Compressive auto-indexing in femtosecond nanocrystallography

Filipe R. N. C. Maia, Chao Yang, Stefano Marchesini

Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Abstract

Ultrafast nanocrystallography has the potential to revolutionize biology by enabling structural elucidation of proteins for which it is possible to grow crystals with 10 or fewer unit cells. The success of nanocrystallography depends on robust orientation-determination procedures that allow us to average diffraction data from multiple nanocrystals to produce a 3D diffraction data volume with a high signal-to-noise ratio. Such a 3D diffraction volume can then be phased using standard crystallographic techniques. "Indexing" algorithms used in crystallography enable orientation determination of a diffraction data from a single crystal when a relatively large number of reflections are recorded. Here we show that it is possible to obtain the exact lattice geometry from a smaller number of measurements than standard approaches using a basis pursuit solver.

Keywords: indexing; crystallography; compressive sensing

1. Introduction

X-ray crystallography is currently the leading method for atomic resolution imaging of macromolecules. Third generation synchrotron sources permit successful structure solution from crystals 5 microns in size or greater. This limits the success rate of structure solution to a few percent.

The Linac Coherent Light Source (LCLS) recently began operation [1] at the SLAC National Accelerator Laboratory in Palo Alto, California, using energetic electrons from a linear accelerator to produce coherent x-rays with an instrument called a free electron laser (FEL). Free Electron Laser sources produce pulses of light that are over 10 orders of magnitude brighter than current third generation synchrotron sources [2]. Several other x-ray laser sources of this type are being built or planned worldwide.

The high number of photons incident on a specimen are expected to produce measurable diffraction patterns from nanocrystals with as few as 10 unit cells, enabling high resolution structure elucidation of systems which can only be crystallized into very small crystals that are not suitable for conventional crystallography. Even for larger crystals, the short pulses can circumvent the radiation damage problem [3, 4] which limits the resolution of many sensitive crystals.

In such an experiment, two-dimensional (2D) diffraction images of randomly oriented nanocrystal of the same type can be captured within femtosecond exposure time. These images can then be used to deduce the 3D structure of the molecule. To see the structure in 3-D, one has to merge the data from all these individual nanocrystals, whose orientations are not known.

Femtosecond nanocrystallography brings new challenges to data processing [5]. One problem is that the orientation of each diffraction image obtained is unknown. Another problem is that a single snapshot of the crystal diffraction pattern may contain very few reflections. In traditional crystallography, a small angular range of integration ensures that many Bragg reflections are recorded while ensuring that overlaps are minimized. This is not possible with ultrafast x-ray pulses. The relentless improvements of these light sources (beam energy, beam divergence and wavelength) will further exacerbate the problem.

These new difficulties make indexing such patterns a hard problem for existing crystallographic software.

2. Structure Determination from Crystal Diffraction

In traditional crystallography, diffraction images are collected while rotating a sample. A small angular range of integration ensures that all Bragg reflections are recorded while ensuring that overlaps are minimized. The strength (structure factor) and orientation of each Bragg reflection is estimated from the diffraction geometry (including source divergence, bandwidth, pixel size and angular average). The diffracted photon flux I (pho-

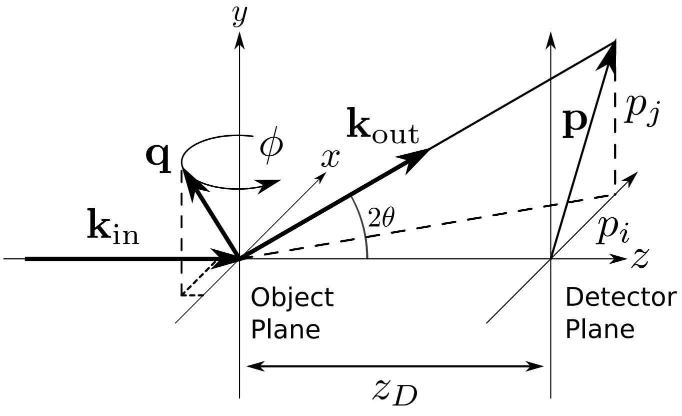


Figure 1: Scattering geometry for coherent x-ray diffraction imaging.

tons/pulse/ pixel) produced by a crystallite is given by

$$I(\mathbf{q}) = J_o r_e^2 P \Delta\Omega \left| F(R_\phi \mathbf{q}) \sum_{h,k,l} (R_\phi \mathbf{q} - S(h\hat{\mathbf{h}} + k\hat{\mathbf{k}} + l\hat{\mathbf{l}}) \right|^2 , \qquad (1)$$

where $F(\mathbf{q})$ is the continuous scattering from one unit cell (molecule), R_ϕ is the 3D rotation matrix of the unknown object orientation, $\mathbf{q}$ a vector that relates the Bragg “reflection” to a point in a three dimensional Fourier space, J_o is the incident photon flux density (photons/pulse/area), r_e^2 is the electron cross section, P is a polarization factor, $\Delta\Omega$ is the solid angle subtended by a detector pixel at the sample, the (h, k, l) integer values are called Miller indices, $(\hat{\mathbf{h}}, \hat{\mathbf{k}}, \hat{\mathbf{l}})$ identify the Bravais lattice characteristic of the crystal periodic structure, and S is the shape transform of the crystallite finite dimensions. For large crystals, S is simply a Dirac δ-function. To determine the unit cell structure factor from nanocrystals truncated after a few unit cells, an additional finite-size effect needs to be accounted for.

In x-ray crystallography, the term *indexing* refers to the task of assigning the measured Bragg peaks to the discrete locations of a periodic lattice. Auto-indexing uses the position of these peaks to deduce the shape and orientation of the lattice, and to identify the lattice coordinates of each measured peak. It is accomplished in several steps

- Bragg peaks are identified in each image and their 2D pixel coordinates $p_{i,j} = p(i\hat{\mathbf{i}} + j\hat{\mathbf{j}})$ are recorded. These coordinates are mapped to

3D coordinates in the 3D reciprocal space according to the geometric description of elastic scattering shown in Figure 1. In this figure, $\mathbf{k}_{\text{in}}$ and $\mathbf{k}_{\text{out}}$ are the incident and scattered wave vectors that satisfy $|\mathbf{k}_{\text{in}}| = |\mathbf{k}_{\text{out}}| = 1/\lambda$, where λ is the wavelength of the x-ray. The distance between the sample, which is placed at the origin, and the detector is assumed to be z_D. An Ewald sphere centered at $(0, 0, -k)$ is drawn in the figure. The radius of the sphere is k. The reciprocal lattice point that contributes to measured Bragg peak at $p_{i,j}$ can be located as the vector $\mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}}$ whose end point lies on the Ewald sphere. The coordinates of this lattice point $\mathbf{q}_{i,j}$ satisfy

$$\mathbf{q}_{i,j} = \mathbf{k}_{\text{out}} - \mathbf{k}_{\text{in}} = \frac{1}{\lambda}\left(\frac{\mathbf{p}_{i,j} + z_D\hat{\mathbf{k}}_{\text{in}}}{\sqrt{|p_{i,j}|^2 + z_D^2}} - \hat{\mathbf{k}}_{\text{in}}\right), \tag{2}$$

where $\mathbf{p}_{i,j} = p(i\hat{\mathbf{i}} + j\hat{\mathbf{j}} + 0\hat{\mathbf{k}})$.

- For the purpose of autoindexing, one can simply assign the value of 1 to $I(\mathbf{q}_{i,j})$ for every peak above a noise threshold. As a result, one obtains a 3D map $I(q)$ in the reciprocal space that contains the values of either 1 or 0.

- Some type of computational analysis is performed on the 3D map to ascertain the orientation and the unit cell parameters of the crystal. The analysis typically proceeds by first determining the reciprocal lattice vectors, and it often makes use of Fourier transform and peak searches. An efficient algorithm that uses many 1D Fourier transform was proposed in [6, 7]. It is used in many existing autoindexing software packages such as MOSFLM [8]. We will provide details of these algorithms in the next section.

Once the orientation and the unit cell parameters associated with a crystal has been determined, one may then proceed to estimate the structure factor of the crystal, from the diffraction geometry (including source divergence, bandwidth, pixel size and angular average). Finally, a phase retrieval algorithm is used to recover the phase of the Fourier transform and subsequently the 3D density map of the crystal.

For the purpose of this paper, we will not discuss the issues of structure factor determination or the phase retrieval problem. Instead, we will focus on the autoindexing problem.

3. Real space autoindexing

Most autoindexing algorithms search for peaks in real space, by applying some form of three dimensional Fourier transform of the binary reciprocal space map. A simple numerical thresholding may reveal the positions of the 3D lattice points in real space. They can subsequently be used to determine the unit cell parameters, crystal orientation and type.

The use of a three-dimensional Fourier transform around the origin for indexing a diffraction pattern was suggested over two decades ago [9]. A similar approach appears to have been incorporated in the program DENZO, which has been distributed as part of the diffraction-image processing suite HKL [10]. A three-dimensional FFT has been used to index diffraction images by calculating a Patterson function from a set of reflections which have all been assigned unit intensity [11]. Efficient implementations make use of the Fourier projection-slice theorem [12], calculating 1D sections of the three dimensional FTs by a series of projections and 1-dimensional FFTs [7] [13]. Indexing software such as MOSFLM [6], LABELIT [14] utilize this approach.

The complexity of the projection-FFT approach is $mn \log n$, where m is the number of direction vectors that must be generated, and n is the number of samples along the projected 1D intensity profile, which is proportional to $N^{1/3}$, where N is the total number of sampled voxels contained in the crystal. A typical value of m is between 5,000 and 20,000. Clearly, this method will not work well if the number of Bragg points on a diffraction pattern is small.

Although the argument used in [7] for abandoning the full 3D FFT is the high cost for performing 3D FFTs of large crystals, this is no longer a serious issue due to the rapid growth in the processing speed and memory capacity of modern multi-core microprocessors. At the time of writing, a 3D (512^3) FFT takes about 0.15 seconds on a GPU processor. Furthermore, there are now algorithms that we may use to take full advantage of the sparsity of the 3D reciprocal space map [15], i.e., there are a few non-zeros in the 3D map constructed from the Bragg reflections, and reduce the complexity of the 3D FFT calculation from the standard $\mathcal{O}(N^2 \log N)$ to that of $\mathcal{O}(N^{2/3} \log N)$, where N is the total number of sampled voxels in the crystal.

As we will show in section 6, when the number of measured Bragg peaks is less than 10, the real space lattice points cannot be easily distinguished from the rest of the sampled voxels based on the intensity of the inverse 3D Fourier transform.

4. Recovering Real Space Lattice via L1 Minimization

An alternative technique for retrieving the positions of the real (and reciprocal) space lattice points associated with a crystal is to use the recently developed compressive sensing methodology [16, 17, 18] and formulate the problem as an L1 minimization problem.

Let x be a vector representation of the 3D density map of a crystal lattice to be determined in real space. Similar to the 3D inverse Fourier transform approach, we will use the magnitude of each component of x to determine whether the 3D voxel associated with that component is a real space lattice point.

The vector x is related to the diffraction measurement through the following equation

$$b_{j_i} = e_{j_i}^T F x, \quad \text{for} \quad i = 1, 2, ..., 2m, \tag{3}$$

where b_{j_i} is the intensity assigned to a sampled 3D reciprocal space voxel that lies on the Ewald sphere, m is the total number of voxels on the Ewald sphere, F is the matrix representation of a 3D discrete Fourier transform, and e_{j_i} is the j_ith column of the identity matrix. To ensure that x is real, Friedel's conjugate symmetry is imposed on b_{j_i}. Therefore, we have $2m$ equations in (3) even though the number of samples on the Ewald sphere is m.

Because the number of sampled voxels on a Ewald sphere is always far fewer than the number of reciprocal lattice points, the linear system defined by (3) is clearly underdetermined. Therefore, x cannot be recovered by simply solving (3). However, because the vector x to be recovered is expected to be "sparse", i.e., it is expected to have nonzero values at a subset of real space voxels, it follows from the recently developed compressive sensing theory [17, 18], that we may be able to recover x by solving the following convex minimization problem

$$\begin{array}{cc} \min_x & \|x\|_1 \\ \text{s.t.} & MFx = b, \end{array} \tag{4}$$

where M is an $2m \times n$ sparse "sensing" matrix that contains $e_{j_i}^T$, $i = 1, 2, ..., 2m$, as its rows, b is a vector representation of the intensity values (0's and 1's) assigned to voxels on the Ewald sphere (and its Friedel symmetric counterpart), $\| \cdot \|_1$ denotes the L_1 norm of a vector.

The equality constraint in (4) can be relaxed to an inequality constraint of the form

$$\|MFx - b\| \leq \sigma,$$

where $\|\cdot\|_2$ denotes the L_2-norm of a vector, for some small constant σ to allow imprecise measurements or noise in the data. The relaxed minimization problem is often known as the basis pursuit denoising (BPDN) problem, and the original L1 minimization problem (4) is also known as the basis pursuit (BP) problem.

5. Algorithms for Solving the L1 Minimization Problem

The L1 minimization problem (4) and its BPDN relaxed form can be solved in a number of ways. In the software package SPGL1 [19], which we use to perform the numerical experiments shown in the next section, the BPDN problem is reduced to a sequence of what is known as the LASSO [20] problems

$$\begin{array}{ll} \min_x & \|MFx - b\|_2 \\ \text{s.t.} & \|x\|_1 \leq \tau, \end{array} \tag{5}$$

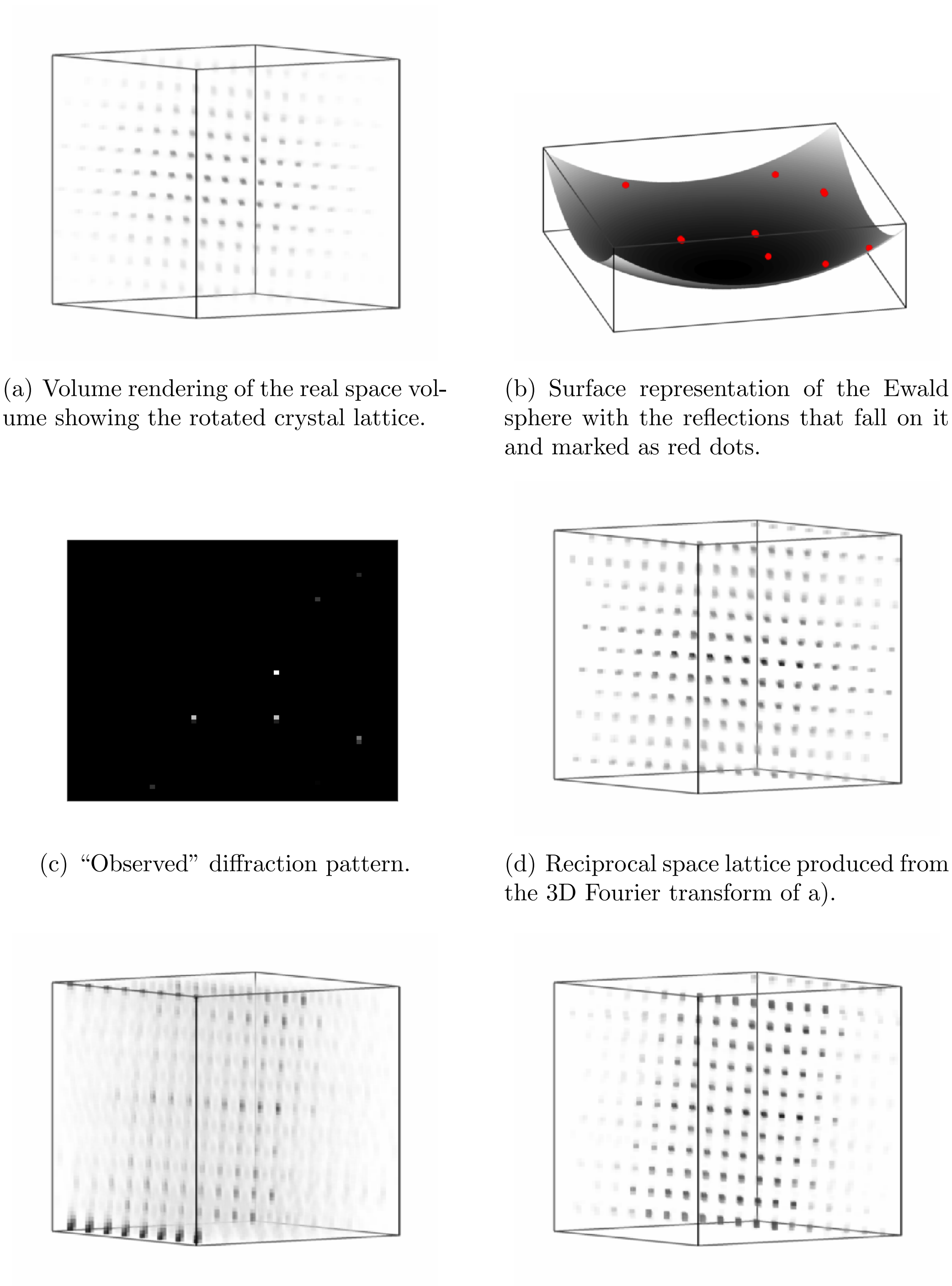
where τ is a parameter that is determined in an iterative process that involves finding the root of nonlinear equation $\phi(\tau) = \sigma$, where $\phi(\tau)$, which is the optimal value of the objective function in (5) for a given τ, is known as the Pareto curve. The LASSO problem is solved by a spectral projected gradient method [21, 22, 23] in SPGL1.

An alternative approach for solving the BPDN problem is to apply a first-order method developed by Y. Nesterov [24, 25] to solve (4) directly. A software package based on this approach is called NESTA [26].

The computational cost of both SPGL1 and NESTA is dominated by the calculation of Fx and F^*x, i.e., 3D fast Fourier and inverse Fourier transforms required in each iteration. Therefore, the overall cost of an autoindexing scheme based on L1 minimization formulation of the problem is higher compared to the existing approaches. However, as we will see in the next section, the advantage of the method is that it can recover the real space (and reciprocal space) lattice points reliably even when only a few Bragg peaks can be identified on a diffraction image.

6. Computational Experiment

To test the algorithm we created a 64^3 real space volume which was then populated with a cubic lattice that contains $8 \times 8 \times 8$ voxels. A rotation was then applied to the lattice (fig. 2(a)) and the result was Fourier transformed to generate the 3D diffraction volume (fig. 2(d)).



(a) Volume rendering of the real space volume showing the rotated crystal lattice.

(b) Surface representation of the Ewald sphere with the reflections that fall on it and marked as red dots.

(c) "Observed" diffraction pattern.

(d) Reciprocal space lattice produced from the 3D Fourier transform of a).

(e) Reconstructed real space volume using the 3D inverse Fourier transform of the observed data.

(f) Reconstructed real space volume using the L1 minimization method.

Figure 2:

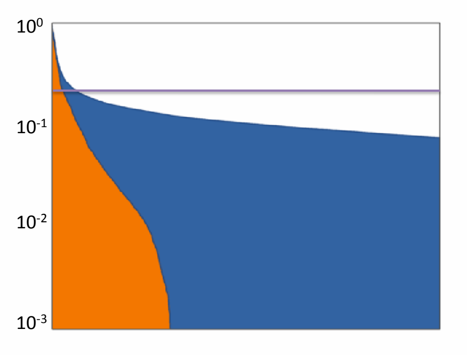


Figure 3: Normalized intensity values of the voxels reconstructed, sorted in descending order. The blue plot corresponds to the 3D Fourier transform method while the orange plot corresponds to the L1 minimization method.

The simulated diffraction data was calculated by selecting those voxels which are crossed by the Ewald sphere (fig. 2(b)) and projecting them onto the detector plane according to the geometry shown in Figure 1. Figure 2(c) shows the simulated 2D diffraction pattern. The detector plane is uniformly sampled with 64×64 pixels, and we set the distance between the crystal and the detector to 64 pixels.

We then tried to recover the real space lattice using two methods. In the first method we simply took the inverse 3D FFT of the diffraction volume in which the voxels that were not "observed" were set to zero. The intensity of the transformed volume is shown in Figure 2(e). In the second approach, we solve the L1 minimization problem (4) discussed in the previous section by using the SPGL1 software provided by the authors of [19]. The intensity of the solution x to (4) is shown in Figure 2(f).

It is clearly from Figures 2(e) and 2(e) that the latter approach results in a much sharper image from which the unit cell parameters can be easily extracted. To quantify this difference we normalized recovered real space intensity values of both methods and sorted them in decreasing order. We plotted the sorted values as 1D curves in Figure 3. The curve that separates the red and blue region of the plot is associated with the solution to the L1 minimization problem. The curve that separates the blue region and the white area above it is associated with the sorted intensities obtained from a direct 3D inverse FFT. Clearly, the intensity associated with the solution to the L1 minimization problem decreases much more more rapidly, thereby making it easy to select a threshold (shown as the magenta line in Figure 3) that can be used to identify real space lattice points.

7. Concluding Remarks

We presented a new technique for autoindexing nanocrystal diffraction images. The technique is based on formulating the indexing problem as an L1 minimization (or BP) problem and solving the problem by an efficient and robust numerical algorithm. Although the algorithm is more costly than the existing approach because it is iterative and performs multiple 3D FFTs, it has the advantage of recovering crystal lattice reliably when only a few Bragg peaks can be measured. We demonstrate the feasibility of the technique with a simple example. More studies are needed to test the efficacy of the method on different types of Bravais lattices and on datasets that may be contaminated with noise. However, we believe that our preliminary results already indicate that compressive sensing based autoindexing a promising tool for ultrafast nanocrystallography. Moreover, this type of technique allows other constraints to be easily incorporated into L1 minimization formulation to improve the reliability of indexing. It may even be possible to extent this approach to index powder diffraction data.

Acknowledgement

This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. Funding for FRNCM was provided by American Recovery and Reinvestment Act Computational Science and Engineering Petascale Initiative.

References

[1] 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, HeringPh, 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, J. Galayda, First lasing and operation of an ångstrom-wavelength free-electron laser, Nature Photonics 4 (2010) 641–647.

[2] W. Ackermann, et al., Operation of a free-electron laser from the extreme ultraviolet to the water window, Nat. Photon. 1 (2007) 336–342.

[3] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert, J. Hajdu, Potential for biomolecular imaging with femtosecond x-ray pulses, Nature 406 (2000) 752–757.

[4] H. N. Chapman, A. Barty, M. J. Bogan, S. Boutet, M. Frank, S. P. Hau-Riege, S. Marchesini, B. W. Woods, S. Bajt, H. W. Benner, R. A. London, E. Plonjes, M. Kuhlmann, R. Treusch, S. Dusterer, T. Tschentscher, J. R. Schneider, E. Spiller, T. Moller, C. Bostedt, M. Hoener, D. A. Shapiro, K. O. Hodgson, D. van der Spoel, F. Burmeister, M. Bergh, C. Caleman, G. Huldt, M. M. Seibert, F. R. Maia, R. W. Lee, A. Szoke, N. Timneanu, J. Hajdu, Femtosecond diffractive imaging with a soft-x-ray free-electron laser, Nature Physics 2 (2006) 839–843.

[5] R. A. Kirian, X. Wang, U. Weierstall, K. E. Schmidt, J. C. H. Spence, M. Hunter, P. Fromme, T. White, H. N. Chapman, J. Holton, Femtosecond protein nanocrystallography—data analysis methods, Opt. Express 18 (2010) 5713–5723.

[6] I. Steller, R. Bolotovsky, M. G. Rossmann, An algorithm for automatic indexing of oscillation images using Fourier analysis, J. Appl. Cryst. 30 (1997) 1036–1040.

[7] M. G. Rossmann, C. G. van Beek, Data processing, Acta Cryst. D 55 (1999) 1631–1640.

[8] A. G. W. Leslie, Recent changes to the MOSFLM package for processing film and image plate data., volume 26, Daresbury Laboratory, Warrington, U.K., 1992.

[9] G. Bricogne, in: Proceedings of the EEC cooperative Workshop on Position-Sensitive Detector Software (Phase III), Paris.

[10] Z. Otwinowski, W. Minor, Processing of X-ray diffraction data collected in oscillation mode, in: C. W. C. Jr., R. M. Sweets (Eds.), Methods in Enzymology, pp. 307–326.

[11] J. W. Campbell, *XDL VIEW*, an X-windows-based toolkit for crystallographic and other applications, Journal of Applied Crystallography 28 (1995) 236–242.

[12] F. Natterer, F. Wübbeling, Mathematical Methods in Image Reconstruction, SIAM, 2001.

[13] H. R. Powell, The Rossmann Fourier autoindexing algorithm in MOSFLM, Acta Cryst. D 55 (1999) 1690–1695.

[14] N. K. Sauter, R. W. Grosse-Kunstleve, P. D. Adams, Robust indexing for automatic data collection, Journal of Applied Crystallography 37 (2004) 399–409.

[15] L. Ying, Sparse Fourier transform via butterfly algorithm, SIAM J. Sci. Comput. 31 (2009) 1678–1694.

[16] D. L. Donoho, Compressive sensing, IEEE Trans. on Inf. Theory 52 (2006) 1289–1306.

[17] E. Candès, J. Romberg, T. Tao, Stable singal recovery from incomplete and inaccurate measurements, Comm. Pure Appl. Math 59 (2006) 1207–1223.

[18] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reeconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2006) 489–509.

[19] E. Van den Berg, M. P. Friedlander, Probing the Pareto frontier for basis pursuit solutions, SIAM J. Sci. Comput. 31 (2009) 890–912.

[20] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Statist. Soc. Ser. B 58 (1996) 267–288.

[21] E. G. Birgin, J. M. Martínez, M. Raydan, Nonmonotone spectral projected gradient methods on convex sets, SIAM J. Optim. 10 (2000) 1196–1211.

[22] E. G. Birgin, J. M. Martínez, M. Raydan, Inexact spectral projected gradient methods on convex sets, IMA J. Numer. Anal. 23 (2005) 539–559.

[23] Y. H. Dai, R. Fletcher, Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming, Numer. Math. 100 (2005) 21–47.

[24] Y. Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$, Dokl. Acad. Nauk SSSR 269 (1983) 543–547.

[25] Y. Nesterov, Smooth minimization of non-smooth functions, Math. Program. 103 (2005) 127–152.

[26] S. Becker, J. Bobin, E. J. Candès, NESTA: A Fast and accurate first-order method for sparse recovery, Technical Report, California Institute of Technology, 2009.