

This is the accepted manuscript made available via CHORUS. The article has been published as:

Automatic recovery of missing amplitudes and phases in tilt-limited electron crystallography of two-dimensional crystals

Bryant R. Gipson, Daniel J. Masiel, Nigel D. Browning, John Spence, Kaoru Mitsuoka, and
Henning Stahlberg

Phys. Rev. E **84**, 011916 — Published 22 July 2011

DOI: [10.1103/PhysRevE.84.011916](https://doi.org/10.1103/PhysRevE.84.011916)

Classification:

Soft Matter, Biological, and Interdisciplinary Physics

Title:

Automatic Recovery of Missing Amplitudes and Phases in Tilt-Limited Electron Crystallography of 2D Crystals

Authors:

Bryant R. Gipson^{1,5,*}, Daniel J. Masiel², Nigel D. Browning², John Spence³,
Kaoru Mitsuoka⁴, and Henning Stahlberg^{1,*}

Author affiliation:

¹ Center for Cellular Imaging and Nano Analytics (C-CINA), Biozentrum, University Basel, WRO-1058 Mattenstrasse 26, CH-4058 Basel, Switzerland.

² Dep. Chemical Engineering and Materials Sciences, University of California at Davis, Davis, CA 95616, USA.

³ Department of Physics, Arizona State University, Tempe, AZ 85287, USA.

⁴ Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan.

⁵ present address: Department of Computer Sciences, Duncan Hall, Rice University, Houston, TX 77030, USA

* Corresponding authors: E-mails:

Bryant.Gipson@rice.edu, Henning.Stahlberg@unibas.ch

Correspondence:

Henning Stahlberg

Center for Cellular Imaging and Nano Analytics (C-CINA)

Biozentrum, University Basel

at the Department of Biosystems Science and Engineering (D-BSSE)

Mattenstrasse 26, WRO-1058

CH-4058 Basel, Switzerland

Tel: +41-61-387 32 62

Fax: +41-61-387 39 86

mailto:Henning.Stahlberg@unibas.ch

Abstract:

Electron Crystallography of 2D protein crystals provides a powerful tool for the determination of membrane protein structure. In this method, data is acquired in the Fourier domain as randomly sampled, uncoupled, amplitudes and phases. Due to physical constraints on specimen tilting, those Fourier data show a vast un-sampled “missing cone” of information, producing resolution loss in the direction perpendicular to the membrane plane. Based on the flexible language of projection onto sets, we provide a full solution for these problems with a projective constraint optimization algorithm that, for sufficiently oversampled data, produces complete recovery of unmeasured data in the missing cone. We apply this method to an experimental dataset of Bacteriorhodopsin and show that, in addition to producing superior results compared to traditional reconstruction methods, full, reproducible, recovery of the missing cone from noisy data is possible. Finally, we present an automatic implementation of the refinement routine as open source, freely distributed, software that will be included in our 2dx software package.

I. Introduction

Electron crystallography of two-dimensional membrane protein crystals records transmission electron microscopy (TEM) images of thin sheet-like crystals, which are oriented in the microscope horizontally (i.e. non-tilted), or at a certain tilt angle of typically up to 60° . The Fourier transformation of such images contains information about the amplitudes and phases of the diffraction orders, which are modulated by the instrument's contrast transfer function (CTF). Alternatively, electron diffraction patterns can be recorded with the TEM, which only yield information about the amplitudes of the structure factors. Image processing conventionally tries to evaluate and merge the amplitude and phase information from several recorded images and diffraction patterns, yielding a dataset in the Fourier domain that is void of measurements in a so-called missing cone region around the Z-axis, due to the limited tilting possibilities of the samples in the microscope. This "missing cone" problem leads to 3D reconstructions of the protein that may have a good resolution in the X/Y membrane plane, but usually have a much worse resolution in the Z-direction perpendicular to the membrane plane. Horizontal protein segments such as surface loops or horizontally arranged amino acid side chains are then strongly smeared out vertically, and are therefore often not resolved in electron crystallography structures. Here we present a new algorithm to reconstruct the data in the missing cone.

The central problem facing electron crystallography of two-dimensional membrane protein crystals can also be formulated with the question: "What protein structure best generates the data observed, given what is known?" Such data populate a continuous 3D Fourier domain as independent amplitudes (from electron diffraction patterns), independent phases (from real-space images, when amplitudes are taken from diffraction patterns), or fully coupled amplitude and phase terms (from real-space images, when diffraction patterns are not available). The terms are irregularly sampled in the Z-direction perpendicular to the membrane and, in TEM, suffer from changes in the amplitude profile due to uncertainties in the determination of the microscope's CTF. Uncorrelated noise also introduces variation of certainties in data quality.

Iterative enforcement of known constraints allows the above question to be answered in full, using the measured data as well as additional prior knowledge. We present here both simulations and an experimental application of projection onto non-convex sets using a pseudo-inverse mapping between regularized real space and non-uniform Fourier space. We show that this mapping, combined with the application of Lions-Mercier [1] style projections, permits the native use of uncoupled amplitudes and phases, in addition to allowing a complete recovery of the missing cone.

The algorithm is completely general and, without modification, should be applicable to any constrained tomographic problem including, for example, 2D electron crystallography, single particle protein analysis by electron microscopy, thick-specimen (electron-)tomography and seismic tomography.

II. Algorithm

A. Projection onto Convex and Non-Convex Sets

Presenting a solution for the missing cone problem is not trivial. There are an infinite number of unconstrained real-space solutions that could generate any given incomplete 3D Fourier dataset (i.e. those cropped by a missing cone), and the situation is made worse in the presence of noisy data. Application of reasonable constraints to the real or Fourier domain, however, can reduce the size or dimension of the search space of solutions dramatically. Introducing this a-priori information about a protein under study in an explicit, non-iterative, algebraic manner can be difficult, particularly in the case where constraints are only loosely known, and data are noisy.

Projection onto convex sets is a robust and rigorous theory of optimization, which is applied using a straightforward alternating iterative application of known convex constraints to a dataset. In each iteration the dataset is sequentially modified to represent the closest solution that satisfies a given constraint. If these constraints represent convex sets, such that one uniquely defined closest solution (projection) can be found for each step, iterative application will move the dataset towards a guaranteed optimal solution.

Closely related is the study of projections onto non-convex sets, which governs constraint sets with ambiguous projections. An example of such a non-convex set would be that defined by the points on the circumference of a circle. Here the point at the center of the circle for example has an undefined projection, as all points in the set are equidistant from it. In spite of this, extensions to known convex projection algorithms (in particular Lions-Mercier [1]) have proven convergence properties (subject to satisfiability) in theory [2, [3], numerical simulation [4], and experiment [5].

The language of projection operators is well suited to most constraints and is often “human readable”. In the case of electron crystallography, constraints exist in the real and Fourier domains, which can be translated into projection operators. For example, a 2D membrane protein crystal has no densities above and below the membrane, implying that the real space volume of this membrane bound structure has to be strictly contained within a slab of finite thickness. This knowledge can then be enforced as a constraint, specifying that the real-space volume is non-zero only within a known horizontal slab region. The associated projection is unique for all structures (and therefore convex) as it simply sets to zero all those regions that are known to be empty.

In the Fourier domain, the constraint of known amplitudes or phases (namely the collected data) at particular points has an equally simple projection operator: the enforcement of the appropriate phase or amplitude at the point in question. Changing an existing phase value in the current dataset to an experimentally known (measured) phase represents a unique, convex constraint. The case of changing a zero amplitude to a non-zero amplitude, however, presents an ambiguous projection as it is undefined which phase should be assigned to the new amplitude. As a result, the amplitude constraint is non-convex.

The method of projection onto convex sets as answer to the missing-cone problem was already investigated by Agard and Stroud [6, [7], and by Barth *et al.* [8, [9], who combined this with a maximum entropy approach. Here we present an application of

projection onto convex sets, in combination with a linear algebra solution for the transformation between real space and non-regularly sampled Fourier space, fractional steps with individual point-wise weighting of applied constraints, and a shrinkwrap boundary constraint.

B. Truncated singular value decomposition

In the presented method we alternate between real and Fourier space, applying constraints in each space. Alternating projections between sets requires an invertible mapping between projection spaces. In the case of a simulated dataset, the values in Fourier space can be placed at regular sampling positions, and a Discrete Fourier Transform (DFT) suffices as a one-to-one mapping between the real and Fourier domains. In the case of experimentally obtained electron crystallography data, however, due to the typically random orientation of crystals on tilted support films in the electron microscope, the available data are usually distributed in Fourier space at an arbitrary location in the vertical direction along the Fourier lattice lines, resulting in a 3D Fourier dataset of irregular sampling. Such a dataset then requires either regularization by (re-)interpolation, or an operator that takes this non-uniform sampling into account. Traditional methods [10, [11] have used SINC interpolation as an initial regularizing step of the lattice line data, before using the DFT as the mapping. We present here a direct method that, instead of SINC interpolation, uses a non-uniform transform from a regularly sampled real space volume to an irregularly sampled Fourier space, which then provides a method to find the best unbiased real linear estimator for a given Fourier dataset. The advantage of this method is that repeated interpolation errors are avoided and the raw data is used directly as a constraint.

While sampling irregular points along a sub-Nyquist interval may seem to create a problem, it is precisely this extra information gained by additional sampling that can be leveraged to determine un-sampled data. Historically, the corresponding "oversampling" along the Fourier domain rods normal to the slab have been found to greatly reduce the number of high-resolution cryo-EM images needed, at the experimentally difficult high tilt angles, to phase a dataset [12]. Here we present a method to use such oversampled data to fully reconstruct both missing amplitudes and phases.

The relationship between a discretely sampled real-space 2D-crystal structure and an irregular, finite sampling of the continuous Fourier domain is a linear one,

$$\mathfrak{S}_{r,s}x_r = \hat{x}_s \quad (1)$$

where $\mathfrak{S}_{r,s}$ is the discrete to s -sampled Fourier transform, x_r is an unknown regularly spaced real space solution, and \hat{x}_s is the Fourier domain data sampled at s . In most experimental cases, depending on the number and distribution of the data samples, this represents a severely ill-conditioned matrix. A standard tactic in ill-conditioned inverse problems is to use a pseudo-inverse matrix calculated by truncated singular value decomposition [13] (SVD), which provides a stabilized least squares solution to the problem.

$$x_r = \mathfrak{S}_{r,s}^+ \hat{x}_s \quad (2)$$

Applying constraints in the Fourier or real domain may then be performed directly on the data, with numerical accuracy dependent only on the quality of the data and the number of assumed discrete real samples [14].

As zero terms in the real domain (i.e. those values outside of the 2D-crystal slab) do not contribute to the Fourier sum, they may be removed, resulting in a $m \times n$ transform with an implicit support constraint. In cases where this is the only real-space constraint, the above may be re-written as

$$\mathfrak{I}_{r,s} \mathfrak{I}_{r,s}^+ \hat{x}_s = \Gamma \hat{x}_s = \hat{x}_s \quad (3)$$

or that all Fourier domain solutions lie in the Eigenvector space of Γ , describing all finite slabs that generate the observed data.

Both the sensitivity of Γ to noise and the accuracy of solutions depend heavily on the choice of the regularization parameter ν in the SVD truncation. High condition value matrices are extremely sensitive to noise, while low condition numbers yield a less accurate reconstruction. The reliance on ν can be lessened by calculating an estimated error for Γ :

$$W = \text{diag}^+(abs(\Gamma)\epsilon J) \quad (4)$$

Here, J is the N dimensional vector of all ones, ϵ is the estimated error level, $\text{diag}^+(v)$ indicates the diagonal matrix formed from vector v , where the non-zero elements have been inverted, and $abs(\Gamma)$ is the element-wise absolute value of the matrix Γ . This weighting function can additionally be normalized, combined with other known confidence values, or otherwise altered to fit the purposes of a given dataset. While this error function could in principle be used as a least squares weighting function

$$\Gamma_W = \mathfrak{I}_{r,s} (W \mathfrak{I}_{r,s})^+ W \quad (5)$$

it will later be shown in Eq. 11 that this is not optimal for our purposes.

In the most general case of electron crystallography of 2D crystals, Fourier amplitudes (A) and phases ($e^{i\Theta}$) are collected as uncoupled values

$$\begin{pmatrix} A \\ * \end{pmatrix}, \begin{pmatrix} * \\ e^{i\Theta} \end{pmatrix} \quad (6)$$

which form the basis for the Fourier constraints. A projection operator P_3 (Eq. 8) enforcing the constraints in Eq. 6 can be applied to an arbitrary image in the form of Eq. 7

$$v = \left[\begin{pmatrix} \alpha \\ \Lambda \end{pmatrix} \circ \begin{pmatrix} e^{i\Phi} \\ e^{i\Theta} \end{pmatrix} \right] \quad (7)$$

$$P_3(v) = \left[\begin{pmatrix} A \\ * \end{pmatrix} \circ \begin{pmatrix} e^{i\Phi} \\ e^{i\Theta} \end{pmatrix} \right] \quad (8)$$

where α and e^{it} are discarded and replaced by the known amplitudes (A) and phases ($e^{i\Theta}$). Here, \circ represents the Hadamard (element-wise) product of two vectors or matrices. Using Eq. 3 and Eq. 8, iterative application of these Fourier constraints can be performed.

$$\Gamma P_3(v_n) = v_{n+1} \quad (9)$$

C. Iterations in fractional time steps

Due to the non-convex nature of the system, however, strictly enforcing these constraints as in Eq. 9 can lead to pseudo-solutions and exponentially slow convergence (stagnation) [3]. These problems can be alleviated using fractional time-step parameters as used in Lions-Mercier [1], Fienup's Hybrid Input Output [4], and the Douglas-Rachford Algorithm [15].

$$(1 - \beta) v_n + \beta \Gamma P_3(v_n) = v_{n+1} \quad | \quad 0 \leq \beta \leq 1 \quad (10)$$

While such fractional parameters have historically been globally enforced as in Lions and Mercier (1979) [1], we have used the weights of Eq. 4 in a point-wise fashion independently operating, subject to known confidence levels W , on each value.

$$(1 - W) v_n + W \Gamma P_3(v_n) = v_{n+1} \quad (11)$$

This has the simultaneous benefit of introducing a per-point fractional time step for convergence purposes, as well as allowing values that are more certain to guide the convergence process as a whole.

Combining the concepts of Eq. 2 and Eq. 11, we then form a completely general optimization routine using both real and Fourier constraints. The data in real space (x) and Fourier space (v) are iteratively obtained by applying Eq. 12-14.

$$x_n = \mathfrak{S}_{r,s}^+(v_n) \quad (12)$$

$$\mathfrak{S}_{r,s}((1 - \beta) x_n + \beta P_{\mathfrak{R}}(x_n)) = v_{n+1} \quad (13)$$

$$(1 - W) v_{n+1} + W P_3(v_{n+1}) = v_{n+2} \quad (14)$$

Here the final $P_{\mathfrak{R}}$ and P_3 represent all real and Fourier space constraints, respectively, that are appropriate for the system. For example, $P_{\mathfrak{R}}$, the real-space constraint operator, might include (but by no means be limited to) real-valuedness, non-negativity and additional support information, as well as more vague concepts such as appropriately defined regional continuity or non-diffuse (i.e. sharp) boundaries between zero and non-zero regions. Equally, P_3 , the Fourier space constraint operator, might include symmetry and spectral profile (e.g. from a small-angle X-Ray scattering experiment, or from general spectral expectations about the amplitude for the dataset), as well as optical constraints such as that from a known point-spread function. The presence of a global real-space weight β in Eq. 13 is necessary here, as point-wise confidence about the real-

domain is typically unknown. Nevertheless, Eq. 13 could obviously be modified according to Eq. 14, if such were available.

D. Shrinkwrap optimization

Shrinkwrap [16], a final added optimization that has recently gained increased theoretical support [17], can be implemented by setting to zero all real-space values that are found below an assumed noise threshold, as an iterative extension to the known support.

$$x_n(i) = \begin{cases} x_n(i) & ; x_n(i) > \varepsilon \\ 0 & ; x_n(i) \leq \varepsilon \end{cases} \quad (15)$$

Knowledge that the 2D-crystal exists as a slab of finite thickness provides an initial estimate for the support of the object; empty regions found within this slab are then iteratively refined and can be used in the above procedure by appropriately modifying the real space projection operator $P_{\mathfrak{R}}$.

We refer to the above-described algorithm, with or without Shrinkwrap, as Projective Constraint Optimization (PCO) for the remainder of the text.

E. Cylindrical Ring Correlation (CRC) as convergence measure

In addition to using standard resolution measures for analysis, such as Fourier Shell correlation [18], we developed a cylindrical ring correlation measure (CRC), which calculates the Fourier Ring Correlation (normalized dot-product for corresponding points on a ring of Fourier pixels [33]) for each horizontal 2D slab of 3D Fourier space between the current dataset and a reference dataset. Alternatively, the CRC can be calculated between the current dataset, and its symmetry-rotated copy, if the membrane protein has an internal symmetry that was not exploited in the processing. This CRC measure produces a 2D-map that describes resolution in ρ, z space. The ring correlation value is a function of radius ρ and height z of a ring of constant thickness (e.g. 1 Fourier pixel) in Fourier space. Using this measure it is possible to measure point-wise ring-correlation for all points inside and outside of the missing cone, and observe the convergence of the algorithm in real time [33].

III. Application

A. Application to underdetermined simulated experimental data

We applied Eq. 12-14 with Eq. 15 to a series of test cases based on computed datasets of actual 3D models of biological proteins (described in [33]). For all these noise-free datasets we found, in agreement with previous results [12], a near complete convergence (greater than 99% Fourier shell correlation out to Nyquist). The CRC for simulations showed that the convergence process is being spatially localized in Fourier space, growing from known constraint regions outward into unknown regions in Fourier space, such as the missing cone [33]. During this process, a starting point, x_0 , for

undetermined Fourier pixel values has to be chosen at the beginning of the iterative processing. In terms of time to convergence, setting x_0 equal to the minimum energy solution to Eq. 1 (or to zero, a closely related solution path) consistently out-performed guesses with random initial amplitudes and phases for the entire Fourier domain (supplemental movie 3) and also outperformed hybrid-guesses, where only unknown values in the Fourier domain were set to random values.

Random starting points will typically partially satisfy constraints in both the real and Fourier domain. These semi-consistent solutions produce densities in the Fourier domain that persist through iterations, and can guide the refinement of neighboring densities into wrong pseudo-solutions. This appears to result in slowly growing "solution fronts" in Fourier space, both from pseudo-solution densities and constraint densities, which eventually collide and slowly attempt to reconcile. Starting with the unknown densities set to zero instead insures that a single, self-consistent, solution front is allowed to propagate through "resistance free" zero-valued solution space. This was found to be a rapid process. In the absence of noise however, all starting positions converged to essentially complete correlation, differing only by required iterations –typically ~ 30 for a zero starting point to more than 1000 rounds for a random guess. Starting with zeros in the missing cone, however, led to a dilution of the "energy" in the dataset during the iterations, so that the amplitude values from the known regions "bled" into the formerly missing cone region. Multiplication of the amplitudes of the entire dataset by a constant factor (e.g. 2) once, after a certain number of iterations, was found helpful to scale the intensities back to the original experimental values.

B. Application to an experimental dataset of Bacteriorhodopsin

As also done with a limilar approach by Agard and Stroud [7], the routine was applied to an experimental data-set of Bacteriorhodopsin (BR) [19]. This dataset was used in 1997 to generate a 3D reconstruction at 3.5Å resolution and contains both electron diffraction intensities and approximate phases derived from real-domain transmission electron microscope projection images. Data were available from -60° to $+60^\circ$ sample tilt angle, leaving an empty, vertically aligned "missing cone" of 60° angular diameter. These data were the basis of our refinement processing with the projective constraint optimization (the full details of processing are available online [33]). For comparison we used an atomic model of Bacteriorhodopsin 1BRR [20] and 1C3W [21], both of which stem from 3D crystals grown of BR in the presence of lipids, and X-ray diffraction structure determination at 2.9Å and 1.55Å resolution respectively. These models were only used for comparative quality analysis of the results, and were not used during the processing at any time.

The above-described algorithm was applied to this raw electron crystallography dataset, running 24 rounds with conventional constraints (without Shrinkwrap, for example see supplemental movies 8-9 at [33]), followed by 122 rounds of iterations including the Shrinkwrap constraint (supplemental movies 4-7 at [33]) and with the inclusion of a sharp-boundary constraint [33], until convergence was reached. The raw measured amplitudes and phases from the images were directly used as the input of this algorithm.

Ideally, the above algorithm would be applied between 3D domains of regular real space and irregularly sampled Fourier space. Due to the large size of the dataset, however, we divided the algorithm into two steps: A first, one-dimensional PCO (1D-PCO) refinement was performed on individual lattice line data from regular real space to irregular Fourier space, strictly in the sampled regions and restricted to one lattice line at a time. This was followed by a second, three-dimensional PCO (3D-PCO) refinement, operating between the full (sampled plus unsampled) regular 3D real space to regular 3D Fourier space (see supplemental online figure 1). These steps were identical in form, following Eq. 12-14 iteratively, differing only in the applied constraints (i.e. 1D-PCO could not assume real-valuedness) and number of iterations (1D-PCO was performed for 20000 iterations vs the ~ 150 rounds for 3D-PCO). This was strictly due to computational limits on storage (i.e. a maximum size for the matrices of $\sim 200,000 \times 200,000$ elements) and the fact the data was processed serially on a single machine. Future versions of this algorithm plan to incorporate large parallelized SVD computation, as in Hernandez *et al.* (2005) [22], which is expected to additionally improve results. Though initial 1D-PCO operates on lattice lines, as does MRC image processing of 2D crystal data [10], it is important to note that PCO is algorithmically distinct from this method and can additionally be generally applied in any number of dimensions.

Maximum structure correlation [23] of better than 0.82 of the obtained refined map relative to the atomic model 1BRR [20] and better than 0.80 when compared to 1C3W [21] was achieved when filtered to 2.5\AA (Table 1).

A more complete picture of angle dependent resolution is available from the full CRC (Fig. 1), which shows correlation out to beyond the data resolution limit. Comparison with the CRC in the region of the missing cone also shows significant correlation and a generally isotropic distribution of resolution, with the exception of characteristic semi-regular absences in correlation in the vertical direction. While these absences could be attributed to the low number of pixels at radii close to the center of the cylinder (leading to less meaningful correlation factors) or the tightness of the 2D slab mask, it may also be a result of "striation" errors typical [24] of iterated projection solutions. It is also possible that high-frequency "ringing" effects, resulting from discontinuities at the edges of the 2D slab, introduce such artifacts – a problem a carefully chosen real-space "soft" Gaussian edge constraint might alleviate.

Real-space correlation values (Table 1) show significant improvement for the here described refinement algorithm out to 2.5\AA . Already a comparison of the PCO-refined result strictly within the experimentally sampled region (i.e. outside of the missing cone) shows improvement out to 3.0\AA . Interestingly, however, the calculated R-Factors appear to get worse and increase commensurately with the amount of purely refined (i.e. unobserved) data in the case of full 3D-PCO. A careful reading of the CRC explains this apparent contradiction.

As iterative application of real-space boundaries equate to a series of convolutions on Fourier space, unmeasured amplitudes derive their energy iteratively from neighboring constrained or previously refined regions. Without additional energetic Fourier constraints or consideration [31], refined amplitudes remain down-weighted relative to their distance to constraint densities and thus yield higher R-factors. This leads to significant non-linear (yet smoothly predictable) scaling effects as distance from

constraint regions increase. Thus phases, which have a significant impact on real-space structure and therefore structural correlation, are quickly refined to near accurate values, while amplitudes require more iterations to come to equilibrium as inaccuracies here violate real-space constraints far less. This results in high structural correlation values (and real-space quality) with increasingly worse R-factors, as the latter rely on correct amplitude scaling that may at most differ by the non-linear (exponential) effect of a wrong B-factor.

To test this we performed analysis on the results that had only been processed by 1D-PCO and had therefore not received missing cone refinement. It was found in this case that R-factors were dramatically better, reaching their best values when the dataset was actually limited to 50° (Table 1); an effect likely due to the unreliability of measured amplitudes, phases, and error estimates at high tilt ranges as well as sparsity of sampling in these regions (see supplemental materials for details). As can be seen, in this case R-factors are considerably better than MRC refined values, while the structural correlation was correspondingly worse, the opposite of the case of full 3D-PCO refinement. Fortunately, these refinements are not mutually exclusive and this issue could in principle be solved with the careful tuning of refinement parameters, the enforcement of local Fourier energy constraints, such as constant radial averages, or the application of known amplitude profiles, such as those derived from low-angle scattering experiments. Generally, as this method simultaneously refines both amplitude and phase, here real-space structure correlation becomes the more reliable indicator of map quality.

Visual inspection of the map (Fig. 2) shows that the refinement procedure produces a reconstruction with a well-resolved helical pitch and side chain densities that were not as clearly seen after application of traditional lattice line interpolation methods [32]. In particular, horizontal features such as extracellular loops, most dependent on vertical resolution, are in our refined reconstruction clearly visible at a single map-wide iso-surface threshold.

C. Evaluation of the tolerance of limited tilt-range

We evaluated the robustness of the PCO reconstruction with respect to the size of the missing cone (Fig. 4). The BR dataset was further limited in tilt, so that copies of the dataset with tilt-ranges between $\pm 1^\circ$ to $\pm 60^\circ$ were created. Full 3D-PCO reconstructions were performed on these datasets and the average 3D resolution of the final maps was estimated by Fourier Shell Correlation (FSC) with the 0.5 threshold with the 1BRR map. The preliminary 1D-PCO refined reconstructions (see supplemental online materials) were compared to measure the benefit of 3D constraints. The resulting resolutions are shown in Fig. 4: The original MRC-based 3D reconstruction from this dataset showed an FSC-0.5 resolution of 4.4\AA . PCO refined results from data from the low tilt-range between $\pm 10^\circ$ produced a map at 9\AA resolution, while data from the tilt-range of $\pm 35^\circ$ were found sufficient for PCO to reconstruct the almost complete map with an average FSC estimated resolution of 4.5\AA . Inclusion of higher tilted data between 40° and 60° (Fig. 3) for PCO refinement did not significantly improve the isotropic resolution of the final map. This means that with the PCO algorithm available, data collection in the tilt-range between 40° and 60° would not have been required for this dataset.

The best statistics for the final reconstruction were obtained after removing the top 10° tilt range from the bacteriorhodopsin dataset, i.e. by not using the data with a specimen tilt of 50° or higher. Even though in pure simulations, including higher tilt data always produced faster convergence rates for solutions, for an experimental dataset, the high-tilt data may present both larger uncertainties as well as lower sampling density. In this specific case, the given phase incoherence, noise level, and the distribution of sampled data in different tilt ranges, showed the highest errors and lowest sampling density at the highest tilts. By removing the top 10° tilt range, the resolution of the final reconstruction was improved, highlighting algorithmic sensitivity both to data certainty and, importantly, sampling density. A better estimation of error levels might improve this situation. In the current situation, however, a sufficiently dense sampling is highly influential in balancing the effects of noise (see Fig. S3 and S4 at [33]).

IV. Discussion

We have presented an algorithm for the PCO refinement of electron crystallography data. This was applied to several simulated (noise-free) and an experimental (noisy) dataset of Bacteriorhodopsin, and resulted in the reconstruction of data inside the missing cone region. Noise was found to typically lead to an increased density of local minima and stagnation of the iterative refinement procedure [5]. Eq. 2 offers a solution to this problem as data are collected in the continuously oversampled Fourier domain, causing neighboring values (relative to real-space support) to correlate. In this way, global noise contribution can be controlled by the addition of more data. The amount of data necessary and the chances for achieving a unique solution were seen to be most sensitive to the ability to estimate error. When error was not considered, the algorithm effectively proceeded as assuming an error of zero, which typically corresponded to the enforcement of conflicting constraints, which, as can also easily be shown through simulations, produced radically diverging densities within a few tens of rounds. Even only approximate estimations of error, however, applied as in Eq. 11 in the form of individual data-point weights, were found to stabilize convergence.

Over-fitting was found to be a problem in cases where Fourier reconstruction was attempted well beyond the limits of sampled resolution (e.g. out to 1 Å). As both phases and amplitudes are unbounded in the refinement regime, spurious high frequency Fourier terms began to dominate the reconstruction, likely counter-balancing known Fourier constraints in an attempt to satisfy both Fourier-domain noise and the real-space support constraint. This appeared in real space as either high-valued, apparently randomly scattered, point densities or 1-pixel wide slabs touching the boundaries of the allowed support regions. When a real-space sharp region-boundary constraint was applied, however, all above-mentioned over-fitting effects were almost completely absent (see supplementary online materials for description, and supplementary movies 4-7 for real-time application at [33]).

V. Conclusion

Iterative application of known projective constraints presents a dynamic and completely general framework for the refinement of incomplete data, which automatically reduces solution search space sizes for a wide range of problems. The procedure defined here has shown that even in experimental cases where irregular sampling and variable levels of noise are present, complete structure recovery is possible. The presented refined bacteriorhodopsin dataset shows the power of this algorithm for general-purpose electron crystallography of 2D crystals.

Additionally, we show that an increase in the amount of data can be leveraged into an increase in resolution. As recent results [25] demonstrate, given sufficiently oversampled data, almost any tilt-range suffices as the basis for full 3D reconstruction. The PCO algorithm presented here allows a reduction in tilt-range requirements, if a sufficiently large dataset can be collected, e.g. through automation of electron crystallography data collection. In the utilized bacteriorhodopsin data set, inclusion of data up to a tilt range of only 35° allowed the almost full reconstruction of the 3D structure, implying that data collection in the tilt range of 40° and higher could have been avoided. Since lower tilt angle electron crystallography data can be collected with much higher success rates and usually contain higher resolution data, the presented algorithm not only allows pushing the resolution limit in the direction perpendicular to the membrane plane, but also offers significant improvements for speed and overall resolution of the structure determination of membrane proteins by electron crystallography.

This software is released open source under the GPL. It currently exists as a stand-alone C++ program and individual Octave [26] modules, and has been optimized for a CUDA ready GPU, if present. The stand-alone software will be incorporated into future versions of 2dx [27], and will be available online [28]. The PCO refined map is available at the EMDB under accession code EMD-1856.

VI. Acknowledgements

The authors gratefully acknowledge Y. Fujiyoshi, Kyoto University, for the providing the Bacteriorhodopsin dataset. We thank T. Schirmer, University Basel, and T. Strohmer, UC Davis, for fruitful discussions. This work was in part supported by SNF grant 205321_12649, the Swiss initiative for Systems Biology (SystemsX.ch: CINA), and the NSF grant MCB-0447860, NIH grant U54-GM074929, DOE award DE-FG03-02ER45996, and DOE grant DE-FG52-06NA26213.

References

- [1] P. Lions and B. Mercier, *SIAM J. Numer. Anal.* **16**, 964 (1979).
- [2] A. Levi and H. Stark, *J. Opt. Soc. America A* **1**, 932 (1984).
- [3] H. Bauschke, P. Combettes, and D. Luke, in *IEEE International Conference on Image Processing*, Rochester, NY, 2002), Vol. II, p. 841.
- [4] J. Fienup, *Journal of the Optical Society of America A* **4**, 118 (1987).
- [5] J. Spence, *Science of Microscopy* **2**, 1196 (2007).
- [6] R. M. Stroud and D. A. Agard, *Biophys J* **25**, 495 (1979).
- [7] D. A. Agard and R. M. Stroud, *Biophysical Journal* **37**, 589 (1982).
- [8] M. Barth, R. K. Bryan, R. Hegerl, and W. Baumeister, *Scanning Microsc Suppl* **2**, 277 (1988).
- [9] M. Barth, R. K. Rryan, and R. Hegerl, *Ultramic.* **31**, 365 (1989).
- [10] D. A. Agard, *J Mol Biol* **167**, 849 (1983).
- [11] R. A. Crowther, R. Henderson, and J. M. Smith, *J Struct Biol* **116**, 9 (1996).
- [12] J. C. Spence, U. Weierstall, T. T. Fricke, R. M. Glaeser, and K. H. Downing, *J Struct Biol* **144**, 209 (2003).
- [13] P. C. Hansen, *BIT* **27**, 534 (1987).
- [14] A. D. Schenk, D. Castano-Diez, B. Gipson, M. Arheit, X. Zeng, and H. Stahlberg, *Methods Enzymol* **482**, 101 (2010).
- [15] J. Douglas Jr and H. Rachford Jr, *Transactions of the American Mathematical Society* **82**, 421 (1956).
- [16] S. Marchesini, H. N. Chapman, S. P. Hau-Riege, A. Noy, M. R. Howells, U. Weierstall, and J. C. H. Spence, *Phys. Rev. B* **68**, 140101 (2003).
- [17] Y. Jin, Y.-H. Kim, and B. D. Rao, *arXiv:1003.0888v1 cs.IT* (2010).
- [18] M. van Heel, *Ultramic.* **21**, 95 (1987).
- [19] Y. Kimura, D. G. Vassilyev, A. Miyazawa, A. Kidera, M. Matsushima, K. Mitsuoaka, K. Murata, T. Hirai, and Y. Fujiyoshi, *Nature* **389**, 206 (1997).
- [20] L. Essen, R. Siebert, W. D. Lehmann, and D. Oesterhelt, *Proc Natl Acad Sci U S A* **95**, 11673 (1998).
- [21] H. Lücke, B. Schobert, H. T. Richter, J. P. Cartailier, and J. K. Lanyi, *J. Mol. Biol.* **291**, 899 (1999).
- [22] V. Hernandez, J. E. Roman, and V. Vidal, *ACM Transactions on Mathematical Software* **31**, 351 (2005).
- [23] A. A. Vaguine, J. Richelle, and S. J. Wodak, *Acta Crystallogr D Biol Crystallogr* **55**, 191 (1999).
- [24] J. Spence, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **360**, 875 (2002).
- [25] K. S. Raines, S. Salha, R. L. Sandberg, H. Jiang, J. A. Rodríguez, B. P. Fahimian, H. C. Kapteyn, J. Du, and J. Miao, *Nature* **463**, 214 (2010).
- [26] J. W. Eaton, *GNU Octave Manual* (Network Theory Limited, 2002).
- [27] B. Gipson, X. Zeng, Z. Zhang, and H. Stahlberg, *Journal of Structural Biology* **157**, 64 (2007).
- [28] <http://2dx.org>
- [29] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, *Acta Crystallogr D Biol Crystallogr* **53**, 240 (1997).

- [30] P. Emsley and K. Cowtan, Acta Crystallogr D Biol Crystallogr **60**, 2126 (2004).
- [31] The one-time scaling of all purely refined (i.e. not experimentally derived) values by a multiplicative factor (e.g. 1.5) after several rounds of iteration was found to improve refined amplitudes by the end of processing, though appeared to have little effect on the CRC or real space result.
- [32] All unrefined results (lattice line fit, or MRC) refer to only the MRC processed dataset used prior to CNS/CCP4 refinement in Kimura et al. 1997. No non-MRC refinement has been performed on the comparisson dataset.
- [33] Supplemental material is available online at [*URL inserted by publisher*]. This contains detailed information on the developed algorithm, the application of the algorithm to the processed Bacteriorhodopsin dataset, and additional data showing the resolution performance of PCO in dependence of the signal-to-noise level, and the amount of sampling. The Online Material also contains movies, illustrating the performance of PCO on simplated and experimental Bacteriorhodopsin data.

Figures & Tables

	<i>MRC (Outside of Cone)</i>		<i>3D-PCO (Total Volume)</i>			<i>3D-PCO (Outside of Cone)</i>		<i>1D-PCO (Outside of Cone)</i>	
Resolution limit	4.00	2.98	4.00	2.98	2.50	4.00	2.98	4.00	2.98
Number of used reflections	3171	6512	3531	8339	11891	2877	5192	5643	13108
% observed	89.8	78.1	100	100	82.8	81.1	61.0	83.8	82.5
Overall R factor (1C3W)	0.324	0.344	0.348	0.463	0.503	0.348	0.398	0.321	0.344
Overall R factor (1BRR)	0.305	0.335	0.351	0.456	0.490	0.325	0.388	0.287	0.329
Structure correlation (1C3W)	0.651	0.691	0.646	0.755	0.807	0.646	0.715	0.673	0.644
Structure correlation (1BRR)	0.677	0.711	0.670	0.769	0.820	0.665	0.731	0.690	0.648

Table 1: Amplitude Comparison - refmac5 [29] generated table describing comparison of atomic structure (1BRR and 1C3W) with the unrefined MRC processed data that contain a missing cone region, Projective Constraint Optimization (PCO) refinement data, and refinement result data with Fourier pixels from within the formerly missing cone region excluded. The last column describes 1D-PCO refinement of the 50° tilt-limited dataset. sfcheck [23] was used to produce the real-space structure correlation coefficient values (last row). Full refinement details available in [33].

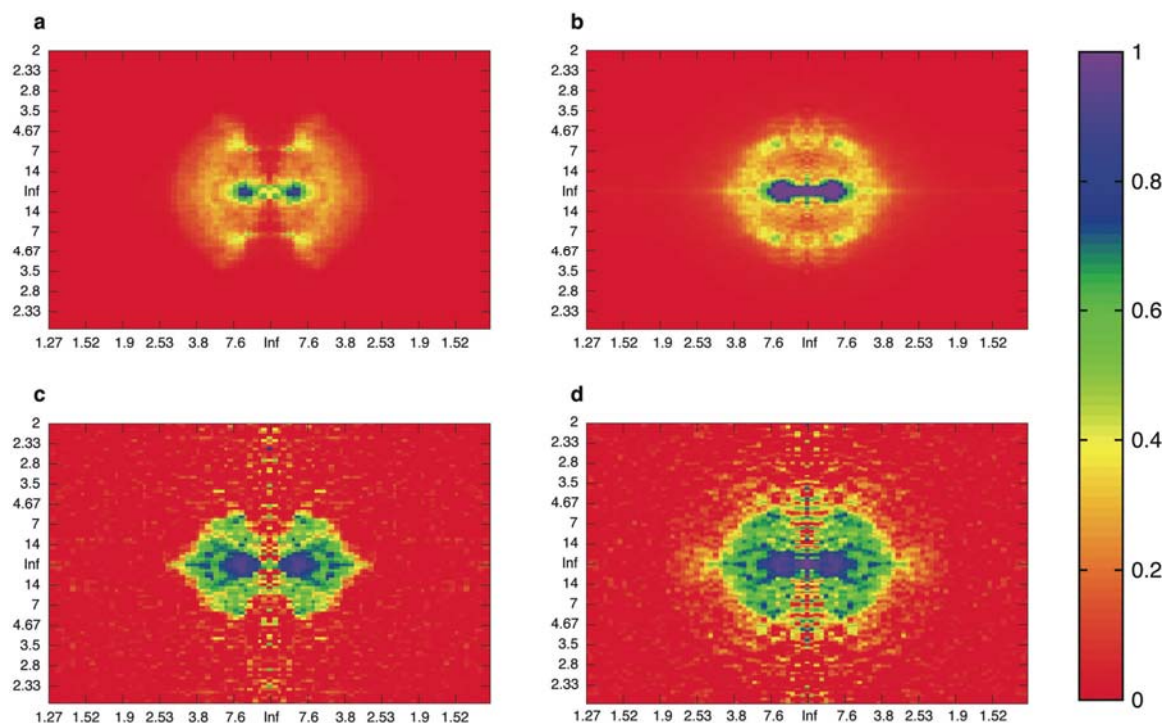


Figure 1: (Color) CRC Comparison - Normalized Intensity profiles (top: a, b) and Cylindrical Ring Correlation (CRC) plots (bottom: c, d) calculated relative to an atomic Bacteriorhodopsin crystallographic model (1BRR), without (left: a, c) and with (right: b, d) refinement by the here described PCO refinement algorithm. Plots are given as a function of ring of radius ρ (horizontal axis, in Å), and height z (in the vertical axis, in Å) in Fourier space. Initial rigid body fitting performed for CRC comparison (see supplemental online materials) were performed by the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081). Figure 2 was generated by COOT [30].

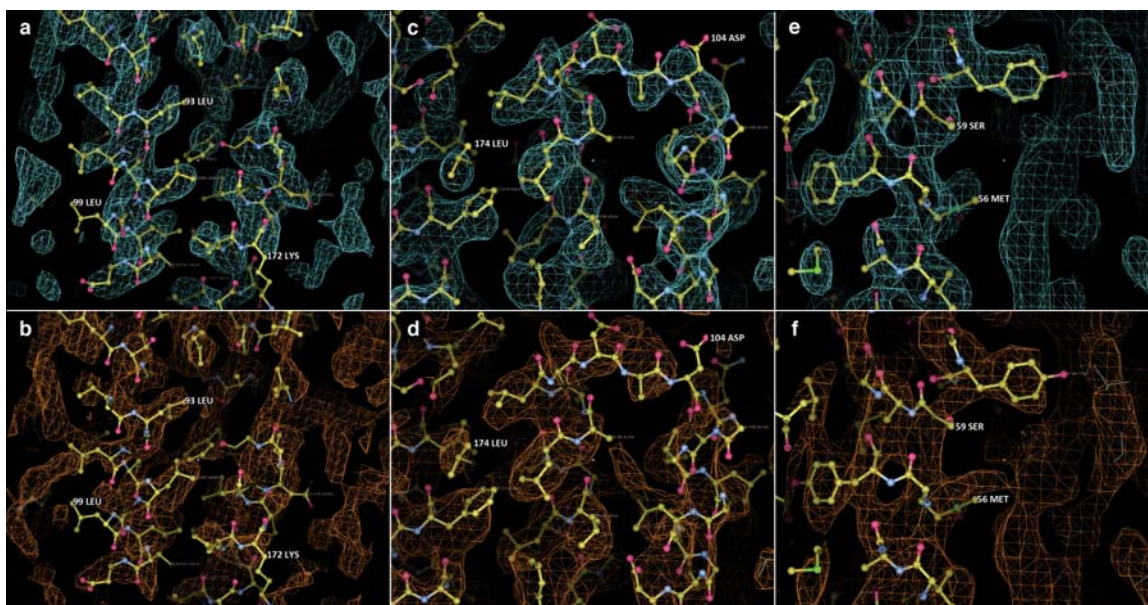


Figure 2: (Color) Structural Comparison - Close-up views of selected map positions of the bacteriorhodopsin dataset after refinement by PCO (top), and before refinement (bottom). Helix backbone (a, b) is well defined, and the loop at 103 ALA (c, e) is connected and well resolved relative to the rest of the structure. All prominent side chains (e, f) are accounted for.

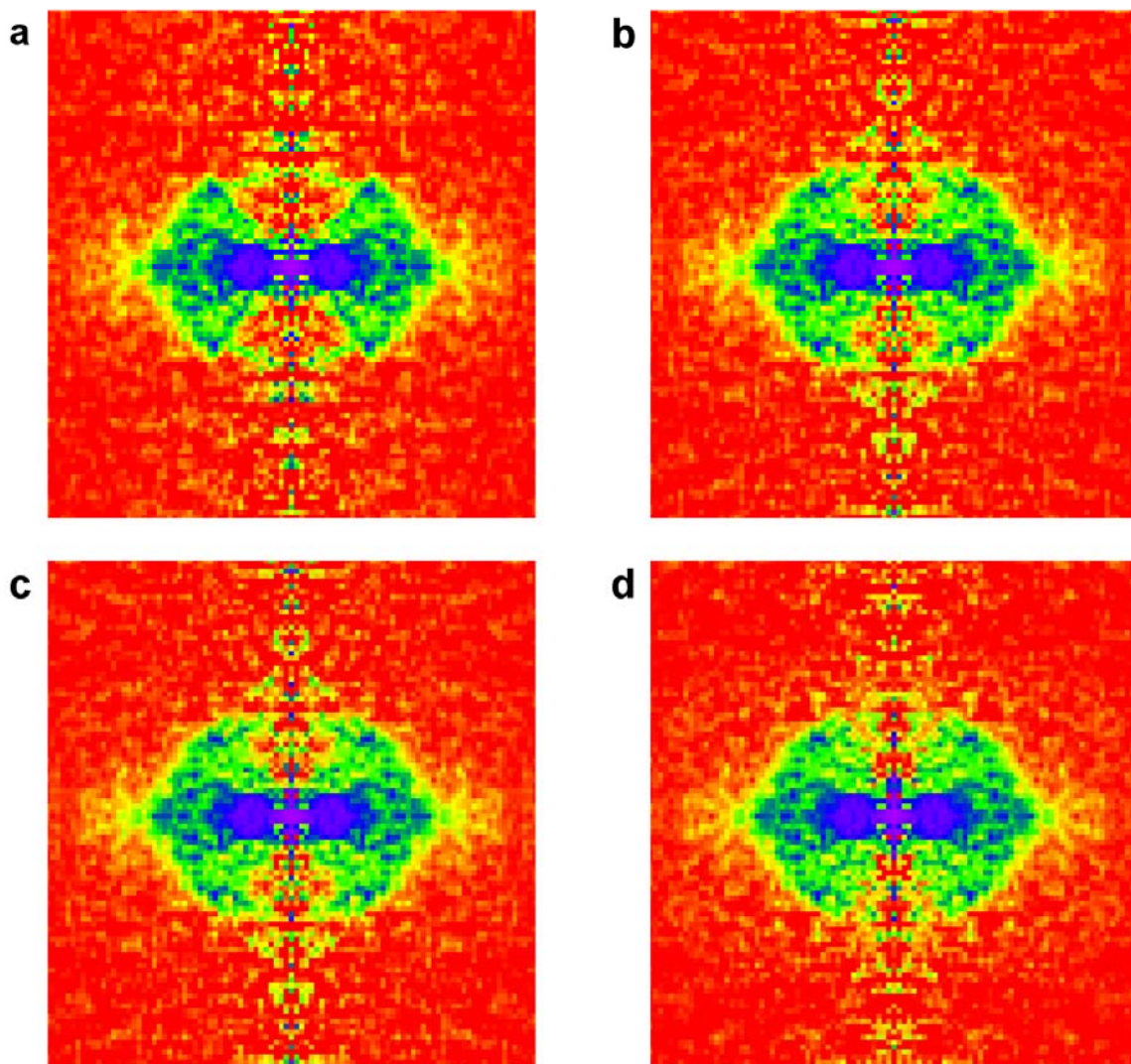


Figure 3: (Color) CRC Comparison of Tilt Limited Data – Calculated CRC for 3D-PCO reconstructions of the kimura-97 dataset tilt limited to (a) 30°, (b) 40°, (c) 50° and (d) 60° against 1BRR. Note the sharpest transition occurs between 30° and 40°, with CRC roughly constant there-after. Plots are given as a function of ring of radius ρ (horizontal axis, in Å), and height z (in the vertical axis, in Å) in Fourier space (see figure S2 in [33]).

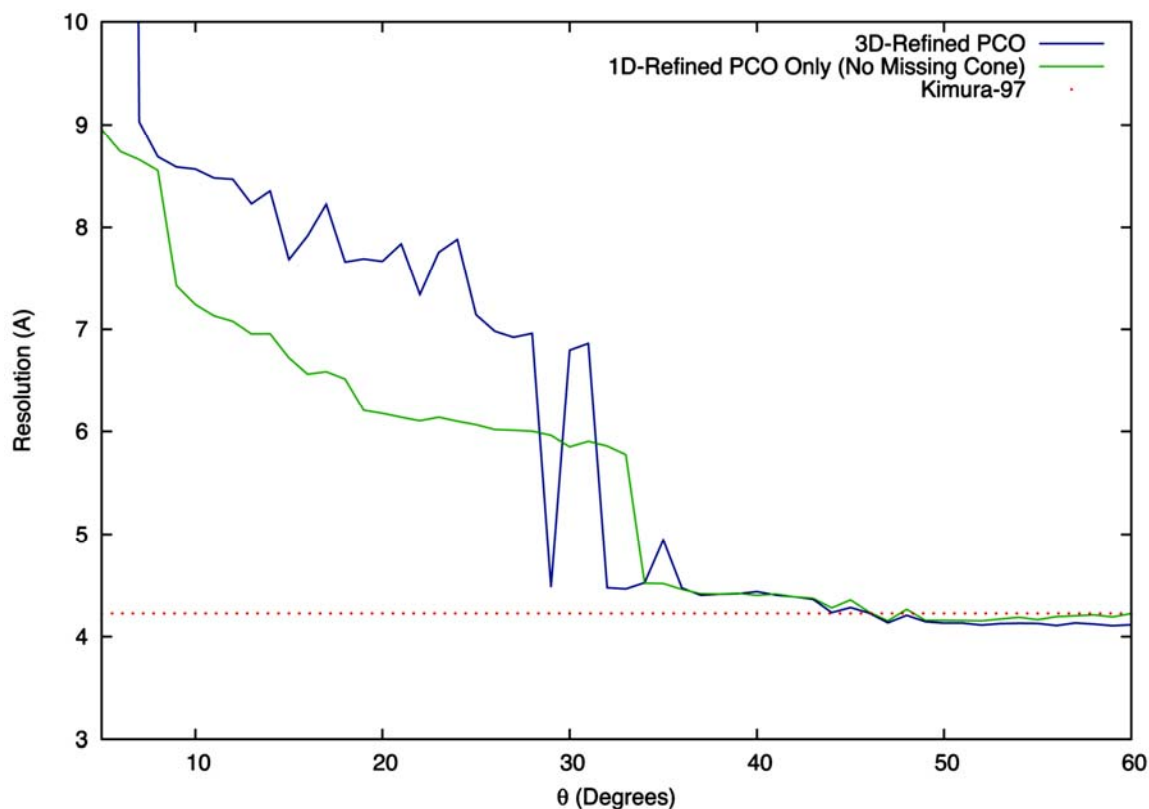


Figure 4: (Color) Resolution performance of PCO in dependence of available tilt-range data

– The Bacteriorhodopsin dataset was tilt-range limited at 1 degree intervals, each fully subjected to structure reconstruction by 200 rounds of PCO [33]. The average resolution of the final dataset was estimated by Fourier Shell Correlation (FSC, 0.5 threshold) with the 1BBR model. Shown is the resolution of the full 3D PCO refined dataset (blue line) relative to a one dimensional PCO lattice line refinement only (green line). The original MRC reconstruction was based on data up to 60° specimen tilt, and showed a 4.4Å resolution when evaluated by FSC-0.5. The plot shows that a tilt limitation to only $\pm 10^\circ$ sample tilts still allows a 9Å reconstruction, additionally verified by CRC and visual inspection. Inclusion of data up to only 35° tilts effectively allowed the full reconstruction of the structure up to a resolution of 4.5Å. Inclusion of data beyond a 45° tilt-range did not bring any further improvements.