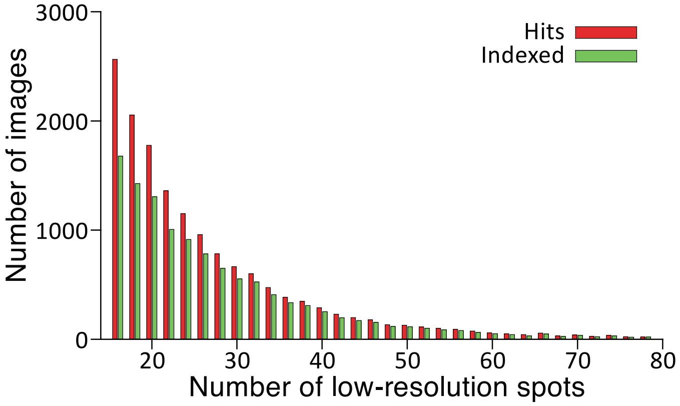
# Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers

Johan Hattne, Nathaniel Echols, Rosalie Tran, Jan Kern, Richard J. Gildea, Aaron S. Brewster, Roberto Alonso-Mori, Carina Glöckner, Julia Hellmich, Hartawan Laksmono, Raymond G. Sierra, Benedikt Lassalle-Kaiser, Alyssa Lampe, Guangye Han, Sheraz Gul, Dörte DiFiore, Despina Milathianaki, Alan R. Fry, Alan Miahnahri, William E. White, Donald W. Schafer, M. Marvin Seibert, Jason E. Koglin, Dimosthenis Sokaras, Tsu-Chien Weng, Jonas Sellberg, Matthew J. Latimer, Pieter Glatzel, Petrus H. Zwart, Ralf W. Grosse-Kunstleve, Michael J. Bogan, Marc Messerschmidt, Garth J. Williams, Sébastien Boutet, Johannes Messinger, Athina Zouni, Junko Yano, Uwe Bergmann, Vittal K. Yachandra, Paul D. Adams, Nicholas K. Sauter

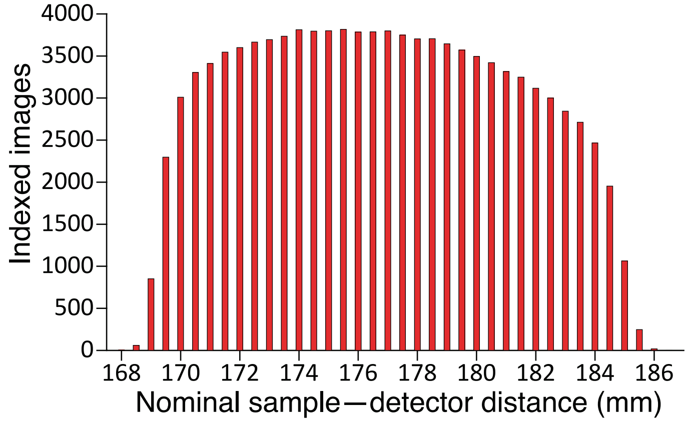
**Supplementary Figure 1 | Indexing success.**



| Interval | Indexed/Hits | [%] |
|---|---|---|
| 16-17 | 1674 / 2561 | [65] |
| 18-19 | 1423 / 2051 | [70] |
| 20-21 | 1302 / 1773 | [73] |
| 22-23 | 1003 / 1357 | [74] |
| 24-25 | 911 / 1148 | [79] |
| 26-27 | 778 / 955 | [82] |
| 28-29 | 646 / 778 | [83] |
| 30-31 | 549 / 662 | [83] |
| 32-33 | 522 / 597 | [87] |
| 34-35 | 405 / 470 | [86] |
| 36-37 | 332 / 381 | [87] |
| 38-39 | 305 / 344 | [89] |
| 40-41 | 250 / 286 | [87] |
| 42-43 | 194 / 226 | [86] |
| 44-45 | 168 / 195 | [86] |
| 46-47 | 151 / 175 | [86] |
| 48-49 | 114 / 130 | [88] |
| 50-51 | 112 / 126 | [65] |
| 52-53 | 98 / 109 | [70] |
| 54-55 | 84 / 96 | [73] |
| 56-57 | 77 / 88 | [74] |
| 58-59 | 59 / 72 | [79] |
| 60-61 | 48 / 55 | [82] |
| 62-63 | 40 / 47 | [83] |
| 64-65 | 28 / 39 | [83] |
| 66-67 | 45 / 53 | [87] |
| 68-69 | 22 / 28 | [86] |
| 70-71 | 33 / 36 | [87] |
| 72-73 | 20 / 23 | [89] |
| 74-75 | 28 / 33 | [87] |
| 76-77 | 16 / 20 | [86] |
| 78-79 | 18 / 18 | [86] |
| All | 11455/14932 | [77] |

Within the thermolysin dataset, 14932 diffraction patterns were identified as having between 16 and 79 candidate Bragg spots at low angles out to 4.0 Å resolution. Of these, 11455 were ultimately indexed and integrated.

**Supplementary Figure 2 | Distance calibration.**



| Distance (mm) | Indexed | Distance (mm) | Indexed |
|---|---|---|---|
| 168.0 | 5 | 177.5 | 3749 |
| 168.5 | 60 | 178.0 | 3704 |
| 169.0 | 852 | 178.5 | 3705 |
| 169.5 | 2297 | 179.0 | 3645 |
| 170.0 | 3010 | 179.5 | 3571 |
| 170.5 | 3304 | 180.0 | 3495 |
| 171.0 | 3411 | 180.5 | 3420 |
| 171.5 | 3546 | 181.0 | 3315 |
| 172.0 | 3601 | 181.5 | 3248 |
| 172.5 | 3665 | 182.0 | 3117 |
| 173.0 | 3695 | 182.5 | 3000 |
| 173.5 | 3735 | 183.0 | 2845 |
| 174.0 | 3813 | 183.5 | 2713 |
| 174.5 | 3796 | 184.0 | 2466 |
| 175.0 | 3800 | 184.5 | 1954 |
| 175.5 | 3817 | 185.0 | 1064 |
| 176.0 | 3785 | 185.5 | 248 |
| 176.5 | 3788 | 186.0 | 18 |
| 177.0 | 3798 | 186.5 | 1 |

Thermolysin data from run 21 were reprocessed with different trial distances, with the final value of 175.5 mm chosen on the basis of indexing success rate.  After this calibration, a beamline encoder provided relative offsets for data collected at different detector distances.

## Supplementary Table 1 | Data collection and refinement statistics.

| | **Thermolysin processed with *cctbx.xfel*** | **Lysozyme, 40fs exposure processed with *cctbx.xfel*** | **Lysozyme, 40 fs exposure processed with *CrystFEL*** |
|---|---|---|---|
| **Data collection** | | | *as reported in ref. 12:* |
| Mean wavelength (Å) | 1.269 ± 0.001 (*N* = 12,692)<br>1.297 ± 0.001 (*N* = 912) | 1.320 ± 0.002 (*N* = 21,743) | 1.32 |
| Space group | $P6_122$ | $P4_32_12$ | $P4_32_12$ |
| Cell dimensions[a] | | | |
| *a*, *c* (Å) | 92.9 ± 0.3, 130.4 ± 0.6 | 79, 38 | 79, 38 |
| Resolution[b] (Å) | 68.5–2.10 (2.18–2.10) | 39.5–1.90 (1.97–1.90) | 35.3–1.90 |
| No. collected images | 651,793 | 1,507,834 (runs 305-327)[c] | 1,471,615 (runs 305-327) |
| No. images used | 11,647 | 21,743 | 12,247 |
| No. lattices merged | 13,371 | 23,929 | 12,247 |
| No. total reflections | 19,854 (1,781) | 9923 (948) | 9921 |
| $R_{\mathrm{split}}$[d](%) | 24.4 (79.2) | 13.0 (25.0) | 15.8 |
| $CC_{1/2}$[e](%) | 84.5 (17.0) | 96.0 (75.2) | n.a. |
| $CC_{\mathrm{iso}}$[f](%) | 85.5 (36.9) | 82.9 (84.3) | n.a. |
| $I$ / $\sigma$ ($I$)[g] | 50.2 (5.6) | 97.1 (22.8) | 7.4 (2.8) |
| Completeness (%) | 99.1 (91.2) | 100.0 (100.0) | 98.3 (96.6) |
| Multiplicity | 209.0 (4.5) | 587.0 (125.7) | n.a. |
| Wilson *B* factor (Å$^2$) | 14.3 | 14.1 | 28.3 |
| | | | |
| **Refinement**[h] | | | *as re-refined by us:* |
| $R_{\mathrm{work}}$ / $R_{\mathrm{free}}$ (%) | 22.2 / 26.5 (32.2 / 37.4) | 18.7 / 22.9 (19.2 / 26.2) | 17.7 / 22.0 (20.2 / 33.7) |
| No. atoms | | | |
| Protein | 5094 | 1001 | 1001 |
| Ligand/ion | 5 | 2 | 2 |
| Water | 452 | 137 | 84 |
| *B*-factors (Å$^2$) | | | |
| Protein | 14.7 | 13.7 | 30.6 |
| Ligand/ion | 14.6 | 18.3 | 38.0 |
| Water | 24.1 | 28.1 | 40.2 |
| R.m.s. deviations | | | |
| Bond lengths (Å) | 0.003 | 0.004 | 0.008 |
| Bond angles (°) | 0.69 | 1.04 | 1.25 |
| Clashscore[i] | 0.86 | 3.06 | 3.06 |
| Ramachandran statistics | | | |
| Favored (%) | 96 | 98 | 99 |
| Outliers (%) | 0 | 0 | 0 |

[a] For thermolysin the distribution of unit cell dimensions was experimentally determined, leading to the population standard deviation given here. Unit cell dimensions for lysozyme were freely determined during indexing but constrained during merging to the same values reported in ref. 12, to facilitate a comparison between *cctbx.xfel* and *CrystFEL*.

[b] The high-resolution cutoff for thermolysin was determined as described in the text. For lysozyme it was constrained to the value (1.9 Å) reported in ref. 12 to facilitate the comparison. Statistics reported in parentheses represent values computed for the highest resolution shells (see **Supplementary Tables 2 and 3**).

[c] The run numbers represent the raw-data file serial numbers we believe were actually used to derive the ref. 12 structure factors, based on the diffraction image list deposited in the Coherent X-ray Imaging Data Bank (http://www.cxidb.org/data/17/40fs_5fs_indexed.txt) and the number of images (12,247) reported in the ref. 12 paper. The number of collected images we report (1,507,834) is the number actually present in those data files, which differs slightly from that (1,471,615) listed in the paper.

[d] $R_{\mathrm{split}}$ measures the percent difference between half-datasets as defined in ref. 12.

[e] $CC_{1/2}$ is the correlation coefficient between half-datasets defined in ref. 21.

[f] $CC_{\mathrm{iso}}$ is the correlation coefficient with a reference set of structure factor intensities. See **Supplementary Tables 2 and 3** for details.

[g] $I$ / $\sigma$ ($I$) values from the *cctbx.xfel* and *CrystFEL* programs can not be directly compared; see the **Online Methods**.

[h] In the lysozyme/*CrystFEL* column the refinement statistics represent our re-processing of the ref. 12 structure factors (as deposited in Protein Data Bank entry 4ET8) using our protocols, in order to control for any differences between our structure solution procedures and those employed by the other group. For both lysozyme refinements we used the same $R_{\mathrm{free}}$ flags as were originally used in ref. 12.

[i] Clashscore is the number of bad all-atom clashes per thousand atoms from *MolProbity*[55].

**Supplementary Table 2 | Thermolysin merging statistics by resolution bin.**[a]

| Resolution range (Å) | # Lattices[b] | # Measurements | # Unique reflections | Completeness (%) | <Multiplicity> | $\langle I/\sigma(I)\rangle$ | $R_{\mathrm{split}}$(%) | $CC_{1/2}$ (%) | $CC_{\mathrm{iso}}$[c] (%) |
|---|---|---|---|---|---|---|---|---|---|
| ∞ – 4.53 | 13,371 | 1,589,278 | 2,196 | 100.0 | 723.7 | 123.4 | 15.4 | 90.6 | 86.0 |
| 4.5 – 3.59 | 11,678 | 1,008,350 | 2,050 | 100.0 | 491.9 | 123.8 | 16.2 | 86.7 | 89.8 |
| 3.59 – 3.14 | 10,300 | 627,340 | 2,014 | 100.0 | 311.5 | 78.7 | 18.1 | 85.6 | 88.5 |
| 3.14 – 2.85 | 8,326 | 418,197 | 1,998 | 100.0 | 209.3 | 55.4 | 19.3 | 82.9 | 86.6 |
| 2.85 – 2.65 | 6,770 | 249,160 | 1,978 | 100.0 | 126.0 | 36.9 | 23.8 | 79.8 | 86.6 |
| 2.65 – 2.49 | 5,482 | 130,094 | 1,983 | 100.0 | 65.6 | 24.2 | 29.7 | 70.0 | 83.0 |
| 2.49 – 2.37 | 4,416 | 66,763 | 1,955 | 100.0 | 34.2 | 16.6 | 39.9 | 55.5 | 76.6 |
| 2.37 – 2.26 | 3,556 | 34,321 | 1,953 | 100.0 | 17.6 | 11.8 | 52.6 | 27.7 | 64.0 |
| 2.26 – 2.18 | 2,648 | 17,632 | 1,946 | 99.2 | 9.1 | 8.3 | 62.7 | 36.1 | 59.9 |
| 2.18 – 2.10 | 1,700 | 8,009 | 1,781 | 91.2 | 4.5 | 5.6 | 79.2 | 17.0 | 36.9 |
| ∞ – 2.10 | 13,371 | 4,149,144 | 19,854 | 99.1 | 209.0 | 50.2 | 24.2 | 84.5 | 85.5 |
| 2.10–2.03[d] | 529 | 3,300 | 1,246 | 63.8 | 2.7 | 4.3 | 86.6 | 14.5 | 23.2 |
| 2.03 – 1.98 | 225 | 1,935 | 910 | 47.4 | 2.1 | 3.6 | 90.3 | 53.9 | 28.9 |
| 1.98 – 1.92 | 209 | 1,241 | 696 | 35.4 | 1.8 | 3.0 | 91.4 | 2.8 | 27.6 |
| 1.92 – 1.88 | 191 | 740 | 484 | 25.1 | 1.5 | 2.8 | 85.4 | 41.9 | n.a. |
| 1.88 – 1.84 | 149 | 409 | 303 | 15.6 | 1.4 | 1.8 | 99.8 | 6.5 | n.a. |
| 1.84 – 1.80 | 23 | 59 | 59 | 3.1 | 1.0 | 1.5 | n.a. | n.a. | n.a. |
| 1.80 – 1.76 | 9 | 14 | 13 | 0.7 | 1.1 | 1.9 | n.a. | n.a. | n.a. |
| ∞ – 1.76 | 13,371 | 4,156,853 | 23,565 | 70.1 | 176.4 | 42.8 | 26.0 | 79.2 | 81.1 |

[a] In **Supplementary Tables 2 and 3**, resolution bins are chosen so as to have equal reciprocal space volumes, which is the default option in the *cctbx* toolkit.

[b] This column reports the number of lattices that have been successfully indexed and integrated and that contain any integrated observations in the indicated resolution range. The full set of 13,371 lattices was used to produce the merged thermolysin data in **Supplementary Table 1**. Successively smaller nested subsets of this collection contained data in successively higher resolution bins, with the smallest nested subset (of only 1700 lattices) extending to 2.1 Å. We refer to "lattices" instead of "images" because some images contain diffraction lattices from two distinct microcrystals (**Fig. 1d**) that can be usefully integrated with *cctbx.xfel*.

[c] Correlation coefficient between the XFEL-observed structure factor intensities and the previously published structure factors from Protein Data Bank entry 2TLI. The $\alpha$–carbon r.m.s.d. between the two respective models is 0.41 Å. Both structures were determined at room temperature.

[d] Data in the 2.10 – 1.76 Å resolution ranges were not included in the final averaged structure factor set, because they do not satisfy our inclusion criterion (**Fig. 2e**). They are listed here simply to document how many images were observed to contain high-resolution spots in the detector corners. Note that the scarce data in these resolution ranges have insufficient multiplicities (1.0 – 2.7) for accurate averaging, which is also reflected in the poorly-behaved $CC_{1/2}$ and $R_{\mathrm{split}}$ statistics that compare merged half-datasets.

**Supplementary Table 3 | 40 fs-pulse lysozyme merging statistics by resolution bin, for the *cctbx.xfel*-processed data.**[a]

| Resolution range (Å) | # Lattices[b] | # Measurements | # Unique reflections | Completeness (%) | <Multiplicity> | $\langle I/\sigma(I) \rangle$ | $R_{split}$(%) | $CC_{1/2}$ (%) | $CC_{iso}$[c] (%) |
|---|---|---|---|---|---|---|---|---|---|
| ∞ – 4.09 | 23,929 | 1,443,414 | 1,089 | 99.9 | 1325.5 | 215.5 | 10.8 | 95.8 | 78.8 |
| 4.09 – 3.25 | 23,260 | 1,153,093 | 1,022 | 100.0 | 1128.3 | 208.4 | 10.5 | 94.0 | 95.0 |
| 3.25 – 2.84 | 22,014 | 838,213 | 999 | 100.0 | 839.1 | 137.3 | 12.3 | 92.9 | 95.8 |
| 2.84 – 2.58 | 19,768 | 584,099 | 989 | 100.0 | 590.6 | 96.5 | 13.4 | 91.7 | 94.3 |
| 2.58 – 2.39 | 17,230 | 529,545 | 978 | 100.0 | 541.5 | 82.0 | 12.9 | 91.4 | 93.6 |
| 2.39 – 2.25 | 15,513 | 449,561 | 977 | 100.0 | 460.1 | 67.6 | 15.8 | 88.1 | 89.1 |
| 2.25 – 2.14 | 13,672 | 314,469 | 967 | 100.0 | 325.2 | 49.5 | 17.1 | 89.3 | 90.7 |
| 2.14 – 2.05 | 11,486 | 231,022 | 986 | 100.0 | 234.3 | 39.1 | 18.7 | 86.8 | 90.6 |
| 2.05 – 1.97 | 9,599 | 161,713 | 968 | 100.0 | 167.1 | 29.7 | 21.5 | 81.7 | 89.6 |
| 1.97 – 1.90 | 7,925 | 119,195 | 948 | 100.0 | 125.7 | 22.8 | 25.0 | 75.2 | 84.3 |
| ∞ – 1.90 | 23,929 | 5,824,324 | 9,923 | 100.0 | 587.0 | 97.1 | 13.0 | 96.0 | 82.9 |

[a] In **Supplementary Tables 2 and 3**, resolution bins are chosen so as to have equal reciprocal space volumes, which is the default option in the *cctbx* toolkit.

[b] The full dataset contained 21,743 images successfully indexed and integrated. Some contained diffraction lattices from two distinct microcrystals, thus giving a total of 23,929 lattices.

[c] Correlation coefficient between the *cctbx.xfel*-processed structure factor intensities and the published *CrystFEL*-processed structure factors from Protein Data Bank entry 4ET8. The $\alpha$–carbon r.m.s.d. between the two respective models is 0.08 Å.

## Supplementary Note | Targeting the exact pixels that contain signal

The concept of tailoring the Bragg spot shape model to the data at hand is highlighted by the recent work of Kirian *et al.*[14] that analyzes photosystem I (PS I) XFEL XRD[4]. The observed diffraction fringes connecting each low-angle Bragg spot with its six nearest neighbors arise when the crystallite contains only a small number of unit cells along each crystal axis. Therefore a main challenge in that work was to select just the central portion of each Bragg peak for data integration. The width of the central peak scales as $\lambda/a$ (where $\lambda$ is the wavelength of the incident light, and $a$ is the width of the crystal). Accordingly, the *CrystFEL* program was developed with a single adjustable parameter, $\delta \approx 1/a$, and it considers a pixel to contain signal if the lattice model places the corresponding reciprocal point within reciprocal distance $\delta$ of the sphere of reflection (the Ewald sphere). A single $\delta$ radius is chosen that best models the entire ensemble of diffraction images. The treatment of Bragg spots as reciprocal space spheres instead of complicated fringe functions is an approximation well-suited to the 8.5 Å resolution PS I data[4]. In contrast, the high-resolution experiments described here use hard X-rays ($\lambda$ ~ 1.3 Å, instead of 7 Å), and crystal sizes are larger by a factor of ~10, reducing the $\lambda/a$ broadening to less than one pixel. The size and shape of Bragg spots is determined by other effects, familiar from SR crystallography, including the presence of microscopic "mosaic" crystal domains, which cause Bragg spots to appear as concentric arcs if there is a distribution of orientations, or as large circles if there is a distribution of cell dimensions. Also, XFEL experiments performed with self-amplified spontaneous emission (SASE) pulses have a considerably larger bandwidth than SR experiments, causing Bragg spots to be streaked in the radial direction. The result is that Bragg spots in high-resolution XFEL experiments are markedly different from the integration masks of constant radius currently used by *CrystFEL,* particularly at higher diffraction angles.