nature physics

ARTICLES

PUBLISHED ONLINE: 9 NOVEMBER 2008 | DOI: 10.1038/NPHYS1129

Structure from fleeting illumination of faint spinning objects in flight

Russell Fung, Valentin Shneerson, Dilano K. Saldin and Abbas Ourmazd*

Moves are afoot to illuminate particles in flight with powerful X-ray bursts, to determine the structure of single molecules, viruses and nanoparticles. This would circumvent important limitations of current techniques, including the need to condense molecules into pure crystals. Proposals to reconstruct the molecular structure from diffraction 'snapshots' of unknown orientation, however, require ~1,000 times more signal than available from next-generation sources. Using a new approach, we demonstrate the recovery of the structure of a weakly scattering macromolecule at the anticipated next-generation X-ray source intensities. Our work closes a critical gap in determining the structure of single molecules and nanoparticles by X-ray methods, and opens the way to reconstructing the structure of spinning, or randomly oriented objects at extremely low signal levels.

The X-ray signal scattered by a single molecule is so faint that crystals containing at least 10^8 molecules have had to be used to determine molecular structure. Diffraction-quality crystals are hard to produce, and complicate retrieval of the information of interest: the structure of the molecule. Ideally, one would like to dispense with the need for crystals. This has led to proposals to use powerful next-generation X-ray sources, such as X-ray free-electron lasers (XFELs), to determine the structure of individual (that is, not crystallized) macromolecules and nanoparticles[1–7]. A train of identical objects would be successively exposed to powerful X-ray pulses, and diffraction 'snapshots' collected from single objects of unknown orientation. The diffraction patterns would be oriented relative to each other and used to reconstruct the three-dimensional (3D) diffracted intensity distribution (the diffraction volume) in reciprocal space. In general, the object structure can then be determined through iterative 'phasing algorithms'[8–11].

The key difficulties with this approach stem from the small number of photons scattered by a single molecule. A 500 kD biological molecule in the beam of a next-generation XFEL focused to 0.1 μm diameter, for example, scatters $\sim 4 \times 10^{-2}$ photons per pulse into a pixel at 1.8 Å resolution[12] (Fig. 1). Determining the orientations of the individual low-signal diffraction patterns and hence reconstructing the diffraction volume by proposed approaches require 1,000 times more signal than available[12]. To circumvent this difficulty, suggestions have been made first to orientationally classify and average the individual diffraction patterns in each class to improve the signal-to-noise ratio, and then determine the orientation of each class. The per-shot dose needed to 'classify' exceeds the available photon flux by two orders of magnitude[12,13]. These problems stem, in essence, from reliance on the very limited information available in one, or a few low-signal-to-noise diffraction patterns. No algorithm capable of reconstructing the structure of such faint objects to high resolution has been demonstrated.

Here, we present a new approach, which exploits the correlations in the entire scattered photon ensemble to recover the structure of a faint object to high resolution from scattering snapshots of random orientation, and demonstrate its power by recovering the structure of a single biological molecule to 1.8 Å. This serves as a proof of principle, and also closes a key conceptual and algorithmic gap in planned 'single-molecule' experiments.

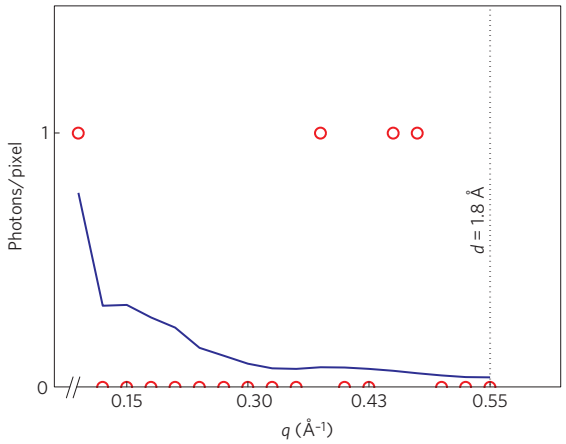

Figure 1 | Diffracted intensity. Number of photons scattered versus wave vector q by the test molecule chignolin. Solid curve: Noise-free signal averaged over all pixels at magnitude q, scaled to give a mean photon count of 4×10^{-2} per pixel at 1.8 Å resolution. Red circles: Actual signal along a typical wave vector **q** with shot noise. The signal level corresponds to that expected from a 500 kD biological molecule exposed to a single pulse of an XFEL. The higher-intensity, information-poor regions at small wave vector are not shown.

We begin by outlining the conceptual framework of our approach, first non-mathematically, then in more detail. This is followed by demonstrating the ability to orient simulated low-signal diffraction patterns from a small test molecule, chignolin[14] (Protein Data Bank designation: 1UAO) down to a scattered mean photon count (MPC) of 4×10^{-2} per diffraction pattern pixel at 1.8 Å with shot noise—the signal-to-noise level expected for a 500 kD molecule. Finally, by recovering the structure of the test molecule from a collection of simulated noisy diffraction patterns of unknown orientation, it is shown that the orientational accuracy we achieve is sufficient for structure determination to high resolution.

Department of Physics, University of Wisconsin Milwaukee, 1900 E. Kenwood Blvd, Milwaukee, Wisconsin 53211, USA. *e-mail: ourmazd@uwm.edu.

Conceptual framework

A loose but illuminating analogy can be used to illustrate our approach. Consider the eroded fragments of an ancient Greek vase recovered in a dig. The vase can be reconstructed from the correlations between the fragments. The most likely shape is obtained when the eroded pieces are maximally correlated with each other. For best results, the correlations considered should not be limited to the shapes of neighbouring fragments, but include the elaborate patterns spanning all of the fragments. In other words, correlations in the entire data set must be considered simultaneously. This is the basis of our approach: we exploit the correlations in the entire ensemble of diffracted photons to reconstruct the 3D diffraction volume. An iterative phasing algorithm is then used to recover the molecular structure.

A compact nomenclature is needed to describe our approach in more detail. Although the algorithm does not rely on any particular data representation, consider the ensemble of scattered photons as a collection of diffraction patterns, each emanating from a random orientation of the object. (In reality, the diffraction patterns stem from randomly oriented members of a set of objects assumed to be identical.) The nomenclature consists of representing each snapshot by a vector, the components of which are the measured intensity values at the pixels of the snapshot (see Supplementary Information, Fig. S1). The diffracted photon ensemble is then a matrix consisting of the ensemble of diffraction pattern vectors. As described in the Methods section, individual pixels in each diffraction pattern span the interval needed to ensure optimum information capture—'oversampling' in the sense of refs 3,8–11 . (We call a gauge satisfying the sampling requirement an 'appropriate sampling gauge.')

A molecule in a specific orientation gives rise to a vector in the so-called 'manifest space' of measured pixel intensities (Fig. 2). As the molecular orientation is changed in the hidden or 'latent space' of orientations, the vector representing the diffraction pattern traces out a path in the p-dimensional manifest space of measured intensities. Because the molecule resides in 3D space, it has only three orientational degrees of freedom. Thus, the tip of the vector in the p-dimensional intensity space is confined to a 3D manifold. To translate a particular position on the manifold in the manifest intensity space to a specific orientation, that is, a specific point in the 3D latent space of orientations, we must determine the mapping between the manifold and the latent space of orientations. This can be done by embedding a 3D manifold in the manifest space so as to include all vector tips to within noise, subject to the constraints imposed by the geometry of the latent space. Once this function is known, the position of each vector in the manifest intensity space can be directly related to a point in the latent space of orientations, that is, to a specific molecular orientation.

A number of manifold-embedding techniques are available[15,16]. We use generative topographic mapping (GTM), a Bayesian nonlinear factor-analytical approach originally developed for data projection and visualization[17–20] and neural network (see, for example, ref. 21 and references therein) applications. This approach determines the maximum likelihood manifold in the manifest space of experimental intensity measurements by fitting the correlations in the diffracted photon ensemble, subject to the constraints imposed by the geometry of the latent space. Through its discrete treatment of the latent and manifest spaces, GTM enables natural classification of similar patterns into orientational classes, and thus noise reduction through averaging. We note, however, that averaging is carried out after the orientation of each diffraction pattern 'snapshot' has been determined. In other words, GTM functions at the actual experimental signal-to-noise level without the need for prior classification and averaging. This is a key attribute. Further details are provided in Supplementary Information.

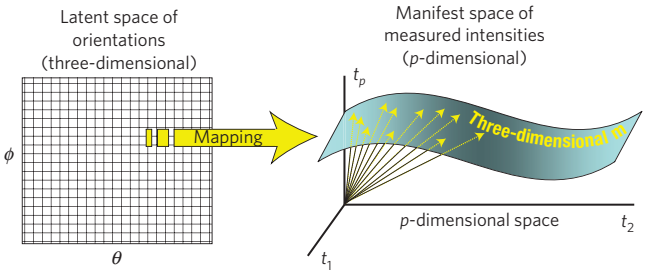

Figure 2 | Latent and manifest spaces. The relationship between the three-dimensional latent (or hidden) space of orientations and the p-dimensional manifest (or accessible) space of measured intensities. The molecular orientation has only three degrees of freedom. As the molecule rotates, the tip of the vector representing its diffraction pattern is confined to a 3D manifold in the pD manifest space. This manifold is a nonlinear mapping of the space of orientations. The mapping function can be determined by well-known manifold-embedding techniques, and used to relate a particular diffraction pattern to a specific orientation of the molecule.

Determining the orientations of diffraction patterns

We now demonstrate the ability to recover the orientations of simulated diffraction patterns of the protein chignolin at signal-to-noise levels corresponding to the data shown by the red circles in Fig. 1, using no information other than the dimensionality of the orientational space. Consider first the case where the molecule can assume any orientation about one axis. In this case, the appropriate sampling angle, the natural length scale for the orientational accuracy needed for 1.8 Å resolution is 3.2°. Figure 3 shows a plot of the determined versus actual orientations for a collection of 3,000 simulated diffraction patterns (1) with no noise (infinite signal) and (2) at the signal-to-noise corresponding to the data shown by the red circles in Fig. 1. The noise-free case produces a root-mean-square (r.m.s.) orientation error of 1.4°. (The minimum error due to discretization of orientational space into 3.2° bins is 0.9°.) When the signal is reduced to that of the data shown by the red circles in Fig. 1, the r.m.s. orientation error amounts to 3.8°. The same accuracy was achieved when 72,000 diffraction patterns were analysed. Diffraction patterns were therefore oriented to within 1.2 appropriate sampling angles. As shown below, this is ample for structure determination to high resolution.

Next, consider the case where the molecule can assume any orientation in 3D space. The possible orientations are now represented by points on the surface of the unit 4-sphere[22]. With appropriate sampling for 1.8 Å resolution, $\sim 10^5$ distinct orientations must be recognized, requiring $\sim 10^6$ diffraction patterns. This exceeds our current desktop computational capabilities. We have therefore limited our simulations to random orientations over $30° \times 30° \times 30°$ patches of the surface of the unit 4-sphere. For a set of 10^3 diffraction patterns with a signal level of the data shown by the red circles in Fig. 1, the r.m.s. error in orientation determination is 5.2°. (For further details, see Supplementary Information, Fig. S7.) When the molecule is free to assume any orientation in 3D, the appropriate sampling angle for 1.8 Å resolution is 5°. The orientational accuracy achieved is thus 1.04 times the appropriate sampling angle.

Recovering the structure of the molecule

As noted earlier, a full 3D orientational recovery is beyond our current computing resources. In principle, however, the 3D molecular structure can be deduced from diffraction patterns obtained when the molecule is free to assume any orientation about a single axis. In practice, the curvature of the Ewald sphere means that only part of the diffraction volume is covered by

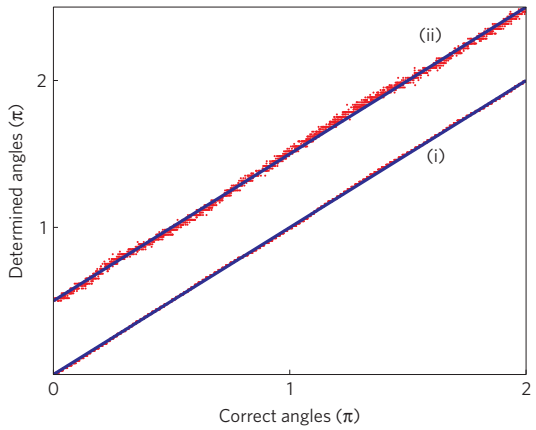

Figure 3 | Orientation recovery. Plot of determined versus actual orientations (modulo 2π) for 3,000 simulated diffraction patterns. (i) No noise (infinite signal). (ii) MPC of 4×10^{-2} per pixel at 1.8 Å resolution, with shot noise. The red dots represent actual results, the blue lines best linear fits. The *y*-intercepts represent unimportant rigid rotations of the molecule.

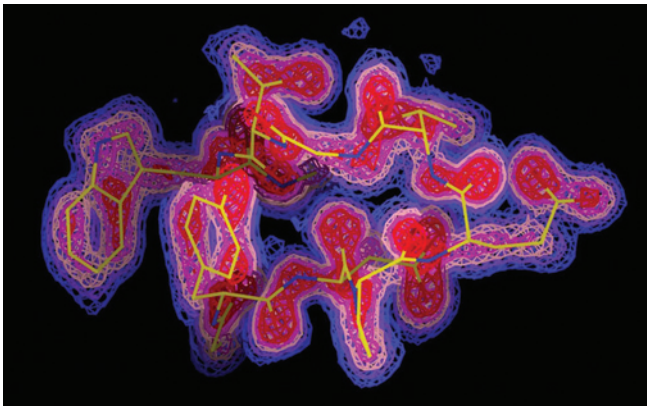

Figure 4 | Structure recovery. Isosurfaces of electron density of the protein chignolin, recovered from 72,000 diffraction patterns of unknown orientation at a MPC of 4×10^{-2} per pixel (see text). The molecular model is represented by the stick figure, with C bonds shown in yellow, N in blue and O in red. The 1, 2 and 3σ electron density contours are shown in blue, pink and red, respectively, with σ denoting the r.m.s. deviation from the mean electron density.

the ensemble of diffraction patterns, leaving regions devoid of diffraction data. These gaps can be eliminated by allowing the molecule to assume any orientation about each of two orthogonal axes in turn. We therefore used the following procedure to recover the 3D structure of the test molecule chignolin. (1) With the beam along the positive z direction and the molecule free successively to assume any orientation about the x and the y axis, a total of 72,000 diffraction patterns were simulated for random orientations of the molecule. (2) The orientations of the molecule were determined from the diffraction patterns. At the signal level corresponding to the data shown by the red circles in Fig. 1, the r.m.s. orientational error was 1.2 times the appropriate sampling angle for 1.8 Å resolution. (3) The diffraction patterns belonging to the same orientational classes, each spanning the appropriate sampling angle were averaged. (4) The data were combined to produce a diffraction volume on a regular Cartesian grid of points in reciprocal space. (5) An iterative phasing algorithm[9] with 'charge flipping' of low electron densities[23] and 'phase shifting' of weak reflections[24] was used to recover the electron density shown in Fig. 4. It is clear that the orientational accuracy achieved is ample for high-resolution structure recovery at very low signal levels.

Implications and outlook

It is important to estimate the range of particle sizes amenable to our approach. The lower limit is set by the number of photons scattered to large angle. This varies as $N_{\mathrm{atoms}}^{1/3}$ (ref. 12), where N_{atoms} is the number of (non-hydrogen) atoms in the molecule, in effect, the molecular weight. A 500 kD molecule scatters $\sim 4 \times 10^{-2}$ photons per pixel at 1.8 Å resolution, with heavier molecules producing larger signals. The MPC of 10^{-2} per pixel, the smallest signal level at which we can at present recover orientation, corresponds to a molecular weight of $500\,\mathrm{kD} \times (10^{-2}/4 \times 10^{-2})^3 \sim 8\,\mathrm{kD}$. This represents our current lower limit for the molecular weight, with the upper limit unbounded by intensity considerations.

Available computational resources set the upper bound. This stems from the increasingly tight orientational accuracy needed for larger objects. To image a particle with diameter D to resolution d, the number of independent orientations to be recognized is given by $N_{\mathrm{nodes}} = (8\pi^2/S)(D/d)^3 \approx (8\pi^2 a^3 N_{\mathrm{atoms}}/S d^3)$, where S denotes the number of asymmetric units in the particle, and a the interatomic spacing. We have characterized the computational requirements of the elementary steps in our approach, and conducted a feasibility study of the resources needed for large molecules and nanoparticles. Assuming appropriate modifications to the present code, a 100-node computing cluster (each node a 2.33 GHz Intel Core 2 Duo with 4 GB random-access memory), and no symmetry in the molecule ($S = 1$), it should be possible to recover the structure of a 500 kD molecule to 3 Å, a 1 MD molecule to 4 Å and a 2 MD molecule to 5 Å, respectively. This range includes important macromolecules, nanoparticles and colloids. Clearly, the computational cost for recovering the structure of symmetric particles is lower. For example, reconstructing the diffraction volume for the satellite tobacco necrosis virus (Protein Data Bank designation: 2buk) to 1 nm is no more difficult than doing so for chignolin to 1.8 Å, the example used in this article.

The approach we have outlined accurately determines the orientations of diffraction patterns at very low signal levels, thus filling a critical gap in the proposed single-particle experiments. These proposals have assumed a single conformational state for the molecules exposed to X-ray pulses. No means have been suggested to deal with cases where this assumption is not valid. In our approach, if the beam of molecules consists of a number of distinct conformations (or a number of different molecular types), each should produce a different manifold in the manifest intensity space. It should be possible to fit a manifold to each type separately to determine the structure of a number of molecular (sub)types, or include structural variations as an extra latent variable and study structural variations within a given species. This might mitigate the need for conformational and/or chemical homogeneity, and potentially enable the study of reactions.

Our algorithm exploits the entire diffracted photon ensemble. So far, we have been able to determine particle orientations at MPCs as low as 10^{-2} per pixel using a total of 10^5 scattered photons. A 500 kD particle in an XFEL beam can scatter $\sim 10^9$ photons to high angle in a few minutes[12]. It may therefore be possible to trade the per-pulse dose against the total number of diffraction patterns recorded, enabling the former to be reduced. If the per-pulse dose can indeed be reduced to below the single-molecule damage threshold, the data collection window increases from the 20 fs (refs 25,26) anticipated at present to the 100 ps–10 ns range, depending on the molecular size and rotational energy. Lower per-pulse doses might also bring single-particle structure determination within range of non-FEL X-ray sources, albeit at reduced resolution. Two questions then arise. (1) What is the lowest practical per-pulse dose needed for structure recovery? (2) Is this dose below the 'acceptable' damage threshold of a single molecule?

A practical application of our approach also requires the ability to maintain sensitivity in the presence of background scattering, which is instrument dependent. These issues highlight important directions for future work.

Finally, our approach reconstructs an object from sections of any shape and dimension with no orientational information. Other potential applications therefore include, but are not limited to, the reconstruction of faint, radiation-sensitive objects by ultralow-dose electron microscopy, diffraction imaging of nanoparticles and colloids and rapid tomography of faint macroscopic objects.

Methods

Unless otherwise stated, the pixel size is the 'appropriate sampling pixel,' which 'oversamples' the diffracted amplitude by a factor of $\sqrt[s]{2}$, with s being the dimensionality of the space considered (1, 2 or 3). The angle subtended by such a pixel is the 'appropriate sampling angle' $\Delta\theta = (d/D\sqrt[s]{2})$, where d is the resolution, and D the diameter of the particle. For our test molecule, the 1D and 3D appropriate sampling angles for 1.8 Å resolution are 3.2° and 5°, respectively. With the molecule free to assume any orientation about one axis, the MPC is 4×10^{-2} per pixel at 1.8 Å. The 1D sampling pixel size was used in the iterative phasing algorithm.

At high angle, in the so-called Wilson statistics regime, the signal averaged over a diffraction shell at a given distance from the origin is essentially independent of the detailed structure of the molecule. We scale the entire intensity distribution such that its value at 1.8 Å resolution is 4×10^{-2} photons per pixel (with $s = 1$). Note that the photon count of 1.4×10^{-2} at 'large angle' in ref. 12 corresponds to 4×10^{-2} at 1.8 Å.

No thermal broadening was included in the calculations, because the appropriate value for single molecules is not known. Using the value typical of biological crystals reduces the calculated intensity at 1.8 Å by less than a factor of 2, well within the accuracy of diffracted intensity estimates[12]. (40×40) pixel diffraction patterns of the protein chignolin in random orientations were simulated out to a scattering angle corresponding to 1.8 Å resolution using an incident photon wavelength of 1 Å. Shot noise was incorporated as Poisson statistics. The incident photon intensity was successively reduced so as to produce down to 10^{-2} scattered photons per pixel at 1.8 Å. The innermost central pixels were not used for analysis, because, despite their higher photon counts, they contain little orientational information. Up to $\sim 10^3$ pixels from each diffraction pattern were provided to the algorithm with no information other than the dimensionality of the orientational space. The pixels stemmed from rectangular strips at the perimeter, annuli excluding the innermost central pixels or pixels with the highest variance across all diffraction patterns. Typically, the 10^3 pixels used by the program contained a total of ~100 photons.

Received 16 June 2008; accepted 9 October 2008; published online 9 November 2008

References

1. Solem, J. C. & Baldwin, G. C. Microholography of living organisms. *Science* **218**, 229–235 (1982).
2. Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752–757 (2000).
3. Miao, J., Hodgson, K. O. & Sayre, D. Extending the methodology of X-ray crystallography to allow imaging of micrometer-sized non-crystalline specimens. *Proc. Natl Acad. Sci.* **98**, 6641–6645 (2001).
4. Huldt, G., Szőke, A. & Hajdu, J. Diffraction imaging of single particles and biomolecules. *J. Struct. Biol.* **144**, 219–227 (2003).
5. Gaffney, K. J. & Chapman, H. N. Imaging atomic structure and dynamics with ultrafast X-ray scattering. *Science* **316**, 1444–1448 (2007).
6. Chapman, H. N. *et al.* Femtosecond diffractive imaging with a soft-X-ray free-electron laser. *Nature Phys.* **2**, 839–843 (2006).
7. Chapman, H. N. *et al.* Femtosecond time-delay X-ray holography. *Nature* **448**, 676–679 (2007).
8. Gerchberg, R. W. & Saxton, W. O. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972).
9. Fienup, J. R. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.* **3**, 27–29 (1978).
10. Sayre, D. in *Imaging Processes Coherence in Physics* Vol. 112 (eds Schlenker, M. *et al.*) 229–235 (Lecture Notes in Physics, Springer, 1980).
11. Miao, J., Charalambous, P., Kirz, J. & Sayre, D. Extending the methodology of X-ray crystallography to allow imaging of micrometer-sized non-crystalline specimens. *Nature* **400**, 342–344 (1999).
12. Shneerson, V. L., Ourmazd, A. & Saldin, D. K. Crystallography without crystals. I. The common-line method for assembling a three-dimensional diffraction volume from single-particle scattering. *Acta Crystallogr. A* **64**, 303–315 (2008).
13. Bortel, G. & Faigel, G. Classification of continuous diffraction patterns: A numerical study. *J. Struct. Biol.* **158**, 10–18 (2007).
14. Honda, S., Yamasaki, K., Sawada, Y. & Morii, H. 10 residue folded peptide designed by segment statistics. *Structure* **12**, 1507–1518 (2004).
15. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
16. Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl Acad. Sci.* **100**, 5591–5596 (2003).
17. Bishop, C. M. *Neural Networks for Pattern Recognition* (Oxford Univ. Press, 1995).
18. Bishop, C. M. in *Learning in Graphical Models* (ed. Jordan, M. I.) 371–403 (MIT Press, 1999).
19. Svensen, J. F. M. *The Generative Topographic Mapping*. Thesis, Aston Univ. (1998).
20. Bishop, C. M. & Tipping, M. E. A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Recognition Machine Intelligence* **20**, 281–293 (1998).
21. Howard, R. E., Jackel, L. D. & Graf, H. P. Electronic neural networks. *J. Am. Telephone Telegraph Co.* **67**, 58–64 (1988).
22. Kuipers, J. B. *Quaternions and Rotation Sequences* (Princeton Univ. Press, 1999).
23. Oszlányi, G. & Sütő, A. *Ab initio* structure solution by charge flipping. *Acta Crystallogr. A* **60**, 134–141 (2004).
24. Oszlányi, G. & Sütő, A. *Ab initio* structure solution by charge flipping. II. Use of weak reflections. *Acta Crystallogr. A* **61**, 147–152 (2005).
25. Neutze, R., Huldt, G., Hajdu, J. & van der Spoel, D. Potential impact of an X-ray free electron laser on structural biology. *Rad. Phys. Chem* **71**, 905–916 (2004).
26. Jurek, Z., Faigel, G. & Tegze, M. Dynamics in a cluster under the influence of intense femtosecond hard X-ray pulses. *Eur. Phys. J. D* **29**, 217–229 (2004).

Acknowledgements

We acknowledge valuable discussions with M. Schmidt and P. Schwander. We are grateful to V. Elser for stimulating us to think about general methods for determining orientations, and to D. Starodub for the suggestion to consider the application of our approach to multicrystalline materials.

Additional information

Supplementary Information accompanies this paper on www.nature.com/naturephysics. Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions. Correspondence and requests for materials should be addressed to A.O.