensated by increasing the number of snapshots.

Our results thus offer the tantalizing possibility of reducing the snapshot dose in 3D reconstruction techniques using ionizing radiation, in some cases by at least an order of magnitude. This would significantly mitigate the limits set by radiation damage. As a benchmark, the essentially unfulfilled promise of the costly transition of cryo-EM to liquid He temperatures was aimed at improving dose tolerance by a factor of two. The superior signal extraction capability offered by manifold mapping could also be used to obtain images at smaller defocus values in order to reconstruct the object to higher resolution.

4.4. Time-series (movies) from ultra-low-signal random-sequence snapshots

Our knowledge of the precise time at which an experimental snapshot of a dynamic system was obtained is corrupted by inevitable uncertainties, which can substantially exceed the intrinsic time resolution of the observation technique. Modern pump-probe experiments, for example, can now be performed with pulses as short as a few femtoseconds, but their time-resolution is often determined by timing jitter, which can be up to two orders of magnitude larger [41, 42]. When the state of a system under observation is not synchronized with the observation windows, a sequence of snapshots can represent random sightings of the system during its evolution. It is thus important to develop means for deducing "time stamps" directly from the data, either to reduce jitter-induced uncertainty, or to order a sequence of snapshots according to the intrinsic evolution of the system under observation. Here, we demonstrate this capability at SNRs as low as -21 dB. Specifically, we show that: (1) Randomized movie sequences can be time-ordered, even when the signal is extremely low; and (2) Frames generated (lifted) from the manifold produce movies of superior quality.

Movies of a pirouette and a pas de deux were downloaded from the web. These represent optical snapshots of a conformationally rigid body in rotation, and the evolution of two flexible bodies in interaction, respectively. In order to reduce the SNR, a constant background was added, and shot noise incorporated at each pixel depending on its intensity value, as described in Ref. [31]. For the pirouette, a sequence of 16 turns consisting of 268 frames (210×160 pixels each) was replicated 132 times, a background $5\times$ the mean intensity added, and shot

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

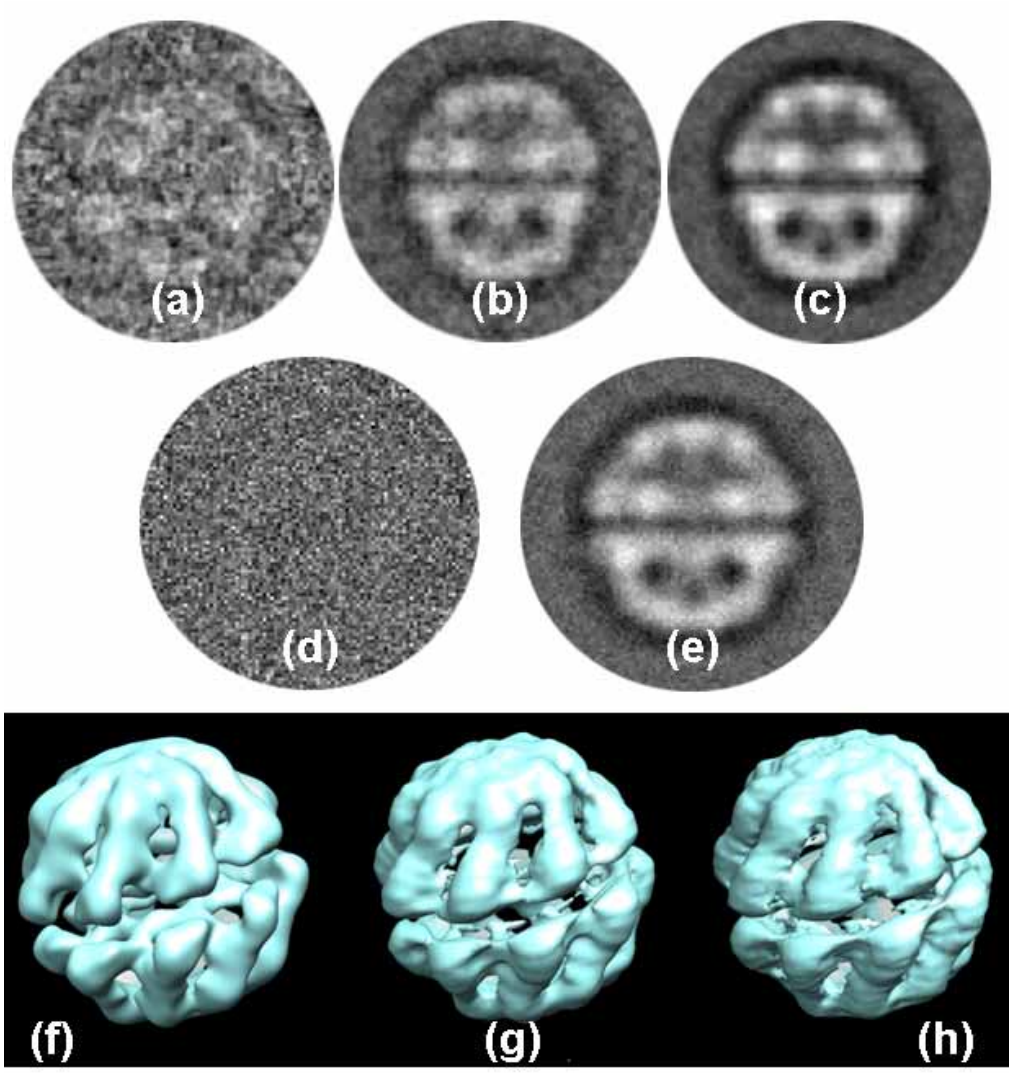

Fig. 4. (a) Experimental cryo-electron micrograph of a chaperonin molecule at a mean electron count of $20/\text{Å}^2$ (SNR $= -6$ dB). (b) Image obtained by averaging the members of an orientation class. (c) Image generated (lifted) from the data manifold. (d) Experimental micrograph of chaperonin molecule processed to reflect a mean electron count of $1.7/\text{Å}^2$ (SNR $= -16$ dB). (e) Image lifted from the data manifold. (f) 3D reconstruction with simulated images. (g) 3D reconstruction with lifted images at a snapshot dose of $20/\text{Å}^2$. (h) 3D reconstruction with lifted images at a snapshot dose of $1.7/\text{Å}^2$.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

noise incorporated to produce an effective mean photon count per pixel of 0.08 and a SNR of -21 dB (see Eq. (4)). For the pas de deux, a sequence of 870 frames (265×305 pixels each) was replicated 12 times with an added background of twice the mean intensity, and shot noise incorporated to produce a mean photon count of 0.8 and a SNR of -11 dB. For both movies, camera motion was corrected by reference to a stationary marker.

Each random sequence was ordered by a suitable manifold-embedding technique (Diffusion Map or Isomap). Using the generative property of GTM, images were then lifted from the manifold. As described in Ref. [4] and demonstrated in Fig. 4(c) and 4(e) , this procedure uses the information content of the entire dataset to generate each snapshot, producing images of significantly higher quality than possible by traditional classifying and averaging techniques. It is also more robust against non-uniform sampling and jitter. Table 2 summarizes the lifting procedure.

Table 2. Manifold-lifting algorithm based on GTM

Inputs:

Noisy snapshots $\mathscr{M}_I = \{\underline{I}_1, \ldots, \underline{I}_s\}$
Estimated quaternions $\mathscr{T} = \{\tau_1, \ldots, \tau_s\}$
number of nodes K
number of basis functions M
basis function width σ

Outputs:

Manifold-lifted images $\mathscr{M} = \{\underline{a}_1, \ldots, \underline{a}_s\}$

1: Generate the grid of latent points $\{x_1, \ldots, x_K\}$.
2: Generate the grid of basis function centers $\{\mu_1, \ldots, \mu_M\}$.
3: Compute the matrix of basis function activations Φ such that

$$\Phi_m(x) = \exp(-(x-\mu_m)^2/2\sigma^2).$$

4: Initialize a set of weights W using principal component analysis.
5: Initialize inverse Gaussian noise variances α and β.
6: Compute a set of responsibilities R that assigns each snapshot to a node from the results of manifold embedding.
7: Compute the diagonal matrix G using R, where $G_{kk} = \sum_{i=1}^{s} R_{ki}$.
8: **repeat**
9: $W \leftarrow (\Phi^{\mathrm{T}} G \Phi + \lambda I)^{-1} \Phi^{\mathrm{T}} R \mathscr{M}_I$, where the regularization parameter λ may be zero.
10: Compute Δ, where $\Delta_{kn} = \|\underline{I}_n - \Phi_k W\|^2$.
11: Compute γ from λ and α.
12: Update α and β using γ, R and Δ.
13: **until** convergence of W
14: $\mathscr{M} \leftarrow \Phi W$
15: **return** $\mathscr{M}$

For the pirouette, Diffusion Map was used to recover the object orientation in each frame during the dance, and hence the sequence order (number of nearest neighbors in the sparse distance matrix $d = 5896$; Gaussian kernel bandwidth $\varepsilon = 200$.) Snapshots were then lifted

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

from the manifold (number of nodes $K = 28$; number of basis functions $M = 14$, basis function width $\sigma = 2$.) Reconstructed images are shown in Fig. 5 together with a sequence of unsorted, unprocessed snapshots. For a movie see [7]. The randomized pas de deux sequence was ordered with Isomap (number of nearest neighbors $d = 33$.) To compile the movie, images were lifted from an ordered sequence of 870 points (nodes), corresponding to uniform sampling on the Isomap manifold. Reconstructed images are shown in Fig. 6 together with a sequence of unsorted snapshots. For a movie see [8].

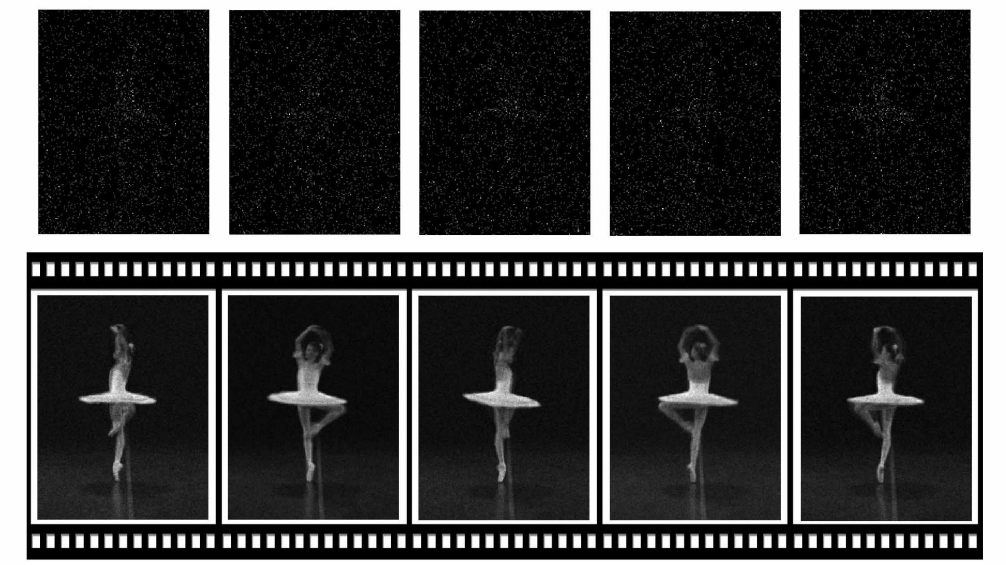

Fig. 5. Top row: First five frames of 35,000 randomly-sequenced snapshots of a pirouette preprocessed to reflect a mean photon count of 0.08 per pixel with added background and shot noise, corresponding to a signal-to-noise ratio of -21 dB. Bottom row: Five evenly-spaced images extracted from the Diffusion Map manifold. (See also [7] for a movie.)

Figures 5 and 6, and the associated movies clearly show that our approach is able to determine the correct frame sequence and generate high quality snapshots at signal levels as low as 0.08 photon/pixel for the pirouette (modulo one revolution), and 0.8 photon/pixel for the pas de deux, both with added background and non-additive noise corresponding to signal-to-noise ratios in the range -11 to -21 dB. These examples demonstrate the capability to determine the time evolution of systems from unsorted random sightings at extremely low signal. They also highlight the potential to correct timing jitter in pump-probe experiments, and reconstruct the evolution of dynamic systems from random sightings of members of a heterogeneous set, each at a different stage of its evolution. These possibilities will be described in detail elsewhere. The general implications for signal extraction and image processing are clear.

5. Conclusions

We have shown that manifold mapping, as described in Paper I, augmented with modest noise reduction measures, is able to extract structural and timing information from simulated and experimental snapshots at extremely low signal. The ability to orient simulated and experimental diffraction and image snapshots confirms the accessibility of the homogeneous manifold expected from our theoretical framework for a wide range of objects and imaging scenarios, including crystalline samples. The capability to recover 3D structure at extremely low signal is on

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

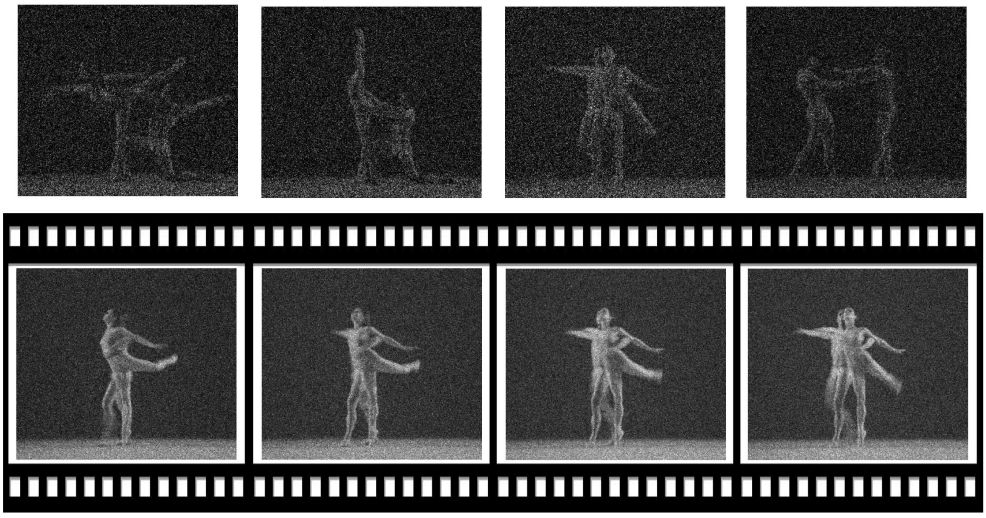

Fig. 6. Top row: First four frames of 10,000 randomly-sequenced snapshots of a pas de deux preprocessed to reflect a mean photon count of 0.8 per pixel with added background and shot noise, corresponding to a signal-to-noise ratio of -11 dB. Bottom row: Four evenly-spaced images extracted from the Isomap manifold. (See also [8] for a movie.)

par with the more expensive Bayesian approaches, but offers greater reach in terms of sample size and resolution, as demonstrated in Paper I. The noise-robustness of our approach substantially exceeds what has been demonstrated with comparable graph-theoretic approaches without restrictive, application-specific assumptions. The manifold itself offers a powerful route to image reconstruction at low signal, because snapshots reconstructed from the manifold achieve higher signal-to-noise ratios than possible by traditional approaches based on classification and averaging. Taken together, these offer a physically-based, computationally efficient, noise-robust route to analyzing the large and varied datasets generated by existing and emerging structure recovery methods. In the longer term, it should be possible to use these approaches to recover or improve timing information in pump-probe experiments, and construct 3D movies (4D maps) from random sightings of members of structurally heterogeneous and dynamically evolving ensembles.

A. Treatment of noise

In the manifold picture, noise can be described as a perturbation of the noise-free manifold $\mathscr{M} = \{\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_s\}$, where $\underline{a}_i$ is a snapshot vector of measured pixel amplitudes, viz.

$$\underline{a}_i \mapsto \tilde{\underline{a}}_i = \kappa \underline{a}_i + \delta \underline{a}_i, \tag{5a}$$

$$\delta \underline{a}_i = (\delta a_{i1}, \delta a_{i2}, \ldots, \delta a_{in})^{\mathrm{T}}, \quad \delta a_{ij} = I_{ij}^{1/2} - \kappa a_{ij}\,. \tag{5b}$$

This causes the observed, noisy, dataset

$$\tilde{\mathscr{M}} = \{\tilde{\underline{a}}_1, \tilde{\underline{a}}_2, \ldots, \tilde{\underline{a}}_s\} \tag{6}$$

not to lie exactly on the manifold $\mathscr{M}$ (up to a global scaling by κ). One would expect that if the perturbation norm $\|\delta \underline{a}_i\|$ becomes comparable to the kernel bandwidth $\varepsilon^{1/2}$, the computed eigenfunctions $\underline{\psi}_k$ are distorted to the point that the embedded manifold no longer has the

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

topology of SO(3) [18, 43]. Indeed, as illustrated in Fig. 1 direct application of the algorithm for noise-free data (see Paper I Table III) to a noisy dataset $\tilde{\mathscr{M}}$ at an MPC $= \mathrm{O}(10^{-2})$ results in poor accuracy. In order to be practically useful, the noise-free orientation-recovery scheme must be augmented by a suitable denoising method.

Conceptually, we denoise the data in three steps: (1) Low-pass filtering of each snapshot by convolution of pixel intensities with a 2D Gaussian; (2) Variance-stabilizing transformation (VST); and (3) Local averaging over nearest neighbors in data space prior to embedding. In practice, we combine (1) and (2), known to be effective for shot noise [44–46], into a single step, and follow the iterative procedure described below.

A.1. Low-pass Gaussian filtering and variance-stabilizing transformation (VST)

Convolution with a low-pass Gaussian filter of bandwidth σ is represented by:

$$\underline{I} \mapsto (I * H_\sigma)(\vec{r}) = \int d\vec{r}\,' I(\vec{r}\,') H_\sigma(\vec{r} - \vec{r}\,'), \tag{7a}$$

with

$$H_\sigma(\vec{r}) = \exp(-\|\vec{r}\|^2/2\sigma^2)/(2\pi\sigma^2)^{1/2}. \tag{7b}$$

VST, proposed by Guan [46] for low-intensity data, can be written as:

$$\underline{I} \mapsto \underline{I}^{1/2} + (\underline{I} + 1)^{1/2}. \tag{8}$$

$I(\vec{r})$ denotes the (discretely sampled) intensity pattern on the detector plane obtained by "unpacking" the column vector of intensities $\underline{I}$.

Equations (7) and (8) are combined into a single operation:

$$\mathrm{VST}(I;\sigma) = (I * H_\sigma)^{1/2} + [(I * H_\sigma) + 1]^{1/2}. \tag{9}$$

Given an intensity-pattern dataset $\mathscr{M}_I$ consisting of s samples and an index matrix of nearest neighbors N of dimensions $s \times l$, we introduce a combined VST and aggregation operation taking $\mathscr{M}_I$ to a dataset $\tilde{\mathscr{M}} = \mathrm{VSTL}(\mathscr{M}_I;\sigma,\mathsf{N})$ such that

$$\tilde{\mathscr{M}} = \{\underline{\tilde{a}}_1, \underline{\tilde{a}}_2, \ldots, \underline{\tilde{a}}_s\}, \quad \underline{\tilde{a}}_i = \mathrm{VST}\left(\sum_{k=1}^{l} \underline{I}_{N_{ik}};\sigma\right). \tag{10}$$

Noise robustness can be further enhanced by a so-called self-tuning Gaussian kernel introduced by Zelnik-Manor and Perona [47]. Here, instead of the isotropic Gaussian kernel $\mathscr{K}(\underline{a}_i,\underline{a}_j) = \exp(-\|\underline{a}_i - \underline{a}_j\|^2/\varepsilon)$ of Eq. (B3) in Paper I, one uses an anisotropic Gaussian kernel with local scaling parameters, ε_i, given by

$$\mathscr{K}(\underline{a}_i,\underline{a}_j) = \exp(-\|\underline{a}_i - \underline{a}_j\|^2/(\varepsilon_i\varepsilon_j)^{1/2}). \tag{11}$$

A canonical choice for the scaling parameters, which we adopt throughout, is $\varepsilon_i = \|\underline{a}_i - \underline{a}_{N_{il}}\|^2$, where, as usual, N_{il} denotes the index of the l-th nearest neighbor of datapoint i.

A.2. Iterative local averaging

If the true nearest neighbors of a point in data space are known, and noise produces no systematic bias, local averaging approaches the true manifold. To see this, let ε be an error tolerance in data space, and consider a reference orientation R with corresponding noise-free snapshot $\underline{a}$. For any $\varepsilon > 0$ it is possible to find a ball B_ε in data space centered at $\underline{a}$, such that for any countable set $\{\underline{a}_1, \ldots, \underline{a}_l\}$ of noise-free snapshots lying in B_ε the mean, $\overline{\underline{a}} = \sum_{i=1}^{l} \underline{a}_i/l$, has error

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

$\|\bar{\underline{a}} - \underline{a}\| < \varepsilon$. In the presence of noise, the $\underline{a}_i$ are replaced by the random variables in Eq. (5a); i.e., $\underline{a}_i \mapsto \tilde{\underline{a}}_i$, where $\tilde{\underline{a}}_i$ are statistically independent, have expectation value $\kappa^{1/2}\underline{a}_i$ proportional to $\underline{a}_i$, and finite variance $\Delta\underline{a}_i^2$. Moreover, the sample mean within B_ε becomes a random variable $\hat{\underline{a}} = \sum_{i=1}^{l} \tilde{\underline{a}}_i / l$ with expectation value $\kappa^{1/2}\bar{\underline{a}}$ and variance $\Delta\underline{a}^2 = \sum_{i=1}^{l} \Delta\underline{a}_i^2 / l$. By the law of large numbers, in the limit of an infinite data set with infinite snapshots in B_ε (i.e., $l \to \infty$), $\hat{\underline{a}}_i$ is equal to $\bar{\underline{a}}_i$ (up to an unimportant proportionality constant) with probability one. Thus, recovery of the data manifold with error ε is possible almost surely.

In practice, noise corrupts the local neighborhood relations, and without *a priori* information, it is not possible to identify which of the snapshots in a noisy data set are associated with the ball B_ε of the underlying noise-free system. Therefore, it cannot be guaranteed that local averaging leads to the correct manifold. We therefore exploit our knowledge of the natural eigenfunctions of scattering manifolds to monitor the effect of local averaging, and terminate the procedure before substantial deviations have occurred.

Specifically, we follow the algorithm described in Table 3. First, we apply the VST operation Eq. (9) to the intensity data $\{\underline{I}_i\} = \mathscr{M}_I$, setting the filter width σ to a relatively small value (e.g., in Sec. 4.1, σ is set to $7/10$ of a pixel width). The autotuning version of the orientation-recovery method (Table 4) is executed using the VST-filtered intensities as input data. The nearest-neighbor indices N_0 obtained in the course of the calculation of the sparse distance matrix then become our initial estimate for the true nearest-neighbor indices. We also record the residual of the nonlinear least-squares output, $\mathscr{G}_0^*$, and choose a value $l \leq d$ for the number of nearest neighbors for local averaging.

Next, we enter an iteration loop, where in the i-th step the dataset

$$\tilde{\mathscr{M}}_i = \mathrm{VSTL}(\mathscr{M}_I; \sigma, \mathsf{N}_{i-1}) \tag{12}$$

is computed, and the algorithm in Table 4 executed using $\tilde{\mathscr{M}}_i$ as input data. The resulting quaternion estimates, nearest-neighbor indices, and least-squares residual are respectively designated $\mathscr{T}_i$, N_i, and $\mathscr{G}_i^*$. Note that the residual $\mathscr{G}_i^*$ is a measure of the difference between the eigenfunctions obtained by embedding and the natural eigenfunctions (Wigner D-functions) expected on the basis of symmetry (see Paper I).

If, after an iteration i, $\mathscr{G}_i^*$ is larger than the residual $\mathscr{G}_{i-1}^*$ encountered in the previous step, the loop is terminated. Otherwise, the iteration is repeated using N_i as an updated estimate of the true nearest-neighbor indices. Our final orientation (quaternion) assignment is the one corresponding to the minimum least-squares residual, reached in the iteration prior to the termination step. As shown in Figs. 7 and 8, the effects of this denoising procedure are benign. The histogram of the Pearson correlations between two million noisy and noise-free snapshots and the associated mean and standard deviation (Fig. 7) show a smooth evolution toward noise-free snapshots as denoising proceeds. This is also evident from visual inspection of the snapshots themselves (Fig. 8).

The empirical evidence in Sec. 4.1 clearly shows that, given a sufficiently large number of sample points, the scheme, applied only a handful of times and terminated using the value of $\mathscr{G}_i^*$ as a criterion, provides noise reduction sufficient for accurate orientation recovery at $\mathrm{MPC} = \mathrm{O}(10^{-2})$.

A potentially fruitful way of interpreting mathematically the success of the process (which lies outside the scope of the present paper) would be to explore its connections with mutually reinforcing models (MRMs) for graph filtering [48]. This type of model involves iteratively replacing vertices of graphs with weighted averages, whereby the vertex itself and its local neighborhood exert an influence on the vertex in the course of iterative updates. In certain applications, the iterative process in an MRM is terminated after only a small number of iterations. Both of these two features are present in the scheme presented here.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

Table 3. Orientation-recovery for noisy snapshots

Inputs:

Noisy snapshots $\mathscr{M}_I = \{\underline{I}_1, \ldots, \underline{I}_s\}$
Number of retained nearest neighbors d
Number of nearest neighbors for local averaging
Number of datapoints in the least-squares fit, r
Number of nearest neighbors for autotuning, n
Gaussian filter bandwidth σ

Outputs:

Estimated quaternions $\mathscr{T} = \{\tau_1, \ldots, \tau_s\}$
Estimated nearest-neighbor index matrix N
Least-squares residual $\mathscr{G}^*$

for $i = 1, \ldots, s$ **do**
    $\tilde{\underline{a}}_i \leftarrow \mathrm{VST}(\underline{I}_i; \sigma)$
end for
$\tilde{\mathscr{M}}_0 \leftarrow \{\tilde{\underline{a}}_i\}$ ▷ initial iterate for Diffusion Map input data
Execute the algorithm in Table 4 with input data $\tilde{\mathscr{M}}_0$; store the returned nearest-neighbor index matrix as N_0 and the least squares residual as $\mathscr{G}_0^*$.
$i \leftarrow 1$ ▷ initialize iteration counter.
terminate $\leftarrow$ false ▷ initialize termination flag.
while *terminate* $\equiv$ false **do**
    $\tilde{\mathscr{M}}_i \leftarrow \mathrm{VSTL}(\mathscr{M}_I; \sigma, \mathsf{N}_{i-1})$ ▷ current iterate for Diffusion Map input-data
    Execute the algorithm in Table 4 with input data $\tilde{\mathscr{M}}_i$; store the outputs as $\mathscr{T}_i$, N_i, and $\mathscr{G}_i^*$.
    terminate $\leftarrow \mathscr{G}_i^* > \mathscr{G}_{i-1}^*$ ▷ terminate if the residual has increased.
    if *terminate* $\equiv$ false **then**
        $i \leftarrow i+1$ ▷ increment iteration counter.
    end if
end while
$\mathscr{T} \leftarrow \mathscr{T}_{i-1}$ ▷ set outputs to the values corresponding to minimum residual.
$\mathscr{G}^* \leftarrow \mathscr{G}_{i-1}^*$
$\mathsf{N} \leftarrow \mathsf{N}_{(i-1)}$
return $\mathscr{T}$, $\mathscr{G}^*$, N.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

Table 4. Orientation-recovery using a self-tuning kernel

Inputs:

Snapshots $\mathscr{M} = \{\underline{a}_1, \ldots, \underline{a}_s\}$
Number of retained nearest neighbors d
Number of datapoints in the least-squares fit, r
Number of nearest neighbors for autotuning, n

Outputs:

Estimated quaternions $\mathscr{T} = \{\tau_1, \ldots, \tau_s\}$,
Nearest-neighbor index matrix N
Least-squares residual $\mathscr{G}^*$

1: Compute the $s \times d$ matrices N and S such that

$$N_{ij} = \text{index of } j\text{-th nearest neighbor to snapshot } \underline{a}_i,$$
$$S_{ij} = \|\underline{a}_i - \underline{a}_{N_{ij}}\|.$$

2: **return** N
3: Rescale the distance data by the n-th nearest neighbors:

$$S_{ij} \leftarrow \left(S_{i,N_{i,n}} S_{j,N_{j,n}}\right)^{1/2}.$$

4: Compute the sparse transition probability matrix P using the algorithm in Table 5 with inputs S, N, ε, and $\alpha = 1$.
5: Solve the sparse eigenvalue problem $\mathsf{P}\underline{\psi}_k = \lambda_k \underline{\psi}_k$ for $0 \le k \le 9$ and $1 = \lambda_0 < \lambda_1 \le \cdots \le \lambda_9$.
6: Solve the nonlinear least-squares problem

$$\mathscr{G}(\{c_{ijk}\}) = \sum_{l=1}^{r} \|\tilde{\mathsf{R}}_l^{\mathrm{T}} \tilde{\mathsf{R}}_l - \mathsf{I}\|^2 + |\det(\tilde{\mathsf{R}}_l) - 1|^2 \text{ with } [\tilde{\mathsf{R}}_l]_{ij} = \sum_{k=1}^{9} c_{ijk} \psi_{lk}.$$

7: **return** $\mathscr{G}^*$
8: **for** $i = 1, \ldots, s$ **do**
9: Compute an approximate SO(3) matrix $\tilde{\mathsf{R}}_i$ for snapshot $\underline{a}_i$
10: Project $\tilde{\mathsf{R}}_i$ to an orthogonal matrix
11: Convert R_i to a unit quaternion τ_i
12: **return** τ_i
13: **end for**

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

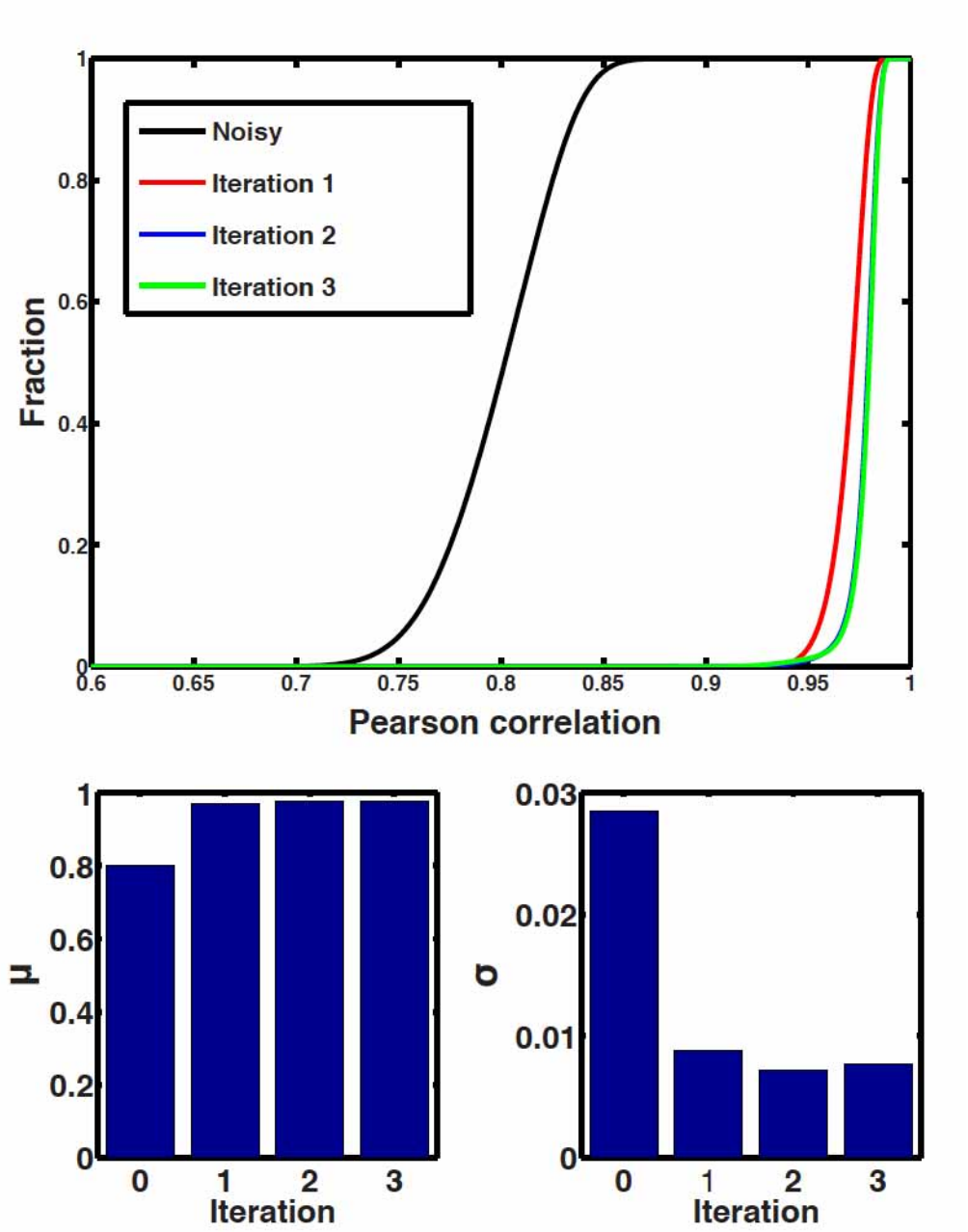

Fig. 7. The effect of denoising on two million images of chignolin at at MPC = 0.04 per Shannon pixel at 1.8 Å resolution. The evolution of the cumulative distribution function of the Pearson correlations with the noise-free images and the associated means and standard deviations demonstrate that noisy images tend toward their noise-free counterparts.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

Table 5. Calculation of the sparse transition probability matrix P_ε in Diffusion Map, reproduced from Paper I for convenience.

Inputs:

$s \times d$ distance matrix S
$s \times d$ nearest-neighbor index matrix N
Gaussian width ε
Normalization parameter α

Outputs:

$s \times s$ sparse transition probability matrix P

1: Construct an $s \times s$ sparse symmetric weight matrix W, such that

$$W_{ij} = \begin{cases} 1, & \text{if } i = j, \\ \exp(-S_{ik}^2/\varepsilon), & \text{if } j = N_{ik}, \\ W_{ji}, & \text{if } W_{ij} \neq 0, \\ 0, & \text{otherwise.} \end{cases}$$

2: Evaluate the $s \times s$ diagonal matrix Q with nonzero elements $Q_{ii} = \sum_{j=1}^{s} W_{ij}$.
3: Form the anisotropic kernel matrix $\mathsf{K} = \mathsf{Q}^{-\alpha}\mathsf{W}\mathsf{Q}^{-\alpha}$.
4: Evaluate the $s \times s$ diagonal matrix D with nonzero elements $D_{ii} = \sum_{j=1}^{s} K_{ij}$.
5: **return** $\mathsf{P}_\varepsilon = \mathsf{D}^{-1}\mathsf{K}$

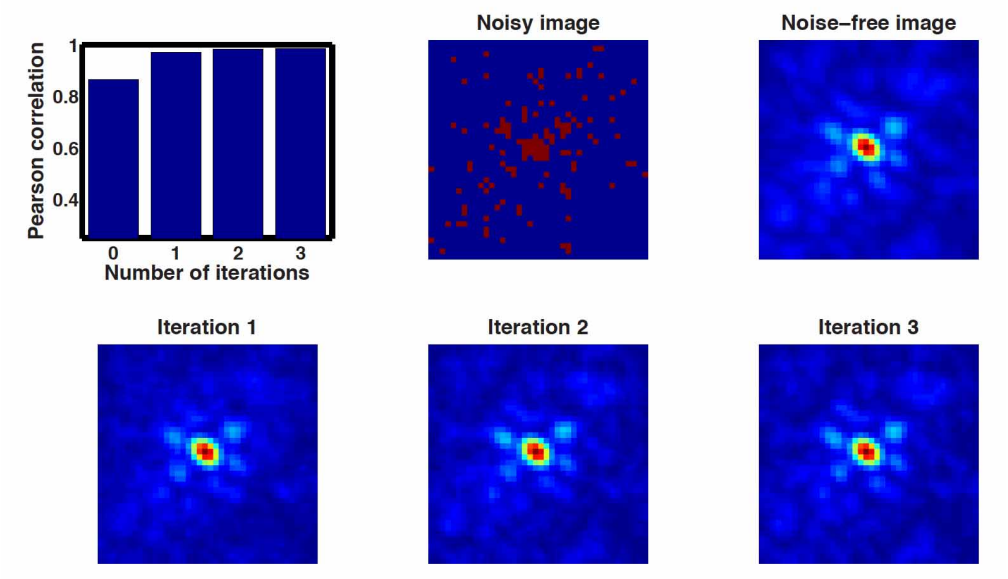

Fig. 8. Evolution of a noisy snapshot of chignolin toward its noise-free counterpart with denoising iteration.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012

B. Computational resources

The calculations reported in this work were primarily performed on a Rocks cluster with 30 nodes, each consisting of two 2.5 GHz Quad-Core Intel Xeon CPUs with 16 GB RAM. Algorithms were usually implemented in MATLAB R2009b with the Parallel Computing Toolbox together with the MATLAB Distributed Server using up to 120 workers (parallel processes). For less intensive calculations, a Linux workstation with a 2.66 GHz Quad-Core Intel Xeon CPU, 32 GB RAM and/or a Mac Pro 2×2.8 GHz Quad-Core Intel Xeon CPUs, 10 GB RAM were used. In Diffusion Map, by far the most CPU-intensive calculations are: (1) the determination of the Euclidean distances of snapshots; and (2) setting up of the sparse distance matrix of nearest neighbors. These calculations were performed in parallel using 100 workers. Such a distance calculation involving 2×10^6 snapshots typically takes 7 hours for chignolin and 48 hours for adenylate kinase (ADK) in Paper I. Other calculations, including the eigenvector determination and the estimation of the orientation matrices were performed on the Linux workstation in about 8 hours altogether. In total, the orientation determination for 2×10^6 snapshots requires 56 hours for noise-free ADK and 33 hours for noisy chignolin with 5 local-averaging iterations. Compiling a 3D diffraction volume consisting of a uniform Cartesian grid was implemented in parallel code, with an execution time of less than three hours for two million ADK diffraction snapshots using 80 workers. Diffraction patterns and cryo-EM images were simulated on the cluster and on the Mac Pro. The GTM and phasing algorithms were performed on the Linux workstation and/or the Mac Pro. The CHIMERA package [49] was used to visualize electron density maps.

Acknowledgments

We are grateful for experimental data and advice to W. Chiu and J. Zhang (cryo-EM images) and M. Schmidt (crystallographic data), and acknowledge discussions with G. N. Phillips, Jr., R. Rosner, W. Schröter, and members of the UWM Physics Department. We are grateful to J. Frank and H. Y. Liao for assistance with 3D reconstruction by back-projection. One of us (DG) is grateful to the Random Shapes Program held in 2007 at the Institute for Pure & Applied Mathematics. This work was partially supported by the U.S. Department of Energy Office of Science (SC-22, BES) award #DE-SC0002164 and a UWM Research Growth Initiative award. PS and DG contributed equally to this work.

#162852 - $15.00 USD Received 9 Feb 2012; revised 11 May 2012; accepted 16 May 2012; published 23 May 2012