Protein crystal structure obtained at 2.9 Å resolution from injecting bacterial cells into an X-ray free-electron laser beam

Michael R. Sawaya[a,b,1], Duilio Cascio[a,b,1], Mari Gingery[a,b,1], Jose Rodriguez[a,b], Lukasz Goldschmidt[a,b], Jacques-Philippe Colletier[c,d,e], Marc M. Messerschmidt[f,2], Sébastien Boutet[f], Jason E. Koglin[f], Garth J. Williams[f], Aaron S. Brewster[g], Karol Nass[h], Johan Hattne[g], Sabine Botha[h], R. Bruce Doak[h,i], Robert L. Shoeman[h], Daniel P. DePonte[f], Hyun-Woo Park[j,3], Brian A. Federici[j,k], Nicholas K. Sauter[g], Ilme Schlichting[h], and David S. Eisenberg[a,b,l,4]

[a]UCLA–DOE Institute for Genomics and Proteomics, [b]Department of Biological Chemistry, and [l]Howard Hughes Medical Institute, University of California, Los Angeles, CA 90095-1570; [c]Université Grenoble Alpes, [d]Centre National de la Recherche Scientifique, and [e]Commissariat à l'Energie Atomique, Institut de Biologie Structurale, F-38044 Grenoble, France; [f]Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, CA 94025; [g]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [h]Max Planck Institute for Medical Research, 69120 Heidelberg, Germany; [i]Department of Physics, Arizona State University, Tempe, AZ 85287; and [j]Department of Entomology and [k]Graduate Program in Cell, Molecular and Developmental Biology, University of California, Riverside, CA 92521

Contributed by David S. Eisenberg, July 23, 2014 (sent for review April 22, 2014)

It has long been known that toxins produced by *Bacillus thuringiensis* (Bt) are stored in the bacterial cells in crystalline form. Here we describe the structure determination of the Cry3A toxin found naturally crystallized within Bt cells. When whole Bt cells were streamed into an X-ray free-electron laser beam we found that scattering from other cell components did not obscure diffraction from the crystals. The resolution limits of the best diffraction images collected from cells were the same as from isolated crystals. The integrity of the cells at the moment of diffraction is unclear; however, given the short time (~5 μs) between exiting the injector to intersecting with the X-ray beam, our result is a 2.9-Å-resolution structure of a crystalline protein as it exists in a living cell. The study suggests that authentic in vivo diffraction studies can produce atomic-level structural information.

XFEL | Cry3A insecticidal toxin | serial femtosecond crystallography

The advent of X-ray free-electron lasers (XFELs) has made it possible to obtain atomic resolution macromolecular structures from crystals with sizes approximating only 1/60th of the volume of a single red blood cell. Brief, intense pulses of coherent X-rays, focused on a spot of 3-μm diameter, have produced 1.9-Å-resolution diffraction data from a stream of lysozyme crystals, each crystal no bigger than 3 μm^3 (1). A stream of crystals, not just one crystal, is required to collect the many tens of thousands of diffraction patterns that compose a complete data set. No single crystal can contribute more than one diffraction pattern because the XFEL beam is so intense and the crystals so small that the crystals are typically vaporized after a single pulse. Impressively, a photosystem I crystal no bigger than 10 unit cells (300 nm) on an edge produced observable subsidiary diffraction peaks between Bragg reflections, details which would be unobservable from conventionally sized crystals (2). With this new ability to collect diffraction patterns from crystals of unprecedentedly small dimensions, it is conceivable that high-resolution diffraction data could be collected from crystals in vivo. The structure obtained in this manner would be unaltered from that occurring naturally in a living cell, free from distortion that might otherwise potentially arise from nonphysiological conditions imposed by recrystallization. A practical advantage would also be gained by eliminating the need for a protein purification step, whether the in vivo grown crystals were naturally, or heterologously expressed (3).

The nascent field of serial femtosecond crystallography (SFX) has published results on nine different macromolecular systems since its inception in 2009 (Table 1). One system in particular, cathepsin B, marks an advancement toward in vivo crystallography (3, 9). The crystals for this study were *not* grown in artificial crystallization chambers as has been the protocol of conventional macromolecular crystallography since the 1950s. Instead, crystals were grown in cells. Specifically, they were grown in Sf9 insect cells, heterologously expressing *Trypanosoma brucei* cathepsin B. These in vivo-grown crystals were used for the XFEL diffraction experiment. To this end, the cells were lysed and the crystals were extracted before injecting them in the XFEL beam for data collection. This last purification step seems to be the only major departure from our goal of obtaining high-resolution structural information from crystal inclusions in vivo, without requiring the crystal to be extracted from the cell that assembled it. Here we attempt to go one step further than previous studies—to record diffraction from crystals within living cells.

Significance

In vivo microcrystals have been observed in prokaryotic and eukaryotic cells. With rare exception, however, the ~100,000 biological structures determined by X-ray crystallography to date have required the macromolecule under study to be extracted from the cells that produced it and crystallized in vitro. In vivo crystals present a challenge for structure determination and pose the question of the extent to which in vivo macromolecular structures are similar to those of extracted and recrystallized macromolecules. Here we show that serial femtosecond crystallography enabled by a free-electron laser yields the structure of in vivo crystals, as they exist in a living cell, and in this case the in vivo structure is essentially identical to the structure of extracted and recrystallized protein.

Author contributions: M.R.S., D.C., M.G., and D.S.E. designed research; M.R.S., D.C., M.G., J.R., M.M.M., S. Boutet, J.E.K., G.J.W., S. Botha, R.B.D., R.L.S., and D.P.D. performed research; J.R., L.G., R.B.D., H.-W.P., B.A.F., and N.K.S. contributed new reagents/analytic tools; M.R.S., D.C., J.R., J.-P.C., A.S.B., K.N., J.H., N.K.S., I.S., and D.S.E. analyzed data; and M.R.S., D.C., M.G., A.S.B., K.N., N.K.S., I.S., and D.S.E. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The atomic coordinates and structure factors have been deposited in the Protein Data Bank, www.pdb.org (PDB ID codes 4QX0, 4QX1, 4QX2, and 4QX3).

[1]M.R.S., D.C., and M.G. contributed equally to this work.

[2]Present address: National Science Foundation BioXFEL Science and Technology Center, Buffalo, NY 14203.

[3]Present address: Department of Natural and Mathematical Sciences, California Baptist University, Riverside, CA 92504.

[4]To whom correspondence should be addressed. Email: david@mbi.ucla.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413456111/-/DCSupplemental.

Table 1. SFX publications from XFEL sources to date

Publication date	System	Product	Resolution (Å)	Title of publication	Authors	Reference
Feb 2011*	Photosystem I	Structure	8.7	Femtosecond X-ray protein nanocrystallography	Chapman et al.	2
Dec 2011*	Lysozyme	Structure	8.7	Radiation damage in protein serial femtosecond crystallography using an X-ray free-electron laser	Lomb et al.	4
Jan 2012*	Photosystem I-Ferredoxin	Data	11	Time-resolved protein nanocrystallography using an X-ray free-electron laser	Aquila et al.	5
Jan 2012*	Cathepsin B	Data	7.5	In vivo protein crystallization opens new routes in structural biology	Koopman et al.	3
Jan 2012*	Photosynthetic Reaction Center	Structure	7.4	Lipidic phase membrane protein serial femtosecond crystallography	Johansson et al.	6
Jun 2012	Photosystem II	Structure	6.6	Room temperature femtosecond X-ray diffraction of photosystem II microcrystals	Kern et al.	7
Jul 2012	Lysozyme	Structure	1.9	High-resolution protein structure determination by serial femtosecond crystallography	Boutet et al.	1
Nov 2012	Thermolysin	Data	4.0	Nanoflow electrospinning serial femtosecond crystallography	Sierra et al.	8
Jan 2013	Cathepsin B	Structure	2.1	Natively inhibited *Trypsanosoma brucei* cathepsin B structure determined by using an X-ray laser	Redecke et al.	9
Apr 2013	Photosystem II	Structure	5.7	Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature	Kern et al.	10
May 2013	Lysozyme	Structure	3.2	Anomalous signal from S atoms in protein crystallographic data from an X-ray free-electron laser	Barends et al.	11
Sept 2013	Ribosome	Data	<6	Serial femtosecond X-ray diffraction of 30S ribosomal subunit microcrystals in liquid suspension at ambient temperature using an X-ray free-electron laser	Demirci et al.	12
Dec 2013	Photosynthetic Reaction Center	Structure	3.5	Structure of a photosynthetic reaction center determined by serial femtosecond crystallography	Johansson et al.	13
Dec 2013	Serotonin receptor	Structure	2.8	Serial femtosecond crystallography of G protein-coupled receptors	Liu et al.	14
Jan 2014	Lysozyme + Gd	Structure	2.1	De novo protein crystal structure determination from XFEL data	Barends et al.	15
This study	Cry3A toxin, isolated crystals and whole cells	Structure	2.8, 2.9	2.9 Å-Resolution protein crystal structure obtained from injecting bacterial cells into an X-ray free-electron laser beam	Sawaya et al.	This study

*The available XFEL energy was limited to 2 keV (6.2 Å wavelength) when these experiments were conducted.

Our target for in vivo crystal structure determination is the insecticidal Cry3A toxin from *Bacillus thuringiensis* (Bt). The bacterium naturally produces crystals of toxin during sporulation (16). Presumably, the capacity for in vivo crystallization evolved in Bt as a mechanism to store the toxin in a concentrated, space-efficient manner. Since the 1920s, farmers have used the crystalline insecticidal proteins to control insect pests; its production as a natural pesticide is now a commercial enterprise. Attempts to structurally characterize the toxins date back to more than 40 y ago with the first report of diffraction from isolated crystals that were packed together in powder form to obtain a measurable signal; X-ray sources available at the time were relatively weak (17). More than 20 y later, the structure was determined at 2.5-Å resolution by single crystal diffraction using a synchrotron X-ray source (18). However, to achieve this result, the authors dissolved the naturally occurring microcrystals and recrystallized the toxin using the hanging drop vapor diffusion method. To date, more than a dozen Bt toxin structures have been reported from various strains [Protein Data Bank (PDB) ID codes 1cby, 1ciy, 1i5p, 1ji6, 1w99, 2d42, 2c9k, 2rci, 3eb7, 2ztb, 3ron, 4d8m, 4ato, 4ary, and 4arx], but none using naturally occurring crystals, and all of the crystals had lost their native context.

In pursuit of in vivo diffraction, we took advantage of the Bt subsp. *israelensis* strain 4Q7/pPFT3As to produce the largest in vivo crystals achievable. This strain contains the plasmid pPFT3As, which increases expression of Cry3A by 12.7-fold over wild type by using strong promoters and an mRNA stabilizing sequence (19). The level of Cry3A production is such that the cell essentially distorts to take on the shape of the enclosed crystal. The calculated average crystal volume is 0.7 μm^3 (19), almost accounting for the volume of the cell. To explore the possibilities for in situ data collection of in vivo microcrystals, we injected both the crystals in cells and crystals that we isolated from cells in the XFEL beam and collected SFX diffraction data. Our experiments revealed that the cell wall and other cellular components are not an obstacle to achieving 2.9-Å-resolution diffraction, and analogous studies in other systems might be similarly successful.

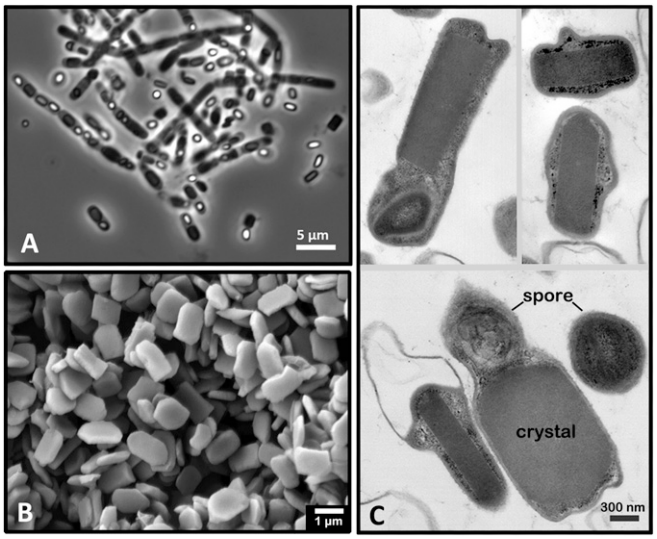

Fig. 1. Samples used for XFEL diffraction studies. (*A*) Phase contrast light micrograph of sporulating Bt cells (rod shaped). The dark rectangular shapes inside (and a few outside) cells correspond to the Cry3A toxin crystals. The bright white oval shapes correspond to spores. The micrograph shows that the cells, suspended in pure water, remain intact with no added buffers. (*B*) Scanning electron micrograph of Cry3A crystals isolated from cells. The image shows that the sample is free of large cell debris and that the crystals have a relatively uniform size. (*C*) Transmission electron micrograph of thin-sectioned Bt cells showing that the crystals (rectangular objects with uniform electron density) are so large that the cells are distended to the shape of the crystals. The rounded objects in the cells (and free-floating; *Lower*) are spores.

Results

Production and Isolation of in Vivo-Grown Cry3A Crystals. Cry3A crystals were produced by acrystalliferous Bt subsp. *israelensis* strain 4Q7 containing plasmid pPFT3As harboring the Cry3A gene from DSM 2803, a wild-type isolate of Bt *subsp. morrisoni* (strain *tenebrionis*) (4Q7/pPFT3As). The Cry3A gene in pPFT3As produces rectangular plate crystals in cells several-fold larger than in wild-type strains (19), with approximate dimensions of $1.5 \times 1.0 \times 0.5$ μm (Fig. 1). Crystals were isolated as described in *Materials and Methods*.

Data Collection. SFX experiments were carried out in March 2013 at the CXI instrument (Coherent X-ray Imaging) at the SLAC Linac Coherent Light Source (LCLS) (20). The photon energy of the X-ray pulses was 8.52 keV (1.45 Å). Each 40-fs pulse contained up to 6×10^{11} photons at the sample position, taking into account a beamline transmission of 60%. The diameter of the beam was ∼1 μm. The in vivo-grown crystals were injected into the XFEL beam using a liquid jet injector and a gas dynamic virtual nozzle (21).The micro jet width was ∼4 μm, and the flow rate was 20–50 μL/min. After emerging from the injector tip, the isolated Cry3A crystals or Bt cells travel in a liquid jet through a vacuum chamber for ∼200 μm before they are intercepted by the X-ray pulse. The crystal concentration was adjusted to compromise between maximizing the hit rate and minimizing the observation of multiple crystals per diffraction image, as described in *Materials and Methods*. Diffraction patterns of these crystals or cells were recorded by a Cornell-SLAC pixel array detector (22). The repetition rate of the XFEL pulses was 120 Hz. The sample to detector distance varied from 110 to 180 mm, and the resolution at the edge of detector varied from 2.3 Å to 3.0 Å, depending on the distance to the sample. A total of 380,688 diffraction images were collected for isolated Cry3A crystals and 736,360 images for the Bt cells (Fig. 2 and Table 2).

Data Processing and Refinement. The SFX diffraction data collected from isolated crystals and from whole cells were each processed using two different programs, CrystFEL (23) and cctbx.xfel (26, 27), yielding four data sets total (Table 2). Cry3A models were refined against the four data sets. In all cases the starting model for refinement was the structure of Cry3A (PDB code 1DLC) obtained from recrystallized material (18), with water molecules removed. We found that although the data statistics differ in some aspects (Table 2 and Figs. S1 and S2), the quality of the models obtained from data processed by each of the two programs is similar, as judged by the similarity of the refined models (Fig. S3); the rmsd between α-carbon positions of the models was only 0.10 Å in the case of data collected from isolated crystals and 0.12 Å in the case of data collected from cells.

Comparison of Structures of Cell-Grown Crystals and Reconstituted Crystals. The crystal structure of Cry3A determined using the in vivo-grown crystals showed no significant structural differences from that previously determined from recrystallized Cry3A (18). The rmsd of 584 α-carbon positions is small, 0.14 Å. In fact, the crystals are isomorphous (28) (recrystallized: a = 117.1, b = 134.2, c = 104.5; in vivo-grown: a = 116.9 ± 1.0, b = 135.8 ± 0.7, c = 105.2 ± 0.5; Table 2). The diffraction limit of the recrystallized Cry3A toxin was higher (2.5 Å) than that of the cell-grown crystals (2.8 Å) at LCLS. Recrystallization outside the boundaries of the cell permitted growth of much larger crystals, which more than compensated for the relatively lower brilliance of the second-generation synchrotron source (Deutsches Elektronen-Synchrotron storage ring, DORIS) that was used to collect the data (18). In addition, the in vivo-grown microcrystals may have suffered from increased disorder owing to an ∼10% impurity of unprocessed proprotein, containing an additional 57 residues at the N terminus (29). The impurity was lacking in the recrystallized material owing to exogenous papain treatment of the starting material before crystallization (18). Furthermore, with only 10% of the toxin remaining uncleaved in vivo (29), it was not surprising to find that no electron density was observed for these 57 residues in the maps calculated from any of our data sets.

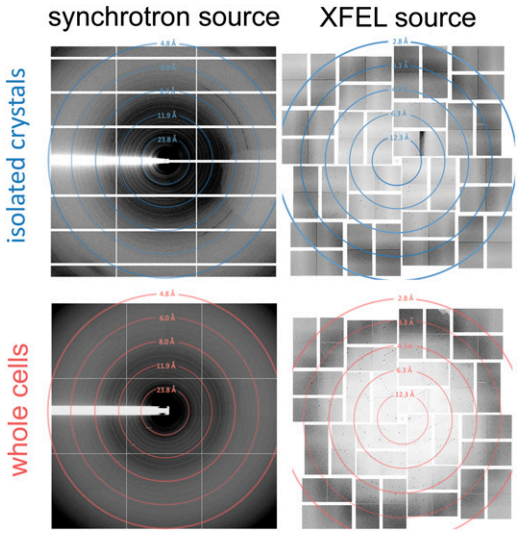

Fig. 2. Diffraction images from isolated Cry3A crystals (*Upper*) and cells (*Lower*). The XFEL experiment (LCLS, CXI station) permitted single crystal diffraction to be observed (*Right*), whereas synchrotron sources produced powder diffraction patterns (*Left*). The powder pattern from Cry3A crystals was collected at the Advanced Photon Source Northeastern Collaborative Access Team (APS NECAT) beamline 24-ID-C on a Dectris Pilatus 6M detector. The powder pattern from Bt cells was collected at APS NECAT beamline 24-ID-E using an ADSC Q-315 detector.

Table 2. Cry3A XFEL data collection and refinement statistics using isolated crystals and whole Bt cells

Parameter / Software	Sample: Crystals isolated from Bt cells		Sample: Whole Bt cells	
	cctbx.xfel	CrystFEL	cctbx.xfel	CrystFEL
Data collection				
Space group	$C222_1$	$C222_1$	$C222_1$	$C222_1$
a (Å)	116.9 ± 1.0	117.1 ± 0.9	117.1 ± 1.1	117.3 ± 1.1
b (Å)	135.8 ± 0.7	135.4 ± 1.0	134.8 ± 0.8	135.3 ± 1.2
c (Å)	105.2 ± 0.5	105.6 ± 0.9	104.9 ± 0.6	105.3 ± 1.1
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Wavelength(Å)*	1.454 ± 0.002	1.456 ± 0.002	1.457 ± 0.002	1.457 ± 0.002
Resolution (Å)	56.7–2.8 (2.88–2.80)	88.6–2.8 (2.90–2.80)	52.4–2.9 (2.99–2.90)	88.64–2.9 (3.00–2.90)
Total patterns	380,650	380,688	736,312	736,360
Indexed patterns	78,642	76,308	30,008	30,952
Indexing rate (%)†	20.7	20.0	4.1	4.2
Total observations	14,279,911	23,731,501	4,383,931	9,174,339
Multiplicity‡	717.8 (1.5)	1128.0 (691.5)	252.5 (1.3)	484.3 (444.5)
Unique reflections	19,894	21,038	17,360	18,944
Completeness (%)	95.6 (55.7)	100.0 (100.0)	92.3 (39.2)	100.0 (100.0)
R_{split}(%)§	12.2 (75.3)	15.9 (41.2)	21.6 (90.6)	24.4 (48.9)
$CC_{1/2}$(%)‡	97.7 (27.3)	91.6 (57.7)	90.6 (13.6)	81.7 (40.1)
I/σ(I)¶	101.2 (1.8)	9.5 (1.6)	59.2 (2.3)	5.2 (0.8)
Wilson B ($Å^2$)‡	32.4	86.4	38.0	84.6
Refinement				
Resolution (Å)	56.8–2.8 (2.95–2.80)	44.3–2.8 (2.94–2.80)	52.4–2.9 (3.08–2.90)	44.3–2.9 (3.06–2.90)
Total reflections	19,894	20,560	18,583	18,934
R_{work} (%)	16.5 (28.2)	17.8 (25.8)	16.8 (24.8)	17.7 (26.5)
R_{free} (%)	19.2 (30.8)	19.7 (26.4)	20.1 (28.2)	19.4 (28.5)
Protein atoms	4,659	4,659	4,659	4,659
Water atoms	26	0	0	0
Protein *B*-factors($Å^2$)‡	38.5	75.7	38.4	83.9
Water *B*-factors($Å^2$)‡	18.7	N/A	N/A	N/A
rmsd bond lengths (Å)	0.010	0.010	0.010	0.010
rmsd bond angles (°)	1.0	1.0	1.0	1.0
Errat score (%)‖	97.2	96.8	96.8	97.1
Verify3D score (%)**	96.1	95.9	96.9	96.8
PDB ID code	4QX0	4QX1	4QX2	4QX3

*The spectral bandwidth of each X-ray pulse for a self-amplified spontaneous emission free electron laser is approx. 0.2%, and the shot-to-shot rms photon wavelength jitter is approx. 0.2%.

†Indexing rate is defined as the number of indexed images per number of patterns collected. It differs from the previous definition given as the number of indexed images per total hits (1). The revision eliminates dependence on subjective choices of "hit" parameters, such as reflection intensity, threshold values, and minimum acceptable spot sizes. By either definition, the indexing rate does not report on diffraction quality. Rather, diffraction quality is reflected in statistics such as R_{split}, I/σ(I), and $CC_{1/2}$.

‡Program-specific differences in B-factors and outer shell statistics are due to different acceptance criteria for observations in the outer shells. Refer to *SI Text*.

§We substituted R_{merge} with R_{split} as is appropriate for SFX experiments in which all reflection measurements are partial (23).

¶Methods of estimating I/σ(I) are reported in *SI Text*.

‖Overall Quality Factor (24).

**Percentage of residues with score >0.2 (25).

Comparison of Data from in Situ Cell and Isolated in Vivo-Grown Crystals. The resolution limits of the best diffraction images from cells were comparable to those from isolated crystals. However, there is a difference in quality of the data sets (2.8-Å vs. 2.9-Å resolution), largely due to the fewer number of patterns indexed from the in situ cell diffraction experiment. There are less than half as many indexed patterns for the data collected from whole cells compared with isolated crystals by either cctbx.xfel (78,642 from isolated crystals vs. 30,008 from whole cells) or CrystFEL methods (76,308 from isolated crystals vs. 30,952 from whole cells). When equal numbers of indexed images were used, the data set obtained from isolated crystals was slightly better quality than obtained from whole cells (Table S1). The small difference could be due to variation in beam transmission, jet diameter, or scattering from cell components.

Evidence for increased background scattering from cell components in the data collected from whole cells is not obvious. If background scattering from cell components were significant, it might reduce the indexing rate (defined as the ratio of indexed images to total number of images collected) from whole cells compared with isolated crystals. However, this rate is also affected by differences in crystal concentration, for which we do not have an accurate measure. The 16% reduction in indexing rate we observed for whole cells compared with isolated crystals could be due in part to a lower concentration of crystals in the whole cell sample. At first glance, it might seem possible to eliminate the influence of crystal concentration from this ratio by including in the denominator only those images with recorded diffraction events (i.e., "hits"). However, the criteria for defining a hit differs from data set to data set and between indexing algorithms. To find objective evidence of scattering from noncrystalline cell components, we performed a comparison of radial profiles (plots of intensity vs. scattering angle) obtained from analysis of equal numbers of hits from isolated crystals vs. whole cells. It revealed no significant difference in background scattering between the two samples (Fig. S4). It suggests that scattering from cell components in the whole cell sample does not strongly limit the data quality.

Discussion

Cell Integrity at the Moment of Diffraction. The structure of Cry3A toxin described here originates from injecting bacterial cells into an XFEL beam. That is, whole cells were loaded in the injector, and 2.9-Å-resolution diffraction patterns were collected from the crystals injected in the XFEL beam. However, a true in situ experiment would require the crystal to reside inside a cell during the diffraction event. Control experiments have yielded insufficient data to conclude whether this criterion was met (*SI Text*). In the absence of a more definitive experiment, we consider the possibility that lysis occurred before or after the cells reach the XFEL beam position, located only ~200 μm from the nozzle (Fig. 3). At sample flow rates ranging from 20 to 50 μL/min, the cells travel only for 3.0–7.5 μs to reach the XFEL beam from the nozzle exit. Even if all cell walls have ruptured at the instant of exiting the nozzle, a few microseconds may not be sufficient time for the crystals to dislodge from the cells. Even if the crystals do dislodge from cells, a few microseconds are probably insufficient time for the toxin molecules of the cellular crystals to recrystallize into another form. Thus, the crystal diffraction patterns we recorded are not significantly altered from what would be expected in a true in vivo experiment.

Prospects for in Vivo Diffraction. There are other systems for which in vivo diffraction may not only be feasible but highly desirable. These include crystals of seed proteins, secretory granules containing the major basic protein from white blood cells called eosinophils and insulin from the islets of Langerhans, enzyme assemblies of urate oxidase and alcohol oxidase produced by peroxisomes, and the protein HEX-1, which composes the proteinaceous core of woronin bodies to prevent cytoplasmic bleeding in filamentous fungi (30). As technological advances increase the attainable intensity of XFEL pulses, atomic resolution structure determination of smaller, less ordered systems, such as carboxysomes, may become feasible.

The results of our studies suggest that the cell components would not prevent obtaining high-resolution diffraction patterns from crystalline material in vivo. We acknowledge that intensified background scattering might present a more serious barrier in other less favorable systems where the crystals make up a smaller portion of the cell volume. Additionally, a different cell delivery system might be required to guarantee the integrity of the cell at the moment of diffraction. However, this technical distinction is likely irrelevant from the point of view of structural biology. If the crystal has not changed its organization during the few microseconds between exiting the nozzle and intercepting the beam, then structures obtained in this manner would reveal the protein crystal structure as it exists inside the living cell.

A thread running through the history of cell biology is the increasing recognition that cellular components are structured. The crystallization of proteins, starting in the 19th century, showed that these large molecules have a defined structure. Sumner's crystallization of urease in 1926 extended this recognition of order to enzymes. Later work revealed the organization of DNA and nucleosomes and the existence of elaborate molecular machines consisting of numerous ordered components. With the advent of electron microscopy in the mid-20th century, it became evident that cells, far from being bags of freely diffusing molecules, are compartmentalized and ordered. That these ordered structures are dynamic, constantly changing, does not contradict the existence of much order at any given instant. As shown by this work, free-electron lasers offer the prospect of interrogating the extent and nature of this order.

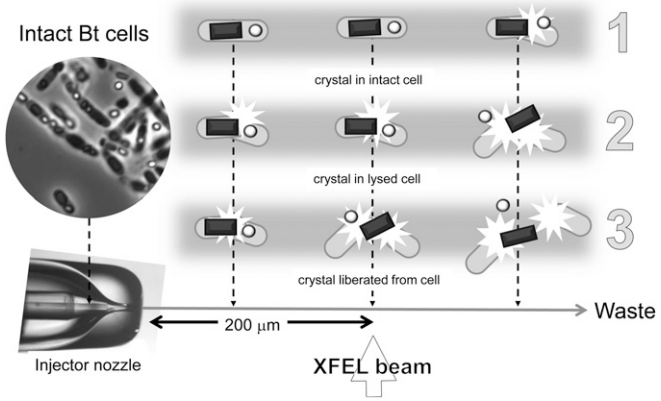

Fig. 3. Three scenarios suggesting how the integrity of the cells might vary at the moment of diffraction. The horizontal arrow depicts the flow of sample from injector to waste collection. The XFEL beam intercepts the sample stream ~200 μm from the nozzle. The left, middle, and right columns depict three different time points along the jet trajectory. Depending on the rate of lysis and the flow rate of the jet, the crystals may arrive at the interaction point either (1) inside intact cells, (2) inside lysed cells, or (3) segregated from lysed cells. The time of travel is estimated to be 3–7.5 μs.

Materials and Methods

Details of Production and Isolation of in Vivo-Grown Cry3A Crystals. Crystals were isolated as follows. Five hundred milliliters of glucose-yeast-salts (GYS) liquid growth medium [0.1% glucose, 0.2% yeast extract, 0.05% K_2HPO_4, 0.2% $(NH_4)_2SO_4$, 0.002% $MgSO_4$, 0.005% $MnSO_4$, and 0.008% $CaCl_2$] supplemented with 25 μg/mL erythromycin was prefiltered through a 0.22-μm membrane to eliminate dust and suspended contaminants and sterilized by autoclave in a 2-L baffled flask. Media was inoculated with spores (from a lyophilized 3-d lysate) of Bt subsp. *israelensis* strain 4Q7 containing plasmid pPFT3As (4Q7/pPFT3As) (19) and incubated for 3 d at 30 °C with shaking at 250 rpm. Cultures were monitored by phase contrast light microscopy, until sporulation and cell lysis were observed, then spores, crystals, cells, and cell debris were pelleted by centrifugation at 6,000 × *g* for 30 min. The culture pellet was resuspended in 50 mL filtered water and sonicated for 3 min on ice [1 s on, 1 s off (6 min elapsed time); 60% intensity] to lyse remaining cells. The lysate was pelleted at 6,000 × *g* for 30 min at 4 °C, washed in 50 mL filtered water to remove soluble material and some of the spores, then repelleted before being resuspended in 15 mL filtered water. The crystals remained intact and did not dissolve in the absence of ions or buffer (Fig. 1*B*). The crystals can be induced to dissolve if exposed to alkaline conditions (pH 10) as exist in the larval gut. Such were the conditions used to solubilize Cry3A for recrystallization (28). Crystals were separated from other cellular components on sucrose step gradients (11 mL each of filtered 67%, 72%, and 79% wt/vol sucrose solutions) formed in 25 × 89-mm transparent, thin-wall tubes (Beckman). Each gradient was overlaid with 5 mL of lysate and centrifuged in a Beckman SW28 rotor at 35,000 × *g* for 1 h at 4 °C. Crystals formed a wide band above the interface of the 72% and 79% sucrose layers and were recovered from each gradient in 8–10 mL of sucrose solution using a BioComp Gradient Fractionator (BioComp Instruments). Recovered gradient bands were pooled and serially dialyzed six times into 100 volumes of filtered water at 4 °C for ≥1 h to remove sucrose. Dialyzed crystals were pelleted at 6,000 × *g* for 15 min at 4 °C, resuspended in 10 mL filtered water, and stored at 4 °C. After settling, excess liquid was removed to leave a milky-white slurry of suspended crystals.

Preparation of *Bt* Cells Containing Cry3A Crystals. Crystal-containing 4Q7/pPFT3As cells were grown from spores inoculated into 500 mL filtered GYS medium supplemented with 25 μg/mL erythromycin and incubated for 1.5–2 d at 30 °C, shaking at 250 rpm. Cells were monitored by phase contrast light microscopy until sporulation and then harvested by centrifugation at 6,000 × *g* for 15 min at 4 °C. All remaining steps were done at 4 °C and all liquids filtered through a 0.22-μm membrane. The cell pellet was washed with 50 mL water, repelleted at 6,000 × *g* for 30 min, and resuspended in 15 mL water. Sucrose step gradients (11 mL each of 67%, 72%, and 79% wt/vol sucrose) in 25 × 89-mm transparent, thin-wall tubes were each loaded with 5 mL of washed cells and centrifuged in an SW28 rotor at 35,000 × *g* for 1 h at 4 °C. Cells forming a broad band in the 72% sucrose step were recovered with a BioComp Gradient Fractionator (BioComp Instruments). Gradient bands were extensively dialyzed into water to remove sucrose (six changes of 100 volumes of filtered water at 4 °C for ≥1 h), then pelleted at 6,000 × *g* for 30 min and resuspended in 2 mL water. No salts, sucrose, or other materials were added. The cells and crystals did not seem to lose integrity because of the absence of ions or buffer (Fig. 1*A*). Samples were passed through a 10-μm stainless steel frit (Upchurch Scientific, part A-107X) to remove any large particles that might clog the sample injector. The frit was seated in an HPLC filter holder (Upchurch

Scientific, part A-356), which was adapted to a 3-mL Luer-Lok tip disposable syringe for convenience of filtering. There was virtually no resistance to pushing the sample through the frit.

Adjustment of Crystal Concentration. To maximize the chances of crystals intercepting the X-ray pulses, we aimed for a concentration of 10^{11} to 10^{12} crystals/mL. To estimate the crystal count, we pelleted the isolated crystals by a 30-s spin in a tabletop centrifuge. We started with 25 μL of crystal pellet diluted to 1 mL with water. Estimating 0.7 μm^3 per crystal, the calculated concentration corresponded to 3.6×10^{10} crystals/mL. Even though this concentration is likely to be an overestimate for lack of accounting for the space between crystals, many of the diffraction patterns showed multiple lattices. We interpreted the appearance of multiple lattices per exposure to signify that the crystal slurry was too concentrated. The sample was then diluted to 2.4×10^{10} crystals/mL for the remaining runs. In retrospect, we realize that most of the multiple lattices could have been the result of crystals clumping together, a physical attachment that cannot be broken by dilution. The entire 25-μL crystal pellet was consumed over the course of the crystal diffraction experiment (58 min).

Adjustment of Cell Concentration. For the in-cell experiments, the cells were pelleted in the same way as the isolated crystals. We used 25 μL of cell pellet/mL. We did not dilute the cells because there were relatively fewer instances of multiple lattices per diffraction image.

Algorithms for Processing Serial Femtosecond Crystal Diffraction Images. The data were processed with cctbx.xfel and CrystFEL, as described in *SI Text*.

Comparison of SFX Data Processed Using Two Independent Algorithms. Please refer to *SI Text*.

Atomic Refinement. Because the crystals isolated from cells and crystals within cells are isomorphous with the published structure determined using conventional means (18), the atomic refinement was started with a rigid body refinement in phenix.refine (31), followed by atomic refinement, manual rebuilding of the models in COOT (32), and TLS refinement. The final cycles of atomic refinement were performed using BUSTER (33), and COOT was used to add solvent molecules. We used coordinates of the Cry3A model (PDB code 1DLC) as a source of external geometric restraints when refining with Buster. Table 2 shows the results of the atomic refinement. The low values obtained for R and R_{free} are likely related to the high quality of the starting model (1DLC), which was determined at a higher resolution (2.5 Å) than the diffraction data obtained and used for refinement in these experiments. The structure validation was performed using the SAVES server (http://nihserver.mbi.ucla.edu/SAVES/), which validates the models using the programs PROCHECK (34), WHAT_CHECK (35), ERRAT (24), and VERIFY 3D (25) to assess the stereochemical quality, nonbonded interactions, and the compatibility of each amino acid in its local environment. In addition, an analysis of of CC* and CC_{work} offers further evidence that we have not overfit our model. Fig. S5 shows a plot of CC* vs. resolution for each of the four refinements. CC* is an estimate of the correlation between the measured data and hypothetical noise-free signal (36). It is derived mathematically from $CC_{1/2}$, which measures the correlation between two randomly chosen halves of the unmerged data set. Plotted with CC* is CC_{work}, the correlation between the measured structure factors in the working set and the corresponding structure factors calculated from the model coordinates. If we had overfit our model, it would be indicated by CC_{work} having a larger value than CC*. It would indicate that the model agrees better with the experimental data than does the true signal. In none of the four refinements do these statistics indicate overfitting.

ACKNOWLEDGMENTS. We thank M. Capel, K. Rajashankar, N. Sukumar, J. Schuermann, I. Kourinov, and F. Murphy [Northeastern Collaborative Access Team Beamline 24-ID at the Advanced Photon Source, which is supported by National Center for Research Resources Grant 5P41RR015301-10 and National Institute of General Medical Sciences Grant 8 P41 GM103403-10 from the National Institutes of Health (NIH)]. Use of the Advanced Photon Source is supported by the US Department of Energy (DOE) under Contract DE-AC02-06CH11357. We also thank Harold Aschmann and the University of California, Los Angeles (UCLA)–DOE X-ray Crystallography Core Facility, which is supported by DOE Grant DE-FC02-02ER63421; and Heather McFarlane and Daniel Anderson at UCLA for help with cell preparation and filtration. Portions of this research were carried out at the Linac Coherent Light Source, a National User Facility operated by Stanford University on behalf of the DOE Office of Basic Energy Sciences. The CXI instrument was funded by the Linac Coherent Light Source Ultrafast Science Instruments project funded by the DOE Office of Basic Energy Sciences. This work was supported by Keck Foundation Grant 2843398, NIH Grant AG-029430, National Science Foundation Grant MCB 0958111, DOE Grant DE-FC02-02ER63421, NIH Grants GM095887 and GM102520 for data-processing methods (to N.K.S.), NIH Grant AI45817 (to B.A.F.), Howard Hughes Medical Institute, and the Max Planck Society.

1. Boutet S, et al. (2012) High-resolution protein structure determination by serial femtosecond crystallography. *Science* 337(6092):362–364.
2. Chapman HN, et al. (2011) Femtosecond X-ray protein nanocrystallography. *Nature* 470(7332):73–77.
3. Koopmann R, et al. (2012) In vivo protein crystallization opens new routes in structural biology. *Nat Methods* 9(3):259–262.
4. Lomb L, et al. (2011) Radiation damage in protein serial femtosecond crystallography using an x-ray free-electron laser. *Phys Rev B* 84(21):214111.
5. Aquila A, et al. (2012) Time-resolved protein nanocrystallography using an X-ray free-electron laser. *Opt Express* 20(3):2706–2716.
6. Johansson LC, et al. (2012) Lipidic phase membrane protein serial femtosecond crystallography. *Nat Methods* 9(3):263–265.
7. Kern J, et al. (2012) Room temperature femtosecond X-ray diffraction of photosystem II microcrystals. *Proc Natl Acad Sci USA* 109(25):9721–9726.
8. Sierra RG, et al. (2012) Nanoflow electrospinning serial femtosecond crystallography. *Acta Crystallogr D Biol Crystallogr* 68(Pt 11):1584–1587.
9. Redecke L, et al. (2013) Natively inhibited *Trypanosoma brucei* cathepsin B structure determined by using an X-ray laser. *Science* 339(6116):227–230.
10. Kern J, et al. (2013) Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature. *Science* 340(6131):491–495.
11. Barends TR, et al. (2013) Anomalous signal from S atoms in protein crystallographic data from an X-ray free-electron laser. *Acta Crystallogr D Biol Crystallogr* 69(Pt 5):838–842.
12. Demirci H, et al. (2013) Serial femtosecond X-ray diffraction of 30S ribosomal subunit microcrystals in liquid suspension at ambient temperature using an X-ray free-electron laser. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 69(Pt 9):1066–1069.
13. Johansson LC, et al. (2013) Structure of a photosynthetic reaction centre determined by serial femtosecond crystallography. *Nat Commun* 4:2911.
14. Liu W, et al. (2013) Serial femtosecond crystallography of G protein-coupled receptors. *Science* 342(6165):1521–1524.
15. Barends TR, et al. (2014) De novo protein crystal structure determination from X-ray free-electron laser data. *Nature* 505(7482):244–247.
16. Höfte H, Whiteley HR (1989) Insecticidal crystal proteins of *Bacillus thuringiensis*. *Microbiol Rev* 53(2):242–255.
17. Holmes KC, Monro RE (1965) Studies on the structure of parasporal inclusions from *Bacillus thuringiensis*. *J Mol Biol* 14(2):572–581.
18. Li JD, Carroll J, Ellar DJ (1991) Crystal structure of insecticidal delta-endotoxin from *Bacillus thuringiensis* at 2.5 A resolution. *Nature* 353(6347):815–821.
19. Park HW, Ge B, Bauer LS, Federici BA (1998) Optimization of Cry3A yields in *Bacillus thuringiensis* by use of sporulation-dependent promoters in combination with the STAB-SD mRNA sequence. *Appl Environ Microbiol* 64(10):3932–3938.
20. Boutet S, Williams GJ (2010) The Coherent X-Ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS). *New J Phys* 12:035024.
21. Weierstall U, Spence JC, Doak RB (2012) Injector for scattering measurements on fully solvated biospecies. *Rev Sci Instrum* 83(3):035108.
22. Philipp HT, Hromalik M, Tate M, Koerner L, Gruner SM (2011) Pixel array detector for X-ray free electron laser experiments. *Nucl Instrum Methods A* 649(1):67–69.
23. White TA, et al. (2012) CrystFEL: A software suite for snapshot serial crystallography. *J Appl Cryst* 45(2):335–341.
24. Colovos C, Yeates TO (1993) Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci* 2(9):1511–1519.
25. Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356(6364):83–85.
26. Sauter NK, Hattne J, Grosse-Kunstleve RW, Echols N (2013) New Python-based methods for data processing. *Acta Crystallogr D Biol Crystallogr* 69(Pt 7):1274–1282.
27. Hattne J, et al. (2014) Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers. *Nat Methods* 11(5):545–548.
28. Li J, Henderson R, Carroll J, Ellar D (1988) X-ray analysis of the crystalline parasporal inclusion in *Bacillus thuringiensis* var. tenebrionis. *J Mol Biol* 199(3):543–544.
29. Carroll J, Li J, Ellar DJ (1989) Proteolytic processing of a coleopteran-specific delta-endotoxin produced by *Bacillus thuringiensis* var. tenebrionis. *Biochem J* 261(1):99–105.
30. Doye JPK, Poon WCK (2006) Protein crystallization in vivo. *Curr Opin Colloid Interface Sci* 11(1):40–46.
31. Afonine PV, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68(Pt 4):352–367.
32. Emsley P, Cowtan K (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2126–2132.
33. Blanc E, et al. (2004) Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1):2210–2221.
34. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Cryst* 26(2):283–291.
35. Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. *Nature* 381(6580):272.
36. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. *Science* 336(6084):1030–1033.

Supporting Information

Sawaya et al. 10.1073/pnas.1413456111

SI Text

Cell Integrity at the Moment of Diffraction. Evidence suggests that the cells might have lysed at some point after exiting the nozzle. We have observed that lysis is not triggered by the elevated pressure within the injector itself; cells that were pressurized in the injector but not passed through the nozzle were retrieved and observed to have remained intact. However, in a mockup experiment conducted at atmospheric pressure, without X-rays, we observed by inspection under a microscope that the recovered cells were lysed after passing through the nozzle. One interpretation of these results is that lysis was triggered by a sudden change in pressure as the cells exit the nozzle. Another interpretation is that the surrounding sheath of helium gas mitigated the pressure change, but lysis was triggered by the collision between the cells and the tube used for collection.

Algorithms for Processing Serial Femtosecond Crystal Diffraction Images. The demands of processing serial femtosecond crystal diffraction images have exceeded the capabilities of established, conventional crystallography data processing programs. These demands originate principally from a combination of two unconventional obstacles in the experimental setup. First, there is insufficient time for the crystal to rotate within the length of each exposure (40 fs). Second, only a single exposure can be collected per crystal.

The first obstacle means that the reflection intensities measured from individual images are only partial intensities. That is, each measurement contributes an unknown fraction of the total intensity of the reflection. To integrate the full intensity of each individual reflection, many tens of thousands of diffraction patterns are collected from as many crystals in different orientations. Then, an average is taken over many measurements of the same reflection (1). If there are an insufficient number of recorded measurements, then the data set becomes too inaccurate for structure determination. Accuracy can be improved by scaling images with respect to an external reference data set (such as one collected by conventional means); however, such a reference is not always available.

The second obstacle means there is no defined relationship between the orientations of crystals in successive diffraction patterns. Because of the large number of hits required to assemble a complete data set, human intervention in indexing individual patterns is not practical; indexing must be performed entirely automatically, requiring indexing algorithms to be exceptionally robust, fast, and accurate.

In addition to these fundamental issues, there are obstacles to data processing associated with the unique design of the experiment. For example, before indexing can be accomplished, the position and orientation of the 64 independent Cornell-SLAC pixel array detector (CSPAD) tiles must be determined with sufficient accuracy. Additionally, the processing programs must be able to filter out the enormous number of blank images recorded and stored along with the hits. Data are stored in a format (hdf5) previously unfamiliar to crystallographers, chosen for its ability to manage extremely large and complex data collections.

Comparison of Serial Femtosecond Crystallography Data Processed Using Two Independent Algorithms. Two new processing packages, CrystFEL and cctbx.xfel, have been developed to overcome the obstacles involved in serial femtosecond crystallography (SFX); they use different approaches to filtering and merging data, as discussed below. We used both packages to investigate how the differences between algorithms might be manifested in data and refinement statistics. Although the data statistics differ, such as completeness, $CC_{1/2}$, and redundancy (Table 2), we found that the models obtained from data processed with CrystFEL and cctbx.xfel are good quality and similar to each other as reported above. Both programs are continuing development.

The most notable difference in the results obtained using the two data processing algorithms concerns the appearance of ordered water molecules in electron density maps. Although both cctbx.xfel and CrystFEL permit confident identification of water molecules on the protein surface, there is a marked difference in the appearance of the density of the water molecules in $2mF_o$-DF_c maps calculated after the putative waters are included in the model. For example, positive F_o-F_c density with spherical shape was found located within 3.0-Å distance from three potential hydrogen bonding partners (backbone atoms of Leu299, Arg301, and His495) in maps generated from both cctbx.xfel- and CrystFEL-processed data (Fig. S3 *A* and *B*, respectively) collected from isolated Cry3A crystals. The fact that this level of detail can be observed in a 2.8-Å-resolution map is evidence of the quality of data processing by both algorithms. However, after including the water in an additional round of refinement, the $2mF_o$-DF_c simulated annealing composite omit density surrounding the water appeared only in cctbx.xfel-generated maps. No $2mF_o$-DF_c density could be seen for the water molecule in the CrystFEL-generated simulated annealing composite omit map at the 1.2 σ level. The map was calculated with the program Phenix using a starting temperature of 10,000 K (2). Density covering this water could be seen only when the contour level was dropped to 0.7 σ, and then it appeared as a protuberance from the main chain density rather than as a separate sphere.

Our confidence that these positive residual peaks correspond to actual water molecules and not just random noise is supported by the correlation of these peak positions with water molecules modeled in the structure of the recrystallized Cry3A protein [Protein Data Bank (PDB) ID 1DLC]. The 1DLC model was refined at 2.5-Å resolution and contains 106 ordered water molecules. Although fewer water molecules are expected to be visible at the 2.8-Å-resolution limit of diffraction from our isolated Cry3A crystals, some correspondence in water positions is expected given the isomorphism in unit cell parameters between the two sources of crystals. Using the $2mF_o$-DF_c simulated annealing composite omit map calculated in the absence of water, we observed 49 peaks above the 4.0 σ threshold using the cctbx.xfel refined coordinates and data. Of these peaks, 19 were located within 1.0 Å of an ordered water molecule in the 1DLC model. That is, 38% of the residual positive peaks above the 4.0 σ threshold corresponded to ordered water molecules. Moreover, for the CrystFEL refinement, 40 of the 74 residual positive peaks above the 4.0 σ threshold corresponded to ordered water molecules. This is 54% of the residual peaks. The correlation of the positive peak positions with modeled water molecules in the 1DLC model is unlikely to be due to phase bias because water molecules corresponding to those in 1DLC were not included in the refinement up to this point, and the model had undergone simulated annealing dynamics before calculation of this difference map. These observations support our conclusion that both cctbx.xfel and CrystFEL algorithms have produced integrated diffraction intensities with a sufficient amount of accuracy to locate ordered water molecules.

With confidence in our ability to locate ordered water molecules, 26 waters were included in the model refined against

cctbx.xfel processed data. Each of these putative water molecules was inspected for spherical density in subsequent $2mF_o$-DF_c maps to validate their assignment. These were ultimately included in the cctbx.xfel-derived model deposited in the PDB. Similarly, many waters could be identified using the CrystFEL-processed data, and many of these overlapped the water sites identified in cctbx.xfel-derived maps. However, the lack of spherical $2mF_o$-DF_c density near these putative water molecules in maps generated from subsequent rounds of refinement prompted us to remove these waters from the model. Therefore, these putative waters were not included in the CrystFEL-derived model deposited in the PDB, even though there was clear evidence for their presence in simulated annealing omit F_o-F_c maps discussed above.

These results do not imply that there is a difference in accuracy between the structure factor amplitudes derived from CrystFEL- and cctbx.xfel-processed data. The difference in level of detail in the $2mF_o$-DF_c density can be attributed primarily to the difference in Wilson B-factors characterizing the data sets produced by the two algorithms. Indeed, for data collected from isolated crystals, the Wilson B for CrystFEL-processed data (86.4 $Å^2$) is remarkably higher than for cctbx.xfel data (32.4 $Å^2$) (Fig. S1). The trend is the same for data collected from cells, where CrystFEL and cctbx.xfel yield 84.6 $Å^2$ and 38.0 $Å^2$, respectively. The Wilson B-factor characterizes the intensity distribution of a data set with respect to resolution. The higher the Wilson B-factor, the steeper is the falloff in average intensity with increasing resolution. High values of the Wilson B-factor are associated with loss of detail in electron density maps because structure factor amplitudes in the high-resolution shell are relatively smaller (compared with data sets with low Wilson B-factors) and so contribute fewer features to the electron density maps. In cases such as observed here with CrystFEL processed data, detail can be restored to the maps by a process known as "sharpening." Sharpening is simply the application of a resolution-dependent scale factor, which scales up the amplitude of high-resolution bins of a data set with respect to low-resolution bins. In the case of the water molecule discussed above, sharpening by application of a −8.5 $Å^2$ B-factor revealed the presence of a separate sphere of $2mF_o$-DF_c simulated annealing composite omit density surrounding the water molecule at the 1.2 σ level (Fig. S3).

The systematically higher Wilson B-factors observed for CrystFEL- compared with cctbx.xfel-processed data sets are likely due to the different ways that resolution limits are introduced by the two algorithms. The program cctbx.xfel adjusts the resolution limit for each diffraction pattern to accommodate the variation in diffraction quality associated with shot-to-shot variation in size and diffraction limit of the crystals, as well as fluctuations in the intensity of the X-ray free-electron lasers (XFEL) beam. In contrast, CrystFEL uses a single resolution limit for all diffraction patterns. The consequence of the CrystFEL strategy is that the higher-resolution shells of the integrated data have higher completeness and multiplicity because CrystFEL makes no effort to filter out measurements in the high-resolution shells of weaker diffraction patterns originating from relatively small crystals or low XFEL intensity. As a further consequence of accepting all measurements regardless of shot-to-shot variations in diffraction strength, CrystFEL data have lower $I/\sigma(I)$ in the high-resolution shell compared with cctbx.xfel data. Hence, acceptance of a greater proportion of weak measurements in the outer resolution shells of the CrystFEL-processed data leads to a higher Wilson B-factor than observed in cctbx.xfel data.

In summary, the difference in appearance of the electron density associated with ordered water molecules in cctbx.xfel- vs. CrystFEL-generated maps is likely due to the large (~50 $Å^2$) difference in Wilson B-factors between the two data sets. The magnitude of the difference raises the question, which value of the Wilson B more accurately reflects the degree of order in the crystals? The complexity of this question is analogous to that involved in deriving a meaningful Wilson B-factor for a highly anisotropic data set. In the case of refinement with anisotropic data, a single Wilson B-factor must reconcile the different diffraction strengths associated with each of the three principal directions of the reciprocal lattice. In the case discussed here, we are trying to reconcile the Wilson B-factors from many thousands of crystals with different diffraction strengths. Although we may not be able to answer this question here, we note that we have obtained similar models from data processed by either algorithm. The lower Wilson B-factor obtained by cctbx.xfel makes it more convenient to locate and refine the water molecules. A similar water model can be obtained using CrystFEL processed data. However, accurate refinement of the water molecules would benefit from some degree of map sharpening.

Another notable distinction between the two algorithms concerns different levels of reliance on an external reference data set as an aid in merging intensities recorded from separate images. As implemented here, the program cctbx.xfel used a set of scale factors, one per image, to improve the accuracy of the merged intensities. These scale factors were formulated to maximize agreement with a reference set of intensities calculated from the deposited coordinates of the recrystallized protein (PDB ID code 1DLC). In contrast, no such external scaling was used by the CrystFEL algorithm; as implemented here, CrystFEL relied solely on averaging as a means of merging intensities. This distinction between algorithms raises two points. First, it illustrates a level of robustness in the CrystFEL algorithm in so much as it produced an accurate set of intensities by methods independent of external data sources. Such sources may not always be available for XFEL projects. Second, it raises the question whether the cctbx.xfel algorithm might introduce an elevated risk of model bias relative to CrystFEL. Unlike the common type of model bias that is caused by introducing incorrect features into an atomic model which then propagate in subsequent maps through model-derived phases, we are concerned here with a bias propagated through structure factor amplitudes. To address this question we analyzed the relationship between the set of F_{obs} produced by cctbx.xfel with the set of F_{calc} from 1DLC. Our analysis revealed no undue bias. In fact, the degree of correlation between these structure factors was no greater than that between the set of F_{obs} produced by CrystFEL and the set of F_{calc} from 1DLC (Fig. S2). Furthermore, the sets of F_{obs} produced by the two algorithms, cctbx.xfel and CrystFEL, agree with each other more closely than either agrees with the set of F_{calc} from 1DLC.

The difference in the statistics of the high-resolution shell data and Wilson B-factors are due to the different approaches used to determine the high-resolution cutoff. CrystFEL estimates the resolution limit for the entire data set whereas cctbx.xfel determines a resolution limit for each diffraction pattern before inclusion into the data set. For example, completeness = 100% and $I/\sigma(I) = 0.8$ are not conflicting attributes of the high-resolution shell of CrystFEL-processed data, considering that CrystFEL draws a single resolution boundary for all images. By this algorithm, intensities will be integrated for every pixel predicted to correspond to a reflection regardless of diffraction intensity or lack thereof.

Last, we note in comparing cctbx.xfel and CrystFEL-processed data sets that significant differences in quality reported in the high-resolution shell (Table 2) may call into question our choice to use the same resolution limit for both algorithms. The choice of resolution limit is a well-known point of contention among crystallographers, owing to the inadequacies of any one statistical measure and the lack of a single, widely accepted standard (3–5). Indeed, even among the authors of this article there are conflicting views on the value of a high-resolution shell with R_{split} >50%, redundancy <2, or completeness ≤55%, as is reported for cctbx.xfel-processed data (Table 2). Our choices of resolution limits were made by considering a combination of input from those authors most familiar with each algorithm and a desire to make

the comparison as straightforward as possible. We also consider the recent evidence from Karplus and Diederichs (3), which suggest that the resolution limits reported in Table 2 are not overly generous. In their experiments they found that high-resolution shells can contain valuable signal even though conventional statistics are poor, for example $I/\sigma(I) = 0.3$ and R_{merge} >400% (in SFX R_{split} is the analog of R_{merge}). More concerning is the expectation that a resolution shell with average multiplicity <2 could contain valuable information. A multiplicity <2 is routinely accepted for data sets collected by conventional means, but in the field of SFX in which each image is a "still" and all intensity measurements are necessarily partial, such a low value is a reason for concern. Some of this concern might be relieved by noting that the $CC_{1/2}$ statistics in the high-resolution shell of cctbx.xfel-processed data (0.27 and 0.14 for isolated crystals and whole cells, respectively; Table 2) are within the limits found significant in controlled experiments (3–5). Additionally, we note that others have found that "adding weak higher-resolution data beyond the commonly used limits may make some improvement and does no harm" (5).

Data Processing with cctbx.xfel. X-ray diffraction patterns were recorded on the CSPAD installed at the CXI instrument of the Linac Coherent Light Source (LCLS) (6) and processed with the package *cctbx.xfel* (7). The version used is dated March 28, 2013 and is accessible at http://cci.lbl.gov/xfel. After subtraction of a dark-run average image, Bragg spots were counted with the *Spotfinder* component of the package (8), with settings being adjusted by trial and error specifically for these data (e.g., the minimum spot area was set at 2 square pixels, and the criteria for accepting spots was relaxed to allow spot picking to an outer resolution limit of approximately 3.0 Å).

Images containing ≥40 candidate Bragg spots were indexed with the Rossmann DPS algorithm (9, 10) as implemented in *LABELIT* (11). Later we found that distinction of "hits" from "non-hits" based on detection of a threshold number of low-resolution spots did not aid in identifying indexable images with cctbx.xfel (12). Therefore, cctbx.xfel regards images either as indexed or nonindexed; the concept of a hit is not used here. The success rate for indexing was increased by requiring that the resulting unit cell be approximately similar to that of the presumed isomorphous structure from PDB entry 1DLC (13). From the ensemble of possible unit cell basis vectors identified by DPS, groups of three vectors were evaluated to find the highest agreement with the target cell lengths and angles. Similar approaches have been used previously by others to identify the diffracted lattice within noisy data (14). One difficulty with the *Bacillus thuringiensis* (Bt) data sets is the presence of more than one lattice on a considerable fraction of images. We found that 11% of images contained a second indexable lattice, identified by running the DPS algorithm on the candidate Bragg spots remaining after the first-indexed lattice is removed, as described previously (15). Only the dominant lattice on each image was retained for data analysis (Table 2). Bragg spots were sufficiently separated so that spot overlaps among multiple lattices were rare.

Indexing success was critically dependent on modeling the position and orientation of the 64 application-specific integrated circuit (ASIC) tiles serving as 185 × 195-pixel readouts for the CSPAD (16). This geometric calibration was performed using Bragg diffraction patterns from 12,818 thermolysin crystals, previously collected and analyzed with the same protocol as that used for Bt. After indexing, observed *Spotfinder* spot coordinates were compared with coordinates predicted by the indexing model. The resulting observed vs. predicted residual was minimized by jointly adjusting both the ASIC tile positions and the indexing model (crystal to detector distance, unit cell, and crystal orientation). This resulted in an rms residual of 100 μm (1 pixel = 110 μm). Modeled tile positions were judged to be accurate to within 10 μm according to the modeled spacing of ASIC pairs bump-bonded to the same silicon sensor chip, which in actuality should be perfectly aligned. ASIC pairs had an rms rotation of 0.22° with respect to the overall detector axes.

Bragg spots were integrated by the extension of existing (17) synchrotron methods. Pixel masks covering the size and shape of each spot were constructed by combining the shapes formed by the nearest 10 *Spotfinder* spots. Intensity signal (I) was integrated after subtraction of a background plane derived from a surrounding region twice the area of the spot (and allowing for a buffer zone two pixels wide) and corrected for polarization (18). Individual error estimates for each spot [$\sigma(I)$] were estimated using Poisson counting statistics, assuming that the detector's high-sensitivity gain is 7.5 ADU (analog to digital units) per photon. Error estimates from each diffraction pattern were then inflated by assuming that negative values of $I/\sigma(I)$ are actually decoy measurements (noise only) with a Gaussian distribution centered at zero and with an SD of 1, thus providing a lower bound on modeling errors. When later merging multiple measurements of the same Miller index, the error was modeled simply by propagating the $\sigma(I)$ values in quadrature. Because the systematic error contributions for XFEL data are not fully understood, no other error normalization was attempted.

Bragg spot intensities on separate images were scaled to an isomorphous reference data set calculated from PDB structure 1DLC. Images having very low correlation with the reference (<10%) were rejected as outliers, as were indexing solutions that failed to conform to orthorhombic symmetry. On the remaining images (listed in Table 2 as "indexed images"), separate resolution cutoffs were computed by determining at what resolution the mean $I/\sigma(I)$ falls below a cutoff value (0.2). Because this cutoff value was deliberately chosen to be very low, it is presumed that there is no significant signal beyond the cutoff. Allowing separate resolution cutoffs for each image leads to a final merged data set with high multiplicity of observation at low resolution and lower multiplicity at high resolution (Table 2), yet there is confidence that the highest-resolution shell contains significant signal. The quality of the merged reflections was assessed by calculating the correlation coefficient of semidatasets merged from odd- and even-numbered images ($CC_{1/2}$) as previously described (3).

Data Processing with CrystFEL and Cheetah. Single-shot diffraction patterns from cells and isolated crystals were preanalyzed using Cheetah (https://github.com/antonbarty/cheetah) (19). Preanalysis included identification of crystal hits. A diffraction pattern was identified as a crystal hit if it contained more than 10 regions of 550 or more ADUs and a signal to noise ratio equal to or greater than 6 (see Cheetah documentation section "Hit finding algorithms – Algorithm 6"). There were a total of 140,911 hits recorded from the isolated crystal sample and 68,218 hits recorded for the whole cell sample.

Diffraction patterns identified as crystal hits by Cheetah were subjected to further analysis using CrystFEL (20), version 0.4.3. Peak finding on single diffraction patterns was performed using a built-in method based on gradient detection. After confirming that the in vivo-grown Cry3A crystals are isomorphous with the macroscopic crystals of purified Bt protein (13, 21), indexing was performed using an interface to Mosflm (22) and DirAx (23) including the cell reduction step with the known unit cell parameters (13). If the diffraction pattern was successfully indexed by one of the two programs, then spot intensities were integrated from a ring of five pixels centered on the Bragg spot. The integrated Bragg spot intensity was corrected for local background, which was estimated from an annulus of six to nine pixels from the Bragg spot center. Integration of Bragg spot intensities was performed from all predicted Bragg peak positions located on each single-shot diffraction pattern. Merged Bragg spot intensities (I) were calculated from all individual measurements using Monte

Carlo averaging (1) without using any prior knowledge. In other words, individual images were not scaled to the set of intensities calculated from the Cry3A model available in the PDB (ID code 1DLC), as was done in the cctbx.xfel processing. Merged sigma values [σ(I)] were estimated using Eq. **3** from ref. 20. Numbers of indexed diffraction patterns, merged signal to noise ratios [I/σ(I)], and other statistical measures for all data sets processed using CrystFEL are presented in Table 2.

1. Kirian RA, et al. (2010) Femtosecond protein nanocrystallography-data analysis methods. *Opt Express* 18(6):5713–5723.
2. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2): 213–221.
3. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. *Science* 336(6084):1030–1033.
4. Diederichs K, Karplus PA (2013) Better models by discarding data? *Acta Crystallogr D Biol Crystallogr* 69(Pt 7):1215–1222.
5. Evans PR, Murshudov GN (2013) How good are my data and what is the resolution? *Acta Crystallogr D Biol Crystallogr* 69(Pt 7):1204–1214.
6. Hart P, et al. (2012) The CSPAD megapixel x-ray camera at LCLS. *Proc SPIE* 8504: 85040C.
7. Sauter NK, Hattne J, Grosse-Kunstleve RW, Echols N (2013) New Python-based methods for data processing. *Acta Crystallogr D Biol Crystallogr* 69(Pt 7):1274–1282.
8. Zhang Z, Sauter NK, van den Bedem H, Snell G, Deacon AM (2006) Automated diffraction image analysis and spot searching for high-throughput crystal screening. *J Appl Cryst* 39(1):112–119.
9. Steller I, Bolotovsky R, Rossmann MG (1997) An algorithm for automatic indexing of oscillation images using Fourier analysis. *J Appl Cryst* 30(6):1036–1040.
10. Rossmann MG, van Beek CG (1999) Data processing. *Acta Crystallogr D Biol Crystallogr* 55(Pt 10):1631–1640.
11. Sauter NK, Grosse-Kunstleve RW, Adams PD (2004) Robust indexing for automatic data collection. *J Appl Cryst* 37(Pt 3):399–409.
12. Kern J, et al. (2014) Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nat Commun* 5:4371.
13. Li JD, Carroll J, Ellar DJ (1991) Crystal structure of insecticidal delta-endotoxin from *Bacillus thuringiensis* at 2.5 A resolution. *Nature* 353(6347):815–821.
14. Paithankar KS, et al. (2011) Simultaneous X-ray diffraction from multiple single crystals of macromolecules. *Acta Crystallogr D Biol Crystallogr* 67(Pt 7):608–618.
15. Sauter NK, Poon BK (2010) Autoindexing with outlier rejection and identification of superimposed lattices. *J Appl Cryst* 43(Pt 3):611–616.
16. Philipp HT, Hromalik M, Tate M, Koerner L, Gruner SM (2011) Pixel array detector for X-ray free electron laser experiments. *Nucl Instrum Methods A* 649(1):67–69.
17. Leslie AGW (1999) Integration of macromolecular diffraction data. *Acta Crystallogr D Biol Crystallogr* 55(Pt 10):1696–1702.
18. Kahn R, et al. (1982) Macromolecular crystallography with synchrotron radiation: Photographic data collection and polarization correction. *J Appl Cryst* 15(3):330–337.
19. Barty A, et al. (2014) Cheetah: software for high-throughput reduction and analysis of serial femtosecond X-ray diffraction data. *J Appl Cryst* 47(Pt 3):1118–1131.
20. White TA, et al. (2012) CrystFEL: A software suite for snapshot serial crystallography. *J Appl Cryst* 45(2):335–341.
21. Li J, Henderson R, Carroll J, Ellar D (1988) X-ray analysis of the crystalline parasporal inclusion in *Bacillus thuringiensis* var. tenebrionis. *J Mol Biol* 199(3):543–544.
22. Leslie AGW, Powell HR (2007) *Evolving Methods for Macromolecular Crystallography*, eds Read RJ, Sussman J (Springer, Dordrecht), pp 41–51.
23. Duisenberg AJM (1992) Indexing in single-crystal diffractometry with an obstinate list of reflections. *J Appl Cryst* 25(2):92–96.

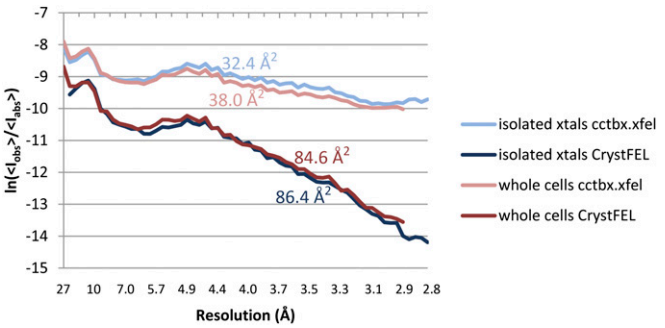

Fig. S1. Wilson plots for data sets collected from isolated crystals (blue hues) and whole cells (red hues), each independently processed by cctbx.xfel (light shade) and CrystFEL (dark shade). The Wilson B-factor characterizes how reflection intensity falls off with resolution. Its value, labeled adjacent to the individual trace, is proportional to the slope of the line plotted here. The large differences in Wilson B-factors are associated with systematic differences in data processing algorithms rather than the type of the sample (isolated or whole cell). The average reflection intensity falls off more rapidly with resolution in CrystFEL-processed data than in cctbx.xfel-processed data. The reason for systematic bias is related to the use of image-dependent resolution limits in cctbx.xfel but not in CrystFEL.

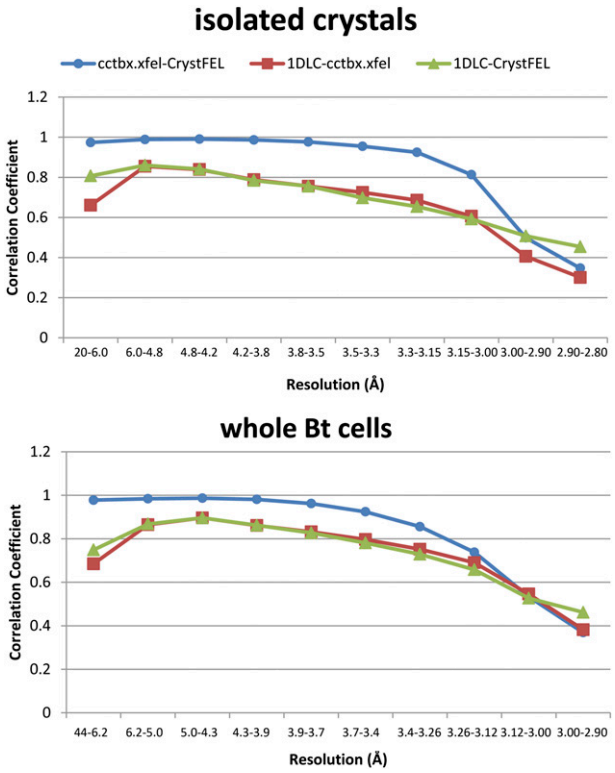

Fig. S2. Plot of correlation coefficients vs. resolution. Serial femtosecond crystal diffraction data were collected from isolated Cry3A crystals and processed by cctbx.xfel and CrystFEL, which use different algorithms for merging data; the implementation of cctbx.xfel used an external reference (calculated structure factors from PDB ID 1DLC) to aid in filtering and scaling together measurements recorded from separate images, whereas CrystFEL did not. The use of 1DLC as an external reference prompted us to test whether bias exists between the cctbx.xfel-processed data set and the set of structure factors used for scaling. The plot reveals that no model bias exists between the cctbx.xfel-processed data set (F_{obs}) and the set of F_{calc} generated from 1DLC. The correlation between the set of F_{obs} obtained from cctbx.xfel and the set of F_{calc} from 1DLC (red trace) indicate that cctbx.xfel-processed data are no more biased toward 1DLC F_{calc} than is CrystFEL data (green trace). The F_{obs} produced by the two algorithms agree with each other (blue trace) more closely than either agrees with the 1DLC F_{calc} (red and green traces). The trends shown here for data collected from isolated crystals are the same trends we observed for data sets collected from whole cells.

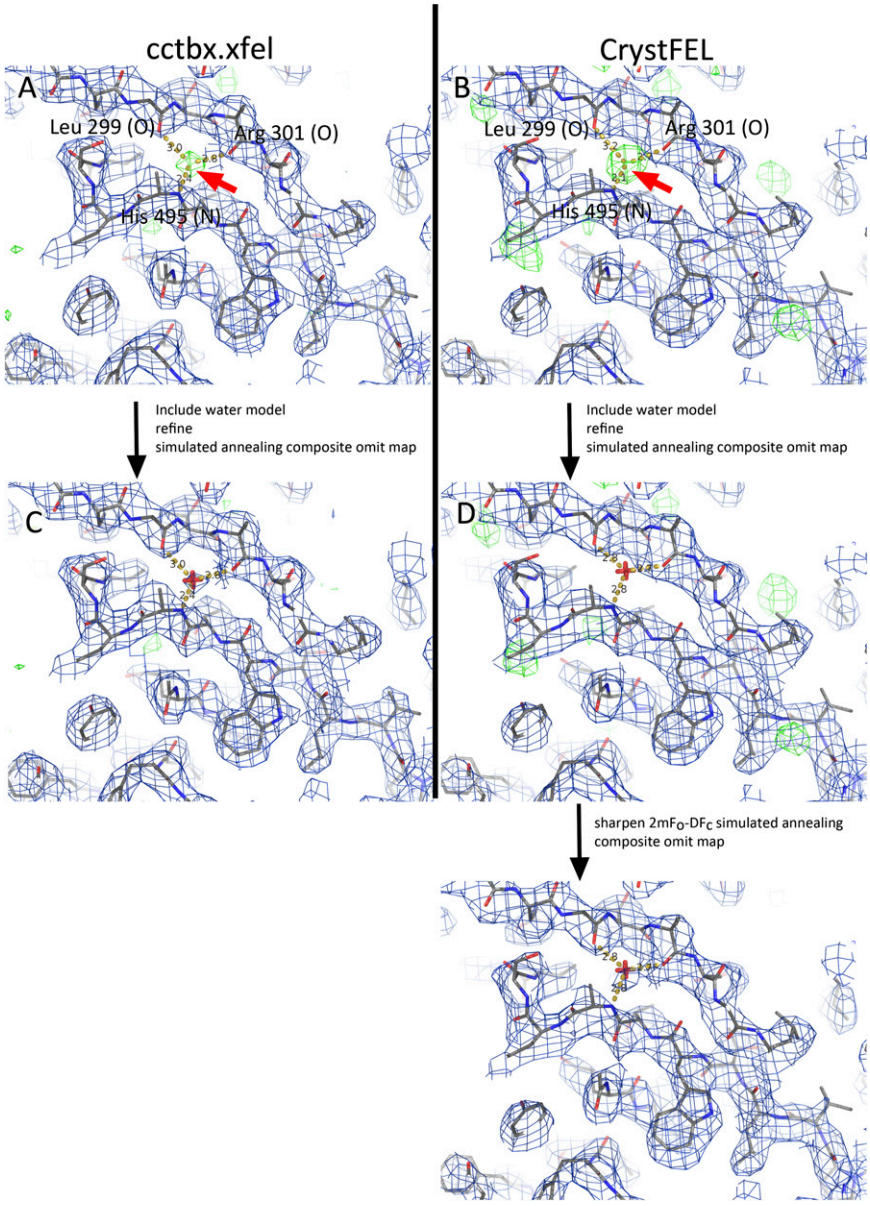

Fig. S3. Simulated annealing composite omit maps indicate the quality of SFX data collected from isolated Cry3A crystals and processed by (*A*) cctbx.xfel and (*B*) CrystFEL. The blue mesh represents $2mF_o$-DF_c density contoured at 1.2 σ. The green mesh represents positive F_o-F_c density contoured at 3.2 σ. The sticks represent the atomic models that were refined against the individual data sets. Overall, the $2mF_o$-DF_c omit maps calculated from diffraction intensities obtained by the two processing algorithms share a similar appearance and fit the models well. Evidence for the higher Wilson B-factor in the CrystFEL data is subtle but noticeable in the smoother appearance near some side chains and missing carbonyl bumps. A red arrow pointed at a peak of positive difference density (green) marks the location of a putative water molecule, which was not included in the model at the time of calculation of the simulated annealing composite omit map. Interpretation of this positive density as a water molecule is supported by the peak's spherical shape and its placement with respect to three hydrogen bonding partners (backbone atoms of Leu299, Arg301, and His495, labeled). The fact that this level of detail can be observed in a 2.8-Å-resolution map is evidence of the quality of data processing by both algorithms. After the putative waters were included in the models, simulated annealing composite omit maps were again calculated using cctbx.xfel processed data (*C*) and CrystFEL processed data (*D*). Again, blue mesh represents $2mF_o$-DF_c density contoured at 1.2 σ, and green mesh represents positive F_o-F_c density contoured at 3.2 σ. Spherical $2mF_o$-DF_c density covers the water molecule in *C*, validating the water assignment for the cctbx.xfel processed data. However, the $2mF_o$-DF_c density near the putative water molecule in *D* was weak and not spherical, probably owing to the higher Wilson B-factor of the CrystFEL processed data. Therefore, these putative waters were not included in the CrystFEL-derived model deposited in the PDB, even though there was clear evidence for their presence in simulated annealing F_o-F_c composite omit maps. Later, we found that if sharpened by application of a -8.5 $Å^2$ B-factor, a separate sphere of $2mF_o$-DF_c simulated annealing composite omit density surrounding the water molecule became apparent at the 1.2 σ level.

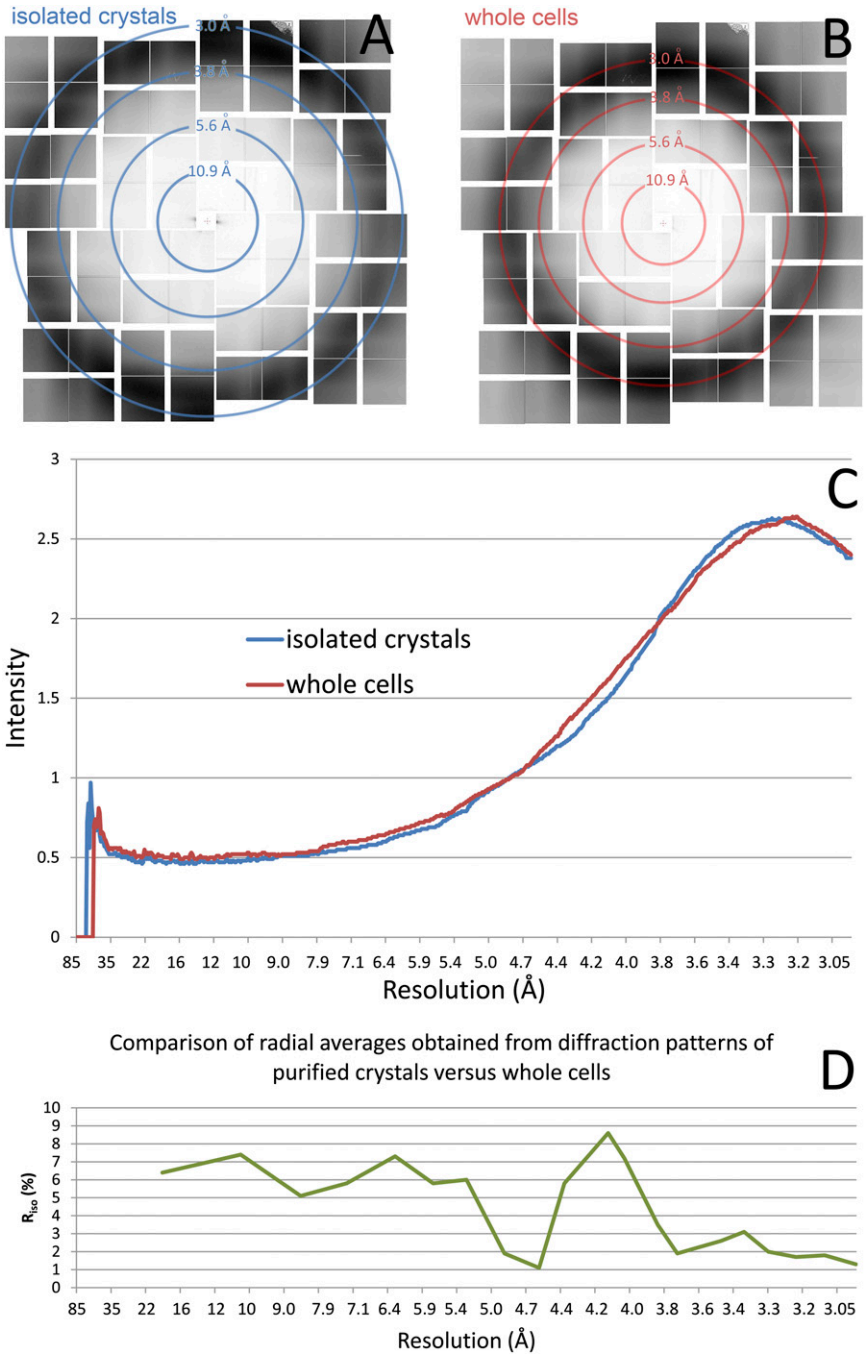

Fig. S4. The difference in background scattering between diffraction patterns collected from whole cell and isolated crystals is small. (*A*) Virtual powder pattern obtained from 645 diffraction patterns recorded from isolated crystals and indexed by cctbx.xfel software. The intensity value at each pixel in the virtual powder pattern averaged among the 645 individual patterns. (*B*) The analogous virtual powder pattern obtained from 458 diffraction patterns (indexed) recorded from whole cells. (*C*) Radial profiles comparing runs collected from isolated crystals (blue trace) and whole cells (red trace). The profiles were scaled by the CCP4 program SCALEIT (1) using two scaling parameters: a constant and a resolution-dependent exponential (B-factor). The discrepancy between traces appears insignificant throughout the resolution range, indicating that the scattering from cell components does not limit data quality. These two runs lasted approximately the same time (5 min 24 s) and contain approximately the same number of indexed patterns. The sample flow rate for whole cells was higher (30 μL/min) compared with that for isolated crystals (22 μL/min), suggesting that the diameter of the sample jet was larger for whole cells, but it does not significantly increase the background scattering. (*D*) Isomorphous R-factor calculated between the radial profiles depicted in *C*. The variation in R-factor with resolution is small, indicating that the scattering from cell components does not limit data quality.

1. Winn MD, et al. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* 67(Pt 4):235–242.

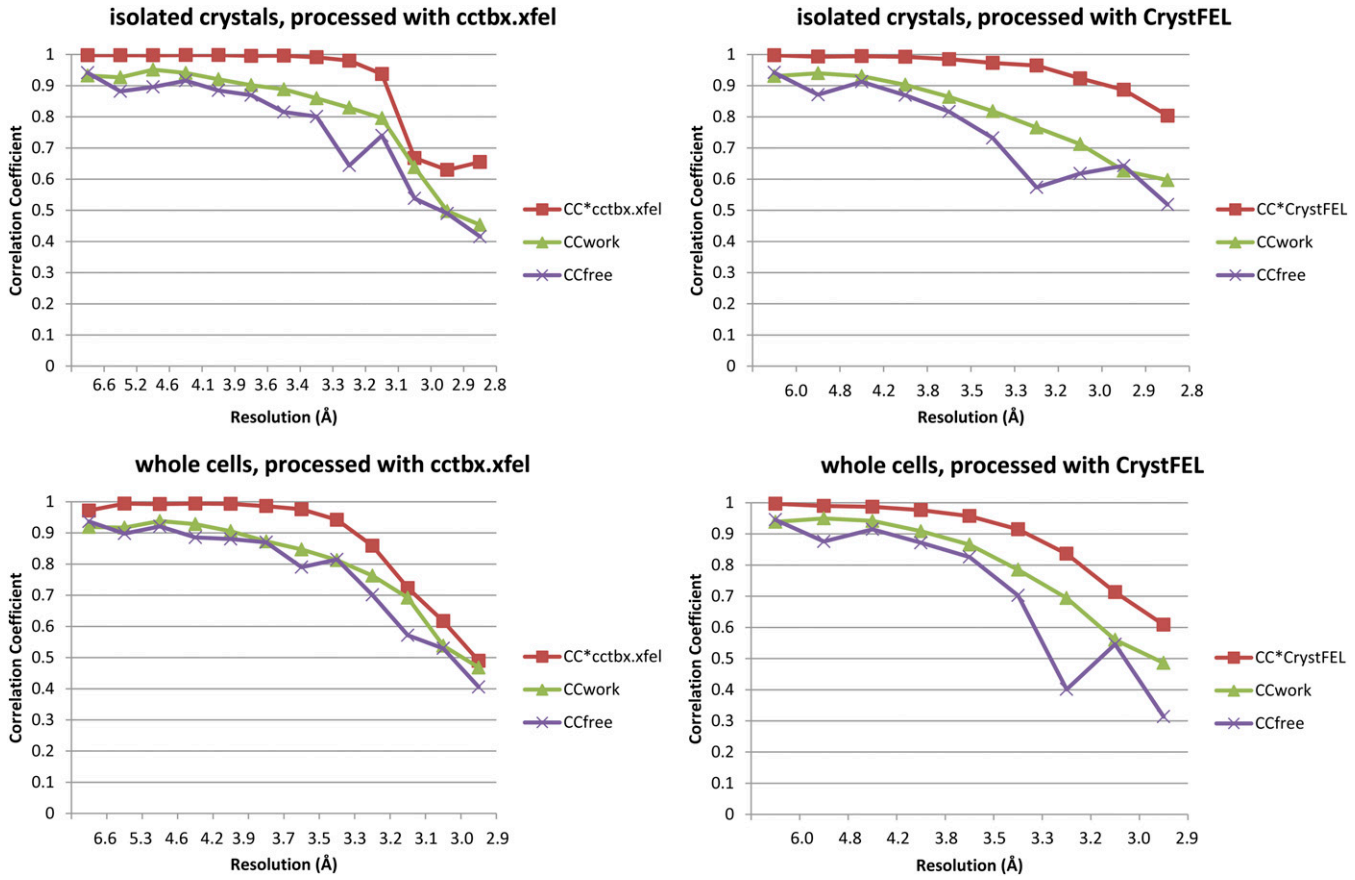

Fig. S5. Plots of CC*, CC_{work}, and CC_{free} reveal no indication of overfitting in any of the four refinements conducted here. CC* is an estimate of the correlation between the measured data and hypothetical noise-free signal (1). It is derived mathematically from $CC_{1/2}$, which measures the correlation between two randomly chosen halves of the unmerged data set. CC_{work} is the correlation between the measured structure factors in the working set and the corresponding structure factors calculated from the model coordinates. If we had overfit the model, CC_{work} would be larger than CC* in the highest-resolution shell. It would indicate that the model agrees better with the experimental data than the true signal does. In none of the four refinements do these statistics indicate overfitting. Observation of a gap at high resolution between CC_{free} and CC* in the CrystFEL refinements suggests that there is signal in the CrystFEL-processed data that is not accounted for in the CrystFEL-produced model. We speculate that this unaccounted signal might correspond to the ordered water molecules that we were unable to include in CrystFEL-produced model due to the high Wilson B-factor.

1. Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. *Science* 336(6084):1030–1033.

Table S1. Comparison of data collection and refinement statistics for isolated Cry3A crystals and whole Bt cells at CXI-LCLS processed with cctbx.xfel and CrystFEL to 2.9-Å resolution

Parameter	Isolated crystals (cctbx.xfel)	Isolated crystals (CrystFEL*)	Whole cells (cctbx.xfel)	Whole cells (CrystFEL*)
Measurements	5,404,411	10,746,964	4,383,931	11,084,440
Multiplicity	294.1 (1.9)	575.3 (438.6)	242.5 (1.3)	593.3 (569.0)
Unique reflections	18,312	18,635	17,360	18,635
R_{split} (%)	11.9 (62.6)	16.15 (36.18)	21.6 (90.6)	21.95 (47.60)
$CC_{1/2}$ (%)	96.2 (20.1)	92.20 (71.81)	90.6 (13.6)	83.17 (39.78)
I/σ(I)	67.5 (2.2)	8.79 (1.11)	59.2 (2.3)	6.65 (0.71)

The statistics in this table offer a more straightforward comparison of data quality over those reported in Table 2 because the number of images included for processing here was controlled to be the same for each sample (30,000 images each). The statistics suggest that the data from isolated crystals are slightly better. The difference may be due to variation in liquid jet diameter, beam transmission, or scattering from cell components present in the whole cell sample, but not the isolated crystal sample.
*The version of CrystFEL used for this analysis was 0.5.3a.