Investigation of protein structure determination

using X-ray free-electron lasers

Dissertation
zur Erlangung des Doktorgrades
des Fachbereich Physik
der Universität Hamburg

vorgelegt von
Karol Jan Nass
aus Chorzów

Hamburg
2013

Gutachter der Dissertation:	Prof. dr. Henry Chapman Prof. dr. Edgar Weckert
Gutachter der Disputation:	Prof. dr. Henry Chapman Prof. dr. Franz Kärtner
Datum der Disputation:	25 February 2013
Vorsitzender des Prüfungsausschusses:	Prof. dr. Michael A. Rübhausen
Vorsitzender des Promotionsausschusses:	Prof. dr. Peter Hauschildt
Dekan der MIN-Fakultät:	Prof. dr. Heinrich Graener

Author email: karol.nass@gmail.com

To my father.

Abstract

With the advent of fourth generation radiation sources, X-ray free-electron lasers (X-FEL's), several fields of research, including atomic and molecular sciences, matter at extreme conditions, X-ray imaging, obtained a tool that allows the realization of experiments at conditions and time scales previously inaccessible for researchers. Time scales of atomic motions and extreme temperatures and pressures that occur naturally in the cores of the largest stars became available to the researchers disposal.

X-FEL radiation generation is based on the self-amplified spontaneous emission principle (SASE). Timing properties of an X-FEL, combined with femtosecond optical or infra-red laser that is used to trigger specific states in the sample, enabled studies of processes that occur at time scales of atomic motion in a time-resolved manner. One of the promising applications of this new radiation source is the possibility to expand the area of biomolecular structure determination by using intense, ultra-short X-ray pulses [Neu12]. The first step in the direction of single biomolecule imaging was made by retrieving the structure of a known protein arranged in a nano-sized crystal [Bou12, Cha11].

In this dissertation an exhaustive study about the methodology and capabilities of Serial Femtosecond Crystallography (SFX) for novel protein structure determination is presented. Two unknown protein structures were solved from tens of thousands of *in-vivo* grown micro-crystals using X-FEL radiation.

Comprehensive verification and assessment of SFX for novel protein structure determination from micro-crystals using homologous structures to solve the phase problem is presented and necessary improvements to this method are discussed. Description and influence of the variation of X-ray pulses and protein crystals properties are presented. Possibilities of *de-novo* protein structure determination using the high intensity multi-wavelength anomalous dispersion (MAD) method [Son11] or by direct phase retrieval using information encoded in shape transforms around Bragg peaks measured from protein nano-crystals [Spe11] are also discussed.

Acknowledgements

Many people have given me valuable encouragement, support and constructive criticism during the graduate work and writing of this thesis. Thank you to: Henry Chapman, Christian Betzel, Martin Aepfelbacher, Markus Perbandt, members of the Coherent Imaging Division in the Center for Free-electron Laser Science and graduate students of the Hamburg School for Structure and Dynamics in Infection (SDI). Special thanks to our group secretary Irmtraud Kleine for help in resolving administrative issues.

I'm very grateful to scientific and administrative personnel at the SLAC National Laboratory in Stanford for help and resources during the experiments at the Linac Coherent Light Source.

Hamburg, *Karol Nass*
February 2013

List of papers

This coherent monograph is based on my graduate work that has been a part of a collaborative effort of many people and institutions from different countries. This manuscript is based on the following papers.

All results presented in the manuscript are based on my individual analysis of data acquired by the collaboration at LCLS and on my individual work before experiments at LCLS in preparation laboratory, except for the production of *in-vivo* crystals.

Data evaluation was performed by myself using software developed by the collaboration with my contribution or using software commonly used for that type of analysis.

I Redecke, L.[1], **Nass, K.**[1], DePonte, D. P., White, T. A., Rehders, D., Barty, A., Stellato, F., Liang, M., Barends, T. R. M., Boutet, S., Williams, G. J., Messerschmidt, M., Seibert, M. M., Aquila, A., Arnlund, D., Bajt, S., Barth, T., Bogan, M. J., Caleman, C., Chao, T., Doak, R. B., Fleckenstein, H., Frank, M., Fromme, R., Galli, L., Grotjohann, I., Hunter, M. S., Johansson, L. C., Kassemeyer, S., Katona, G., Kirian, R. A., Koopmann, R., Kupitz, C., Lomb, L., Martin, A. V., Mogk, S., Neutze, R., Shoeman, R. L., Steinbrener, J., Timneanu, N., Wang, D., Weierstall, U., Zatsepin, N. A., Spence, J. C. H., Fromme, P., Schlichting, I., Duszenko, M., Betzel, C., Chapman, H. N. (2013) Natively inhibited Trypanosoma brucei cathepsin B structure determined by using an X-ray laser. *Science* vol. 339:pp. 227–230.

II Koopmann, R.[1], Cupelli, K.[1], Redecke, L.[1], **Nass, K.**, DePonte, D. P., White, T. A., Stellato, F., Rehders, D., Liang, M., Andreasson, J., Aquila, A., Bajt, S., Barthelmess, M., Barty, A., Bogan, M. J., Bostedt, C., Boutet, S., Bozek, J.

[1]These authors contributed equally to the publication.

D., Caleman, C., Coppola, N., Davidsson, J., Doak, R. B., Ekeberg, T., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Graafsma, H., Gumprecht, L., Hajdu, J., Hampton, C. Y., Hartmann, A., Hartmann, R., Hauser, G., Hirsemann, H., Holl, P., Hunter, M. S., Kassemeyer, S., Kirian, R. A., Lomb, L., Maia, F. R. N. C., Kimmel, N., Martin, A. V., Messerschmidt, M., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schlichting, I., Schulz, J., Seibert, M. M., Shoeman, R. L., Sierra, R. G., Soltau, H., Stern, S., Struder, L., Timneanu, N., Ullrich, J., Wang, X., Weidenspointner, G., Weierstall, U., Williams, G. J., Wunderer, C. B., Fromme, P., Spence, J. C. H., Stehle, T., Chapman, H. N., Betzel, C., Duszenko, M. (2012) In vivo protein crystallization opens new routes in structural biology. *Nature Methods,* vol. 9(3):pp. 259-262.

III White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., **Nass, K.**, Barty, A., Chapman, H. N. (2012) CrystFEL: a software suite for snapshot serial crystallography. *Journal of Applied Crystallography,* vol. 45(2):pp. 335-341.

IV DePonte, D. P., **Nass, K.**, Stellato, F., Liang, M., Chapman, H. N. (2011) Sample Injection for Pulsed X-ray Sources. *Proceedings of SPIE, Advances in X-ray Free- Electron Lasers: Radiation Schemes, X-ray Optics, and Instrumentation,* 80780.

V Boutet, S., Lomb, L., Williams, G. J., Barends, T. R. M., Aquila, A., Doak, R. B., Weierstall, U., DePonte, D. P., Steinbrener, J., Shoeman, R. L., Messerschmidt, M., Barty, A., White, T. A., Kassemeyer, S., Kirian, R. A., Seibert, M. M., Montanez, P. A., Kenney, C., Herbst, R., Hart, P., Pines, J., Haller, G., Gruner, S. M., Philipp, H. T., Tate, M. W., Hromalik, M., Koerner, L. J., van Bakel, N., Morse, J., Ghonsalves, W., Arnlund, D., Bogan, M. J., Caleman, C., Fromme, R., Hampton, C. Y., Hunter, M. S., Johansson, L. C., Katona, G., Kupitz, C., Liang, M., Martin, A. V. **Nass, K.**, Redecke, L., Stellato, F., Timneanu, N., Wang, D., Zatsepin, N. A., Schafer, D., Defever, J., Neutze, R., Fromme, P., Spence, J. C. H., Chapman, H. N., Schlichting, I. (2012) High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. *Science,* vol. 337(6092):pp. 362-364.

VI Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., Doak, R. B., Maia, F. R. N. C., Martin, A. V., Schlichting, I., Lomb, L., Coppola, N., Shoeman, R. L., Epp, S. W., Hartmann, R., Rolles, D., Rudenko, A., Foucar, L., Kimmel, N., Weidenspointner, G., Holl, P., Liang, M., Barthelmess, M., Caleman, C., Boutet, S., Bogan, M. J., Krzywinski, J., Bostedt, C., Bajt, S., Gumprecht, L., Rudek, B., Erk, B., Schmidt, C., Homke, A., Reich, C., Pietschner, D., Struder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G.,

Schopper, F., Soltau, H., Kuhnel, K., Messerschmidt, M., Bozek, J. D., Hau-Riege, S. P., Frank, M., Hampton, C. Y., Sierra, R. G., Starodub, D., Williams, G. J., Hajdu, J., Timneanu, N., Seibert, M. M., Andreasson, J., Rocker, A., Jonsson, O., Svenda, M., Stern, S., **Nass, K.**, Andritschke, R., Schroter, C., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton, J. M., Barends, T. R. M., Neutze, R., Marchesini, S., Fromme, R., Schorb, S., Rupp, D., Adolph, M., Gorkhover, T., Andersson, I., Hirsemann, H., Potdevin, G., Graafsma, H., Nilsson, B., Spence, J. C. H. (2011) Femtosecond X-ray protein nanocrystallography. *Nature,* vol. 470(7332):pp. 73-77.

List of additional papers

Additionally, I present a list of publications to which I also contributed during my graduate work. My contribution here can be generally expressed by: (i) taking part in preparations for experiments, (ii) contribution in experimental tasks necessary for data acquisition during experiments at LCLS, (iii) partial contribution in data analysis.

VII Aquila, A., Hunter, M. S., Doak, R. B., Kirian, R. A., Fromme, P., White, T. A., Andreasson, J., Arnlund, D., Bajt, S., Barends, T. R. M., Barthelmess, M., Bogan, M. J., Bostedt, C., Bottin, H., Bozek, J. D., Caleman, C., Coppola, N., Davidsson, J., DePonte, D. P., Elser, V., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Frank, M., Fromme, R., Graafsma, H., Grotjohann, I., Gumprecht, L., Hajdu, J., Hampton, C. Y., Hartmann, A., Hartmann, R., Hau-Riege, S., Hauser, G., Hirsemann, H., Holl, P., Holton, J. M., Homke, A., Johansson, L., Kimmel, N., Kassemeyer, S., Krasniqi, F., Kuhnel, K., Liang, M., Lomb, L., Malmerberg, E., Marchesini, S., Martin, A. V., Maia, F. R. N. C., Messerschmidt, M., **Nass, K**., Reich, C., Neutze, R., Rolles, D., Rudek, B., Rudenko, A., Schlichting, I., Schmidt, C., Schmidt, K. E., Schulz, J., Seibert, M. M., Shoeman, R. L., Sierra, R., Soltau, H., Starodub, D., Stellato, F., Stern, S., Struder, L., Timneanu, N., Ullrich, J., Wang, X., Williams, G. J., Weidenspointner, G., Weierstall, U., Wunderer, C., Barty, A., Spence, J. C. H., Chapman, H. N. (2012) Time-resolved protein nanocrystallography using an X-ray free-electron laser. *Optics Express,* vol. 20(3):pp. 2706-2716.

VIII Barty, A.[1], Caleman, C.[1], Aquila, A., Timneanu, N., Lomb, L., White, T. A., Andreasson, J., Arnlund, D., Bajt, S., Barends, T. R. M., Barthelmess, M., Bogan, M. J., Bostedt, C., Bozek, J. D., Coffee, R., Coppola, N., Davidsson,

[1]These authors contributed equally to the publication.

J., DePonte, D. P., Doak, R. B., Ekeberg, T., Elser, V., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Fromme, P., Graafsma, H., Gumprecht, L., Hajdu, J., Hampton, C. Y., Hartmann, R., Hartmann, A., Hauser, G., Hirsemann, H., Holl, P., Hunter, M. S., Johansson, L., Kassemeyer, S., Kimmel, N., Kirian, R. A., Liang, M., Maia, F. R. N. C., Malmerberg, E., Marchesini, S., Martin, A. V., **Nass, K.**, Neutze, R., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Scott, H., Schlichting, I., Schulz, J., Seibert, M. M., Shoeman, R. L., Sierra, R. G., Soltau, H., Spence, J. C. H., Stellato, F., Stern, S., Struder, L., Ullrich, J., Wang, X., Weidenspointner, G., Weierstall, U., Wunderer, C. B., Chapman, H. N. (2011) Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nature Photonics,* vol. 6(1):pp. 35-40.

IX Johansson, L. C., Arnlund, D., White, T. A., Katona, G., DePonte, D. P., Weierstall, U., Doak, R. B., Shoeman, R. L., Lomb, L., Malmerberg, E., Davidsson, J., **Nass, K.**, Liang, M., Andreasson, J., Aquila, A., Bajt, S., Barthelmess, M., Barty, A., Bogan, M. J., Bostedt, C., Bozek, J. D., Caleman, C., Coffee, R., Coppola, N., Ekeberg, T., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Graafsma, H., Gumprecht, L., Hajdu, J., Hampton, C. Y., Hartmann, R., Hartmann, A., Hauser, G., Hirsemann, H., Holl, P., Hunter, M. S., Kassemeyer, S., Kimmel, N., Kirian, R. A., Maia, F. R. N. C., Marchesini, S., Martin, A. V., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schlichting, I., Schulz, J., Seibert, M. M., Sierra, R. G., Soltau, H., Starodub, D., Stellato, F., Stern, S., Struder, L., Timneanu, N., Ullrich, J., Wahlgren, W. Y., Wang, X., Weidenspointner, G., Wunderer, C., Fromme, P., Chapman, H. N., Spence, J. C. H. Neutze, R. (2012) Lipidic phase membrane protein serial femtosecond crystallography. *Nature Methods,* vol. 9(3):pp. 263-265.

X Kassemeyer, S., Steinbrener, J., Lomb, L., Hartmann, E., Aquila, A., Barty, A., Martin, A. V., Hampton, C. Y. Bajt, S., Barthelmess, M., Barends, T. R. M., Bostedt, C., Bott, M., Bozek, J. D., Coppola, N., Cryle, M., DePonte, D. P., Doak, R. B., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Graafsma, H., Gumprecht, L., Hartmann, A., Hartmann, R., Hauser, G., Hirsemann, H., Homke, A., Holl, P., Jonsson, O., Kimmel, N., Krasniqi, F., Liang, M., Maia, F. R. N. C., Marchesini, S., **Nass, K.**, Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schmidt, C., Schulz, J., Shoeman, R. L., Sierra, R. G., Soltau, H., Spence, J. C. H., Starodub, D., Stellato, F., Stern, S., Stier, G., Svenda, M., Weidenspointner, G., Weierstall, U., White, T. A., Wunderer, C., Frank, M., Chapman, H. N., Ullrich, J., Struder, L., Bogan, M. J., Schlichting, I., (2012) Femtosecond free-electron laser X-ray diffraction data sets for algorithm development. *Optics Express,* vol. 20(4):pp. 4149-4158.

XI Loh, N. D., Hampton, C. Y., Martin, A. V., Starodub, D., Sierra, R. G., Barty,

A., Aquila, A., Schulz, J., Lomb, L., Steinbrener, J., Shoeman, R. L., Kassemeyer, S., Bostedt, C., Bozek, J., Epp, S. W., Erk, B., Hartmann, R., Rolles, D., Rudenko, A., Rudek, B., Foucar, L., Kimmel, N., Weidenspointner, G., Hauser, G., Holl, P., Pedersoli, E., Liang, M., Hunter, M. M., Gumprecht, L., Coppola, N., Wunderer, C., Graafsma, H., Maia, F. R. N. C., Ekeberg, T., Hantke, M., Fleckenstein, H., Hirsemann, H., **Nass, K.**, White, T. A., Tobias, H. J., Farquar, G. R., Benner, W. H., Hau-Riege, S. P., Reich, C., Hartmann, A., Soltau, H., Marchesini, S., Bajt, S., Barthelmess, M., Bucksbaum, P., Hodgson, K. O., Strder, L., Ullrich, J., Frank, M., Schlichting, I., Chapman, H. N., Bogan, M. J., (2012) Fractal morphology, imaging and mass spectrometry of single aerosol particles in flight. *Nature,* vol. 486(7404):pp. 513-517.

XII Lomb, L., Barends, T., Kassemeyer, S., Aquila, A., Epp, S., Erk, B., Foucar, L., Hartmann, R., Rudek, B., Rolles, D., Rudenko, A., Shoeman, R., Andreasson, J., Bajt, S., Barthelmess, M., Barty, A., Bogan, M., Bostedt, C., Bozek, J., Caleman, C., Coffee, R., Coppola, N., DePonte, D., Doak, R. B., Ekeberg, T., Fleckenstein, H., Fromme, P., Gebhardt, M., Graafsma, H., Gumprecht, L., Hampton, C., Hartmann, A., Hauser, G., Hirsemann, H., Holl, P., Holton, J., Hunter, M., Kabsch, W., Kimmel, N., Kirian, R., Liang, M., Maia, F. R. N., Meinhart, A., Marchesini, S., Martin, A., **Nass, K.**, Reich, C., Schulz, J., Seibert, M. M., Sierra, R., Soltau, H., Spence, J. C., Steinbrener, J., Stellato, F., Stern, S., Timneanu, N., Wang, X., Weidenspointner, G., Weierstall, U., White, T., Wunderer, C., Chapman, H., Ullrich, J., Struder, L., Schlichting, I. (2011) Radiation damage in protein serial femtosecond crystallography using an X-ray free-electron laser. *Physical Review B,* vol. 84(21).

XIII Martin, A. V., Andreasson, J., Aquila, A., Bajt, S., Barends, T. R. M., Barthelmess, M., Barty, A., Benner, W. H., Bostedt, C., Bozek, J. D., Bucksbaum, P., Caleman, C., Coppola, N., DePonte, D. P., Ekeberg, T., Epp, S. W., Erk, B., Farquar, G. R., Fleckenstein, H., Foucar, L., Frank, M., Gumprecht, L., Hampton, C. Y., Hantke, M., Hartmann, A., Hartmann, E., Hartmann, R., Hau-Riege, S. P., Hauser, G., Holl, P., Hoemke, A., Jonsson, O., Kassemeyer, S., Kimmel, N., Kiskinova, M., Krasniqi, F., Krzywinski, J., Liang, M., Loh, N. D. Lomb, L., Maia, F. R. N. C., Marchesini, S., Messerschmidt, M., **Nass, K.**, Odic, D., Pedersoli, E., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schmidt, C., Schultz, J., Seibert, M. M., Shoeman, R. L., Sierra, R. G., Soltau, H., Starodub, D., Steinbrener, J., Stellato, F., Struder, L., Svenda, M., Tobias, H., Ullrich, J., Weidenspointner, G., Westphal, D., White, T. A., Williams, G., Hajdu, J., Schlichting, I., Bogan, M. J., Chapman, H. N. (2011) Single particle imaging with soft X-rays at the Linac Coherent Light Source. *Proceedings of SPIE, Advances in X-ray Free-Electron Lasers: Radiation Schemes, X-ray Op-*

tics, and Instrumentation, 8078.

XIV Martin, A. V., Wang, F., Loh, N. D., Ekeberg, T., Maia, F. R. N. C., Hantke, M., van der Schot, G., Hampton, C. Y., Sierra, R. G. Aquila, A., Bajt, S., Barthelmess, M., Bostedt, C., Bozek, J. D., Coppola, N., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Frank, M., Graafsma, H., Gumprecht, L., Hartmann, A., Hartmann, R., Hauser, G., Hirsemann, H., Holl, P., Kassemeyer, S., Kimmel, N., Liang, M., Lomb, L., Marchesini, S., **Nass, K.**, Pedersoli, E., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schulz, J., Shoeman, R. L., Soltau, H., Starodub, D., Steinbrener, J., Stellato, F., Strder, L., Ullrich, J., Weidenspointner, G., White, T. A., Wunderer, C. B., Barty, A., Schlichting, I., Bogan, M. J., Chapman, H. N. (2012) Noise-robust coherent diffractive imaging with a single diffraction pattern. *Optics Express,* vol. 20(15):pp. 16650-16661.

XV Martin, A. V., Loh, N. D., Hampton, C. Y., Sierra, R. G., Wang, F., Aquila, A., Bajt, S., Barthelmess, M., Bostedt, C., Bozek, J. D., Coppola, N., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Frank, M., Graafsma, H., Gumprecht, L., Hartmann, A., Hartmann, R., Hauser, G., Hirsemann, H., Holl, P., Kassemeyer, S., Kimmel, N., Liang, M., Lomb, L., Maia, F. R. N. C., Marchesini, S., **Nass, K.**, Pedersoli, E., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schulz, J., Shoeman, R. L., Soltau, H., Starodub, D., Steinbrener, J., Stellato, F., Strder, L., Ullrich, J., Weidenspointner, G., White, T. A., Wunderer, C. B., Barty, A., Schlichting, I., Bogan, M. J., Chapman, H. N. (2012) Femtosecond dark-field imaging with an X-ray free electron laser. *Optics Express,* vol. 20(12):pp. 13501-13512.

XVI Yoon, C. H., Schwander, P., Abergel, C., Andersson, I., Andreasson, J., Aquila, A., Bajt, S., Barthelmess, M., Barty, A., Bogan, M. J., Bostedt, C., Bozek, J., Chapman, H. N., Claverie, J., Coppola, N., DePonte, D. P., Ekeberg, T., Epp, S. W., Erk, B., Fleckenstein, H., Foucar, L., Graafsma, H., Gumprecht, L., Hajdu, J., Hampton, C. Y., Hartmann, A., Hartmann, E., Hartmann, R., Hauser, G., Hirsemann, H., Holl, P., Kassemeyer, S., Kimmel, N., Kiskinova, M., Liang, M., Loh, N. D., Lomb, L., Maia, F. R. N. C., Martin, A. V., **Nass, K.**, Pedersoli, E., Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schlichting, I., Schulz, J., Seibert, M., Seltzer, V., Shoeman, R. L., Sierra, R. G., Soltau, H., Starodub, D., Steinbrener, J., Stier, G., Struder, L., Svenda, M., Ullrich, J., Weidenspointner, G., White, T. A., Wunderer, C., Ourmazd, A. (2011) Unsupervised classification of single-particle X ray diffraction snapshots by spectral clustering. *Optics Express,* vol. 19(17):pp. 16542–16551.

XVII Starodub, D., Aquila, A., Bajt, S., Barthelmess, M., Barty, A., Bostedt, C., Bozek, J. D., Coppola, N., Doak, R. B., Epp, S. W., Erk, B., Foucar, L.,

Gumprecht, L., Hampton, C. Y., Hartmann, A., Hartmann, R., Holl, P., Kassemeyer, S., Kimmel, N., Laksmono, H., Liang, M., Loh, N. D., Lomb, L., Martin, A. V., **Nass, K.**, Reich, C., Rolles, D., Rudek, B., Rudenko, A., Schulz, J., Shoeman, R. L., Sierra, R. G., Soltau, H., Steinbrener, J., Stellato, F., Stern, S., Weidenspointner, G., Frank, M., Ullrich, J., Strder, L., Schlichting, I., Chapman, H. N., Spence, J. C. H., Bogan, M. J. (2012) Single-particle structure determination by correlations of snapshot X-ray diffraction patterns. *Nature Communications,* vol. 3, 1276

Most publications listed above contain, within their content, a paragraph with detailed description of each authors contribution.

Contents

Chapter 1

Introduction

1.1 X-ray free-electron lasers for structural biology

For more than fifty years synchrotron based light sources have been used to produce significant results in a variety of research fields. Many of them are considered as scientific breakthroughs in physics, materials science, biological and life sciences, medicine, chemistry, environment and earth science. For example the giant magnetoresistance (GMR) effect has been studied using synchrotron radiation [Bai88, Bin89] which is used in our modern hard drives or biosensors. Synchrotron X-ray beams allow detailed analysis and modelling of strain, cracks and corrosion as well as in situ study of materials during production processing [Noy87]. This research performed usin synchrotrons is vital to the development of high-performance materials and their use in innovative products and structures. Pharmaceutical companies and medical researchers are making increasing use of macromolecular X-ray crystallography implemented at synchrotron beam lines. Structural information obtained from protein crystals allows structural biologists to develop compounds that are used as drug candidates. Both the anti-flu drug Tamiflu and Herceptin, used to treat breast cancer, benefited from synchrotron experiments. In addition to that scientists are using non-destructive synchrotron techniques to find answers to questions in palaeontology, archaeology, art history and forensics.

With the recent development and realization of new generation of synchrotron light source, the so called X-ray Free-Electron Laser (X-FEL), came the undeniable promise that more important discoveries can be made in the area of basic natural sciences, especially in biological and related pharmaceutical sciences, previously inaccessible for researchers due to technical limitations. The X-ray FEL provides femtosecond pulses orders of magnitude brighter than what is typically achieved at synchrotron sources.

Structural biology concentrates on retrieving molecular structures of macro-

molecules such as proteins or nucleic acids. Structures of proteins are of great interest to biologists, pharmacologists and medical researchers because proteins are the molecular basis of processes that occur in every living organism. Models of protein structures are required to understand how proteins work. Researchers have been using bright sources of X-rays to retrieve three dimensional models of proteins from their crystallized forms. One of the latest prominent examples was to retrieve the structure of ribosome, thus to understand the mechanism of protein translation on a molecular level. By knowing protein structures one can study the molecular basis of processes that are driven by proteins. One can also deduce, from protein structure, how to effectively influence protein function, for example by choosing or designing specific molecules (ligands) that change the native protein structure or protein interactions with other components of the cell. This approach is often applied in the search of effective drugs and treatments against diseases.

Another example of protein structure to function relationship, which is being investigated by many researchers, is to discover the molecular basis of enzymatic photosynthesis. That goal can be achieved only by understanding what is the underlying mechanism of photosynthesis on the molecular level in time resolved manner. Implications of expanding the field of structural biology might be tremendous and great amount of effort has been made to improve the methodology of protein structure determination.

The primary method for protein structure determination is X-ray crystallography performed using synchrotron sources. To date more than 70,000 protein structures have been solved and deposited to the RCSB PDB data bank [Ber00], which accounts for almost 90% of all solved protein structures by all available methods. Completing methods are Nuclear Magnetic Resonance (NMR), Electron Microscopy (EM) and Neutron Diffraction. Since X-rays interact weakly with matter, many identical copies of a protein have to be assembled into a crystal which enhances the scattered signal enough to permit structure determination. From protein crystal X-ray diffraction data it is possible to retrieve the electron density map which can be used to build an atomic model of the protein arranged into crystal.

X-ray photons are highly energetic, for that reason they are able to ionize atoms in the specimen. This effect is the primary cause of radiation induced chemical and structural changes that occur in protein crystal exposed to X-rays. This effect introduces the, so called, radiation induced damage to protein crystals which are used for protein X-ray macromolecular crystallography at synchrotron [Hol09] and X-FEL sources [Lom11] and is one of the main limitations for both methods. In fact the time scales on which radiation damage occurs and the effects in those two cases are different but the primary mechanism of damage (ionization of atoms) is preserved. Radiation induced damage in protein crystallography at synchrotrons manifests itself as the global loss of scattering power of protein crystal with increas-

ing exposure time (Bragg spots are fading out with time) and as specific changes to side chains of amino acid residues which then loose definition in the reconstructed electron density map. That means, the exposure to X-rays causes changes in the sample which are not desirable when one wants to decipher the three dimensional structure of a protein crystal. X-ray photons carry energy which is absorbed by the protein crystal of given mass. Energy absorbed by medium of certain mass is called absorbed dose. The dose is expressed in SI units as grays (1 Gy = 1 J/kg). Larger crystals withstand longer exposures at synchrotrons to X-rays before the radiation damage prevents further data collection from a particular protein crystal because, the dose delivered at the same rate as for smaller crystals is distributed over larger mass for larger crystals [Hol10] while the dose limit that cryogenically cooled protein crystal is able to tolerate is the same in both cases. Consequently protein crystals of very small sizes (from hundreds of nanometres to few micrometers) are useless for conventional macromolecular crystallography, even when using micro-focused synchrotron beams [Rie04], because of insufficient scattering power of such crystals that is required to record usable diffraction signal and because of its sensitivity to radiation damage. Recently it has been shown that if one can apply to protein crystals X-ray exposures shorter than the time scale of such destructive radiation damage processes (few femtoseconds), for example by using X-FEL femtosecond pulses, the radiation induced damage to protein crystals can be outrun [Bar11] and the information from undamaged crystal can be recorded before the inset of it. Our other recent study presents research performed using X-ray FEL radiation applied to protein crystals of sub micrometer sizes which shows that such small crystals have the potential to yield sufficient diffraction signal to retrieve its three dimensional structure [Cha11].

Progress in the area of X-ray protein crystallography encounters two main bottlenecks: (i) growing crystals of large enough sizes necessary to obtain diffraction data before the onset of radiation damage [Hol10] and of high enough quality (or high degree of order) for extending the diffraction to high-resolution [Hel88], and (ii) solving the phase problem to obtain a real-space electron density map of the macromolecule. Very often the growth of large crystals is difficult or almost impossible. Only 3% of all cloned protein target genes pass through all steps necessary to obtain a interpretable structure [Che04]. These steps include expression of a protein, purification, crystallization and finally data collection and structure solution [Grä08, Cha08]. Large macromolecular complexes, typically proteins that are functionally embedded in lipid bilayer for example potassium channel [Doy98], are in particular difficult to crystallize into crystals of sufficient size that withstand radiation damage and of high quality to produce high-resolution structures at conventional synchrotron sources. It may take several years of tedious protein expression, purification and crystallization trials in order to find the conditions necessary to grow large crystals

suitable for macromolecular crystallography applications at synchrotrons. On the other hand It is often observed that micrometer crystals occur in crystallization trials as so called crystal showers. Additionally, the fact that nanometre objects escape from attention in typical screening procedures during crystallization plate inspection may indicate that very tiny protein crystals of only few unit cells on sides occur more frequently than large crystals. Because of their relatively low absorption and small size, nano-crystals are well suited for time resolved studies of light-induced or solution-mixing reactions. An experimental tool and well established methodology are necessary to enable structure determination of difficult protein targets.

New generation synchrotron light sources provide a new "disruptive technology" to overcome these bottlenecks. They produce X-ray pulses 10 billion times brighter than any other type of existing X-ray source. The key to that unprecedented brightness is the ultra short pulses. Duration of a typical X-FEL pulse is less than 100 femtoseconds (1 fs = 10^{-15} s). Such intense radiation could produce strong scattering signal even from one pulse exposure of a single biomolecule completely vaporizing it after the pulse. Enhancement of the integrated Bragg spot intensity that appears when multiple copies of the same molecule are arranged in a crystal lattice is proportional to the number of unit cells [Hol09]. This is extremely advantageous because even crystals that have few unit cells along the sides enhance the scattered signal many times than compared to signal scattered from single molecule. It has been demonstrated that diffraction from protein crystals as small as 100 nm can be used to record interpretable molecular structure factors using X-ray FEL [Cha11]. These crystals are more easily grown than large crystals, for example crystals of large membrane proteins that are especially difficult to crystallize. Such large crystals are necessary at synchrotron X-ray diffraction studies. On the other hand, nano- or micro-sized protein crystals overcome the crystallization bottleneck and limitations of radiation damage by literally outrunning it when exposed to ultra intense and short X-ray pulses [Bar11]. Thus protein crystals that appear to be too small or radiation sensitive for studies at conventional synchrotron sources may produce useful and interpretable data at X-ray free-electron laser sources. Additionally, data analysis methods and phasing techniques that have been developed for conventional crystallography during almost hundred years of its history can be transferred to femtosecond protein nanocrystallography at X-FEL's.

1.2 Thesis Context and Overview

At the end of 2009 the first free-electron laser that reach the hard X-ray wavelength regime became operational. The Linac Coherent Light Source (LCLS) located at the Stanford Linear Accelerator Center (SLAC) a National Accelerator Labolatory enabled the realization of experiments at time scales and intensities that have been

long awaited by researchers [Emm10].

This thesis presents research in the area of Serial Femtosecond Crystallography (SFX) of protein nano- and micro-crystals at X-ray Free-Electron Lasers (X-FEL's). Comprehensive verification and assessment of that method in novel protein structure determination is presented, including two examples of new protein structures solved during this work.

Chapter 1 (*Introduction*) gives a brief general discussion about the possibilities that opened up for structural biology with the realization of an X-FEL.

Chapter 2 (*X-ray Free-Electron Lasers*) provides short theoretical introduction to underlying physics behind the ultra bright X-ray radiation produced by fourth generation synchrotron sources.

Chapter 3 (*Introduction to X-ray crystallography*) briefly introduces theoretical basis of macromolecular X-ray crystallography and X-ray powder diffraction techniques.

Chapter 4 (*In-vivo crystallization*) briefly introduces emerging opportunities of *in-vivo* protein crystallization and it's application to novel protein structure determination in combination with SFX.

Chapter 5 (*Serial femtosecond crystallography*) describes methodology of SFX experiments, data analysis methods and two novel protein structures solved during this work.

Chapter 6 (*Verification and assessment of SFX*) describes differences and similarities between SFX and conventional macromolecular crystallography and verifies the quality of obtained protein structures.

Chapter 7 (*Outlook*) Gives an outlook and concludes this thesis.

To provide the reader with a description and short documentation of software used for analysis of SFX data Appendix A and B are included to the manuscript. Appendix C contains reprints of main articles that are listed at the beginning of this manuscript.

Chapter 2

X-ray Free-Electron Lasers

Electromagnetic radiation that is produced by an X-ray Free-Electron Laser (X-FEL) has similar properties to that from a conventional optical laser, such as high power, narrow bandwidth and coherence, however there are also some significant differences. For example in the distributions of spatial and temporal intensity profiles and in the distribution of spectral bandwidth and in spatial and temporal coherence properties. These differences arise due to distinct mechanisms of radiation generation (SASE vs. stimulated emission and optical amplification) [Hua07]. The underlying processes of radiation generation in the SASE X-FEL case is called Self Amplified Spontaneous Emission (SASE) which is enhancing or boosting the power of an electromagnetic (EM) light wave emitted by radially accelerated electrons, whereas in the optical lasers the process of radiation generation is based on stimulated emission of light, population inversion and optical amplification. Gain media in both lasers types are different. X-FEL radiation is generated by unbound (free) electrons that propagate in vacuum along curved paths on which the electrons emit electromagnetic radiation [Mad71], unlike optical lasers where the emission of light requires electrons bound to atoms [Sch58], either in a crystal, liquid dye or a gas.

Free electrons originate from an electron emitter (electron gun) and are accelerated to relativistic velocities in a linear particle accelerator. Electrons from the electron gun are emitted in bunches. Accelerated electrons from the accelerator section enter into the undulator section. Electrons propagate through the undulator section that consists of a periodic array of magnetic dipoles arranged along the undulator length. Magnets arranged this way produce constant spatially periodic transverse magnetic field. While the high energetic electrons are traversing through the long alternating magnetic dipole structure, they change their paths due to the Lorentz force and therefore emit electromagnetic radiation on every bend of their paths [Eli76]. The Lorentz force that acts between electrons propagating through an undulator and the electromagnetic field of radiation emitted by these electrons leads

to a positive feedback and an exponential growth of the radiation intensity emitted by electrons [Bon85]. This amplification of radiation is initiated by an increasingly pronounced longitudinal density modulation of the electron bunch [Ber79]. The initial radiation field can be external, for example from a seed laser [Yu91], or coming form the emission of electromagentic radiation by electrons travelling through the undulator. In the latter case the realized free-electron laser is called SASE X-FEL [Kon80]. Because the electrons in the FEL are not bound to atoms and thus not limited to specific state transitions, the wavelength of the FEL is tunable over a wide range depending on accelerator energy and undulator parameters.

2.1 Synchrotron radiation

When a charged particle (e.g. electron) moves in a magnetic field its path is curved due to the Lorentz force. Such particle is accelerated radially ($\vec{a} \perp \vec{v}$) due to this force. The electromagnetic radiation that is emitted by radially accelerated electrons is called synchrotron radiation. In the classical case, when the speed of an electron is much lower than the speed of light ($v \ll c$), the frequency of emitted radiation is equal to $\omega = eB_0/m_0$, where B_0 is the magnetic field magnitude; e, m_0 are the electron charge and mass. This frequency is determined by the magnitude of the Lorentz force $F = evB_0$. When the speed of an accelerated electron approached the speed of light ($v \approx c$) the Lorentz transformation of the B-field transforms the Lorenz force magnitude to become $F \approx ec\gamma B_0$. The corresponding frequency of emitted radiation in the electron frame is $\omega = \gamma eB_0/m_0$. In the laboratory frame the relativistic Doppler effect shifts this frequency to

$$\omega = \frac{2\gamma^2 eB_0}{m_0}. \tag{2.1}$$

When the velocity of electrons approaches that of the speed of light the factor $2\gamma^2 = 2(1 - v^2/c^2)^{-1}$ becomes very large and boosts the frequency of emitted radiation towards the X-ray regime. Relativity also improves the collimation of the emission. In the reference frame of the electron, the angular distribution of emitted radiation occurs in a broad angular range. The Lorentz transformation of the transverse velocity component to the laboratory frame contains a factor $1/\gamma$ thus making the angular distribution of emitted radiation very narrow. The typical opening angle of the radiation is

$$1/\gamma = \frac{m_0 c^2}{E_0}, \tag{2.2}$$

where m_0, E_0 are electron mass and energy.

2.2 Undulator radiation

The undulator is a long periodic arrangement of short dipole magnets with alternating polarity, as depicted in figure 2.1. Electrons travelling along the undulator encounter periodic changes of the polarity of the magnetic field. The length of one magnetic period of the undulator is defined as λ_u. The deflection of the electrons by the Lorentz force from it's forward direction of movement is comparable to the opening angle of the synchrotron radiation cone. Thus the radiation generated by the electrons while travelling along the individual magnetic periods overlaps. This interference effect is reflected in the formula for the wavelength λ_{ph} of the first harmonic of the spontaneous, on-axis undulator radiation emission

$$\lambda_{ph} = \frac{\lambda_u}{2\gamma^2}(1 + \frac{K^2}{2}), \tag{2.3}$$

where

$$K = \frac{eB_0\lambda_u}{2\pi m_0 c} \tag{2.4}$$

is the *undulator parameter*. Power of the emitted spontaneous radiation is proportional to the number of electrons ($P \propto N_e$) because there is no phase correlation between the fields emitted from individual electrons and proportional to the number of undulator periods squared ($P \propto N_u^2$).

Synchrotron based light sources produce X-ray radiation that already has many characteristics of the laser like radiation such as collimation and the high degree of coherence. Still missing is the mechanism of optical amplification that would boost the emission intensity and brightness.

2.3 The principle of SASE

To obtain exponential amplification of the radiation power emitted by highly energetic electrons while they are travelling through an undulator, some special characteristics of the electrons have to be guaranteed. In order to generate pulsed X-ray FEL beam, electrons must come out from the accelerator section in bunches, meaning that electrons have to be structured into compressed, short pulses that posses high electron density. Typically the number of electrons in one bunch is $N_e \geq 10^9$. The energy spread of the electrons must be low but the charge density must be extremely high in order to create the SASE effect. Also the emittance must be very low, less than 10^{-6} mrad.

Electrons from the bunch emit electromagnetic waves when they traverse through periodic magnetic field in an undulator. These initial waves interact with electrons in the bunch and are causing the formation of electron microbunches (very narrow

sheets of electrons perpendicular to the electron banch direction of motion). If this process continues along the undulator length, an unstructured electron bunch is further transformed into a mico-sliced electron bunch with the distance between each microbunch equal to λ_{ph}. Microbunching and the interaction of emitted electromagentic waves with the travelling electrons are responsible for the optical amplification of the radiation emitted by electrons. See figure 2.1.

The detailed explanation of the microbunching mechanism is the following. Before entering the periodic magnetic field of an undulator the electron bunch does not have microbunches. When electrons begin to emit waves, the transverse magnetic field of an emitted wave (B_w) and the transverse velocity (v_T) of an electron traversing that undulator cause longitudinal Lorentz force that modifies the longitudinal velocity of that electron. When the magnetic field (B_w) and the transverse electron velocity (v_T) have the right phase difference the electron is pushed into a region where the magnetic field (B_w) is zero. This process occurs for all electrons in the bunch and results in creation of microbunches within the electron bunch with the separation length of λ_{ph}. Microbunching occurs in an electron bunch with randomly distributed electrons that just left the accelerator module and entered the undulator module. The more intense the electromagnetic field of emitted radiation becomes, the more pronounced the longitudinal density modulation of electrons in the electron bunch and vice versa. However, to sustain this process electrons have to be a little bit slower than the emitted waves ($v < c$) otherwise after one-half magnet period the transverse velocity and the Lorentz force would be reversed, acting against microbunching. With complete micro-bunching, all electrons radiate almost in phase. This leads to emitted radiation power proportional to the number of electrons squared ($P \propto N_e^2$) and thus amplification of many orders of magnitude with respect to the spontaneous emission of the undulator. The radiation power $P(z)$ grows exponentially with the distance z along the undulator

$$P(z) \propto P_{in} exp(z/L_g), \tag{2.5}$$

where L_g is the field gain length, P_{in} the input power. To estimate of the input power P_{in} one can use the spontaneous radiation power of the first gain length inside a coherence angle and within the FEL bandwidth. The exponential growth takes place until the electron beam is completely bunched (saturation point).

The radiation from a SASE X-FEL is almost fully transversely coherent [Sal10]. As explained above, the amplification process starts from the shot noise in the electron beam. Additionally, any random fluctuation in the electron beam will correspond to an intensity modulation of the beam current at all frequencies simultaneously. The fluctuations of current density in the electron beam are uncorrelated not only in time but also in space. Thus, a large number of transverse radiation modes are excited when the electron bunch enters the undulator. These radiation

modes have different gains. As undulator length progresses, the high gain modes start to predominate more and more. For enough long undulator, the emission will emerge in a high degree of transverse coherence. The numerical analysis of transverse coherence properties of LCLS is resented in [Din10]

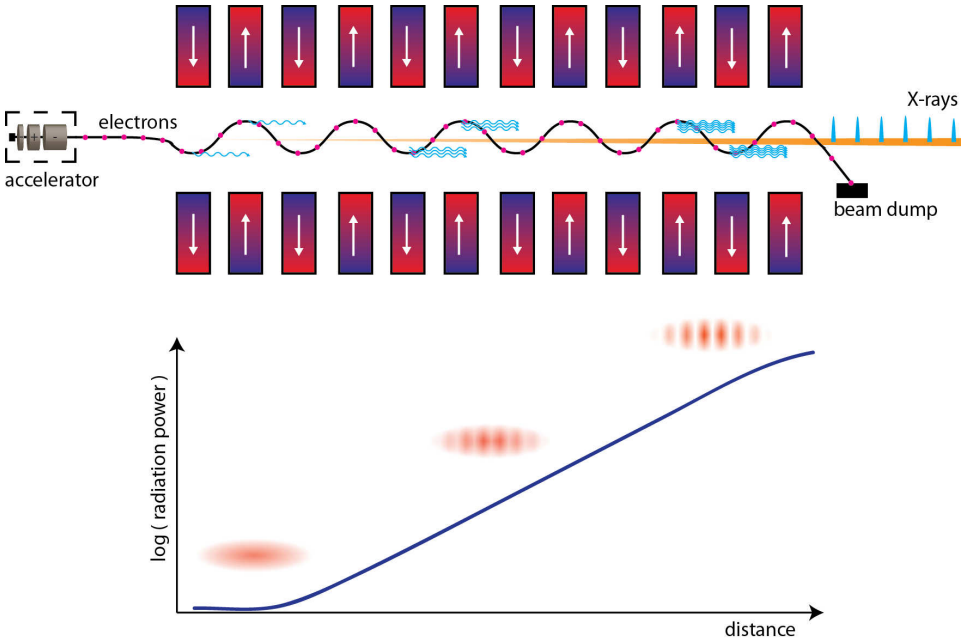


Figure 2.1: *Principle of a SASE X-ray free-electron laser. Electron bunches from an electron gun are further compressed and accelerated to relativistic velocity in a linear accelerator. After that, electrons enter a long undulator where they emit electromagnetic radiation on every bend of their path. As the electron bunch travels along the undulator, waves interact with electrons from the bunch causing microbunching. Microbunching motivates electrons to emit radiation in phase therefore it is responsible for the exponential growth of the radiation power emitted by electrons along the undulator length. At the end of the undulator electrons are discarded into a beam dump. Intense and short X-ray pulses are then guided to the experimental station.*

2.4 Other types of X-ray FEL sources

The fact that the SASE X-FELs have been successfully realized and that they are extensively used by worldwide users in a range of applications proves the significance of this great achievement in the photon science. In addition to the SASE X-FEL there exist other schemes that allow creation of ultrashort and ultrabright X-ray pulses. These schemes make use of for example (i) the oscillator cavity made with

crystal mirrors, (ii) external or self 'seeding' stimulation to engage the FEL radiation generation process or (iii) even the wake field acceleration in plasma in which the electric field of particle or laser beam sets up waves in a plasma, which trap and accelerate charged particles.

2.4.1 X-FEL oscillator cavity with crystal mirrors

The key components for an X-ray FEL oscillator (X-FELO) are: a continuous sequence of ultra low emittance electron bunches from a multi-GeV energy-recovery linac (ERL) and a low-loss optical cavity constructed from high-reflectivity crystals. The electron bunches from an ERL are not suitable for high-gain X-FELs due to its relatively small charge density. However, realization of an X-ray FEL is possible in an oscillator configuration which takes the advantage of repeated low-gain amplifications [Kim08, Lin11].

The most advantageous property of an X-FELO is its fully temporal coherence in contrast to a high-gain SASE FEL [Hua06]. Also compared to a SASE high-gain FEL, the pulse intensity of a proposed X-FELO is lower by approximately three orders of magnitude, but its spectral bandwidth is narrower by more than thousand times. The pulse repetition rate is at least 1 MHz, which is higher by at least two orders of magnitude than that of the high-gain, high-repetition-rate FEL using a super conducting linear accelerator. With these characteristics, X-FELOs for X-rays in the range from 5 to 20 keV may open up new scientific opportunities in various research fields, such as inelastic scattering, nuclear resonance scattering or X-ray imaging.

2.4.2 X-FEL from laser wake field acceleration

The theory for laser wake field acceleration has been introduced in [Taj79]. In brief, the principle of wake field acceleration of charged particles can be described as follows. A wave packet of electromagnetic radiation (photons), for example coming from a pulsed laser, injected into plasma excites an electrostatic wake behind the photons. In other words it introduces extreme charge-density modulations in plasma. These modulations may set up field gradients of more than 100 GV/m and propagate like waves through the plasma medium with velocities close to the speed of light. Thus, they are suited perfectly for the acceleration of a co-propagating, charged particle beam.

The advantage of plasma acceleration is that its acceleration field can be much stronger than that of conventional radio-frequency (RF) accelerators allowing the construction of much shorter linear particle accelerators than the RF linear accelerator, for example its realization for soft X-rays can be found in [Fuc09].

2.4.3 Seeded X-FELs

The poor temporal coherence of a SASE X-FEL pulse can be improved by some form of 'seeding' [Hua07]. Seeding generally involves two scenarios: self-seeding and external seeding. Since a proper coherent seed does not exist at X-ray wavelengths, a high-gain harmonic generation (HGHG) FEL relies on a coherent seed at subharmonic wavelengths of optical laser. In this scheme [Yu91], a small energy modulation is imposed on the electron bunch by the interaction with the seed laser. The energy modulation is converted to a coherent spatial density modulation as the electron beam traverses further. An optical pulsed laser tuned to a higher harmonic of the seed frequency, causes the microbunched electron beam to emit coherent radiation at that harmonic frequency. This shorter-wavelength radiation may then be used as the coherent seed to the next stage HGHG. Realization of an seeded in the extreme ultraviolet regime has been demonstrated in [All12].

In addition to better temporal coherence than in a SASE X-FEL, a seeded FEL displays narrower spectral bandwidth of generated radiation. A self seeding scheme [Fel97] consists of two undulators (of the same undulator period and strength) and an X-ray monochromator located between them. The first undulator operates in the exponential gain regime of a SASE FEL. After the exit of the first undulator, the electron beam is guided through a dispersive bypass that smears out the microbunching induced in the first undulator. The SASE output enters the monochromator, which selects a narrow band of radiation. At the entrance of the second undulator the monochromatic x-ray beam is combined with the electron beam and is amplified up to the saturation level. Realization of a self-seeded hard X-ray free-electron laser has been recently demonstrated in [Ama12].

2.5 Properties of the LCLS electron and photon beams

In this section parameters of the LCLS electron and X-ray pulses measured during the serial femtosecond crystallography (SFX) experiment with *in-vivo* grown protein crystals are presented and their importance for serial femtosecond cystallography is discussed.

SFX experiments on *in-vivo* grown crystals of TbCatB and TbIMPDH (see Chapter 4, *In-vivo protein crystallization*) were carried out at the Coherent X-ray Imaging (CXI) beam line [Bou10] at the Linac Coherent Light Source (LCLS) [Emm10], operated by the SLAC National Accelerator Laboratory.

LCLS is a hard X-ray SASE FEL that produce highly intense pulses that are typically from 10 to 200 fs in duration with X-ray pulse energy of up to 3 mJ, which gives approximately 10^{12} X-ray photons per single pulse. Transmission achieved for the CXI beam line from the source to the sample during the experiments was

approximately 20%. The X-ray pulses produced by LCLS are almost fully spatially coherent and are quasi-monochromatic with spectral bandwidth of about 0.2%. The wavelength may jitter by 0.3% from shot to shot. This instability effects in the wavelength and energy of the X-ray pulses are reflected by the fluctuations of the electron bunch energy and the SASE process. The intensity fluctuations are expected to follow a gamma distribution [Hog98]. That has been captured in the histograms in the figure 2.2 and in the table 2.1 related to those histograms, which present shot-by-shot measured electron bunch parameters and X-ray pulse energies from 2284 measurements.

The electron bunch length and its temporal profile were measured on a per-shot basis using RF deflection cavities in conjunction with an electron beam energy spectrometer. That information can be then used to simulate the X-ray pulse temporal profile [Din11]. The electron beam position, charge and energy were measured using other beam profile monitors (BPMs) such as wire scanners, optical transition radiation monitors (OTR), transverse RF deflection cavities or fluorescent screens [Fri09]. X-ray pulse energies were measured using a gas monitor detector (XGMD) similar to the one presented in [Ric03]. Measurements of the electron and X-ray beams parameters were performed during our experiments with *in-vivo* crystals on a per-shot basis by using the LCLS instrumentation implemented at the LCLS facility.

X-ray pulse energies, in [keV], were calculated using transformed undulator equation 2.3. In the equation 2.6 photon pulse energy is expressed in terms of electron bunch energy E_{el}, where K is the undulator parameter and λ_u is the undulator period. E_{el} is measured before the electron bunch enters to the undulator section. In that equation a correction to the electron beam energy E_{el} for the wakefield energy loss, which depends on the electron beam peak current ($peakCurrent[A]$) and a correction for the spontaneous energy loss due to emission in the undulator segments is applied [lcl] – sections 4.5 and 4.6 of that reference. The emission of spontaneous radiation by electrons in the undulator decreases the average electron energy and increases the relative energy spread [Sal96]. Additionally, the sources of wakefields within the undulator vacuum chamber are: resistive walls, geometry of the design, and the surface roughness. The wakefields in the undulator vacuum pipe can also have an important effect on the lasing process, reducing the output power and changing the temporal structure of the X-ray pulse [lcl]. Exact values for the losses in electron bunch energy and for other parameters used for calculation of X-ray photon energy were obtained from private communication with LCLS machine physicist James Welch [Cas] and are presented in equations 2.8 – 2.12.

The fact that the photon energy of every X-ray pulse varies from shot to shot due to the stochastic nature of SASE process and the uncertainty in estimation of electron bunch energy losses introduces difficulties in data evaluation for serial

femtosecond crystallography. The fluctuations of the photon beam energy is one of the sources of errors in the Monte Carlo integration of every structure factor (see Chapter 3 and section 5.2.2 in Chapter 5). The number of photons per pulse and the very short pulse duration also play a crucial role in the experiment with the smallest protein crystals, because the scattered signal depends on the number of incident photons and the size of the crystal. This is the main limitation of protein crystallography at synchrotrons – small protein crystals simply do not give detectable scattering signal before the onset of radiation damage and researchers are struggling to obtain larger crystals in their tedious protein crystallization trials. This limitation may be overcomed by using serial femtosecond crystallography method (SFX). The spectral bandwidth of the single pulse also should not be forgotten when designing a SFX experiment. This parameter is advantageous in terms of the faster convergence of the Monte Carlo method due to the fact that larger number of reflections will be in the Bragg condition and reflections will be less "partial" when recorded on a detector. See next chapters for reference.

$$E_{ph}[keV] = 0.950\frac{E_{el}^2[GeV]}{(1+K^2/2)\lambda_u[cm]} \tag{2.6}$$

$$E_{el}[GeV] = E_{measured} - E_{wakefield\ loss} - E_{spontaneous\ loss} \tag{2.7}$$

$$LTUWakeLoss = 0.0016293 * peakCurrent[A] \tag{2.8}$$

$$SRLossPerSegment = 0.63 * E_{measured} \tag{2.9}$$

$$WakeLossPerSegment = 0.0003 * peakCurrent[A] \tag{2.10}$$

$$\begin{aligned} EnergyLossPerSegment = SRLossPerSegment+ \\ WakeLossPerSegment = 0.63 * E_{measured} + 0.0003 * peakCurrent[A]. \end{aligned} \tag{2.11}$$

$$\begin{aligned} E_{el}[GeV] = E_{measured} - 0.001 * (0.0016293 * peakCurrent[A])- \\ -0.0005 * (0.63 * E_{measured} + 0.0003 * peakCurrent). \end{aligned} \tag{2.12}$$

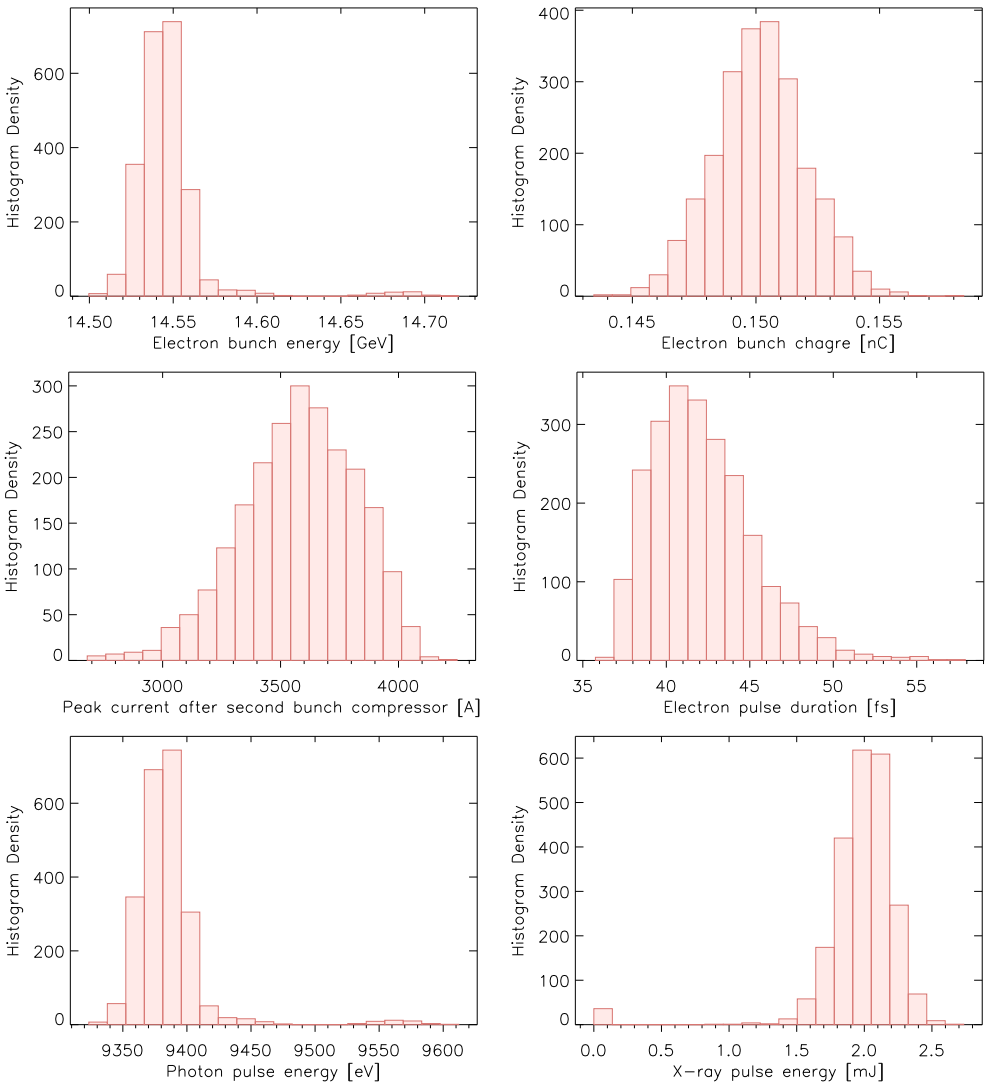


Figure 2.2: *Histograms of photon and electron pulse parameters for ≈ 40 fs 9.4 keV X-ray pulses. X-ray pulse duration has been estimated from the measurement of the electron bunch length. Data presented in histograms have been obtained from 2284 single-shot measurements. Energy of the electron bunch that enters into the undulator module. Electron bunch charge corresponds to the number of electrons in a single bunch. Peak current of a bunch after second bunch compressor in the accelerator module. Photon pulse energy calculated by the equation that is explained in the text above. X-ray pulse energy (mJ) measured by a gas detector (XGMD) upstream from the experimental station, beam line transmission was only about 20%. Mean and standard deviation values of electron and photon beam parameters presented in the table 2.1.*

Table 2.1: *Mean and standard deviation values of electron and photon beam parameters presented in the histograms in the figure 2.2.*

Parameter	Value	
Electron bunch energy [MeV]	14546.0	± 22.0
Electron bunch charge [nC]	0.150	± 0.002
Peak current [A]	3577.0	± 246.0
Elecron pulse duration [fs]	42.2	± 3.0
Photon pulse energy [eV]	9385.0	± 30.0
X-ray pulse energy [mJ]	1.95	± 0.31

Chapter 3

Introduction to X-ray crystallography

A crystal is a solid material whose atoms, molecules, or ions are arranged in an ordered manner extending in all three dimensions. In addition to their microscopic structure, macroscopic crystals can be usually identified by their geometrical shape, consisting of flat faces with specific, characteristic orientations. The scientific study of crystals and crystal formation is known as crystallography.

X-ray crystallography is a method of determining the atomic structure of a crystal, in which the crystalline atoms cause a beam of X-rays to diffract into many specific directions. By measuring the angles and intensities of these diffracted beams, a crystallographer can produce a three-dimensional picture of the density of electrons within the crystal. From this electron density, the mean positions of the atoms in the crystal can be determined, as well as their chemical bonds, their disorder and various other information. Structural biology extensively use X-ray crystallography for their studies on structures and structure to function relationship of macromolecules such as proteins or viruses.

The smallest volume which, when re-produced by close packing in three dimensions, gives the whole crystal is called the unit cell. Using common nomenclature the lengths of unit cell faces are named a, b, c and the angles between them are named α, β, γ. Crystals can be divided into classes depending on the arrangement of the unit cell axes and angles, see the table 3.1. These seven crystallographic systems can be then further divided into 32 crystallographic point groups depending on the different degrees of symmetry of crystals which belong to the same system. Additionally, symmetry elements such as centre of symmetry, mirror plane, glide plane, rotation axis, screw axis, can be combined in groups and it can be shown that 230 distinctive arrangements are possible. Each of these arrangements is called a space group and they are all listed and described in volume A of the *International Tables for Crystallography*. In three dimensions a plane may always be found, parallel to a crystal face, which makes intercepts a/h, b/k and c/l on the unit-cell edges, where

Table 3.1: *The seven crystal systems with their names and relationships between dimensions of their unit cells.*

System	Dimensions relationships
Triclinic	None
Monoclinic	$a \neq b \neq c; \quad \beta \neq \alpha = \gamma = 90^\circ$
Orthorhombic	$a \neq b \neq c; \quad \alpha = \beta = \gamma = 90^\circ$
Tetragonal	$a = b \neq c; \quad \alpha = \beta = \gamma = 90^\circ$
Trigonal	$a = b = c; \quad \alpha = \beta = \gamma \neq 90^\circ$
Hexagonal	$a = b \neq c; \quad \alpha = \beta = 90^\circ, \gamma = 120^\circ$
Cubic	$a = b = c; \quad \alpha = \beta = \gamma = 90^\circ$

h,k,l are integers and are typically called Miller indices.

3.1 The scattering of X-rays

The scattering is an interaction between X-rays as electromagnetic waves and the electrons, typically those in atoms. In X-ray diffraction the electrons in an atom can be regarded, to good approximation, as free electrons. The wave scattered by the crystal may be described as a summation of waves each scattered by one electron in the crystal. The scattering is dependent on the number of electrons and their positions in the electron cloud of an atom. The electron density at position r is denoted by $\rho(r)$. The atomic scattering factor is:

$$f = \int_r \rho(r)\ exp[2\pi i r \cdot S]\, dx, \tag{3.1}$$

where S is the scattering vector, a difference between the incident wave vector s_0 with length equal to $1/\lambda$ and the scattered wave vector s ($S = s - s_0$). The magnitude of the scattering vector is given by:

$$|S| = \frac{2\sin\theta}{\lambda}, \tag{3.2}$$

where θ is the reflecting angle between the incident wave vector and the reflecting plane, this plane is perpendicular to the direction of scattering vector S.

The unit cell has n atoms at positions r_j ($j = 1, 2, 3, ..., n$) with respect to the origin of the unit cell. The total scattering factor from an unit cell is

$$F(S) = \sum_{j=1}^{n} f_j\ exp[2\pi i r_j \cdot S]. \tag{3.3}$$

$F(S)$ is called the structure factor.

Suppose the crystal has translation vectors a, b, c and is built by a large number of unit cells. There are n_1 unit cells in the direction of the vector a, n_2 unit cells in the direction of the vector b and n_3 unit cells in the direction of the vector c. To obtain the scattering by the crystal we must add scattering from all unit cells with respect to the origin. For a given crystal consisting of $n_1 \times n_2 \times n_3$ unit cells with origins at the positions $t \cdot a + u \cdot b + v \cdot c$ with respect to the crystal origin, where t, u, v are whole numbers describing the n-th unit cell, the total scattered wave is

$$\begin{aligned} K(S) = F(S) \times \sum_{t=0}^{n_1} exp[2\pi ita \cdot S] \times \\ \times \sum_{t=0}^{n_2} exp[2\pi iub \cdot S] \times \sum_{t=0}^{n_3} exp[2\pi ivc \cdot S]. \end{aligned} \tag{3.4}$$

The summation $\sum_{t=0}^{n_1} exp[2\pi ita \cdot S]$ and the other two summations alre almost always equal to zero, unless $a \cdot S$, $b \cdot S$, $c \cdot S$ are integers. These are the Laue conditions for X-ray scattering from a crystal:

$$\begin{aligned} a \cdot S &= h \\ b \cdot S &= k \\ c \cdot S &= l, \end{aligned} \tag{3.5}$$

where the h, k, l are whole numbers. The amplitude of the wave scattered by a crystal is proportional to the structure factor $F(S)$ and the number of unit cells in the crystal.

3.1.1 The Bragg law

The direction of the scattering vector S is perpendicular to the reflecting plane of the crystal. The reflecting plane can be considered as the lattice plane (h, k, l). The incident s_0 and the scattered wave s make an angle θ with this plane. By transforming the first Laue condition we obtain

$$\frac{a}{h} \cdot S = 1 \tag{3.6}$$

and similar for other two. The scalar product of two vectors a/h and S is equal to 1. In other words, the vector a/h projected on the vector S and multiplied by $|S|$ gives unity. This projection has a length of $1/|S|$ and is also equal to d, which is the distance between the reflecting planes. From $|S| = 1/d$ and $|S| = 2(\sin\theta/\lambda)$ the Bragg law can be derived

$$\frac{2d \sin\theta}{\lambda} = 1. \tag{3.7}$$

3.1.2 The reciprocal lattice and Ewald construction

For the lattice planes (100), (010), and (001) the scattering vectors $S(100)$, $S(010)$, $S(001)$ are perpendicular to those lattice planes if those planes are in the reflecting (Laue) conditions and have the length of $1/d(100), 1/d(010), 1/d(001)$. Lets name those vectors as a^*, b^* and c^*.

Because a^* is perpendicular to the reflecting plane (100) it is perpendicular to the b- and c-axis, thus $a^* \cdot b = a^* \cdot c = 0$, but $a \cdot a^* = a \cdot S(100) = h = 1$. The same for other two vectors. The end point of the vectors $S(hkl)$, where h, k, l are whole numbers, are located in the lattice points of a lattice constructed with the unit vectors a^*, b^* and c^*.

Therefore $S = h \cdot a^* + k \cdot b^* + l \cdot c^*$. The crystal lattice based on vectors a, b, c is called the direct lattice and the latice based on vectors a^*, b^*, c^* is called the reciprocal lattice. It can be shown that if the volume of the unit cell in direct space is V the volume of the unit cell in the reciprocal space V^* is equal to $1/V$.

The reciprocal lattice is a very useful concept used for constructing the directions of diffraction from a crystal. Scattering occurs only when the scattering vectors $S(hkl)$ have their end points (reciprocal lattice points for planes h, k, l) on the sphere of the radius $1/\lambda$ with respect to the origin. This is called the Ewald construction, see the figure 3.1. Crystal lattice planes that do not lie on the Ewald sphere can be brought into the reflection condition by rotating the reciprocal lattice.

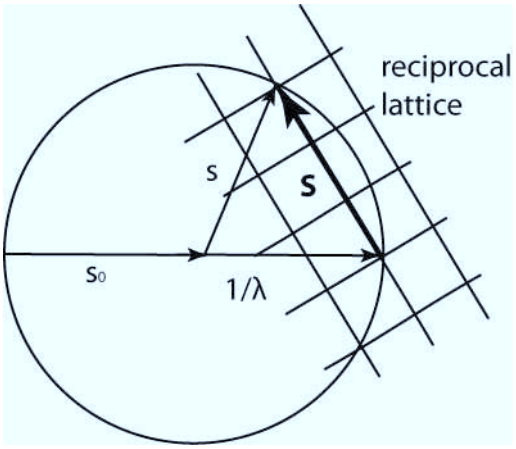


Figure 3.1: *The Ewald construction. A part of a reciprocal lattice is represented as a set of perpendicular lines intercepting at reciprocal lattice points, s_0 indicates the direction of the incident X-ray beam, s indicates the direction of the scattered beam. The sphere has radius $1/\lambda$. The scattering vector S connects two lattice points that intercept with the Ewald sphere.*

3.1.3 The temperature factor

All atoms in a crystal lattice vibrate around theirs equilibrium positions. The X-rays do not encounter atoms in identical positions in successive unit cells. This fact reduces the scattered intensity, especially at high scattering angles. Therefore the atomic scattering factor must be multiplied by a temperature dependent factor.

The vibration component in the plane perpendicular to the reflecting plane (h, k, l) has an effect on the scattered intensity of reflection (h, k, l). In the simplest case, where vibrations are the same in all directions the vibration is called *isotropic*. The component perpendicular to the reflecting plane, thus along to the scattering vector S, is equal for each reflection (h, k, l) and the correction factor for the atomic scattering factor is

$$T(iso) = exp\left[-B\frac{\sin^2\theta}{\lambda^2}\right] = exp\left[-\frac{B}{4}\left(\frac{2\sin\theta}{\lambda}\right)^2\right] = exp\left[-\frac{B}{4}\left(\frac{1}{d}\right)^2\right]. \quad (3.8)$$

The thermal parameter B is related to the mean square displacement $\overline{u}^2$ of the atomic vibration:

$$B = 8\pi^2 \times \overline{u}^2. \quad (3.9)$$

For the *anisotropic* vibration the temperature factor is much more complicated. Usually determination of the *isotropic* temperature factors is sufficient in X-ray macromolecular crystallography.

3.1.4 Calculation of the electron density

The intensity of the diffracted X-ray beam is proportional to the square of the structure factor $F(S)$ or $F(hkl)$. The structure factor is a function that describes the electron density in the unit cell of a crystal, see equation 3.3. Instead of summation we can integrate over all atoms in the unit cell

$$F(S) = \int_{cell} \rho(r)\ exp[2\pi i r \cdot S]\, dv, \quad (3.10)$$

where $\rho(r)$ is the electron density at position r in the unit cell and V is the unit cell volume then

$$dv = V\dot{d}xdydz, \quad (3.11)$$

where x, y, z are the fractional coordinates that describe position in the unit cell. Using direct representation of the position vector r and Laue conditions we obtain

$$\begin{aligned} r \cdot S &= (a \cdot x + b \cdot y + c \cdot z) \cdot S = a \cdot S \cdot x + b \cdot S \cdot y + c \cdot S \cdot z \\ &= hx + ky + lz. \end{aligned} \quad (3.12)$$

Therefore $F(S)$ can be rewritten as $F(hkl)$.

$$F(hkl) = V \int_x \int_y \int_z \rho(xyz)\ exp[2\pi i(hx + ky + lz)]\, dx\, dy\, dz \tag{3.13}$$

$F(hkl)$ is the Fourier transform of $\rho(xyz)$ and the reverse is also true: $\rho(xyz)$ is the Fourier transform of $F(hkl)$. The integrals have been replaced by summations because Laue conditions tell us that diffraction occurs only in discrete directions.

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l F(hkl)\ exp[-2\pi i(hx + ky + lz)]. \tag{3.14}$$

Because $F = |F|\ exp(i\alpha)$ we can also write

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |F(hkl)|\ exp[-2\pi i(hx + ky + lz) + i\alpha(hkl)]. \tag{3.15}$$

Despite the fact that the $|F(hkl)|$ can be derived from intensities measured experimentally, the phase $\alpha(hkl)$ cannot be obtained directly from the diffraction patterns. This is called the *phase problem* but fortunately there exist few ways to obtain the phase of scattered waves and retrieve the electron density from diffraction images.

3.1.5 Molecular replacement

Molecular replacement (MR) among many other methods enables the solution of the phase problem by providing initial estimates of the phases of the new structure from a previously known structure. Molecular Replacement also provides an initial starting model for refinement. Other methods for obtaining the unknown phases are called experimental methods or direct methods.

The MR is now often used to solve new structures that are homologous to previously determined structures. MR is currently used to solve approximately 70% of all structures deposited to the protein data bank [Ber00].

In principle MR is very simple, we have a known model that is approximately similar to our unknown model. We know that because those two proteins display similarity in their amino acid sequence or function or belong to the same class of proteins which present very similar folding of their polypeptide chain. We then try all possible orientations and positions of the known model in the unknown crystal and find where the calculated set of structure factors matches the set of structure factors obtained from experiment with our unknown crystal structure. The known model at this point is the best fit to the target structure. The phases for the reflections of the unknown crystal are then borrowed from the phases calculated

from the known model as if it was the model that had crystallized in the unknown crystal and an initial electron density map is calculated with these borrowed phases and the experimentally observed amplitudes. The crystallographer therefore relies on the measured amplitudes to supply the information for rebuilding of the model so that it more closely resembles the target structure. At this point, the MR problem becomes a crystallographic refinement problem [Eva08].

The work on molecular replacement was pioneered by M. Rossmann and D. Blow [Ros62]. Placement of the molecule in the target unit cell requires two steps: rotation and translation. Rotation finds the spatial orientation of the known and unknown molecules whereas translation superimposes the correctly oriented molecule onto the other molecule. There are two ways to calculate the rotation and translation functions. The first one uses the Patterson function which then scores the overlap between observed and model Pattersons. The second way is to use the probability approach, the maximum-likelihood method to find the orientation and translation functions.

The Patterson function is the Fourier transform of the squared structure amplitude $|F|^2$ with phases set to zero. Patterson functions are useful because they can be calculated directly from the observed data. They can also be calculated from the model by ignoring the phase component of the calculated structure factor. If the Patterson function from the observed data is superimposed on a correctly rotated version of Patterson function from the model a maximum overlap between two Patterson function will occur. In this step a translation is not incorporated. However, for the final solution of MR the translation required to overlap one molecule onto the other in real space must be determined, after it has been oriented in the correct way with the rotation function. To accomplish this task the molecule is moved through the asymmetric unit and structure factors are calculated (F(calc)) and compared to the observed set of structure factors by calculating an R factor:

$$R_{MR} = \frac{\sum_{hkl} ||F(obs)| - k|F(calc)||}{\sum_{hkl} |F(obs)|}, \tag{3.16}$$

where k is the scale factor for the intensities. The lowest R factor will indicate the correct translation. Or the correlation coefficient

$$C = \frac{\sum_{hkl}(|F(obs)|^2 - \overline{|F(obs)|^2}) \times (|F(calc)|^2 - \overline{|F(calc)|^2})}{\sqrt{\sum_{hkl}(|F(obs)|^2 - \overline{|F(obs)|^2})^2 \sum_{hkl}(|F(calc)|^2 - \overline{|F(calc)|^2})^2}}. \tag{3.17}$$

Correlation coefficient is usually more useful than R factor because is gives better MR solution even if the model has errors typically encountered during molecular replacement.

3.1.6 Wilson plots

It is possible to obtain a good estimate of the temperature factor and of the factor required for putting the intensities $I(S)$ on absolute scale. On absolute scale the intensity is given by

$$F(abs,S) = F(S) \cdot F^*(S) = |F(S)|^2 = \sum_i \sum_j f_i f_j \; exp[2\pi i(r_i - r_j) \cdot S] \tag{3.18}$$

For a set of evenly distributed reflections the values of the angles $[2\pi(r_i = r_j)]$ vary over the range $0 - 2\pi$ for $i \neq j$. The average intensity for $i \neq j$ will be zero, only the terms with $i = j$ remain.

$$\overline{|F(S)|^2} = \overline{I(abs,S)} = \sum_i f_i^2 \tag{3.19}$$

$$f_i^2 = \; exp\left[-2B\frac{\sin^2\theta}{\lambda^2}\right] \times (f_i^0)^2, \tag{3.20}$$

where $(f_i^0)^2$ is the scattering factor of an atom i at rest.

In order to compare the calculated values $\overline{I(abs,S)}$ with the experimental data $\overline{I(S)}$ requires a scale factor C.

$$\overline{I(S)} = C \times \overline{I(abs,S)} = C \; exp\left[-2B\frac{\sin^2\theta}{\lambda^2}\right] \sum_i (f_i^0)^2 \tag{3.21}$$

To determine B and C the above equation can be written

$$ln\frac{\overline{I(S)}}{\sum_i (f_i^0)^2} = lnC - 2B\frac{\sin^2\theta}{\lambda^2} \tag{3.22}$$

When plotted on a logarithmic scale the result should be a straight line for high scattering angles. This is the so called Wilson plot, from which both the temperature factor and the absolute scale factor can be derived.

3.1.7 Crystallographic data and model refinement metrics

The completeness of the data is the first parameter that needs to be taken for consideration when solving and refining a protein structure from X-ray crystallographic data. The completeness is defined as the percentage of the number of measured reflections or structure factors ($F(hkl)$) to the number of all possible reflections that can be produced by a protein crystal of a given unit cell dimensions an up to certain scattering angle (resolution). Usually the completeness should be very high (100% – 98%) to obtain a good quality data set. Other important data quality metric

is the $I/\sigma(I)$ value. I is the average intensity of a group of reflections divided by the average standard deviation $\sigma(I)$ of the same group of reflections. Usually it is reported per resolution shell. Typically, in conventional protein crystallography at synchrotrons the value of $I/\sigma(I) \geq 3$ indicates the resolution limit of the diffraction data.

In serial femtosecond crystallography that uses the Monte Carlo method for obtaining structure factors from measured data, see section 5.2.2 from the Chapter 5 (*Serial femtosecond crystallography*) another parameter connected with completeness, called *redundancy*, is used. Redundancy measures how many times a structure factor has been measured in the Monte Carlo integration.

One of the commonly used metrics of the agreement between the measured data and the protein model during model refinement is R-factor and R free [Brü92]. The refinement of the model is a procedure that a protein crystallographer applies to the protein structure model in order to correct for any errors in the model such as wrong amino acid conformation. The model refinement usually leads to improvements of R factors.

The R-factor is defined as follows

$$R = \frac{\sum_{hkl} ||F_{hkl}^{obs}| - |F_{hkl}^{calc}||}{\sum_{hkl} |F_{hkl}^{obs}|} \times 100\%, \tag{3.23}$$

where the sums run over all h,k,l indices and F's are the values of structure factors of observed and calculated sets.

3.1.8 Effect of crystal shape on diffracted intensities

For the smallest nanocrystals, the crystal shape effects will also influence the recorded Bragg intensities.

For plane-polarized monochromatic incident radiation with wave vector k_i ($|k_i| = 1/\lambda$) and negligible beam divergence, the diffracted photon flux I (counts/pulse) at $\Delta k = k_i - k_o$ produced by the nth parallelepiped crystallite, consisting of $N = N1 \times N2 \times N3$ unit cells, is given in the kinematic theory as [Kir10]

$$\begin{aligned} I_n(\Delta k, k_o, \alpha, \beta, \gamma, N_i) = J_o |F(\Delta k)|^2 r_e^2 P(k_o) \frac{\sin^2(N_1\Psi_1)}{\sin^2(\Psi_1)} \times \\ \frac{\sin^2(N_2\Psi_2)}{\sin^2(\Psi_2)} \frac{\sin^2(N_3\Psi_3)}{\sin^2(\Psi_3)} \Delta\Omega = c|F(\Delta k)|^2 S(\Psi) \end{aligned} \tag{3.24}$$

where $F(\Delta k)$ is the structure factor of the unit cell. J_o is the incident photon flux

density (counts/pulse/area) and $\Delta\Omega$ is the solid angle subtended by a detector pixel.

$$\begin{aligned} \Psi_1 &= 2\pi a \sin(\theta)\cos(\alpha)/\lambda, \\ \Psi_2 &= 2\pi b \sin(\theta)\cos(\beta)/\lambda, \\ \Psi_3 &= 2\pi c \sin(\theta)\cos(\gamma)/\lambda, \end{aligned} \tag{3.25}$$

where θ is half of the scattering angle, and α, β and γ define the crystal orientation as the angles which the scattering vector Δk makes with the directions of the real-space unit cell vectors a, b and c. Δk is defined by the position of the detector pixel and X-ray wavelength, and defines a point in reciprocal space where the Ewald sphere intersects the shape transform. r_e is the classical radius of the electron, equal to 2.82×10^{-5} Å and $S(\Psi)$ is a function describing the Fourier Transform of the external shape of the nanocrystal (the "shape transform"). $P(k_o)$ is the polarization factor.

The 3D integration over the triple product in equation 3.24 is proportional to $N_1 \times N_2 \times N_3$ and the volume of the unit cell, so the measured diffracted counts are therefore proportional to the number of electrons in the crystal, thus for a nanocrystal of just 10 molecules on a side, one has a thousand times more signal than from a single molecule, due to the coherent amplification of Bragg diffraction.

3.1.9 X-ray powder diffraction

In X-ray powder diffraction the diffraction data is collected from many crystals in different orientations at once. The sample is prepared in such way that it contains many crystals in small volume. The sample prepared in this way is then exposed to X-ray radiation and the diffraction image is recorded typically using an area detector. Ideally, the orientations of crystals are distributed equally in a powdered sample. The resulting orientation averaging causes the three-dimensional reciprocal space that is studied in single crystal diffraction to be projected onto a single dimension. The plot of diffracted intensity versus the scattering angle has characteristic peaks at positions determined by the crystal lattice. In a two dimensional plot, see the figure 3.2, diffracted waves recorded on a detector make diffraction rings. The rotational averaging leads to diffraction rings around the beam axis recorded on a detector, rather than the discrete spots observed in single crystal X-ray diffraction. At high scattering angles the diffraction rings overlay making it very difficult to extract information of the structure factors, thus retrieve the structure of crystal made by large macromolecules by using X-ray powder diffraction method.

For this thesis I used X-ray powder diffraction method to characterize samples used in serial femtosecond crystallography experiments at LCLS. In order to confirm the crystalline nature of nano- and micro-crystals of proteins expressed and grown *in-vivo*, see the Chapter 4, and other protein crystals. Liquid suspensions with

crystals were placed in a thin walled glass capillaries typically used for SAXS measurements and placed under the synchrotron X-ray beam. A typical X-ray powder diffraction pattern, obtained from many proteinase K [Ebe74] micro-crystals in different orientations, is presented in the figure 3.2.

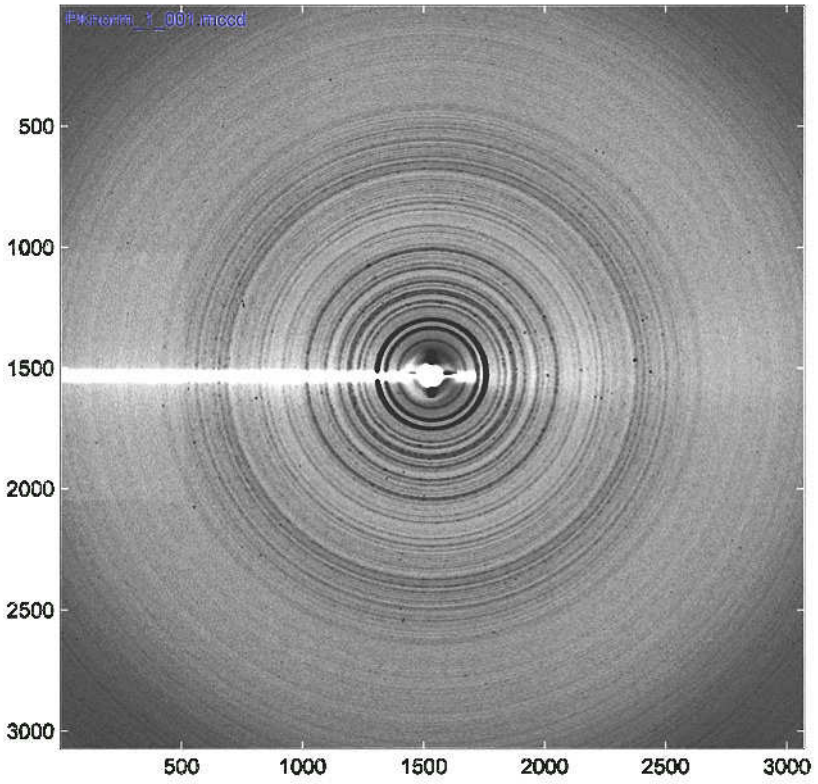


Figure 3.2: *The protein powder pattern. Sample: proteinase K micro-crystals in a glass capillary. Diffraction pattern was recorder at the X12 eam line at the DORIS light source in Hamburg. The wavelength of incident radiation was 1.54 Å(8 keV) with approximately* 10^9 *photons/sec. The beam spot size at the sample was approximately 300 µm x 300 µm. The exposure time was 60 seconds. The detector to sample distance was 384 mm, this gives a resolution at the corner of the image of 4 Å. Measurements were performed at room temperature using wavelength similar to this used at LCLS.*

Chapter 4

In-vivo protein crystallization

In this chapter a short description of the spontaneous occurrence of small protein crystals in living cells is presented. This phenomenon is an example of self-assembly that occurs in living cells. It has been observed that protein crystallisation occurs spontaneously in living cells in their natural environment and after stimulated overexpression of a target gene in a controlled cell culture [Doy05, Koo12].

It has been known for many years that protein crystallization occurs naturally *in-vivo* and has a number of applications for living organisms. Eminent examples of this biological self-assembly include storage proteins in seeds, enzymes within peroxisomes and insulin within secretory granules [Doy05].

In vivo crystallization has been considered as peculiar behaviour, therefore it has been neglected in comparison to the considerable efforts devoted to understanding and optimizing *in-vitro* protein crystallization for X-ray structure determination at conventional synchrotron sources. So far *in-vivo* crystallized proteins were not considered as suitable for structural investigations due to their small dimensions. The size of *in-vivo* grown crystals is limited by the cell size and by the amount of material that one cell can produce. The quality i.e. the degree of the internal order of *in-vivo* grown crystals has never been studied before. They seemed, however to be very well suited for serial femtosecond crystallography experiments at X-ray free-electron laser facilities. Their small size was not a disadvantage in this case and the fact that a cell culture can produce large quantities of highly concentrated crystal suspension was also profitable. During the work for this thesis the assumption that *in-vivo* grown protein crystals are well suited fro SFX experiments has been proved.

In-vivo grown crystals were proven experimentally that they are well suited for serial femtosecond crystallography (SFX) applications. These crystals can be extracted from a cell culture by the process of lysis of cells and then by applying different centrifugation steps and in order to separate the cell debris from crystals. The purified crystals can be then concentrated and re-suspended in water and easily

flown trough thin glass capillaries of the liquid-jet injector used in the experimental setup, see section 5.1 in the Chapter 5. In contrast, highly concentrated slat solutions or addition of viscous substances that are common in *in-vitro* crystallization trials, often display disability or difficulties to flow in a liquid-jet injector, making the SFX experiment very difficult to perform. To support the statement of simple applicability, two protein structures from micro-sized, *in-vivo* grown crystals have been determined using SFX and are briefly described at the end of the next chapter.

Application of protein crystals grown *in-vivo* in SF9 insect cells to SFX measurements have been demonstrated [Koo12] and this field is anticipated to grow rapidly when X-FEL facilities will become more accessible for users. Sections in this chapter are based on the above publication to which I contributed by taking part in performing the SFX experiment and analysing the SFX data, and by characterizing *in-vivo* grown protein crystals prior to the experiment. The results of my contribution to the above mentioned publication are presented in this chapter.

4.1 Baculovirus-Sf9 expression system

The Baculovirus Expression Vector System (BEVS) has been widely used for the production of post-translationally modified (for example glycosylation), biologically active and functional recombinant proteins [Alt99]. It is based on the introduction of a gene of a protein of interest into the viral genome region via homologous recombination using a transfer vector containing that target gene. The resulting recombinant *Baculovirus* lacks one of non-essential genes which is replaced by a gene encoding heterologous protein which can be then expressed by infected insect cells in large quantities [Kos99].

The baculovirus-mediated expression of recombinant proteins hosted in insect cells is a very well established protocol for the production of recombinant glycoproteins [Hit09]. Most commonly used lines of host cells are *Spodoptera frugiperda* (SF9) insect cells. The Baculovirus-SF9 expression system has been used to express and to crystallize recombinant proteins: cathepsin B and inosine monophosphate dehydrogenase from *Trypanosoma brucei* parasite. Those crystals were used during SFX experiments at LCLS and the results of the data analysis are presented in this thesis and in the publications listed in the front matter of this manuscript.

4.2 *In-vivo* crystallisation of *Trypanosoma brucei* Cathepsin B

The *Trypanosoma brucei* parasite causes human African trypanosomiasis (HAT), one of the most important neglected diseases, affecting over 60 million people in central Africa [Mac04]. The *Trypanosoma brucei* Cathepsin B (TbCatB) enzyme

has been shown to be important in the life cycle of the parasite. When the function of the TbCatB enzyme is deactivated it leads to death of the parasite. The cathepsin B belongs to the class of enzymes called cysteine proteases, which degrade polypeptides. Efficient and cost-effective drugs against sleeping sickness are not yet available, but the cysteine proteases such as TbCatB have been identified as potential drug targets in protozoan parasites [Bry09].

The *in-vivo* crystallization of polyhedrin-free, glycosylated enzyme cathepsin B from *Trypanosoma brucei* (TbCatB) in Sf9 insect cells infected with modified *baculovirus* was observed during routine inspection of the cell colony. Approximately 70 h after infection, the formation of needle-shaped microstructures were visible by light microscopy in Sf9 cells infected with recombinant baculovirus containing the gene encoding the pre-pro form of TbCatB. Cysteine proteases are synthesized by cells as inactive protein precursors (pre-pro form) with typically a 60 - 110 residue N-terminal propeptide. The propeptides are native inhibitors of their parent proteases. In our case, TbCatB was expressed in Sf9 cells in the pre-pro form. From the solved crystals structure it is apparent that the propeptide remains attached to the main body of the enzyme and the carbohydrate chains are in the structure as well when the the whole construct is crystallized. Detailed analysis of cells containing crystals by electron microscopy revealed damaged cell surface of the Sf9 cells and sharp, needle-like crystals 10 - 15 μm in length and 0.5 - 1 μm in width spiking out of the cells. Crystals appeared as dark areas in the transmission electron micrographs (TEM) with sharp square edges within the cytoplasm. An ordered lattice structure observed at higher magnification on TEM micrographs identifying these particles as protein crystals. See figure 4.1 for reference.

During the progressing process of infection, the number of observed cells with crystals inside continuously increased, until more than 70% of the cells contained one or more crystals. If released during cell lysis, crystals either remained attached to cell remnants or floated freely within the medium. The TbCatB crystals displayed very high rigidity and resistivity to mechanical stress. To isolate TbCatB *in-vivo* grown crystals, Sf9 cells were lysed and the resulting lysate containing crystals was subjected to differential centrifugation in order to extract pure crystal suspension. Lysis is a process in which the cells are degraded either by enzymatic or osmotic mechanisms that degrades the cell integrity. The efficiency of TbCatB crystal formation in SF9 cells was $\sim 5 \cdot 10^5$ purified crystals from $\sim 10^6$ cells [Koo12] .

4.3 *In-vivo* crystallisation of *Trypanosoma brucei* IMPDH

Similar protocol has been applied to crystallize inosine monophosphate dehydrogenase from *Trypanosoma brucei* (TbIMPDH) as in the TbCatB case. This enzyme catalyses the key reaction in guanine nucleotide biosynthesis: the conversion of in-

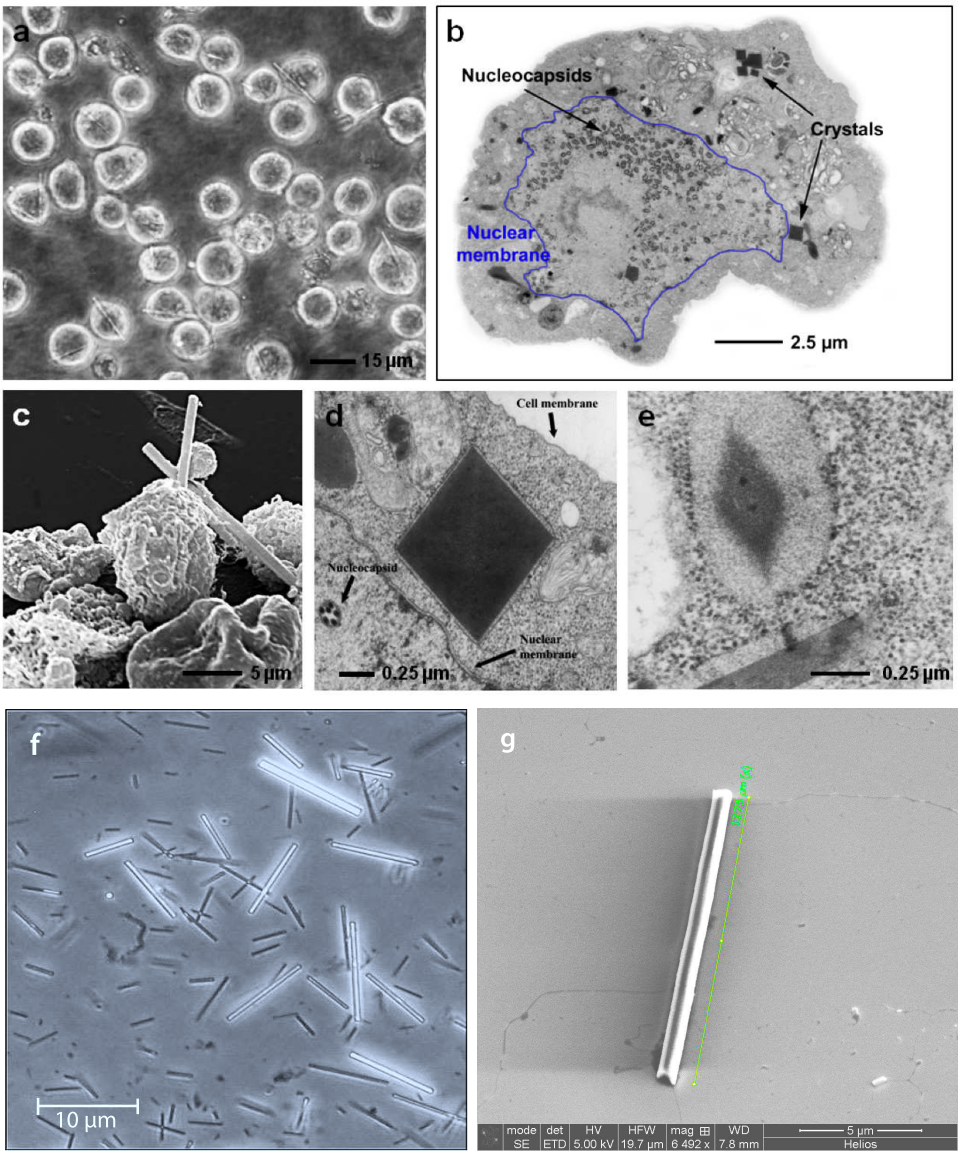


Figure 4.1: *Light microscopic and EM analysis of Sf9 insect cells with embedded TbCatB in-vivo crystals. (a) Light micrograph of Sf9 cells infected with TbCatB virus 90 h after infection. (b) Transmission EM micrograph of an embedded and sectioned infected Sf9 cell with crystals cut perpendicular to the long axis of the needle. (c) Scanning EM micrograph of a group of Sf9 cells infected with TbCatB virus 80 h after infection. (d) TEM micrograph of a sectioned sample, showing a crystal cut perpendicular to the long axis of the needle. (e) TEM micrograph showing the lattice structure of a crystal. (f) Optical micrograph showing crystals extracted from cells. (g) SEM micrograph of a crystal deposited onto a silicon nitride membrane, nano-sized structures are visible in the lower-right corner, possibly crumbled from one of the larger ones or just grew small.*

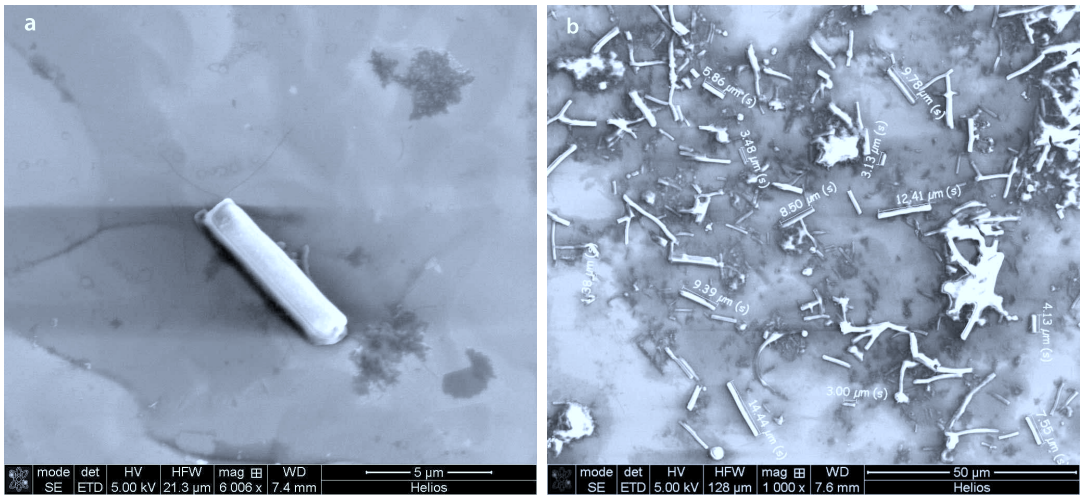


Figure 4.2: *Scanning electron micrographs of in-vivo grown IMPDH crystals. The IMPDH crystals were grown in Sf9 insect cells. (a) Image of a single crystal of IMPDH which dried on a silicon nitride membrane after extraction from a SF9 cell. (b) View of multiple crystals that shows characteristic distribution in crystal sizes and remaining cell debris.*

osine monophosphate (IMP) into xanthosine monophosphate (XMP) with the reduction of nicotinamide adenine dinucleotide (NAD) [Hed03]. The guanine nucleotide is a substrate for the synthesis of nucleic acids. As this enzyme is responsible for biosynthesis of vital components of *Trypanosoma brucei* parasite it is also a possible target for drug design against HAT.

After 80h of infection, SF9 cells were lysed and crystals were extracted from the lysate resulting in a purified suspension containing $\sim 10^5$ needle-like crystals in one ml of volume. Crystals of TbIMPDH appeared to be on average two times larger than *in-vivo* crystals of TbCatB. See the figure 4.2. Exact measurements and characterization of those crystals are presented in next section.

4.4 Characterization of *in-vivo* grown crystals

4.4.1 Light and electron microscopy

In order to characterize the size and shape and the approximate number of particles per volume of in-vivo grown crystals I used optical light microscopy and scanning electron microscopy. For the light microscopy an aliquot of 1 μl was placed on a glass microscopy slide, covered with a thin glass cover slip and then imaged before the solution dried out. For the electron microscopy an aliquot of 1 μl of diluted

suspension containing crystals was placed on a silicon nitride membrane and then imaged using a scanning electron microscope.

4.4.2 Second order non-linear optical imaging

Second Order Nonlinear Imaging of Chiral Crystals (SONICC) relies on the principle of Second Harmonic Generation (SHG) where two low energy photons combine to form a higher energy photon under intense electric field. Nonlinear effects such as SHG require high electric fields, thus require the use of a femtosecond laser. The laser operates with a pulse duration of 200 fs and has high peak power resulting in nonlinear effects but its short pulse is terminates before the damage associated with localized heating destroys the sample. The SHG process occurs also in non-centrosymmetric ordered protein crystals. Nearly all chiral protein crystals fall into symmetry classes, except icosahedral and octahedral, that are symmetry allowed for bulk SHG. Approximately 99% of known protein crystals (data taken from the Protein Data Bank [Ber00]) can be expected to generate bulk-allowed SHG [Wam08]. The signal that is generated in the presence of chiral crystals is easily distinguishable from solubilized or aggregated proteins, which produce absolutely zero SGH signal. The resulting images have extremely high contrast between crystalline material and other material in the solution.

The second-order nonlinear optical imaging of chiral crystals was demonstrated to be a sensitive and selective detection method for protein crystallization, with detection limits for the onset of crystallization corresponding to crystal dimensions well below the optical diffraction-limit. SONICC can be used to detect crystalline protein material invisible for light microscopy or hidden in the mush of aggregated protein [Kis11, Hau11]. Therefore SONICC used as a characterization tool could be ideal for protein crystallization trials aimed for SFX measurements.

The fact that *in-vivo* grown crystals of TbCatB and TbIMPDH were large enough to be observed by the light microscopy during standard inspection of SF9 cell culture used to over-express a protein of interest for further conventional purification and crystallization trials, did not clearly prove their crystalline composition. There was an indication given by transmission electron microscopy, that crystals of TbCatB posses ordered structure visible in one of the insets of the figure 4.1. To prove that, I applied SONICC (manufacturer: Formulatrix) to TbCatB and TbIMPDH crystals in order to detect SHG signal and to prove their crystalline nature. In figure 4.3 we can observe detected SHG signal generated by needle-shaped *in-vivo* crystals after excitation of a 1064 nm femtosecond laser. The laser power was 300mW, acquisition time was set to 1 second for the entire image. The size of the objects visible in the figure 4.3 is comparable to the sizes obtained using SEM. The SONICC measurements proved also that the needle-like structures of TbIMPDH are crystalline, and not the other material identified as cell debris as visible in figure 4.2, inset "b".

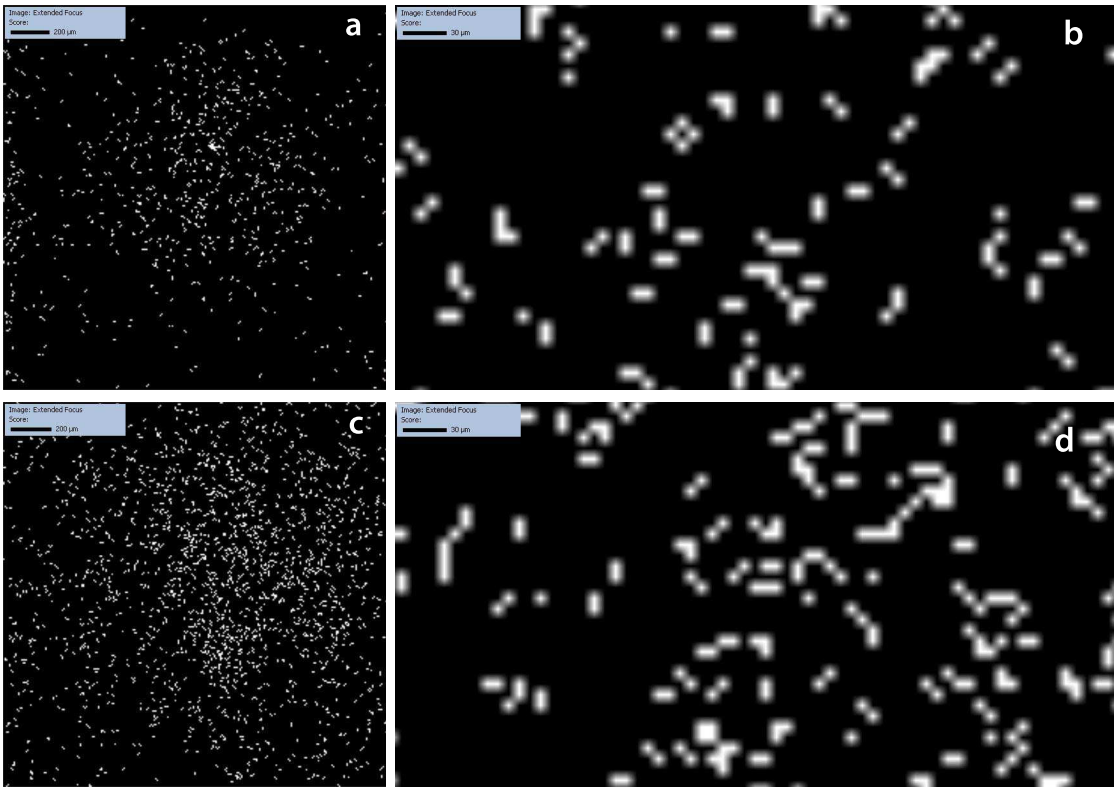


Figure 4.3: *Second-order nonlinear optical imaging of chiral crystals (SONICC) was used to image drops containing solution of (a,b) TbCatB and (c,d) TbIMPDH in-vivo crystals. Clear second harmonic (at $\lambda = 532nm$) signal has been detected confirming the crystalline nature of both in-vivo grown crystals.*

4.4.3 X-ray powder diffraction measurements at synchrotrons

The X-ray powder diffraction (XRD) technique has been applied to *in-vivo* grown TbCatB protein crystals in order to characterize their ability to diffract X-rays. This experiment was necessary to ensure if the material is of crystalline nature, therefore if it is appropriate for SFX measurements. This experiment was realized before the application of SONICC to *in-vivo* grown crystals became possible. An aliquot of the crystal suspension was placed into a thin (1mm inner diameter) borosilicate capillary, then closed using melted wax, and left in vertical position for 20 minutes to settle.

Sample prepared in this way was placed under the strong synchrotron X-ray beam of one of the PX beam lines (X10SA) at the Swiss Light Source. The energy of incident radiation was 8keV (λ=1.55Å). Exposure time was set to 3 seconds. The Pilatus 6M detector was placed 800 mm from the interaction region. Photon flux

was $1.6 \cdot 10^{11}$ photons/sec measured using the silicon pin diode [Owe09].

Clear powder diffraction rings were observed from TbCatB micro-crystals contained in a glass capillary as displayed in the figure 4.4. Strong powder diffraction signal obtained from nano- and micro-sized protein crystals which proved the crystalline nature of TbCatB needle-shaped assemblies. Additionally, a few single, strong Bragg peaks extended to resolution of ≈ 8.5 Å under this experimental conditions which indicated that during the SFX measurement those crystals would have the ability to diffract at least to that resolution. In the next section a preliminary SFX measurement of TbCatB crystals is described in which the crystals also gave detectable Bragg diffraction signal from which I could determine their unit cell dimensions.

4.4.4 Preliminary SFX measurements of TbCatB

A detailed explanation of the experimental setup, its components and the procedure of the serial femtosecond crystallography (SFX) measurement and the SFX data analysis method are described in the Chapter 5, section 5.1 (*Experimental setup for SFX*).

In this section I describe our preliminary measurement of *in-vivo* grown TbCatB crystals. This experiment was performed as a short test during one of the other SFX measurements performed at LCLS by my and others from our collaboration. Nevertheless, the data collected during this test led to significant improvement of the publication: Koopmann, R. et al.*"In vivo protein crystallization opens new routes in structural biology", Nature Methods (2012).* This preliminary measurement was performed before the main TbCatB and TbIMPDH SFX experiments, that enabled the high-resolution structure determination of these proteins. Here I present work that was my contribution to the above mentioned publication [Koo12].

The previously purified and characterized by myself *in-vivo* grown crystals of TbCatB were injected across the X-FEL beam a thin stream of water of $\approx 4\mu m$ in diameter using a liquid jet injector. A HPLC machine was used to supply the sample line of that injector. Single-shot diffraction patterns were recorded at up to 7.5 Å resolution using CAMP multi-purpose chamber [Str10]. Resolution limit was set by the available photon energy of the X-ray pulses, 2.0 keV, ($\lambda = 6.2$ Å) and the detector geometry. During 23.1 minutes, 83,224 detector frames were obtained. After background subtraction, 988 of the frames were identified to contain Bragg diffraction signal from TbCatB crystals. Each pattern was corresponding to a snapshot diffraction pattern from a different and randomly oriented TbCatB crystal. For a reference about the software and algorithms used to find protein diffraction patterns and to clean them from background see appendix A (*Data pre-analyser: Cheetah*).

During this preliminary measurement the time available for data collection was limited to 23 minutes. However, 879 (89%) of the 988 recorded diffraction patterns

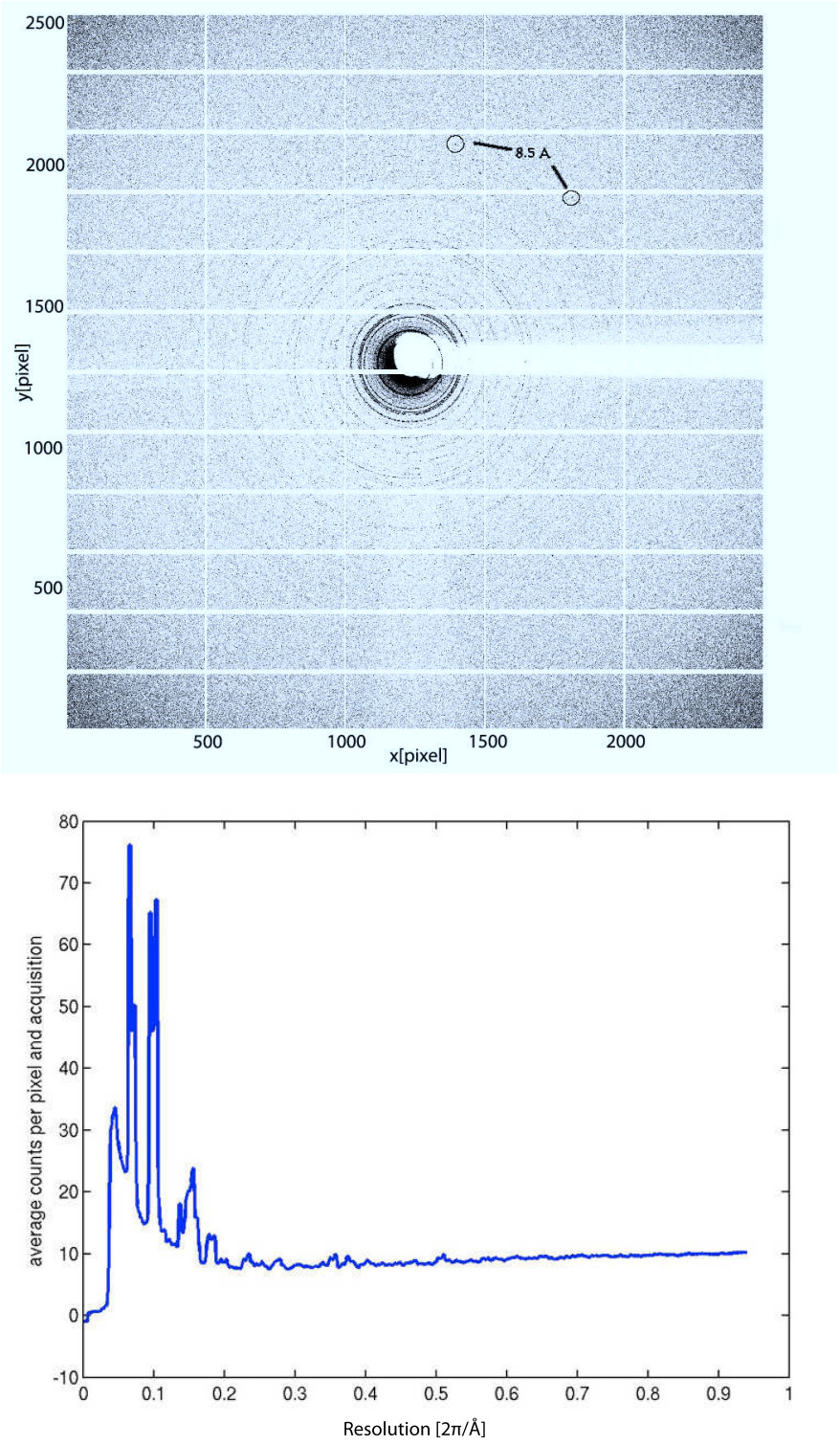


Figure 4.4: *Characterization of TbCatB in-vivo crystals by X-ray powder diffraction. Strong powder diffraction rings were recorded on a Pilatus 6M detector (upper panel) from crystals in different orientations settled at the bottom of a glass capillary. Single, strong Bragg peaks extended to resolution of ≈8.5 Å, at this experimental conditions, as marked with lines and circles. The sample to detector distance was 800 mm, the pixel size was 0.172 mm, the wavelength of incident radiation was* $\lambda = 1.55$Å*. The radial average (lower panel) of TbCatB powder diffraction pattern presents clear diffraction spikes up to ≈6.3* Å.

were indexed without previously knowing their unit cell constraints. For a reference about the software and algorithms used for indexing of protein diffraction patterns recorded during a SFX measurement see the appendix B (*CrystFEL software suite*).

From 879 measurements of unit cell parameters, the TbCatB crystal lattice parameters were determined to be $a = 122.9$ Å, $b = 123.6$ Å, $c = 53.4$ Å, $\alpha = 90.3\,^\circ$, $\beta = 90.2\,^\circ$ and $\gamma = 90.3\,^\circ$. Therefore, the unit cell was assigned to be tetragonal, with $a = b = 123.3$ Å and $c = 53.4$ Å. See histograms of measured unit cell parameters in figure 4.5.

It was not possible to obtain a complete three dimensional dataset during this short test, as indicated by the granularity in the summed single diffraction images, see the figure 4.6. This image should contain continuous X-ray powder diffraction rings if the data set was complete. The above mentioned figure presents the so called "virtual powder pattern", which in principle is almost identical with conventional X-ray powder pattern. The difference is that many diffraction patterns from single crystallites are measured separately and then are artificially added, whereas in the conventional X-ray powder pattern, the image is recorded from many crystals in random orientations at once. Other difference between "virtual powder pattern" from an SFX experiment at X-FEL source and a conventional powder pattern is that the incoming femtosecond X-FEL pulses posses distribution in their spectral bandwidth in a single X-ray pulse and between subsequent pulses due to the SASE process whereas conventional powder pattern is usually recorded at synchrotron sources, where the wavelength is well defined.

The result presented in this section reports the first SFX measurement of unknown *in-vivo* grown TbCatB crystals from which the values of unknown unit cell parameters were extracted and the SFX method was proven to be well suited for such kind of crystals.

It is interesting to note that the virtual powder patterns and their radial averages from a different experiment performed by our collaboration were used to make the main conclusions of the following articles: [Lom11] and [Bar11]. The main finding from this studies was to confirm that the Bragg diffraction at femtosecond X-FEL sources is gated by the radiation induced damage ("diffraction-before-destruction") to protein nano- and micro-crystals.

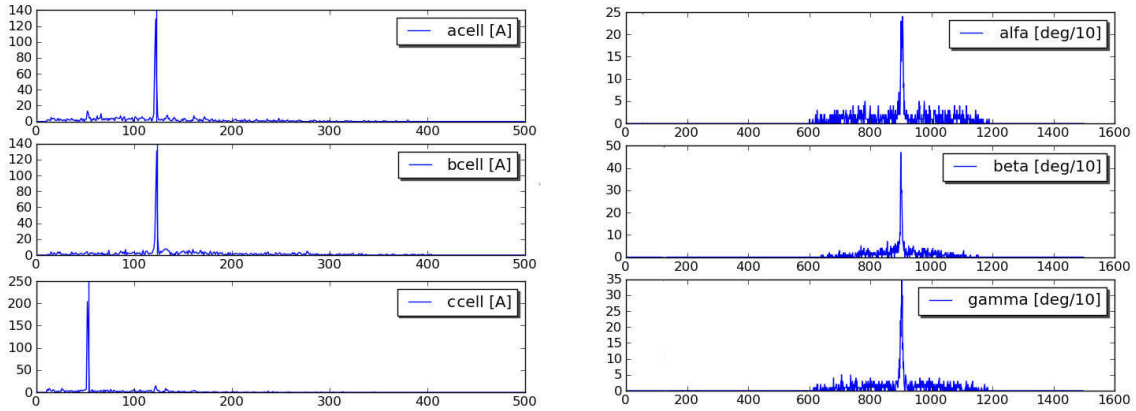


Figure 4.5: *TbCatb unit cell histograms from preliminary SFX measurements performed at LCLS using 2keV photons. A sum of 879 diffraction patterns of different, single TbCatB crystals in random orientations were indexed without restriction on unit cell parameters. Resulting 879 unit cell constants are presented in the histograms. The unit cell was assigned to be tetragonal, with* $a = b = 123.3$ Å *and* $c = 53.4$ Å.

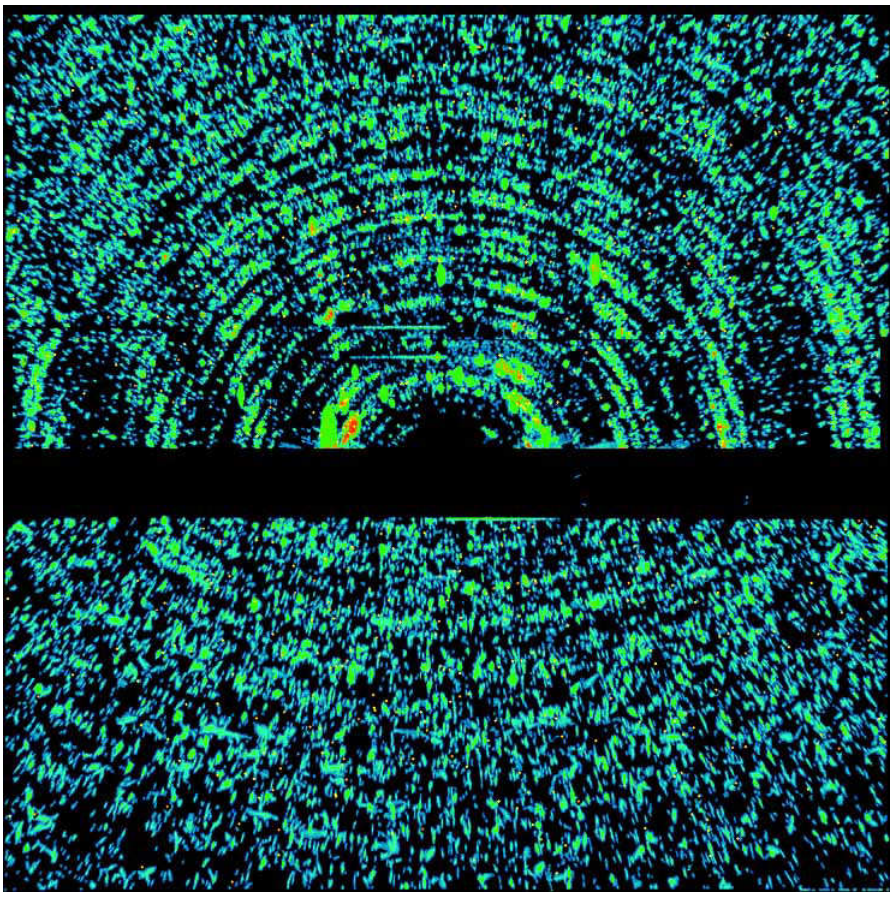


Figure 4.6: *A sum of 988 single-shot FEL diffraction patterns from TbCatB crystals in different orientations from preliminary SFX measurements using 2keV photons. The measurement was performed using the CAMP multi-purpose chamber. The lower panel of the pnccd detector was shifted to achieve higher resolution. At the edge of the detector, a maximum resolution of 7.5* Å *was obtained.*

Chapter 5

Serial femtosecond crystallography

In this chapter the methodology of serial femtosecond crystallography (SFX) for X-ray free-electron laser sources is described. The basic layout of the experimental setup is presented. A discussion about data collection and data analysis techniques used in SFX is included. Important differences to standard crystallography techniques, such as radiation damage effects or integration of partial reflections are addressed in this chapter. Two protein structures that have been solved by SFX during this work are presented at the end of this chapter.

The method of serial femtosecond protein crystallography (SFX) at the X-ray free-electron laser source has been presented for the first time in [Cha11]. In that first study the available photon energy of LCLS limited the resolution of recorded images. The modification of LCLS that enabled its operation at the hard X-ray regime enabled the high-resolution protein structure investigations. The first report of a high resolution protein structure determination using SFX method has been presented in [Bou12]. In this experimental method a concentrated suspension of nano- or micro-sized protein crystals flows across X-ray pulses generated by the X-FEL source in a liquid micro-jet. Whenever a crystal intersects with an X-ray pulse, a "snap-shot" diffraction pattern is recorded using a fast-readout detector. Many single crystal diffraction patterns from different and randomly oriented crystals are then indexed and merged into a three dimensional set of crystallographic structure factors using standard indexing methods and Monte Carlo approach for integration of partial reflections. Partial, or not fully integrated reflections occur in SFX case because of the perfectly still diffraction patterns, finite size of the crystals and the fact that the Ewald sphere may not intersect with exact centre of the Bragg peak. Molecular replacement or other phasing techniques known from macromolecular crystallography (MX) may by then applied to the merged data in order to retrieve the crystal structure of a protein of interest.

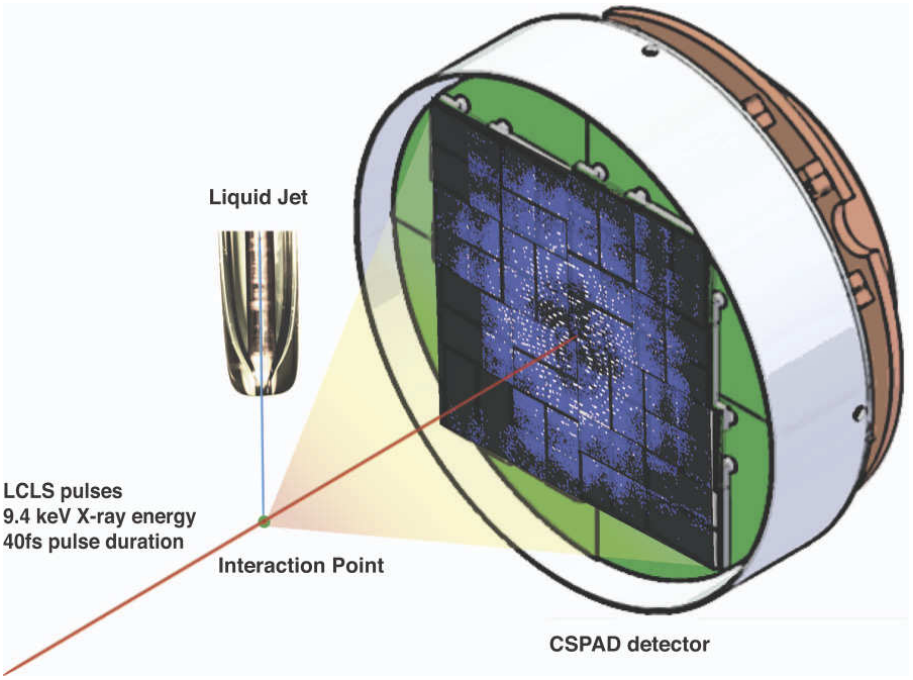


Figure 5.1: *Experimental arrangement for SFX at the CXI beam line of LCLS. A flowing suspension of in-vivo grown TbCatB and TbIMPDH microcrystals was introduced into vacuum in a liquid micro-jet. The liquid column was* $\approx 4\mu m$ *in diameter and flowed at* $10\mu l/min$*. More advanced designs allow for sub-micron liquid jets that operate at much lower flow rates, decreasing sample consumption and background scattering from water. Single-crystal diffraction data were recorded on a CSPAD detector operating at the 120 Hz repetition rate of the Xray pulses. The crystals remained fully hydrated and at room temperature in the interaction region.*

5.1 Experimental setup for SFX

During the experiments with *in-vivo* grown crystals we were using an experimental setup as schematically displayed in figure 5.1. Experiments were carried out in vacuum. The protein crystal suspension was introduced into interaction region using a gas dynamic virtual nozzle (GDVN), as explained below. It produced a thin column of liquid suspension of protein crystals that was continuously flown across the pulsed X-ray beam.

Diffraction patterns were read out on a CSPAD detector after each pulse, 120 times per second. A gap in the detector was required to allow the unscattered, intense X-ray beam to pass to a down-stream beam dump.

5.1.1 Liquid jet injector

One of the technical challenges for SFX experiments is the need for a suitable injector device that will be able to continuously introduce and refresh the sample under the highly energetic X-FEL pulses. In our experiment with hydrated protein crystals a gas dynamic virtual nozzle was used to deliver the sample to the interaction region [DeP08]. The GDVN guarantees continuous operation during data collection. The size of the liquid column produced by the GDVN matches the size of the crystals and the X-ray beam. Due to the continuous mode of operation of GDVN, some portion of the protein material is wasted between the X-ray pulses. This can be however decreased by manufacturing nozzles that produce liquid jets of sub-micron diameter, therefore reducing the flow rate to $\approx 1\mu l/min$. Other injection schemes currently being under development, appropriate for hydrated biological particles such as protein crystals, are electrospun liquid microjet [Sie12] or a pulsed injector – the "drop-on-demand" (DoD) dispenser (model MJ-AT-01-020) commercially available from MicroFab Technologies, Inc.

The thin liquid column that is produced by the GDVN moves with a velocity of approximately 10 m/s or higher depending on the jet diameter and the pressure of the liquid. Liquid jet is introduced into vacuum environment thus cools rapidly by evaporative cooling at high rate [Smi06]. However, the jet is kept in a high pressure He_2 gas sleeve which lasts over a distance of around few tens to hundred micrometers from the nozzle orifice. This high pressure He_2 gas sleeve is used to focus the liquid stream and additionally to reduce the evaporative cooling rate of the water jet by sustaining the high pressure environment in that region. The X-ray exposure is carried out about 100 μm from the nozzle tip where the temperature drop is calculated to be much less than 1 K [Sel12].

Figure 5.2 presents a micrograph of a GDVN in operation that was constructed by myself. This image was obtained using an Environmental Scanning Electron Microscope (ESEM). Imaging of a GDVN in operation was possible in the ESEM because of the special design of the microscope. In ESEM the specimen chamber sustains the high-pressure gaseous environment which is separated from the high vacuum environment of the electron optics column with at least two small orifices – pressure-limiting apertures (PLA). The gas leaking through the first aperture is quickly removed from the system with a pump that maintains a much lower pressure above the aperture [Dan85]. This is called differential pumping and is widely used in many experimental setups. An example of a sub-micron liquid jet, obtained by using a nozzle of also my construction, is presented in figure 5.3. A demonstration of the operation and the use of the sub-micrometre liquid jets in the ETEM (Enviromental Transmission Electron Microscopy) imaging was presented in [DeP11a].

The manufacturing process of GDVN consists of three steps (i) grinding a sharp tip of a typically 50μm ID glass capillary, used for flowing liquid suspension of

crystals, (ii) shrinking one of the ends of a larger ID glass capillary, used for co-axially flowing Helium gas, by fire polishing (iii) matching the shrunk end of the larger ID capillary with the sharp tip of a smaller capillary in order to achieve good alignment. Step (i) is usually performed by holding a tip of a borosilicate fiber capillary, that is commercially available, of typically 50 μm inner diameter and 320 μm outer diameter at an angle of 45 degrees to a rotating plate of sand paper with grain thickness of 1 μm. Other diameters of the fiber capillary are also available from the manufacturer and were found to create good quality nozzles. This process is performed until a sharp, regular tip is formed at the end of the capillary. Inspection of the quality of the tip during this process is usually performed by using optical microscope. Step (ii) is performed by holding and rotating a tip of a glass capillary with inner diameter of 600 μm and outer diameter of 1 mm above the tip of a flame until glass starts to melt and forms a symmetrical aperture. The size of this aperture must be match to the size of the tip of the thinner capillary manufactured in the step (i) so that the sharp tip of the thinner capillary can be aligned very close to the tip of the larger capillary, see figure 5.2 and 5.3. Step (iii) is necessary in order to achieve good alignment between two capillaries. When good alignment is achieved the liquid jet that is made by pushing liquid trough the inner capillary and Helium gas trough outer capillary is straight and thin. Different nozzles have distinct characteristics, such as the driving gas and liquid pressures, straightness or jet thickness. The pressure used to push the liquid trough the inner capillary is typically delivered by an HPLC machine. The Helium gas is typically delivered from a gas bottle, the pressure is regulated by electronic pressure regulator.

Important for consideration is the parameter called "hit rate" [DeP11b]. The hit rate measures how often a protein crystal from a protein crystal suspension that flows continuously across the X-FEL beam is hit by X-ray pulses generated by a pulsed X-FEL source. The hit rate depends on the average number of particles per interaction volume. Interaction volume is defined as the product of the square of the X-ray beam diameter and the diameter of the liquid column. The hit rate, for a given X-ray repetition rate and particle number density is then defined as a the product of the interaction volume (V), repetition rate (R) and the average number density of particles (n) in the interaction volume:

$$hitrate = nVR. \tag{5.1}$$

Hit rate decreases as the X-FEL beam diameter is made smaller and increases with repetition rate. Hit rate is often expressed as a percentage of shots collected that contain useful diffraction patterns.

During the experiments with *in-vivo* crystals we recorded average hit rate efficiency of the order of $\approx$ 10%. In order to consider the best use of an injector for hydrated bioparticles, such as protein nano-crystals, one should think about

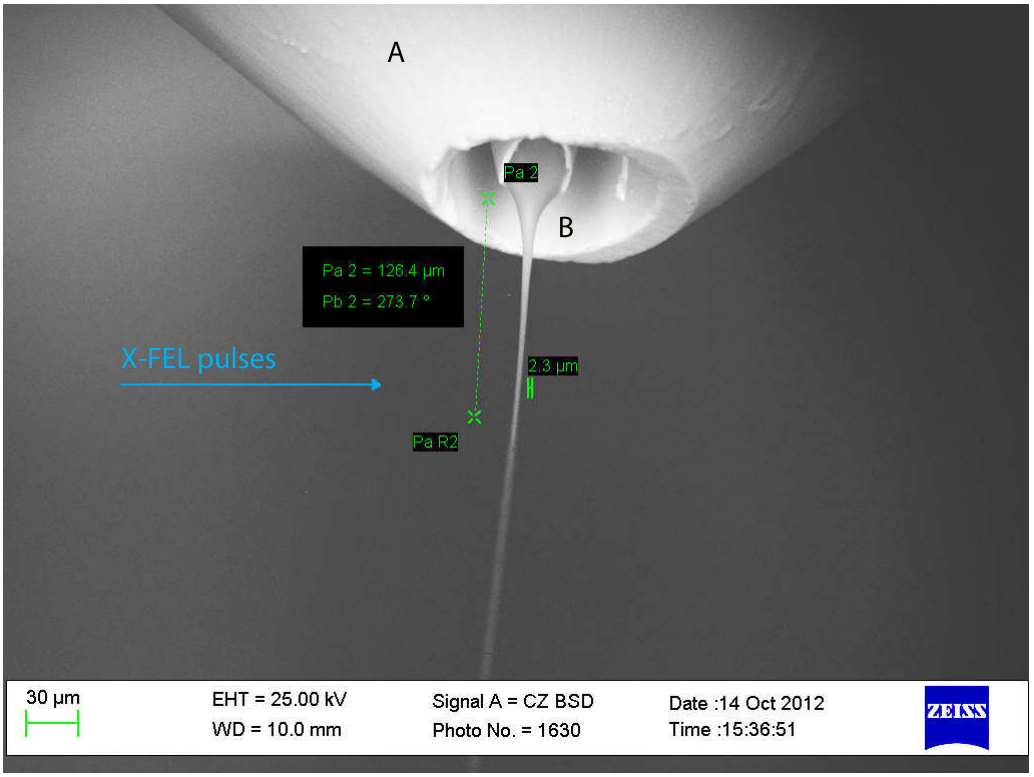


Figure 5.2: *Micrograph of a tip of gas dynamic virtual nozzle in operation taken by Environmental Scanning Electron Microscope (ESEM). A) presents the cone of an outer glass capillary that has been grind away in order to allow the high-resolution diffraction scattering to be recorded. B) marks the area where the liquid jet is formed and then focused to approximately* $2.3\mu m$ *in this case by co-axially flown Helium gas. The sharp tip of inner glass capillary, which is used to carry the crystal liquid suspension, is visible in the exit hole of outer capillary that carries the Helium gas. The positions of the X-FEL beam is marked. The jet brakes up into droplets after typically few tens of micrometers (droplets freeze over a certain distance with a certain cooling rate as they cool by evaporation into vacuum).*

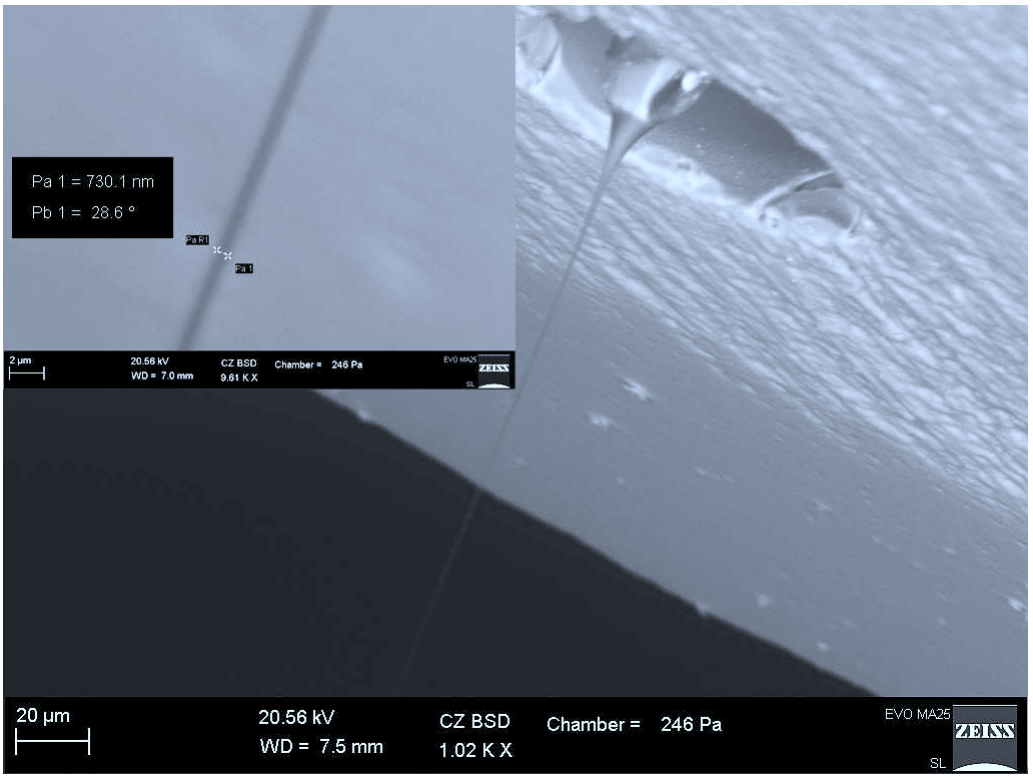


Figure 5.3: *Environmental Scanning Electron micrograph of a tip of gas dynamic virtual nozzle in operation. Inset in the upper right corner presents measurement of the jet diameter ($D = 730nm$) measured approximately 40 μm from the nozzle tip. Flow rate of that nozzle has been measured using Sensirion flow sensor, at that liquid pressure conditions (40psi) it was 1.2 $\mu l/min$.*

synchronisation with the X-ray pulses. The particle injector should be also able to introduce as many particles in a best collimated beam with optimum spacing between them and with optimum thickness of a water layer. This is essential to not waste too much of the sample nor to produce much scattering background from water that may introduce problems in data analysis. However, some water layer around bioparticles is necessary to prevent them from loosing their most "physiological" properties. More efficient and reliable injectors are needed which waste less sample, such as those that are synchronised to the repetition rate of the X-ray laser or those that produce jets smaller in diameter, therefore operate at lower flow rates.

5.1.2 Detector

Single shot diffraction patterns of randomly oriented *in-vivo* grown crystals were recorded at 120 Hz repetition rate (7,200 patterns per minute) by a Cornell-SLAC Pixel Array Detector (CSPAD). This detector consists of 64 independently controlled panels of 192×185 pixels each, forming a 1516×1516 pixel array (arranged into 1702×1702 pixels with small gaps between the panels and with a hole in the centre to let the intense x-ray pulses pass through [Phi10, Bou10, Bou12]. The pixel size is $110 \times 110 \mu m^2$.

Such detector was developed especially for the ultra bright LCLS source for the purpose of single molecule imaging at the Coherent X-ray Imaging (CXI) beam line [Bou10]. Many of this detector characteristics are tailored specifically for the demands of the single molecule imaging experiment at X-FEL source. Specific technical requirements to achieve that goal include the ability to distinguish single X-ray photon detection event in any given pixel, while maintaining the ability to detect thousands ($\geq$ 2500) of X-rays per pixel in other parts of the scattering pattern (high dynamic range). For an X-ray pulse from FEL source that approaches 10 fs in length, the instantaneous count rate for some pixels will be greater than 10^{17} photons per second. This count rate excludes the possibility of using a photon counting detector. The photon counting detectors are often used at synchrotrons because the count rates are manageable by such detectors. The X-FEL detector requires a photon integrating design, in which the detector has the capability of detecting several thousands of photons simultaneously [Ber10]. An other technical requirement, related to the ability to distinguish single photons, is a pixel-limited point spread function. In addition, the detector must be able to sustain a continuous frame-rate of 120 Hz and must have low electronic noise [Phi10].

There is a space for improvements for future X-FEL detectors. They should have increased dynamic range, reduced background electronic and readout noise and more pixels. Modular layout is highly desirable and the possibility to easily replace the front ends of such detectors also. For the upcoming new X-FEL sources with higher repetition rate, improvements in the read out time of detectors are also desirable [Gra09].

5.1.3 Data collection at LCLS

In order to retrieve high-resolution structures from *in-vivo* crystals, the following steps were necessary for the SFX data collection at LCLS. The schematic representation of the experimental setup is displayed in the figure 5.1. Samples were prepared as described previously in the Chapter 4 ("*In-vivo protein crystallization*"). Samples consisting of crystal suspension of approximately $10^8/ml$ *in-vivo* grown crystals of TbCatB and TbIMPDH at room temperature were used during experiments.

For the sample delivery or injection of suspensions of nano- and micro-crystals trough the liquid jet injector described previously, an anti-settling device [Lom12] in combination with a HPLC system was used to supply necessary liquid pressure and prevent crystalline material from settling on the bottom of the sample reservoir. HPLC stands for High-performance liquid chromatography. The gas dynamic virtual nozzle (GDVN) was introduced into the CXI experimental chamber [Bou10] using a steel tube supplied with in vacuum camera and a catcher-exhaust system developed at the Arizona State University [Wei12]. A 2 μm filter in the sample line connected to the GDVN injector was used in order to prevent any dust particles from clogging the nozzle. Necessary alignment of the liquid beam to the X-ray pulsed beam was carried out using pure water jet in order to save the protein sample. Data collection was carried out by our team and beamline scientists for approximately 9 hours for the TbCatB case and for approximately 2.5 hours for the TbIMPDH case. The recorded crystal hit rates were $\approx 7\%$ and $\approx 3\%$ respectively. Parameters at which the liquid jet injector was operating during the experiments were: approximately 4 μm in diameter and 8 – 10 μl/minute liquid flow rate.

5.2 Data analysis method

Reconstruction of a three-dimensional model from diffraction patterns recorded using X-FEL pulses from many identical copies of the same protein arranged into a nano-crystal requires slightly different approach than usually applied in conventional macromolecular crystallography performed at synchrotrons, where "large" protein crystals are typically used. See the Chapter 3 (*Introduction to X-ray crystallography*) for a reference.

In a SFX experiment a combination of thousands of crystals is required to complete full three dimensional data set of structure factors [Kir10, Kir11]. The small size of the crystals and perfect still X-ray diffraction snapshots introduce difficulties in data analysis and interpretation.

The indexing (i.e. assigning the Miller indices to a set of Bragg spots in diffraction patter) and merging of thousands of snapshot diffraction patterns from randomly oriented protein crystals provide technical challenges involving difficulties in quick analysis of many terabytes of raw data, which resulted from many hours of data collection using a CSPAD detector at a rate of 120 diffraction patterns per second. Currently there is no automated procedure implemented at CXI that would distinguish "on the fly" and save to disk only crystal hits. There would be a need for similar automated procedure that could distinguish hit from non hits from other types of samples, such as single biomolecules or viruses where the diffraction pattern is continuous and may contain only few scattered photons per pattern. In the case of crystal diffraction pattern the procedure to distinguish a hit from non

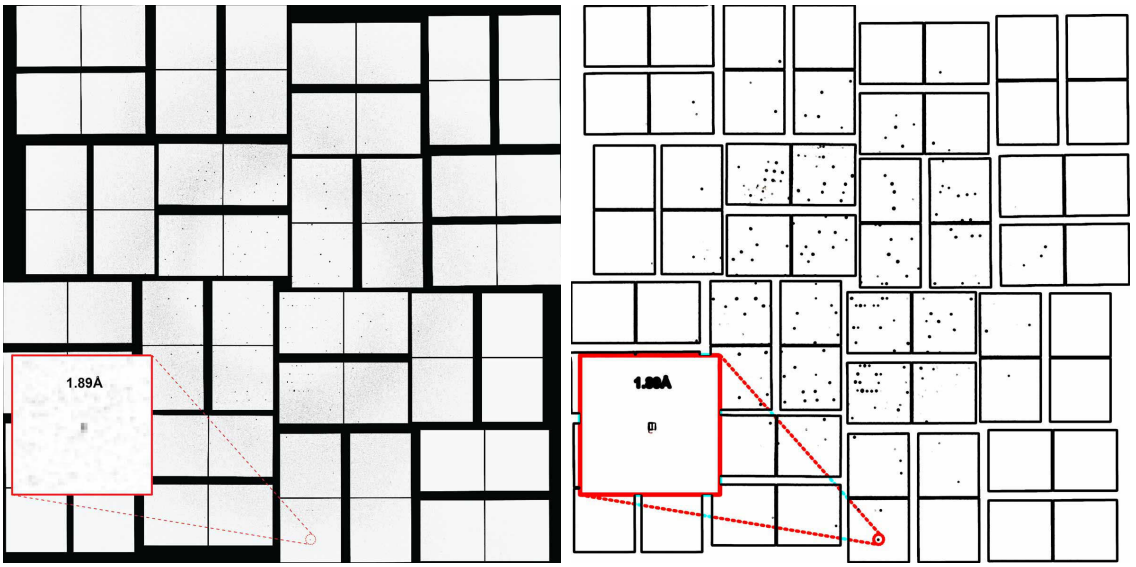


Figure 5.4: *A typical single-pulse X-ray diffraction pattern from individual TbCatB crystal (left). This diffraction pattern was recorded single X-ray pulse of approximately 40 fs duration. Bragg peaks were recorded on a CSPAD detector. The Bragg scattering signal extended to more than 2 Å resolution. The inset shows magnification of an individual high-resolution Bragg reflection at 1.89 Å resolution. On the right an enhanced version of diffraction pattern from the left side is presented in order to better visualize the Bragg spots on the diffraction pattern. Filters applied to the image from left: Gaussian blur (1px radius) and variance filter (5px radius), which highlights edges in the image by replacing each pixel with the neighbourhood variance.*

hit is much simpler. In our experiment this procedure was applied after collecting every detector frame. Efficient and fast filters were applied before indexing and merging, in order to extract diffraction patterns that contain crystal diffraction, see the Appendix A ("*Data pre-analyser: Cheetah*"). When that was performed, resulting single-shot exposures taken from crystals that were flown across X-ray pulsed beam in a liquid jet require determination of the relative orientation of diffraction patterns from thousands of randomly orientated crystals. This process is realized using automated crystallographic indexing schemes, where the orientation of the crystal relative to the laboratory frame is easily determined, see the Appendix B ("*CrystFEL software suite*").

5.2.1 Indexing

Indexing of crystal diffraction patterns was performed using the CrystFEL Software suite [Whi12]. Indexing is performed using conventional algorithms, such as the

"DPS" algorithm implemented in MOSFLM [Pow99, Les07] or the DirAx algorithm [Dui92]. Indexing is successful if the lattice parameters found by the autoindexing tool match the known unit cell of the crystal structure. It is also possible to determine the correct unit cell parameters and index the pattern when the unit cell parameters are not know previously i.e. from an unknown crystal structure. This procedure relies on the ability of the autoindexing tool to find unit cell parameters consistent across many diffraction patterns of the same crystal structure in different orientations. Such procedure I applied to our preliminary SFX test on TbCatB crystals in order to extract the unit cell dimensions, see histograms in figure 4.5 from section 4.4.4 of the Chapter 4 ("*In-vivo protein crystallization*")

Indexing ambiguities occur in many crystal structures which have less symmetry than their lattices. For example a rotation by 180° might overlay the lattice with itself but not the structure, which will be in different orientation. Conventional indexing relies only on the geometrical description of the lattice and is unable to distinguish orientations of the underlying structure which will change only the intensities of the Bragg spots. In a serial femtosecond crystallography experiment all indexing ambiguities allowed by the crystal symmetry will be present which will result in a problem for extraction of structure factors by the Monte Carlo integration.

As in conventional MX, this is a serious problem which relates to twinning [LP84]. While individual nanocrystals may not be twinned, indexing alone does not provide sufficient information in space groups that support merahedral twinning to allow merging of data from different microcrystals, without a 50% chance that they are merged in twin-related orientations. This problem is even worst when working with partial reflections from crystals of different size. For example in the hexagonal space group P 63 , which supports merahedral twinning, a rotation by 180° normal to the c axis brings the indexed reciprocal lattice into coincidence with itself, but not the structure factors, so that there are two ways to combine patterns from two different, untwinned nanocrystals. (The twinning operation takes reflection (h,k,l) to (k,h,l) in this case.)

5.2.2 Monte Carlo approach to extraction of structure factors

The duration of a single X-FEL pulse is in the order of few tens of femtoseconds, therefore the nanocrystal does not move during the exposure. Resulting snapshot diffraction pattern contain "partial" Bragg reflections, unlike those recorded at a synchrotron, where continuous crystal rotation introduced by a goniometer provides the angular integration across the Bragg condition needed to obtain a full structure factor. When collecting data using oscillation photography, it is necessary to distinguish between those reflections which have passed completely through the Ewald sphere ("full reflections") and those whose penetration is incomplete ("partial reflections") [Ros79]. Crystals used for SFX also have characteristic size distribution.

Which adds another complication factor in extraction of structure factors from many different crystals, because those crystals would have different scattering power which depends on the number of unit cells [Hol10]. For the smallest nanocrystals, the crystal shape effects will also influence the recorded Bragg intensities. The analysis of this problem is presented in Chapter 3 (*"Introduction to X-ray crystallography"*).

When the diffraction patterns contain partially integrated reflections the extraction of full squared structure factors $|F(\Delta k)|^2$ from equation 3.24 requires a 3D integration over crystal orientation around each Bragg condition. If sufficient redundancy in the tens of thousands of recorded patterns is available, this may be achieved by summing all intensity from many randomly oriented nanocrystals within a small volume around each reciprocal lattice point, thus adopting a "Monte Carlo" approach to integration [Kir11]. The integration assumes that the variance of crystal sizes is small. A suite of programs (CrystFEL) has been developed for the indexing and Monte Carlo merging of protein nanocrystal diffraction patterns [Whi12]. It also contain a set of useful scripts and programs that help in extraction figures of merit for a given data set.

An experimental result of the Monte Carlo integration of squared structure factors obtained from diffraction patterns recorded at a SFX experiment from thousands of small crystals of different shapes and orientations is presented in the next chapter. The quality of structure factors improves (crystallographic R-factors and R_{split} factor improve) when more data is added to the integration. See section 6.6 ("*Effect on R factors of the number of indexed patterns*") in the next chapter.

$$R_{split} = 2^{-1/2} \frac{\sum |I_{even} - I_{odd}|}{\frac{1}{2} \sum (I_{even} + I_{odd})}, \tag{5.2}$$

5.3 Radiation damage in SFX

X-rays are ionizing radiation. That means their energy is sufficient to remove electrons bound to atoms when a photon is photo-absorbed. Especially using X-FEL beams, where the beam intensity is large enough to multiply ionize significant fraction of atoms in the sample [Rud12]. Interaction of X-rays with matter depends on the cross section and for our case as well as in conventional macromolecular crystallography (MX) about 98% of 12keV X-rays photons are transmitted through the sample without interaction. Of the remaining 2%, 84% of the photons are annihilated in the production of a damaging photo-electron cascade, 8% undergo Compton scattering, and only the remaining 8% generate useful Bragg scattering. [Spe12, Hen93].

A measure of the energy deposited in a medium by ionizing radiation per unit mass is called absorbed dose. It is equal to the energy deposited per unit mass of medium, which may be measured as joules per kilogram and represented by the

equivalent SI unit, gray (Gy) The radiation induced damage matters for structure determination of protein crystals because it destroys the internal order of the crystalline lattice and changes the chemical composition of the sample which results in decreased scattering signal and incomplete reconstructed electron density. The effects of radiation damage influence also achievable resolution in standard macromolecular crystallography. Finer details of the retrieved electron density map are destroyed first [How09]. The tolerable dose for cryo-cooled protein crystals at synchrotrons to obtain a macromolecular structure was determined experimentally and is about 30 MGy [Owe06]. However, the rate at which the dose is delivered is important, at least for MX data collection at room temperatures. That means that a higher dose rate leads to the crystal being able to withstand a higher absorbed X-ray dose [Hol09, SD07].

The diffract-before-destroy principle for protein nanocrystals has been demonstrated at LCLS [Bar11]. The dose delivered to protein nanocrystals in that study exceeded 3GGy in few tens of femotseconds, but the diffraction was recorded before the onset of significant radiation damage occurred in the sample. The principal mechanism of radiation damage is the same in both cases (synchrotrons and X-FELs), i.e. photo-ionization by energetic X-rays. In SFX case much more ionization events occur in the sample per volume and time, causing the system literally to explode after few tens of femtoseconds. In contrast, during conventional MX measurements at synchrotrons using cryo-cooled protein crystals, the radiation induced damage does not cause the crystal to explode but manifests itself differently. It destroys the internal order of the crystalline lattice and changes the chemical composition of the sample which results in decreased scattering signal and incomplete reconstructed electron density. There are two kinds of radiation damage in conventional macromolecular crystallography at synchrotrons (MX), global and specific damage. For MX case the radiation induced processes that occur in the protein crystal are described in [O'N02]. Radiation induced damage in the MX can be reduced by freezing the sample and reducing radicals mean free path in the crystal, which increases the tolerable dose from the limit that has been observed at room temperatures.

SFX experiments performed for this thesis on TbCatB and TbIMPDH *in-vivo* grown protein crystals were carried out using the gas dynamic virtual nozzle (GDVN) injector. Thus were carried out at room temperature. The liquid with crystals that exits the inner capillary remains in thermal equilibrium with the co-axially flown helium gas. That prevents the liquid from evaporative cooling until X-ray pulse arrives, which happens only few tens of micrometers down from the nozzle exit, see section 5.1.1 for detailed description of the liquid jet injector. In our case of protein *in-vivo* crystals of TbCatB and TbIMPDH the dose was calculated to be of the order of around 30 MGy for our experimental parameters, see section 2.5. The dose

was calculated using the program RADDOSE [Pai10]. It is important to note that the dose limit for protein crystallography at conventional X-ray sources at room temperature is about 1 MGy [SD07].

We have not observed any radiation induced damage in the high-resolution electron density maps of TbCatB and TbIMPDH. Comparison of electron density maps obtained from TbCatB data collected at two different pulse lengths, 10 and 40 fs, didn't reveal any global or specific radiation induced damage. We conclude that the delivered dose was too low in that cases. X-ray pulses did not carry enough energy to create significant fraction of ionized atoms that would cause displacement of atoms and disorder in our crystalline structures before the X-ray pulse terminated. The effects of chemical changes due to photo-ionization, which are causing the radiation induced changes to the crystal structure during conventional MX studies, occur at longer time scales and we have not observed it in our SFX experiments due to very short pulse duration and exploding nature of the "diffract-before-destroy" principle. That was confirmed by simulations performed using plasma modelling code 'CRETIN' for protein-like assembly of atoms [Sco94, Sco01]. CRETIN is a a multi-dimensional non-local thermodynamic equilibrium (NLTE) radiation transfer code which can be used to model the ionisation dynamics in protein serial femtosecond crystallography. The modelled behaviour is well reflected in the measurements [Bar11]. This population kinetics plasma code takes into account radiation transfer for possible deviations from local thermal equilibrium during the X-ray exposure and follow the ionization and heating of the target sample in the photon pulse as it turns into a warm/hot dense plasma. It also follows changes in the cross sections and in the optical properties of the sample during exposure [Ber08]. Simulation parameters that were used for simulating the ionization levels and the root mean square displacement of atoms were set to match the experimental conditions of SFX measurements on *in-vivo* grown crystals. The sample resemble organic material of 1 μm in depth, with composition similar to protein without water layer around.

Simulations were performed for 10 and 40 fs X-ray pulses of 0.2 mJ and 0.6 mJ pulse energy respectively, focused to 10 μm^2 and photon energy of 9.4 keV. This gives the power density of $6kJ/cm^2$ for 40 fs pulses and $2kJ/cm^2$ for 10 fs pulses. The result is shown in the figure 5.5. As visible on the left hand side of that figure, the fraction of ionized oxygen atoms is very low at this pulse energy and intensity, which gives negligible values of the average ion displacement $\sigma(t)$ (right hand side of that figure). The average ion displacement $\sigma(t)$ was calculated from the diffusion coefficient that was obtained from the ion temperatures and the ion collision frequencies in the function of time obtained from CRETIN simulations.

$$\sigma(t) = \sqrt{2ND(t)}, \tag{5.3}$$

where N is the number of dimensions, in our case $N = 1$, $D(t)$ is the average

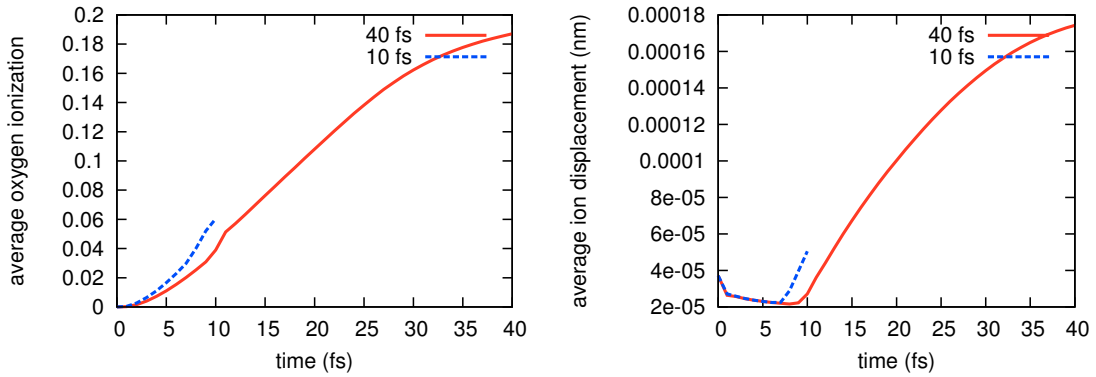


Figure 5.5: *Simulations using the 'CRETIN' code. Performed for "protein" like composition of atoms using parameters similar to experimental conditions of our SFX measurements on in-vivo grown crystals. Pulse durations 10 and 40 fs, 0.2 mJ and 0.6 mJ pulse energy respectively, focused to 10 μm^2 and photon energy of 9.4 keV. This gives the power density of $6kJ/cm^2$ for 40 fs pulses and $2kJ/cm^2$ for 10 fs pulses. The fraction of ionized oxygen atoms (left) is very small and would not cause much atomic displacement (right) for the high-resolution Bragg peaks to turn off as explained in [Bar11].*

diffusion coefficient

$$D(t) = \frac{k_B T(t)}{mv(t)}, \tag{5.4}$$

where k_B is Boltzmanns constant and m is the weighted mean mass of the ions, $T(t)$ it the ions temperature, $v(t)$ is the ion collision frequency [Bar11].

5.3.1 Bragg diffraction termination

If the X-FEL pulse has enough irradiance (W/m^2) to ionize a significant fraction of the atoms in the specimen thus to cause significant ion displacement during the pulse due to the repulsive force between ions, the so called Bragg termination effect occurs. The Bragg diffraction process may self-terminate before the end of the X-ray pulse due to the loss of crystalline order [Bar11]. Atoms in a crystal are ionized when exposed to the intense X-ray radiation from X-FEL. Photoelectrons that are ejected from atomic shells ionize further atoms. Empty electron shells create Auger electron cascades that either escape from the nanocrystal or contribute in creation of larger number of ionized atoms. After few or few tens of femtoseconds the system becomes charged due to ionization and then explodes due to repelling Coulomb forces occurring between the atoms [Cal12]. The short-range order is destroyed first when the atomic displacement increases over the time, which gives rise to the diffuse background scattering and reduction in the intensity of Bragg peaks [Lom11]. High

resolution Bragg peaks turn off quicker than low resolution ones.

5.4 *Trypanosoma brucei* Cathepsin B structure

This section is based on the article "Natively inhibited Trypanosoma brucei cathepsin B structure determined using an x-ray laser" published in Science [Red12]. This article has been selected one of the top 10 achievements in the year 2012. I have contributed to this article by characterizing the sample, taking part in performing the experiment at LCLS and in preparation of liquid jet injectors, analysing and all diffraction data, calculating figures of merit of the data, solving and refining the structure, modelling unknown parts into the structure and interpreting the results.

The Trypanosoma brucei cysteine protease cathepsin B (TbCatB), which is involved in host protein degradation [Mac04], is a promising target to develop new treatments against sleeping sickness, a fatal disease caused by this protozoan parasite. By combining two recent innovations, in vivo crystallization and serial femtosecond crystallography, we obtained the room-temperature 2.1 Åresolution structure of the fully glycosylated precursor complex of TbCatB. The structure reveals the mechanism of native TbCatB inhibition and demonstrates that new biomolecular information can be obtained by the diffraction before destruction approach of X-ray free-electron lasers from hundreds of thousand of individual microcrystals.

Over 60 million people are affected by human African trypanosomiasis (HAT), also known as sleeping sickness, which causes approximately 30,000 deaths per year [Fre10]. The knock-down of TbCatB in T. brucei resulted in clearance of parasites from the blood of infected mice and cured the infection [Abd08], qualifying cathepsin B as a suitable drug target. Cysteine proteases are synthesized as inactive precursors with N-terminal propeptides that act as potent and selective intrinsic inhibitors until the proteases enter the lysosome [Lec02] where the propeptide is released, forming the mature active enzyme.

5.4.1 Experimental details

The Coherent X-ray Imaging (CXI) beamline [Bou10] at the LCLS enables high-resolution data collection using the SFX approach. We used this instrument to obtain diffraction data from *in-vivo* grown crystals of TbCatB produced in the baculovirus-Sf9 insect cell system. Crystals with average dimensions of approx. $0.9 \times 0.9 \times 11 \mu m^3$ as determined by SEM imaging of dried crystals on silicon nitride membrane, were flowed in a 4 μm diameter column of buffer fluid at room temperature, at a flow rate of 10 μl/minute, using a liquid microjet. The amount of sample used was approximately 4 ml of crystal suspension in water of concentration approximately 10^9 crystals/ml. X-ray pulses from the FEL were focused onto this column

to a spot of $4\mu m$ diameter, prior to the breakup of the jet into drops. Single-pulse diffraction patterns of randomly oriented crystals, that by chance were present in the interaction region, were recorded at 120 Hz repetition rate by a Cornell-SLAC pixel array detector (CSPAD) at 9.4 keV photon energy (1.3 Åwavelength). An average pulse energy of 0.6 mJ at the sample (4×10^{11} photons per pulse) with a duration of less than 40 fs gave an x-ray intensity above $10^{17} W/cm^2$ and a maximum dose of about 31 MGy per crystal. This dose exceeds that tolerable at room temperature using conventional data collection approaches due to the radically different time scales and dose rates.

5.4.2 Data analysis details

Almost 4 million individual snap-shot diffraction patterns were collected. Of these, 293,195 snapshots contained crystal diffraction, which was identified using the data pre-analyser *Cheetah*, see Appendix A for reference. From all crystals hits exactly 178,875 (61 %) diffraction patterns were indexed using CrystFEL software suite, see Appendix B, and combined into a three-dimensional dataset of structure factors by Monte Carlo integration of partial reflections from each randomly oriented micro-crystal. The resulting complete set of structure factors contains 25,969 reflections in a resolution range from 20 to 2.1 Å. The high quality of the merged dataset is indicated by a R_{split} of 10.2%. Data statistics are summarized in table 5.1 and table 5.2. The data quality metric $I/\sigma(I)$ for SFX is called $Merged\ I/\sigma(I)$ and is different to the standard definition. In SFX I means the intensity of reflection extracted by Monte Carlo integration and $\sigma(I)$ is the standard deviation of the distribution of error measurements of I. Figures 5.6 and 5.7 present the Wilson Plot and R_{split} in function of resolution confirming the good quality of merged structure factors up to 2.1 Åresolution. The structure was solved by molecular replacement using the coordinates of the previously determined in vitro crystallized mature TbCatB structure (PDB ID: 3MOR) [Koo12] as a search model.

Figure 5.8 presents the molecular replacement model used to solve the structure of *in-vivo* greown crystals of TbCatB. The green mesh visible in that figure represents the omit ($F_{obs} - F_{calc}$) electron density map observed directly after the molecular replacement phasing procedure. This is where the unknown parts of the protein model, the propeptide and carbohydrates were located. That parts of the electron density map obtained form TbCatB crystals are the first new biological information from protein crystals obtained using an X-FEL. Figure 5.9 presents TbCatB surface model with the electron density where the propeptide and carbohydrate chains were placed into the observed density basing on the previously known amino-acid sequence. Final model after model building and refinement is presented in the figure 5.10.

Table 5.1: *SFX data collection and refinement statistics for TbCatB. Numbers in brackets represent values for the highest resolution shell.*

Data collection	
Wavelength (Å)	1.32
Maximum dose per crystal (MGy)	31
Space Group	$P4_22_12$
Cell dimensions a, b, c (Å)	125.4, 125.4, 54.6
V_M(Å^3/Da) / solvent content (%)	3.2 / 61
Number of crystal hits	293,195
Number of indexed patterns	178,875
Number of unique reflections	25,969
Resolution (Å)	20 - 2.1 (2.175 - 2.1)
Completeness (%)	100 (100)
$I/\sigma(I)$	11.92 (2.37)
R_{split}	0.10 (0.35)
Redundancy	7,807 (7,060)
Refinement	
No. reflections used in refinement	24,648
No. reflections used for R_{free}	1,321
R_{work}/R_{free}	0.181 / 0.214
No. of non-hydrogen atoms	
Protein	2,386
Carbohydrate	67
Water	98
B-factors (Å^2)	
Wilson B-factor	49.0
Protein (main chain / side chain)	46.1 / 47.7
Carbohydrate	65.7
Water	50.0
Diffraction precision index (DPI) (Å)	0.146
R.m.s. deviations	
Bond lengths (Å)	0.013
Bond angles (°)	1.578
Av. r.m.s. B-factor main/side chain atoms (Å^2)	1.491 / 1.937
Ramachandran plot (% of residues)	
Most favoured	96.7
Allowed	3.3
Disallowed	0

Table 5.2: *Quality indicators for the individual resolution shells of the TbCatB dataset.*

Resolution shell (Å)	**Number of unique reflections**	**Redundancy**	**Merged** $I/\sigma(I)$	R_{split} (%)
20.000 – 4.509	2,793	7,541	32.91	3.0
4.509 – 3.585	2,648	8,094	27.57	3.3
3.585 – 3.134	2,609	8,353	19.12	3.5
3.134 – 2.848	2,588	7,656	12.04	5.8
2.848 - 2.645	2,588	7,656	8.22	11.5
2.645 - 2.489	2,568	7,968	5.93	16.2
2.489 - 2.365	2,552	8,212	4.55	19.6
2.365 - 2.262	2,560	7,899	3.59	24.4
2.262 - 2.175	2,536	7,505	2.93	28.5
2.175 - 2.100	2,540	7,060	2.37	35.3

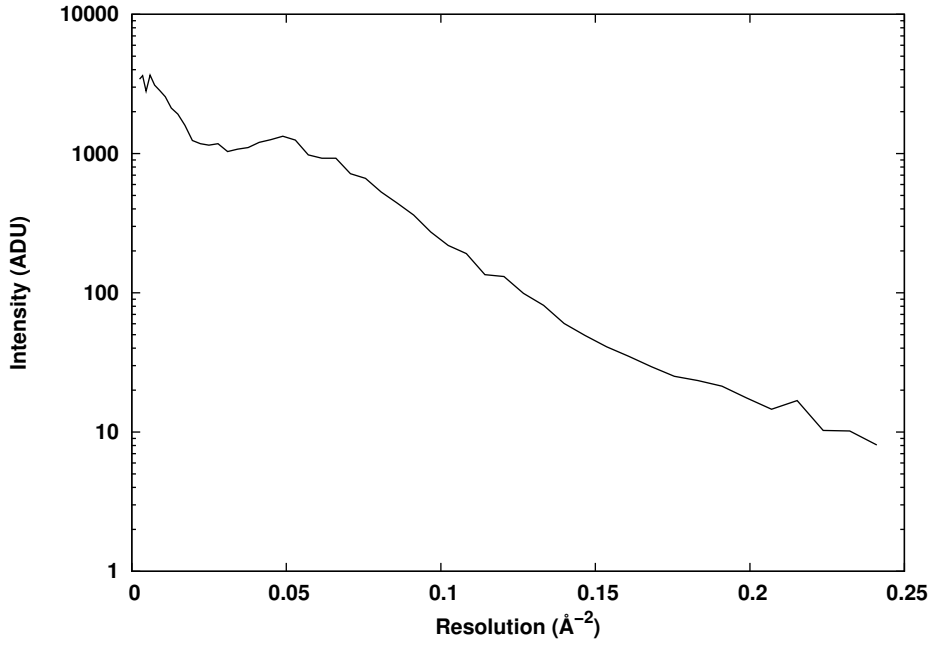


Figure 5.6: *TbCatB Wilson plot.*

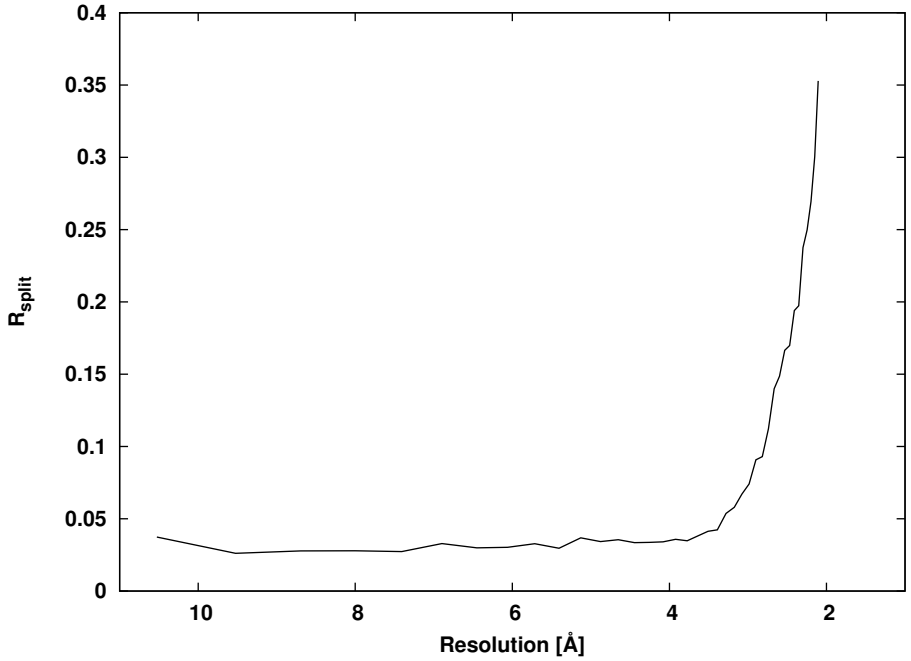


Figure 5.7: *R_{split} in function of resolution for TbCatB data set.*

5.5 *Trypanosoma brucei* IMPDH structure

In this section I describe my contribution to the experiment during which another unknown *in-vivo* crystallized protein structure was solved. I performed in the characterization of the sample, as described in previous chapters, I took part in performing the experiment at LCLS and in preparation of liquid jet injectors, I performed the analysis of diffraction data, calculation of figures of merit, structure solution and refinement of the structure.

The role of TbIMPDH is briefly explained in the section 4.3 of chapter 4. Crystallization was performed *in-vivo* by SF9 cells after cells were infected by baculovirus vector containing the target gene in place of the polyhedrin gene. For details explaining the expression and crystallization systems see chapter 4.

The structure of TbIMPDH was unknown until our successful application of SFX to *in-vivo* grown crystals of TbIMPDH. Experiments were performed using the CXI Coherent X-ray Imaging instrument at LCLS using experimental equipment and procedures similar to those applied to TbCatB crystals.

After data collection (2 hours 15 minutes), 29,247 hits were identified using data reducer *Cheetah*, see appendix A. This corresponds to overall hit rate of 3%. From identified crystal hits 9,724 could be indexed, which gives 33 % indexing yield. Data

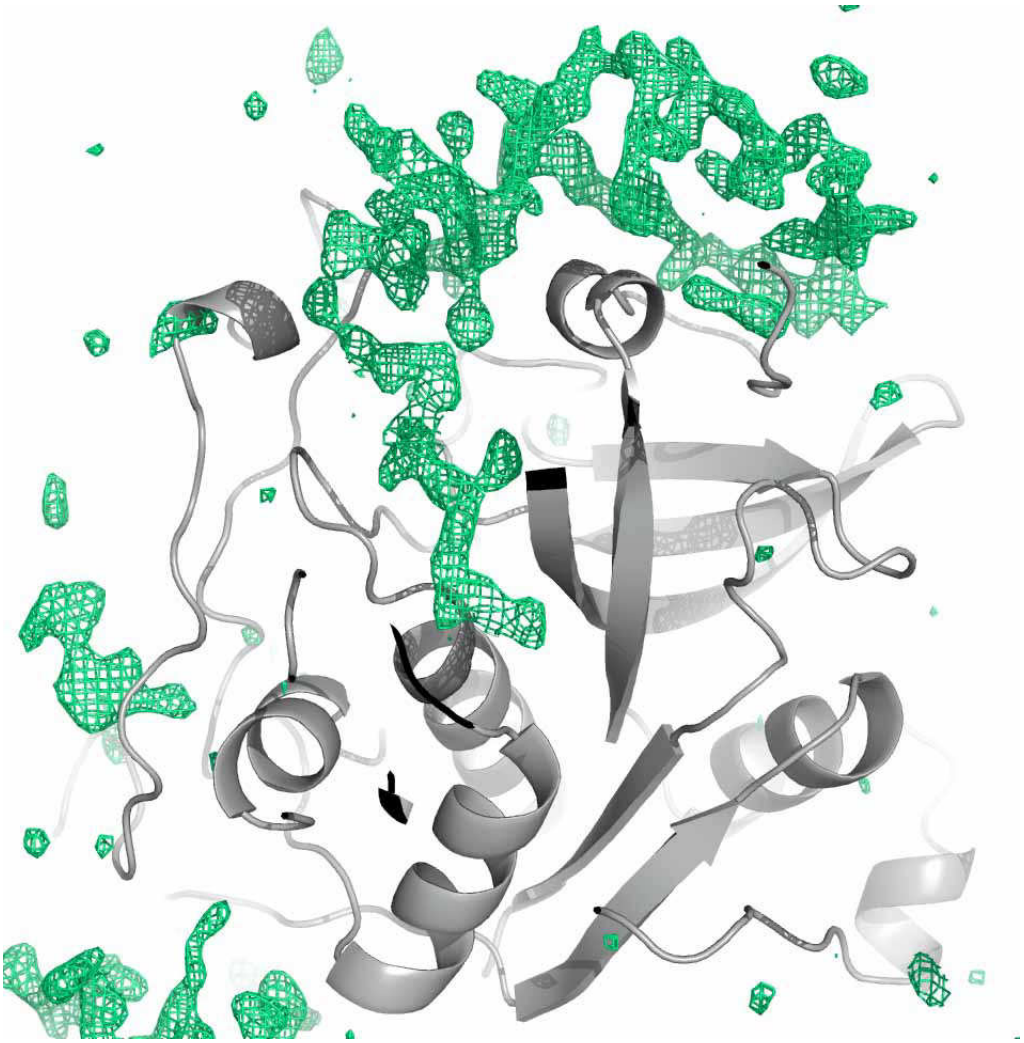


Figure 5.8: *Molecular replacement search model (PDB ID: 3MOR) used to solve the structure of TbCatB in-vivo crystals (gray) with an omit ($F_{obs} - F_{calc}$) electron density map (green) presenting the propeptide and carbohydrate sites. This is the first unknown biologoical information revealed by X-FEL from protein crystals.*

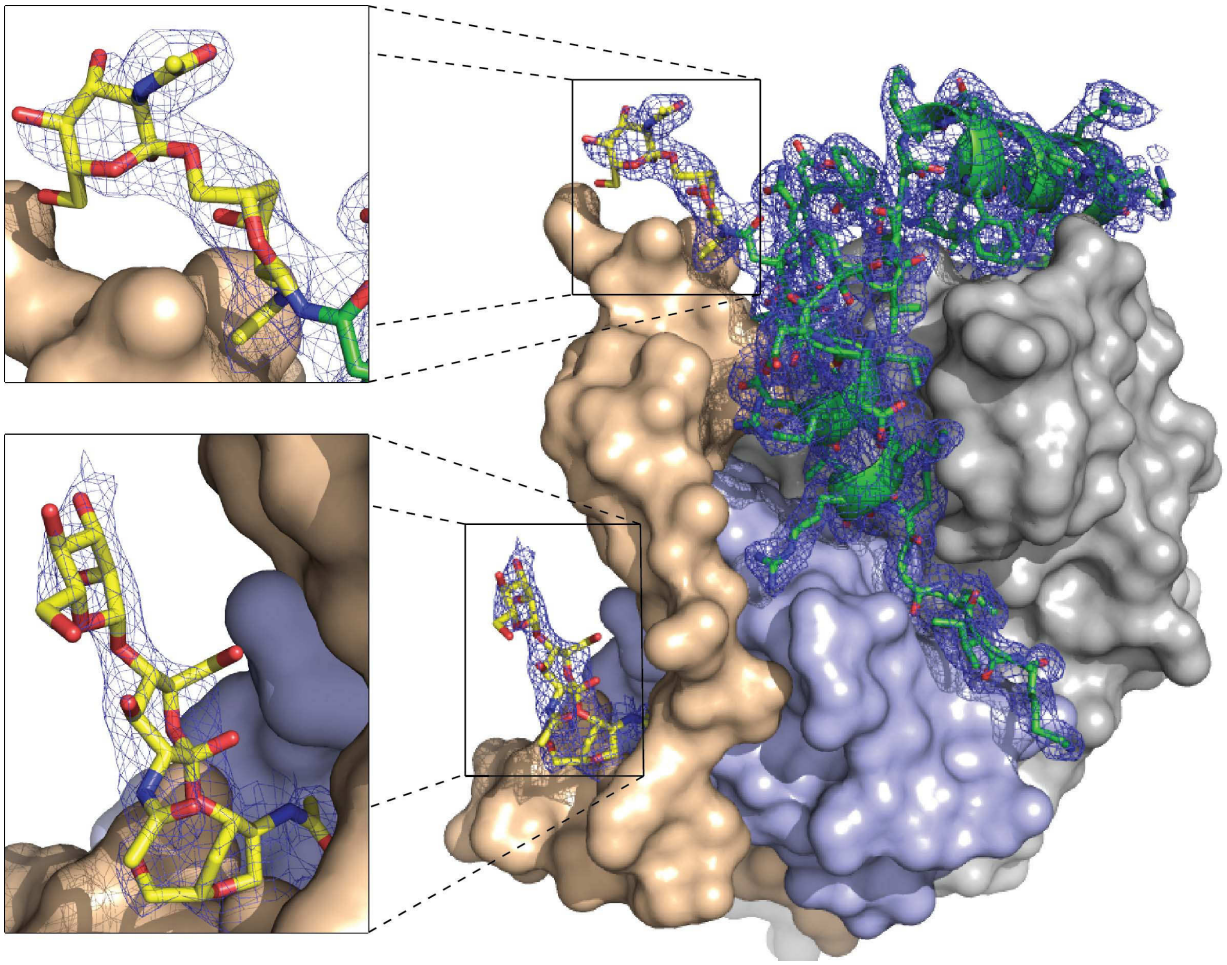


Figure 5.9: *TbCatB surface model with* $2F_{obs} - F_{calc}$ *electron densities displayed for the propeptide and carbohydrate chains. Models of propeptide and carbohydrate chains were modelled into electron density.*

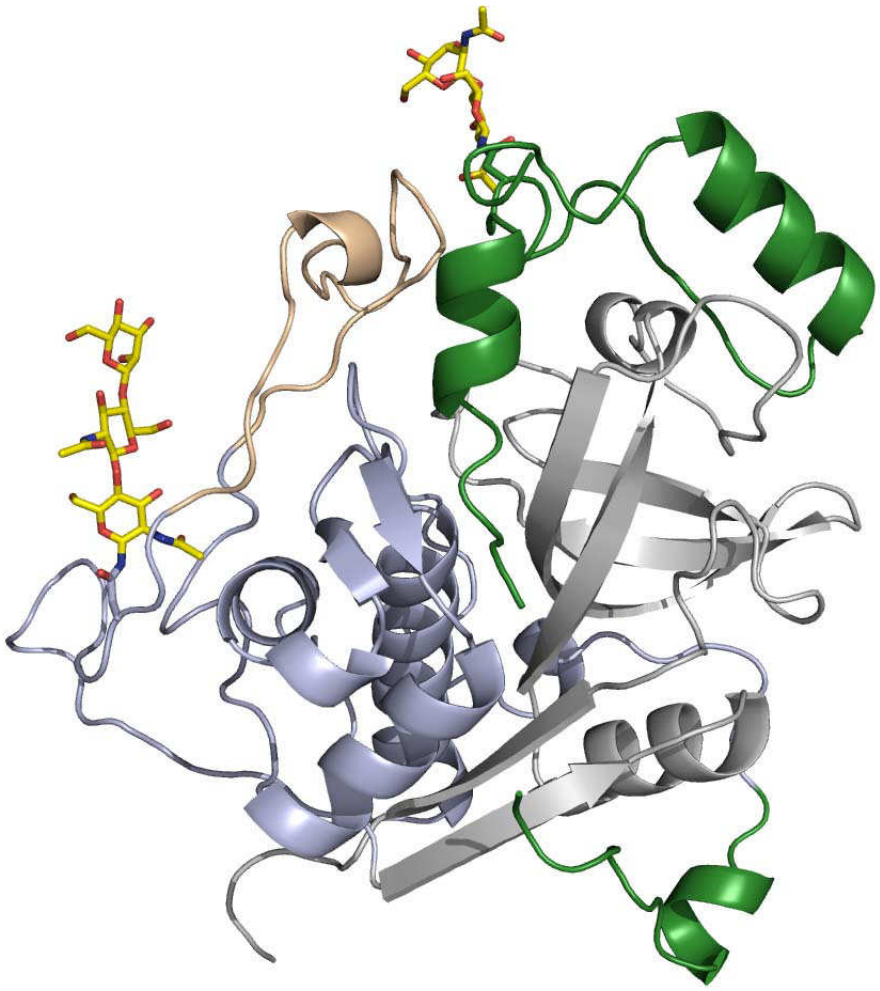


Figure 5.10: *Final TbCatB model in cartoon representation where lines represent the carbon backbone, spirals and large arrows represent the secondary protein structure, alpha helices and beta sheets. Carbohydrate atoms are represented by stick representation. The green chain is representing the inhibitory propeptide, yellow sticks represent the carbohydrate chains – the first new biological information revealed by an X-ray free-electron laser.*

Table 5.3: *Quality measures of TbIMPDH data set in resolution shells up to 3.2* Å *resolution. Mean redundancy is 184 measurements per reflection. Mean signal to noise ratio is 4.14 (1.01 in the highest resolution shell).*

Resolution shell (Å)	**Redundancy**	**Merged** $I/\sigma(I)$
46.5 - 6.9	300	7.28
6.9 - 5.4	232	6.30
5.4 - 4.7	201	6.19
4.7 - 4.3	187	5.76
4.3 - 4.0	175	4.78
4.0 - 3.8	169	3.86
3.8 - 3.6	157	2.81
3.6 - 3.4	140	1.95
3.4 - 3.3	144	1.45
3.3 - 3.2	140	1.01

quality indicators suggest good statistics up to 3.2 Å resolution.

The structure was determined using molecular replacement method. Monomer of homologue structure of human IMPDH was used as a search model (PDB ID: 1NFB). The crystal structure of TbIMPDH is a tetramer with four subunits related by a crystallographic 4-fold axis. For comparison, detailed structural and biochemical analysis of bacterial IMPDH is presented in [Zha99]. After molecular replacement was successful, amino acid sequence was changed to the Trypanosoma brucei sequence. Manual model refinement was performed in order to match positions of flexible loops regions to electron density. Statistics of the TbIMPDH data set are displayed in the table 5.3. Due to limited beam time we collected only a fraction of data in comparison to TbCatB which means reduced redundancy and reduced $Merged\ I/\sigma(I)$. Nevertheless, this data was of sufficient quality to determine the structure of TbIMPDH up to 3.2 Å resolution. If more diffraction data was collected, better statistics for higher resolution could be obtained and the resolution limit of the useful diffraction data could be set higher. Current refinement statistics are presented in the table 5.4. Tetramer of TbIMPDH structure is presented in the figure 5.12. The Wilson plot for TbIMPDH is displayed in figure 5.11 which presents the good quality of merged diffraction intensity up to the resolution limit mentioned above.

Table 5.4: *SFX data collection and refinement statistics for TbIMPDH. Numbers in brackets represent values for the highest resolution shell.*

Data collection	
Wavelength (Å)	1.3
Space Group	$P42_12$
Cell dimensions a, b, c (Å)	215.0, 215.0, 92.5
$V_M(\text{Å}^3/Da)$ / solvent content (%)	3.2 / 61.6
Number of crystal hits	29,247
Number of indexed patterns	9,724
Number of unique reflections	34,541
Resolution (Å)	46.5 - 3.2 (3.1 - 3.2)
Completeness (%)	100 (100)
$I/\sigma(I)$	4.14 (1.0)
Redundancy	184 (140)
Refinement	
R_{work}/R_{free}	0.27 / 0.30

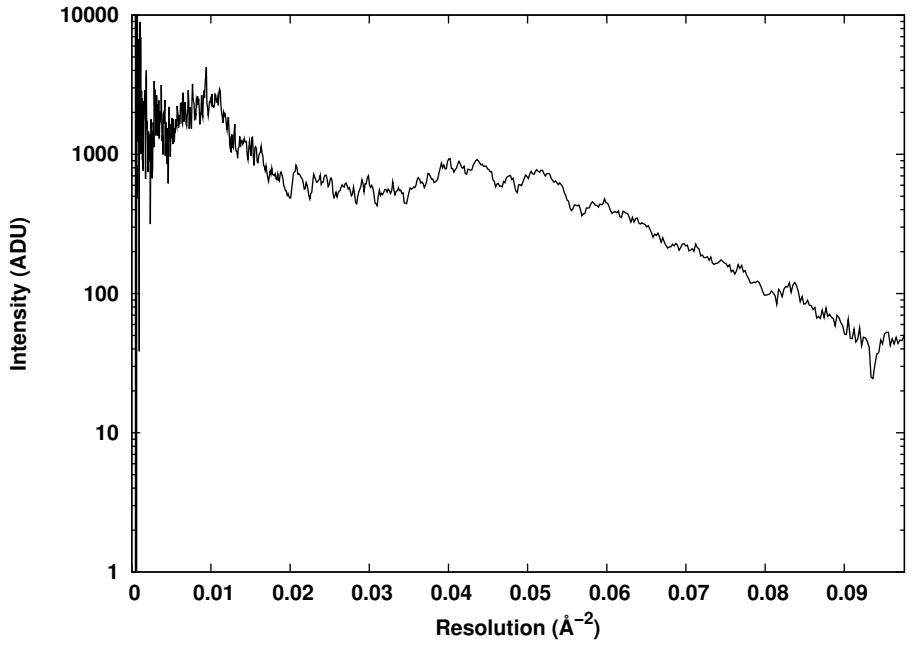


Figure 5.11: *TbIMPDH Wilson plot up to 3.2* Å*resolution.*

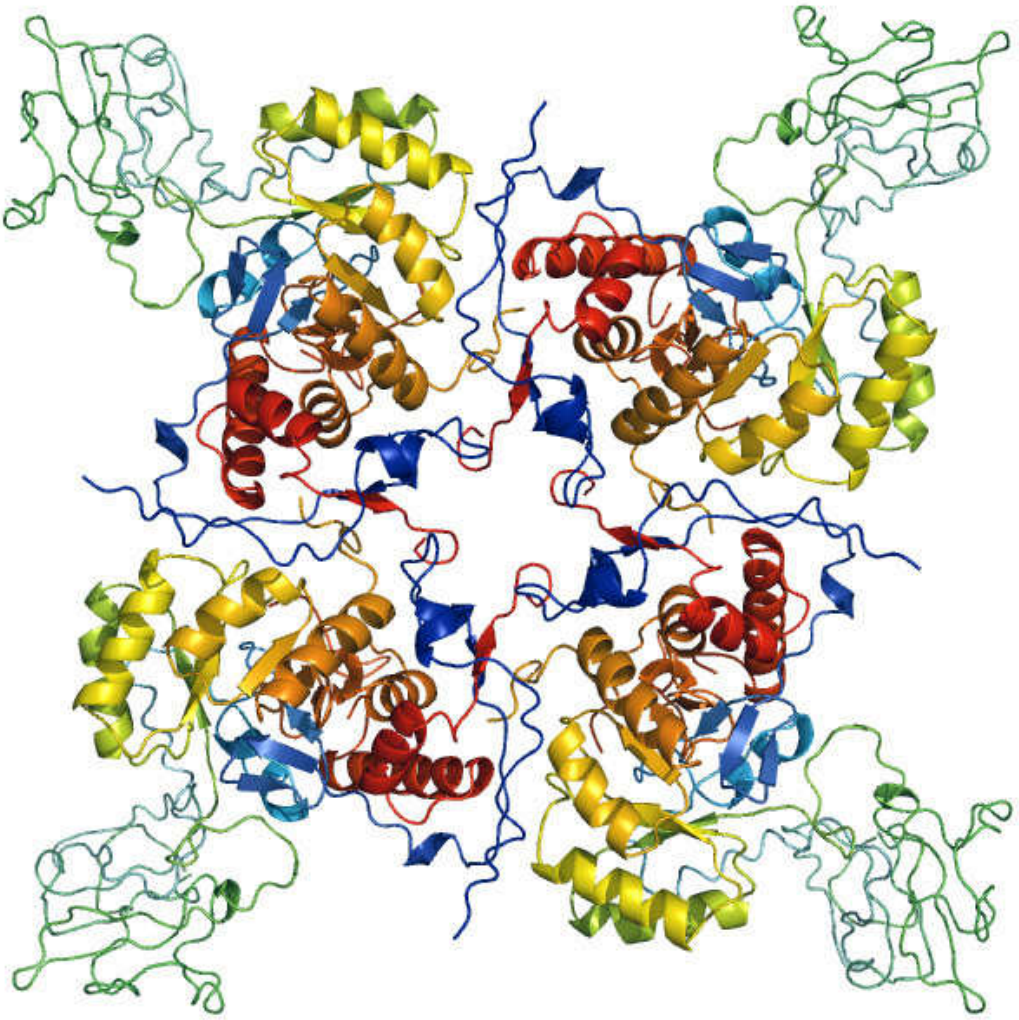


Figure 5.12: *Tetramer of TbIMPDH crystal structure solved using SFX and in-vivo crystallization. This structure is the first completely unknown structure of a biological macromolecule solved using X-ray radiation from a free-electron laser source.*

Chapter 6

Verification and assessment of SFX

In this chapter an exhaustive investigation on how well serial femtosecond crystallography works for small *in-vivo* grown crystals is presented. In general this method can be extended to any other macromolecular structure solved from thousands nano- or micro-crystals using pulses from an X-ray free-electron laser (X-FEL).

SFX is a new method for obtaining structural information from protein crystals that are too small for conventional X-ray diffraction studies at synchrotrons. X-FEL provides many orders of magnitude more brighter X-ray radiation in a single few femtosecond pulse than any existing synchrotron source. This enabled the possibility of using nano-sized protein crystals for X-ray crystallography at X-FELs.

At first I present a study about radiation damage to TbCatB protein crystals, quantify if there was any model bias in TbCatB electron density map obtained from SFX data due to molecular replacement. I identify what are the main sources of errors in Monte Carlo approach of merging the data from randomly thousands of diffraction snap-shots obtained from different, randomly oriented crystals. I present results obtained for different indexing schemes, detector effects and sizes of crystallines. Lastly, an interesting investigation on unit cell distribution of almost 180 thousands TbCatB crystals is presented and its connection to preferred orientation of crystals due to flow alignment in the liquid jet injector.

6.1 X-FEL radiation induced specific damage to TbCatB

The TbCatB model has six disulphide bonds between cysteine residues, as shown in the table 6.1 and in the figure 6.1. Disulphide bonds and carboxyl groups of acidic residues (Glu, Asp) are particularly sensitive to radiation induced damage in conventional protein crystallography. The effects of specific (loss of definition of electron density in sensitive parts of the protein) and global (Bragg spots vanishing over time, starting from high resolution) radiation damage in conventional protein

Table 6.1: *Disulphide bonds in TbCatB molecule. Indicators of local radiation damage.*

Number	First of a pair	Second of a pair
1	CYS 107	CYS 136
2	CYS 119	CYS 162
3	CYS 154	CYS 215
4	CYS 155	CYS 158
5	CYS 184	CYS 219
6	CYS 192	CYS 205

crystallography at synchrotrons are caused by photo-ionization events and following chemical changes that occur in the sample under exposure. Cryo-cooled samples withstand larger radiation dose than at room temperature. For details see section 5.3 in chapter 5.

The mechanism of radiation damage to protein crystals during synchrotron measurements is described in [O'N02]. Study about specific radiation damage to in protein crystals is presented in [Wei00] [Fio07]. The effects of radiation induced damage to protein crystals at synchrotrons take place on a much longer time scales than in SFX. In SFX the intense radiation ionizes the atoms of the specimen. The resulting photoelectrons, Auger electron cascades and the secondary ionization processes may influence the quality of the retrieved electron density map by modification to the electronic structure of atoms in the crystal. This effect could be visible in the electron density and mostly around disulphide bonds in the TbCatB electron density map, due to their higher cross section for absorption and ionization than other atoms in the structure.

In order to check if TbCatB electron density map obtained from micro-crystals at room temperature from pulses orders of magnitude brighter than from any synchrotron available presents signs of specific damage, visualization and calculation of volumes of electron density around disulphide bonds is presented. TbCatB crystals were inaccessible for synchrotron studies due to small size and sensitivity to conventional radiation damage processes. Maps obtained using SFX approach at X-ray FEL do not present any observable specific radiation induced damage. This may be due to the low ionization levels during the exposure, as predicted by CRETIN simulations presented in previous chapter.

In figure 6.2 a comparison of measured and simulated electron density maps is presented. Figures and calculations of volumes and contour levels were made using the "UCSF Chimera" program. The experimental electron density map (yellow surface) in that figure was set to show the surface contoured at 0.8306 $e^{-}/Å^{3}$ which corresponds to contour level of 4.53 σ. The simulated electron density map (blue

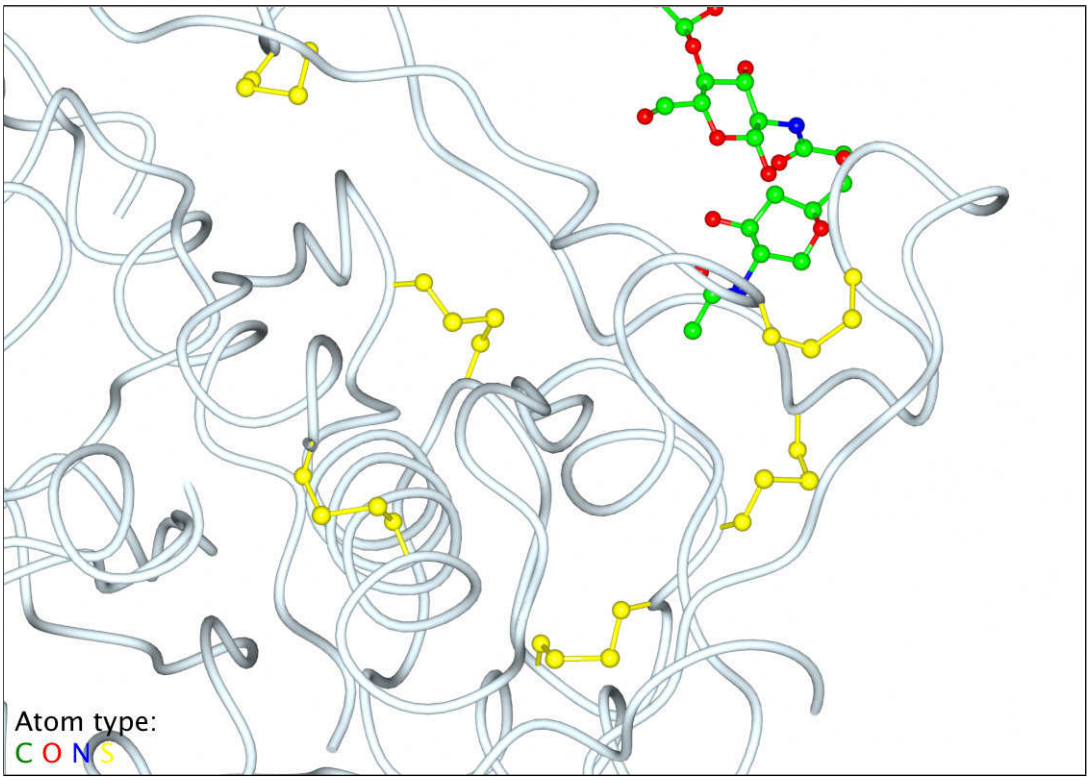


Figure 6.1: *TbCatb model with disulphide bonds. Carbon atoms backbone of the structure is shown in white, disulphide bonds are presented in yellow ball and stick representation, carbohydrate chain is located in the upper right corner also in ball and stick representation. Atom colour legend is located in the lower left corner.*

surface) was set to show the surface contoured at 1.1985 $e^-/Å^3$ which corresponds to contour level of 4.56 σ. The contour range for the experimental map is <-0.68, 1.22> $e^-/Å^3$. The contour range for the simulated map is <-0.206, 1.76> $e^-/Å^3$. Both maps are comparable at the same contour level, which means that there are no signs of radiation damage in the experimental data around disulphide bonds at this resolution. Simulations of the electron density maps were performed by standard crystallographic software used in X-ray crystallography using the known TbCatB model.

Additionally, maps used for calculation and displaying contain the same volume of space – asymmetric unit of TbCatB crystal. The volume of the experimental map is 41.34 $Å^3$ and of the simulated map is 52.92 $Å^3$. The area of the experimental map is 186.9 $Å^2$ and of the simulated map is 274.4 $Å^2$. The simulated map has larger volume and area at the same contour level as measured map, because other residues

like methionine and cysteine also display some electron density around sulphur atoms at this sigma level and resolution, whereas experimental map doesn't. It is depicted in figure 6.3.

Carboxyl (acidic) residues may also be affected by radiation damage in conventional X-ray crystallography at synchrotrons, for example ASP, HIS, ASN, GLU, GLN residues. The mechanism of radiation damage to acidic residues is associated with water radiolysis and creation of free radicals. TbCatB electron density doesn't present any observable damage to acidic residues. This could be explained in such way that on the time scale of the X-ray pulse duration there was no such radiation mechanism present in our experiment.

Now I will concentrate on only one randomly selected disulphide bond, the Cys-154 – Cys-215. Figure 6.4 presents comparison of two electron densities around this selected disulphide bond. There is no difference between measured and simulated densities, contoured at the same $\sigma = 1$ level, indicating that there was no specific damage to that disulphide bond observed at this resolution.

To check if there is no bias from the model in the electron density maps around disulphide bonds ($2F_{obs} - F_{calc}$), I mutated all cysteine residues involved in disulphide bonds to alanine residues and performed rigid body refinement. Result is displayed in the figure 6.5. The difference ($F_{obs} - F_{calc}$) electron density shows up as positive (green) mesh indicating that there was no bias from the model during refinement of disulphide bonds described above.

6.2 Errors in Monte Carlo merging of structure factors

In serial femtosecond crystallography a Monte Carlo approach of averaging thousands of measurements of the same partial reflection from many crystals of similar size but different orientation is performed. One of the main experimental sources of errors comes form the fact that crystals are flowed across the X-ray beam in a water jet. This introduces fluctuating amount of water background scattering to diffraction patterns. The so called "water ring" is apparent in every snap-shot diffraction pattern and influences the analysis. Incorrect treatment of fluctuating amount of background introduces one of the biggest errors in the extracted structure factors.

Background scattering from water is different shot-to-shot because crystals have a variable size distribution. We assume that the variation in the thickness of the liquid jet itself is less contributing to the amount of water background than the variation in the size of crystals. Because the water jets that were imaged by the ESEM performed very stable and produced little variation in the jet thickness. Therefore the amount of background in diffraction patterns fluctuates because of the different size of the crystals that displace different amounts of water in the jet for every single exposure and introduces errors to structure factors after merging.

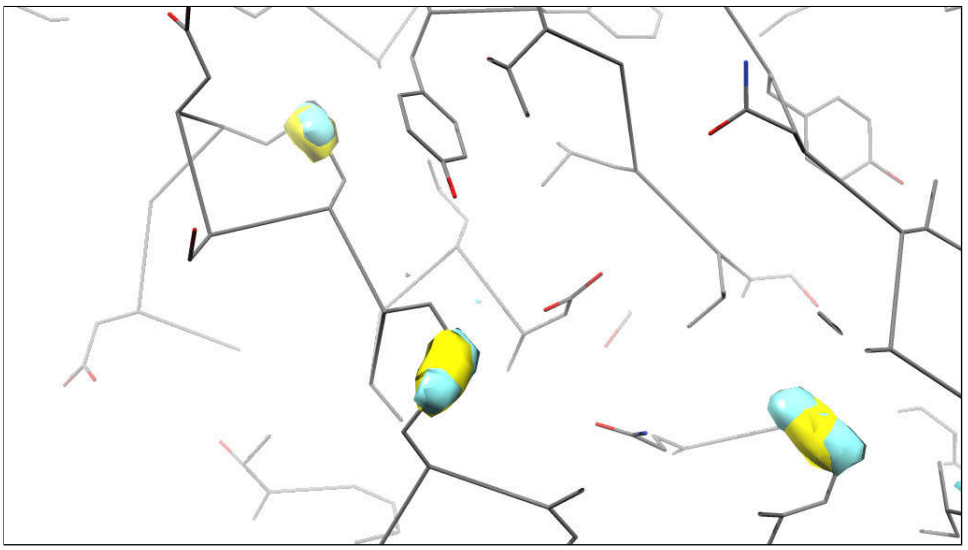


Figure 6.2: *Disulphide bonds electron density of TbCatB. Electron density is represented as opaque surface coloured yellow (experimental data) and blue (simulated data). Contour of the maps is at the same relative level. From left to right Cys-155 – Cys-158; Cys-154 – Cys-215; Cys-184 – Cys-219.*

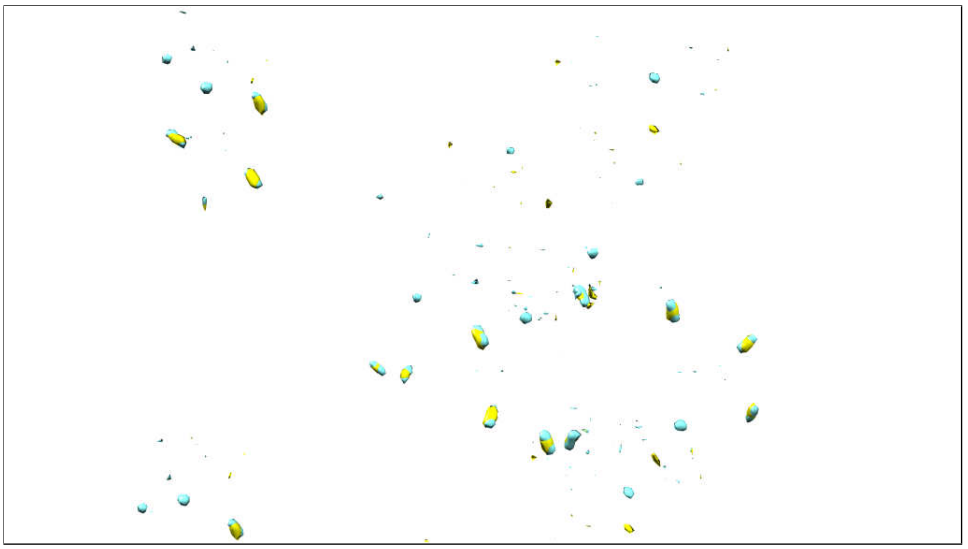


Figure 6.3: *Electron density of TbCatB asymmetric unit. Electron density is represented as opaque surface coloured yellow (experimental data) and blue (simulated data). Contour of the maps is at the same relative level of 4.5 σ.*

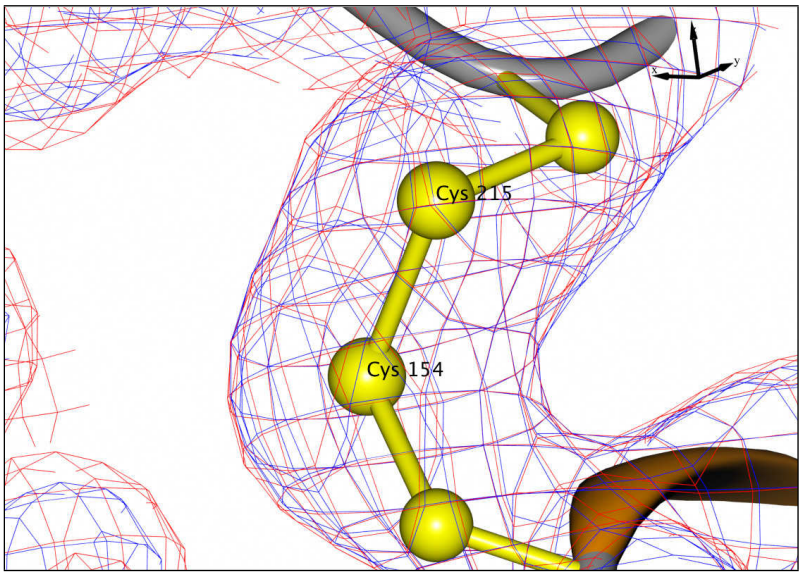


Figure 6.4: *Disulphide bond between Cys-154 and Cys-215 residues. Blue mesh represents the measured data, red mesh represents the simulated data. Contour level at 1 σ. Both experimental and simulated are very similar.*

In order to clean diffraction patterns from disturbing water background a median filter is applied during the hit finding step in *Cheetah*. The value of every pixel in an image is reduced by a median calculated from values of surrounding pixels within the radius of two pixels, see Appendix A. The median filter is better than the mean filter because pixels with high number of counts due to Bragg peaks do not affect the median as much as they affect the mean. Resulting diffraction images have the background suppressed while the Bragg peaks are maintained. This procedure also helps to reduce the number of false positive hits because Bragg peaks become better recognizable for the hit finding algorithm. Indexing of clean diffraction patterns is also improved in comparison to noisy patterns. The resulting intensities obtained from diffraction patterns subjected to background cleaning procedure have very similar values to those without background subtraction but the signal to noise ratio is improved when the background subtraction is applied, see the tables 6.2 and 5.2 for comparison.

For the analysis presented in table 6.2, the analysis of the whole TbCatB data set was reprocessed without the median background filter in the hit finder step. Hit finder step is necessary in order to extract hits from non-hits and to save individual cleaned images. For that case with background in the images the statistics of the merged data set are worst than in tables 5.1 and 5.2 from chapter 5 where the median background cleaning filter was applied. In particular the number of hits found is

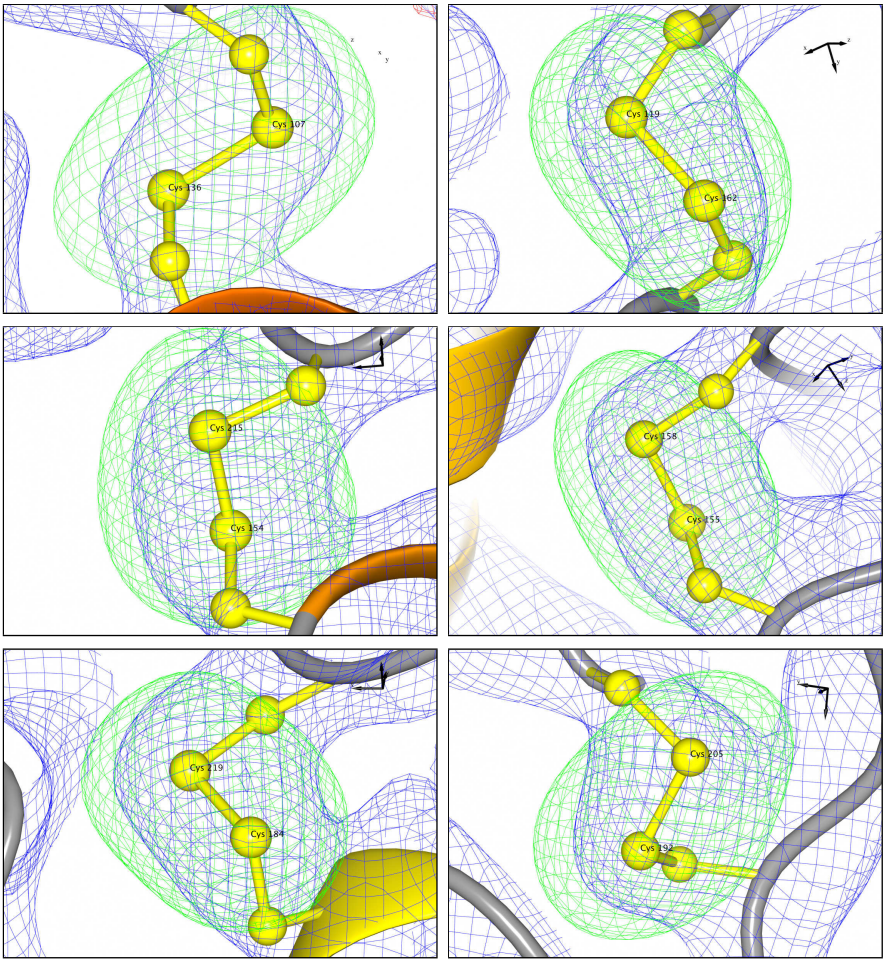


Figure 6.5: *All six disulphide bonds from the TbCatB model with electron densities. Cysteine residues were mutated to alanines to avoid bias in the electron density from the model. Difference ($F_{obs} - F_{calc}$) electron density shows as green (positive) mesh where there should be sulphur atoms. Blue mesh is the electron density of the measured data ($2F_{obs} - F_{calc}$) contoured at 1 σ, green mesh is contoured at 3 σ.*

Table 6.2: *Statistics on the TbCatB data set without median background subtraction. For comparison to the background subtracted data set see table 5.2.*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.509	3553.1	26.0
4.509 – 3.585	3824.9	23.9
3.585 – 3.134	3928.5	18.8
3.134 – 2.848	3642.5	11.9
2.848 - 2.645	3660.6	7.4
2.645 - 2.489	3755.4	5.2
2.489 - 2.365	3901.9	4.0
2.365 - 2.262	3781.9	2.9
2.262 - 2.175	3528.7	2.4
2.175 - 2.100	3264.3	1.3

lower and number of indexed patterns is also lower by approximately half in the non-subtracted background case. This can be attributed to sensitivity of our hit finding algorithms to water background in the diffraction patterns. There were 174,019 crystal hits of which 84,624 could be indexed for the data set without background subtraction. Less crystal hits were found due to larger background oscillations that suppress the peaks for hit finder routine. Less patterns were indexed for the same reason. Merged $I/\sigma(I)$ in resolution shells is lower than in data set of only 59 thousands indexed patterns taken at random from the cleaned, final TbCatB data set used to solved the structure, see table 6.10 meaning that non-cleaned background introduced more noise than in the cleaned case. This gives a hint that when the peaks are weak at high resolution median background subtraction improves the signal to noise ratio. Also R factors from refinement, when using the same model and rigid body refinement, are worst in the "non-cleaned background" case (R factor = 0.209, R free = 0.2371) than in the final refinement using cleaned diffraction patterns, which can be also attributed by lower signal to noise ratio in the non-cleaned case.

To conclude, the median filter applied to diffraction images improves the number of identified diffraction patterns that contain Bragg diffraction signal, improves indexing rate, signal to noise statistics and R factors from refinement.

The metric that is used in SFX and shows the Monte Carlo integration convergence is the R_{split} factor. It is defined like

$$R_{split} = 2^{-1/2} \frac{\sum |I_{even} - I_{odd}|}{\frac{1}{2} \sum (I_{even} + I_{odd})}, \tag{6.1}$$

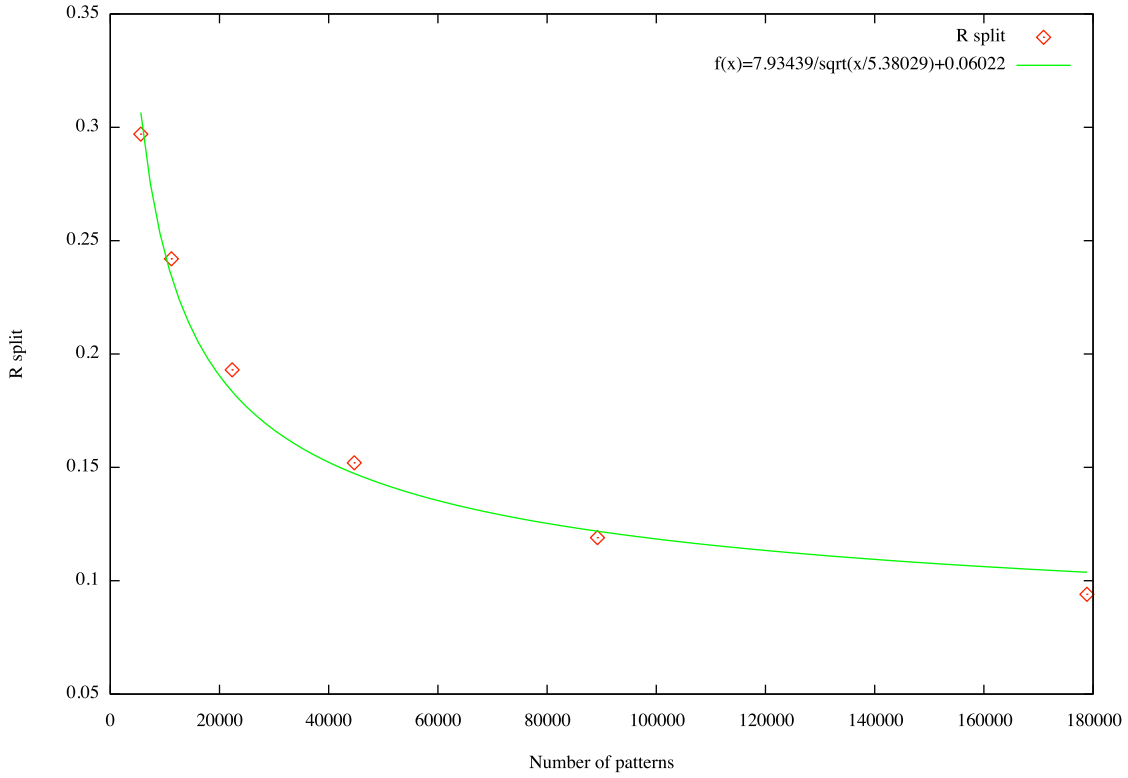


Figure 6.6: *R_{split} in the function of the number of indexed patterns. Red points present the experimental values of R_{split} for TbCatB crystals and the green line represents the fit to the data of the function $f(x) = a/\sqrt{(x)}$ where a is the parameter.*

where I_{even} represents the intensity of a reflection produced by merging even-numbered patterns, I_{odd} represents the intensity of the equivalent reflection from the odd-numbered patterns and the sum is over all reflections [Whi12].

In the figure 6.6 the R_{split} factor is presented in the function of the number of indexed patterns. In general the error in approximation follows $1/\sqrt{N}$ dependence in Monte Carlo sampling, where N is the number of samples in the integration.

Other sources of errors are described below. The SASE X-FEL radiation generation process and the resulting properties of the X-ray pulses also may influence the resulting data quality. The duration and the mean of the wavelength distribution of X-ray pulses changes from pulse to pulse. Additionally each pulse has a different distribution in the spectrum and the intensity varies for different wavelengths in the pulse.

Detector related errors also contribute to the resulting data quality. The detector has a response characteristics which become non linear at high photon signals. This

can be corrected if the non-linear response function is known. Also incorrect detector geometry, for example a tilt in the horizontal or vertical direction may introduce errors in the resulting data quality.

Other source of errors may come from crystal defects or non-isomorphism between crystals. Historically crystals were called isomorphous if they had closely similar shapes. Shape was defined by measuring angles between crystal faces using goniometer. Nowadays crystals are called isomorphous if they belong to the same space group and have the same unit cell parameters and their positions of atoms are the same. Flash-cooling procedures used in synchrotron MX studies may change crystal dimensions [Jue01][Hal04], therefore reduce the success in merging data sets from different cryo-cooled crystals. Room temperature crystals are better for structure solution due to lower isomorphism [Dun05]. Most of the protein crystals remain the same at room temperature and can be used for multi crystal data collection (Blundell and Johnson, "Protein crystallography", 1976). This fact works for our advantage because we merge data collected from many thousands different crystals at room temperature. However, as apparent in following sections, TbCatB crystals have a variation in their unit cell dimensions that affects the quality of the data set.

6.3 Effects of TbCatB unit cell parameters variation

The TbCatB data set is unique in terms of number of crystals used in the analysis. No records exists in the history of macromolecular crystallography of such large number of single crystals of the same protein used for one study. The fact that TbCatB data set consists of 178,875 single crystal diffraction patterns enables the analysis of distribution in unit cell parameters. Non-isomorphic crystal classes could be selected based on the information about unit cell parameters and cluster analysis. Preferred orientation for the alignment of needle-like particles flown in the water jet can be obtained with full orientation information coming from indexing every individual crystal diffraction pattern. Measurements of the average crystalline domain size based on the width of the Bragg spots and the virtual powder patterns can be performed and statistical information extracted. Finally the data set can be divided into subsets depending on the unit cell parameters distribution where outliers are excluded from averaging, or depending on the strength of diffraction signal, resolution, etc. Effects of division into smaller subsets can be verified in terms of goodness factors from protein model refinement.

Merging one data set of structure factors from distinct protein crystals has been found to be beneficial in terms of the quality of the resulted data set if the crystals were identical [Gio12]. This is particularly important for the anomalous dispersion data sets, where *ab initio* structure solution depends strongly on the quality of the data set which often can be completed only using multiple crystals in conventional

MX data collection techniques. In previous section, small comments were made about the isomorphism of different crystals composed of one protein type. Room temperature crystals were found to be more isomorphic than the ones that were flash-frozen. Cryogenically cooled protein crystals become different to each other due to differences in crystal size and cooling rates for that crystals. Cryo-cooling in many cases introduce non-isomorphism. It has been observed that non-isomorphism in cryo-cooled crystals reduces the quality of merged data set.

6.3.1 Variation of TbCatB unit cell parameters

In this section the mean and standard deviation values for TbCatB unit cell parameters obtained by indexing 178,875 single crystal diffraction patterns are presented. TbCatB crystals belong to the $P4_22_12$ space group therefore the crystal lattice type is tetragonal with the unit cell lengths $a = b \neq c$ and angles $\alpha = \beta = \gamma = 90°$. The mean unit cell parameters and their relative errors are presented in table 6.3. Histograms of TbCatB unit cell parameters are presented in figure 6.7. The raw unit cell parameters were obtained by indexing every single-shot TbCatB diffraction pattern recorded during the experiment without previously knowing unit cell parameters of TbCatB crystals. Those unit cells are the raw unit cells that were found by the autoindexing tools used for indexing. The wavelength that was used for indexing was measured on per shot basis and this information was used for indexing each diffraction pattern. For comparison, this table presents the relative errors of unit cell parameters calculated from indexing 8,666 diffraction patterns of tetragonal hen egg lysozyme single crystals. Data collection on lysozyme crystals was performed during the same experiment as TbCatB. Relative errors in estimation of the unit cell parameters from indexing TbCatB and lysozyme diffraction patterns are different, which suggests that the main contribution to these errors comes from differences of unit cell dimensions across many crystals of the same kind. Experimental errors that may contribute to the determination of the unit cell parameters are: incorrect detector geometry (tilt in horizontal or vertical direction) or the change in the sample to detector distance due to movement of the liquid jet during the experiment or the finite pixel size.

The expected error in the estimation of the unit cell lengths due to the finite pixel size can be estimated from:

$$\begin{aligned} \lambda &= 2d \sin\theta \\ d &= \frac{\lambda}{2\sin\theta} \\ \frac{\Delta d}{d} &= \frac{\Delta\theta}{\tan\theta} \end{aligned} \tag{6.2}$$

where $\Delta\theta = P\cos 2\theta / L$ is the angular difference due to the pixel size. In our case

Table 6.3: *Mean values and standard deviations of TbCatB unit cell parameters measured from 178,875 TbCatB micro-crystals with relative errors. Last column presents relative errors in the unit cell parameters determined for lysozyme protein crystals during the same experiment.*

Parameter	Mean	St. dev.	Error %	Error (lysozyme) %
a (Å)	126.12	0.62	0.49	0.26
b (Å)	126.64	0.68	0.53	0.25
c (Å)	54.32	0.23	0.42	0.30
α (°)	90.01	0.43	0.47	0.16
β (°)	89.98	0.36	0.40	0.17
γ (°)	90.00	0.43	0.47	0.16

the pixel size was P=0.11 mm and the distance between the detector and the liquid jet was L=127 mm. The 2θ at the resolution of 2.5 Å is equal to 15.3 degrees. The relative error $\Delta d/d = 0.62\%$.

6.3.2 Data points and Principal Component Analysis

In order to reveal if there is any special subset of TbCatB crystals that has unit cell parameters systematically different to other crystals in the whole TbCatB data set, we could look at plots of (a,b,c) and (α, β, γ) unit cell parameters or use Principal Component Analysis (PCA), a generic approach used to deduce relationships among multiple datasets. See figures 6.8, 6.9, 6.10 and 6.11.

In order to calculate principal components I performed Singular Value Decomposition (SVD) on TbCatB unit cell parameters. In SVD, a $m \times n$ real or complex matrix A is decomposed into $A = USV^T$, where U is a $m \times n$ real or complex unitary matrix whose columns are called left singular vectors of A, S is a $m \times n$ diagonal matrix with non negative entries σ_{ii} called singular values of A, ordered on the diagonal with decreasing magnitude, and V^T is a conjugate transpose V, which is a $n \times n$ real or complex unitary matrix whose columns are called right singular vectors of A.

Singular values obtained after SVD performed on unit cell lengths and angles separately are presented in the first column of the table 6.4 and 6.6 respectively. These values indicate the variance of the linearly independent components along each dimension. A normalized eigenvalue will indicate the percentage of total variance. This is shown in the second column of the tables 6.4 and 6.6 as the cumulative sum of percentage contributions for each component.

The matrix V^T from SVD analysis is presented in tables 6.5 and 6.7 for lengths

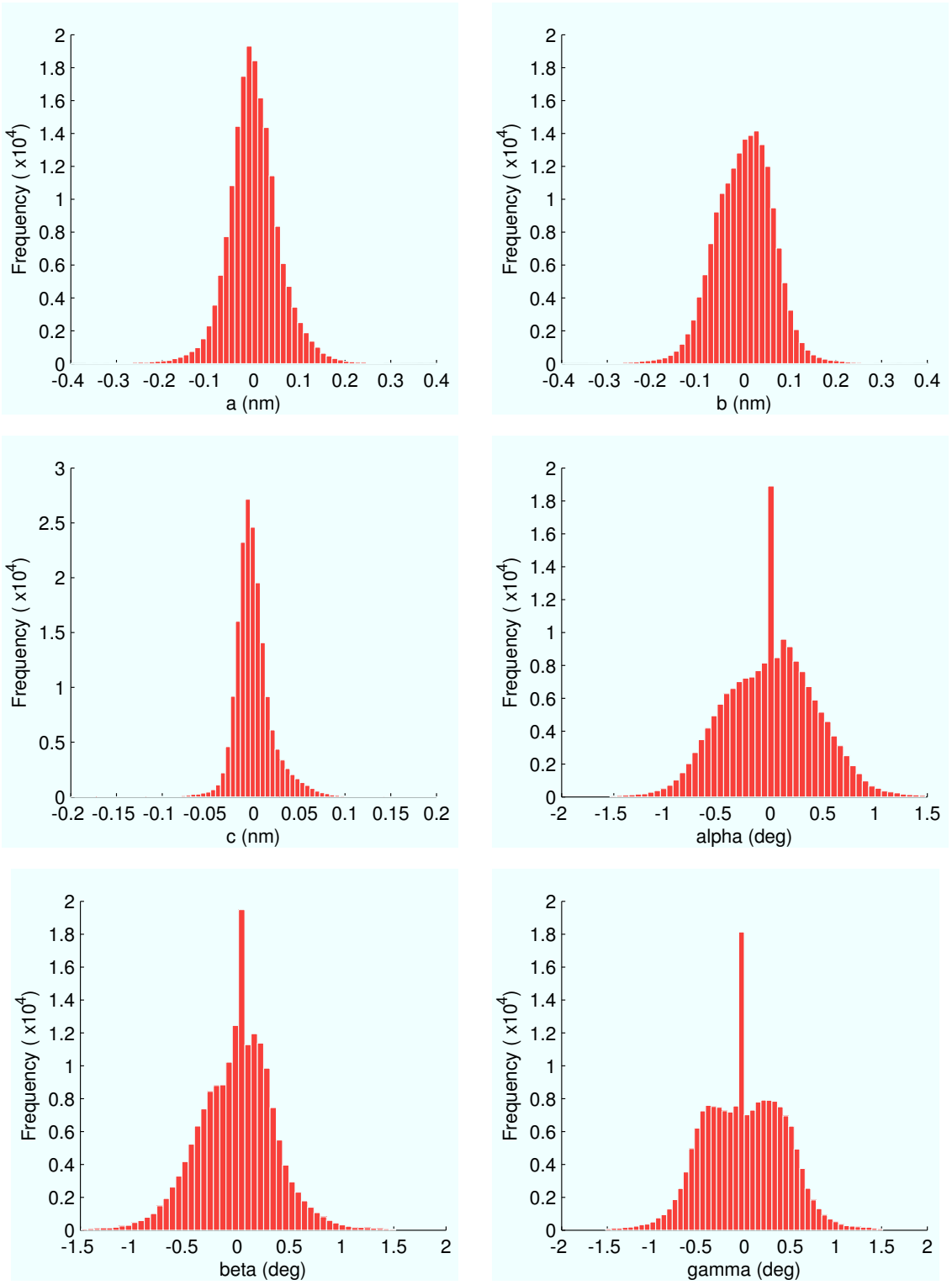


Figure 6.7: *Histograms of TbCatB unit cell parameters measured from 178,875 TbCatB micro-crystals. Mean values were subtracted from every measurement.*

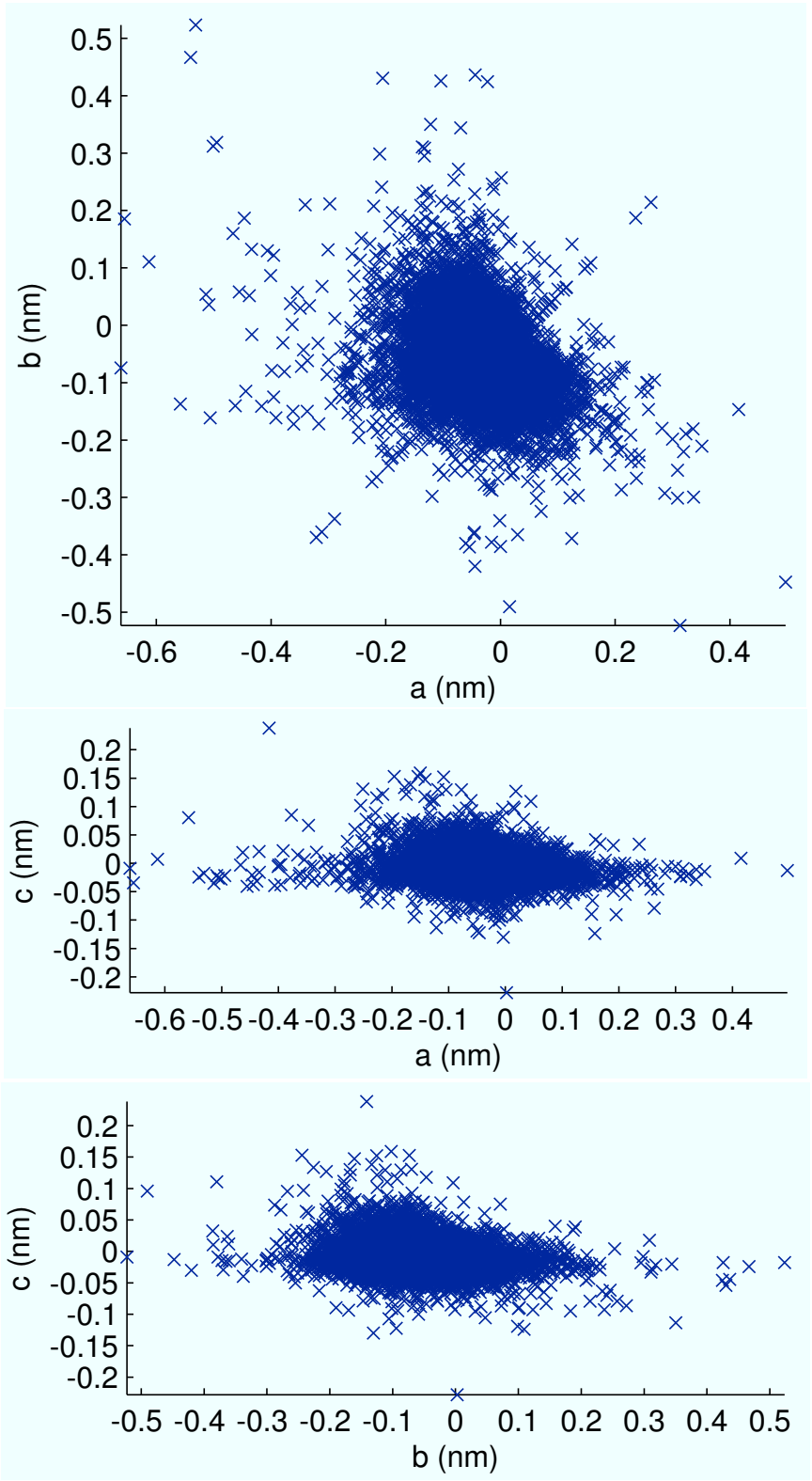


Figure 6.8: *Data points of 178,875 TbCatB a,b,c unit cell parameters, axis scales in (nm). Mean values were subtracted from every measurement.*

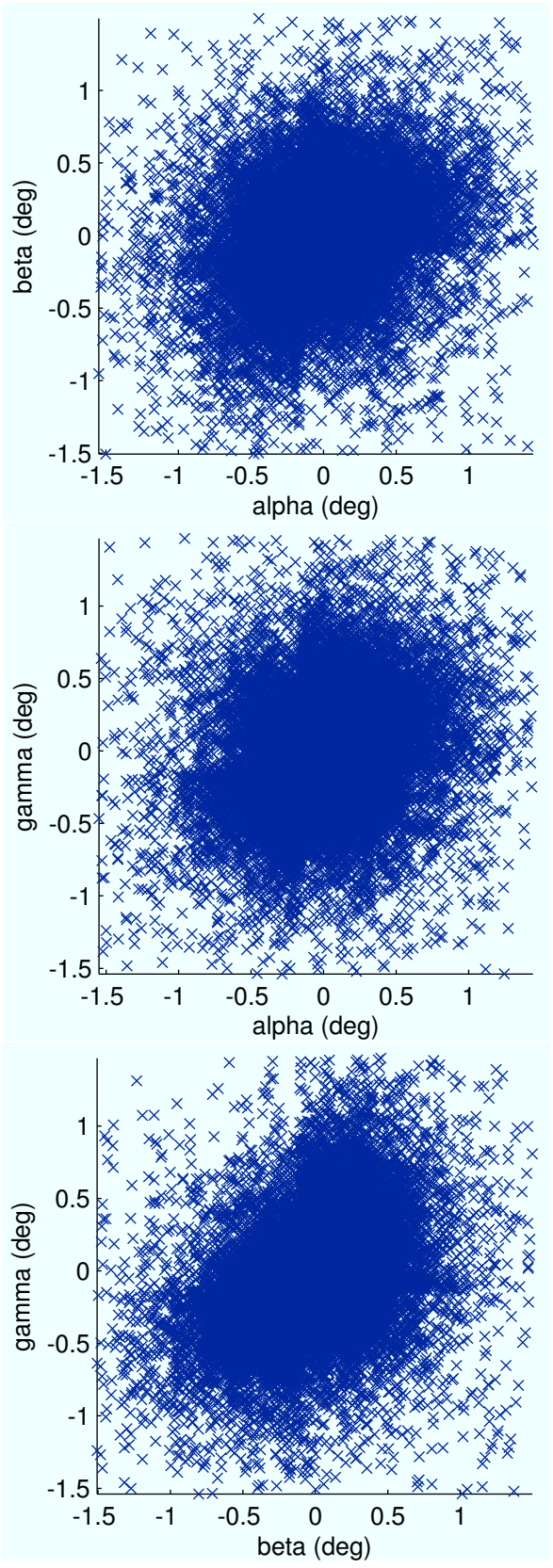


Figure 6.9: *Data points of 178,875 TbCatB α, β, γ unit cell parameters, axis scales is in ($^\circ$). Mean values were subtracted from every measurement.*

Table 6.4: *Singular values from SVD on a,b,c unit cell parameters.*

Singular values	Cumulative percentage
24.3502	0.5876
19.1067	0.9493
7.1526	1.0000

Table 6.5: V^T *matrix from SVD on a,b,c unit cell parameters*

0.5564	0.8304	-0.0290
0.8301	-0.5570	-0.0262
0.0379	0.0095	0.9992

Table 6.6: *Singular values from SVD on* α, β, γ *unit cell parameters.*

Singular values	Cumulative percentage
168.6687	0.5872
107.1565	0.8242
92.2925	1.0000

Table 6.7: V^T *matrix from SVD on* α, β, γ *unit cell parameters.*

-0.6009	-0.7449	0.2899
-0.5251	0.0944	-0.8458
-0.6027	0.6604	0.4479

and angles respectively. Singular vectors are in rows of those matrices. The data points of a,b,c and α, β, γ projected on principal components is presented in figures 6.10 and 6.11.

The plots with data points and especially those that present projections of data on singular vectors show that any specific subset of unit cell lengths or angles is apparent.

6.3.3 R-factor analysis of sub sets of TbCatB data set

The whole TbCatB data set was divided into smaller subsets of indexed diffraction patterns basing on the information of unit cell lengths obtained from indexing. The subsets are characterized by gradually smaller variance of unit cell lengths from their mean values. For comparison the same number of indexed patterns was chosen

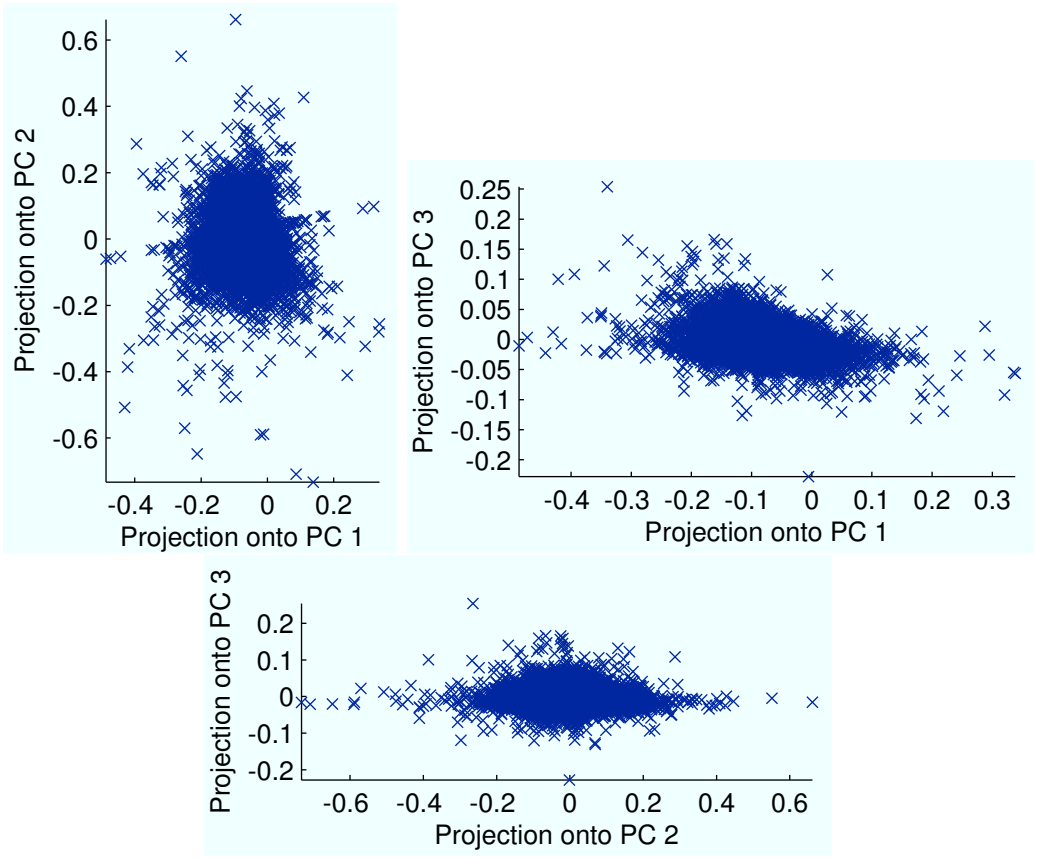


Figure 6.10: *PCA of a,b,c unit cell parameters, scale in (nm). Mean value was subtracted from each dimension.*

randomly from the whole data set.

Crystals of the same kind that have similar unit cell parameters should have the most isomorphic internal structure, therefore the merged structure factors according to the Monte Carlo method of integration, from subsets of patterns that were indexed with similar unit cell parameters should have higher quality than the ones where all crystals were included in the integration. Later, by using the final TbCatB model for rigid body refinement on subsets of data, we can obtain information on how good the data sets are by comparison of values of R factors.

A set of indexed diffraction patterns with values of the unit cell lengths very close to mean values consist of 14,363 patterns. A set of indexed diffraction patterns with values of the unit cell lengths with larger variation from mean values consist of 52,238 patterns. Those subsets were chosen basing on the following criteria: In the program *indexamajig* from the CrystFEL software suite, one of the user defined parameters

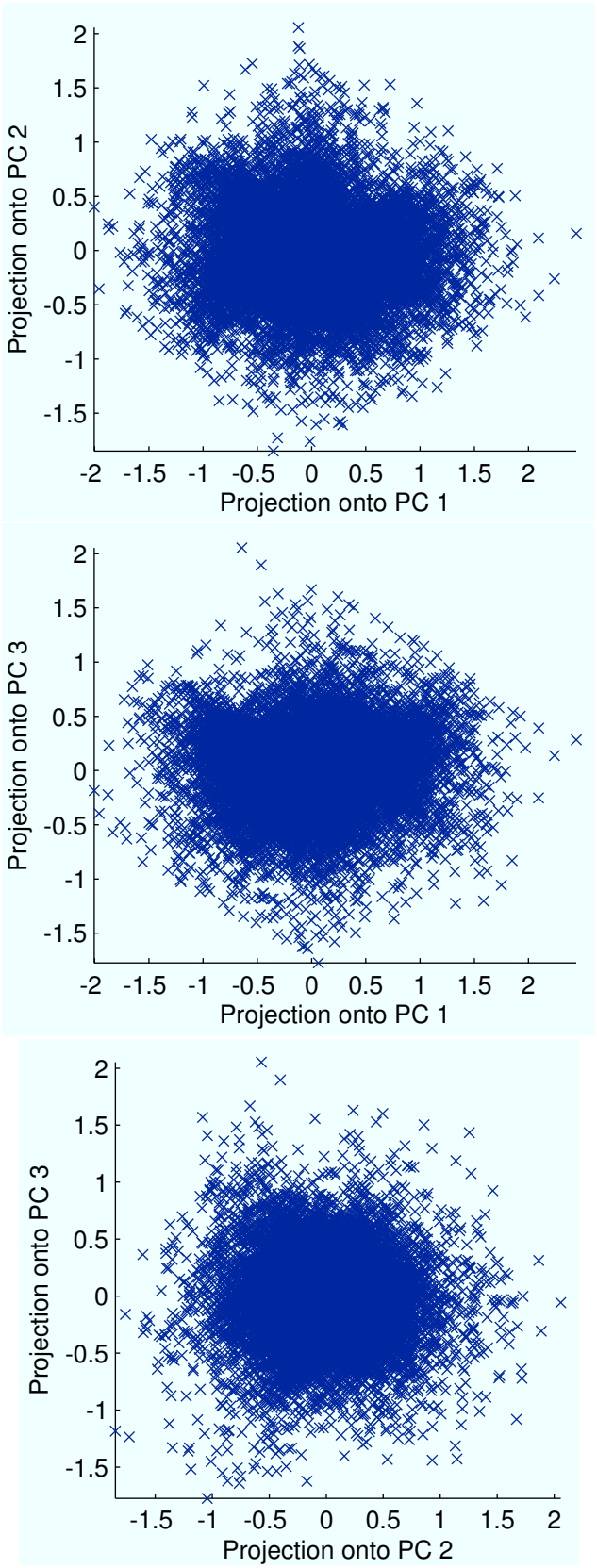


Figure 6.11: *PCA of α, β, γ unit cell parameters, scale in ($^\circ$). Mean value was subtracted from each dimension.*

sets the percentage of allowed variation of the reciprocal unit cell lengths when comparing the found unit cell parameters after indexing the diffraction pattern to the set of parameters provided by the user. In this case the given unit cell parameters were set to be the mean values of the unit cell parameters obtained from histograms from the figure 6.7).

In order to compare the quality metrics (R factors from refinement) of those subsets I selected the same number of patterns at random from the whole TbCatB data set and performed rigid body refinement to calculate the R and R free factors and also calculated the redundancies and signal to noise ratios in ten resolution shells using the CrystFEL software suite.

The whole TbCatB data set consists of 178,875 indexed diffraction patterns. The R factor and R free for this data set equal 18.10 % and 21.41 % respectively. This set has been indexed using tolerances for change in the reciprocal unit cell parameters of 5% for reciprocal lengths and 1.5 degrees for angles. Redundancies and signal to noise levels for this data set are presented in the table 6.8.

Subset that consists of 52,238 diffraction patterns was selected by reducing the tolerance for change in the reciprocal unit cell parameters to 0.5% for reciprocal lengths and 0.15 degrees for angles. Redundancies and signal to noise levels for this data set are presented in the table 6.9. After refinement I obtained this R factors: 19.71 %, R free: 22.27 %. In order to compare those values I selected at random similar number (52,150) of patterns from the whole data set and performed refinement using the same model and refinement procedure. With this data set I obtained R factor of 20.02 % and R free of 22.51 %. Redundancies and signal to noise levels for this data set are presented in the table 6.10.

I have reduced the tolerance for change in the reciprocal unit cell parameters even further, to 0.3% for reciprocal lengths and 0.15 degrees for angles. The resulting subset of TbCatB diffraction patterns consists of 14,363 indexed patterns. Redundancies and signal to noise levels for this data set are presented in the table 6.11. Refinement on this subset gave that R factors: 20.34 % and R free 23.85 %. The same number of patterns (14,439) taken at random from the whole data set gives that R factor: 22.03 % and R free 24.77 %. It is important to note that the subset of data with unit cell parameter closest to mean values contain significantly higher signal at high resolution in comparison to the subset taken at random from the full data set.

Summary of the above mentioned analysis can be found in the table 6.13. The above analysis suggests that the subset of TbCatB data merged from crystals with unit cell parameters with the smallest variance from the mean values presented here is better in terms of the R factors from refinement and signal to noise ratio than the subset of the same number of patterns selected randomly from the whole TbCatB data set. This fact also suggests that the convergence of the Monte Carlo

Table 6.8: *Quality measures of the whole TbCatB data set that consist of 178,875 indexed diffraction patterns.*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.509	7466	33.00
4.509 – 3.585	8086	27.66
3.585 – 3.134	8350	19.19
3.134 – 2.848	7659	12.09
2.848 - 2.645	7785	8.23
2.645 - 2.489	7967	5.94
2.489 - 2.365	8204	4.55
2.365 - 2.262	7910	3.59
2.262 - 2.175	7501	2.93
2.175 - 2.100	7061	2.37

integration could be reached faster by using patterns that were index with similar unit cell parameters, at least for the TbCatB case. The higher values of the mean integrated intensities presented in the last columns of tables 6.11 and 6.12 suggests that those crystals give more counts in the Bragg peaks recorded on the detector, which suggests that they are either better ordered or bigger than the ones which have unit cell parameters more distinct than the mean values. Nevertheless, the best R factors and signal to noise statistics are obtained from the whole data set which suggests that the the best strategy for SFX data collection is to use as many indexable patterns as the experiment allows to collect in order to improve the signal to noise ratio to the highest resolution that the crystals give detectable signal, which then allows to improve the quality of resulting electron density.

6.3.4 Investigation on flow alignment of TbCatB crystals

TbCatB crystals grew inside living SF9 insect cells. They acquired needle-like shapes with some distribution of sizes as presented in previous sections. The mean of that distribution of crystal sizes is $0.9 \times 0.9 \times 11.0 \mu m^3$. It was calculated from 300 measurements of width and length of TbCatB crystals performed by using SEM imaging. The fact that TbCatB crystals are have needle like shapes with the aspect ration of approximately 11, introduces the occurrence of directional alignment of crystals when flown throughout liquid jet of 4μm in diameter at the flow rate of 10μl/min.

The subset of 14,000 TbCatB diffraction patterns that was indexed using the

Table 6.9: *Quality measures of the subset of 52,238 diffraction patterns indexed with reduced tolerances (as explained in the text of this section).*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.509	2204	19.24
4.509 – 3.585	2326	16.42
3.585 – 3.134	2451	12.39
3.134 – 2.848	2227	8.00
2.848 - 2.645	2295	5.50
2.645 - 2.489	2322	3.87
2.489 - 2.365	2337	2.81
2.365 - 2.262	2246	2.22
2.262 - 2.175	2216	1.82
2.175 - 2.100	2138	1.47

Table 6.10: *Quality measures of the subset of 52 thousands diffraction patterns selected at random from the whole TbCatB data set.*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.509	2201	17.89
4.509 – 3.585	2288	14.62
3.585 – 3.134	2417	10.09
3.134 – 2.848	2280	6.55
2.848 - 2.645	2276	4.41
2.645 - 2.489	2298	3.10
2.489 - 2.365	2363	2.24
2.365 - 2.262	2304	1.77
2.262 - 2.175	2187	1.44
2.175 - 2.100	2006	1.27

Table 6.11: *Quality measures of the subset of 14 thousands diffraction patterns indexed with reduced tolerances (as explained in the text of this section). The "Mean(I)" column contain the mean values if integrated intensities from reflections that in given resolution shell. The units of the last column are analog-to-digital units (ADU).*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**	**Mean(I)**
20.000 – 4.509	619	10.42	2580.76
4.509 – 3.585	614	8.65	1495.81
3.585 – 3.134	679	6.85	659.35
3.134 – 2.848	635	4.66	274.47
2.848 - 2.645	641	3.27	136.06
2.645 - 2.489	633	2.38	70.63
2.489 - 2.365	619	1.69	41.68
2.365 - 2.262	614	1.36	30.48
2.262 - 2.175	618	1.10	21.60
2.175 - 2.100	603	0.88	16.07

Table 6.12: *Quality measures of the subset of 14 thousands diffraction patterns selected at random from the whole TbCatB data set. The "Mean(I)" column contain the mean values if integrated intensities from reflections that in given resolution shell. The units of the last column are analog-to-digital units (ADU).*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**	**Mean(I)**
20.000 – 4.509	613	9.58	1620.65
4.509 – 3.585	616	7.64	885.23
3.585 – 3.134	635	4.44	299.92
3.134 – 2.848	658	2.77	106.57
2.848 - 2.645	598	1.68	46.32
2.645 - 2.489	613	0.98	18.66
2.489 - 2.365	682	0.70	10.64
2.365 - 2.262	688	0.54	6.54
2.262 - 2.175	573	0.50	6.24
2.175 - 2.100	434	0.42	5.99

Table 6.13: *Summarized table of the analysis of subsets of TbCatB data set.*

Data set name	R(%)	R free(%)	Redundancy	Merged $I/\sigma(I)$ (at 2.1 Å)
Full TbCatB	18.10	21.41	7799	2.37
52k tolerances	19.71	22.27	2277	1.47
52k random	20.02	22.51	2262	1.27
14k tolerances	20.34	23.85	628	0.88
14k random	22.03	24.77	612	0.42

smallest tolerance for the change of unit cell parameters from mean values of the unit cell parameters distribution and the subset of the same number of TbCatB diffraction patterns selected at random from the whole TbCatB data set, see previous section, was used to investigate the flow alignment of TbCatB crystals.

The number of measurements (redundancy) per h,k,l indices was visualised in two dimensions using *render_hkl* program from CrystFEL software suite. One section from the three dimensional set of indices from both subsets of TbCatB data was chosen for displaying. This section presents the cut trough the middle of the three dimensional set of indices with "a" crystal axis normal to the plane of the paper, "b" crystal axis aligned horizontally to the paper plane and "c" crystal axis aligned vertically to the paper plane. Those sections are presented in figures 6.12 and 6.13. The colour scale has the same relative values in both images, from 0 (black) to 1 (white).

In the figure 6.12 it is possible to notice that there are missing wedges in the number of measurements per given h,k,l value. The same situation occurs in the data set selected at random from the full TbCatB data set, see figure 6.13, but the missing wedge is smaller in that case. The colour scale is equal in both figures and represents number of measurements of particular h,k,l value that was performed during indexing. It can bee seen that the random data set has lower degree of flow alignment.

In conclusion, the subset of data with unit cell parameters closer to the mean values of the unit cell parameters, comes from crystals with higher aspect ratio (long needles) that had more directional flow alignment during the X-ray exposure than the other subset of data chosen randomly from the whole data set. One can also notice that the integrated signal measured by the detector is greater in the flow aligned case, see tables with mean values of detector counts in function of resolution shells in the previous section. A clear sign of flow alignment can be seen in the virtual 2D powder pattern created by summing 2,700 subsequent diffraction patterns, see figure 6.14. Therefore, larger crystals are better flow aligned in the water jet, give stronger scattering signal but in order to measure a complete data set from needle

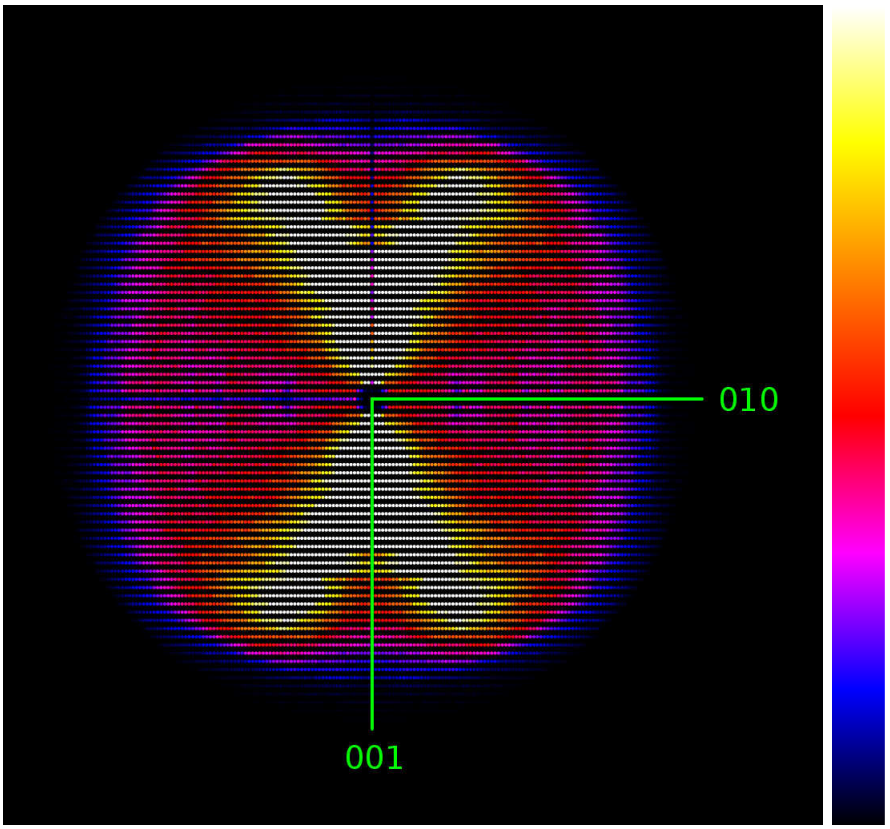


Figure 6.12: *Directional flow alignment of 14,000 TbCatB crystals from the subset of data that was indexed with unit cell parameters close to mean values of the unit cell parameters distribution obtained by indexing the whole data set.*

like crystal systems that additionally have little symmetry internal one should consider this effect. One solution could be for example to perform measurement further down from the exit of the nozzle, where the liquid jet brakes up into droplets, which could disturb the directional flow alignment obtained in the region where the thin liquid jet is formed.

6.4 Estimation of the size of ordered domain

If the assumption is made that ordering of the crystalline lattice is limited to small regions irregularly spaced, the linear dimension of the ordered domains may be estimated by the FWHM (full width at half maximum) or integral breadth of the Bragg peak. According to the Scherrer equation [Sch18, Pat39], a crystal block of dimension D produces a diffraction peak having an approximate integral breadth B

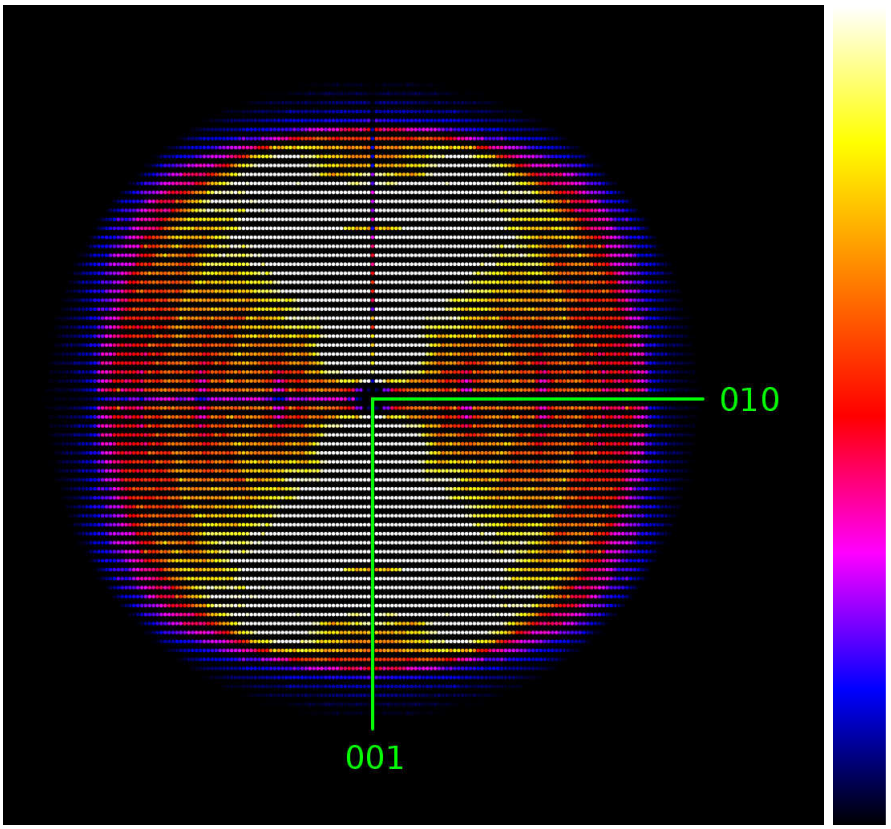


Figure 6.13: *Directional flow alignment of 14,00 TbCatB crystals with unit cell parameters selected at random from the whole distribution of TbCatB unit cell parameters. Less flow alignment is apparent in comparison to the previous figure.*

given by

$$B \cong \frac{K \cdot \lambda}{D \cdot \cos\theta_B}, \tag{6.3}$$

where K is the dimensionless shape factor typically equal to 0.96, λ is the wavelength of incident radiation, θ_B is the Bragg angle. Integral breadth B of a reflection is defined as the ratio of the peak area to the peak maximum. Integral breadth of the peak is inversely proportional to the crystal size.

Using equation 6.3 and the geometry for our TbCatB experiment (110 μm^2 pixel size, 144 mm detector distance, the beam spot size 10 μm^2 and $\lambda = 1.32$Å) I estimated the FWHM of a Bragg peak to be of the order of 1 pixel for 1 μm ordered domain size. For 0.5 μm domain size the FWHM of the Bragg peak should have the size of 2 pixels as seen in the figure 6.15.

Calculations of the integral breadths of Bragg peaks recorded during the ex-

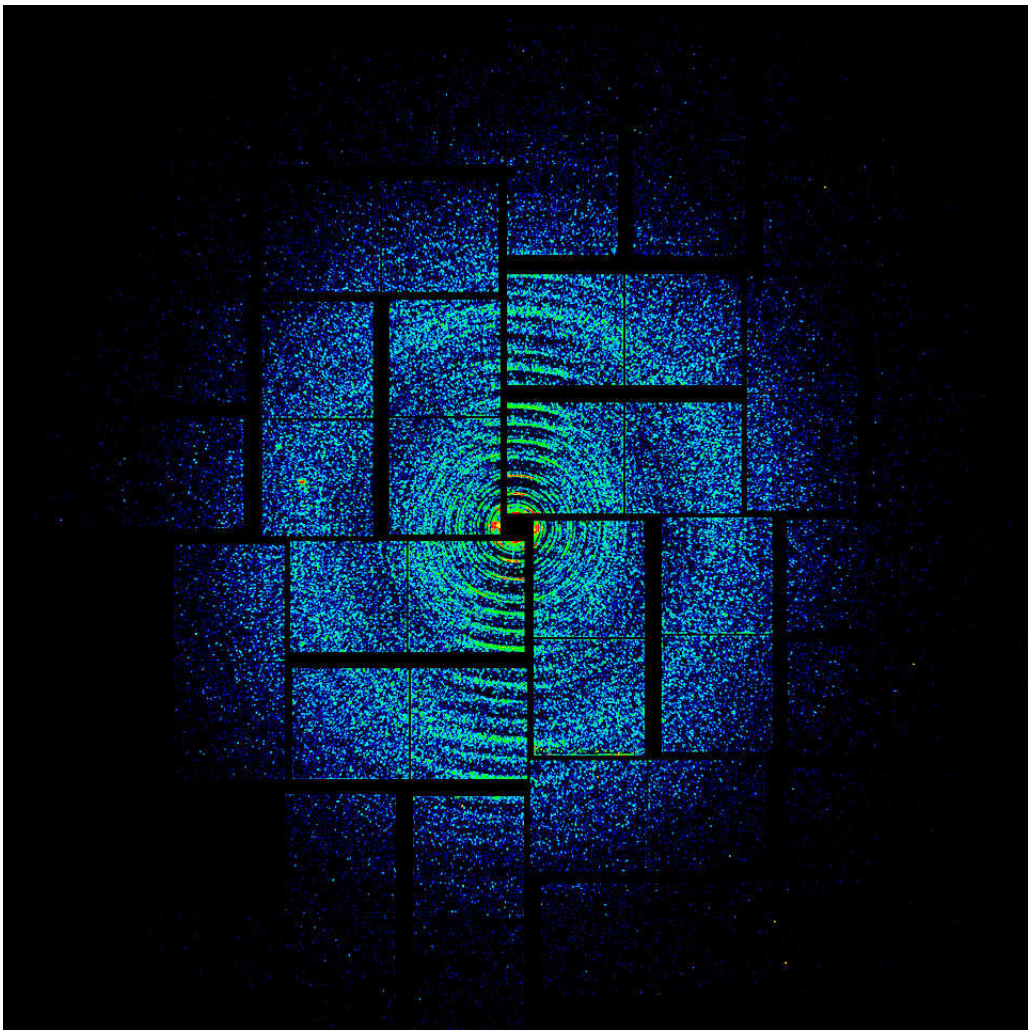


Figure 6.14: *TbCatB powder pattern created by summing 2,700 subsequent diffraction patterns. Signs of directional flow alignment are apparent in different spacings of the powder rings. This suggests that the smallest unit cell dimension along the c crystal axis is aligned along the direction of the liquid flow.*

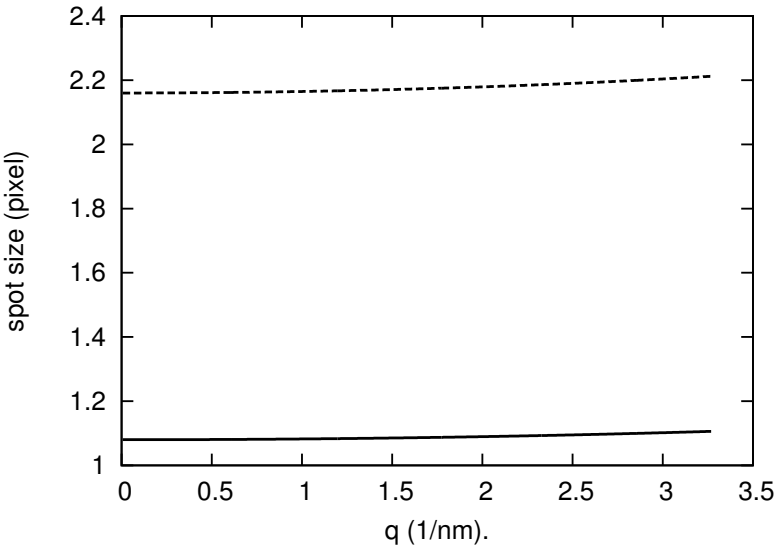


Figure 6.15: *Estimation of a Bragg peak FWHM using Scherrers formula. The Bragg peak FWHM is expressed in no. of pixels calculated for TbCatB experimental parameters in the function of the scattering vector q (1/nm). The solid line corresponds to the Bragg peak FWHM calculated for the ordered domain of the size of 1μm and the dashed line corresponds to the Bragg FWHM calculated for the ordered domain of the size of 0.5μm.*

periment were performed for 360 TbCatB diffraction patterns selected at random from the whole data set. A sum of 3,763 Bragg peaks were found and their integral breadths were calculated using modified CrystFEL routines. Resulting estimations of ordered domain size of TbCatB crystals by applying the Scherreers formula are presented in figure 6.16. The mean size of the ordered domain was estimated to be equal to 138.4 nm with standard deviation of 33.8 nm. The parameters of the Gaussian fit to the histogram presented in figure 6.17 as a green line are equal to: $A = 17904.1$ with the asymptotic standard error of ± 705.8, $\mu = 114.5$nm with the asymptotic standard error of ± 1.3, $\sigma = 28.4$nm with the asymptotic standard error of ± 1.3 and the fitted function is presented in equation 6.4. From this analysis one can deduce the expected average width of the recorded Bragg peak and then use this information to optimise the width of the region from which CrystFEL integrates the Bragg peaks. In the case of TbCatB the integration radius was set to 4 pixels.

$$Gauss(x) = \frac{A}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \tag{6.4}$$

I performed the same analysis, as described above, to other protein crystals that were investigated during other SFX experiment performed by our collaboration. This additional analysis aims to prove that by using integral breadths and the

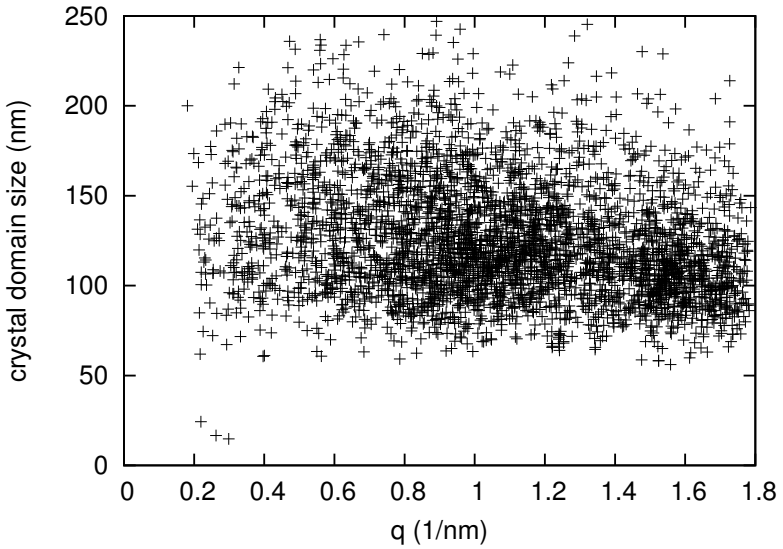


Figure 6.16: *TbCatB ordered domain size in function of the scattering vector q (1/nm), estimated from 3,763 measurements of integral breadths of Bragg peaks by using the Scherrer formula.*

Scherrer equation we can retrieve the average size of the ordered crystalline domain size with good accuracy. During that experiment we were using nano-crystals of Photosystem I protein (PSI) prepared by our collaborator. The dynamic light scattering (DLS) measurement, that is able to retrieve the average radius of a particle in aqueous solution, was performed on PSI nano-crystals in solution shortly before they were injected to the FEL beam. The size distribution of PSI nano-crystals as measured by DLS is presented in figure 6.19. The average PSI crystal size that was measured by DLS is equal to 767±117nm.

The experimental parameters during SFX data collection of PSI nano-crystals were as follows: the wavelength was set to λ =2.06Å (λ=6keV), the sample to detector distance was set to L=2.6 meters and we also used the CSPAD the detector to record diffraction patterns of PSI nano-crystals. One of the PSI Bragg spots is presented in figure 6.20.

The figure 6.18 presents analysis performed for photosystem I (PSI) nano-crystals in the same way as described above. The number of PSI diffraction patterns used in the analysis was equal to 421 and the number of measurements of Bragg peaks integral breadths was equal to 1,133. The average size of the ordered domain of PSI nano-crystals was estimated using the Scherrers formula to be equal to 902.2 nm with standard deviation of 245.7 nm. Both estimations, of the average ordered crystalline size from DLS and from measurements of integral breadths of PSI Bragg peaks, are

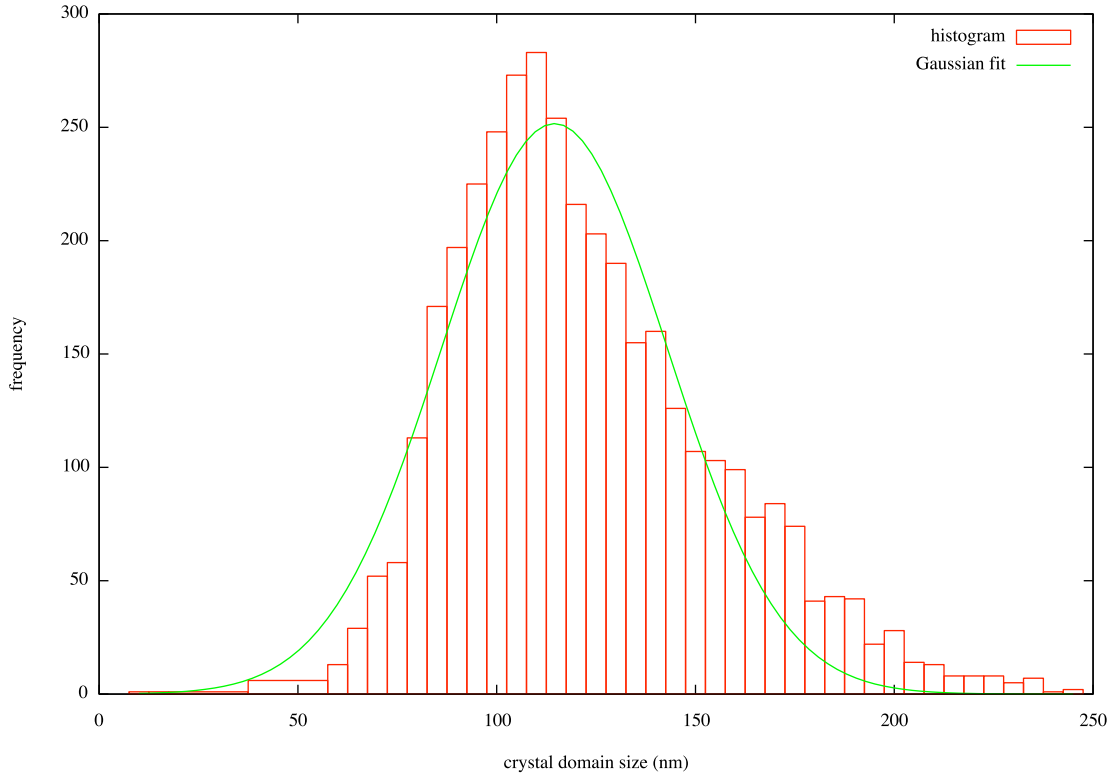


Figure 6.17: *Histogram with fitted Gaussian curve of TbCatB ordered domain sizes estimated from 3,763 measurements of integral breadths of Bragg peaks by using the Scherrer formula.*

comparable within the estimated standard deviations of the distributions. This fact may suggest that the PSI nano-crystals consist of one ordered domain whereas in TbCatB crystals the average domain size is smaller than the average dimensions of the crystals.

6.4.1 Size of ordered domain by virtual powder analysis

By adding up single crystal diffraction patterns one can achieve a diffraction pattern that is very similar to the X-ray powder diffraction pattern collected from many crystallites at once. Powder diffraction patterns contain rings which widths depend on the average ordered domain size as given previously by the Scherrer's equation.

Protein powder diffraction patterns contain many overlapping rings. In order to retrieve only one strong powder ring from the TbCatB data set I performed the following analysis. From the whole data set of indexed TbCatB diffraction patterns

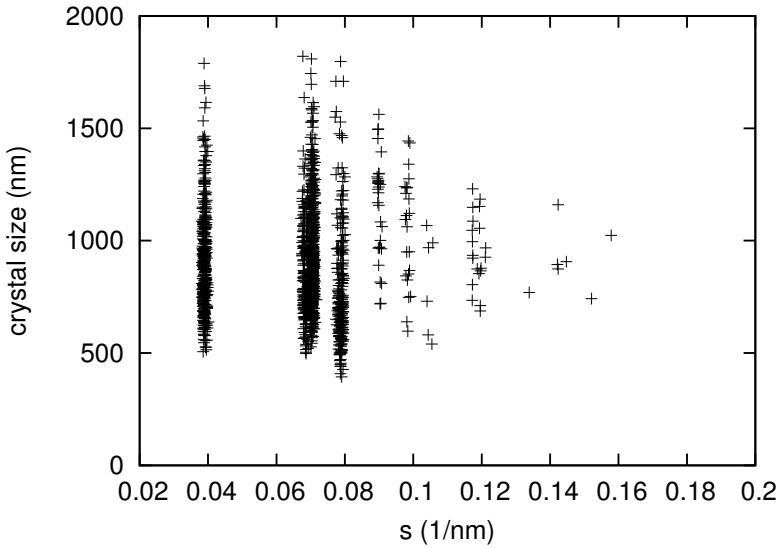


Figure 6.18: *PSI ordered domain size in function of the scattering vector q (1/nm), estimated from 1,133 measurements of integral breadths of Bragg peaks by using the Scherrer formula.*

I selected those that contain only (6,0,0) reflection and added them up, previously correcting for the wavelength fluctuations that occur on per shot basis due to the SASE effects. Wavelengths were recorded on per shot basis. There were 1,555 diffraction patterns that contained that reflection. I selected patterns with this particular reflection because reflections (5,0,0) and (7,0,0) are not occurring due to the space group symmetry, which means that this ring is very well separated from others in the virtual powder pattern. The resulting radial average of this virtual powder pattern is presented in figure 6.21 which shows one peak separated from others and having much larger intensity than the other peaks.

A magnified region of that virtual powder pattern presenting the data in the resolution range from 0.02 1/Å to 0.065 1/Å is presented in the figure 6.22. The FWHM of the (6,0,0) peak was estimated to be equal to 0.00153 1/Å. Which according to the Scheerer's formula presented above corresponds to the average ordered domain size of D=65.3 nm.

The width of TbCatB powder ring is influenced by the average ordered domain size and the variation of unit cell parameters. The equation 6.2 ($\Delta d/d = \Delta\theta/tg\theta$) can be converted to $\Delta d/d = \Delta s/s$, where $s = 1/d$ (Å^{-1}). After an easy transformation we obtain that $\Delta d = \Delta s \times d^2$. Because the width of the powder ring is a convolution of the widths of the distribution of unit cell lengths and the average ordered domain size we can estimate what is the width of the unit cell lengths

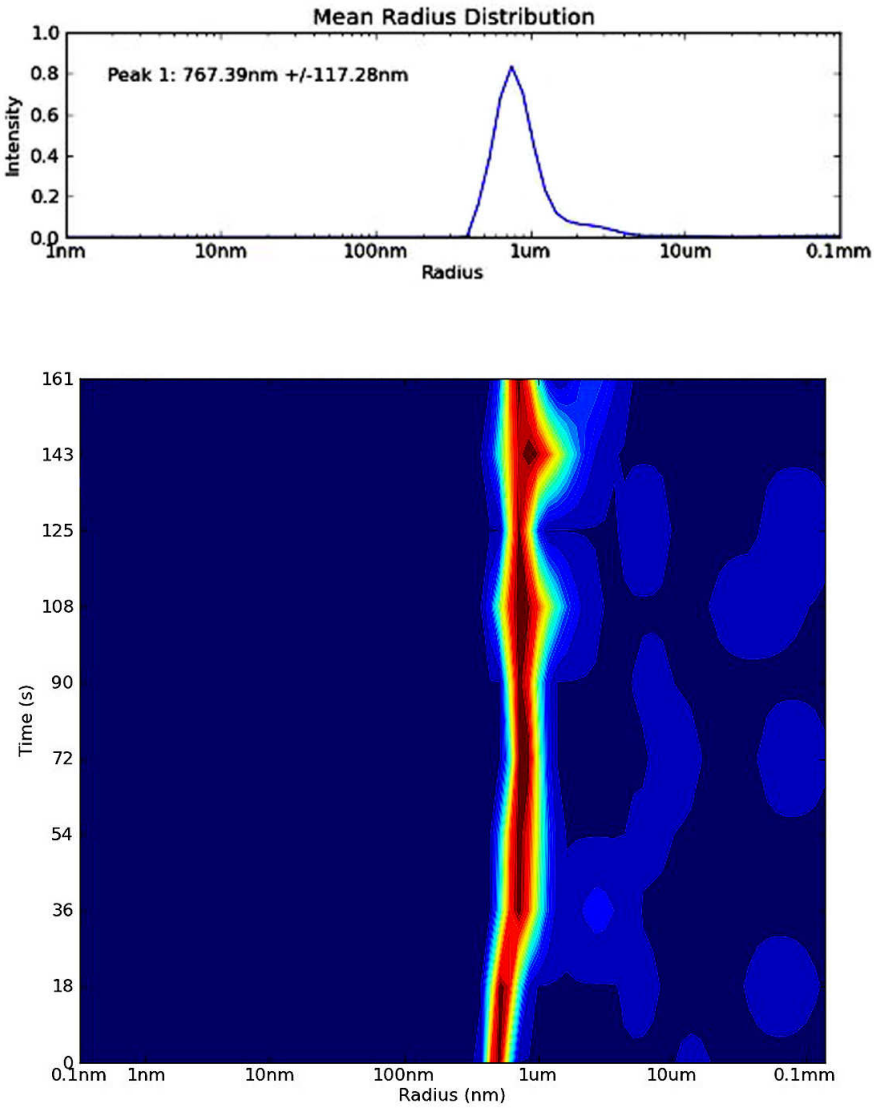


Figure 6.19: *PSI crystal size. From dynamic light scattering measurements (DLS). Mean of the distribution is around 770 nm and is stable over time as visible in the lower part of the figure.*

Figure 6.20: *A single PSI Bragg spot as recorded on one of the panels of the CSPAD detector selected from the distribution of PSI Bragg spots.*

distribution from:

$$\sigma_{cell}^2 = \sigma_{ring}^2 - \sigma_{size}^2, \tag{6.5}$$

where σ_{cell} is the value that we want to obtain from the knowledge of the width of the powder ring, $\sigma_{ring} = 0.00153$ 1/Å. The $\sigma_{size} = 0.000724$ (1/Å) is the average domain size as estimated from the integral breadth analysis from previous section. The resulting $\sigma_{cell} = 0.0013$ (1/Å) which we when put into the equation $\Delta d = \Delta s \times d^2$ in order to obtain the width of the unit cell length distribution. For the (6,0,0) reflection and the unit cell length $a = 126$ Å, the d=126/6=21 Å.

$$\Delta d = 0.0013 \times 21^2 = 0.57 \text{ Å} \tag{6.6}$$

which is very close to the value of the standard deviation of the unit cell length "a" obtained from indexing.

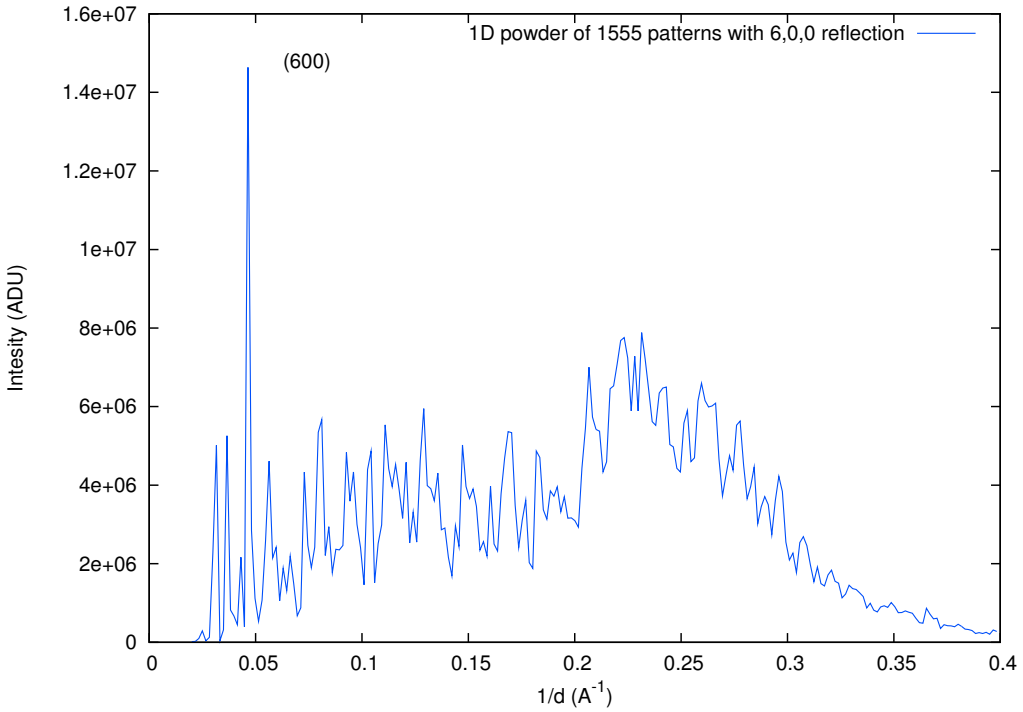


Figure 6.21: *A virtual powder pattern created by summing up 1,555 indexed single crystal TbCatB diffraction patterns that contain the (6,0,0) reflection. This has been realized in order to select one diffraction ring for further analysis of its width and to ensure that this ring will not overlap with other rings in the virtual powder pattern. The correction for per pulse wavelength instabilities was applied.*

6.5 Indexing schemes

Diffraction patterns collected during our SFX measurements were indexed using the CrystFEL software suite as explained in Chapters 3 and 5 and Appendix B. There are three methods implemented in the CrystFEL for indexing single crystal diffraction patterns basing on the positions of Bragg peaks found in the diffraction pattern. The first scheme uses Mosflm [Les07], the second scheme uses DirAx [Dui92]. The first scheme uses the conventional "DPS" autoindexing algorithm, the second uses the conventional DirAx autoindexing algorithm and the third uses ReAx, an experimental algorithm, similar to the fast Fourier transform (FFT) autoindexing algorithm described in [Ste97]. In the ReAx method, the three dimensional positions of the peaks in the diffraction pattern are calculated by mapping the peaks onto the Ewald sphere in reciprocal space. Once the three dimensional positions of the peaks in a diffraction pattern are established, the FFT searches for the interplanar vector

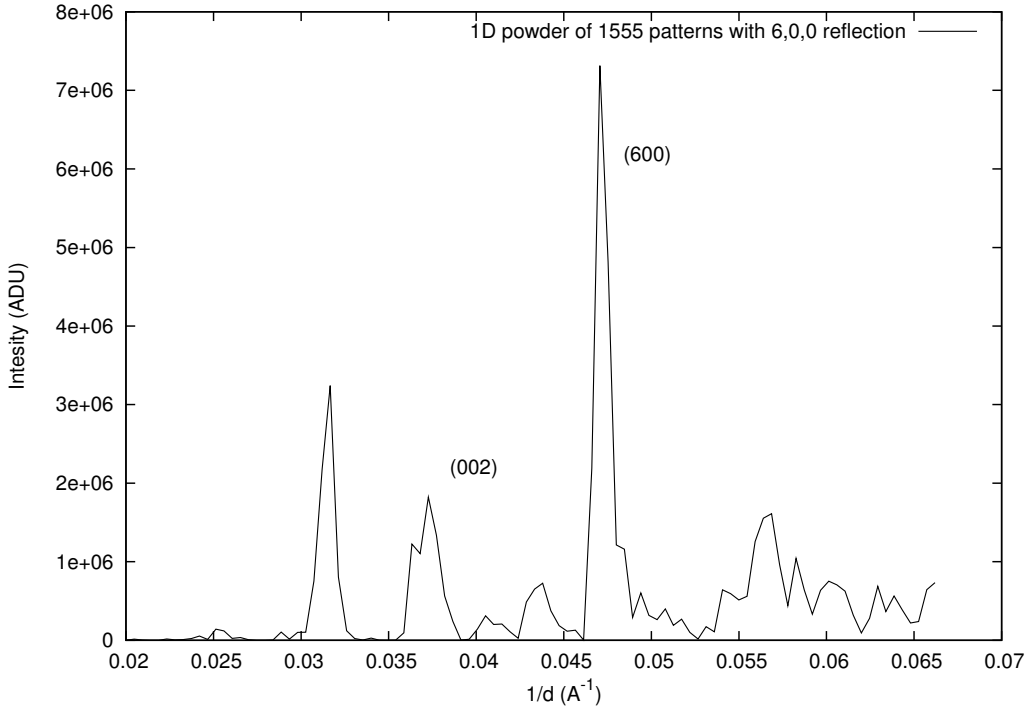


Figure 6.22: *A virtual powder pattern created by summing up 1,555 indexed single crystal TbCatB diffraction patterns that contain the (6,0,0) reflection. This has been realized in order to select one diffraction ring for further analysis of its width and to ensure that this ring will not overlap with other rings in the virtual powder pattern. Correction of per pulse wavelength instabilities was applied.*

candidates of the lengths within a 10% tolerance of the reference unit cell parameters. The vectors that match the largest number of calculated reciprocal positions of the Bragg peaks from the diffraction pattern are selected to be the unit cell vectors.

The full TbCatB data set was indexed using the known unit cell parameters of TbCatb crystals determined previously, using both DirAx and Mosflm indexing schemes at once in CrysFEL, i.e. when one failed to index a given diffraction pattern with the provided unit cell parameters the second scheme was used. If any of the schemes were not successful in indexing then the patterns was rejected as non-indexable. The indexing relies strongly on the ability of finding the positions of the Bragg peaks on diffraction pattern and then on translating them into unit cell vectors. The indexing yield of those two indexing methods using known unit cell parameters on the whole TbCatB data set was 61%. In this section I will describe which combination of the three above mentioned indexing schemes is the best for TbCatB data set.

Table 6.14: *Mean and standard deviations of unit cell parameters measured from 104,115 TbCatB crystals indexed using only Mosflm. The space group of TbCatB crystals is* $P4_22_12$*. Tetragonal primitive:* $a = b \neq c$, $\alpha = \beta = \gamma = 90°$.

Parameter	Mean	Standard dev.
a (Å)	126.096	0.537
b (Å)	126.096	0.537
c (Å)	54.237	0.288
α (°)	90.000	1.756e-05
β (°)	90.000	1.767e-05
γ (°)	90.000	2.261e-05

6.5.1 Indexing TbCatB data set using only Mosflm

In this study case only the Mosflm scheme in CrystFEL was used to index the full TbCatB data set. Resulting indexed data set consists of 104,155 indexed diffraction patterns, which gives indexing yield of 33%, much lower than for the final, full TbCatB data set which is 61%. Mosflm imposes unit cell constraints according to the space group of the crystals. Standard deviations of the resulting unit cell lengths are similar as in indexing with DirAx and Mosflm but for unit cell angles it gives much lower deviation, see the table 6.14. This is because of the space group constraint. R factor analysis on the data set indexed only using Mosflm: R factor equals 19.44 %, R free equals 22.20 %. R factors from refinement using the data set indexed only by Mosflm are worst than form the whole TbCatB data set indexed using Mosflm and Dirax, see table 5.1 for comparison. Signal to noise ratios and redundancies in resolution shells are presented in the table 6.15. The signal to noise ratios and redundancies are also worst than in the whole TbCatB data set indexed using Mosflm and Dirax, see table 5.2 for comparison. Indexing using only the Mosflm scheme is not as good as indexing using both Mosflm and Dirax schemes at once.

6.5.2 Indexing TbCatB data set using only DirAx

In this study case only the DirAx scheme in CrystFEL was used to index the full TbCatB data set. Resulting indexed data set consists of 170,972 indexed diffraction patterns, which gives indexing yield of 55%. Combining both indexing schemes, DirAx and Mosflm gives indexing yield of 61%. DirAx does not impose the unit cell constraints according to the space group, as it is done by Mosflm. Standard deviations of the resulting unit cell lengths are similar as in indexing with DirAx and Mosflm, see the table 6.16. R factors from refinement using the data set indexed

Table 6.15: *Quality measures of the TbCatB data set indexed using only Mosflm.*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.500	4375	25.52
4.509 – 3.585	4669	21.52
3.585 – 3.134	4832	15.34
3.134 – 2.848	4423	9.46
2.848 - 2.645	4520	6.24
2.645 - 2.489	4616	4.28
2.489 - 2.365	4741	2.99
2.365 - 2.262	4538	2.36
2.262 - 2.175	4361	1.94
2.175 - 2.100	4085	1.61

Table 6.16: *Mean and standard deviations of unit cell parameters measured from 170,972 TbCatB crystals indexed using DirAx only. The space group of TbCatB crystals is $P4_22_12$. Tetragonal primitive: $a = b \neq c$, $\alpha = \beta = \gamma = 90°$.*

Parameter	**Mean**	**Standard dev.**
a (Å)	125.992	0.525
b (Å)	126.812	0.617
c (Å)	54.3221	0.222
α (°)	90.022	0.466
β (°)	89.985	0.356
γ (°)	90.004	0.447

only by DirAx are worst than from the whole TbCatB data set. The R factor equals 19.04 %, R free equals 21.40 %. Signal to noise ratio in resolution shells is presented in the table 6.17 and are also worst than in the whole TbCatB data set. Indexing using only the DirAx scheme is not as good as indexing using both Mosflm and Dirax schemes at once but better than in the case where only the Mosflm scheme was used.

6.5.3 Indexing using only ReAx

Similar analysis was performed using only the ReAx indexing method which is currently under development in the CrystFEL software suite. Resulting indexed data

Table 6.17: *Quality measures of the TbCatB data set indexed using only DirAx.*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.500	7191	32.09
4.509 – 3.585	7726	26.87
3.585 – 3.134	7962	18.73
3.134 – 2.848	7283	11.95
2.848 - 2.645	7433	8.06
2.645 - 2.489	7614	5.85
2.489 - 2.365	7849	4.51
2.365 - 2.262	7507	3.58
2.262 - 2.175	7172	2.91
2.175 - 2.100	6688	2.37

set consists of 262,425 indexed diffraction patterns, which gives indexing yield of 84%. This is 23% more indexed patterns than by using Mosflm and DirAx combined methods. However, the R factor analysis on the data set indexed only using ReAx reveals lower quality data set than the final TbCatB data set indexed by using Mosflm and DirAx together. The R factor in the ReAx case equals to 20.09 %, R free equals to 23.72 %. The higher than previously obtained R free value indicates that the merged set of structure factors is of lower quality than the sets obtained by using other indexing schemes. This is also apparent in the Wilson plots in figure 6.23. This suggests that the highest usable resolution for that data set indexed only using DirAx is lower than 2.1 Å. Another indicator of that fact is the R_{split} plot in function of resolution, see the figure 6.24. Indeed, as visible in the table 6.18 the "Merged $I/\sigma(I)$" is lower than in previous cases.

In conclusion, the indexing scheme that was the best from all currently available in terms of the lowest R factors from refinement and the best redundancy with the highest signal to noise ratios in resolution shells was the combination of two indexing methods that use Mosflm and DirAx.

6.6 Effect on R factors of the number of indexed patterns

In this section I am comparing the effects on R_{split}, R work, R free, signal to noise and redundancies in function of the number of indexed patterns by both methods Mosflm and DirAx. The R_{split} quality factor is defined like:

Table 6.18: *Quality measures of the TbCatB data set indexed using only ReAx.*

Resolution shell (Å)	**Redundancy**	**Merged $I/\sigma(I)$**
20.000 – 4.500	11150	24.09
4.509 – 3.585	11863	15.91
3.585 – 3.134	12334	8.51
3.134 – 2.848	11360	4.40
2.848 - 2.645	11401	2.84
2.645 - 2.489	11691	1.96
2.489 - 2.365	11830	1.50
2.365 - 2.262	11555	1.22
2.262 - 2.175	11157	1.18
2.175 - 2.100	10850	1.20

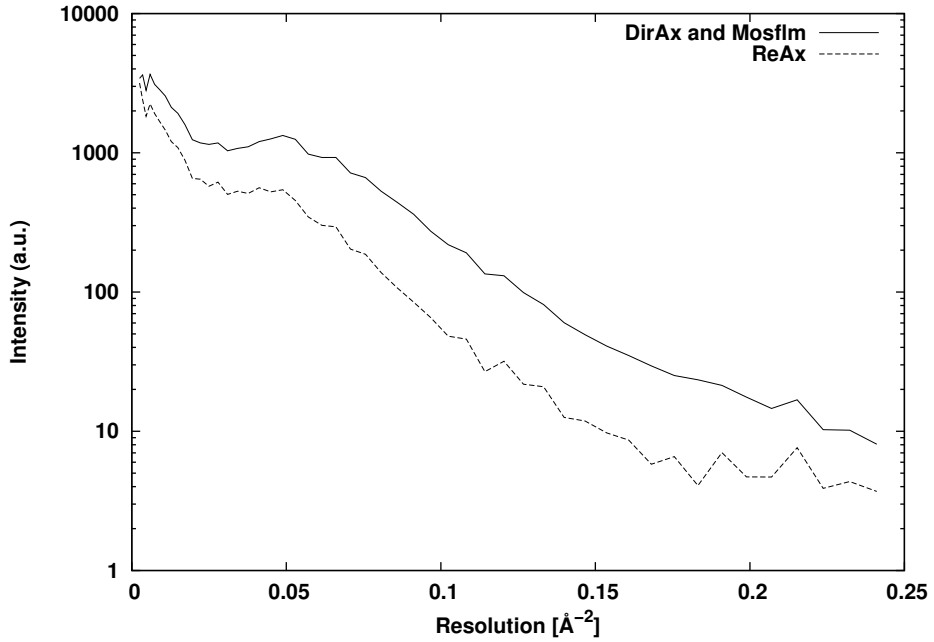


Figure 6.23: *Wilson plot of of TbCatB data set indexed using only ReAx method (solid) and for the final TbCatB data set indexed using DirAx and Mosflm (dashed).*

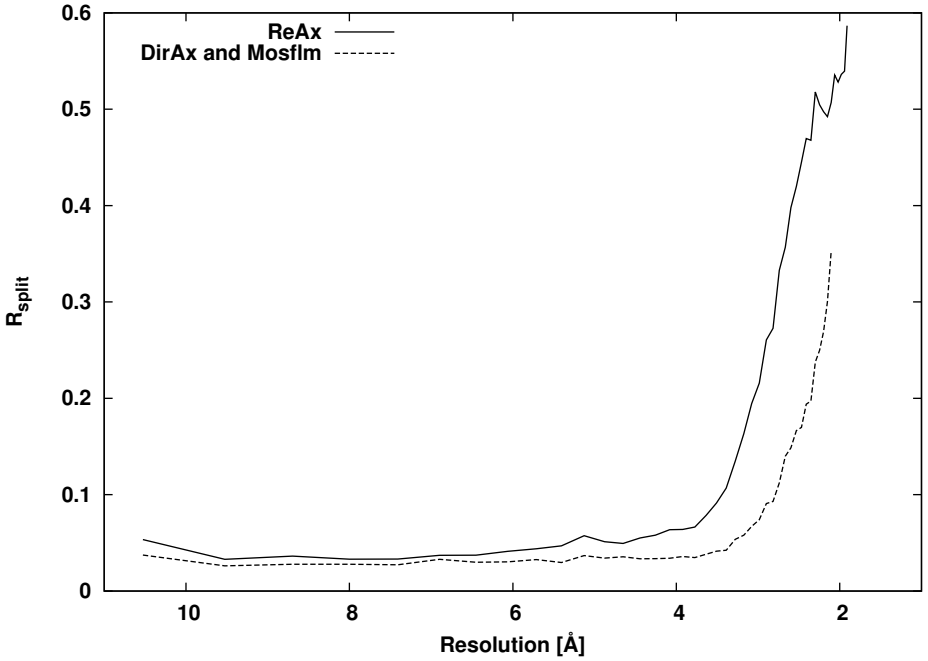


Figure 6.24: *Plot of R_{split} of TbCatB data set indexed using only ReAx method (solid) and for the final TbCatB data set indexed using DirAx and Mosflm (dashed).*

$$R_{split} = 2^{-1/2} \frac{\sum |I_{even} - I_{odd}|}{\frac{1}{2} \sum (I_{even} + I_{odd})}, \tag{6.7}$$

where I_{even} represents the intensity of a reflection produced by merging even-numbered patterns, I_{odd} represents the intensity of the equivalent reflection from the odd-numbered patterns and the sum is over all reflections [Whi12] in given resolution limit. Here I use the resolution limits of the final TbCatB data set (20.0Å–2.1Å). All data set quality indicators become worst when the number of indexed patterns decreases, see the table 6.19 and the plot of R_{split} in the function of number of indexed patterns 6.25 for reference.

Table 6.19: *Table with data for TbCatB data set for R_{split}, merged signal to noise ratio, redundancy, R factor and R free factor in function of the number of indexed patterns of TbCatB crystals.*

No. of patt.	**R$_{split}$ (%)**	**Merged$I/\sigma(I)$**	**Redun.**	**R (%)**	**R free (%)**
178,875	9.4	8.68	7780	19.62	22.28
89,223	11.9	6.15	3884	20.07	23.14
44,684	15.2	4.36	1946	20.58	24.50
22,314	19.3	3.11	976	21.19	24.47
11,186	24.2	2.22	488	22.10	25.28
5,586	29.7	1.62	244	23.22	27.36

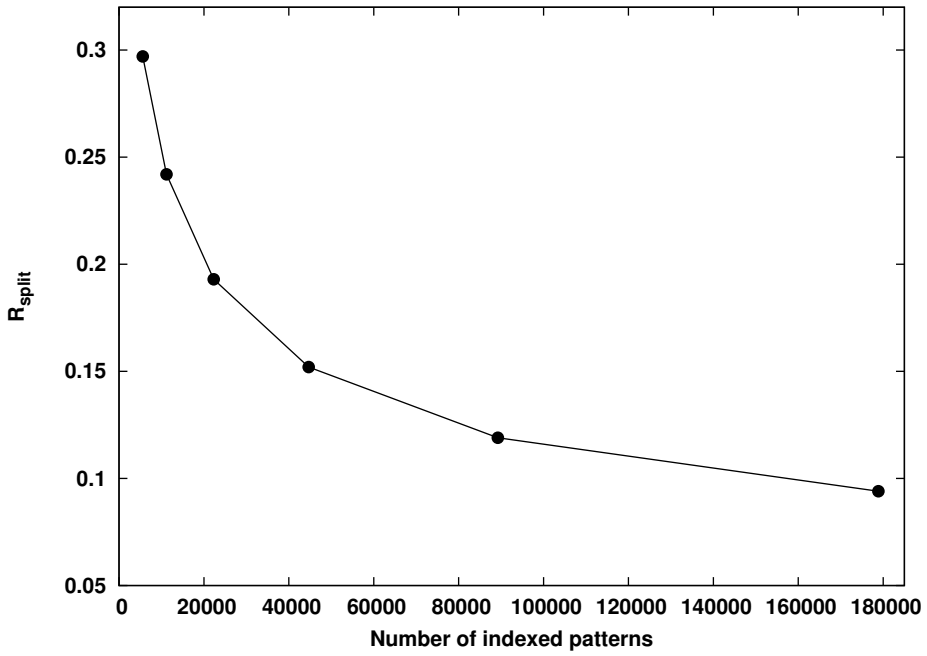


Figure 6.25: *R_plit vs. number of indexed TbCatB patterns.*

Chapter 7

Outlook

This dissertation presented research performed on the experimental data collected by a large scientific collaboration during the first serial femtosecond crystallography (SFX) experiments using the first available hard X-ray free-electron laser source, the SLAC Linac Coherent Light Source (LCLS) in California and thousands of *in-vivo* grown protein crystals of unknown structures and sizes smaller than acceptable for similar studies at third generation X-ray sources. The research performed by our collaboration hopefully will encourage others to perform similar type measurements on their favourite and important biological macromolecules that would lead to the solution of so far unsolved protein structures and understanding of their structure to function relations.

The research presented in this thesis lead to the discovery of the first and second high-resolution structures of protein crystals, which were previously unknown, by using the X-FEL radiation source. Thesis presents also the assessment of the serial femtosecond crystallography method. The work performed for this thesis and presented in the form of scientific publications in highly ranked scientific journals confirms the usefulness of X-ray FEL sources in retrieving new structural information from biological macromolecules arranged in the form of nano- and micro-crystals. Additionally, the structure of the *Trypanosoma brucei* Cathepsin B (TbCatB) solved during this work, revealed previously unknown structural details that are connected with the native inhibitory mechanism of that enzyme that can help others to develop selective inhibitor compound for this enzyme. This could lead to the development of a new drug against Human African Sleeping Sickness.

7.1 Summary of results

The serial femtosecond crystallography experiments were performed using a custom built sample delivery system, which was designed uniquely for those experiments.

The experimental layout is extensively described in this thesis. The sample delivery system produced a thin stream of liquid in which the nano-sized protein crystals were delivered to the FEL pulsed beam. It has been shown in this thesis and in other sources that the diameter of the liquid jet can be made smaller than one micron, therefore achieve the match between the size of the sub micrometer protein crystals, the liquid jet in which they are delivered, and the X-ray beam that can be focused to sub micrometer dimension. The resulting diffraction patterns from protein crystals were indexed using conventional methods and merged using new for protein crystallography Monte Carlo approach. This method has been experimentally verified to be useful for obtaining very good quality set of structure factors extending to high-resolution, from protein nano- and micro-crystals as seen in previous chapters, that enabled the structure determination of those protein crystals.

Those *in-vivo* grown crystals that were used during the SFX measurement described in this thesis were extensively characterized in the preparation laboratory before the experiment using optical and electron scanning microscopy (SEM), as well as a new method called SONICC (Second Order Non-linear of Chiral Crystal) which enables the detection and characterization of nano-crystalline protein material. This method of detection and characterization has a big chance to be extensively used in the future by users of the FEL facilities that would like to perform SFX experiments on their protein targets. Also a preliminary X-ray powder diffraction analysis was performed on TbCatB at a synchrotron source, in order to characterize their ability to diffract X-ray radiation. This step was necessary in order to confirm that those crystals can be used at the X-FEL source.

The resulting protein structures presented no signs of radiation induced damage in the retrieved electron density at the maximal achieved resolution of 2.1Å using exposures to ultra intense X-ray pulses of 40fs in duration. This suggests that the structures obtained by SFX present no observable radiation induced damage. And confirmed for protein crystals the "diffract-before-destroy" principle that was demonstrated earlier at the soft X-ray FEL source for non-periodic, inorganic samples.

One of the problems that I encountered in the data analysis was the presence of the background scattering from liquid jet that resulted in a bright "water ring" structure in every diffraction pattern, which intensity fluctuated from shot to shot. In order to clean the diffraction patterns from background scattering I contributed to development of background cleaning procedures implemented in the software used for the data analysis. Without this step, the indexing was difficult and the Monte Carlo integration of intensities extracted from thousands of indexed diffraction patterns was not of very good quality that could not enable the solution of the underlying crystal structure or discovery of the new structural features of TbCatB crystals (pro-peptide and carbohydrate chains). In the beginning of the work on

the structure solution the water background problem was causing the problem of very poorly defined electron density of those new features therefore making then not recognizable as parts of the protein structure. The solution of the background problem enabled the possibility to index more diffraction patterns by making the Bragg peas easier to find by the peak finding algorithm, therefore to improve the statistics of the merged, final set of structure factors which improved the resulting electron density to the level where the new parts of the protein structure became apparent and data set quality indicators also improved.

This fact suggests that the largest contribution to the error in estimation of the structure factors are: the uncertainty of estimation of the amount of background that has to be subtracted on per shot basis. And the necessity to perform a large number of measurements per every reflection that is required to fully integrate partials and to build up enough signal from weak reflections, that are detected in diffraction patterns at high resolution, in order to improve the statistics of the merged data set at the highest possible resolution. The requirement of collecting a large number of diffraction patterns in order to fully integrate partial reflections resulting from still diffraction images and build up a signal from weak reflections at high resolution can be accounted by using X-FEL pulses with larger spectral bandwidth, which would allow to fully integrate a reflection in a single exposure and would result in a larger number of Bragg peaks recorded per one diffraction pattern.

Also the results presented here suggests that the SFX experiments on the smallest nano-crystals should be performed using the most intense X-ray pulses that are available in order to record detectable signal at the highest resolution that the crystals permit and using the shortest pulses available in order to outrun the radiation damage induced by the intense X-ray radiation.

Lastly, the intriguing effects of the distribution of unit cell parameters from this unique and large X-ray diffraction data set collected from thousands of protein crystals of the same kind and its connection to the observed flow alignment of needle shaped TbCatB protein crystals in the liquid jet injector were investigated. The analysis on that subject suggests that large protein crystals display the ability to flow align during the injection of crystal suspension by using the liquid jet injector and present the low deviation from the mean unit cell parameters that were calculated from 178,875 measurement of the unit cell parameters. Those large crystals give on average stronger Bragg scattering signal to higher resolution than the averaged Bragg scattering signal measured from the same (as of the flow aligned crystals) number of diffraction patterns selected at random from the whole data set.

7.2 Future Work

One of the biggest challenges that the field of structural biology is facing at the moment, is the development of experimental technique that would allow determination of biological macromolecules without the need of crystallization. The successful realization of SFX experiment that enabled the determination of the structure of an unknown large protein determined from thousands of crystals smaller that previously usable in structural investigations of proteins at synchrotrons is a large step forward in this direction. However, before this aim can be realized, there is some field of improvements left for serial femtosecond crystallography.

7.2.1 De novo phasing of SFX data

To retrieve the three dimensional model of a protein from single diffraction patterns of protein nanocrystals recorded at X-FEL after the extraction of structure factors by Monte Carlo method described previously, the information about missing phases must be obtained. The percentage of structure solution methods in the PDB data bank tells that molecular replacement (MR) method is the most favourable. However, *de novo* structure determination is the most desirable, since many of the most interesting and difficult targets aimed for SFX do not have homologous structures already deposited in the PDB data bank, that could be used for phasing using MR.

Multi-wavelength anomalous dispersion (MAD) method is one of the methods that can be used to determine the structure of protein crystal without any previous information about the homologue model. It has been recently presented that the MAD method is valid for structure determination at high intensity X-ray sources [Son11]. The application of MAD methods at short wavelengths to SFX introduces some interesting effects due to high intensity of X-ray pulses. In this method, anomalous contribution to scattering from heavy atoms, which X-ray absorption edge is close to the incident wavelengths, allows for the phase to be determined using two data sets recorded at two different wavelengths or interestingly, two different fluences. Data sets recorded below and above absorption edge of a heavy atom or using more and less intense radiation of the same wavelength that introduces changed to the absorption coefficients for heavy atoms that are buried into structure of protein that form crystals, can be used to localize positions of heavy atoms and then straightforwardly phases for other atoms in the structure. Common in conventional MX is a protocol that replaces methionine residues by the selenomethionine, thus allows use of the selenium absorption edge for MAD method. This approach could also be introduced at XFEL's.

Intense X-FEL radiation of longer wavelengths than absorption edges of metal atoms present in protein crystal structures could be also used to exploit small differences in anomalous scattering factors of sulphur or phosphor atoms for two different

pulse intensities. Therefore, enabling native, *de novo* protein structure determination using XFEL's very intense radiation at wavelengths tunable to around 5 keV. This relies on the preferential ionization of S or P atoms over lighter elements. As the vast majority of proteins contain either methionine and/or cysteine residues (that have one sulphur atom), the method has a huge potential. In the case of Sulphur, differences in the signal are very small compared to heavy atom anomalous signal, but the use of very redundant data sets may be sufficient.

SFX has created another new opportunity to solve the phase problem, by using the interference fringes between Bragg peaks recorded from protein nanocrystals. This method relies on the use of the smallest nanocrystlas for which the shape transform plays a sufficient role in recorded diffracted intensities so that it can be measured and divided out from the equation 3.24 and the molecular transform modulus $|F(\Delta k)|$, might be extracted from the measured intensity [Spe11]. Having recovered the molecular transform modulus $|F(\Delta k)|$, the complex molecular transform $F(\Delta k)$ (effectively: content of the unit cell) may be obtained by iterative phasing methods if sufficient sampling between Bragg spots.

7.2.2 Possible improvements to sample consumption issues

The large amount of protein needed for these first SFX experiments described in this thesis could be reduced. Advances in data analysis methods that need to be developed could provide converged structure factors with less data. Less data is also required if the bandwidth of the X-FEL pulses could be made larger, so that Bragg peaks are fully integrated on each shot. Secondly, the micro fluidic devices could be used as sample delivery method that works on lower flow rate than current liquid jets, see for example the recent development of the electro spun injector device [Sie12]. The development of pulsed sample delivery system, synchronized to the X-FEL could also reduce the sample consumption. Finally, the repetition rate of the European XFEL will be over 200 times higher at the LCLS, which would potentially reduce the amount of sample that is wasted between the X-ray pulses if the liquid jet injector is used. It may also become important to develop stationary methods of sample environment as for example presented in [ZA12].

Appendix A

Data pre-analyser: Cheetah

Cheetah is a piece of software that is used to offline pre-analyse raw data stream collected by using CSPAD detector from the CXI instrument located at CXI beam line at LCLS. It was developed by Anton Barty and colleagues from CFEL and SLAC. It is based on the *myana* frame developed by SLAC scientists. *Myana* was created for handling the detector raw data streams. Recently, *Cheetah* has been updated in order to handle the new SLAC *psana* libraries which enable online (during the experiment) and offline access to the detector raw data stream. It is written in C++ language. It is multi-threaded for quicker data analysis.

Cheetah enables extraction of single detector frames (diffraction patterns) and other information regarding the experiment from the LCLS data stream that is stored in XTC file format and saving it into popular HDF5 file format.

The simplest task achieved by this program is to create a HDF5 image of every single detector frame with other information such as wavelength, peak current, etc. associated with it in the HDF5 file structure and save it to local hard drive. The actual purpose of *Cheetah* is to identify and save single diffraction images from the detector raw data stream that contain usable information, for example Bragg diffraction from protein crystals. A simple threshold algorithm is used to localise Bragg peaks in the image. If sufficient number of Bragg peaks is localised in the image it is considered as a protein crystal hit. In the later version of *Cheetah* a more advanced spot finder algorithm was implemented. It takes the information about signal to background ratio around the Bragg peak. This approach was found to be more reliable in localising Bragg peaks in diffraction patterns. Identification of usable detector frames performed online, while the experiment is running, is highly desirable for any "serial" type of experiments at high-repetition rate FEL sources.

In addition to hit-finding task, *Cheetah* can apply corrections to raw detector frames necessary to extract diffraction images used in further analysis that are free from noise and background scattering from for example water jet. At first, subtrac-

tion of an average readout value measured in ADUs (analogue to digital units) when no X-rays are present is performed from every pixel value. The so called "darkcal" calibration. It is in the range of approximately 1200 to 1600 ADUs. It consists of (i) pedestal contribution uniform for the pixels within each application-specific integrated circuit (ASIC) of the CSPAD detector and (ii) the electronic noise of each individual pixel. For well behaved pixels its standard deviation is smaller than the detector response of a single photon (typically 9 ADUs for 9.4 keV photons in high gain mode). Secondly, the gain calibration is applied. It rescales the detector response on a pixel basis using a normalized gain correction. The gain correction is calculated from a linear fit of the average readout value of each individual pixel of the CSPAD detector as a function of photon flux from flat-field Cu fluorescence with varying attenuation. It can also apply geometrical correction and assemble individual detector tiles in a two dimensional image. Pixels that have atypical behaviour can be masked out from analysis.

One of the important corrections of *Cheetah* is per-shot background subtraction. It substantially cleans the diffraction image from scattering of water molecules of the liquid jet that fluctuates on per-shot basis and disturbs further analysis of merged data set. Background subtraction filter calculates a median value from a box of typically two pixels radii taken around every pixel in an image and subtracts it from that pixel. Bragg peaks remain preserved after this treatment.

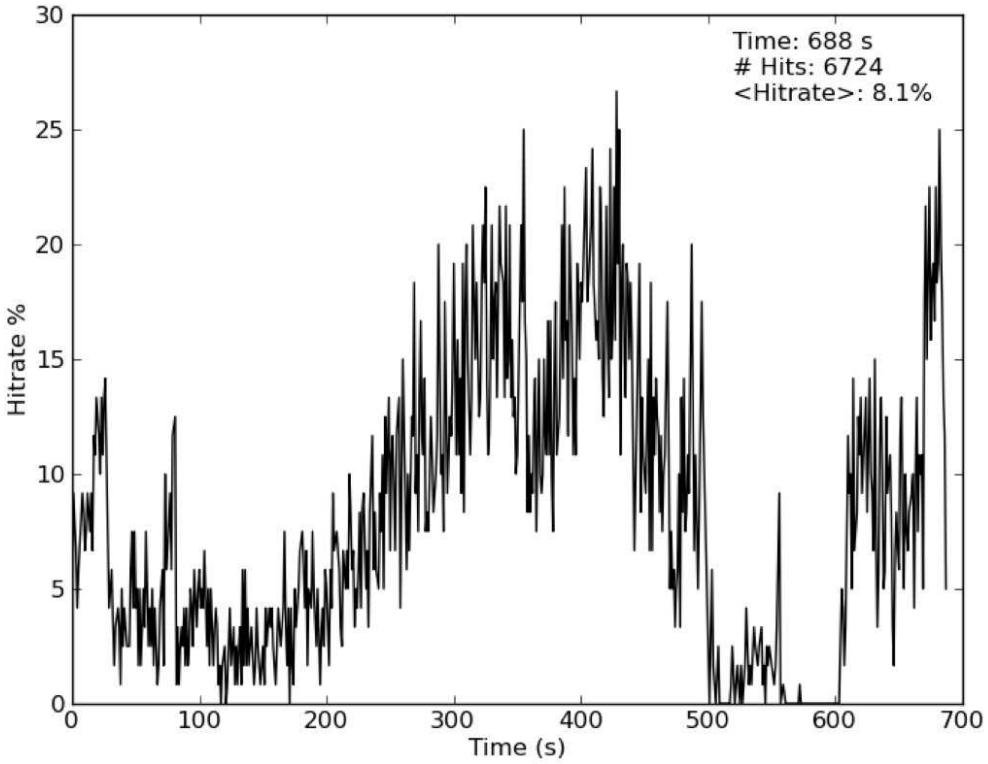


Figure A.1: *Presentation of a hit rate detected by Cheetah in TbCatB crystals experiment. The data stream of 11 minutes displays hit rate of approximately 8.1 %. Hit rate is varying during the experiment due to settlement of crystals in feed lines and re-suspending them again by using anti settling device [Lom12].*

Appendix B

CrystFEL software suite

CrystFEL is a software suite designed to analyse diffraction data obtained by Serial Femtosecond Crystallography (SFX) method [Whi12]. It can also be used to simulate diffraction from crystals. In SFX method diffraction patterns are acquired in serial manner from thousands of protein crystals in random orientations. CrystFEL tries to find orientation of every crystal diffraction pattern and add its contribution to the set of structure factors that converge to the solution with high enough redundancy.

CrystFEL consists of shared library (libcrystfel) and contain few programs written in C language and scripts that can be executed from the command line. CrystFEL uses HDF5 library for handling the data format issues. Primary program of CrystFEL is called *indexamajig*. It is used to locate Bragg peaks in an image and to parse the information of peak positions to autoindexing software like DirAx and Mosflm. If the pattern could be indexed by one of those programs then it performs check whether the predicted peak positions correspond to at least 10% of detected Bragg peaks. If this check is true then the pattern is considered as indexed. Intensities of peaks from predicted positions are then integrated and corrected by the surrounding background estimation. Indexed image is added to the pool (so called "stream") which program (*process_hkl*) uses to merge one set of intensities according to the Monte Carlo method.

Check_hkl is a simple program of the CrystFEL suite that calculates figures of merit of the merged set of intensities. Such as completeness and average signal strengths, in resolution shells.

Compare_hkl is an another program of the CrystFEL suite that is used to compare two sets of merged intensities and to calculate R-factors and correlation coefficients between them.

Very useful simple program *render_hkl* is often used to visualise the integrated intensities as circles or spheres on a common colour scale. All three programs de-

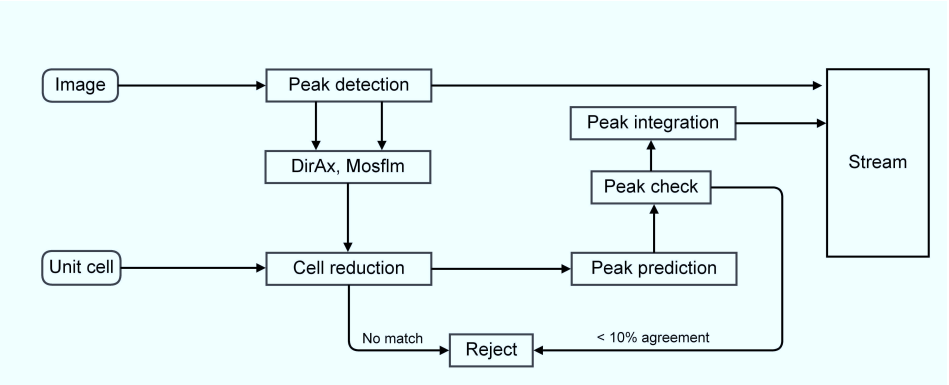


Figure B.1: *Flow diagram of diffraction pattern processing in indexamajig. Adapted from [Whi12].*

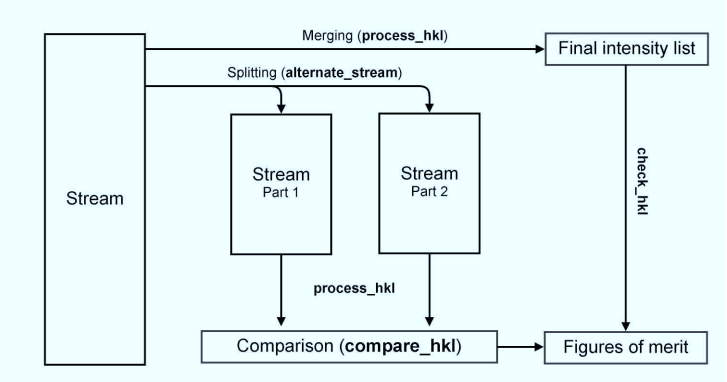


Figure B.2: *Flow diagram of final intensity list merging and data evaluation in CrystFEL. Adapted from [Whi12].*

scribed above are helpful in inspecting the quality of a merged data set.

CrystFEL contain also simple HDF5 image viewer called *hdfsee*, two programs used to simulate diffraction data (*pattern_sim, partial_sim*) and set of useful scripts used to check, convert, find or generate information from processed data set in order to evaluate its quality.

Figure B.1 presents overall processing flow of diffraction pattern in *indexamjig*. Figure B.2 presents flow diagram of the final intensity list merging form "stream" file created in previous step and evaluation of the quality of data.

Appendix C

Reprints of published articles

In this appendix I include reprints of main articles that are listed in the front matter of this thesis, which were published in scientific journals during my graduate work.

Natively Inhibited *Trypanosoma brucei* Cathepsin B Structure Determined by Using an X-ray Laser

Lars Redecke,[1,2]† Karol Nass,[3,4]† Daniel P. DePonte,[3] Thomas A. White,[3] Dirk Rehders,[1] Anton Barty,[3] Francesco Stellato,[3] Mengning Liang,[3] Thomas R.M. Barends,[5,6] Sébastien Boutet,[7] Garth J. Williams,[7] Marc Messerschmidt,[7] M. Marvin Seibert,[7] Andrew Aquila,[3] David Arnlund,[8] Sasa Bajt,[9] Torsten Barth,[10] Michael J. Bogan,[11] Carl Caleman,[3] Tzu-Chiao Chao,[12] R. Bruce Doak,[13] Holger Fleckenstein,[3] Matthias Frank,[14] Raimund Fromme,[12] Lorenzo Galli,[3,4] Ingo Grotjohann,[12] Mark S. Hunter,[12]* Linda C. Johansson,[8] Stephan Kassemeyer,[5,6] Gergely Katona,[8] Richard A. Kirian,[3,13] Rudolf Koopmann,[10] Chris Kupitz,[12] Lukas Lomb,[5,6] Andrew V. Martin,[3] Stefan Mogk,[10] Richard Neutze,[8] Robert L. Shoeman,[5,6] Jan Steinbrener,[5,6] Nicusor Timneanu,[15] Dingjie Wang,[13] Uwe Weierstall,[13] Nadia A. Zatsepin,[13] John C. H. Spence,[13] Petra Fromme,[12] Ilme Schlichting,[5,6] Michael Duszenko,[10] Christian Betzel,[16]‡ Henry N. Chapman[3,4]‡

The *Trypanosoma brucei* cysteine protease cathepsin B (TbCatB), which is involved in host protein degradation, is a promising target to develop new treatments against sleeping sickness, a fatal disease caused by this protozoan parasite. The structure of the mature, active form of TbCatB has so far not provided sufficient information for the design of a safe and specific drug against *T. brucei*. By combining two recent innovations, in vivo crystallization and serial femtosecond crystallography, we obtained the room-temperature 2.1 angstrom resolution structure of the fully glycosylated precursor complex of TbCatB. The structure reveals the mechanism of native TbCatB inhibition and demonstrates that new biomolecular information can be obtained by the "diffraction-before-destruction" approach of x-ray free-electron lasers from hundreds of thousands of individual microcrystals.

Over 60 million people are affected by human African trypanosomiasis (HAT), also known as sleeping sickness, which causes ~30,000 deaths per year (*1*). The protozoan parasite *Trypanosoma brucei*, transmitted by tsetse flies, infects the blood and the lymphatic system before invading the brain. Severe clinical manifestations occur within weeks or months. Current treatments of HAT rely on antiparasitic drugs developed during the last century, without knowledge of the biochemical pathways. These treatments are limited in their efficacy and safety, and drug resistance is increasing (*2*–*4*). Thus, new compounds that selectively inhibit vital pathways of the parasite without adverse affects to the host are urgently required. A promising strategy is to target lysosomal papainlike cysteine proteases that are involved in host-protein degradation, such as cathepsin B (*5*). The knockdown of this essential enzyme in *T. brucei* resulted in clearance of parasites from the blood of infected mice and cured the infection (*6*), which qualify cathepsin B as a suitable drug target. Cysteine proteases are synthesized as inactive precursors with N-terminal propeptides that act as potent and selective intrinsic inhibitors until the proteases enter the lysosome (*7*), where the propeptide is released and forms the mature active enzyme. Such native propeptide-inhibited structures have been used to develop species-specific protease inhibitors against proteases of other *Trypanosoma* species, e.g., cruzipain of *T. cruzi* (causing human Chagas disease in America) and congopain of *T. congolense* (causing nagana in cattle) (*8*, *9*). This approach could not be explored for *T. brucei* cathepsin B (TbCatB) because of the lack of structural information on the mode of propeptide inhibition and the large extent of structural conservation at the active site between mammalian and trypanosome cathepsin B (*10*–*12*). Previously solved mature *T. brucei* and human CatB structures show differences at the S2 and in part of the S1′ subsite of the substrate-binding cleft (Fig. 1C) and have been suggested as possible targets for the development of species-specific CatB inhibitors (*10*). Together with the natively inhibited human procathepsin B structure (*13*), our work fills the gap to understand the structural basis for species-specific inhibition.

The growth of large well-ordered protein crystals is one of the major bottlenecks in structure determination by x-ray crystallography—with important biological targets, such as integral membrane proteins and posttranslationally modified proteins, proving particularly challenging to crystallize (*14*). Sizable crystals are required to obtain measurable high-resolution diffraction data within an exposure that is limited by the accumulation of radiation damage (*15*). Although microfocus beamlines enable the collection of diffraction data from micron-sized protein crystals (*16*), the tolerable dose limit of less than 30 MGy for cryogenically cooled protein crystals remains, which limits the achievable signal. The tolerable dose for room temperature measurements is about 1 MGy (*15*). We have previously shown that micron-sized crystals of glycosylated TbCatB spontaneously form in insect cells during protein overexpression (*11*). Such crystals are extremely well suited for the new method of serial femtosecond crystallography (SFX) (*17*). X-ray free-electron laser (FEL) pulses of less than 100-fs duration allow the dose to individual crystals to exceed the ~1 MGy limit by over a thousand times because of the "diffraction-before-destruction" principle (*17*, *18*). Diffraction data are recorded for each pulse as crystals are continually replenished by a microcrystal suspension in aqueous buffer flowing across the FEL beam in a vacuum in a fine liquid jet.

The Coherent X-ray Imaging (CXI) beamline (*19*) at the Linac Coherent Light Source (LCLS) enables high-resolution data collection using the SFX approach (*20*). We used this instrument to obtain diffraction data from in vivo grown crystals of TbCatB produced in the baculovirus-infected *Spodoptera frugiperda* (baculovirus-Sf9) insect cell system (*11*) (Fig. 1, A and B). Crystals with average dimensions of about 0.9 by 0.9 by 11 μm^3 (fig. S1) were sent in a 4-μm-diameter column of buffer fluid at room temperature, at a flow rate of 10 μl/minute, by using a liquid microjet (*21*). X-ray pulses from the FEL were focused onto this column to a spot 4 μm in diameter, before the breakup of the jet into drops (fig. S2). Single-pulse diffraction patterns of randomly oriented crystals that, by chance, were present in the interaction region, were recorded at a 120-Hz repetition rate by a Cornell-SLAC pixel array detector (CSPAD) (*19*, *20*) at 9.4-keV photon energy (1.3 Å wavelength). An average pulse energy of 0.6 mJ at the sample (4×10^{11} photons per pulse) with a duration of less than

[1]Joint Laboratory for Structural Biology of Infection and Inflammation, Institute of Biochemistry and Molecular Biology, University of Hamburg, and Institute of Biochemistry, University of Lübeck, at Deutsches Elektronen-Synchrotron (DESY), Notkestrasse 85, 22607 Hamburg, Germany. [2]German Centre for Infection Research, University of Lübeck, 23538 Lübeck, Germany. [3]Center for Free-Electron Laser Science (CFEL), DESY, Notkestrasse 85, 22607 Hamburg, Germany. [4]Department of Physics, University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany. [5]Max-Planck-Institut für medizinische Forschung, Jahnstrasse 29, 69120 Heidelberg, Germany. [6]Max Planck Advanced Study Group, Center for Free-Electron Laser Science (CFEL), DESY, Notkestrasse 85, 22607 Hamburg, Germany. [7]Linac Coherent Light Source, Stanford Linear Accelerator Center (SLAC) National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA. [8]Department of Chemistry and Molecular Biology, University of Gothenburg, SE-405 30 Gothenburg, Sweden. [9]Photon Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany. [10]Interfaculty Institute of Biochemistry, University of Tübingen, Hoppe-Seyler-Strasse 4, 72076 Tübingen, Germany. [11]Photon Ultrafast Laser Science and Engineering (PULSE) Institute, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA. [12]Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287, USA. [13]Department of Physics, Arizona State University, Tempe, AZ 85287, USA. [14]Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA. [15]Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, SE-75124 Uppsala, Sweden. [16]Institute of Biochemistry and Molecular Biology, University of Hamburg, at DESY, Notkestrasse 85, 22607 Hamburg, Germany.

*Present address: Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA.
†These authors contributed equally to this study.
‡To whom correspondence should be addressed. E-mail: henry.chapman@desy.de (H.N.C.) or christian.betzel@uni-hamburg.de (C.B.)

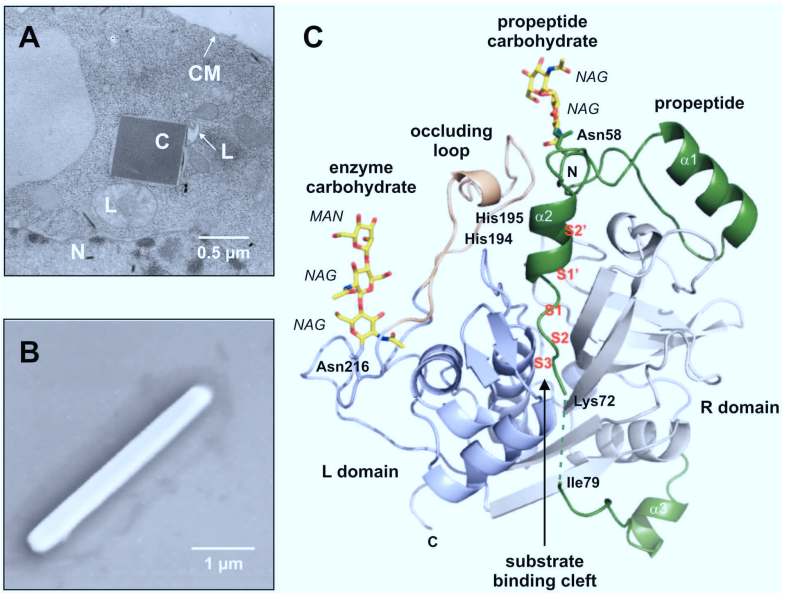


Fig. 1. In vivo grown crystals and three-dimensional structure of the TbCatB-propeptide complex. (**A**) Transmission electron microscopy (EM) of an infected Sf9 insect cell showing a crystal of overexpressed TbCatB inside the rough endoplasmic reticulum that is cut perpendicular to its long axis. N, nucleus; L, lysosome; C, crystal; CM, cell membrane. (**B**) Scanning EM of a single TbCatB crystal after isolation. (**C**) Cartoon plot of the TbCatB-propeptide complex exhibiting the typical papain-like fold of cathepsin B–like proteases (supplementary text S1). Gray, R domain; blue, L domain; beige, occluding loop. The native propeptide (green) blocks the active site. The subsites of the substrate-binding cleft N-terminal (nonprime: S2, S3) and C-terminal (prime: S1′, S2′) to the active site (S1) have been identified by comparison with the human CatB structure (*13*) and labeled (red) according to Schechter and Berger (*27*). Two N-linked carbohydrate structures (yellow) consist of *N*-acetylglucosamine (NAG) and mannose (MAN) residues (yellow, carbon atoms; blue, nitrogen atoms; red, oxygen atoms).

40 fs gave an x-ray intensity above 10^{17} W/cm^2 and a maximum dose of about 31 MGy per crystal. This dose exceeds that tolerable at room temperature with conventional data collection approaches because of the radically different time scales and dose rates. The electron and photon beam parameters are summarized in table S1. Almost 4 million individual "snap-shot" diffraction patterns were collected. Of these, 293,195 snapshots contained crystal diffraction (fig. S3), from which 178,875 (61%) diffraction patterns were indexed and combined into a three-dimensional data set of structure factors by "Monte Carlo" integration of partial reflections from each randomly oriented microcrystal (*22*, *23*). The resulting complete set of structure factors contains 25,969 reflections in a resolution range from 20 to 2.1 Å. The high quality of the merged data set is indicated by an R_{split} of 10.2% (which is a quality measure for SFX instead of R_{merge}) (*23*). Data statistics are summarized in table S2, table S3, and fig. S4. The structure was solved by molecular replacement using the coordinates of the previously determined in vitro crystallized mature TbCatB structure (Protein Data Bank ID, 3MOR) (*11*) as a search model.

The refined SFX TbCatB structure (*R* factor = 18.1%, R_{free} = 21.4%) shares the papainlike fold that is characteristic of cathepsin B–like proteases (Fig. 1C and supplementary text S1) (*24*), with a root mean square deviation of 0.4 Å for equivalent Cα atoms of the mature TbCatB structure determined at 100 K and refined to 2.55 Å resolution (*11*). The molecular replacement solution reveals electron density that is not part of the search model, which we identified as the coordinated, cleaved main part of the propeptide (residues 26 to 72) (Fig. 2A), and as two-carbohydrate

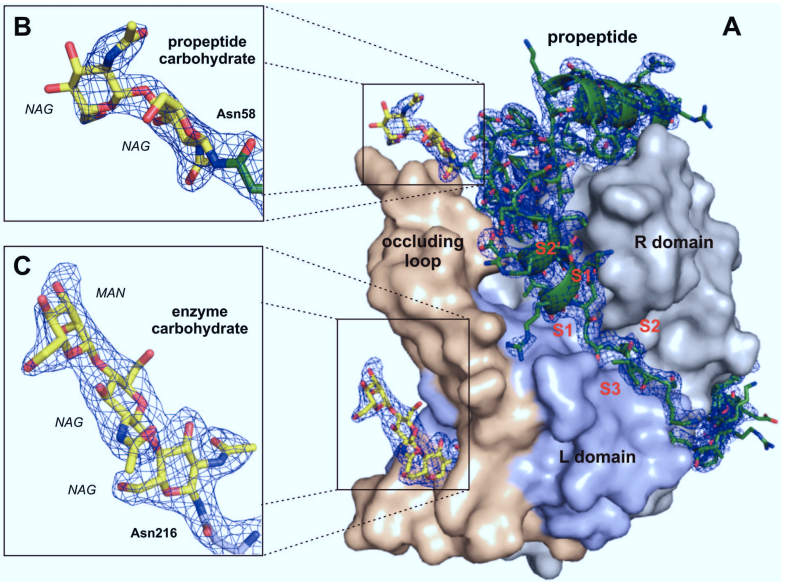
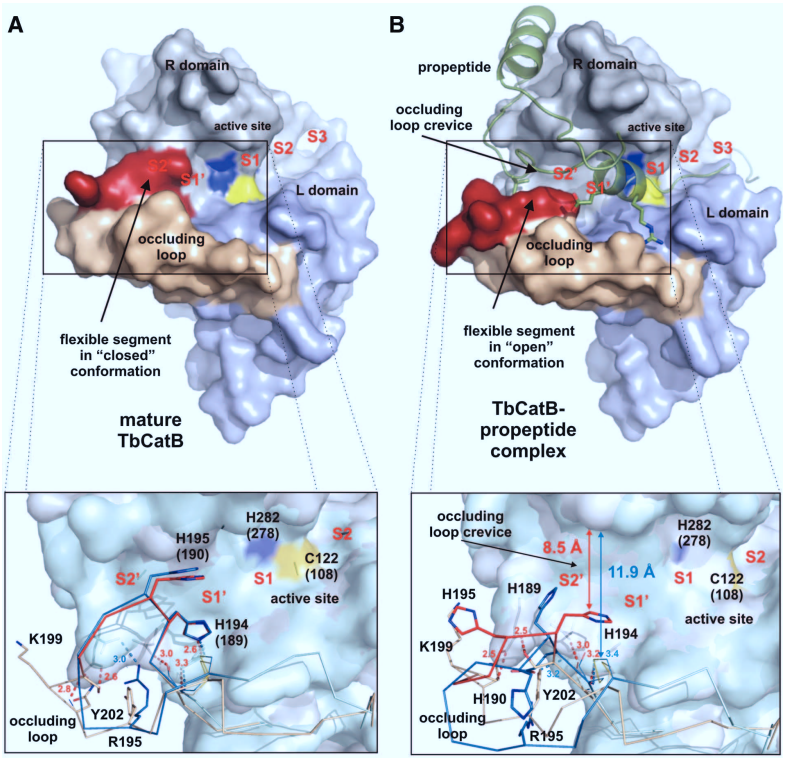


Fig. 2. Quality of the calculated electron density. (**A**) Surface representation of the TbCatB-propeptide complex solved by molecular replacement using the mature TbCatB structure (*11*) as a search model. The solution revealed additional electron density ($2F_{obs} - F_{calc}$, 1σ, blue) of the propeptide (green) that is bound to the V-shaped substrate-binding cleft and of two carbohydrate structures (yellow) N-linked to the propeptide (**B**) and to the mature enzyme (**C**). The propeptide, as well as both carbohydrates, are well-defined within the electron density map (blue), which confirms that the phases are not biased by the search model. Color codes correspond to Fig. 1C.

structures (Fig. 2, B and C). Proteolytic cleavage of the expressed precursor occurs within the propeptide between Ser^{78} and Ile^{79}, as revealed by mass spectrometry, which leaves 15 propeptide residues bound to the N terminus of mature TbCatB (supplementary text S2). The preceding residues Lys^{73} to Ser^{78} are disordered in the crystal structure, owing to a rise in flexibility, which shows up as a gap in the electron density in this region of the propeptide. The cleavage may be



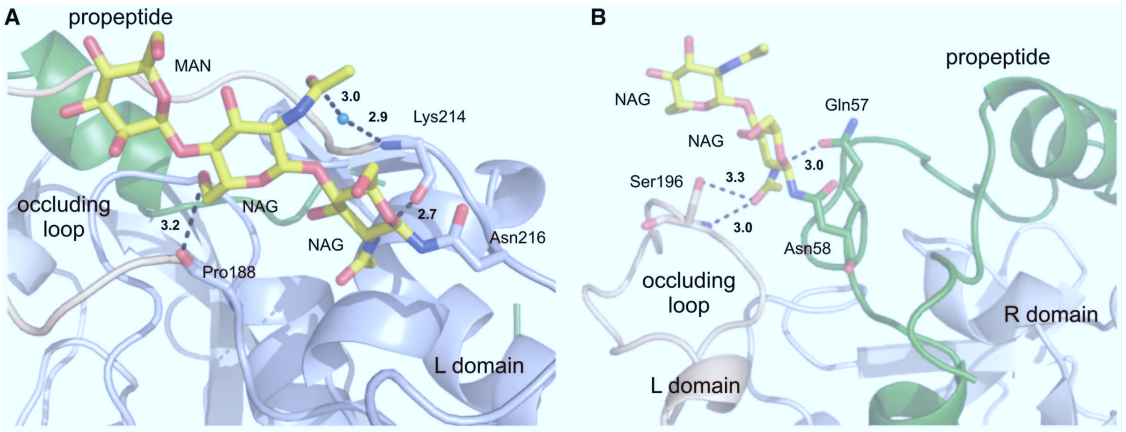


Fig. 3. Occluding loop conformations of mature and propeptide-inhibited TbCatB. (**A**) Surface representation of mature TbCatB (*11*) showing the occluding loop (rigid part, beige; flexible part, red) in the closed conformation. A loop segment blocks the S2′ site and part of the S1′ site of the substrate-binding cleft. (**B**) Propeptide binding (green) shifts the flexible occluding loop segment (red) into an open conformation, which exposes the entire prime subsite. The insets compare the Tb (gray) and human (cyan) surface representations of mature and propeptide bound CatB. Almost superimposable Cα chains of the human (blue) and Tb (beige/red) occluding loops indicate a conservation of the closed loop conformation in the mature enzymes (A), whereas the open conformations show significant differences (B). Four H bonds maintained during conformational transition restricted the flexible loop segment to four residues (red) in TbCatB, which opens a crevice of ~8.5 Å. In contrast, only one H bond is maintained in the open loop conformation of human procathepsin B (*13*). Thus, the mobile segment of the human occluding loop (blue) comprises 10 residues and exposes an enlarged occluding loop crevice (supplementary text S5). Residues are labeled according to TbCatB, referring to the corresponding human CatB numbering in parentheses, if applicable.

Fig. 4. Glycosylation of the TbCatB-propeptide complex. (**A**) Enzyme carbohydrate structure comprising two NAG and one MAN residue (yellow) N-linked to Asn^{216} C-terminal of the occluding loop (beige). The carbohydrate structure connects both occluding loop strands by two direct and one water-bridged H bond (black dashed lines). (**B**) Propeptide glycosylation site comprising two NAG units (yellow) at Asn^{58} within the kinked region of the propeptide (green). The propeptide carbohydrate structure forms an H bond to Gln^{57} of the propeptide and two H bonds to Ser^{196} at the tip of the occluding loop (beige), which stabilize its open conformation. Color codes correspond to Fig. 1C.

part of the initial maturation step within the activation process of TbCatB. The final model of glycosylated TbCatB in complex with its processed, but still-bound, propeptide contains 62 propeptide residues and 247 mature enzyme residues, as well as 98 solvent and 5 carbohydrate molecules. No electron density is observed for 11 flexible amino acid side chains or the eight atoms of the carbohydrate structures.

The SFX TbCatB structure shows that the inhibitory mechanism observed for mammalian papainlike protease-precursors remains largely conserved in *T. brucei*, including the overall conformation of the propeptide (supplementary text S3 and fig. S5). The active site of TbCatB is blocked by the propeptide, which tightly binds in a direction the reverse of the substrate's (fig. S6) (*25*). A detailed comparison of the propeptide-

enzyme contact area with that observed for human procathepsin B (Protein Data Bank ID, 3PBH) (*13*) indicates an interface enlarged by ~310 Å^2 within the TbCatB-propeptide complex (supplementary text S4). Tight binding of the *T. brucei* propeptide to the enzyme interface through three conserved epitopes is maintained by 21 intermolecular polar and ionic interactions (fig. S7). These are eight fewer interactions than for human procathepsin B.

The most significant difference between the structures of mature TbCatB and the natively inhibited propeptide complex occurs in the "occluding loop" region (residues 193 to 207) (fig. S8). This highly flexible loop is a structural element characteristic of cathepsin B–like enzymes that confers exopeptidase activity (removal of dipeptide units from the C terminus of the substrate), which supplements the endopeptidase (nonterminal substrate cleavage) activity common to all papainlike proteases (*26*). In mature CatB, the occluding loop is in the "closed" conformation and buries an essential part of the prime subsite (S1′ and S2′ positions) at the substrate cleft (Fig. 3A) (*27*) and competes for binding with large substrates with an affinity that depends sensitively on pH (*28*). As a consequence of propeptide binding, the occluding loop is reoriented into an "open" conformation, exposing the entire S1′ and S2′ subsite of the substrate-binding cleft in TbCatB (Fig. 3B). This mirrors the open and closed confirmations observed in human CatB; however, the trypanosomal occluding loop is more rigid. The displaced loop segment comprises only 4 residues rather than the 10 observed in human CatB (*13*). This results in a narrower exposed S2′ subsite ~8.5 Å wide compared with ~11.9 Å for human CatB (supplementary text S5). In particular, the side chain of His^{194} is only slightly shifted compared with the closed loop conformation and still extends into the open cleft. Thus, His^{194} not only establishes steric constraints for the substrates but also provides a prominent polar anchor in the otherwise largely hydrophobic S2′ and S1′ subsites that are highly conserved between trypanosome and human CatB (fig. S9). In human CatB, the larger exposed S2′ subsite in the open loop conformation is less restricted by the corresponding His^{189} residue. This suggests that smaller hydrophobic substituents could target the prime site (S1′ and S2′ positions) in TbCatB, which is also supported by the propeptide structures: The bulky Phe residue that sticks into the S2′ subsite of human CatB is replaced by the smaller Met of the *T. brucei* propeptide.

The occluding loop conformation is further stabilized by two carbohydrate structures identified in the TbCatB complex, as shown in Fig. 4. The enzyme carbohydrate chain interacts with both strands of the occluding loop at the loop termini (Fig. 4A), which supports the increased loop rigidity in TbCatB mentioned above. The propeptide carbohydrate connects the tip of the open occluding loop and stabilizes the open conformation (Fig. 4B). Although N-linked oligosaccharide substitution has been detected in human procathepsin B, the predicted glycosylation sites differ from our observations in TbCatB (*28*, *29*). Therefore, it is unlikely that the occluding loop is stabilized in a similar way in the human case (supplementary text S6). Differential glycosylation between the human and *T. brucei* precursors along with the differences in the occluding loop conformation could be exploited for synthetic parasite-specific inhibition.

As illustrated by the room-temperature glycosylated TbCatB-propeptide structure determined here, the combination of in vivo grown microcrystals with the diffraction-before-destruction technique of x-ray FELs provides a compelling path to obtain macromolecular structures from challenging samples. This methodology could vastly speed up structure determination by removing the need for large well-diffracting crystals and providing a suitable amount of crystals of posttranslationally modified proteins, in their biologically functional form.

References and Notes

1. J. A. Frearson *et al.*, *Nature* **464**, 728 (2010).
2. M. P. Barrett, D. W. Boykin, R. Brun, R. R. Tidwell, *Br. J. Pharmacol.* **152**, 1155 (2007).
3. S. Alsford *et al.*, *Nature* **482**, 232 (2012).
4. A. H. Fairlamb, *Trends Parasitol.* **19**, 488 (2003).
5. C. Bryant *et al.*, *Bioorg. Med. Chem. Lett.* **19**, 6218 (2009).
6. M. H. Abdulla *et al.*, *PLoS Negl. Trop. Dis.* **2**, e298 (2008).
7. F. Lecaille, J. Kaleta, D. Brömme, *Chem. Rev.* **102**, 4459 (2002).
8. Y. Yamamoto, M. Kurata, S. Watabe, R. Murakami, S. Y. Takahashi, *Curr. Protein Pept. Sci.* **3**, 231 (2002).
9. G. Dubin, *Cell. Mol. Life Sci.* **62**, 653 (2005).
10. I. D. Kerr, P. Wu, R. Marion-Tsukamaki, Z. B. Mackey, L. S. Brinen, *PLoS Negl. Trop. Dis.* **4**, e701 (2010).
11. R. Koopmann *et al.*, *Nat. Methods* **9**, 259 (2012).
12. K. Tomoo, *Curr. Top. Med. Chem.* **10**, 696 (2010).
13. M. Podobnik, R. Kuhelj, V. Turk, D. Turk, *J. Mol. Biol.* **271**, 774 (1997).
14. R. M. Bill *et al.*, *Nat. Biotechnol.* **29**, 335 (2011).
15. R. J. Southworth-Davies, M. A. Medina, I. Carmichael, E. F. Garman, *Structure* **15**, 1531 (2007).
16. C. Riekel, *J. Synchrotron Radiat.* **11**, 4 (2004).
17. H. N. Chapman *et al.*, *Nature* **470**, 73 (2011).
18. A. Barty *et al.*, *Nat. Photonics* **6**, 35 (2012).
19. S. Boutet, G. J. Williams, *New J. Phys.* **12**, 035024 (2010).
20. S. Boutet *et al.*, *Science* **337**, 362 (2012).
21. U. Weierstall, J. C. Spence, R. B. Doak, *Rev. Sci. Instrum.* **83**, 035108 (2012).
22. R. A. Kirian *et al.*, *Opt. Express* **18**, 5713 (2010).
23. T. A. White *et al.*, *J. Appl. Cryst.* **45**, 335 (2012).
24. M. E. McGrath, *Annu. Rev. Biophys. Biomol. Struct.* **28**, 181 (1999).
25. J. R. Chagas, M. Ferrer-Di Martino, F. Gauthier, G. Lalmanach, *FEBS Lett.* **392**, 233 (1996).
26. C. Illy *et al.*, *J. Biol. Chem.* **272**, 1197 (1997).
27. I. Schechter, A. Berger, *Biochem. Biophys. Res. Commun.* **27**, 157 (1967).
28. O. Quraishi *et al.*, *Biochemistry* **38**, 5017 (1999).
29. Y. Chen, C. Plouffe, R. Ménard, A. C. Storer, *FEBS Lett.* **393**, 24 (1996).

Acknowledgments: Experiments were carried out in February 2011 at the LCLS, a national user facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. This work was supported by the following agencies: the German Federal Ministry for Education and Research (grants 01KX0806 and 01KX0807), the Hamburg Ministry of Science and Research and Joachim Herz Stiftung as part of the Hamburg Initiative for Excellence in Research and the Hamburg School for Structure and Dynamics in Infection (SDI), the Deutsche Forschungsgemeinschaft (DFG) Cluster of Excellence "Inflammation at interfaces" (EXC 306), the DFG, the Landesgraduiertenförderung Baden-Württemberg, the Max Planck Society, the Swedish Research Council, the Swedish Strategic Research Foundation, the Swedish Foundation for International Cooperation in Research and Higher Education, the U.S. Department of Energy Office of Basic Energy Sciences through PULSE Institute at SLAC, the U.S. Department of Energy through Lawrence Livermore National Laboratory under the contract DE-AC52-07NA27344 and supported by the University of California Office of the President Lab Fee Program (award no. 118036), the NSF (award MCB-1021557 and MCB-1120997), and the NIH (award 1R01GM095583). **Author contributions:** H.N.C., J.C.H.S., S.B., P.F., I.S., M.D., L.R., and C.B. conceived the experiment, which was designed with A.B., R.A.K., J.C.H.S., D.P.D., U.W., R.B.D., M.J.B., R.L.S., and H.F.; R.K., S.M., and T.B. performed the in vivo crystallization experiments under the supervision of M.D.; FEL samples were prepared and characterized by F.S., K.N., L.R, and D.R. under the supervision of C.B. and H.N.C.; SFX experiments were carried out by K.N., L.R., H.N.C., D.P.D., F.S., M.L., T.A.W., A.A., M.J.B., U.W., A.B., L.G., S. Bajt, R.A.K., R.B.D., R.L.S., L.L., D.A., L.C.J., C.C., R.N., G.K., C.K., P.F., D.W., I.G., R.F., T.C., N.A.Z., N.T., M.S.H., M.F., J.S., S.B., M.M., M.M.S., and G.J.W. Beamline setup was done by S.B., G.J.W., and M.M. The development and operation of the sample delivery system was performed by R.B.D., D.P.D., U.W., R.L.S, L.L., J.S., and J.C.H.S.; K.N., T.A.W., R.A.K., A.A., A.V.M., L.L., S.K., T.R.M.B., I.S., and H.N.C. analyzed the data. K.N., L.R., and C.B. performed molecular replacement, refined the structure, and calculated the electron density maps. The manuscript was prepared by L.R., C.B., K.N., M.D, I.S., A.B., and H.N.C. with discussions and improvements from all authors. The structure factors and coordinates have been deposited with the Protein Data Bank (accession code 4HWY). The Arizona Board of Regents, acting for and on behalf of Arizona State University and in conjunction with R.B.D., U.W., D.P.D., and J.C.H.S., has filed U.S. and international patent applications on the nozzle technology applied herein. One of the patents was granted on 25 September 2012, U.S. 8,272,576 "Gas dynamic virtual nozzle for generation of microscopic droplet streams."

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1229663/DC1
Materials and Methods
Supplementary Text S1 to S6
Figs. S1 to S10
Tables S1 to S5
References (*30*–*42*)

3 September 2012; accepted 14 November 2012
Published online 29 November 2012;
10.1126/science.1229663

In vivo protein crystallization opens new routes in structural biology

Protein crystallization in cells has been observed several times in nature. However, owing to their small size these crystals have not yet been used for X-ray crystallographic analysis. We prepared nano-sized *in vivo*–grown crystals of *Trypanosoma brucei* enzymes and applied the emerging method of free-electron laser-based serial femtosecond crystallography to record interpretable diffraction data. This combined approach will open new opportunities in structural systems biology.

Protein crystallization occurs as a native process *in vivo*. Prominent examples of this biological self-assembly include storage proteins in seeds, enzymes within peroxisomes and insulin within secretory granules[1]. Cells seem to control interactions of these proteins through changes in the ionic environment, proteolysis of precursor proteins or specific binding partners. Although these structures regulate cellular functions, *in vivo* crystallization has been perceived as atypical behavior and has therefore been neglected in comparison to the considerable efforts devoted to understanding and optimizing *in vitro* protein crystallization for X-ray structure determination.

Heterologously expressed proteins are also able to form crystals within cells. In the baculovirus-Sf9 expression system, virions are coated with a crystalline polyhedrin matrix[2], representing a functional biological crystallization system. Site-specific transposition of an expression cassette into a baculovirus shuttle vector replaces the polyhedrin gene with a gene of interest. The permanent activation of the polyhedrin gene promoter ensures high local protein concentrations, obviously one prerequisite for crystal formation *in vivo*. This system has been used in insect cells to successfully crystallize polyhedrin[3] and chimeric proteins consisting of a polyhedrin part attached to the protein of interest[4], as well as two polyhedrin-free subunits of calcineurin[5]. However, no structural data from *in vivo*–grown, polyhedrin-free crystals have been obtained so far, a gap largely attributed to the small size of the crystals, limited by the maximum diameter of the living cell.

In this study we present an approach for structural biology that combines the recently established method of serial femtosecond X-ray crystallography (SFX)[6] and the use of *in vivo*–grown crystals to record diffraction data. The SFX method uses coherent X-ray pulses produced by an X-ray free-electron laser (FEL) that are over a billion times more brilliant than third-generation synchrotron sources and aims to overcome resolution limits imposed by radiation damage at conventional sources[7]. Using the 'diffraction before destruction' principle, diffraction patterns are collected with single, ultrafast pulses that essentially terminate before the onset of significant structural changes occurs, and the X-ray pulse finally vaporizes the sample[6]. The single-pulse diffraction pattern has been predicted to represent the undamaged crystal injected across the beam in a liquid microjet, depending on the pulse intensity and duration[8]. As a first proof of principle, a recent study has recorded tens of thousands of snapshots from individual crystals of photosystem I (PSI) at room temperature (17–20 °C), ranging in size from 200 nm to 2 m (ref. 6). The successful structural investigation of this protein complex up to a resolution of 8.5 Å demonstrates the viability of using nanocrystals and the SFX method for structure determination.

We observed *in vivo* crystallization of polyhedrin-free, glycosylated cathepsin from *Trypanosoma brucei* (TbCatB) within Sf9 insect cells transfected with bmon14272 bacmid (Invitrogen) created by on site–specific transposition with pFastBac1 expression plasmid (Invitrogen) containing the *Autographa californica* multiple nuclear polyhedrosis virus polyhedrin (PH) promoter. Knockdown of TbCatB has been shown to be lethal for the parasite, which causes human African trypanosomiasis, one of the most important neglected diseases, affecting over 60 million people in central Africa[9]. Efficient and cost-effective drugs are not yet available, but cysteine proteases such as TbCatB have been identified as potential drug targets in protozoan parasites[10].

Approximately 70 h after infection, the formation of needle-shaped microstructures was visible by light microscopy in Sf9 cells infected with recombinant baculovirus containing the gene encoding the pre-pro form of TbCatB (including the TbCatB signal sequence, the pro-peptide and the mature enzyme sequence of TbCatB; **Fig. 1a**). Electron microscopy (EM) revealed a damaged cell surface and sharp, needle-like crystals 10–15 m in length and 0.5–1 m in width spiking out of the cells (**Fig. 1b**). Capsids were visible within the nucleus, and crystals appeared as defined dark areas with sharp square edges within the cytoplasm (**Fig. 1c**). Membranes decorated with ribosomes surrounding the crystals indicate an origin of crystal formation within the endoplasmic reticulum (**Fig. 1d**). An ordered lattice structure observed at higher magnification identified these particles as protein crystals (**Fig. 1e**).

Usually, protein accumulation is inhibited by the 'unfolded protein response' (UPR), in which atypical or misfolded proteins inhibit further protein biosynthesis[11]. This transcriptional regulation fails in the case of the polyhedrin promoter, leading to an enormous increase of TbCatB concentrations that finally provokes crystallization. This interpretation is consistent with the observation that changing the signal sequence from the trypanosomal to the insect cell signal peptide led to secretion of soluble

A full list of authors and affiliations appears at the end of the paper.

RECEIVED 2 MAY 2011; ACCEPTED 21 DECEMBER 2011; PUBLISHED ONLINE 29 JANUARY 2012; DOI:10.1038/NMETH.1859

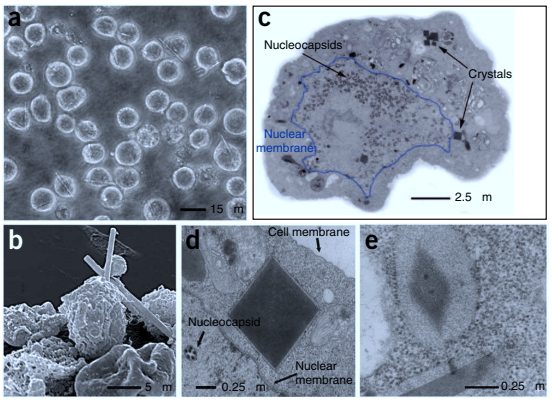


Figure 1 | Light microscopic and EM analysis of Sf9 insect cells with embedded *in vivo* crystals. (**a**) Light micrograph of Sf9 cells infected with TbCatB virus 90 h after infection. (**b**) Transmission EM (TEM) micrograph of an embedded and sectioned infected Sf9 cell with crystals cut perpendicular to the long axis of the needle. Nuclear membrane is outlined in blue. (**c**) Scanning EM micrograph of a group of Sf9 cells infected with TbCatB virus 80 h after infection. (**d**) TEM micrograph of a sectioned sample, showing a crystal cut perpendicular to the long axis of the needle with surrounding membrane between nuclear and cell membrane. (**e**) TEM micrograph showing the lattice structure of a crystal and a longitudinal section of a second crystal (both crystals are surrounded by membrane).

TbCatB without crystal formation (data not shown). In contrast to polyhedrin[3], all crystals appeared without embedded virions. During the progress of infection, the number of crystals continuously increased, until more than 70% of the cells contained one or more crystals. If released during infection-mediated lysis, crystals either remained attached to cell remnants or floated freely within the medium.

To isolate *in vivo* crystals, we lysed cells and subjected them to differential centrifugation, resulting in ~5 × 10^5 purified crystals from 10^6 cells obtained after 8 d in suspension culture. These crystals were stable in deionized water, high- or low-salt solutions and alkaline buffers but became soluble in buffers at pH values below 4. Analysis of the solubilized protein confirmed that the major constituent of the crystals was glycosylated TbCatB comprising the C-terminal residues 63 to 93 of the propeptide and the mature enzyme sequence (residues 94 to 336; **Supplementary Note** and **Supplementary Fig. 1**).

Synchrotron radiation–based experiments showed that the isolated TbCatB *in vivo*–grown crystals were too small to generate suitable X-ray diffraction at beamline X13 of DORIS III (HASYLAB/DESY, Hamburg, Germany), probably owing to technical limitations of the beamline associated with the crystal size rather than to the diffraction quality of the crystal. However, strong diffraction was observed at the Linac Coherent Light Source (LCLS, Menlo Park, California, USA). Purified *in vivo* crystals were crushed by vigorous stirring with glass beads to increase the particle density and injected across the FEL beam in a vacuum in a stream of water using a liquid jet (**Supplementary Fig. 2**). Single-pulse diffraction patterns were recorded at up to 7.5-Å resolution, corresponding to the technical resolution limit determined by the photon energy of the available X-ray pulses (2.0 keV, = 6.2 Å) and the detector geometry. During 23.1 min, 83,224 frames were obtained. After background subtraction, 988 of the frames contained distinct diffraction signals of three or more Bragg spots, each corresponding to a 'snapshot' diffraction pattern from a different randomly oriented crystal (**Fig. 2a**). Quality measures indicate that the FEL dataset conforms to diffraction data from a macromolecular crystal (**Supplementary Fig. 3** and **Supplementary Table 1**). In the limited time available for data collection, it was not possible to obtain a complete dataset, as indicated by the granularity in the summed powder diffraction rings (**Fig. 2b**) and by the statistics of the dataset (**Supplementary Table 1**). As the redundancy was far too low for structure-factor extraction by Monte Carlo integration[12], an overall R_{split} value (T.A.W. *et al.*, unpublished data) was not calculated. Detailed comparison to corresponding values of a similar number of diffraction patterns recorded from photosystem I nanocrystals using SFX[6] showed that the data quality improves with the number of patterns, confirming that additional diffraction data will result in a complete TbCatB dataset. However, 879 (89%) of the 988 recorded diffraction patterns were indexed without unit cell constraints. The raw lattice parameters were a = 122.9 Å, b = 123.6 Å, c = 53.4 Å, = 90.3°, = 90.2° and = 90.3°; therefore, the unit cell was assigned to be tetragonal, with $a = b$ = 123.3 Å and c = 53.4 Å. Multiple measurements of individual Bragg reflections were averaged, and intensities from symmetry-equivalent reflections were merged according to the point group 4/*mmm*; a pseudo-precession pattern of the [001] zone is shown in **Figure 2c**. Our results reveal that structural information can be obtained from *in vivo* grown crystals of a glycosylated protein, providing a second independent proof of the SFX method.

In addition, PNGase F–treated TbCatB from isolated and solubilized *in vivo* crystals was recrystallized *in vitro* using the vapor diffusion technique. X-ray diffraction data were collected at the Swiss Light Source (Villigen, Switzerland). The structure was determined by molecular replacement using the bovine CatB

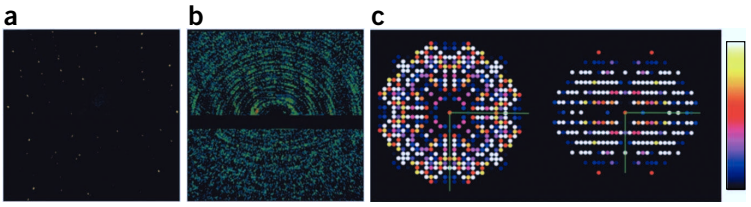


Figure 2 | Serial femtosecond crystallography of *in vivo* TbCatB crystals. (**a**) Diffraction pattern of a TbCatB *in vivo* crystal recorded from a single shot of 70 fs FEL X-rays. (**b**) Sum of 988 single-shot FEL diffraction patterns from TbCatB crystals in different orientations. The lower panel of the detector was shifted to achieve higher resolution (Online Methods). At the edge of the detector, a maximum resolution of 7.5 Å was obtained. (**c**) Precession-style image of the [001] zone for TbCatB, obtained by merging SFX data from 328 *in vivo* crystal patterns indexed with unit cell constraints. Intensities of integrated reflections were normalized to values between 0 and 1 on a linear scale.

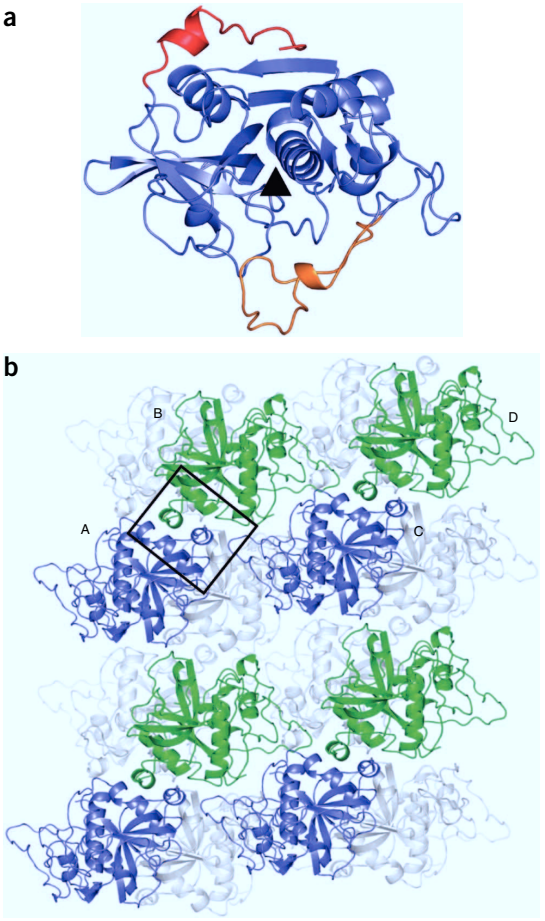


Figure 3 | Structure of solubilized and recrystallized TbCatB solved by conventional X-ray crystallography. (**a**) Ribbon tracing of TbCatB, showing the pro-peptide in red and the occluding loop in orange. The black triangle indicates the position of the catalytic cleft. (**b**) Cartoon representation of the crystal packing. The crystals contain two molecules in their asymmetric unit and four molecules in the unit cell. This diagram of crystal contacts shows four molecules labeled A–D. In the foreground, every second TbCatB molecule is shown in blue or green, respectively. The major contact area is boxed. Gray ribbons represent additional molecules in the crystal packing.

structure (PDB 1QDQ) as a model and refined to 2.55-Å resolution (**Supplementary Table 2**). The final structure of TbCatB (PDB 3MOR) shows the characteristic papain–cathepsin B fold, including 16 of the pro-peptide residues (Ser78 to Pro93; **Fig. 3a**), and is highly similar to a structure of nonglycosylated recombinant TbCatB that has been independently reported[13] (PDB 3HHI). Here we refer to our own structural data, as we prepared our crystals from protein obtained from the *in vivo*–grown crystals.

Despite deglycosylation treatment, our TbCatB structure still contains a glycan side chain at Asn216, clearly identified in the electron density. This suggests that, although deglycosylation was effective as judged from SDS-PAGE results (**Supplementary Fig. 1c**), it removed only the glycan at Asn76. As the essential crystal contacts primarily involve hydrophobic patches within the persisting pro-peptide of TbCatB (**Fig. 3b**), this pro-peptide might also influence crystal formation *in vivo* (**Supplementary Note**). Moreover, the structure shows distinct differences from human CatB (PDB 1GMY), which are particularly important to consider for rational drug discovery investigations (**Supplementary Note**).

The advantages of *in vivo* crystallization are (i) production of post-translationally modified proteins of interest, (ii) easy isolation by spinning down the crystals after cell lysis, (iii) the possibility of preliminarily analyzing crystals by EM and (iv) a narrow crystal size distribution that is ideal for SFX applications. As the use of non-insect cell signal peptides is explored further, it seems likely that more proteins will form *in vivo* crystals within insect cells. Recently, we obtained similar results with an inosine monophosphate dehydrogenase from trypanosomes (TbIMPDH). Under comparable experimental conditions to those reported for TbCatB, TbIMPDH *in vivo* crystals appeared after 5 to 6 d. Besides TbCatB and calcineurin, this is a third example of *in vivo*–grown crystals from a heterologously expressed protein in Sf9 insect cells. TbIMPDH crystals were isolated and subjected to a brief SFX diffraction test at LCLS. The ability of these *in vivo* crystals to diffract up to a resolution of more than 8 Å (data not shown) strongly supports a more general applicability of the approach described in this study.

Although progress has been made in using microfocus beamlines that allow data collection from crystals only 1 m in size, radiation damage is an inherent problem of X-ray crystallography[14]. Typically, large and well-ordered crystals of 20–500 m are required for conventional X-ray crystallography, often a serious challenge to grow[15]. Therefore, the unique combination of *in vivo* crystallization and serial femtosecond crystallography offers notable new possibilities for proteins that do not form crystals suitable for conventional X-ray diffraction *in vitro*, extending the available methods of structural biology particularly when, in due time, shorter wavelengths of the FEL will provide higher-resolution data.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Protein Data Bank: 3MOR (coordinates and structure factors for *in vitro*–recrystallized TbCatB).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

FEL experiments were carried out at LCLS in June 2010 (TbCatB) and in August 2011 (TbIMPDH), a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The X-ray diffraction experiments on recrystallized TbCatB crystals were carried out at beamline X06DA of the Swiss Light Source (Villigen, Switzerland). This work was supported in part by a grant from the Deutsche Forschungsgemeinschaft (DFG), from the Swedish Research Council, from the Knut och Alice Wallenbergs Stiftelse, from the European Research Council, as well as by US National Science Foundation award MCB-1021557. R.K. received a fellowship from the Landesgraduiertenförderung Baden-Württemberg. L.R., D. Rehders and C. Betzel thank the German Federal Ministry for Education and Research for funding (grants 01KX0806 and 01KX0807). Support from the Hamburg Ministry of Science and Research and Joachim Herz Stiftung as part of the Hamburg Initiative for Excellence in Research and the Hamburg School for Structure and Dynamics in infection, and from the DFG Cluster of Excellence "Inflammation at Interfaces" (EXC 306) is gratefully acknowledged.

Funding for the development and operation of the CFEL-ASG multipurpose (CAMP) instrument within the Advanced Study Group at the Center for Free-Electron Laser Science was provided by the Max Planck Society. M.J.B., R.G.S. and C.Y.H. acknowledge funding from the US Department of Energy Office of Basic Energy Sciences through the Photon Ultrafast Laser Science and Engineering (PULSE) Institute at the Stanford Linear Accelerator Center (SLAC) National Accelerator Laboratory.

AUTHOR CONTRIBUTIONS

R.K., K.C. and L.R. contributed equally to this work. R.K. performed the *in vivo* crystallization experiments under the supervision of M.D.; K.C. prepared samples for synchrotron X-ray crystallography and collected and analyzed data under the supervision of T.S.; H.N.C. and J.C.H.S. conceived the SFX experiment, which was designed with P.F., A.B., R.A.K., J.S., D.P.D., U.W., R.B.D., M.J.B., I.S., H.F. and J.H.; FEL samples were prepared by L.R., D. Rehders and C. Betzel; SFX experiments were carried out by L.R., K.N., H.N.C., D.P.D., F.S., M.L., T.A.W., A.A., M.J.B., C.Y.H., R.G.S., U.W., A.B., R.A.K., R.B.D., N.C., R.L.S., L.L., J.D., M.S.H., C. Bostedt, J.D.B., S. Boutet and G.J.W.; beamline setup was done by C. Bostedt, J.D.B., S. Boutet, G.J.W. and M.M. The delivery system was developed and operated by R.B.D., D.P.D., U.W., J.C.H.S., P.F., L.L. and R.L.S.; S.W.E., B.E., L.F., H.G., A.H., R.H., G.H., H.H., P.H., N.K., C.R., D. Rolles, B.R., A.R., H.S., L.S., J.U., C.G.W. and G.W. operated the CAMP instrument and the pn junction charge-coupled devices and developed the software for pnCCD readout. Diffraction instrumentation was developed and calibrated by H.N.C., A.B., A.A., J.S., D.P.D., U.W., R.B.D., M.J.B., L.G., J.H., M.M.S., N.T., J.A., S.S., S. Bajt, M.B. and J.C.H.S. Data were analyzed by T.A.W., K.N., F.S., A.B., R.A.K., A.A., F.R.N.C.M., A.V.M., L.L., N.C., L.F., N.K., G.W., P.H., C.C., I.S., T.E., J.H., S.K., X.W., H.N.C. and J.C.H.S. The manuscript was prepared by L.R., M.D., C. Betzel and T.S. with discussion and improvements from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index/html.

1. Doye, J.P.K. & Poon, W.C.K. *Curr. Opin. Colloid Interface Sci.* **11**, 40–46 (2006).
2. Rohrmann, G.F. *J. Gen. Virol.* **67**, 1499–1513 (1986).
3. Coulibaly, F. *et al. Nature* **446**, 97–101 (2007).
4. Ijiri, H. *et al. Biomaterials* **30**, 4297–4308 (2009).
5. Fan, G.Y. *et al. Microsc. Res. Tech.* **34**, 77–86 (1996).
6. Chapman, H.N. *et al. Nature* **470**, 73–77 (2011).
7. Owen, R.L., Rudino-Pinera, E. & Garman, E.F. *Proc. Natl. Acad. Sci. USA* **103**, 4912–4917 (2006).
8. Chapman, H.N. *et al. Nat. Phys.* **2**, 839–843 (2006).
9. Mackey, Z.B., O'Brien, T.C., Greenbaum, D.C., Blank, R.B. & McKerrow, J.H. *J. Biol. Chem.* **279**, 48426–48433 (2004).
10. Bryant, C. *et al. Bioorg. Med. Chem. Lett.* **19**, 6218–6221 (2009).
11. Kitamura, M. *Int. Rev. Immunol.* **30**, 4–15 (2011).
12. Kirian, R.A. *et al. Opt. Express* **18**, 5713–5723 (2010).
13. Kerr, I.D. *et al. PLoS Negl. Trop. Dis.* **4**, e701 (2010).
14. Cusack, S. *et al. Nat. Struct. Biol.* **5** (suppl.) 634–637 (1998).
15. Mueller, M., Jenni, S. & Ban, N. *Curr. Opin. Struct. Biol.* **17**, 572–579 (2007).

Rudolf Koopmann[1,22], Karolina Cupelli[1,22], Lars Redecke[2,22], Karol Nass[3], Daniel P DePonte[4], Thomas A White[4], Francesco Stellato[4], Dirk Rehders[2], Mengning Liang[4], Jakob Andreasson[5], Andrew Aquila[4,21], Sasa Bajt[6], Miriam Barthelmess[6], Anton Barty[4], Michael J Bogan[7], Christoph Bostedt[8], Sébastien Boutet[8], John D Bozek[8], Carl Caleman[4], Nicola Coppola[4,21], Jan Davidsson[5], R Bruce Doak[9], Tomas Ekeberg[5], Sascha W Epp[10,11], Benjamin Erk[10,11], Holger Fleckenstein[4], Lutz Foucar[10,12], Heinz Graafsma[6], Lars Gumprecht[4], Janos Hajdu[5,21], Christina Y Hampton[7], Andreas Hartmann[13], Robert Hartmann[13], Günter Hauser[14], Helmut Hirsemann[6], Peter Holl[13], Mark S Hunter[15], Stephan Kassemeyer[12], Richard A Kirian[9], Lukas Lomb[12], Filipe R N C Maia[16], Nils Kimmel[14,17], Andrew V Martin[4], Marc Messerschmidt[8], Christian Reich[13], Daniel Rolles[10,12], Benedikt Rudek[10,11], Artem Rudenko[10,11], Ilme Schlichting[10,12], Joachim Schulz[4], M Marvin Seibert[5,20], Robert L Shoeman[12], Raymond G Sierra[7], Heike Soltau[13], Stephan Stern[4], Lothar Strüder[10,14,17,18], Nicusor Timneanu[5], Joachim Ullrich[10,11], Xiaoyu Wang[9], Georg Weidenspointner[14,17], Uwe Weierstall[9], Garth J Williams[8], Cornelia B Wunderer[6], Petra Fromme[15], John C H Spence[9], Thilo Stehle[1,19], Henry N Chapman[3,4], Christian Betzel[20] & Michael Duszenko[1]

[1]Interfaculty Institute of Biochemistry, University of Tübingen, Tübingen, Germany. [2]Joint Laboratory for Structural Biology of Infection and Inflammation, Institute of Biochemistry and Molecular Biology, University of Hamburg, and Institute of Biochemistry, University of Lübeck, at Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany. [3]Institute for Experimental Physics, University of Hamburg, Hamburg, Germany. [4]Center for Free-Electron Laser Science, DESY, Hamburg, Germany. [5]Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. [6]Photon Science, DESY, Hamburg, Germany. [7]Photon Ultrafast Laser Science and Engineering (PULSE) Institute, Stanford Linear Accelerator Center (SLAC) National Accelerator Laboratory, Menlo Park, California, USA. [8]Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, California, USA. [9]Department of Physics, Arizona State University, Tempe, Arizona, USA. [10]Max Planck Advanced Study Group, Center for Free-Electron Laser Science, Hamburg, Germany. [11]Max-Planck-Institut für Kernphysik, Heidelberg, Germany. [12]Max-Planck-Institut für Medizinische Forschung, Heidelberg, Germany. [13]PNSensor GmbH, Munich, Germany. [14]Max-Planck-Institut Halbleiterlabor, Munich, Germany. [15]Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona, USA. [16]National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [17]Max-Planck-Institut für extraterrestrische Physik, Garching, Germany. [18]University of Siegen, Siegen, Germany. [19]Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [20]Institute of Biochemistry and Molecular Biology, University of Hamburg, Hamburg, Germany. [21]Present addresses: European XFEL GmbH, Hamburg, Germany (A.A. and N.C.) and Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, California, USA (J.H.). [22]These authors contributed equally to this work. Correspondence should be addressed to M.D. (michael.duszenko@uni-tuebingen.de).

ONLINE METHODS

Cloning. The gene coding for the pre-pro-form of TbCatB (GeneDB Tb927.6.560) was amplified by PCR using primers 5 -TAGGATCCATGCATCTCATGCGTGCCT-3 (sense) and 5 -TAACTCGAGCTACGCCGTGTTGGGTG-3 (antisense), and AccuPrime Taq DNA polymerase (Invitrogen) according to the manufacturer's instructions. After subcloning (TOPO-TA cloning kit, Invitrogen) into XL1-Blue–competent *Escherichia coli* cells (Stratagene), plasmid DNA purification (QIAprep spin miniprep kit, Qiagen) and digestion with BamHI and XhoI, the extracted gel fragment (QIAquick gel extraction kit, Qiagen) was cloned into pFastBac1 expression plasmid (Invitrogen) and transformed into DH10Bac-competent *E. coli* cells (Invitrogen) according to the manufacturer's instructions. The TbCatB gene, containing its own signal peptide sequence, replaced the polyhedrin gene, including the polyhedrin nuclear import signal.

Recombinant Bacmid DNA was purified according to the Bacmid isolation protocol (Invitrogen) using the QIAprep spin miniprep kit (Qiagen) and subsequently used for PCR analysis of the cloned sequence. Correctness of the PCR products was verified by sequencing. Bacmid DNA was then used for lipofection with Sf9 insect cells grown in serum-free medium at 27 °C to generate recombinant virus stock according to the Bac-to-Bac manual (Invitrogen). This stock was used to generate a high-titer virus stock for further infections (titer: 1×10^8 plaque-forming units (p.f.u.) ml^{-1}).

Expression of TbCatB. Recombinant virus stock was used to infect a 70%-confluent monolayer culture of Sf9 insect cells (Invitrogen), grown in serum-free medium at 27 °C with a multiplicity of infection of 0.1 p.f.u. per cell. After 72–96 h (upon visual inspection of the amount of crystals within cells by light microscopy), we harvested the cells by rinsing and scraping them from the surface of the cell flask and then centrifuging the cell suspension at 1,000*g* for 5 min. For greater amounts of protein material, suspension cultures with agitation at 110 r.p.m. at 27 °C were infected and harvested as described above.

Electron microscopy. For TEM, infected insect cells of a 75-mm^2 confluent monolayer culture were fixed using 2% (vol/vol) glutaraldehyde in 0.2 M sodium cacodylate buffer containing 0.12 M sucrose for 1 h at 41 °C. After washing four times (10 min each) and storing overnight in sodium cacodylate buffer, cells were postfixed in 1.5% (wt/vol) osmium tetroxide and stained in 0.5% (wt/vol) uranyl acetate. Dehydration in ethanol, clearing in propylene oxide, embedding in Agar 100 resin and sectioning were performed according to standard procedures. Sections were stained in 5% (wt/vol) uranyl acetate and 0.4% (wt/vol) lead citrate. For scanning electron microscopic analysis, the same fixation and staining protocol was applied up to 70% (vol/vol) ethanol. Cells were then placed on polylysine-covered coverslips and dehydrated with 96% and 100% ethanol for 8 h each. Critical-point drying and gold-palladium sputter staining were performed using standard protocols. TEM was performed using a Zeiss EM10 device; for scanning electron microscopy, a Cambridge Stereo Scan 250 Mk2 device was used.

Isolation and solubilization of crystals. Harvested cells were washed twice in PBS and lysed in RIPA lysis buffer (50 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1% (vol/vol) Nonidet P-40, 0.25% (wt/vol) sodium deoxycholate, 0.1% (wt/vol) SDS, 0.01% (wt/vol) sodium azide, pH 7.4), intensely vortexed and incubated on ice for 15 min. To reduce viscosity, we added 25 units ml^{-1} Benzonase (Novagen) before incubating the mixture for 30 min at 37 °C or overnight at 4 °C. The sample was then washed several times with PBS. The pellet was resuspended in solubilization buffer (50 mM sodium acetate, pH 3.5), incubated for 15 min on ice and centrifuged at 16,000*g* for 15 min. The supernatant was concentrated using 10-kDa NMWL-Centricon devices, thereby changing the buffer to 50 mM sodium acetate, 4 mM DTT, 0.5 mM EDTA (pH 5.5) or another appropriate buffer, depending on the application.

Serial femtosecond crystallography. TbCatB *in vivo* crystals were analyzed by the SFX method[6] at LCLS in the SLAC National Accelerator Laboratory (Menlo Park, California, USA)[16]. The X-ray FEL generated intense quasi-monochromatic X-ray pulses of 67-fs duration with 6.7×10^{12} photons per pulse and a wavelength of 6.2 Å (2 keV). Focused onto the crystal, a single pulse reaches an intensity of ~0.5×10^{18} W cm^{-2} and a dose to the sample of about 1 to 3 GGy, equivalent to 100 times the Garman safe limit[7], depending on the number of TbCatB monomers within the unit cell. The detailed electron and photon beam parameters are summarized in **Supplementary Table 3**. To ensure that there was always, on average, one crystal passing through the 100-fl X-ray interaction volume of the liquid jet, corresponding to the theoretically required optimal volume/hit rate ratio, we had to increase the crystal number density before measurement. Therefore, ~5×10^7 TbCatB crystals isolated from a single 100-ml culture of 2×10^6 virus-infected insect cells per ml were crushed by vigorous vortexing with glass beads for 30 min at room temperature, concentrated by centrifugation for 30 s at 14,500*g* and filtered through a 2- m filter (A-430, Upchurch Scientific). Thus, ~10^9 crystals in 1 ml ddH_2O were obtained.

The experiment was performed at the Atomic, Molecular and Optical Science beamline[17] using the CFEL-ASG multipurpose (CAMP) instrument[18]. A liquid microjet[19] focused by a coaxial flow of gas to a diameter of about 4 m at a flow rate of 15 l min^{-1} was used to introduce crystals stored in a sample loop[6] into the FEL interaction region. The interaction region of the X-rays and the crystals was located in the continuous liquid column, before the Rayleigh break-up of the jet into drops. Testing of TbCatB microcrystal delivery through the gas-focused jet at SLAC in the PULSE Institute enabled selection of the optimum jet nozzle before installation on the CAMP apparatus. In 23.1 min, 83,224 frames were recorded with a pair of movable, high-frame-rate, low-noise pn junction charge-coupled device (pnCCD) detector modules[18] (each panel consisted of 512 pixels × 1,024 pixels, each 75 m^2 × 75 m^2 in area) operating at the 60-Hz repetition rate of the FEL pulses. The upper panel of the detector was located 65 mm from the interaction point; the lower panel was mounted to $z = 68$ mm to record scattering from 3.51° to 49.02° in the vertical plane. The 988 identified single-crystal diffraction patterns were processed to subtract background noise and then indexed by an autoindexing procedure based on DirAx[20] to determine the unit cell parameters and therefore the symmetry of the lattice. Details of data processing are given in ref. 6.

Activity assay. The cathepsin B activity kit (BioVision) was used to determine the enzyme activity of solubilized TbCatB. This assay uses the preferred cathepsin B substrate sequence RR labeled with amino-4-trifluoromethyl coumarin. Fluorescence emission at 495 nm was measured using a PerkinElmer LS 55 fluorescence spectrometer with an excitation wavelength of 400 nm. The chemical identity of the included CatB inhibitor was not released by the manufacturer.

Deglycosylation. Solubilized protein in 10 mM HEPES and 20 mM NaCl (pH 7) was deglycosylated overnight at 25 °C by addition of 1 unit g^{-1} N-glycosidase F (PNGase F, New England Biolabs). Sample was then concentrated using 10 kDa NMWL-Centricon devices to a concentration of 2 mg ml^{-1} in 10 mM HEPES and 20 mM NaCl, pH 7.

Recrystallization and structure determination. We grew all crystals with the sitting-drop vapor diffusion technique by mixing TbCatB (2 mg ml^{-1} in 10 mM HEPES and 20 mM NaCl, pH 7) and precipitant solution (22–30% polyethylene glycol 3000 (wt/vol) and 200–400 mM $(NH_4)_2HPO_4$) in a 1:1 ratio. Crystals were flash-frozen for X-ray data collection at 100 K using 20% (vol/vol) glycerol as cryoprotectant.

Diffraction data were collected at beamline X06DA at the Swiss Light Source (Villigen, Switzerland) using a data collection wavelength of 1.000 Å and a MarCCD detector (Mar Research), and processed with XDS[21]. Initial phases were obtained by molecular replacement with PHASER[22] using the bovine CatB structure[23] as a search model (PDB 1QDQ). The structure was refined using rigid-body refinement and simulated annealing with PHENIX[24]. Subsequent refinement was carried out by alternating rounds of model-building with Coot[25] and restrained refinement using two-fold noncrystallographic symmetry restraints with Refmac[26]. The final model had a low free *R*-factor of 24.3% and good stereochemistry (**Supplementary Table 2**). The structure reported here was aligned with the TbCatB structure determined in ref. 13 (PDB 3HHI) using the secondary structure–matching superposition protocol of the program Coot[25]. The buried surface area was calculated with AREAIMOL[27]. **Figure 3a**,**b** was prepared with PyMOL (DeLano Scientific)[28].

16. Emma, R. *et al. Nat. Photonics* **4**, 641–647 (2010).
17. Bozek, J.D. *Eur. Phys. J. Spec. Top.* **169**, 129–132 (2009).
18. Strüder, L. *et al. Nucl. Instrum. Methods Phys. Res. A* **614**, 483–496 (2010).
19. DePonte, D.P. *et al. J. Phys. D Appl. Phys.* **41**, 195505 (2008).
20. Duisenberg, A.J.M. *J. Appl. Cryst.* **25**, 92–96 (1992).
21. Kabsch, W. *J. Appl. Cryst.* **26**, 795–800 (1993).
22. Read, R.J. *Acta Crystallogr. D Biol. Crystallogr.* **57**, 1373–1382 (2001).
23. Yamamoto, A. *et al. J. Biochem.* **127**, 635–643 (2000).
24. Adams, P.D. *et al. Acta Crystallogr. D Biol. Crystallogr.* **58**, 1948–1954 (2002).
25. Emsley, P. & Cowtan, K. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
26. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240–255 (1997).
27. Collaborative Computational Project, Number 4. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760–763 (1994).
28. DeLano, W.E. *The PyMOL Molecular Graphics System* (DeLano Scientific, San Carlos, California, 2002).

doi:10.1038/nmeth.1859

Journal of
Applied Crystallography
ISSN 0021-8898

Received 20 July 2011
Accepted 18 January 2012

CrystFEL: a software suite for snapshot serial crystallography

Thomas A. White,[a]* Richard A. Kirian,[b] Andrew V. Martin,[a] Andrew Aquila,[a] Karol Nass,[a,c] Anton Barty[a] and Henry N. Chapman[a,c]

[a]Center for Free-Electron Laser Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany, [b]Department of Physics, Arizona State University, Tempe, Arizona 85287, USA, and [c]University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany. Correspondence e-mail: taw@physics.org

In order to address the specific needs of the emerging technique of 'serial femtosecond crystallography', in which structural information is obtained from small crystals illuminated by an X-ray free-electron laser, a new software suite has been created. The constituent programs deal with viewing, indexing, integrating, merging and evaluating the quality of the data, and also simulating patterns. The specific challenges addressed chiefly concern the indexing and integration of large numbers of diffraction patterns in an automated manner, and so the software is designed to be fast and to make use of multi-core hardware. Other constituent programs deal with the merging and scaling of large numbers of intensities from randomly oriented snapshot diffraction patterns. The suite uses a generalized representation of a detector to ease the use of more complicated geometries than those familiar in conventional crystallography. The suite is written in C with supporting Perl and shell scripts, and is available as source code under version 3 or later of the GNU General Public License.

1. Introduction

The new technique of serial femtosecond crystallography (Chapman *et al.*, 2011) involves the illumination of many small crystals of proteins sequentially using an intense fourth-generation light source such as the Linac Coherent Light Source (LCLS; Emma *et al.*, 2010). Each pulse of the X-ray laser lasts for only a few tens or hundreds of femtoseconds and, since the radiation dose is very large, the crystal will be destroyed. Each crystal is used for only one exposure, and there is not sufficient time for any oscillation or rotation of the sample. As a result, the final integrated Bragg intensities must be constructed from 'snapshot' diffraction patterns containing partially recorded intensities, each pattern corresponding to a different crystal, and with no orientational relationships between the crystals. Furthermore, when the crystals are very small, the Fourier transform of the crystal shape itself (the 'shape transform') may come to dominate the sizes of the peak profiles. Each snapshot provides one slice through this shape transform; however, the required reflection intensities are proportional to its volume integral. In addition, the size and form of the shape transform vary from crystal to crystal. The Monte Carlo method of integration (Kirian *et al.*, 2010) provides a theoretical method to process data in this situation but relies on a very high redundancy in the data set. This fact makes it necessary to index and measure intensities from many thousands of patterns, so unsupervised automatic processing of such data is of the utmost importance. The software developed as a result could be applied to any technique in which 'snapshot' diffraction patterns are acquired in such a 'serial' manner.

A new software suite, called *CrystFEL*, has been created to address these concerns. Three programs are central to the suite in the initial version:

(1) *indexamajig*, for quickly indexing and integrating large numbers of diffraction patterns.

(2) *pattern_sim*, for diffraction pattern simulation.

(3) *process_hkl*, for merging Bragg intensities using the Monte Carlo method.

In addition to these, extra programs are provided to help with the individual stages of the data analysis:

(1) *check_hkl*, for calculating figures of merit for merged data.

(2) *compare_hkl*, for examining the differences between two sets of merged intensities.

(3) *render_hkl*, for plotting intensities, structure factors and redundancies in two and three dimensions.

(4) *sum_stack*, for summing diffraction patterns after peak detection to produce a two-dimensional 'virtual powder pattern', which can be used to quickly evaluate the amount of data collected.

(5) *powder_plot*, for summing data from a wider range of formats (image, reflection list or peak-list form) into one-dimensional 'powder' traces.

(6) *get_hkl*, which can perform various manipulations on reflection data, such as artificially 'twinning' their intensities, expanding them out to point groups of lower symmetry, adding noise or filtering reflections according to a template file.

(7) *hdfsee*, an image viewer.

CrystFEL accepts image data contained within a hierarchical data format (HDF5) container. The image viewer, *hdfsee*, may be used to examine images in this format, and can use a calibrated detector geometry or can overlay peak locations from the indexing or peak detection steps. Future versions of the software will incorporate an abstraction layer to allow the use of many more formats, enabling full use of the flexible detector geometry specification system described in the next section.

The software is intended to process 'clean' diffraction patterns, meaning that steps to remove electronic artefacts should have already been performed. Blank images in which no crystal intersected the

computer programs

X-ray beam at the moment of the X-ray pulse should not be included in the input as far as possible, but the inclusion of such images should have no effect on the processing other than the speed. Since crystal hit rates in experiments so far have been around 5%, attempting to index all frames without this selection procedure would increase the indexing time by a factor of around 20. However, since these pre-processing steps are likely to be specific to a particular detector or experimental configuration, they are not implemented by any part of the *CrystFEL* suite. One possible route for performing this pre-processing is to create a piece of software based on the LCLS data analysis tools '*myana*', '*pyana*' or '*psana*'. *CrystFEL*'s indexing component could, in principle, be executed directly by these processing steps to create a streamlined data processing path.

The tightly knit structure of the suite, where file formats and code are shared between programs, has enabled a high degree of code reuse. This not only simplified the structure of the suite but led to a constant re-visitation of many parts of the code. This has resulted in many bugs, and even potential bugs, being removed at an early stage.

2. Detector geometry

Fourth-generation light sources have brought with them new detector technology, required to match the repetition rates called for by the 'serial crystallography' methodology. Many of these detectors, such as the pnCCD detector used in the CAMP instrument (Strüder *et al.*, 2010) or the CSPAD detector used in the CXI instrument at the LCLS, consist of multiple smaller detectors in some fixed or movable arrangement. The small sizes of the panels, combined with separate sets of readout electronics for each panel, help to achieve the required high readout rates. To take this into account, *CrystFEL*'s representation of a detector is broken down into one or more 'regions', each of which has its own camera length, position, resolution and other parameters. Programs forming part of the suite take a description of the detector geometry in a text file, allowing the use of the suite with many and varied detector geometries. To avoid the many problems that can arise from confusion over definitions of geometry – all too familiar to macromolecular crystallographers – a precise specification is used to define the detector geometry, illustrated in Fig. 1 and described below.

The raw data of each panel fit into the array of data taken from the input file, with the relevant range of pixels defined only in terms of minimum and maximum coordinates in the 'fast-scan' (fs) and 'slow-scan' (ss) directions. 'Fast scan' refers to the direction whose coordinate changes most quickly as the bytes in the input file are moved through in order, and 'slow scan' refers to the direction whose coordinate changes most slowly. All pixels in the input data block must be assigned to a panel, but regions of the detector can be marked as 'bad' if required, which means that they will be ignored at all stages of the analysis.

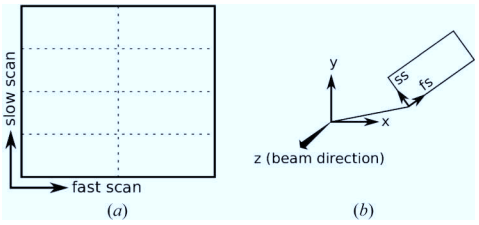


Figure 1
Specification of detector geometry in *CrystFEL*. (*a*) The input data array is broken into rectangular regions by specifying the minimum and maximum coordinates in the 'fast-scan' and 'slow-scan' directions, which are unambiguously defined with respect to the arrangement in memory of the data array itself. (*b*) For each region, the position of the corner (closest to the start of the data array) and the vectors corresponding to the fast-scan (**fs**) and slow-scan (**ss**) directions are specified in terms of a fixed laboratory coordinate system.

The role of the geometry description file is to set up the relationship between pixel locations in the raw image data and in the laboratory coordinate system. The laboratory coordinate system is defined by *CrystFEL* to have $+\mathbf{z}$ being the beam direction, $+\mathbf{y}$ pointing towards the zenith (directly upwards) and $+\mathbf{x}$ completing a right-handed coordinate system. However, this definition does not place any requirements on the representation of the data in the file. For each panel, the geometry description file specifies the coordinates in the laboratory coordinate system of the corner of the panel, meaning the point in the image that would appear first in the raw image array, in units of pixels. The file must then specify the direction, also in the laboratory coordinate system, that corresponds to each of the fast- and slow-scan directions. The direction is defined as a linear combination of the **x** and **y** directions, constituting a transformation matrix, and so arbitrary in-plane rotations of the detector are possible. Since there is no requirement for the direction vectors to have equal moduli, rectangular pixels could be accommodated, if it were to become necessary in the future, by more creative use of the vector combinations such as $\mathbf{fs} = \mathbf{x}$ and $\mathbf{ss} = 2\mathbf{y}$. Hexagonal pixels could also be used within this framework, with some wastage of space in the input data array.

3. Simulation of data: *pattern_sim*

It is important to be able to simulate data in order to test the algorithms. A fast nanocrystal diffraction simulation program, named *pattern_sim*, is included in *CrystFEL*, which simulates patterns in a manner similar to that described in the previous literature (Kirian *et al.*, 2010). The unit-cell dimensions are taken from a Protein Data Bank (PDB; Berman *et al.*, 2000) file; however, the structure factors themselves must be calculated by a separate means, such as using the *sfall* tool in *CCP4* (Winn *et al.*, 2011). The program is, in its initial version, only able to model the crystals as parallelepipeds with a specified number of unit cells along each of the three crystallographic axes, meaning that more complex shapes such as a hexagonal prism or a spherical crystal cannot currently be used. However, this restriction allows the intensity to be calculated as the product of the Laue interference functions and the squared structure factors, which is much more efficient than the equivalent sum over all unit cells:

$$I = \frac{\sin^2(\pi n_a \mathbf{q}\cdot\mathbf{a})}{\sin^2(\pi \mathbf{q}\cdot\mathbf{a})}\frac{\sin^2(\pi n_b \mathbf{q}\cdot\mathbf{b})}{\sin^2(\pi \mathbf{q}\cdot\mathbf{b})}\frac{\sin^2(\pi n_c \mathbf{q}\cdot\mathbf{c})}{\sin^2(\pi \mathbf{q}\cdot\mathbf{c})}\left|\mathbf{F}_\mathbf{q}\right|^2, \qquad (1)$$

where n_a, n_b and n_c are the number of unit cells that the parallelepiped has along the **a**, **b** and **c** directions, respectively. The vector **q** represents the point in reciprocal space at which the diffraction is to be calculated, and **a**, **b** and **c** are the direct-space unit-cell axes (which encode the orientation of the crystal as well as the shape of the unit cell). $\mathbf{F}_\mathbf{q}$ is the complex structure factor at reciprocal space point **q**. Suitable expressions for the scattering vector **q** are given by Kirian *et al.* (2010).

The program calculates the Laue functions in advance for the three required values of n_a, n_b and n_c in terms of $\mathbf{q}\cdot\mathbf{a}$, $\mathbf{q}\cdot\mathbf{b}$ and $\mathbf{q}\cdot\mathbf{c}$, respectively. By storing these values in lookup tables, the values required later can be quickly calculated by interpolation. This method avoids repetitive calculation of trigonometric functions and

has the additional advantage of providing a numerically stable calculation when the scalar product of the reciprocal space vector and a cell axis is zero or very small.

Since the pre-calculated structure factors give only the intensities when Bragg's law is exactly satisfied, further calculation must be performed to find the values between the Bragg peaks. Three methods are available in *pattern_sim*: 'mosaic', where the structure factor at the nearest reciprocal lattice point is taken; 'interpolate', where the intensities at the nearby reciprocal lattice points are interpolated trilinearly; and 'phased', where the relative phases of neighboring structure factors are taken into account. The 'phased' method accounts for the dark region that should appear between two extended Bragg peaks that have structure factors with opposite phases, but requires the most calculation. The 'mosaic' method is expected to be the least computationally demanding. For the highest possible accuracy, a future version of the software could allow a full, oversampled, three-dimensional molecular transform to be input.

The program calculates the absolute scattered intensity by multiplying the result of equation (1) by the incident photon flux density, the square of the Thomson scattering length and the solid angle of the pixel. The calculation is repeated for a number of sub-pixel units and the mean of the resulting values taken, in order to reduce the probability of missing fine structure in the pattern and consequently underestimating the intensity in each Bragg peak. Finite wavelength spread of the incident radiation can also be simulated by a similar method of sampling many different closely spaced wavelengths. In addition, convergence of the incident X-ray beam can be simulated within the limits of a small-angle approximation.

To further accelerate the calculation and enable the simulation of large data sets of tens of thousands of patterns or more, *pattern_sim* can take advantage of a graphics processing unit (GPU) *via* OpenCL, if it is available. The GPU calculation can be enabled by setting a command line switch and gave a speed-up of around 30 times on the test hardware. A separate test program, easily executed amongst other test programs as part of the installation procedure, verifies that no significant differences exist between the two implementations of the calculation. The typical total difference is around 0.3% of the total intensity given by the 'conventional' version, this small difference perhaps being accounted for by the single precision of the GPU's floating-point arithmetic as opposed to the double precision used in the conventional version.

In a final 'post-processing' step, *pattern_sim* can add Poisson noise to the results. It then stores the image in an HDF5 file suitable for input into the indexing stage or other processing pipelines, if required.

4. Pattern indexing and integration: *indexamajig*

The purpose of the indexing component of *CrystFEL*, *indexamajig*, is to facilitate the processing of large numbers of diffraction patterns in an automated and largely unsupervised manner. It does not, in the current version, implement any autoindexing algorithms itself but rather takes advantage of the previous work in this field by executing other indexing programs as sub-processes. In the initial version, *DirAx* (Duisenberg, 1992) and *MOSFLM* (Leslie, 2006) can be used. The user can select more than one indexing method, in which case the program will try the later specified methods should the earlier methods fail to yield a result that passes some basic checks.

Indexamajig takes a list of image filenames as its input. For each image in the list it performs peak detection and sends the peaks to the selected auto-indexing programs in the format required by those programs. Alternatively, peak lists generated by previous processing steps and incorporated in the HDF5 files can be used. If indexing is successful, a unit cell is read back from the auto-indexer and, optionally, compared against a reference cell. Cell comparison can be performed *via* a variety of methods, the default being to check all possible linear combinations of the cell basis vectors for correspondence, within a certain tolerance, to the axes of the reference cell. When indexing using *MOSFLM*, the lattice symmetry (if known) may be used to restrict the number of candidate unit cells returned. *DirAx* has no such option and simply returns a list of possible primitive unit cells.

It is further required by the software that the unit-cell vectors form a right-handed basis after the matching process, meaning that Bijvoet pairs are not confused with one another. This could potentially allow the extraction of an anomalous diffraction signal subject to considerations described in the next section.

If the cell is found to match to the sought unit cell, predicted peak positions are calculated for the image and compared with the initial peak positions sent into the auto-indexing program. A basic check for correctness is performed, where the result is rejected if fewer than 10% of the initial peak positions are close to predicted locations, to within a tolerance defined by the program input. A lower success rate might be expected for the smallest nanocrystals, where the peaks in the diffraction pattern arise from the extended tails of the shape transforms.

The method of unit-cell reduction described above is vulnerable to errors in cases where the length of one unit-cell axis is close to a small multiple of the length of another. This situation is analogous to reticular twinning familiar in conventional crystallography. For such cases, the user may opt instead to compare the lattice vectors without combining them in linear combinations, although this would be expected to result in a lower number of successfully indexed patterns. Alternatively, the cell-matching procedure can be entirely disabled, in which case the result from the auto-indexer is used directly, provided the basic check described above is passed.

Once the pattern has been successfully indexed, intensities are integrated from the predicted peak locations, thereby including peaks that were not found by the initial peak search. In the data evaluated so far, the peak shapes in the image were found to vary widely within a single image (Fig. 2), in a way similar to that seen in simulated patterns for small crystals as a result of the extended shape transforms surrounding each reciprocal lattice point. Therefore, *indexamajig* does not attempt two-dimensional profile fitting in the current version and a method of summing pixel counts within a fixed radius is used instead. The radius of integration can be configured in the detector geometry file. The local background surrounding the peak is estimated and subtracted by measuring the intensity in a thin ring surrounding the integration region. Improved methods for background estimation and subtraction and proper calculation of the errors in the integrated intensities, and for automated rejection of

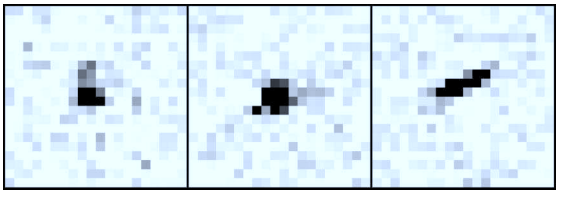


Figure 2
Three peak profiles from a single diffraction pattern from photosystem I, taken from the data described by Chapman *et al.* (2011).

peaks that cannot be satisfactorily integrated, are under development.

The resulting intensities from each image are written to a text file, alongside other information such as the image filename, the reciprocal axis vectors and optionally the peak locations from the initial peak search. The results from all input images are sequentially written to the same text file – which is referred to as a 'stream' – for ease of later handling. The overall flow of diffraction pattern processing by *indexamajig* is shown in Fig. 3.

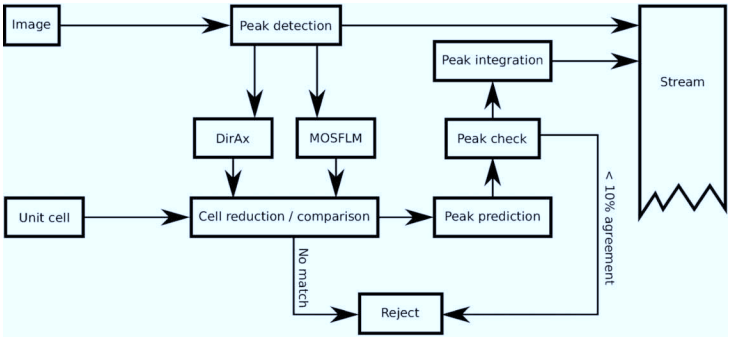


Figure 3
Flow diagram of diffraction pattern processing in *indexamajig*.

The indexing and integration of many thousands of diffraction patterns may take many hours depending on the speed of computer and data retrieval, and *indexamajig* is able to run many indexing jobs in parallel to reduce the time required for the whole data set. *Indexamajig* writes reports of the rate of processing and the 'yield' of the process, defined as the ratio of the number of successfully indexed patterns to the overall number of patterns, to the terminal at intervals of a few seconds.

Once a 'stream' has been compiled for a data set, the results of indexing can be visualized by running a short Perl script, called *check_near_bragg*. This script finds, for each image in sequence, the positions of the peaks predicted by indexing, the coordinates of which are also stored in the stream. The script then runs the image viewer *hdfsee* to show the image and the peak locations (Fig. 4). When the viewer window is closed, the script displays the next image from the stream in the same way. A very similar script called *check_peak_detection* operates in the same way, but displays the results of the initial peak search instead of the predicted peaks.

5. Merging of intensities: *process_hkl*

Merging of individual intensity measurements is performed by the method of Kirian *et al.* (2010), in which the mean of all measurements for each individual reflection is taken. When the number of measurements is large, this procedure constitutes a Monte Carlo integration over the three-dimensional reflection profiles, resulting in values that are proportional to the integrated intensity. A well known feature of Monte Carlo methods is that, in the limit of a sufficiently large quantity of data, all stochastic variables (such as variations in X-ray pulse intensity or crystal size) will be 'integrated out' and become constant factors affecting all intensities equally. For data obtained using a free-electron laser source, further complications arise because of the stochastic nature of the lasing process: the incident beam intensity, wavelength and spectrum are different for each pulse. It is clear that a large number of indexed patterns are necessary for this method to be successful, justifying the highly automated processing of patterns described in the previous section.

The program *process_hkl* performs merging using this method taking into account the symmetry of the structure. It was found to be convenient for the software to neglect information about systematic absences due to glide planes and screw axes. Instead, only the point symmetry of the structure is considered when merging intensities.

The program is also able to perform scaling of the intensities in an attempt to improve the quality of the results. Scaling can be performed by normalizing the intensities according to the mean intensity of the Bragg peaks in each pattern or the overall total intensity in each pattern, or by a two-pass process where the intensities are scaled to most closely fit the values produced by a previous unscaled run of the program. Improved algorithms are under development.

Since each individual diffraction pattern is indexed independently, ambiguities may result if the symmetry of the structure is lower than that of the lattice. These are precisely the same conditions under which the structure may exhibit twinning by merohedry, and the effect of such an ambiguity when combining results from many patterns will be that the data appear to be perfectly twinned. Unfortunately, all attempts so far to resolve such ambiguities by correlating the intensities in the patterns have failed, perhaps owing to the partialities associated with the individual reflection measurements. As a result, the symmetry according to which the intensities must be merged is dictated by the asymmetric unit that the indexing

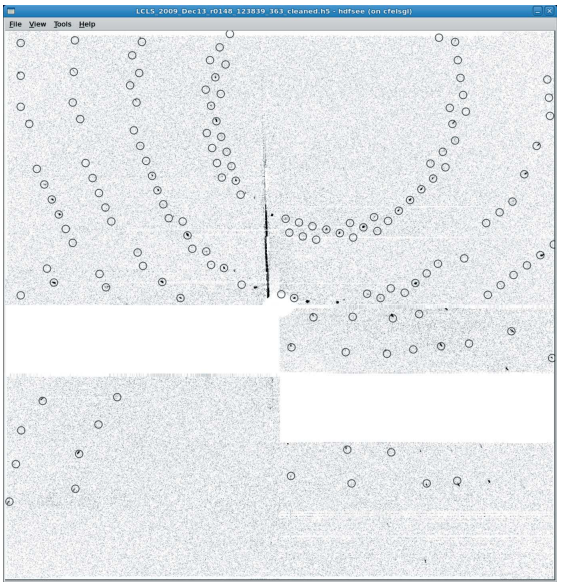


Figure 4
Screenshot of the image viewer *hdfsee* displaying an image from the data described by Chapman *et al.* (2011), with predicted peak locations circled.

procedure can recognize unambiguously. In the merohedral case, this asymmetric unit is reduced in size and so the merging symmetry must be increased. A printable table is provided as part of the documentation for *CrystFEL* listing the 230 space groups according to point group, Laue class and holohedry, with the point groups positioned such that the appropriate merging symmetry can be quickly determined for any given 'true' symmetry.

The above considerations apply also in the case of pseudo-merohedry, for example when an orthorhombic structure has two axis lengths almost equal. In this example, if the difference between the two similar lattice parameters is smaller than the tolerance allowed by the cell-reduction procedure described earlier then the holohedral symmetry for a tetragonal lattice (belonging to Laue class $4/mmm$) must be used instead. It was found useful to introduce the concepts of 'source' and 'target' symmetries when describing the merging process. The source symmetry describes the symmetry that the indexing procedure is able to discern, and the target symmetry describes the symmetry of the true structure. Left coset decomposition (Flack, 1987) of one symmetry group with respect to the other provides the required 'twin laws', which specify the symmetry operations of the ambiguities to be resolved. However, in contrast to the determination of conventional twin laws, mirror and inversion operations are not permitted since the crystallographic axes must form a right-handed basis as described earlier. It should be noted further that no 'twin fraction' is required when describing the ambiguities described here, since the relevant ambiguities can always be identified and merging performed according to the higher symmetry. The contributions from each of the 'sides' of the applicable 'twin laws' are therefore always exactly equal.

In the current implementation, which does not have the means to resolve indexing ambiguities, the target symmetry is always equal to the source symmetry and the resulting intensities must appear twinned. If future versions of the software were to incorporate the required algorithms, the target symmetry could be reduced to the true symmetry of the structure. If the true symmetry were unknown, resolution could be attempted into progressively lower and lower symmetries until no resolution could be successfully performed. If there were multiple ambiguities, a partial resolution into a target group of intermediate symmetry could be performed. This might be useful in specialized cases, perhaps where a full resolution of the ambiguity is difficult and it is considered preferable to have 'twinned' data according to one of the ambiguities than a poor resolution of both. If the intensities are to be merged under the assumption that Friedel's law holds, meaning that Friedel pairs of intensities are to be merged with one another, then the target symmetry can simply be specified as the Laue class corresponding to the appropriate point group.

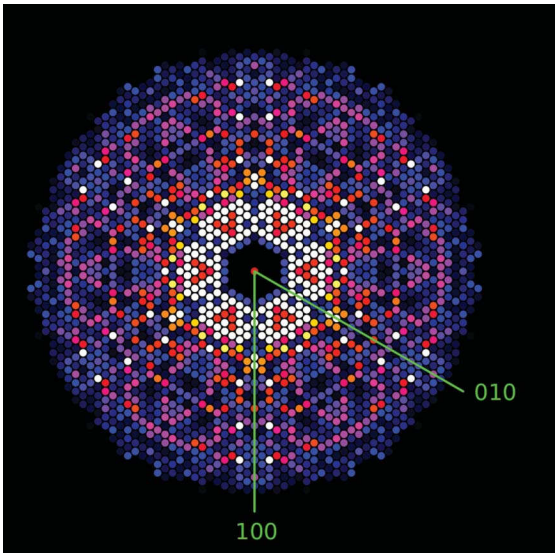


Figure 5
Zone axis structure factors from Chapman *et al.* (2011), plotted using *render_hkl*.

The intensities can be visualized by plotting intensities in flat central sections through reciprocal space (similar to a simulated precession diffraction pattern) using a color scale as shown in Fig. 5. A choice of color scales is available, and the program can also plot the values in three dimensions by creating an input file for the Persistence of Vision ray-tracing program (available from http://www.povray.org/) and invoking it automatically.

A helper script is included which can be invoked once merging has been completed in order to create an MTZ file for import into *CCP4*. This script operates by invoking the *CCP4* program *f2mtz* with a format specification appropriate for *CrystFEL*'s plain text reflection data format.

6. Evaluation of data quality

The evaluation of the data quality is difficult for the type of experiment for which *CrystFEL* has been designed. Since the Monte Carlo merging procedure operates by taking samples from a number of different distributions, it is clear that the individual values to be merged will not share a high degree of similarity. Indeed, it is not desired that they have a high degree of similarity, since the aim is to sample all the underlying distributions as fully as possible, and a full sampling of all the distributions will produce both large and small intensities for any given reflection. As a result, the traditional data quality metrics such as $R_{\rm merge}$ cannot give a meaningful measure of the data quality. A more useful figure can be obtained by splitting the data into two separate interleaved sets, which are merged independently, and then examining the agreement between the two resulting intensity lists. Since the data have been split into two sets, it is expected that the degree of convergence in each subset would be lower and so this method could underestimate the quality of the combined data by a factor of $2^{1/2}$. A suitable figure of merit could therefore be defined as

$$R_{\rm split} = 2^{-1/2}\,\frac{\sum |I_{\rm even} - I_{\rm odd}|}{\frac{1}{2}\sum (I_{\rm even} + I_{\rm odd})}, \qquad (2)$$

where $I_{\rm even}$ represents the intensity of a reflection produced by merging even-numbered patterns, $I_{\rm odd}$ represents the intensity of the equivalent reflection from the odd-numbered patterns and the sum is over all reflections. This method can be performed with *CrystFEL* by using a helper script to split the 'stream' into two, merging the two resulting streams using *process_hkl* and comparing the two merged intensity lists using *compare_hkl*. The program can also calculate the *R* factor as a function of resolution.

A method has been described for estimating the error in the final estimate of the intensity of each reflection in the Monte Carlo method, using

$$\sigma_{hkl} = \left[\sum \left(I_{\rm spot} - \langle I_{hkl}\rangle\right)^2\right]^{1/2} \Big/ N_{hkl}, \qquad (3)$$

where $I_{\rm spot}$ is an individual measurement of the reflection hkl, $\langle I_{hkl}\rangle$ is the mean of all such measurements, N_{hkl} is the number of measurements and the summation is over all measurements of a particular reflection (Chapman *et al.*, 2011, supplementary information). Errors are estimated in this way by *process_hkl* and recorded in the final merged intensity list. By the central limit theorem, the distribution of measured mean intensity values for a given reflection will closely approach a Gaussian provided a sufficiently large number of random samples are taken. However, caution must be used in interpreting the meaning of σ_{hkl} when this condition is not met, in which case the error estimate could over- or underestimate the true error in the intensities.

The program *check_hkl* can calculate statistics, such as the mean $I/\sigma(I)$ or the redundancy and completeness of the data, as a function of resolution.

The overall flow of the indexing, merging and evaluation process is shown in Fig. 6.

7. Creation of 'virtual powder patterns': *powder_plot* and *sum_stack*

Information can be derived from the inspection of serial crystallographic data in 'powder' form, such as when attempting to evaluate the effects of radiation damage on the overall falloff of intensities with resolution. Such analysis can be performed without indexing the individual patterns, for example by summing the patterns themselves (provided the background subtraction is sufficient) or by summing the peaks found by the peak search. *CrystFEL* supports this type of analysis through the programs *sum_stack* and *powder_plot*. The former program, *sum_stack*, performs the usual peak detection on each of its input images and adds pixels within a small circle around each peak to a final image. The latter program, *powder_plot*, creates one-dimensional powder traces from input in many different formats, for example a stream, an individual diffraction pattern or merged intensities. If a stream file is used for input, the user may opt to create the powder plot from the peak locations found by the initial peak search or the integrated intensities after indexing. If the integrated intensities are used, the correct bin to place the intensity in can be calculated by combining the Miller indices either with a provided unit cell or with the unit cell specific to the individual pattern, which can differ from the average unit cell by up to an amount equal to the tolerance of the cell-reduction procedure. A large amount of control can therefore be exercised when performing analysis with virtual powder patterns.

8. Documentation

All programs accept a '`--help`' argument on the command line, which produces a summary of the usage of the program and its various options. Installation instructions detailing the required libraries and other environmental factors are provided, as well as standard 'man' pages describing many of the features in further detail. The symmetry table described earlier is also included, and may also be downloaded separately from the same web site as the software itself. Low-level documentation of the internals of the software, detailing the interfaces available to programs forming part of the suite (such as functions for handling reflection data), is also included.

9. Future work

CrystFEL is a young software project created for use in a very new and rapidly developing field, and so new features and improvements to the analysis pipeline are currently under active development. One continuous development is to improve the yield of the indexing process while filtering out inaccurate indexing results. This will be achieved by implementing new indexing algorithms which operate by searching for *known* reciprocal lattice vector lengths instead of by providing an *ab initio* unit cell for comparison. In addition, interfaces to other auto-indexing programs such as *XDS* (Kabsch, 2010), *DENZO* (Otwinowski & Minor, 1997) and *LABELIT* (Sauter *et al.*, 2004) will be added. The software will also be developed to allow the processing of diffraction patterns corresponding to more than one crystal (Vaughan *et al.*, 2004), which would allow the concentration of crystals in the suspension injected into the path of the X-ray beam to be increased beyond the level at which, on average, one crystal contributes to each pattern.

Improved methods for scaling the intensities will be the subject of much future work. The current method makes no attempt to model the diffraction process, and instead simply averages a large amount of data to obtain an accurate result. By introducing such a model, even a crude one, and fitting parameters such as the incident intensity, the crystal orientation and the wavelength of the radiation, it should be possible to arrive at more accurate estimates of the underlying structure factors, perhaps with a smaller number of patterns. Such a method would be similar to the methods employed in conventional X-ray crystallography, such as post-refinement (Rossmann & van Beek, 1999), but with some complications potentially arising from indexing ambiguities. Initial work in this direction, although not currently usable, is already included in the current version of *CrystFEL* as the program *partialator*.

These improvements will act to reduce the number of diffraction patterns required to obtain an accurate set of intensity data, which is of great importance given the scarcity of experimental time at hard-X-ray free-electron laser sources.

10. Software availability

To ensure the maximum possible use and understanding of the software, *CrystFEL* is available in source code form under version 3 or later of the GNU General Public License (GPL). It can be

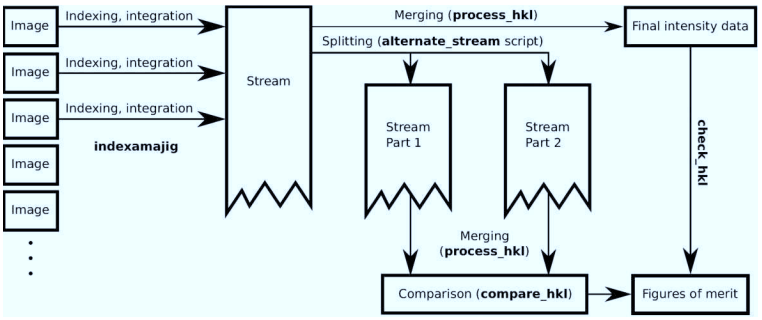


Figure 6
The overall pipeline of the indexing, merging and evaluation workflow. Individual images are indexed and integrated in parallel, and the resulting Bragg intensities written to a single long file known as a 'stream'. The contents of the stream can be merged to produce the final intensity data, or the stream can be split into two smaller streams which can be merged individually. Comparison of the two individually merged results produces figures of merit that can be used to evaluate the data quality. Different figures of merit can be produced from the final intensities themselves.

downloaded from http://www.cfel.de/. Contributions in the form of bug reports, comments and source code patches are actively invited.

Mark Hunter, Francesco Stellato, Linda Johansson, David Arnlund, Nadia Zatsepin and Lorenzo Galli tested the software and provided bug reports as well as feedback on the manuscript. John Spence and Ilme Schlichting also read the manuscript and made corrections and improvements.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Chapman, H. N. *et al.* (2011). *Nature* (*London*), **470**, 73–77.
Duisenberg, A. J. M. (1992). *J. Appl. Cryst.* **25**, 92–96.
Emma, P. *et al.* (2010). *Nat. Photonics*, **4**, 641–647.
Flack, H. D. (1987). *Acta Cryst.* A**43**, 564–568.
Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.
Otwinowski, Z. & Minor, W. (1997). *Methods in Enzymology*, edited by C. W. Carter & R. M. Sweet, Vol. 276, *Macromolecular Crystallography*, part A, pp. 307–326. New York: Academic Press.
Rossmann, M. G. & van Beek, C. G. (1999). *Acta Cryst.* D**55**, 1631–1640.
Sauter, N. K., Grosse-Kunstleve, R. W. & Adams, P. D. (2004). *J. Appl. Cryst.* **37**, 399–409.
Strüder, L. *et al.* (2010). *Nucl. Instrum. Methods Phys. Res. Sect. A*, **614**, 483–496.
Vaughan, G. B. M., Schmidt, S. & Poulsen, H. F. (2004). *Z. Kristallogr.* **219**, 813–825.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Sample Injection for Pulsed X-ray Sources

Daniel P. DePonte*[1] Karol Nass[2], Francesco Stellato[1], Mengning Liang[1], Henry N. Chapman[1,2]
[1]Center for Free-Electron Laser Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany.
[2]University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany.

ABSTRACT

The high intensity of free-electron lasers now allows for the possibility of obtaining measurable diffraction from biological samples with a single X-ray pulse. An important consequence of diffract-before-destroy imaging is that the sample is destroyed and therefore must be replaced preferably at the repetition rate of the FEL. This presents an interesting challenge; the sample must be rapidly replaced within the X-ray focus at the proper particle density and degree of hydration without damaging or denaturing the sample. If particle number density is too high, for example due to clustering or evaporation, the diffraction pattern resulting from coherent illumination of multiple particles may be discarded when sorting for 3D reconstruction. If number density is too low the hit rate, percentage of pulses with measurable scattered intensity, may also be too low to collect a complete data set. Evaporation will also leave behind less volatile material and this change of concentration may be damaging to the sample. On the other hand the similarity in electron density for water and biological material provides poor contrast for fully hydrated material. It is often also necessary to consider sample consumption. While high, near unity, hit rate can be obtained using liquid jets, a liquid flow rate greater then 1 microliter per minute must be maintained. Several sample injection possibilities, drop on demand, aerosols, liquid jets, aerodymamic lenses, have been explored and a review of these results is presented.

Keywords: free electron laser, sample injection

1. INTRODUCTION

The recent development of two techniques in structural biology, X-ray serial crystallography[1] and femtosecond "diffract before destroy" single particle imaging[2, 3] have required new methods of sample delivery to be developed. Common requirements of both techniques are that the samples remain hydrated, undamaged and in their native state, and that the samples are sufficiently thin to permit transmission of X-rays. Additionally it is desired to keep sample consumption low while maintaining a high rate at which usable diffraction patterns are collected. Liquid microjets, rayleigh jets, drop on demand (DOD), gas nebulizers and electrospray have been examined for the purpose of sample delivery and their relative strengths and weaknesses are presented here.

2. INJECTOR TYPES

The bulk of sample injection methods described in the following can be roughly classified as jets or aerosols with some overlap in their description; an aerosol may for example be generated from a jet in a coflowing gas. Liquid jets appear as a continuous column of liquid which results when the speed of liquid flowing from a tube exceeds a critical value. Sprays may be generated and or dispersed for example by electrostatic repulsion or by blasting with gas. An example of a jet is shown in figure 1. A faucet is shown with water dripping and at slightly higher flow rate jetting. In this example the smaller drops are produced at the higher flow rate because disturbances are carried away from the faucet permitting the jet to be accelerated by gravity and reduced in diameter before breakup. Jets can also be accelerated by gas dynamic forces and electrostatic forces to produce much smaller jets as described below. Aerosols are generally too disperse to be of use with micron focus X-ray sources but may be combined with external gas flow or an aerodynamic lenses to increase sample number density.

*daniel.deponte@desy.de; phone +49-40-8998-5784 ; fax +49-40-8998-1958

Advances in X-ray Free-Electron Lasers: Radiation Schemes, X-ray Optics, and Instrumentation,
edited by Thomas Tschentscher, Daniele Cocco, Proc. of SPIE Vol. 8078, 80780M · © 2011
SPIE · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.886780

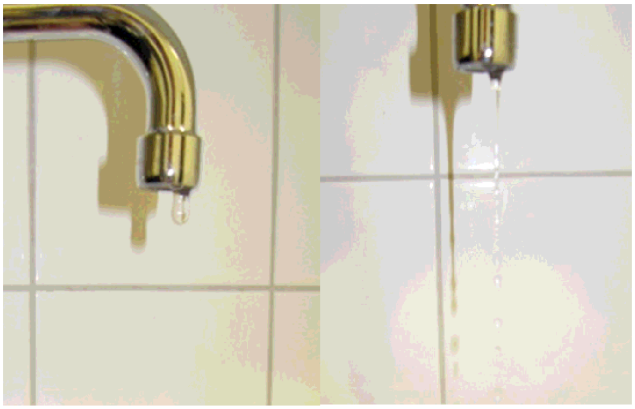


Figure 1. Dripping and jetting from a water faucet.

2.1 Rayleigh jets

Liquid jets have been understood since Raleigh's theoretical studies of jet breakup[4, 5] in the late1800's which showed jets are unstable for disturbances of wavelength greater than π times the jet diameter D_0. Spontaneous breakup occurs at wavelengths around the maximum 4.5 D_0. A more uniform droplet distribution can be produced by driving the jet to breakup at a given frequency using a piezo transducer[6]. An example of a 10μm water droplet stream is shown in figure 2.

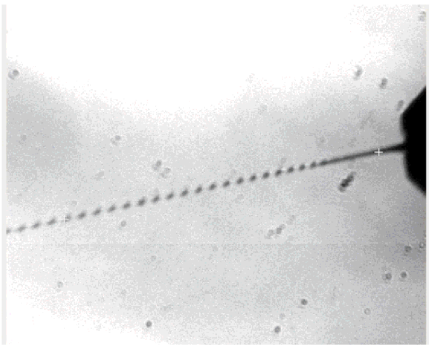


Figure 2. A water jet driven to breakup into 10micron drops by triggering at 200kHz.

Simplicity is the main advantage to working with Rayleigh jets however clogging and sample consumption are two major drawbacks. Rayleigh jets made with converging nozzles require careful cleaning and preparation to produce jets smaller than about ten micron. Sample consumption is also rather large, about 100μl/s for a 10μm jet. A simple way improvement for both these problems is to start with a large nozzle and then accelerate the liquid to form a smaller jet as is done with gas dynamic forces or electrostatic forces.

2.2 Gas dynamic acceleration

By sending a jet through a pressure gradient provided by a coaxially flowing gas, the jet can be accelerated and therefore reduced in diameter. Fabrication of such nozzles is described elsewhere[7]. The nozzle consists of two concentric glass tubes; liquid flows through the inner tube and gas through the outer. Figure 3 shows a typical nozzle consisting of a fifty micron inner diameter tube inside a 1mm OD by 0.75mm ID tube. The exit of the outer tube has been reduced in size by heating with a propane flame. The jet leaving the nozzle is much smaller than the smallest constriction in the nozzle, here fifty micron, and is therefore much less likely to clog than if produced by a traditional rayleigh nozzle. Jet diameters as small as 300nm have been produced by this method.[8]

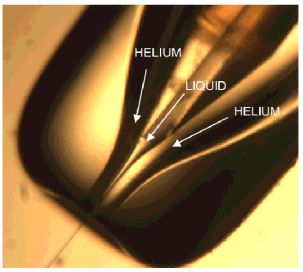


Figure 3. Water flowing through a 50micron inner diameter tube can be seen to form a tapering jet as the liquid is accelerated by a coaxially flowing helium gas sheath.

Liquid low rates for a 10 micron diameter jet would typical be around 20µl/minute for the style of nozzle shown in figure 3. The flow rate can be further reduced by grinding off the end of the outer nozzle as shown in figure 4. Reducing the length of tube through which the jet is accelerated decreases the ultimate speed of the jet. The nozzle on the left hand side of figure 4 produced jets moving at 17m/s but after the end was ground off the jet speed was reduced to 4m/s.

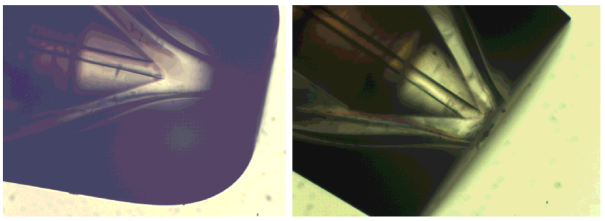


Figure 4. Nozzle with outer capillary ground back nearly flush with the inner capillary. The nozzle requires a minimum jet speed of 17m/s while the nozzle on the right requires only 4m/s.

It may be helpful when trying to produce very small jets to use an asymmetric orifice on the inner (liquid) tube of the nozzle. In this case shear forces can reduce the diameter of the liquid cone before the liquid completely exits the inner tube. Three examples are shown in figure 5. The first from the left shows a jet forming from the tip of a PEEK fiber attached to the inner nozzle tube. The next shows a jet forming from a rough spot on the surface of the inner tube. The last shows an inner tube that was ground off-axis to form a needle shape. In all three cases the jet diameter is already much smaller than the tube orifice before completely exiting the tube.

Flow alignment of rigid gold rods, needle shaped protein microcrystals, and Tobacco Mosaic Virus (TMV) has been observed in liquid jets with gas sheaths. This is to be expected in any extensional flow where the convergence of the streamlines happens sufficiently fast to overcome the thermal rotational energy of the particles. It is not yet know whether the effects of shear or the transition to plug flow will disturb this alignment for a traditional Rayleigh nozzle.

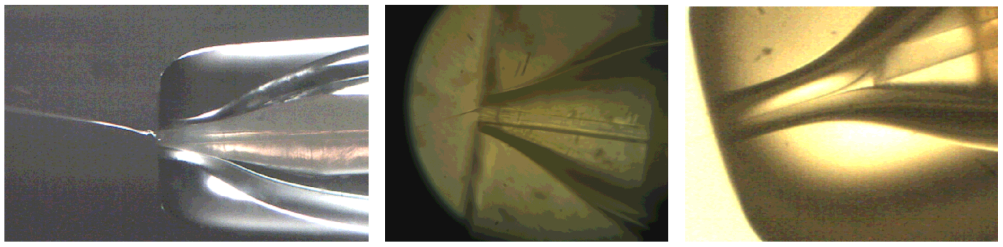


Figure 5. From left to right – A jet forming from a protruding fiber, a jet forming from a rough spot on the end of the inner nozzle tube, a jet forming from an asymmetric, needle-shaped tube.

2.3 Electrospray Ionization

Electrospray Ionization (ESI) is an aerosol generation technique which is used widely as a mass spectrometry analytical technique as well as for sample delivery and deposition. The ESI highly charges an ionic liquid which is then forced through a capillary with a cone shaped tip and placed opposite a counter electrode. The emitted charged liquid forms a Taylor cone at sufficiently high voltages and the charged liquid breaks up into droplets due to Coulombic forces. The aerosol is neutralized with an alpha source, halting breakup, before any applications. The advantages of ESI, such as the wide range of flow rates (10^{-8} liters/min to 10^{-5} liters/min) and droplet sizes (10^{-8} m to 10^{-5} m), make it ideal for low volume samples. Droplet size is determined by the distance between the neutralizer and the Taylor cone because neutralization stops the breakup process. Disadvantages are the lack of droplet size uniformity and the low aerosol concentration compared with the liquid jet or nebulizer/aerodynamic lens stack systems which can result in low hit rates. ESI is currently being investigated as an aerosol source for aerodynamic lens stack focusing and as a sample delivery method. ESI in vacuum has been previously shown for specific geometries and samples[9] but must be further investigated for a wide variety of experimental conditions and samples.

2.4 Nebulizers

While nebulizers are commercially available and easy to use they commonly have high gas flow rates, broad drop size distribution, and a particle density that falls off with distance squared. This last point especially makes them unsuitable as sample injectors to pulsed X-ray sources except for when used with a "focusing" mechanism such as gas dynamic force [10-13] or use of an aerodynamic lens.

A jet operating at high gas pressure can also be used to produce an aerosol. The gas flow rate for a jet-as-nebulizer is typically about 1 liter/hour much less than the 1 liter/minute that could be expected from a medical nebulizer. Jet breakup for spontaneous Plateau-Rayleigh breakup is not mono disperse but still has a fairly narrow size distribution[14] compared to commercial nebulizers. This is especially important when working with high concentration samples where micron or submicron drops are desired. Drops that are larger than the interparticle spacing in the liquid sample may contain multiple particles which will not contribute diffraction patterns useful for single particle 3D reconstruction. The drawback is sample flow rate which is only about 1μl/min for a 1 μm jet.

2.5 Drop on demand

Drop on demand (DOD) is yet another form of convective instability which can produce sample filled liquid drops. Drops are produced from a piezo driven converging nozzle, one drop for each time the piezo is triggered. By synchronizing droplet production to the FEL pulses, sample consumption would be greatly reduced. As with Rayleigh nozzles, clogging becomes problematic for smaller nozzles; 20micron diameter drop production is presently the state of the art, however, DOD tends to form liquid jets behind the primary drop which then break up into much smaller drops.. The small size of the drops will reduce the solution scatter and so provide an attractive target for sample delivery.

While it is possible that fragile crystals may be damaged by DOD we have not observed damage at low resolution. Solutions containing protein crystals too small for traditional crystallography at a synchrotron have been studied at LCLS using a liquid jet[2] at fairly high sample rates of sample consumption, 10μl/min at 10^9 crystal/ml concentration. DOD has the possibility of reducing sample consumption drastically if it can be shown to not damage crystals. Photosystem I

microcrystals were shown to diffract to 3nm at the Swiss Light Source (SLS) after running through a 50µm DOD nozzle at 60Hz. Further studies at high resolution are planned.

2.6 Aerodynamic lens

Aerodynamic lenses can produce particle streams of low divergence[15], with low gas load and variable degree of hydration. Aerodynamic lenses work by flowing a particle laden stream of gas through a series of circular apertures. As the gas moves through each aperture the inertia of the sample particles can cause them to move across the curved gas streamlines either towards or away from the centerline depending on their size and density. This is effectively a 2 dimensional compression of an aerosol. Differential pumping along the lens can reduce the gas load to make the lens stack compatible with high vacuum. Figure 6 shows an aerodynamic lens belonging to the Uppsala group of Janos Hajdu which currently in use at both the LCLS and FLASH.

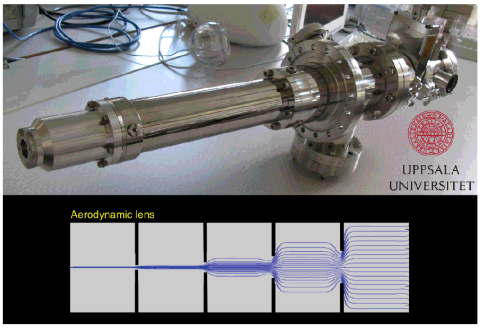


Figure 6. An aerodynamic lens currently in use at LCLS and FLASH. Figure is courtesy of Janos Hajdu.

3. HIT RATE

3.1 Hit probability

A useful number to estimate for any injector is the "hit rate", the rate at which useful diffraction patterns can be collected. To calculate the hit rate, two timescales must be considered: 1 the FEL pulse duration and 2 the transit time of the sample through the interaction volume, where the interaction volume is given as the intersection of the FEL and sample beam. If the transit time is long compared to the pulse duration then the sample can be considered motionless; no sample enters or leaves the interaction volume on the timescale of a single pulse. For example a sample flowing at 100m/s takes 100ns to cross a 10µm X-ray beam. The probability that any given X-ray pulse will hit a particle is therefore the time average probability that there is a particle in the interaction volume. For an X-ray focus smaller than the sample beam the interaction volume is

$$V_{\text{int}} = d_{fel}^2 d_s \tag{1}$$

where d_{fel} and d_s are the effective diameters of the X-ray focus and sample beam respectively. The probability of hitting a particle with any pulse is then

$$P = \rho V_{\text{int}} \tag{2}$$

for particle density ρ in the interaction region. The particle density can be estimated from measurable quantities, namely the rate of sample consumption and the speed and diameter of the sample beam.

$$\rho = \frac{\dot{N}}{v d_s^2} \tag{3}$$

Combining equations 1-3, for each pulse, the probability of hitting a particle is given by

$$P = \frac{\dot{N} d_{fel}^2}{v d_s}. \tag{4}$$

Equation 4 is a first order approximation of hit probability for small ρV_{int}. For larger ρV_{int} the right hand side of equation 2 could be made to take into account Poisson statistics. The probability P_n of an FEL pulse intercepting n particles is then given by

$$P(n) = \frac{1}{n!}\left(\frac{\dot{N} d_{fel}^2}{v d_s}\right)^n e^{\frac{-\dot{N} d_{fel}^2}{v d_s}}. \tag{5}$$

It is also a straightforward matter to modify equation 1 to include more realistic sample and beam profiles however equation 1 seems a reasonable approximation in comparison to other errors introduced for example by estimating sample number density or aligning the sample beam to the X-ray focus.

3.2 Hit rate

The hit rate measured is dependent on the repetition rate of the FEL and the camera readout rate. For the case of one pulse per camera readout, the hit rate which is the product of the FEL repetition rate and the single particle hit probability given in equation 5, can help estimate the time required to collect the desired data set. In this case ideally $\rho V_{\text{int}} = 1$; less means pulses are wasted and more means too many coherently summed multiple particle diffraction patterns. For the case of m pulses per camera readout ideally $\rho V_{\text{int}} = 1/m$ to avoid incoherently summed multiple particle diffraction patterns.

The number density in the interaction region, as in equation 2, is dependent on injector type. For a liquid jet or DOD synchronized with the laser this is just the sample concentration. When working downstream from the jet in the droplet region, the hit rate is reduced due to the drop diameter being larger than the jet diameter. For an aerodynamic lens the number density in the interaction region is much smaller than for a liquid jet although this is partially offset by the greater size of the interaction volume due to the greater width of the particle beam.

4. SAMPLE CONSUMPTION

Sample consumption for a liquid injector is the product of liquid flow rate Q, sample concentration and time needed to collect a complete data set. For similar hit rates, DOD will consume far les sample than jets or aerosol/lens combinations. A liquid jet moves at roughly 10m/s for a jet driven by gas dynamic forces or 100m/s for a rayleigh jet. This implies that for FEL repetition rate and spot size of 100Hz and 10μm only a very small fraction, 1 part in 10^4 to 10^5, of the liquid-containing sample is hit by the FEL. For a DOD stream synchronized to the FEL, there is very little sample flowing through the interaction volume between pulses implying a factor 10^3 to 10^4 lower sample consumption. An aerodynamic lens will consume the most sample - the sample speed is similar to that for a jet but the sample beam diameter is much greater than that of a jet which then requires either a higher sample consumption $\dot{N}$ to get similar hit rate (equation 3) or a longer data collection time.

5. CONCLUSIONS

Existing sample injection methods can deliver samples in various states of hydration and at high repetition rate. Liquid jets are well suited for protein nanocrystals in that they are able to deliver the sample to the interaction region in the same solution in which the crystals are prepared. The high speed of liquid jets make them inefficient in sample consumption for FEL repetition rates of 100 Hz as currently is the case the LCLS and FLASH and so alternate methods such as DOD should be investigated further. The use of an aerodynamic lens to focus an aerosols is better suited for single particle imaging where water removal is desired and higher number density of samples such as, virus, cells, single molecules, are easier to produce. A high number density for protein microcrystals for example is 10^{10} per ml while a typical virus might be produced at concentration up to 10^{14} per ml. The advantage to high speed injection is that it can keep up with high FEL repetition rates. Liquid jets and gas focused aerosols are compatible with FEL repetition rates up to 10Mhz for a micron size X-ray focus.

[1] H. N. Chapman, P. Fromme, A. Barty *et al.*, "Femtosecond X-ray protein nanocrystallography," Nature, 470(7332), 73-U81 (2011).

[2] H. Chapman, "X-ray imaging beyond the limits," Nature Materials, 8(4), 299-301 (2009).

[3] M. M. Seibert, T. Ekeberg, F. Maia *et al.*, "Single mimivirus particles intercepted and imaged with an X-ray laser," Nature, 470(7332), 78-U86 (2011).

[4] L. Rayleigh, "On the Instability of Jets," Proceedings of the London Mathematical Society, s1-10(1), 4-13 (1878).

[5] L. Rayleigh, "On the Capillary Phenomena of Jets," Proceedings of the Royal Society of London, 29, 71-97 (1879).

[6] U. Weierstall, B. Doak, J. Spence *et al.*, "Droplet streams for serial crystallography of proteins," Exp. Fluids, 44(5), 675-689 (2008).

[7] D. P. DePonte, U. Weierstall, K. Schmidt *et al.*, "Gas dynamic virtual nozzle for generation of microscopic droplet streams," Journal of Physics D: Applied Physics(19), 195505 (2008).

[8] D. P. DePonte, M. Hunter, R. B. Doak *et al.*, "Electron Diffraction from Liquid Jets in TEM," Proceedings Microscopy & Microanalysis(2009), (2009).

[9] J. C. Swarbrick, J. B. Taylor, and J. N. O'Shea, "Electrospray deposition in vacuum," Applied Surface Science, 252(15), 5622-5626 (2006).

[10] F. T. Gucker, C. T. O'Konski, H. B. Pickard *et al.*, "A Photoelectronic Counter for Colloidal Particles," J. Am. Chem. Soc., 69(10), 2422-2431 (1947).

[11] R. A. Keller, and N. S. Nogar, "Gasdynamic focusing for sample concentration in ultrasensitive analysis," Appl. Opt., 23(13), 2146 (1984).

[12] B. Y. H. Liu, R. N. Berglund, and J. K. Agarwal, "Experimental studies of optical particle counters," Atmospheric Environment (1967), 8(7), 717-732 (1974).

[13] S. W. Stiller, and M. V. Johnston, "Supersonic jet spectroscopy with a capillary gas chromatographic inlet," Anal. Chem., 59(4), 567-572 (1987).

[14] A. M. Ganan-Calvo, "Generation of steady liquid microthreads and micron-sized monodisperse sprays in gas streams," Physical Review Letters, 80(2), 285-288 (1998).

[15] P. Liu, P. J. Ziemann, D. B. Kittelson *et al.*, "Generating Particle Beams of Controlled Dimensions and Divergence .1. Theory of Particle Motion in Aerodynamic Lenses and Nozzle Expansions," Aerosol Science and Technology, 22(3), 293-313 (1995).

10. T. Voigt, *J. Comp. Neurol.* **289**, 74 (1989).
11. D. H. Rowitch, *Nat. Rev. Neurosci.* **5**, 409 (2004).
12. Y. Muroyama, Y. Fujiwara, S. H. Orkin, D. H. Rowitch, *Nature* **438**, 360 (2005).
13. C. Hochstim, B. Deneen, A. Lukaszewicz, Q. Zhou, D. J. Anderson, *Cell* **133**, 510 (2008).
14. S. W. Levison, J. E. Goldman, *Neuron* **10**, 201 (1993).
15. M. S. Windrem *et al.*, *Cell Stem Cell* **2**, 553 (2008).
16. X. Zhu, R. A. Hill, A. Nishiyama, *Neuron Glia Biol.* **4**, 19 (2008).
17. S. Okada *et al.*, *Nat. Med.* **12**, 829 (2006).
18. S. Robel, S. Bardehle, A. Lepier, C. Brakebusch, M. Götz, *J. Neurosci.* **31**, 12471 (2011).
19. Materials and methods are available as supplementary materials on *Science* Online.
20. N. Masahira *et al.*, *Dev. Biol.* **293**, 358 (2006).
21. C. Shannon, M. Salter, R. Fern, *J. Anat.* **210**, 684 (2007).
22. N. A. Oberheim *et al.*, *J. Neurosci.* **29**, 3276 (2009).
23. J. Cai *et al.*, *Development* **134**, 1887 (2007).
24. M. S. Rao, M. Noble, M. Mayer-Pröschel, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 3996 (1998).
25. H. H. Tsai, W. B. Macklin, R. H. Miller, *J. Neurosci.* **26**, 1913 (2006).
26. Q. Zhou, D. J. Anderson, *Cell* **109**, 61 (2002).
27. N. Kessaris *et al.*, *Nat. Neurosci.* **9**, 173 (2006).
28. F. T. Merkle, Z. Mirzadeh, A. Alvarez-Buylla, *Science* **317**, 381 (2007).
29. S. Magavi, D. Friedmann, G. Banks, A. Stolfi, C. Lois, *J. Neurosci.* **32**, 4762 (2012).
30. P. Rakic, *Science* **241**, 170 (1988).
31. W. P. Ge, A. Miyawaki, F. H. Gage, Y. N. Jan, L. Y. Jan, *Nature* **484**, 376 (2012).

Acknowledgments: We thank M. Wong, S. Kaing, U. Dennehy, M. Grist, and S. Chang for technical help and E. Huillard, V. Heine, and C. Stiles for helpful comments. We thank A. Leiter (University of Massachusetts, Worcester) for *Ngn3-cre* mice. L.C.F. is a Howard Hughes Medical Institute (HHMI) Fellow of the Helen Hay Whitney Foundation. R.T.-M. was funded by the Portuguese Fundação para a Ciência e a Tecnologia. This work was supported by grants from the NIH, UK Medical Research Council, Wellcome Trust, and European Research Council. A.A.-B. holds the Heather and Melanie Muss Chair of Neurological Surgery. D.H.R. is a HHMI Investigator.

Supplementary Materials
www.sciencemag.org/cgi/content/full/science.1222381/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 and S2
References (*32*–*43*)
Movies S1 to S5

26 March 2012; accepted 6 June 2012
Published online 28 June 2012;
10.1126/science.1222381

High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography

Sébastien Boutet,[1]* Lukas Lomb,[2,3] Garth J. Williams,[1] Thomas R. M. Barends,[2,3] Andrew Aquila,[4] R. Bruce Doak,[5] Uwe Weierstall,[5] Daniel P. DePonte,[4] Jan Steinbrener,[2,3] Robert L. Shoeman,[2,3] Marc Messerschmidt,[1] Anton Barty,[4] Thomas A. White,[4] Stephan Kassemeyer,[2,3] Richard A. Kirian,[5] M. Marvin Seibert,[1] Paul A. Montanez,[1] Chris Kenney,[6] Ryan Herbst,[6] Philip Hart,[6] Jack Pines,[6] Gunther Haller,[6] Sol M. Gruner,[7,8] Hugh T. Philipp,[7] Mark W. Tate,[7] Marianne Hromalik,[9] Lucas J. Koerner,[10] Niels van Bakel,[11] John Morse,[12] Wilfred Ghonsalves,[1] David Arnlund,[13] Michael J. Bogan,[14] Carl Caleman,[4] Raimund Fromme,[15] Christina Y. Hampton,[14] Mark S. Hunter,[15] Linda C. Johansson,[13] Gergely Katona,[13] Christopher Kupitz,[15] Mengning Liang,[4] Andrew V. Martin,[4] Karol Nass,[16] Lars Redecke,[17,18] Francesco Stellato,[4] Nicusor Timneanu,[19] Dingjie Wang,[5] Nadia A. Zatsepin,[5] Donald Schafer,[1] James Defever,[1] Richard Neutze,[13] Petra Fromme,[15] John C. H. Spence,[5] Henry N. Chapman,[4,16] Ilme Schlichting[2,3]

Structure determination of proteins and other macromolecules has historically required the growth of high-quality crystals sufficiently large to diffract x-rays efficiently while withstanding radiation damage. We applied serial femtosecond crystallography (SFX) using an x-ray free-electron laser (XFEL) to obtain high-resolution structural information from microcrystals (less than 1 micrometer by 1 micrometer by 3 micrometers) of the well-characterized model protein lysozyme. The agreement with synchrotron data demonstrates the immediate relevance of SFX for analyzing the structure of the large group of difficult-to-crystallize molecules.

Elucidating macromolecular structures by x-ray crystallography is an important step in the quest to understand the chemical mechanisms underlying biological function. Although facilitated greatly by synchrotron x-ray sources, the method is limited by crystal quality and radiation damage (*1*). Crystal size and radiation damage are inherently linked, because reducing radiation damage requires lowering the incident fluence. This in turn calls for large crystals that yield sufficient diffraction intensities while reducing the dose to individual molecules in the crystal. Unfortunately, growing well-ordered large crystals can be difficult in many cases, particularly for large macromolecular assemblies and membrane proteins. In contrast, micrometer-sized crystals are frequently observed. Although diffraction data of small crystals can be collected by using microfocus synchrotron beamlines, this remains a challenging approach because of the rapid damage suffered by these small crystals (*1*).

Serial femtosecond crystallography (SFX) using x-ray free-electron laser (XFEL) radiation is an emerging method for three-dimensional (3D) structure determination using crystals ranging from a few micrometers to a few hundred nanometers in size and potentially even smaller. This method relies on x-ray pulses that are sufficiently intense to produce high-quality diffraction while of short enough duration to terminate before the onset of substantial radiation damage (*2*–*4*). X-ray pulses of only 70-fs duration terminate before any chemical damage processes have time to occur, leaving primarily ionization and x-ray–induced thermal motion as the main sources of radiation damage (*2*–*4*). SFX therefore promises to break the correlation between sample size, damage, and resolution in structural biology. In SFX, a liquid microjet is used to introduce fully hydrated, randomly oriented crystals into the single-pulse XFEL beam (*5*–*8*), as illustrated in Fig.1. A recent low-resolution proof-of-principle demonstration of SFX performed at the Linac Coherent Light Source (LCLS) (*9*) using crystals of photosystem I ranging in size from 200 nm to 2 μm produced interpretable electron density maps (*6*). Other demonstration experiments using crystals grown in vivo (*7*), as well as in the lipidic sponge phase for membrane proteins (*8*), were recently published. However, in all these cases, the x-ray energy of 1.8 keV (6.9 Å) limited the resolution of the collected data to about 8 Å. Data collection to a resolution better than 2 Å became possible with the recent commis-

[1]Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA. [2]Max-Planck-Institut für Medizinische Forschung, Jahnstrasse 29, 69120 Heidelberg, Germany. [3]Max Planck Advanced Study Group, Center for Free-Electron Laser Science, Notkestrasse 85, 22607 Hamburg, Germany. [4]Center for Free-Electron Laser Science, Deutsches Elektronen-Synchrotron (DESY), Notkestrasse 85, 22607 Hamburg, Germany. [5]Department of Physics, Arizona State University, Tempe, AZ 85287, USA. [6]Particle Physics and Astrophysics, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA. [7]Department of Physics, Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853, USA. [8]Wilson Laboratory, Cornell High Energy Synchrotron Source (CHESS), Cornell University, Ithaca, NY 14853, USA. [9]Electrical and Computer Engineering, State University of New York (SUNY) Oswego, Oswego, NY 13126, USA. [10]The Johns Hopkins University Applied Physics Laboratory, 11100 Johns Hopkins Road, Laurel, MD 20723, USA. [11]Nikhef, National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, Netherlands. [12]European Synchrotron Radiation Facility, 38043 Grenoble Cedex, France. [13]Department of Chemistry and Molecular Biology, University of Gothenburg, SE-405 30 Gothenburg, Sweden. [14]PULSE Institute, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA. [15]Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287–1604, USA. [16]University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany. [17]Joint Laboratory for Structural Biology of Infection and Inflammation, Institute of Biochemistry and Molecular Biology, University of Hamburg, and Institute of Biochemistry, University of Lübeck, at DESY, Hamburg, Germany. [18]German Centre for Infection Research, University of Lübeck, 23538 Lübeck, Germany. [19]Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden.

*To whom correspondence should be addressed. E-mail: sboutet@slac.stanford.edu

sioning of the LCLS Coherent X-ray Imaging (CXI) instrument (*10*). The CXI instrument provides hard x-ray pulses suitable for high-resolution crystallography and is equipped with Cornell-SLAC Pixel Array Detectors (CSPADs), consisting of 64 tiles of 192 pixels by 185 pixels each, arranged as shown in Fig. 1 and figs. S1 and S2. The CSPAD supports the 120-Hz readout rate required to measure each x-ray pulse from LCLS (*11*, *12*).

Here, we describe SFX experiments performed at CXI analyzing the structure of hen egg-white lysozyme (HEWL) as a model system by using microcrystals of about 1 μm by 1 μm by 3 μm (*4*, *11*). HEWL is an extremely well-characterized protein that crystallizes easily. It was the first enzyme to have its structure determined by x-ray diffraction (*13*) and has since been thoroughly characterized to very high resolution (*14*). Lysozyme has served as a model system for many investigations, including radiation damage studies. This makes it an ideal system for the development of the SFX technique. Microcrystals of HEWL in random orientation were exposed to single 9.4-keV (1.32 Å) x-ray pulses of 5- or 40-fs duration focused to 10 μm^2 at the interaction point (Fig. 1). The average 40-fs pulse energy at the sample was 600 μJ per pulse, corresponding to an average dose of 33 MGy deposited in each crystal. This dose level represents the classical limit for damage using cryogenically cooled crystals (*15*). The average 5-fs pulse energy was 53 μJ. The SFX-derived data were compared to low-dose data sets collected at room temperature by using similarly prepared larger crystals (*11*). This benchmarks the technique with a well-characterized model system.

We collected about 1.5 million individual "snapshot" diffraction patterns for 40-fs duration pulses at the LCLS repetition rate of 120 Hz using the CSPAD. About 4.5% of the patterns were classified as crystal hits, 18.4% of which were indexed and integrated with the CrystFEL software (*14*) showing excellent statistics to 1.9 Å resolution (Table 1 and table S1). In addition, 2 million diffraction patterns were collected by using x-ray pulses of 5-fs duration, with a 2.0% hit rate and a 26.3% indexing rate, yielding 10,575 indexed patterns. The structure, partially shown in Fig. 2A, was determined by molecular replacement [using Protein Data Bank (PDB) entry 1VDS] and using the 40-fs SFX data. No significant differences were observed in an F_{obs}(40 fs) – F_{obs} (synchrotron) difference electron density map (Fig. 2B). The electron density map shows features that were not part of the model (different conformations of amino acids and water molecules) and shows no

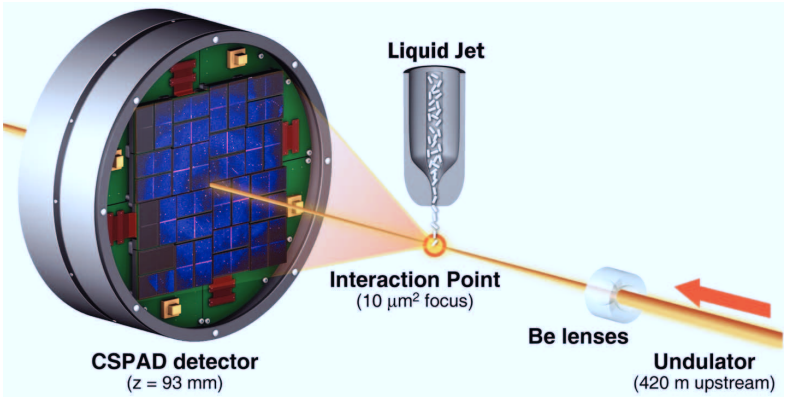


Fig. 1. Experimental geometry for SFX at the CXI instrument. Single-pulse diffraction patterns from single crystals flowing in a liquid jet are recorded on a CSPAD at the 120-Hz repetition rate of LCLS. Each pulse was focused at the interaction point by using 9.4-keV x-rays. The sample-to-detector distance (*z*) was 93 mm.

Table 1. SFX and synchrotron data and refinement statistics. Highest resolution shells are 2.0 to 1.9 Å. R_{split} is as defined in (*16*): $R_{split} = \left(\frac{1}{\sqrt{2}}\right) \cdot \frac{\sum_{hkl} |I_{hkl}^{even} - I_{hkl}^{odd}|}{\frac{1}{2}\sum_{hkl} |I_{hkl}^{even} + I_{hkl}^{odd}|}$. SLS room temperature (RT) data 3 statistics are from XDS (*20*). B factors were calculated with TRUNCATE (*21*). *R* and rmsd values were calculated with PHENIX (*22*). n.a., not applicable. The diffraction patterns have been deposited with the Coherent X-ray Imaging Data Bank, cxidb.org (accession code ID-17).

Parameter	40-fs pulses	5-fs pulses	SLS RT data 3
Wavelength	1.32 Å	1.32 Å	0.9997 Å
X-ray focus (μm^2)	~10	~10	~100 × 100
Pulse energy/fluence at sample	600 μJ/4 × 10^{11} photons per pulse	53 μJ/3.5 ×10^{10} photons per pulse	n.a./2.5 × 10^{10} photons/s
Dose (MGy)	33.0 per crystal	2.9 per crystal	0.024 total
Dose rate (Gy/s)	8.3 × 10^{20}	5.8 × 10^{20}	9.6 × 10^{2}
Space group	$P4_32_12$	$P4_32_12$	$P4_32_12$
Unit cell length (Å), α = β = γ = 90°	*a* = *b* =79, *c* = 38	*a* = *b* = 79, *c* = 38	*a* = *b* = 79.2, *c* = 38.1
Oscillation range/exposure time	Still exp./40 fs*	Still exp./5 fs*	1.0°/0.25 s
No. collected diffraction images	1,471,615	1,997,712	100
No. of hits/indexed images	66,442/12,247	40,115/10,575	n.a./100
Number of reflections	n.a.	n.a.	70,960
Number of unique reflections	9921	9743	9297
Resolution limits (Å)	35.3–1.9	35.3–1.9	35.4–1.9
Completeness	98.3% (96.6%)	98.2% (91.2%)	92.6% (95.1%)
$I/\sigma(I)$	7.4 (2.8)	7.3 (3.1)	18.24 (5.3)
R_{split}	0.158	0.159	n.a.
R_{merge}	n.a.	n.a.	0.075 (0.332)
Wilson B factor	28.3 $Å^2$	28.5 $Å^2$	19.4 $Å^2$
R-factor/*R*-free	0.196/0.229	0.189/0.227	0.166/0.200
Rmsd bonds, Rmsd angles	0.006 Å, 1.00°	0.006 Å, 1.03°	0.007 Å, 1.05°
PDB code	4ET8	4ET9	4ETC

*Electron bunch length

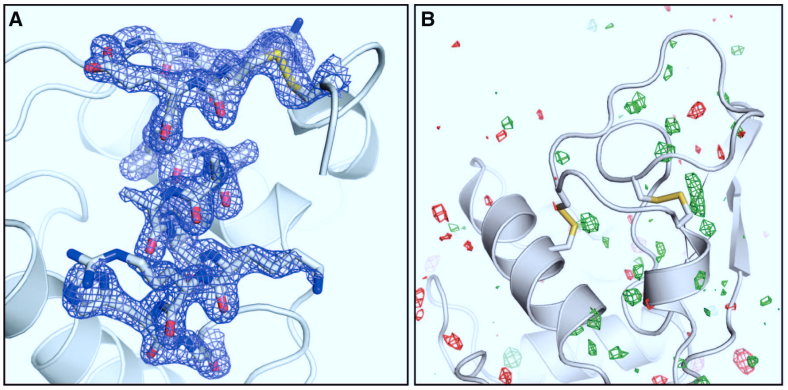


Fig. 2. (**A**) Final, refined $2mF_{obs} - DF_{calc}$ (1.5σ) electron density map (*17*) of lysozyme at 1.9 Å resolution calculated from 40-fs pulse data. (**B**) F_{obs}(40 fs) − F_{obs} (synchrotron) difference Fourier map, contoured at +3 σ (green) and −3 σ (red). No interpretable features are apparent. The synchrotron data set was collected with a radiation dose of 24 kGy.

discernible signs of radiation damage. Also, when the data were phased with molecular replacement by using the turkey lysozyme structure as a search model (PDB code 1LJN), the differences between the two proteins were immediately obvious from the maps (fig. S3).

Even though the underlying radiation damage processes differ because of the different time scales of the experiments using an XFEL and a synchrotron or rotating anode (femtoseconds versus seconds or hours), no features related to radiation damage are observed in difference maps calculated between the SFX and the low-dose synchrotron data (Fig. 2B). In addition to local structural changes, metrics like I/I_0 [the ratio of measured intensities (I) to the ideal calculated intensities (I_0)] and the Wilson B factor are most often used to characterize global radiation damage in protein crystallography (*17*). I/I_0 is not applicable to the SFX data. However, the Wilson B factors of both SFX data sets show values typical for room-temperature data sets and do not differ significantly from those obtained from synchrotron and rotating anode data sets collected with different doses, using similarly grown larger crystals kept at room temperature and fully immersed in solution (*11*) (Table 1 and table S1). The R factors calculated between all collected data sets do not show a dose-dependent increase (fig. S4). However, higher R factors are observed for the SFX data, indicating a systematic difference. This is not caused by nonconvergence of the Monte Carlo integration, because scaling the 40- and 5-fs data together does not affect the scaling behavior. Besides non-isomorphism or radiation damage, possible explanations for this difference could include suboptimal treatment of weak reflections, the difficulties associated with processing still diffraction images, and other SFX-specific steps in the method. SFX is an emerging technique, and data processing algorithms, detectors, and data collection methods are under continuous development.

A simple consideration shows the attainable velocities of atoms in the sample depend on the deposited x-ray energy versus the inertia of those atoms: $\langle v \rangle = \sqrt{3k_B T/m}$, where m is the mass of a carbon atom, for example, T is temperature, and k_B is Boltzmann's constant. For an impulse absorption of energy at the doses of our LCLS measurements, we predict average velocities less than 10 Å/ps, which gives negligible displacement during the FEL pulses. On the time scale of femtoseconds, radiation damage is primarily caused by impulsive rearrangement of atoms and electron density rather than the relatively slow-processes of chemical bond breaking typical in conventional crystallography using much longer exposures at much lower dose rates (the dose rate in this experiment was about 0.75 MGy per femtosecond).

Neither the SFX electron density maps nor the Wilson B factors suggest obvious signs of significant radiation damage. Very short pulses (5-fs electron bunch) are not expected to produce observable damage, according to simulations (*3*). Furthermore, it has been reported that the actual x-ray pulses are shorter than the electron bunches for XFELs, making the pulse duration possibly shorter than the relevant Auger decays (*18*). The agreement between the SFX results using 40-fs pulses and 5-fs pulses suggests similar damage characteristics for the two pulse durations on the basis of the available data. Our results demonstrate that under the exposure conditions used, SFX yields high-quality data suitable for structural determination. SFX reduces the requirements on crystal size and therefore the method is of immediate relevance for the large group of difficult-to-crystallize molecules, establishing SFX as a very valuable high-resolution complement to existing macromolecular crystallography techniques.

References and Notes

1. J. M. Holton, K. A. Frankel, *Acta Crystallogr.* **D66**, 393 (2010).
2. R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, J. Hajdu, *Nature* **406**, 752 (2000).
3. A. Barty *et al.*, *Nat. Photonics* **6**, 35 (2011).
4. L. Lomb *et al.*, *Phys. Rev. B* **84**, 214111 (2011).
5. D. P. DePonte *et al.*, *J. Phys. D* **41**, 195505 (2008).
6. H. N. Chapman *et al.*, *Nature* **470**, 73 (2011).
7. R. Koopmann *et al.*, *Nat. Methods* **9**, 259 (2012).
8. L. C. Johansson *et al.*, *Nat. Methods* **9**, 263 (2012).
9. P. Emma *et al.*, *Nat. Photonics* **4**, 641 (2010).
10. S. Boutet, G. J. Williams, *New J. Phys.* **12**, 035024 (2010).
11. Materials and methods are available as supplementary materials on *Science* Online.
12. H. T. Philipp, M. Hromalik, M. Tate, L. Koerner, S. M. Gruner, *Nucl Instrum. Methods A* **649**, 67 (2011).
13. C. C. F. Blake *et al.*, *Nature* **206**, 757 (1965).
14. J. Wang, M. Dauter, R. Alkire, A. Joachimiak, Z. Dauter, *Acta Crystallogr.* **D63**, 1254 (2007).
15. R. L. Owen, E. Rudiño-Piñera, E. F. Garman, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 4912 (2006).
16. T. A. White *et al.*, *J. Appl. Cryst.* **45**, 335 (2012).
17. R. J. Southworth-Davies, M. A. Medina, I. Carmichael, E. F. Garman, *Structure* **15**, 1531 (2007).
18. L. Young *et al.*, *Nature* **466**, 56 (2010).
19. R. J. Read, *Acta Crystallogr.* **A42**, 140 (1986).
20. W. Kabsch, *J. Appl. Cryst.* **26**, 795 (1993).
21. A. J. McCoy, R. W. Grosse-Kunstleve, L. C. Storoni, R. J. Read, *Acta Crystallogr.* **D61**, 458 (2005).
22. P. D. Adams *et al.*, *Acta Crystallogr.* **D66**, 213 (2010).

Acknowledgments: Portions of this research were carried out at the LCLS, a National User Facility operated by Stanford University on behalf of the U.S. Department of Energy (DOE), Office of Basic Energy Sciences (OBES) and at the Swiss Light Source, beamline X10SA, Paul Scherrer Institute, Villigen, Switzerland. The CXI instrument was funded by the LCLS Ultrafast Science Instruments (LUSI) project funded by DOE, OBES. We acknowledge support from the Max Planck Society, the Hamburg Ministry of Science and Research, and the Joachim Herz Stiftung, as part of the Hamburg Initiative for Excellence in Research (LEXI); the Hamburg School for Structure and Dynamics in Infection; the U.S. NSF (awards 0417142 and MCB-1021557); the NIH (award 1R01GM095583); the German Federal Ministry for Education and Research (grants 01KX0806 and 01KX0807); the Deutsche Forschungsgemeinschaft Cluster of Excellence EXC 306; AMOS program within the Chemical Sciences, Geosciences, and Biosciences Division of the OBES, Office of Science, U.S. DOE; the Swedish Research Council; the Swedish Foundation for International Cooperation in Research and Higher Education. We thank A. Meinhart and E. Hofmann for collecting the synchrotron data set, M. Gebhart for help preparing the crystals, and M. Hayes and the technical staff of SLAC and the LCLS for their great support in carrying out these experiments. Special thanks to G. M. Stewart, T. Anderson, and SLAC Infomedia for generating Fig. 1. The structure factors and coordinates have been deposited with the Protein Data Bank (accession codes 4ET8, 4ET9, 4ETA, 4ETB, 4ETC, 4ETD, and 4ETE). The diffraction patterns have been deposited with the Coherent X-ray Imaging Data Bank cxidb.org (accession code ID-17). The Arizona Board of Regents, acting for and on behalf of Arizona State University and in conjunction with R.B.D., U.W., D.P.D., and J.C.H.S., has filed U.S. and international patent applications on the nozzle technology applied herein.

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1217737/DC1
Materials and Methods
Figs. S1 to S7
Table S1
References (*23*–*26*)

12 December 2011; accepted 21 May 2012
Published online 31 May 2012;
10.1126/science.1217737

LETTER

doi:10.1038/nature09750

Femtosecond X-ray protein nanocrystallography

Henry N. Chapman[1,2], Petra Fromme[3], Anton Barty[1], Thomas A. White[1], Richard A. Kirian[4], Andrew Aquila[1], Mark S. Hunter[3], Joachim Schulz[1], Daniel P. DePonte[1], Uwe Weierstall[4], R. Bruce Doak[4], Filipe R. N. C. Maia[5], Andrew V. Martin[1], Ilme Schlichting[6,7], Lukas Lomb[7], Nicola Coppola[1]†, Robert L. Shoeman[7], Sascha W. Epp[6,8], Robert Hartmann[9], Daniel Rolles[6,7], Artem Rudenko[6,8], Lutz Foucar[6,7], Nils Kimmel[10], Georg Weidenspointner[11,10], Peter Holl[9], Mengning Liang[1], Miriam Barthelmess[12], Carl Caleman[1], Sébastien Boutet[13], Michael J. Bogan[14], Jacek Krzywinski[13], Christoph Bostedt[13], Saša Bajt[12], Lars Gumprecht[1], Benedikt Rudek[6,8], Benjamin Erk[6,8], Carlo Schmidt[6,8], André Hömke[6,8], Christian Reich[9], Daniel Pietschner[10], Lothar Strüder[6,10], Günter Hauser[10], Hubert Gorke[15], Joachim Ullrich[6,8], Sven Herrmann[10], Gerhard Schaller[10], Florian Schopper[10], Heike Soltau[9], Kai-Uwe Kühnel[8], Marc Messerschmidt[13], John D. Bozek[13], Stefan P. Hau-Riege[16], Matthias Frank[16], Christina Y. Hampton[14], Raymond G. Sierra[14], Dmitri Starodub[14], Garth J. Williams[13], Janos Hajdu[5], Nicusor Timneanu[5], M. Marvin Seibert[5]†, Jakob Andreasson[5], Andrea Rocker[5], Olof Jönsson[5], Martin Svenda[5], Stephan Stern[1], Karol Nass[2], Robert Andritschke[10], Claus-Dieter Schröter[8], Faton Krasniqi[6,7], Mario Bott[7], Kevin E. Schmidt[4], Xiaoyu Wang[4], Ingo Grotjohann[3], James M. Holton[17], Thomas R. M. Barends[7], Richard Neutze[18], Stefano Marchesini[17], Raimund Fromme[3], Sebastian Schorb[19], Daniela Rupp[19], Marcus Adolph[19], Tais Gorkhover[19], Inger Andersson[20], Helmut Hirsemann[12], Guillaume Potdevin[12], Heinz Graafsma[12], Björn Nilsson[12] & John C. H. Spence[4]

X-ray crystallography provides the vast majority of macromolecular structures, but the success of the method relies on growing crystals of sufficient size. In conventional measurements, the necessary increase in X-ray dose to record data from crystals that are too small leads to extensive damage before a diffraction signal can be recorded[1–3]. It is particularly challenging to obtain large, well-diffracting crystals of membrane proteins, for which fewer than 300 unique structures have been determined despite their importance in all living cells. Here we present a method for structure determination where single-crystal X-ray diffraction 'snapshots' are collected from a fully hydrated stream of nanocrystals using femtosecond pulses from a hard-X-ray free-electron laser, the Linac Coherent Light Source[4]. We prove this concept with nanocrystals of photosystem I, one of the largest membrane protein complexes[5]. More than 3,000,000 diffraction patterns were collected in this study, and a three-dimensional data set was assembled from individual photosystem I nanocrystals (~200 nm to 2 μm in size). We mitigate the problem of radiation damage in crystallography by using pulses briefer than the timescale of most damage processes[6]. This offers a new approach to structure determination of macromolecules that do not yield crystals of sufficient size for studies using conventional radiation sources or are particularly sensitive to radiation damage.

Radiation damage has always limited resolution in biological imaging using electrons or X-rays[2]. With the recent invention of the femtosecond X-ray laser, an opportunity has arisen to break the nexus between radiation dose and spatial resolution. It has been proposed that femtosecond X-ray pulses can be used to outrun even the fastest damage processes by using single pulses so brief that they terminate before the manifestation of damage to the sample[6]. Experiments at the FLASH free-electron laser (FEL), Germany, confirmed the feasibility of 'diffraction before destruction' at resolution lengths down to 60 Å on test samples fixed on silicon nitride membranes[7]. It was predicted that the irradiance (or power density) of focused pulses from a hard-X-ray FEL such as the Linac Coherent Light Source (LCLS), USA, would be sufficient to produce diffraction patterns at near-atomic resolution[6].

We demonstrate here that this notion of diffraction before destruction operates at subnanometre resolution, using the membrane protein photosystem I as a model system, and establish an approach to structure determination based on X-ray diffraction data from a stream of nanocrystals[6,8]. Membrane proteins have a central role in the functioning of cells and viruses, yet our knowledge of the structure and dynamics responsible for their functioning remains limited. Photosystem I is a large membrane protein complex (1-MDa molecular mass, 36 proteins, 381 cofactors) that acts as a biosolar energy converter in the process of oxygenic photosynthesis. Its crystals display the symmetry of space group $P6_3$, with unit-cell parameters $a = b = 281$ Å and $c = 165$ Å, and consist of 78% solvent by volume. We show that diffraction data can be recorded from these fragile protein nanocrystals before destruction occurs. Furthermore, we demonstrate that structure factors can be extracted from the 'partial' reflections of tens of thousands of single-crystal diffraction snapshots, showing that interpretable high-quality, three-dimensional (3D) structure factor data can be obtained from a suspension of submicrometre crystals.

Our experimental set-up (Fig. 1 and Methods) records single-crystal diffraction data from a stream of crystals carried in a 4-μm-diameter, continuous liquid water jet[9] that flows across the focused LCLS X-ray beam in vacuum at 10 μl min^{-1}. In contrast to cryo-electron microscopy[10,11] or standard crystallography on microcrystals[3], which require cryogenic cooling, these data were collected on fully hydrated, 3D nanocrystals. The crystal located in the interaction region when an X-ray pulse arrives gives rise to a diffraction pattern that is detected on a set of two low-noise, X-ray p–n junction charge-coupled device (pnCCD) modules[12] and read out before the arrival of the next pulse at the FEL repetition rate of 30 Hz, or 1,800 patterns per minute. The

[1]Center for Free-Electron Laser Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany. [2]University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany. [3]Department of Chemistry and Biochemistry, Arizona State University, Tempe, Arizona 85287-1604, USA. [4]Department of Physics, Arizona State University, Tempe, Arizona 85287, USA. [5]Laboratory of Molecular Biophysics, Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 (Box 596), SE-751 24 Uppsala, Sweden. [6]Max Planck Advanced Study Group, Center for Free-Electron Laser Science, Notkestrasse 85, 22607 Hamburg, Germany. [7]Max-Planck-Institut für Medizinische Forschung, Jahnstrasse 29, 69120 Heidelberg, Germany. [8]Max-Planck-Institut für Kernphysik, Saupfercheckweg 1, 69117 Heidelberg, Germany. [9]PNSensor GmbH, Otto-Hahn-Ring 6, 81739 München, Germany. [10]Max-Planck-Institut Halbleiterlabor, Otto-Hahn-Ring 6, 81739 München, Germany. [11]Max-Planck-Institut für Extraterrestrische Physik, Giessenbachstrasse, 85741 Garching, Germany. [12]Photon Science, DESY, Notkestrasse 85, 22607 Hamburg, Germany. [13]LCLS, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA. [14]PULSE Institute, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA. [15]Forschungszentrum Jülich, Institut ZEL, 52425 Jülich, Germany. [16]Lawrence Livermore National Laboratory, 7000 East Avenue, Mail Stop L-211, Livermore, California 94551, USA. [17]Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [18]Department of Chemistry, Biochemistry and Biophysics, University of Gothenburg, SE-405 30 Gothenburg, Sweden. [19]Institut für Optik und Atomare Physik, Technische Universität Berlin, Hardenbergstrasse 36, 10623 Berlin, Germany. [20]Department of Molecular Biology, Swedish University of Agricultural Sciences, Uppsala Biomedical Centre, Box 590, S-751 24 Uppsala, Sweden. †Present addresses: European XFEL GmbH, Notkestrasse 85, 22607 Hamburg, Germany (N.C.); LCLS, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA (M.M.S.).

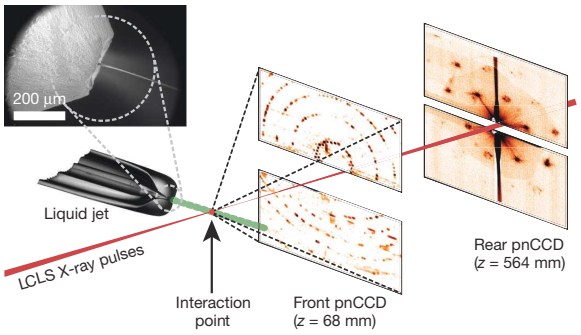


Figure 1 | Femtosecond nanocrystallography. Nanocrystals flow in their buffer solution in a gas-focused, 4-µm-diameter jet at a velocity of 10 m s^{-1} perpendicular to the pulsed X-ray FEL beam that is focused on the jet. Inset, environmental scanning electron micrograph of the nozzle, flowing jet and focusing gas[30]. Two pairs of high-frame-rate pnCCD detectors[12] record low- and high-angle diffraction from single X-ray FEL pulses, at the FEL repetition rate of 30 Hz. Crystals arrive at random times and orientations in the beam, and the probability of hitting one is proportional to the crystal concentration.

photon energy of the X-ray pulses was 1.8 keV (6.9-Å wavelength), with more than 10^{12} photons per pulse at the sample and pulse durations of 10, 70, and 200 fs (ref. 13). An X-ray fluence of 900 J cm^{-2} was achieved by focusing the FEL beam to a full-width at half-maximum of 7 µm, corresponding to a sample dose of up to 700 MGy per pulse (calculated using the program RADDOSE[14]) and a peak power density in excess of 10^{16} W cm^{-2} at 70-fs duration. In contrast, the typical tolerable dose in conventional X-ray experiments is only about 30 MGy (ref. 1). A single LCLS X-ray pulse destroys any solid material placed in this focus, but the stream replenishes the vaporized sample before the next pulse.

The front detector module, located close to the interaction region, recorded high-angle diffraction to a resolution of 8.5 Å, whereas the rear module intersected diffraction at resolutions in the range of 4,000 to 100 Å. We observed diffraction from crystals smaller than ten unit cells on a side, as determined by examining the data recorded on the rear pnCCDs (Fig. 2). A crystal with a side length of N unit cells gives rise to diffraction features that are finer by a factor of $1/N$ than the Bragg spacing (that is, with $N-2$ fringes between neighbouring Bragg peaks), providing a simple way to determine the projected size of the nanocrystal. Images of crystal shapes obtained using an iterative phase retrieval method[15,16] are shown in Fig. 2. The 3D Fourier transform of the crystal shape is repeated on every reciprocal lattice point. However, the diffraction condition for lattice points is usually not exactly satisfied, so each recorded Bragg spot represents a particular 'slice' of the Ewald sphere through the shape transform, giving a variety of Bragg spot profiles in a pattern; these are apparent in Fig. 2. The sum of counts in each Bragg spot underestimates the underlying structure factor square modulus, representing a partial reflection.

Figure 3a shows strong single-crystal diffraction to the highest angles of the front detector. The nanocrystal shape transform is also apparent in many patterns at the high angles detected by the front detector, giving significant measured intensities between Bragg peaks as is noticeable in Supplementary Fig. 3a. These mid-Bragg intensities

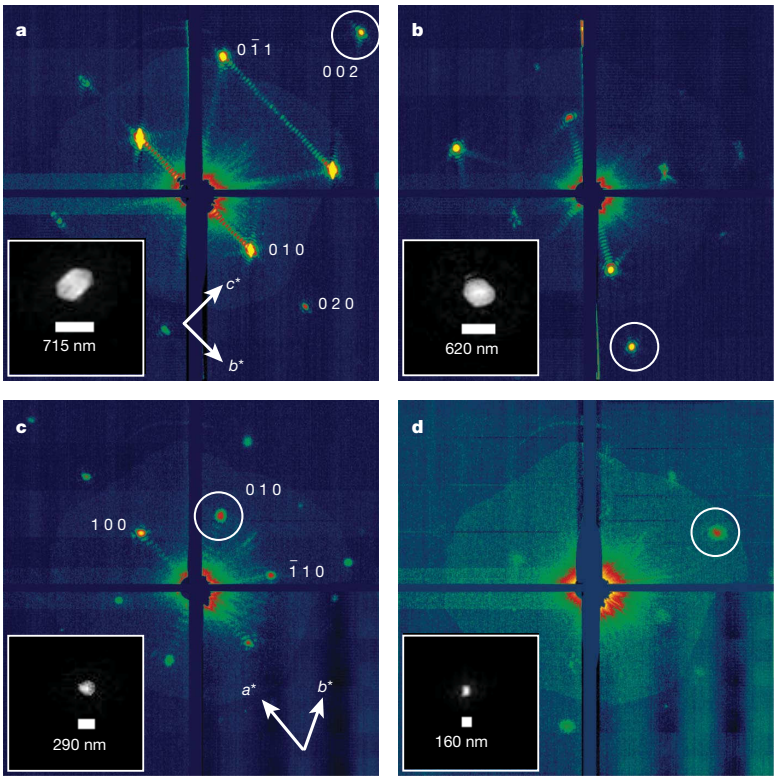


Figure 2 | Coherent crystal diffraction. Low-angle diffraction patterns recorded on the rear pnCCDs, revealing coherent diffraction from the structure of the photosystem I nanocrystals, shown using a logarithmic, false-colour scale. The Miller indices of the peaks in **a** were identified from the corresponding high-angle pattern. In **c** we count seven fringes in the b^* direction, corresponding to nine unit cells, or 250 nm. Insets, real-space images of the nanocrystal, determined by phase retrieval (using the Shrinkwrap algorithm[15]) of the circled coherent Bragg shape transform.

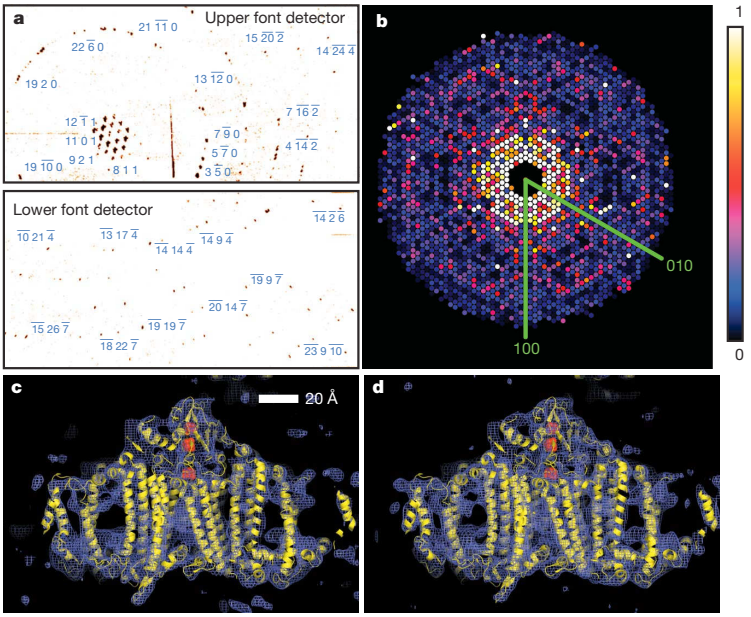


Figure 3 | Diffraction intensities and electron density of photosystem I. **a**, Diffraction pattern recorded on the front pnCCDs with a single 70-fs pulse after background subtraction and correction of saturated pixels. Some peaks are labelled with their Miller indices. The resolution in the lower detector corner is 8.5 Å. **b**, Precession-style pattern of the [001] zone for photosystem I, obtained from merging femtosecond nanocrystal data from over 15,000 nanocrystal patterns, displayed on the linear colour scale shown on the right. **c**, **d**, Region of the $2mF_o - DF_c$ electron density map at 1.0σ (purple mesh), calculated from the 70-fs data (**c**) and from conventional synchrotron data truncated at a resolution of 8.5 Å and collected at a temperature of 100 K (**d**) (Methods). The refined model is depicted in yellow.

oversample the molecular transform, providing a potential route to phasing of the pattern[17,18].

In conventional crystallography, the 'full' Bragg reflection is determined to high precision, for example by integrating counts as the crystal is rotated such that these reflections pass through the diffraction condition. By indexing individual patterns and then summing counts in all partial reflections for each index, we performed a Monte Carlo integration over the reciprocal-space volume of the Bragg reflection and the distribution of crystal shapes and orientations and variations in the X-ray pulse fluence. The result of this procedure converges to the square of the structure factor moduli[18]. We found that over 13% of diffraction patterns with ten or more spots could be consistently indexed using the programs MOSFLM[19] and DirAx[20] (Methods). Merged intensities at 70-fs pulse duration are presented as a precession-style image of the [001]-zone axis in Fig. 3b (see also Supplementary Figs 3 and 4). We tested the reliability of this approach by comparing the LCLS merged data with data collected at 100 K with 12.4-keV synchrotron radiation from a single crystal of photosystem I cryopreserved in 2 M sucrose. These data sets show good agreement, with a difference metric, R_{iso}, of 22.1% computed over the entire resolution range and of less than 13% in the middle resolution shells; see Supplementary Table 1 for detailed statistics.

To complete our proof of principle, we conducted a rigid-body refinement of the published photosystem I structure (Protein Data Bank ID, 1JB0) against the nanocrystal structure factors, yielding $R/R_{free} = 0.25/0.23$. A representative region of the $2mF_o - DF_c$ electron density map at 8.5 Å (Methods) from the LCLS data set is shown in Fig. 3c. This map shows the details expected at this resolution, including transmembrane helices, membrane extrinsic features and some loop structures. For comparison, the electron density refined from the 12.4-keV, single-crystal data set truncated to a resolution of 8.5 Å is given in Fig. 3d.

The dose of 700 MGy corresponds to a K-shell photoabsorption of 3% of all carbon atoms in the protein. This energy is subsequently released by photoionization and Auger decay, followed by a cascade of lower-energy electrons caused by secondary ionizations, taking place on the 10–100-fs timescale[21]. Using a model of the plasma dynamics[22,23], we calculated that by the end of a 100-fs pulse each atom of the crystal was ionized once, on average, and that motion of nuclei had begun. This is expected to give rise to a decrease in Bragg amplitudes, similar to an increase in a Debye–Waller temperature factor[24]. We studied the effects of the initial ionization damage on the diffraction of photosystem I nanocrystals by collecting a series of data sets at pulse durations of 10, 70 and 200 fs. The 10-fs pulses were produced with lower pulse energy: ~10% of the total number of photons of the longer pulses[13], or a 70-MGy dose. Plots of the scattering strength of the crystals versus resolution, generated by selecting and summing Bragg spots from more than 66,000 patterns for each of the three pulse durations measured, are shown in Fig. 4. The 10- and 70-fs traces are very similar, indicating that these pulses are short enough to overcome radiation damage at the observed resolution, 8.5 Å. For 200-fs pulses, there is a decrease in scattering strength at resolutions beyond 25 Å, indicating disordering on this longer timescale. The highest-resolution Bragg peaks for the 200-fs pulses were not broadened or shifted relative to the short-duration data sets, which indicates there was no strain or expansion of the lattice, respectively.

Our next step is to improve resolution by using shorter-wavelength X-rays. Resolution may ultimately be limited by X-ray pulse fluence, the ultrafast radiation damage and the intrinsic disorder within the nanocrystals themselves. Recent experiments[21] at LCLS indicate a brief saturation of the X-ray photoabsorption of atoms in a tightly focused pulse, resulting in a decrease in photoionization damage on a 20-fs timescale without a reduction in the scattering cross-sections that give rise to the diffraction pattern[22]. Planned beamlines at LCLS aim to achieve up to a 10^5-fold increase in pulse irradiance by tighter focusing, allowing data collection with low-fluence, 10-fs pulses or pulses of even shorter duration[25]. This provides a route to further reducing radiation damage and may allow measurements on even smaller nanocrystals,

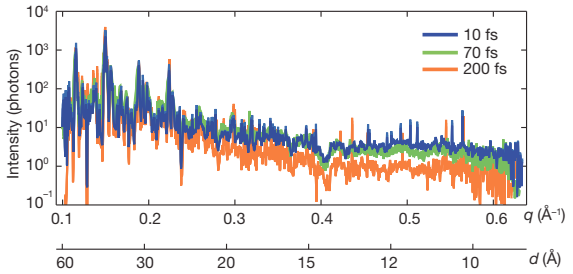


Figure 4 | Pulse-duration dependence of diffraction intensities. Plot of the integrated Bragg intensities of photosystem I nanocrystal diffraction as a function of photon momentum transfer, $q = (4\pi/\lambda)\sin(\theta) = 2\pi/d$ (wavelength, λ; scattering angle, 2θ; resolution, d) for pulse durations of 10, 70 and 200 fs. Averages were obtained by isolating Bragg spots from 97,883, 805,311 and 66,063 patterns, respectively, normalized to pulse fluence. The error in each plot is indicated by the thickness of the line. The decrease in irradiance for 200-fs pulses and $d < 25$ Å indicates radiation damage for these long pulses, which is not apparent for 70-fs pulses and shorter.

down to a single unit cell[6] (that is, a single molecule). As this limit is approached, the ordering of the nanocrystals will become increasingly irrelevant, as each crystal may be treated as a single object and the 'disorder' that conventionally leads to reduced resolution will simply manifest itself as shot-to-shot variability, providing information about not just the average structure but also the range of dynamically accessible conformations.

Data are collected on fully hydrated nanocrystals without cryogenic cooling. We expect that the results presented here will open new avenues for crystallography using X-ray laser pulses that are so short that only negligible X-ray-induced radiation damage occurs during data collection. Significant improvements in sample utilization are expected by exploiting higher X-ray repetition rates or by slowing the liquid flow. For example, the generation, using inkjet technologies, of liquid droplets at a rate that matches the LCLS X-ray pulses would dramatically decrease the total required sample volume by a factor of 25,000, meaning that less than 0.4 µl of nanocrystal suspension would be needed in our particular case, of photosystem I. Further efficiency gains would result from indexing and merging a greater proportion of patterns into the 3D data set, which may be achieved by applying methods for merging continuous diffraction patterns of single molecules[26,27] or by using 'post-refinement'[28] to obtain accurate structure factor estimates from fewer diffraction patterns. These methods will also remove the twinning ambiguity that exists in our current indexing scheme. Our method also has potential application to the study of chemical reactions, such as the processes in photosynthesis or enzymatic reactions.

METHODS SUMMARY

We made our measurements using the CFEL-ASG Multi-Purpose (CAMP) instrument[12] on the Atomic, Molecular and Optical Science beamline[29] at the LCLS[4]. Diffraction data were recorded at the LCLS repetition rate of 30 Hz with a set of two movable, high-frame-rate, low-noise, X-ray pnCCD detector units[12]. The front detector, located 68 mm from the jet, accepts scattering angles up to 47.9°, corresponding to a resolution of 8.5 Å at a wavelength of 6.9 Å. The rear unit was located 564 mm from the jet to record finer sampling of the diffraction pattern at low angles.

The liquid jet was emitted from a capillary with an inner diameter of 40 µm and focused by a coaxial flow of gas to a diameter of about 4 µm (ref. 9), flowing at 10 µl min^{-1}. The low jet diameter constrains the crystals to pass through the most intense part of the focused X-ray beam. Clogging of nanocrystals in the capillary is avoided, and the coaxial gas sheath prevents freezing of the liquid in the vacuum environment. A micropore filter in the fluid delivery line was used to restrict the size of the photosystem I nanocrystals to less than 2 µm. The suspension was diluted to observe a crystal 'hit rate' of 20% (Supplementary Fig. 2) to reduce the occurrence of double hits. The concentration of observed crystals was therefore 0.2 per illuminated volume of $4 \times 4 \times 13$ µm^3, or about 10^9 crystals per millilitre. The overall protein concentration after dilution of the suspension was 1 mg ml^{-1} (1 µM of the photosystem I trimer), and a complete set of structure factors was obtained from 1,850,000 X-ray pulses.

Diffraction peaks from the 70-fs data were identified, indexed and combined into a set of 3D structure factors comprising 3,379 unique reflections from 2,424,394 spots. Statistics of the merged data are given in Supplementary Table 1.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 July; accepted 9 December 2010.

1. Owen, R. L., Rudino-Pinera, E. & Garman, E. F. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proc. Natl Acad. Sci. USA* **103,** 4912–4917 (2006).
2. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28,** 171–193 (1995).
3. Riekel, C. Recent developments in microdiffraction on protein crystals. *J. Synchr. Radiat.* **11,** 4–6 (2004).
4. Emma, P. *et al.* First lasing and operation of an ångstrom-wavelength free-electron laser. *Nature Photon.* **4,** 641–647 (2010).
5. Jordan, P. *et al.* Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature* **411,** 909–917 (2001).
6. Neutze, R., Wout, R., van der Spoel, D., Weckert, E. & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406,** 752–757 (2000).
7. Chapman, H. N. *et al.* Femtosecond time-delay X-ray holography. *Nature* **448,** 676–679 (2007).
8. Spence, J. C. H. & Doak, R. B. Single molecule diffraction. *Phys. Rev. Lett.* **92,** 198102 (2004).
9. DePonte, D. P. *et al.* Gas dynamic virtual nozzle for generation of microscopic droplet streams. *J. Phys. D* **41,** 195505 (2008).
10. Henderson, R. *et al.* Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213,** 899–929 (1990).
11. Wang, D. N. & Kühlbrandt, W. High-resolution electron crystallography of light-harvesting chlorophyll a/b-protein complex in three different media. *J. Mol. Biol.* **217,** 691–699 (1991).
12. Strüder, L. *et al.* Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources. *Nucl. Instrum. Methods Phys. Res. A* **614,** 483–496 (2010).
13. Ding, Y. *et al.* Measurements and simulations of ultralow emittance and ultrashort electron beams in the Linac Coherent Light Source. *Phys. Rev. Lett.* **102,** 254801 (2009).
14. Paithankar, K. S., Owen, R. L. & Garman, E. F. Absorbed dose calculations for macromolecular crystals: improvements to RADDOSE. *J. Synchr. Radiat.* **16,** 152–162 (2009).
15. Marchesini, S. *et al.* X-ray image reconstruction from a diffraction pattern alone. *Phys. Rev. B* **68,** 140101 (2003).
16. Robinson, I. K. & Harder, R. Coherent X-ray diffraction imaging of strain at the nanoscale. *Nature Mater.* **8,** 291–298 (2009).
17. Sayre, D. Some implications of a theorem due to Shannon. *Acta Crystallogr.* **5,** 843 (1952).
18. Kirian, R. *et al.* Femtosecond protein nanocrystallography—data analysis methods. *Opt. Express* **18,** 5713–5723 (2010).
19. Leslie, A. G. The integration of macromolecular diffraction data. *Acta Crystallogr. D* **62,** 48–57 (2006).
20. Duisenberg, A. J. M. Indexing in single-crystal diffractometry with an obstinate list of reflections. *J. Appl. Cryst.* **25,** 92–96 (1992).
21. Young, L. *et al.* Femtosecond electronic response of atoms to ultra-intense X-rays. *Nature* **466,** 56–61 (2010).
22. Hau-Riege, S. P., London, R. A. & Szoke, A. Dynamics of biological molecules irradiated by short X-ray pulses. *Phys. Rev. E* **69,** 051906 (2004).
23. Bergh, M., Huldt, G., Timneanu, N., Maia, F. R. N. C. & Hajdu, J. Feasibility of imaging living cells at subnanometer resolutions by ultrafast X-ray diffraction. *Q. Rev. Biophys.* **41,** 181–204 (2008).
24. Willis, B. & Pryor, A. *Thermal Vibrations in Crystallography* 92 (Cambridge Univ. Press, 1975).
25. Emma, P. *et al.* Femtosecond and subfemtosecond X-ray pulses from a self-amplified spontaneous-emission based free-electron laser. *Phys. Rev. Lett.* **92,** 074801 (2004).
26. Loh, N.-T. D. & Elser, V. Reconstruction algorithm for single-particle diffraction imaging experiments. *Phys. Rev. E* **80,** 026705 (2009).
27. Fung, R., Shneerson, V., Saldin, D. K. & Ourmazd, A. Structure from fleeting illumination of faint spinning objects in flight. *Nature Phys.* **5,** 64–67 (2008).
28. Rossmann, M. G., Leslie, A. G., Sherin, S. A. & Tsukihara, T. Processing and post-refinement of oscillation camera data. *J. Appl. Cryst.* **12,** 570–581 (1979).
29. Bozek, J. D. AMO instrumentation for the LCLS X-ray FEL. *Eur. Phys. J. Spec. Top.* **169,** 129–132 (2009).
30. DePonte, D. P. *et al.* SEM imaging of liquid jets. *Micron* **40,** 507–509 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Experiments were carried out at the Linac Coherent Light Source and the Advanced Light Source, both National User Facilities operated respectively by Stanford University and the University of California on behalf of the US Department of

Energy (DOE), Office of Basic Energy Sciences. We acknowledge support from the DOE through the PULSE Institute at the SLAC National Accelerator Laboratory; the Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344; the Center for Bio-Inspired Solar Fuel Production, an Energy Frontier Research Center funded by the DOE, Office of Basic Energy Sciences (award DE-SC0001016); the Hamburg Ministry of Science and Research and the Joachim Herz Stiftung, as part of the Hamburg Initiative for Excellence in Research (LEXI); the Hamburg School for Structure and Dynamics; the Max Planck Society, for funding the development and operation of the CAMP instrument within the ASG at CFEL; the US National Science Foundation (awards 0417142 and MCB-1021557); the US National Institutes of Health (awards 1R01GM095583-01 (ROADMAP) and 1U54GM094625-01 (PSI:Biology)); the Swedish Research Council; the Swedish Foundation for International Cooperation in Research and Higher Education; Stiftelsen Olle Engkvist Byggmästare; the DFG Cluster of Excellence at the Munich Centre for Advanced Photonics; and the CBST at the University of California under cooperative agreement no. PHY 0120999. We acknowledge discussions with M. Rossmann, E. Snell, R. Stroud and A. Brunger, thank B. Hedman, E. Gullikson, F. Filsinger, A. Berg, H. Mahn and C. Kaiser for technical help and thank the staff of the LCLS for their support in carrying out these experiments.

Author Contributions H.N.C. and J.C.H.S. conceived the experiment, which was designed with P.F., A.B., R.A.K., J.S., D.P.D., U.W., R.B.D., S. Boutet, M.J.B., D.S., I.S., S.M. and J.H. The CAMP instrument was the responsibility of S.W.E., R.H., D. Rolles, A. Rudenko, C.S., L.F., N.K., P.H., B.R., B.E., A.H., Ch.R., D.P., G.W., L.S., G.H., H. Gorke, J.U., I.S., S.H., G.S., F.S., H.S., K.-U.K., R.A., C.-D.S., F.K., M. Bott, S. Schorb, D. Rupp, M.A., T.G., H.H., L.G., G.P., H. Graafsma and B.N., who designed and set up the instrument and/or developed and operated the pnCCD detectors. C.B., J.D.B. and M.M. set up and aligned the beamline. P.F., M.S.H. and I.G. prepared samples; R.B.D., D.P.D., U.W., J.C.H.S., P.F., L.L. and R.L.S. developed and operated the sample delivery system; H.N.C., A.B., A.A., J.S., D.P.D., U.W., R.B.D., S. Bajt, M.J.B., L.G., J.H., M.M.S., N.T., J.A., S. Stern and J.C.H.S. developed diffraction instrumentation; and M. Barthelmess, M.L., A.B. and K.N. designed and/or fabricated calibration samples. J.K., S.P.H.-R., A.B., H.N.C., J.S. and A.V.M. characterized the focus. H.N.C., J.C.H.S., P.F., A.B., T.A.W., R.A.K., A.A., J.S., D.P.D., U.W., R.B.D., I.S., N.C., R.L.S., M.S.H., L.L., M. Bott, S.W.E., R.H., D. Rolles, A. Rudenko, M.L., C.B., J.U., L.F., J.D.B., M.M., M.F., C.Y.H., R.G.S., G.J.W., A. Rocker, M.S., O.J., I.A. and J.H. carried out the experiment. A.B., T.A.W., R.A.K., A.A., F.R.N.C.M., A.V.M., L.L., T.R.M.B., N.C., L.F., N.K., R.N., G.W., P.H., C.C., J.M.H., I.S., J.H., H.N.C. and J.C.H.S. analysed the data. A.V.M. performed the Bragg shape phase retrieval. T.A.W. and R.A.K. merged the 3D data. R.F. collected and evaluated the reference data set; R.A.K., T.A.W., J.M.H. and R.F. refined the structure and calculated the electron density maps; and H.N.C., P.F., J.C.H.S. and I.S. wrote the manuscript with discussion and improvements from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to H.N.C. (henry.chapman@desy.de).

METHODS

Experimental set-up. The experiments were performed at LCLS[4], at SLAC, at the AMO beamline[29] *in vacuo* using the CAMP end station[12]. X-ray pulses, generated at a repetition rate of 30 Hz, were focused to a spot with a full-width at half-maximum of 7 μm (full-width of 13 μm at 10% maximum irradiance) and a pulse fluence of 900 J cm^{-2}, corresponding to a peak power density (irradiance) in excess of 10^{16} W cm^{-2} at 70-fs duration. The pnCCD detectors were read out, digitized and stored at the 30-Hz rate of the delivered LCLS pulses. Each detector panel consists of 512 × 1,024 pixels 75 × 75 μm^2 in area. The rear detectors, located 564 mm from the jet, record low-angle scattering from 0.1° to 4.0° in the vertical scattering plane, and the front detectors, located 68 mm from the jet, cover 4.6° to 40.5° in the same vertical plane. The largest scattering-angle magnitude accepted by the front detector was 47.9°, corresponding to a resolution, *d*, of 8.5 Å at a wavelength of 6.9 Å. X-ray fluorescence from the water jet was filtered by an 8-μm-thick polyimide film in front of the pnCCDs.

A liquid microjet[8,9] was used to inject the nanocrystal suspension into the FEL beam at a flow rate of 10 μl min^{-1}. The microjet was emitted from a 40-μm-diameter capillary and focused to a 4-μm-diameter column by a coaxial flow of helium. The X-ray attenuation in the water was at most 30%. The interaction region of the X-rays and crystals is located in the continuous liquid column, before the Rayleigh break-up of the jet into drops, such that most of the X-ray scattering from the liquid is confined to a narrow vertical streak in reciprocal space.

Crystallization conditions of photosystem I nanocrystals were established by determining the phase diagrams[31,32]. Nanocrystals were grown in batches at 10 mg ml^{-1} protein concentration (30 μM P700, or 10 μM photosystem I trimer) and low ionic strength (8 mM $MgSO_4$, 5 mM MES, pH 6.4, and 0.02% β-dodecylmaltoside) at 4 °C. The photosystem I nanocrystals were then suspended in harvesting buffer (5 mM MES, pH 6.4, and 0.02% β-dodecylmaltoside) to establish a protein concentration of 1 mg ml^{-1}. The crystal suspension was filtered through 2-μm cut-off filters (In-line Filter, Upchurch) and stored at 4 °C until use in the experiment.

The nanocrystals are needles of hexagonal cross-section, with the long axis of the needle along the *c* axis and an aspect ratio (length to maximum diameter of hexagon) ranging from 1:1 to 2:1, as determined from reconstructing single-shot views of the whole crystal from their shape transforms. For example, Fig. 2a shows a view of the crystal almost perpendicular to the *c* axis, where we reconstruct a shape of aspect ratio 1.6:1. A view along the *c* axis (Fig. 2c) shows the hexagonal profile. Large, millimetre-sized, crystals of photosystem I have an aspect ratio of up to 5:1, which is seen to decrease with decreasing crystal size.

The nanocrystal suspension was introduced directly into the microjet through a sample loop (Supplementary Fig. 1). A micropore filter in the fluid delivery line was used to restrict the size of the nanocrystals to less than 2 μm. The suspension was diluted to observe a crystal 'hit rate' of 20% (Supplementary Fig. 2), to minimize the occurrence of double hits. The observed concentration of crystals was therefore 0.2 per illuminated volume of 4 × 4 × 13 μm^3, or 10^9 crystals per millilitre. The overall photosystem I protein concentration after dilution was 1 mg ml^{-1}, and a complete set of structure factors was obtained from 1,850,000 X-ray pulses, or 10 mg of protein. With the current set-up, at the 30-Hz X-ray pulse rate less than 0.004% of the continuously flowing solution was exposed to the X-ray beam, so only one in 25,000 nanocrystals was actually hit by an X-ray pulse.

Details of the acquisition of diffraction patterns and the primary data reduction are given in Supplementary Methods.

3D merging of intensities. Peaks in the processed patterns were located in each pattern using the algorithm of ref. 33, and their locations were mapped into three dimensions according to the curvature of the Ewald sphere, the calibrated detector geometry and the X-ray wavelength. The 3D peak locations for each pattern in turn were presented to the auto-indexing program DirAx[20]. If DirAx succeeded in finding a unit cell for the peaks, linear combinations of the cell basis vectors were checked for correspondence with the photosystem I unit cell[5] from the Protein Data Bank (ID, 1JB0). If a match was found, pixel intensities were summed within a circle of ten-pixel radius centred on the pixel closest to each located Bragg condition. Patterns were rejected if fewer than 10% of the detected peaks were accounted for by unit-cell parameters from DirAx. From 1,850,000 recorded patterns, we identified 112,725 as hits (more than ten detected peaks) and 15,445 were successfully indexed. New peak-finding and -indexing algorithms are under development and are expected to increase significantly the number of patterns that can be indexed, thereby further reducing the number of protein crystals required for a useful data set. The variation of pixel solid angle across the detector plane was accounted for, as was polarization of the X-ray beam assuming complete horizontal polarization. A list of reflection indices and intensities was produced for each individual diffraction pattern, and merging was performed by taking the mean value for the intensity of each unique reflection. Because the indexing algorithm makes use of the positions of the peaks but not their intensities, it was unable to distinguish between crystal orientations related by the symmetry of the lattice. As the symmetry of the lattice is higher than that of the actual structure of photosystem I, an ambiguity exists in that each pattern could correspond to one of two possible orientations. For programmatic convenience, these data (with actual space group symmetry $P6_3$) were merged as $P6_322$ and treated as though merohedrally twinned during refinement (see below). A 3D rendering of the final full data set is shown in Supplementary Fig. 4.

Data quality. Metrics of the merged data quality are shown in Supplementary Table 1 and discussed in Supplementary Information. We carried out a rigid-body refinement of the published photosystem I structure (Protein Data Bank ID, 1JB0) to the merged structure factors using the program REFMAC[34] in twin mode. The refinement *R* and R_{free} values were 0.25 and 0.23, respectively. The $2mF_o - DF_c$ electron density map[35] at a resolution of 8.5 Å is shown in Fig. 3c. The corresponding $2mF_o - DF_c$ electron density map from the conventional synchrotron data, truncated to a resolution of 8.5 Å, is shown in Fig. 3d. The electron density maps show the large subunits PsaA and PsaB, as well as the membrane extrinsic subunits. The transmembrane helices, and even some loop structures, are clearly visible. In these figures, the ribbon representation of the protein model is shown in yellow and the atoms of three iron–sulphur clusters are depicted in red.

31. Fromme, P. & Grotjohann, I. in *Membrane Protein Crystallization* (ed. DeLukas, L.) 192–224 (Curr. Top. Membr. 63, Elsevier, 2009).
32. Hunter, M. S. *et al.* X-ray diffraction from membrane protein nanocrystals. *Biophys. J.* (in the press).
33. Zaefferer, S. New developments of computer-aided crystallographic analysis in transmission electron microscopy. *J. Appl. Cryst.* **33,** 10–25 (2000).
34. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likehood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
35. Praznikar, J., Afonine, P. V., Guncar, G., Adams, P. D. & Turk, D. Averaged kick maps: less noise, more signal...and probably less bias. *Acta Crystallogr. D* **65,** 921–931 (2009).

Bibliography

[Abd08] M.-H. Abdulla, T. O'Brien, Z. B. Mackey, M. Sajid, D. J. Grab, and J. H. McKerrow. RNA Interference of Trypanosoma brucei Cathepsin B and L Affects Disease Progression in a Mouse Model. *PLoS Neglected Tropical Diseases*, vol. 2(9):p. e298, Sep 2008. 57

[All12] E. Allaria, R. Appio, L. Badano, W. A. Barletta, S. Bassanese, S. Biedron, A. Borga, E. Busetto, D. Castronovo, P. Cinquegrana, S. Cleva, D. Cocco, M. Cornacchia, P. Craievich, I. Cudin, G. D'Auria, M. Dal Forno, M. B. Danailov, R. De Monte, G. De Ninno, P. Delgiusto, A. Demidovich, S. Di Mitri, B. Diviacco, A. Fabris, R. Fabris, W. Fawley, M. Ferianis, E. Ferrari, S. Ferry, L. Froehlich, P. Furlan, G. Gaio, F. Gelmetti, L. Giannessi, M. Giannini, R. Gobessi, R. Ivanov, E. Karantzoulis, M. Lonza, A. Lutman, B. Mahieu, M. Milloch, S. V. Milton, M. Musardo, I. Nikolov, S. Noe, F. Parmigiani, G. Penco, M. Petronio, L. Pivetta, M. Predonzani, F. Rossi, L. Rumiz, A. Salom, C. Scafuri, C. Serpico, P. Sigalotti, S. Spampinati, C. Spezzani, M. Svandrlik, C. Svetina, S. Tazzari, M. Trovo, R. Umer, A. Vascotto, M. Veronese, R. Visintini, M. Zaccaria, D. Zangrando, and M. Zangrando. Highly coherent and stable pulses from the FERMI seeded free-electron laser in the extreme ultraviolet. *Nature Photonics*, vol. 6(10):pp. 699–704, Sep 2012. 12

[Alt99] F. Altmann, E. Staudacher, I. B. Wilson, and L. März. Insect cells as hosts for the expression of recombinant glycoproteins. *Glycoconjugate Journal*, vol. 16(2):pp. 109–123, Jan 1999. 32

[Ama12] AmannJ, BergW, BlankV, D. J, DingY, EmmaP, FengY, FrischJ, FritzD, HastingsJ, HuangZ, KrzywinskiJ, LindbergR, LoosH, LutmanA, N. D, RatnerD, RzepielaJ, ShuD, Shvyd'koYu, SpampinatiS, StoupinS, TerentyevS, TrakhtenbergE, WalzD, WelchJ, WuJ, ZholentsA, and ZhuD.

Demonstration of self-seeding in a hard-X-ray free-electron laser. *Nature Photonics*, vol. advance online publication SP - EP -:pp. 1–6, Aug 2012. 12

[Bai88] M. N. Baibich, J. M. Broto, A. Fert, F. N. Van Dau, F. Petroff, P. Etienne, G. Creuzet, A. Friederich, and J. Chazelas. Giant magnetoresistance of (001)fe/(001)cr magnetic superlattices. *Physical Review Letters*, vol. 61(21):pp. 2472–2475, 11 1988. 1

[Bar11] A. Barty, C. Caleman, A. Aquila, N. Timneanu, L. Lomb, T. A. White, J. Andreasson, D. Arnlund, S. Bajt, T. R. M. Barends, M. Barthelmess, M. J. Bogan, C. Bostedt, J. D. Bozek, R. Coffee, N. Coppola, J. Davidsson, D. P. DePonte, R. B. Doak, T. Ekeberg, V. Elser, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, P. Fromme, H. Graafsma, L. Gumprecht, J. Hajdu, C. Y. Hampton, R. Hartmann, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, M. S. Hunter, L. Johansson, S. Kassemeyer, N. Kimmel, R. A. Kirian, M. Liang, F. R. N. C. Maia, E. Malmerberg, S. Marchesini, A. V. Martin, K. Nass, R. Neutze, C. Reich, D. Rolles, B. Rudek, A. Rudenko, H. Scott, I. Schlichting, J. Schulz, M. M. Seibert, R. L. Shoeman, R. G. Sierra, H. Soltau, J. C. H. Spence, F. Stellato, S. Stern, L. Struder, J. Ullrich, X. Wang, G. Weidenspointner, U. Weierstall, C. B. Wunderer, and H. N. Chapman. Self-terminating diffraction gates femtosecond X-ray nanocrystallography measurements. *Nature Photonics*, vol. 6(1):pp. 35–40, Dec 2011. 3, 4, 40, 54, 55, 56

[Ber79] I. B. Bernstein. Amplification on a relativistic electron beam in a spatially periodic transverse magnetic field. *Physical Review A*, vol. 20(4):pp. 1661–1670, Jan 1979. 8

[Ber00] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, vol. 28(1):pp. 235–242, Jan 2000. 2, 24, 36

[Ber08] M. Bergh, G. Huldt, N. Timneanu, F. R. N. C. Maia, and J. Hajdu. Feasibility of imaging living cells at subnanometer resolutions by ultrafast X-ray diffraction. *Quarterly Reviews of Biophysics*, vol. 41(3-4):p. 181, Dec 2008. 55

[Ber10] A. Bergamaschi, R. Dinapoli, B. Henrich, I. Johnson, A. Mozzanica, X. Shi, and B. Schmitt. Beyond single photon counting X-ray detectors. *Nuclear Inst and Methods in Physics Research, A*, pp. 1–4, Jul 2010. 49

[Bin89] G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn. Enhanced magnetoresistance in layered magnetic structures with antiferromagnetic interlayer exchange. *Physical Review B*, vol. 39(7):pp. 4828–4830, 03 1989. 1

[Bon85] R. Bonifacio, C. Pellegrini, and L. Narducci. Collective Instabilities and High Gain Regime in a Free Electron Laser. *Optics Communications*, vol. 50:pp. 373–378, 1985. 8

[Bou10] S. Boutet and G. J Williams. The Coherent X-ray Imaging (CXI) instrument at the Linac Coherent Light Source (LCLS). *New Journal of Physics*, vol. 12(3):p. 035024, Mar 2010. 13, 49, 50, 57

[Bou12] S. Boutet, L. Lomb, G. J. Williams, T. R. M. Barends, A. Aquila, R. B. Doak, U. Weierstall, D. P. DePonte, J. Steinbrener, R. L. Shoeman, M. Messerschmidt, A. Barty, T. A. White, S. Kassemeyer, R. A. Kirian, M. M. Seibert, P. A. Montanez, C. Kenney, R. Herbst, P. Hart, J. Pines, G. Haller, S. M. Gruner, H. T. Philipp, M. W. Tate, M. Hromalik, L. J. Koerner, N. van Bakel, J. Morse, W. Ghonsalves, D. Arnlund, M. J. Bogan, C. Caleman, R. Fromme, C. Y. Hampton, M. S. Hunter, L. C. Johansson, G. Katona, C. Kupitz, M. Liang, A. V. Martin, K. Nass, L. Redecke, F. Stellato, N. Timneanu, D. Wang, N. A. Zatsepin, D. Schafer, J. Defever, R. Neutze, P. Fromme, J. C. H. Spence, H. N. Chapman, and I. Schlichting. High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. *Science*, vol. 337(6092):pp. 362–364, Jan 2012. i, 43, 49

[Brü92] A. T. Brünger. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, vol. 355:p. 472, 1992. 27

[Bry09] C. Bryant, I. D. Kerr, M. Debnath, K. K. Ang, J. Ratnam, R. S. Ferreira, P. Jaishankar, D. Zhao, M. R. Arkin, J. H. McKerrow, L. S. Brinen, and A. R. Renslo. Novel non-peptidic vinylsulfones targeting the S2 and S3 subsites of parasite cysteine proteases. *Bioorganic and Medicinal Chemistry Letters*, vol. 19(21):pp. 6218–6221, 2009. 33

[Cal12] C. Caleman, G. Huldt, F. R. N. C. Maia, C. Ortiz, F. G. Parak, J. Hajdu, D. van der Spoel, H. N. Chapman, and N. Timneanu. On the Feasibility of Nanocrystal Imaging Using Intense and Ultrashort X-ray Pulses. *ACS Nano*, vol. 5(1):pp. 139–146, Jan 2012. 56

[Cas] Correspondence with LCLS machine physicist. http://www.mpi-hd.mpg.de/personalhomes/gitasg/cass/classcass_1_1pp230.html. 14

[Cha08] N. E. Chayen and E. Saridakis. Protein crystallization: from purified protein to diffraction-quality crystal. *Nature Methods*, vol. 5(2):pp. 147–153, Feb 2008. 3

[Cha11] H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall, R. B. Doak, F. R. N. C. Maia, A. V. Martin, I. Schlichting, L. Lomb, N. Coppola, R. L. Shoeman, S. W. Epp, R. Hartmann, D. Rolles, A. Rudenko, L. Foucar, N. Kimmel, G. Weidenspointner, P. Holl, M. Liang, M. Barthelmess, C. Caleman, S. Boutet, M. J. Bogan, J. Krzywinski, C. Bostedt, S. Bajt, L. Gumprecht, B. Rudek, B. Erk, C. Schmidt, A. Homke, C. Reich, D. Pietschner, L. Struder, G. Hauser, H. Gorke, J. Ullrich, S. Herrmann, G. Schaller, F. Schopper, H. Soltau, K.-U. Kuhnel, M. Messerschmidt, J. D. Bozek, S. P. Hau-Riege, M. Frank, C. Y. Hampton, R. G. Sierra, D. Starodub, G. J. Williams, J. Hajdu, N. Timneanu, M. M. Seibert, J. Andreasson, A. Rocker, O. Jonsson, M. Svenda, S. Stern, K. Nass, R. Andritschke, C.-D. Schroter, F. Krasniqi, M. Bott, K. E. Schmidt, X. Wang, I. Grotjohann, J. M. Holton, T. R. M. Barends, R. Neutze, S. Marchesini, R. Fromme, S. Schorb, D. Rupp, M. Adolph, T. Gorkhover, I. Andersson, H. Hirsemann, G. Potdevin, H. Graafsma, B. Nilsson, and J. C. H. Spence. Femtosecond x-ray protein nanocrystallography. *Nature*, vol. 470(7332):pp. 73–77, 02 2011. i, 3, 4, 43

[Che04] L. Chen, R. Oughtred, H. M. Berman, and J. Westbrook. TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, vol. 20(16):pp. 2860–2862, Jan 2004. 3

[Dan85] G. D. Danilatos. Design and construction of an atmospheric or environmental SEM (part 3). *Scanning*, vol. 7(1):pp. 26–42, 1985. 45

[DeP08] D. P. DePonte, U. Weierstall, K. Schmidt, J. Warner, D. Starodub, J. C. H. Spence, and R. B. Doak. Gas dynamic virtual nozzle for generation of microscopic droplet streams. *Journal of Physics D: Applied Physics*, vol. 41(19):p. 195505, 2008. 45

[DeP11a] D. P. DePonte, J. Mckeown, U. Weierstall, R. B. Doak, and J. C. H. Spence. Towards ETEM serial crystallography: Electron diffraction from liquid jets. *Ultramicroscopy*, vol. 111(7):pp. 824–827, Jun 2011. 45

[DeP11b] D. P. DePonte, K. Nass, F. Stellato, M. Liang, and H. N. Chapman. Sample Injection for Pulsed X-ray Sources. *Proceedings of SPIE, the*

International Society for Optical Engineering, Advances in X-ray Free-Electron Lasers: Radiation Schemes, X-ray Optics, and Instrumentation, pp. 80780M–80780M–7, 2011. 46

[Din10] Y. Ding and Z. Huang. Transverse-coherence propoerties of the FEL at the LCLS. *SLAC Publication*, vol. SLAC-PUB-14235, 2010. 11

[Din11] Y. Ding. Femtosecond x-ray pulse temporal characterization in free-electron lasers using a transverse deflector. *Physical Review Special Topics - Accelerators and Beams*, vol. 14(12):pp. 120701 EP –, Jan 2011. 13

[Doy98] D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon. The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity. *Science*, vol. 280(5360):pp. 69–77, Jan 1998. 3

[Doy05] J. P. K. Doye and W. C. K. Poon. Protein crystallization in vivo. *Current Opinion in Colloid & Interface Science*, vol. 11(1):pp. 40–46, Oct 2005. 31

[Dui92] A. Duisenberg. Indexing in single-crystal diffractometry with an obstinate list of reflections. *Journal of Applied Crystallography*, vol. 25(2):pp. 92–96, Jan 1992. 52, 101

[Dun05] K. V. Dunlop, R. T. Irvin, and B. Hazes. Pros and cons of cryocrystallography: should we also collect a room-temperature data set? *Acta Crystallographica Section D: Biological Crystallography*, vol. 61(1):pp. 80–87, Jan 2005. 78

[Ebe74] W. Ebeling, N. Hennrich, M. Klockow, H. Metz, H. D. Orth, and H. Lang. Proteinase K from Tritirachium album Limber. *European Journal of Biochemistry*, vol. 47(1):pp. 91–97, 1974. 29

[Eli76] L. R. Elias. Observation of Stimulated Emission of Radiation by Relativistic Electrons in a Spatially Periodic Transverse Magnetic Field. *Physical Review Letters*, vol. 36(13):pp. 717–720, Jan 1976. 7

[Emm10] P. Emma, R. Akre, J. Arthur, R. Bionta, C. Bostedt, J. Bozek, A. Brachmann, P. Bucksbaum, R. Coffee, F. J. Decker, Y. Ding, D. Dowell, S. Edstrom, A. Fisher, J. Frisch, S. Gilevich, J. Hastings, G. Hays, P. Hering, Z. Huang, R. Iverson, H. Loos, M. Messerschmidt, A. Miahnahri, S. Moeller, H. D. Nuhn, G. Pile, D. Ratner, J. Rzepiela, D. Schultz, T. Smith, P. Stefan, H. Tompkins, J. Turner, J. Welch, W. White, J. Wu,

G. Yocky, and J. Galayda. First lasing and operation of an ångstrom-wavelength free-electron laser. *Nature Photonics*, vol. 4(9):pp. 641–647, Aug 2010. 5, 13

[Eva08] P. Evans and A. McCoy. An introduction to molecular replacement. *Acta Crystallographica Section D*, vol. 64(1):pp. 1–10, Jan 2008. 25

[Fel97] J. Feldhaus, E. L. Saldin, J. R. Schneider, E. Schneidmiller, and M. Yurkov. Possible application of X-ray optical elements for reducing the spectral bandwidth of an X-ray SASE FEL. *Optics Communications*, vol. 140(4–6):pp. 341–352, Jan 1997. 12

[Fio07] E. Fioravanti, F. Vellieux, P. Amara, D. Madern, and M. Weik. Specific radiation damage to acidic residues and its relation to their chemical and structural environment. *Journal of Synchrotron Radiation*, vol. 14:pp. 84 – 91, 2007. 70

[Fre10] J. A. Frearson, S. Brand, S. P. McElroy, L. A. T. Cleghorn, O. Smid, L. Stojanovski, H. P. Price, M. L. S. Guther, L. S. Torrie, D. A. Robinson, I. Hallyburton, C. P. Mpamhanga, J. A. Brannigan, A. J. Wilkinson, M. Hodgkinson, R. Hui, W. Qiu, O. G. Raimi, D. M. F. van Aalten, R. Brenk, I. H. Gilbert, K. D. Read, A. H. Fairlamb, M. A. J. Ferguson, D. F. Smith, and P. G. Wyatt. N-myristoyltransferase inhibitors as new leads to treat sleeping sickness. *Nature*, vol. 464(7289):pp. 728–732, Jan 2010. 57

[Fri09] J. Frisch. Beam Measuremetns at LCLS. *Proceedings of BIW08, Tahoe City, California*, pp. 1–10, Dec 2009. 13

[Fuc09] M. Fuchs, R. Weingartner, A. Popp, Z. Major, S. Becker, J. Osterhoff, I. Cortrie, B. Zeitler, R. Hörlein, G. D. Tsakiris, U. Schramm, T. P. Rowlands-Rees, S. M. Hooker, D. Habs, F. Krausz, S. Karsch, and F. Grüner. Laser-driven soft-X-ray undulator source. *Nature Physics*, vol. 5(11):pp. 826–829, Sep 2009. 12

[Gio12] R. Giordano, R. M. F. Leal, G. P. Bourenkov, S. McSweeney, and A. N. Popov. The application of hierarchical cluster analysis to the selection of isomorphous crystals. *Acta Crystallographica Section D*, vol. 68(6):pp. 649–658, Jun 2012. 78

[Grä08] S. Gräslund, P. Nordlund, J. Weigelt, J. Bray, O. Gileadi, S. Knapp, U. Oppermann, C. Arrowsmith, R. Hui, J. Ming, S. dhe Paganon, H.-w. Park, A. Savchenko, A. Yee, A. Edwards, R. Vincentelli, C. Cambillau, R. Kim,

S.-H. Kim, Z. Rao, Y. Shi, T. C. Terwilliger, C.-Y. Kim, L.-W. Hung, G. S. Waldo, Y. Peleg, S. Albeck, T. Unger, O. Dym, J. Prilusky, J. L. Sussman, R. C. Stevens, S. A. Lesley, I. A. Wilson, A. Joachimiak, F. Collart, I. Dementieva, M. I. Donnelly, W. H. Eschenfeldt, Y. Kim, L. Stols, R. Wu, M. Zhou, S. K. Burley, J. S. Emtage, J. M. Sauder, D. Thompson, K. Bain, J. Luz, T. Gheyi, F. Zhang, S. Atwell, S. C. Almo, J. B. Bonanno, A. Fiser, S. Swaminathan, F. W. Studier, M. R. Chance, A. Sali, T. B. Acton, R. Xiao, L. Zhao, L. C. Ma, J. F. Hunt, L. Tong, K. Cunningham, M. Inouye, S. Anderson, H. Janjua, R. Shastry, C. K. Ho, D. Wang, H. Wang, M. Jiang, G. T. Montelione, D. I. Stuart, R. J. Owens, S. Daenke, A. Schütz, U. Heinemann, S. Yokoyama, K. Büssow, and K. C. Gunsalus. Protein production and purification. *Nature Methods*, vol. 5(2):pp. 135–146, Feb 2008. 3

[Gra09] H. Graafsma. Requirements for and development of 2 dimensional X-ray detectors for the European X-ray Free Electron Laser in Hamburg. *Journal of Instrumentation*, vol. 4(12):pp. P12011–P12011, Dec 2009. 49

[Hal04] B. Halle. Biomolecular cryocrystallography: Structural changes during flash-cooling. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101(14):pp. 4793–4798, Jan 2004. 78

[Hau11] L. M. Haupert and G. J. Simpson. Screening of protein crystallization trials by second order nonlinear optical imaging of chiral crystals (SONICC). *Methods*, vol. 55:pp. 379–386, Dec 2011. 36

[Hed03] L. Hedstrom, L. Gan, Y. G. Schlippe, T. Riera, and M. Seyedsayamdost. IMP dehydrogenase: the dynamics of drug selectivity. *Nucleic Acids Symposium Series*, vol. 3(1):pp. 97–98, Sep 2003. 35

[Hel88] J. Helliwell. Protein crystal perfection and the nature of radiation damage. *Journal of Crystal Growth*, vol. 90(1–3):pp. 259–272, 1988. 3

[Hen93] B. Henke, E. Gullikson, and J. Davis. X-Ray Interactions: Photoabsorption, Scattering, Transmission, and Reflection at E = 50-30,000 eV, Z = 1-92. *Atomic Data and Nuclear Data Tables*, vol. 54(2):pp. 181–342, 1993. 53

[Hit09] R. B. Hitchman, R. D. Possee, and L. A. King. Baculovirus Expression Systems for Recombinant Protein Production in Insect Cells. *Recent Patents on Biotechnology*, vol. 3(1):pp. 46–54, Jan 2009. 32

[Hog98] M. Hogan. Measurements of High Gain and Intensity Fluctuations in a Self-Amplified, Spontaneous-Emission Free-Electron Laser. *Physical Review Letters*, vol. 80(2):pp. 289–292, Jan 1998. 13

[Hol09] J. Holton. A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, vol. 16:pp. 133–142, 2009. 2, 4, 54

[Hol10] J. M. Holton and K. A. Frankel. The minimum crystal size needed for a complete diffraction data set. *Acta Crystallographica Section D: Biological Crystallography*, vol. 66(4):pp. 393–408, Mar 2010. 3, 53

[How09] M. R. Howells, T. Beetz, H. N. Chapman, C. Cui, J. Holton, C. Jacobsen, J. Kirz, E. Lima, S. Marchesini, H. Miao, D. Sayre, D. Shapiro, J. C. H. Spence, and D. Starodub. An assessment of the resolution limitation due to radiation-damage in X-ray diffraction microscopy. *Journal of Electron Spectroscopy and Related Phenomena*, vol. 170(1–3):pp. 4–12, Mar 2009. 54

[Hua06] Z. Huang. Fully Coherent X-Ray Pulses from a Regenerative-Amplifier Free-Electron Laser. *Phys Rev Lett*, vol. 96(14):pp. 144801 EP –, Jan 2006. 11

[Hua07] Z. Huang and K.-J. Kim. Review of x-ray free-electron laser theory. *Physical Review Special Topics - Accelerators and Beams*, vol. 10(3):p. 034801, 2007. 7, 12

[Jue01] D. H. Juers and B. W. Matthews. Reversible lattice repacking illustrates the temperature dependence of macromolecular interactions. *Journal of Molecular Biology*, vol. 311(4):pp. 851–862, Aug 2001. 78

[Kim08] K.-J. Kim, Y. Shvyd'ko, and S. Reiche. A Proposal for an X-Ray Free-Electron Laser Oscillator with an Energy-Recovery Linac. *Physical Review Letters*, vol. 100(24):p. 244802, 2008. 11

[Kir10] R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, J. C. H. Spence, M. Hunter, P. Fromme, T. White, H. N. Chapman, and J. Holton. Femtosecond protein nanocrystallography—data analysis methods. *Optics Express*, vol. 18(6):pp. 5713–5723, Mar 2010. 27, 50

[Kir11] R. Kirian, T. White, J. Holton, H. N. Chapman, P. Fromme, A. Barty, L. Lomb, A. Aquila, F. R. N. C. Maia, A. Martin, R. Fromme, X. Wang, M. S. Hunter, K. Schmidt, and J. C. H. Spence. Structure-factor analysis of femtosecond microdiffraction patterns from protein nanocrystals. *Acta*

Crystallographica Section A: Foundations of Crystallography, vol. 67:pp. 131–140, Feb 2011. 50, 53

[Kis11] D. J. Kissick, D. Wanapun, and G. J. Simpson. Second-Order Nonlinear Optical Imaging of Chiral Crystals. *Annual Review of Analytical Chemistry*, vol. 4:pp. 419–437, Apr 2011. 36

[Kon80] A. M. Kondratenko and E. L. Saldin. Generating Of Coherent Radiation By A Relativistic Electron beam In An ondulator. *Particle Accelerators*, vol. 10:pp. 207–216, 1980. 8

[Koo12] R. Koopmann, K. Cupelli, L. Redecke, K. Nass, D. P. DePonte, T. A. White, F. Stellato, D. Rehders, M. Liang, J. Andreasson, A. Aquila, S. Bajt, M. Barthelmess, A. Barty, M. J. Bogan, C. Bostedt, S. Boutet, J. D. Bozek, C. Caleman, N. Coppola, J. Davidsson, R. B. Doak, T. Ekeberg, S. W. Epp, B. Erk, H. Fleckenstein, L. Foucar, H. Graafsma, L. Gumprecht, J. Hajdu, C. Y. Hampton, A. Hartmann, R. Hartmann, G. Hauser, H. Hirsemann, P. Holl, M. S. Hunter, S. Kassemeyer, R. A. Kirian, L. Lomb, F. R. N. C. Maia, N. Kimmel, A. V. Martin, M. Messerschmidt, C. Reich, D. Rolles, B. Rudek, A. Rudenko, I. Schlichting, J. Schulz, M. M. Seibert, R. L. Shoeman, R. G. Sierra, H. Soltau, S. Stern, L. Struder, N. Timneanu, J. Ullrich, X. Wang, G. Weidenspointner, U. Weierstall, G. J. Williams, C. B. Wunderer, P. Fromme, J. C. H. Spence, T. Stehle, H. N. Chapman, C. Betzel, and M. Duszenko. In vivo protein crystallization opens new routes in structural biology. *Nature Methods*, vol. 9(3):pp. 259–262, Jan 2012. 31, 32, 33, 38, 58

[Kos99] T. A. Kost and J. P. Condreay. Recombinant baculoviruses as expression vectors for insect and mammalian cells. *Current Opinion in Biotechnology*, vol. 10(5):pp. 428–433, October 1999. 32

[lcl] LCLS CONCEPTUAL DESIGN REPORT. `http://www-ssrl.slac.stanford.edu/lcls/cdr/lcls_cdr-ch04.pdf`. 14

[Lec02] F. Lecaille, J. Kaleta, and D. Brömme. Human and Parasitic Papain-Like Cysteine Proteases: Their Role in Physiology and Pathology and Recent Developments in Inhibitor Design. *Chemical Reviews*, vol. 102(12):pp. 4459–4488, Jan 2002. 57

[Les07] A. G. W. Leslie and H. R. Powell. Processing diffraction data with mosflm. In *Evolving Methods for Macromolecular Crystallography*, vol. 245 of *NATO Science Series*, pp. 41–51. Springer Netherlands, 2007. 52, 101

[Lin11] R. R. Lindberg, K.-J. Kim, Y. Shvyd'ko, and W. M. Fawley. Performance of the x-ray free-electron laser oscillator with crystal cavity. *Physical Review Special Topics - Accelerators and Beams*, vol. 14(1):p. 010701, 2011. 11

[Lom11] L. Lomb, T. Barends, S. Kassemeyer, A. Aquila, S. Epp, B. Erk, L. Foucar, R. Hartmann, B. Rudek, D. Rolles, A. Rudenko, R. Shoeman, J. Andreasson, S. Bajt, M. Barthelmess, A. Barty, M. Bogan, C. Bostedt, J. Bozek, C. Caleman, R. Coffee, N. Coppola, D. DePonte, R. B. Doak, T. Ekeberg, H. Fleckenstein, P. Fromme, M. Gebhardt, H. Graafsma, L. Gumprecht, C. Hampton, A. Hartmann, G. Hauser, H. Hirsemann, P. Holl, J. Holton, M. Hunter, W. Kabsch, N. Kimmel, R. Kirian, M. Liang, F. R. N. Maia, A. Meinhart, S. Marchesini, A. Martin, K. Nass, C. Reich, J. Schulz, M. M. Seibert, R. Sierra, H. Soltau, J. C. Spence, J. Steinbrener, F. Stellato, S. Stern, N. Timneanu, X. Wang, G. Weidenspointner, U. Weierstall, T. White, C. Wunderer, H. Chapman, J. Ullrich, L. Struder, and I. Schlichting. Radiation damage in protein serial femtosecond crystallography using an x-ray free-electron laser. *Physical Review B*, vol. 84(21), Dec 2011. 2, 40, 56

[Lom12] L. Lomb, J. Steinbrener, S. Bari, D. Beisel, D. Berndt, C. Kieser, M. Lukat, N. Neef, and R. L. Shoeman. An anti-settling sample delivery instrument for serial femtosecond crystallography. *Journal of Applied Crystallography*, vol. 45(4):pp. 674–678, Jan 2012. 50, 117

[LP84] Y. Le Page, J. Donnay, and G. Donnay. Printing sets of structure factors for coping with orientation ambiguities and possible twinning by merohedry. *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 40(6):pp. 679–684, 1984. 52

[Mac04] Z. B. Mackey, T. C. O'Brien, D. C. Greenbaum, R. B. Blank, and J. H. McKerrow. A Cathepsin B-like Protease Is Required for Host Protein Degradation in Trypanosoma brucei. *Journal of Biological Chemistry*, vol. 279(46):pp. 48426–48433, Jan 2004. 32, 57

[Mad71] J. M. J. Madey. Stimulated emission of bremsstrahlung in a periodic magnetic field. *Journal of Applied Physics*, vol. 42(5):pp. 1906–1913, 1971. 7

[Neu12] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, and J. Hajdu. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, vol. 406(6797):pp. 752–757, Sep 2012. i

[Noy87] I. Noyan and J. Cohen. *Residual Stress: Measurement by Diffraction and Interpretation*. Materials Research and Engineering Series. Springer-Verlag, 1987. 1

[O'N02] P. O'Neill, D. Stevens, and E. Garman. Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective. *Journal of Synchrotron Radiation*, vol. 9:pp. 329–332, 2002. 54, 70

[Owe06] R. L. Owen, E. Rudiño-Piñera, and E. F. Garman. Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103(13):pp. 4912–4917, Jan 2006. 54

[Owe09] R. L. Owen, J. M. Holton, C. Schulze-Briese, and E. F. Garman. Determination of X-ray flux using silicon pin diodes. *Journal of Synchrotron Radiation*, vol. 16(2):pp. 143–151, Feb 2009. 38

[Pai10] K. S. Paithankar and E. F. Garman. Know your dose: RADDOSE. *Acta Crystallographica Section D: Biological Crystallography*, vol. 66(4):pp. 381–388, Jan 2010. 55

[Pat39] A. L. Patterson. The Scherrer Formula for X-Ray Particle Size Determination. *Physical Review*, vol. 56(10):pp. 978–982, Jan 1939. 92

[Phi10] H. T. Philipp, L. J. Koerner, M. Hromalik, M. W. Tate, and S. M. Gruner. Femtosecond Radiation Experiment Detector for X-Ray Free-Electron Laser (XFEL) Coherent X-Ray Imaging. *IEEE Transactions on Nuclear Science*, vol. 57(6):pp. 3795–3799, 2010. 49

[Pow99] H. Powell. The Rossmann Fourier autoindexing algorithm in MOSFLM. *Acta Crystallographica Section D: Biological . . .*, 1999. 52

[Red12] L. Redecke, K. Nass, D. P. DePonte, T. A. White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. M. Barends, S. Boutet, G. J. Williams, M. Messerschmidt, M. M. Seibert, A. Aquila, D. Arnlund, S. Bajt, T. Barth, M. J. Bogan, C. Caleman, T.-C. Chao, R. B. Doak, H. Fleckenstein, M. Frank, R. Fromme, L. Galli, I. Grotjohann, M. S. Hunter, L. C. Johansson, S. Kassemeyer, G. Katona, R. A. Kirian, R. Koopmann, C. Kupitz, L. Lomb, A. V. Martin, S. Mogk, R. Neutze, R. L. Shoeman, J. Steinbrener, N. Timneanu, D. Wang, U. Weierstall, N. A. Zatsepin, J. C. H. Spence, P. Fromme, I. Schlichting, M. Duszenko, C. Betzel, and H. N. Chapman. Natively Inhibited Trypanosoma brucei Cathepsin B

Structure Determined by Using an X-ray Laser. *Science*, vol. advance online publication, Jan 2012. 57

[Ric03] M. Richter, A. Gottwald, U. Kroth, A. A. Sorokin, S. V. Bobashev, L. A. Shmaenok, J. Feldhaus, C. Gerth, B. Steeg, K. Tiedtke, and R. Treusch. Measurement of gigawatt radiation pulses from a vacuum and extreme ultraviolet free-electron laser. *Applied Physics Letters*, vol. 83(14):p. 2970, 2003. 13

[Rie04] C. Riekel. Recent developments in microdiffraction on protein crystals. *Journal of Synchrotron Radiation*, vol. 11:pp. 4–6, Jan 2004. 3

[Ros62] M. G. Rossmann and D. M. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, vol. 15(1):pp. 24–31, Jan 1962. 25

[Ros79] M. G. Rossmann, A. G. W. Leslie, S. S. Abdel-Meguid, and T. Tsukihara. Processing and post-refinement of oscillation camera data. *Journal of Applied Crystallography*, vol. 12(6):pp. 570–581, Jan 1979. 52

[Rud12] B. Rudek, S.-K. Son, L. Foucar, S. W. Epp, B. Erk, R. Hartmann, M. Adolph, R. Andritschke, A. Aquila, N. Berrah, C. Bostedt, J. Bozek, N. Coppola, F. Filsinger, H. Gorke, T. Gorkhover, H. Graafsma, L. Gumprecht, A. Hartmann, G. Hauser, S. Herrmann, H. Hirsemann, P. Holl, A. Homke, L. Journel, C. Kaiser, N. Kimmel, F. Krasniqi, K.-U. Kuhnel, M. Matysek, M. Messerschmidt, D. Miesner, T. Möller, R. Moshammer, K. Nagaya, B. Nilsson, G. Potdevin, D. Pietschner, C. Reich, D. Rupp, G. Schaller, I. Schlichting, C. Schmidt, F. Schopper, S. Schorb, C.-D. Schroter, J. Schulz, M. Simon, H. Soltau, L. Struder, K. Ueda, G. Weidenspointner, R. Santra, J. Ullrich, A. Rudenko, and D. Rolles. Ultra-efficient ionization of heavy atoms by intense X-ray free-electron laser pulses. *Nature Photonics*, vol. advance online publication, Nov 2012. 53

[Sal96] E. L. Saldin, E. Schneidmiller, and M. Yurkov. Calculation of energy diffusion in an electron beam due to quantum fluctuations of undulator radiation. *Nuclear Inst and Methods in Physics Research, A*, vol. 381(2–3):pp. 545–547, 1996. 14

[Sal10] E. L. Saldin, E. A. Schneidmiller, and M. V. Yurkov. Statistical and coherence properties of radiation from x-ray free-electron lasers. *New Journal of Physics*, vol. 12(3):p. 035010, 2010. 10

[Sch18] P. Scherrer. *Gott Nachr*, vol. 2(98), 1918. 92

[Sch58] A. L. Schawlow and C. H. Townes. Infrared and Optical Masers. *Physical Review*, vol. 112(6):pp. 1940–1949, Jan 1958. 7

[Sco94] H. Scott and R. Mayle. GLF - A simulation code for X-ray lasers. *Applied Physics B Lasers and Optics*, vol. 58(1):pp. 35–43, 1994. 55

[Sco01] H. A. Scott. Cretin-a radiative transfer capability for laboratory plasmas. In *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 71, pp. 689–701. October 2001. 55

[SD07] R. J. Southworth-Davies, M. A. Medina, I. Carmichael, and E. F. Garman. Observation of Decreased Radiation Damage at Higher Dose Rates in Room Temperature Protein Crystallography. *Structure*, vol. 15(12):pp. 1531–1541, Dec 2007. 54, 55

[Sel12] J. A. Sellberg. *Temperature-dependent X-ray Scattering of Liquid Water*. Ph.D. thesis, Stockholm University, Stockholm, May 2012. http://www.physto.se/~jose4950/Sellberg-LicThesis2012.pdf. 45

[Sie12] R. G. Sierra, H. Laksmono, J. Kern, R. Tran, J. Hattne, R. Alonso-Mori, B. Lassalle-Kaiser, C. Glöckner, J. Hellmich, D. W. Schafer, N. Echols, R. J. Gildea, R. W. Grosse-Kunstleve, J. Sellberg, T. A. McQueen, A. R. Fry, M. M. Messerschmidt, A. Miahnahri, M. M. Seibert, C. Y. Hampton, D. Starodub, N. D. Loh, D. Sokaras, T.-C. Weng, P. H. Zwart, P. Glatzel, D. Milathianaki, W. E. White, P. D. Adams, G. J. Williams, S. Boutet, A. Zouni, J. Messinger, N. K. Sauter, U. Bergmann, J. Yano, V. K. Yachandra, and M. J. Bogan. Nanoflow electrospinning serial femtosecond crystallography. *Acta Crystallographica Section D: Biological Crystallography*, vol. 68(11):pp. 1584–1587, Oct 2012. 45, 113

[Smi06] J. D. Smith, C. D. Cappa, W. S. Drisdell, R. C. Cohen, and R. J. Saykally. Raman Thermometry Measurements of Free Evaporation from Liquid Water Droplets. *Journal of the American Chemical Society*, vol. 128(39):pp. 12892–12898, Jan 2006. 45

[Son11] S.-K. Son, H. N. Chapman, and R. Santra. Multiwavelength Anomalous Diffraction at High X-Ray Intensity. *Physical Review Letters*, vol. 107(21):p. 218102, Jan 2011. i, 112

[Spe11] J. C. H. Spence, R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, T. White, A. Barty, H. N. Chapman, S. Marchesini, and J. Holton. Phasing of coherent femtosecond X-ray diffraction from size-varying nanocrystals. *Optics Express*, vol. 19(4):pp. 2866–2873, Jan 2011. i, 113

[Spe12] J. C. H. Spence, U. Weierstall, and H. N. Chapman. X-ray lasers for structural and dynamic biology. *Reports on Progress in Physics*, vol. 75(10):p. 102601, Sep 2012. 53

[Ste97] I. Steller, R. Bolotovsky, and M. G. Rossmann. An Algorithm for Automatic Indexing of Oscillation Images using Fourier Analysis. *Journal of Applied Crystallography*, vol. 30(6):pp. 1036–1040, Jan 1997. 101

[Str10] L. Struder, S. Epp, D. Rolles, R. Hartmann, P. Holl, G. Lutz, H. Soltau, R. Eckart, C. Reich, K. Heinzinger, C. Thamm, A. Rudenko, F. Krasniqi, K.-U. Kuhnel, C. Bauer, C.-D. Schroter, R. Moshammer, S. Techert, D. Miessner, M. Porro, O. Hälker, N. Meidinger, N. Kimmel, R. Andritschke, F. Schopper, G. Weidenspointner, A. Ziegler, D. Pietschner, S. Herrmann, U. Pietsch, A. Walenta, W. Leitenberger, C. Bostedt, T. Möller, D. Rupp, M. Adolph, H. Graafsma, H. Hirsemann, K. Gärtner, R. Richter, L. Foucar, R. L. Shoeman, I. Schlichting, and J. Ullrich. Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 614(3):pp. 483–496, Mar 2010. 38

[Taj79] T. Tajima and J. M. Dawson. Laser Electron Accelerator. *Physical Review Letters*, vol. 43(4):pp. 267–270, Jan 1979. 12

[Wam08] R. D. Wampler, D. J. Kissick, C. J. Dehen, E. J. Gualtieri, J. L. Grey, H.-F. Wang, D. H. Thompson, J.-X. Cheng, and G. J. Simpson. Selective Detection of Protein Crystals by Second Harmonic Microscopy. *Journal of the American Chemical Society*, vol. 130(43):pp. 14076–14077, Jan 2008. 36

[Wei00] M. Weik, R. B. G. Ravelli, G. Kryger, S. McSweeney, M. L. Raves, M. Harel, P. Gros, I. Silman, J. Kroon, and J. L. Sussman. Specific chemical and structural damage to proteins produced by synchrotron radiation. *Proceedings of the National Academy of Sciences*, vol. 97(2):pp. 623–628, Jan 2000. 70

[Wei12] U. Weierstall, J. C. H. Spence, and R. B. Doak. Injector for scattering measurements on fully solvated biospecies. *Review of Scientific Instruments*, vol. 83(3):p. 035108, 2012. 50

[Whi12] T. A. White, R. A. Kirian, A. V. Martin, A. Aquila, K. Nass, A. Barty, and H. N. Chapman. CrystFEL: a software suite for snapshot serial crys-

tallography. *Journal of Applied Crystallography*, vol. 45(2):pp. 335–341, Jan 2012. 51, 53, 77, 107, 119, 120

[Yu91] L. H. Yu. Generation of intense uv radiation by subharmonically seeded single-pass free-electron lasers. *Physical Review A*, vol. 44(8):pp. 5178–5193, Jan 1991. 8, 12

[ZA12] A. Zarrine-Afsar, T. R. M. Barends, C. Müller, M. R. Fuchs, L. Lomb, I. Schlichting, and R. J. D. Miller. Crystallography on a chip. *Acta Crystallographica Section D: Biological Crystallography*, vol. 68(3):pp. 321–323, Jan 2012. 113

[Zha99] R.-g. Zhang, G. Evans, F. J. Rotella, E. M. Westbrook, D. Beno, E. Huberman, A. Joachimiak, and F. R. Collart. Characteristics and Crystal Structure of Bacterial Inosine-5‘-monophosphate Dehydrogenase. *Biochemistry*, vol. 38(15):pp. 4691–4700, Apr 1999. 65