Acta Cryst A
FOUNDATIONS
ADVANCES

ISSN 2053-2733

Algorithm for systematic peak extraction from atomic pair distribution functions

L. Granlund,[a]* S. J. L. Billinge[b,c]* and P. M. Duxbury[a]

[a]Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan, 48824, USA, [b]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027, USA, and [c]Condensed Matter Physics and Materials Science Department, Brookhaven National Laboratory, Upton, New York, 11973, USA. *Correspondence e-mail: luke.r.granlund@gmail.com, sb2896@columbia.edu

Received 8 August 2014
Accepted 14 March 2015

Edited by D. A. Keen, STFC Rutherford Appleton Laboratory, UK

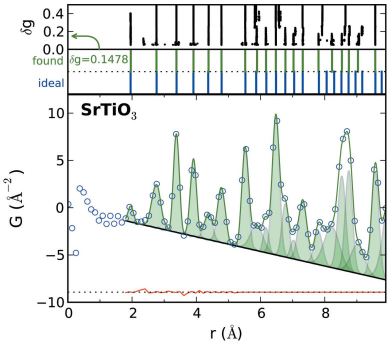
Keywords: pair distribution function; peak extraction; model selection; Akaike information criterion; computer program.

Supporting information: this article has supporting information at journals.iucr.org/a

The study presents an algorithm, ParSCAPE, for model-independent extraction of peak positions and intensities from atomic pair distribution functions (PDFs). It provides a statistically motivated method for determining parsimony of extracted peak models using the information-theoretic Akaike information criterion (AIC) applied to plausible models generated within an iterative framework of clustering and chi-square fitting. All parameters the algorithm uses are in principle known or estimable from experiment, though careful judgment must be applied when estimating the PDF baseline of nanostructured materials. ParSCAPE has been implemented in the Python program *SrMise*. Algorithm performance is examined on synchrotron X-ray PDFs of 16 bulk crystals and two nanoparticles using AIC-based multimodeling techniques, and particularly the impact of experimental uncertainties on extracted models. It is quite resistant to misidentification of spurious peaks coming from noise and termination effects, even in the absence of a constraining structural model. Structure solution from automatically extracted peaks using the Liga algorithm is demonstrated for 14 crystals and for C_{60}. Special attention is given to the information content of the PDF, theory and practice of the AIC, as well as the algorithm's limitations.

1. Introduction

Determination of all atomic coordinates and identities in materials at the nanoscale, also known as the nanostructure problem, is a major challenge in materials science and engineering (Billinge & Levin, 2007). Standard powder diffraction techniques are very successful for periodic systems (David *et al.*, 2002), but the increased prominence of semi-ordered and disordered materials, including nanoparticles, requires advances in acquiring and analyzing structural information (Billinge & Levin, 2007). One approach utilizes the atomic pair distribution function (PDF), which is a one-dimensional real-space function usually obtained from powder diffraction patterns (Egami & Billinge, 2012; Warren, 1990). PDF studies historically concentrated on amorphous, glassy and liquid systems (Warren, 1934; Wright, 1998), and this approach remains popular (Ma *et al.*, 2009). More recently, sufficiently small or partially ordered nanostructured materials inaccessible to crystallographic techniques have been successfully studied (Petkov *et al.*, 2002). Although refinement of an assumed structural model is the usual approach in modern PDF analysis, fitting individual features within the PDF remains an informative complement (Božin *et al.*, 2010). The principal motivation for the present study is to enable the *ab initio* creation of structural models starting from interatomic distances extracted from peak positions and intensities in the



measured PDF, as demonstrated by the recent Liga algorithm structure solution for some nanostructured materials (Juhás *et al.*, 2006, 2010). In these studies the peak positions were extracted manually, making this approach impractical as a more widely used general method for nanostructure determination. This work describes a robust algorithm for unbiased extraction of peaks from the PDF in the absence of a structural model, an important next step in structure determination from PDF materials.

This is similar to Le Bail *et al.* (1987) or Pawley (1981) fitting in powder diffraction. However, in the case of the PDF it is much more difficult since the number and position of the peaks are not known in general, as is the case with Pawley or Le Bail fitting. We introduce the PDF ParSimonious Clustering Algorithm for Peak Extraction (ParSCAPE), which combines standard chi-square fitting and a simple clustering technique to generate a collection of peak models using the Akaike information criterion (Akaike, 1973) to determine model parsimony. This algorithm has been implemented in a software code, *SrMise*, which is also described. Apart from the structurally dependent PDF baseline, all the input parameters to *SrMise* are known or reliably estimable from experiment, and basic tools to aid baseline estimation are considered. We summarize *SrMise*'s assumptions, operation and limitations as well as its performance on several experimental crystalline and nanoparticle systems. Attention is given to the effect of termination ripples, the PDF baseline, uncertainties and information content in the PDF, and multimodeling techniques. In particular, accurate experimental uncertainties are an important element of constraining model complexity, but historically the uncertainties obtained for PDFs from integrating detectors are unreliable or simply not calculated (Egami & Billinge, 2012). We investigate techniques to mitigate this issue, and we also test a PDF with correctly propagated uncertainties. Finally, we demonstrate structure solution using the Liga algorithm from ParSCAPE-extracted peaks, reproducing previously published results (Juhás *et al.*, 2006, 2010) with significantly reduced user intervention.

2. Definitions

For clarity we distinguish several types of peaks. Peaks which are fit to the PDF without the benefit of a structural model are *descriptive peaks*, and in their final form are *extracted peaks*. Unqualified use of the term peak generally refers to these. In contrast, *intrinsic peaks* are peaks which collectively constitute the PDF in its ideal form prior to the effect of noise, instrument effects, artifacts of data reduction *etc*. A *peak function* refers to the (perhaps approximate) mathematical form of some peak, *e.g.* a Gaussian. Although a model in the context of the PDF usually refers to a *structure model* of the system's atomic structure, when speaking of the PDF we freely use model to mean a parametric *peak model* which is a collection of descriptive peaks plus the PDF baseline.

3. The pair distribution function

The basic features of the PDF are summarized below. A detailed derivation of the PDF, valid for periodic and nanoparticle systems, is given in Farrow & Billinge (2009). A comprehensive overview of the PDF method is found in Egami & Billinge (2012). The PDF is a one-dimensional real-space function which characterizes all atomic pairs in a sample. The reduced PDF,

$$G(r) = \frac{2}{\pi} \int\limits_{Q_{\min}}^{Q_{\max}} Q[S(Q) - 1] \sin(Qr)\, \mathrm{d}Q, \qquad (1)$$

is the Fourier transform of the structure-dependent total scattering structure function $S(Q)$, where Q is the momentum transfer. The reduced PDF is a measure of the probability that an atom pair occurs with separation r, weighted by the scattering factors of the atoms in that pair, and these are observed as fluctuations, which decay as r increases, about $G(r) = 0$. Conveniently, experimental uncertainties in $G(r)$ do not scale with r and have approximately equal magnitude across the whole function (Egami & Billinge, 2012). Termination ripples, equivalent to the convolution of the 'true' $G(r)$ with a sinc function, are present in all experimental PDFs due to the finite $Q \le Q_{\max}$ measurement range. These become smaller with increasing $Q_{\max}$, though statistical noise on the data, also convoluted with the same sinc function, increases with increasing $Q_{\max}$ and the presence of peak-like spurious ripples in measured PDFs is inevitable. These are generally not a problem when fitting highly constrained structural models, but may be misinterpreted as peaks in an unbiased peak extraction. In general, we would like to extract reliable intrinsic peaks in the presence of these ripples in a measured PDF.

A related function is the radial distribution function (RDF) $R(r)$, which is the interatomic distance probability distribution. In the harmonic approximation of atomic interactions the contribution to the RDF from each atom pair is a Gaussian (the intrinsic peak) located at the mean separation of the atoms, with an integrated area given by the absolute value of the products of the scattering factors for the corresponding atoms. Given a properly normalized PDF, the area of an extracted peak in an accurate peak model thus is a scattering-power-weighted measure of the occurrences of that distance between those atom pairs. Refinements to peak shape are known, such as from angle averaging in powder diffraction (Dimitrov *et al.*, 2001), finite Q resolution (Thorpe *et al.*, 2002), anisotropic crystals (Thorpe *et al.*, 2002) and quantum vibrational modes (Levashov *et al.*, 2007). These corrections are almost always small, and we do not consider them further.

Momentarily ignoring finite $Q_{\max}$, $G(r)$ and $R(r)$ are related by

$$G(r) = \frac{R(r)}{r} - 4\pi\rho_0\gamma_0(r)r, \qquad (2)$$

where ρ_0 is the average density and $\gamma_0(r)$ is the characteristic function of the sample shape (Guinier *et al.*, 1955; Farrow & Billinge, 2009), also known as the nanoparticle form factor in the PDF literature. The $4\pi\rho_0\gamma_0(r)r$ term is a baseline, and we may informally think of $G(r)$ as the baseline plus peaks. For bulk systems of constant density $\gamma_0(r) \sim 1$, giving the familiar linear baseline found in most definitions of the PDF in the literature. The case is considerably more complex for discrete nanoparticles, as $\gamma_0(r)$ is the orientational average of the normalized autocorrelation of the particle's shape such that $\gamma_0(0) = 1$ and $\int_0^\infty \gamma_0(r)\,\mathrm{d}r = V$, where V is the nanoparticle's volume. Typically the nanoparticle baseline is linear at small r and smoothly goes to 0 at the size of the nanoparticle. Its exact form depends on the details of the nanoparticle's shape and size. In fact, the PDF baseline is due to unmeasured scattering from structure at length scales corresponding to $Q < Q_{\rm min}$, and $\gamma_0(r)$ can be determined in principle from small-angle scattering experiments (Farrow & Billinge, 2009). Without these data it can only be approximated due to uncertainty about the nanoparticle's actual structure, polydispersity of the prepared sample, and the contribution of interparticle correlations to $G(r)$ for r less than the nanoparticle size.

4. Statistical considerations

4.1. Motivation

For the last few decades the PDF community has had considerable success focusing on structural modeling, starting from an initial guess structure and refining to fit the PDF. However, key differences exist between structural modeling and peak extraction. For example, since uncertainties in $G(r)$ have approximately equal magnitude δG, chi-square refinement of a given initial structural model produces nearly identical refined parameter values (though not estimated uncertainties) regardless of the actual magnitude of the uncertainties. In contrast, the impact of a particular value of δG on peak extraction, which should consider whether a PDF feature is a peak or a fluctuation, is more direct, with the same δG introducing greater uncertainty on a small peak than a large one. Furthermore, it is common practice to sample $G(r)$ on a very fine grid, introducing strongly correlated uncertainties between nearby points (Farrow *et al.*, 2011), which for many optimization algorithms violates the assumption that each sample point is statistically independent. This is often harmless, but at worst can create the illusion of support for unjustified conclusions.

Of central interest in peak extraction and cluster analysis is how many peaks/clusters ought to be found, and how they can best be refined and interpreted (Bock, 1996; Tibshirani, 2001; Shao & Wu, 2005). As one of the fundamental issues in scientific modeling, it is a question with deep underpinnings in the philosophy of science (*e.g.* Occam's razor), and frequent interaction with statistics and information theory. Practically speaking, peak extraction packages often require the user to identify suspected peaks by hand, specify their number directly, or specify the latter indirectly with a parameter such as a smoothing factor or definition of a 'threshold' residual. Such techniques are often thoroughly *ad hoc*.

4.2. Information and uncertainties in the PDF

The literature considering powder diffraction and the PDF from the viewpoint of information theory is small, although there is a long history of uncertainty analysis. Recent examples include David & Shankland (2008), Farrow *et al.* (2011), Mullen & Levin (2011) and Toby & Billinge (2004). We summarize the most basic properties.

Sampling the experimental $G(r)$ more frequently than the Nyquist rate $Q_{\rm max}/\pi$ (equivalently $dr < \pi/Q_{\rm max}$) introduces no new information (Farrow *et al.*, 2011). This is due to the famous Nyquist–Shannon sampling theorem for 'band-limited' signals (Shannon, 1949). Indeed, the Whittaker–Shannon interpolation formula defines a continuous function identical for all $G(r)$ sampled above the Nyquist rate. This result does not apply in the ideal case where $Q_{\rm max} \to \infty$, as the signal is no longer band limited.

Given sufficient counting statistics, uncertainties in $G(r)$ are approximately normally distributed, but are also correlated due to the Fourier transform from reciprocal space, which produces an oscillating contribution to the uncertainty at each point. The contributions from terms separated by greater than the Nyquist rate are nearly out of phase and approximately cancel (Farrow *et al.*, 2011; Toby & Billinge, 2004). Thus, Nyquist sampling gives the least correlation without losing information. Although this is not true statistical independence, it is the best approximation to independent, normally distributed uncertainties that can be obtained for $G(r)$.

The uncertainty discussed above arises solely from counting statistics, but other uncertainties are present in $G(r)$. These include systematic errors due to data reduction, instrument effects *etc.* and are especially prominent in the low-r region where no physical peaks are present (Egami & Billinge, 2012). See also Appendix *A*, which discusses a problem propagating the uncertainties of integrating detectors in popular PDF reduction software. This issue is resolved in the upcoming generation of tools (Yang *et al.*, 2014).

Note that the Nyquist rate and the PDF uncertainty (*ergo* information content) are intimately connected *via* $Q_{\rm max}$, which determines the former directly and contributes to the latter due to decreasing signal to noise as Q increases. Consequently, the optimal number of parameters in a 'best fit' to the PDF is poorly defined, although the number of statistically independent points in reciprocal space is clearly an upper bound. The limit for structural modeling is especially unclear, since the information in the PDF necessary for reliable refinement also depends on whether the structural features of interest are discernible and within the fitted range (Farrow *et al.*, 2011), while by definition these are the only features that may be investigated in peak extraction. In Rietveld refinement, where the relevant 'independent' quantities are actually the integrated intensities of individual reflections, current best practices suggest at least three to five times as many independent points as refinable parameters are needed for a stable and accurate refinement, but this rule of thumb does not have a rigorous justification (McCusker *et al.*, 1999).

4.3. The Akaike information criterion

Information theory has developed tools which can help address these difficulties. One such tool is the Akaike information criterion (AIC). The AIC and its offshoots appear often in the ecological and biological sciences (Arnold, 2010), signal processing (Stoica & Sel, 2004), artificial intelligence (Zhao *et al.*, 2008), and increasingly in astrophysics (Liddle, 2007; Wei, 2010). The AIC has also appeared in a peak extraction algorithm developed within the metabolomics community (Morohashi *et al.*, 2007).

The AIC is developed from the Kullback–Leibler (K–L) information

$$I(f,g)=\int f(x)\log\frac{f(x)}{g(x)}\,\mathrm{d}x, \tag{3}$$

which is a measure of information loss when a distribution g approximates distribution f (Kullback & Leibler, 1951). Akaike's result establishes a relationship between the maximized log-likelihood of a model (strictly speaking the distribution it induces) and K–L information. Namely, the maximized log-likelihood of model g given the data is a biased estimate of the expected relative K–L information, and in the asymptotic limit of many independent data points the bias is the number of parameters in g (Akaike, 1973; Kotz & Johnson, 1992). The qualifiers *expected* and *relative* are both a consequence of our ignorance of the 'true' model, f, possessing only the data instead. It is an expected measure because the parameters which maximize the log-likelihood are themselves estimated from the data. It is a relative measure because, for the models $\{g_1, g_2, \ldots\}$ of interest, the unknown contribution of f to $I(f, g_i)$ can be treated as a constant that cancels from $I(f, g_i) - I(f, g_j)$ (Bozdogan, 1987). Since the bias due to parameters is estimated, the AIC itself is asymptotically unbiased, and comparison of quite different models (even non-nested models) is possible. Crucially, the 'true model' is not assumed to be among those available to the investigator (Burnham & Anderson, 2002; McQuarrie, 1998). The approximation of descriptive peaks for the large number of intrinsic peaks is therefore well tolerated in principle.

The AIC has the general form

$$\mathrm{AIC}=-2\ln L+2k \tag{4}$$

where L is the maximized likelihood function and k is the number of estimated parameters in the model. If the uncertainties in the data are independent and normally distributed, the AIC has the convenient form $\mathrm{AIC}=\chi^2+2k$ up to an ignorable model-independent constant, where $\chi^2=\sum_i \varepsilon_i^2/\sigma_i^2$ is the usual chi-square error of the model with residuals ε_i and uncertainties σ_i. Adding parameters to a model will tend to reduce χ^2 but increase the contribution from the parameter bias, therefore suppressing both overfitting and underfitting. In addition, since χ^2 for a given model increases if the uncertainties decrease, the tolerated complexity of a model depends on the actual uncertainties in a natural way.

A lower AIC indicates a more plausible model, but the value of the AIC has no absolute interpretation – only differences in AIC among models compared to the same data have meaning. Furthermore, if we define $\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}$ for a given set of models, where $\mathrm{AIC}_{\min}$ is the minimum AIC among the set, then the relative likelihood for the ith model (given the data) is $\exp(-\Delta_i/2)$. These may be normalized with respect to all the considered models to give the 'Akaike weights':

$$w_i=\frac{\exp(-\frac{\Delta_i}{2})}{\sum_j \exp(-\frac{\Delta_j}{2})}, \tag{5}$$

which is the likelihood that the ith model is the K–L best model from among the models considered. By definition that model exists within the set, and so $\sum_i w_i = 1$. As with the AIC itself, the Akaike weights do not measure 'correctness', but how each model fares amongst its peers. This includes determining if a small subset of models are favored or if none distinguish themselves, but the Akaike weights may also be used to estimate properties over the entire set of models. For example, if a parameter appears in many plausible models (*e.g.* an atomic displacement parameter in structural modeling) the uncertainty of that parameter may be estimated considering all the models rather than just a single one. This feature underlies the AIC's strength in multimodel comparison and inference, and provides tremendous flexibility.

An important related criterion known as $\mathrm{AIC_c}$ includes a second-order correction for sample size. It is

$$\mathrm{AIC_c}=-2\ln L+2k+2k\frac{(k+1)}{n-k-1} \tag{6}$$

where n is the number of independent data points (Hurvich & Tsai, 1989; Sugiura, 1978). It is clear from the final term that $\mathrm{AIC_c}$ penalizes parameters more heavily than the AIC for given n, and is asymptotically equivalent to the AIC as $n/k \to \infty$. For this reason many authors suggest using $\mathrm{AIC_c}$ instead of the AIC whenever possible. However, ParSCAPE does not use $\mathrm{AIC_c}$ because the correction term only appears if the uncertainty in the data is estimated along with the model (*i.e.* as an additional parameter), whereas PDF uncertainties are directly estimated from experiment. A comprehensive analysis of the AIC and related criteria is found in Burnham & Anderson (2002).

Although the AIC is simple, powerful and well understood, we mention in passing other model selection methods one might consider. Adjusted R^2, reduced chi-squared and repeated F-tests have significant weaknesses for model selection (McQuarrie, 1998; Andrae *et al.*, 2010), particularly for non-nested models that may arise from, for example, different PDF baselines. Full Bayesian modeling, bootstrapping and cross-validation are computationally intense and not suited to our approach, although AIC has been shown to be asymptotically equivalent to the latter (Stone, 1977). The post-AIC literature abounds with new criteria and proposed refinements under various assumptions. Examples include the Bayesian information criterion (Schwarz, 1978), quasi-AIC (Lebreton *et al.*, 1992; Anderson *et al.*, 1994), Takeuchi information criterion (Takeuchi, 1976), deviance information criterion (Spiegelhalter, 2002), focused information criterion (Claeskens & Hjort, 2003) and information complexity (Bozdogan, 2000). Development of an information criterion tailored to the particulars of the PDF (*e.g.* accounting for correlations present even at Nyquist sampling) may be an avenue of future research.

5. Algorithm description

5.1. Overview

ParSCAPE, described here for the particular case of the PDF (Fig. 1), uses a simple clustering method as a framework for finding descriptive peaks over peak-like regions, where the clusters are defined below. As the fit progresses, the clusters are carefully combined as they meet using a recursive call to find obscured peaks in model residuals. The eventual result is a single cluster, the *global cluster*, from which a greedy pruning subroutine attempts to remove the least justified peaks. The peak function is initially a Gaussian over r, but termination ripples are applied with a modified pruning step to remove spurious peaks before completion. Finally, the PDF baseline, which must be estimated or known before extraction, is fit along with extracted peaks.

In ParSCAPE the AIC takes the specific form

$$\mathrm{AIC} = \sum_i \frac{[G_i - (P_i + B_i)]^2}{\delta G_i^2} + 2 \times 3 \times \#P \qquad (7)$$

where G_i (δG_i) is the ith data point (uncertainty) within the cluster, P_i the value of the sum of model peaks at the corresponding point, B_i the value of the baseline at this point and $\#P$ the number of peaks in the model. The factor of 3 in the last term is due to the number of fit parameters in the peak function. The baseline is considered fixed during clustering, and its contribution to the number of refinable parameters is ignored until the baseline is fit in the final step. Therefore, the AIC of the final model has an additional contribution equal to twice the number of fit parameters in the baseline.

The algorithm has three key features. First, all inputs other than the PDF baseline are in principle known or estimable solely from experiment, and a reasonable PDF baseline is frequently estimable with minimal structural assumptions. Second, the smooth growth from many low-parameter models to a single many-parameter model mitigates a major weakness of local optimization methods, namely the need for increasingly good initial conditions as the number of parameters increases (Transtrum *et al.*, 2010). Finally, and most significantly, model parsimony is addressed in a consistent and statistically motivated fashion using the AIC, and is primarily dependent on the extent and uncertainty of the data.

5.2. Baseline estimation and oversampling

The first step is baseline estimation and, as this estimate biases the extracted peaks, the best available structural information should be applied. Fitting an empirical or appropriate analytical function is often sufficient. This subtle issue is discussed at length in §8.1.

The data are then moderately oversampled (approximately five times the Nyquist rate) for the beginning and intermediate stages of peak extraction. Although this adds no information, it makes some peaks more apparent while clustering, and we prefer to err gently on the side of overfitting in the early stages.

5.3. Clustering

5.3.1. Growing clusters. A ParSCAPE cluster is a set of contiguous data along the r axis of $G(r)$ which partitions the PDF into regions with peak-like features, and is associated with a peak model fit only to the data within that cluster. To begin, all sample points are ordered by $G(r) - B(r)$, where $B(r)$ is the estimated baseline, from greatest to least. The data are clustered in that order, starting from an initial cluster containing the first point in the list, and each additional point creates a new cluster or is added to an existing one. Which of these actions is performed is determined by the distance parameter d_c equal to the Nyquist interval $\pi/Q_{\max}$. Let $[r_p, G(r_p)]$ denote the largest point not yet added to a cluster, and $[r_n, G(r_n)]$ be the already clustered point which minimizes $d = |r_p - r_n|$ (*i.e.* the nearest point in the nearest cluster). If $d > d_c$ a new cluster containing the point is created, otherwise the point is added to the nearest cluster. From the perspective of cluster analysis this procedure is very similar to agglom-

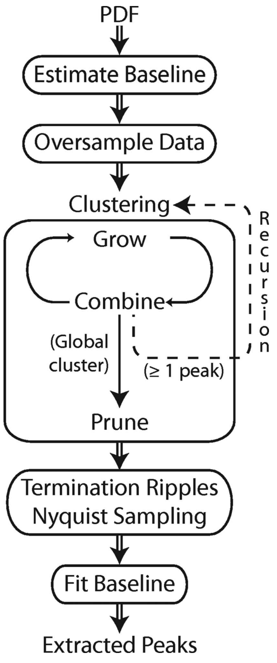


Figure 1
Summary of ParSCAPE in the context of the PDF. Clusters grow (cover greater regions of the data) while adding single peaks, and combine immediately if no unclustered data remain between them, eventually leading to the global cluster. Recursion over the residuals of the existing fit (near where clusters meet) occurs if there is at least one peak in a newly combined cluster. Pruning greedily removes peaks with the least statistical support from the global cluster. Removal of termination ripples is a modification of the pruning process, which simultaneously returns the data to Nyquist sampling. Adding or removing peaks is governed at all stages with Akaike's information criterion.

erative hierarchical methods using single linkage (Gan *et al.*, 2007), with the unusual feature that the data are considered in an order not defined by cluster distance. It also has some similarity to the very recent general clustering method of Rodriguez & Laio (2014).

After a cluster is created, but before it has been combined with another cluster (§5.3.2), its model contains at most one peak. This first peak is created only if a peak estimated from and then fit to the data in the cluster has a lower AIC than the cluster with no peaks at all. As more points are added to the cluster the model is occasionally refit to detect hints of obscured peaks with the following heuristic. Consider a hypothetical model that perfectly fits the data ($\chi^2 = 0$) with a single additional parameter. If the AIC of this hypothetical model would be less than that of the existing model (*i.e.* the hypothetical model is 'better'), then it is possible there exists a model with additional parameters that would fit the data in the cluster better than the current model. Fig. 2 shows how the AIC weighs the evidence for simple models on peaks with simulated noise. A likely source of these parameters are peaks in a nearby cluster, or a hidden peak within this one. In either case it is advantageous not to fit the existing model to data indicative of peak overlap until alternate models can be explored. Therefore, the model reverts to its previous value, and fitting of this cluster ceases until it combines with another.

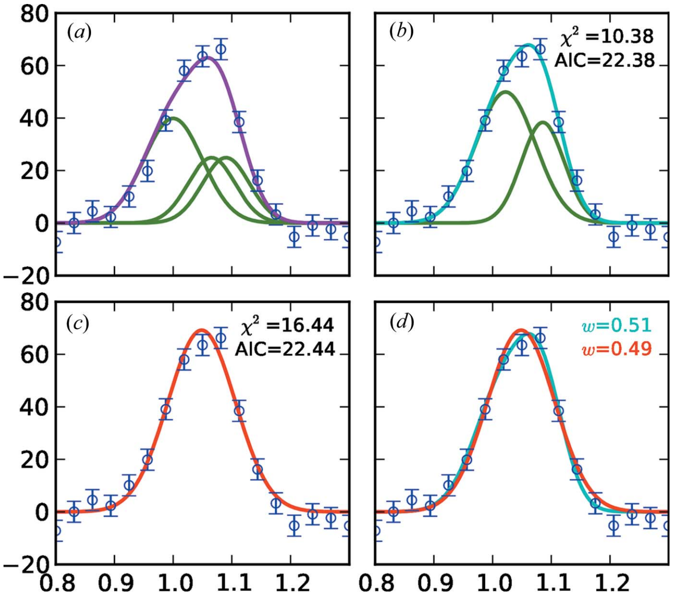


Figure 2
Comparison of two simple models with the AIC. (*a*) The 'experimental' data in blue incorporate simulated Gaussian noise ($\sigma = 4$) added to the sum (magenta) of the three intrinsic Gaussian peaks (green). (*b*) The sum (cyan) of a model with two Gaussian peaks (green) fit to the data. (*c*) A model with a single Gaussian peak (red) fit to the data. This model has larger chi-square error, but requires three fewer parameters. (*d*) The sum of both models compared to the data. Considering just these two models, the Akaike weights w are right at the boundary of favoring the more complex model, however weakly. The AIC decisively rejects (relative to these two models) a model with a third peak for these data. Although the latter is 'correct', AIC judges the slightly improved chi-square error as insufficient to justify additional parameters. In PDFs with significant overlap the same is often true. In addition, recovering the two intrinsic peaks which nearly overlap presents convergence challenges. Pawley-style fitting addresses this by artificially constraining those peak intensities to be equal, but this requires prior knowledge of peak positions.

5.3.2. Combining clusters. When two clusters have no unclustered points between them, they are combined. Any 'interlaced' peaks (*e.g.* a peak in the left cluster located to the right of the leftmost peak from the right cluster) are removed. Though rare, this may indicate an unreliable fit, or that a feature shared by both clusters, such as a small or obscured peak not identified as a separate cluster, has been found independently in each.

Recursion is the next step in combining clusters, and searches for peaks in a boundary region near where the clusters meet, which extends from the positions of the second-nearest-neighbor peaks on either side of the boundary, or to the edge of a cluster that does not contain at least two peaks. If neither cluster contains peaks recursion is not performed. There are two separate preparatory cases before recursion, which are applied under different conditions. The first simply fits the existing model within the boundary region and then calculates the residuals. The second is more involved, and adjusts model parameters without chi-square fitting in a way which attempts to reduce large overlapping contributions from peaks initially from different clusters while preserving hints of obscured peaks in the residual. Namely, if the contribution of these peaks at the boundary $r = r_b$ is greater than $G(r_b)$ the parameters of each peak are the solution (if it exists) to a system of equations such that their sum at r_b is exactly $G(r_b)$, retaining their relative proportions at r_b, while the locations and magnitudes of each peak maximum are unchanged (see the supporting information for details). Only then is the residual calculated.

If at least one peak exists in either cluster, recursion is performed using the first preparatory case. If both peaks contain at least one cluster, a separate instance of recursion using the second preparatory step is also performed. In either case, recursion includes the clustering steps only. Peaks found during recursion are added only if they improve the AIC in the newly combined cluster, and if recursion was performed twice the model which most improved the AIC is retained.

5.3.3. Pruning clusters. When all clusters are combined into a single global cluster containing all the data a pruning step is performed. The motivation for pruning is that some peaks that initially improved the AIC become unfavorable by that same measure as extraction progresses, principally due to well fit peaks originally identified in other clusters. Convergence issues that can lead to unphysical results, such as peaks no longer responsive to changes in parameters or constrained by the data, occasionally appear as well. Pruning is a greedy heuristic that attempts to remove the least justified peaks from the global cluster until the most favorable AIC is obtained. This process, which restricts model complexity, is critical to ParSCAPE.

Pruning creates multiple copies of a model, with one peak removed from each copy. These are each fit and the single model with the best AIC is retained. This is repeated until no improvement is observed. Clearly this is computationally intensive if every peak is a potential candidate for removal, especially since many of these trial models will be unable to converge if an important peak is removed. ParSCAPE combats this in several ways. First, if removing a peak does not lead to improvement compared to the original model that peak will not be tested in future iterations. Second, only peaks near the removed one are fit, and all other peaks in the cluster are temporarily held fixed, though their contributions to the AIC are still calculated and the best trial model is fit with all parameters after each iteration. Third, models that do not converge rapidly are treated as offering no improvement.

5.4. Termination ripples, resampling and baseline

If this is not a recursive call to ParSCAPE then, after pruning the global cluster, termination ripples are applied to each peak, the data are Nyquist sampled and the model is again pruned. If a model has more parameters than samples at the Nyquist rate then an intermediate sampling rate is used, and pruning/downsampling is repeated until either Nyquist sampling is achieved, or pruning at the least oversampling possible for a given model shows no improvement in the AIC. This process opportunistically removes peaks that are likely termination artifacts and ensures the model adheres to our best statistical justifications.

By this time the model is highly conditioned on the assumed baseline, as its parameters have been effectively fixed since baseline subtraction. Nevertheless, a final model including the extracted peaks as well as the baseline (treating its earlier parameters as initial values) is fit to $G(r)$.

6. Multimodel selection

The procedure of §5 selects a single model based on local comparisons, but in the context of a complex nonlinear problem this does not necessarily yield the globally optimal solution. Even locally the AIC may not strongly favor a particular model (Fig. 2), so retaining only the most favored model can be misleading. In view of these issues, AIC-based multimodel selection over a population of plausible but physically distinct models evaluated over all the experimental data is advisable. The strategy ParSCAPE employs to generate a suitable population is multiple trials of peak extraction over a range of assumed PDF uncertainties. The uncertainty assumed during a single ParSCAPE trial effectively becomes a parameter, denoted δg to distinguish it from the true experimental uncertainty δG. Peak extraction as δg decreases will tend to produce relatively more complex models because the fractional contribution to the AIC due to the number of parameters decreases as χ^2 increases, and contrariwise for increasing δg.

The next step in multimodeling is calculating the Akaike weights for each model when compared to the data, which aids the investigator in identifying models meriting further study and even prioritizing which models to examine first. However, this cannot be done immediately because the Akaike weights are improperly normalized if there are redundant models in the population. To account for this, ParSCAPE groups models into 'similarity classes'. Each model in a class has the same number of peaks and essentially identical values for peaks and baselines. For this purpose peaks (or baselines) p and p' are considered 'identical' if they have the same peak function and simultaneously satisfy

$$\frac{|\sum_i [p(r_i)^2 - p'(r_i)^2]|}{\sum_i p(r_i)^2} \leq t \tag{8}$$

as written and under exchange of labels $p \leftrightarrow p'$, where t is the fractional tolerance. This heuristic is sensitive to differences when either compared peak is large, and insensitive when both are small. A model is added to an existing class if all its peaks and its baseline are 'identical' to the model which first defined the class. If no such class exists, the model defines a new one.

By definition all models in a class should have similar AIC, but some variance is inevitable. ParSCAPE defines the properties of a class to be identical with those of the constituent model with least AIC. Misclassifying relatively poor models will usually have very limited impact on conclusions, as such models make only minor contributions to the Akaike weights unless they far outnumber good models. However, failing to group many essentially identical good models can significantly change results. Furthermore, limiting comparisons to the first member assigned to a class may be non-optimal. Since these classifications are not unambiguous the interpretation of the Akaike weights in the context of ParSCAPE should be primarily qualitative.

The AIC-based multimodel selection framework does not define the method by which the population of models is generated, and ParSCAPE is only one possibility for peak extraction from PDFs. Investigator knowledge about what models are physically plausible can be naturally incorporated within this framework by excluding or adding models to the population over which the Akaike weights are calculated.

7. Testing

7.1. Implementation and availability

ParSCAPE has been implemented using the Python programming language, versions 2.6–2.7, in the program *SrMise*, which is available from the DiffPy website http://www.diffpy.org under a Berkeley Software Distribution-style licence. Efficient numerical computation is provided by the *NumPy* and *SciPy* packages, which include the MINPACK implementation of the Levenburg–Marquardt algorithm. Basic functionality is provided at the command line and advanced functionality with Python scripting. Additional features include templates for easy extensibility to other peak and baseline functions, the ability to specify known peaks with some or all parameters fixed, standard uncertainty reporting, and basic AIC multimodeling functionality. Rudimentary support for automatic determination of $Q_{\max}$ and crystal baseline estimation are also included.

The source code, installation instructions, a user's guide and example scripts can be found at http://www.diffpy.org.

For more information, please contact Simon Billinge (sb2896@columbia.edu) or Luke Granlund (luke.r.granlund@gmail.com).

7.2. Methods

Using *SrMise*, we examined the performance of ParSCAPE on experimental X-ray PDFs of 16 bulk crystals, C_{60} and a PbTe nanoparticle sample. Peaks were extracted up to 10 Å for crystal structures, 7.5 Å for C_{60} and 20 Å for the PbTe nanoparticle. All PDFs were obtained using the *PDFgetX2* software (Qiu *et al.*, 2004). As discussed in Appendix *A*, experimental uncertainties are often not correctly propagated to the PDF from integrating detectors by popular data reduction software, and that is the case for these PDFs. Methods to cope with this shortcoming are described below. However, we also performed a new data reduction for the $SrTiO_3$ sample using *SrXplanar*, which can estimate and propagate uncertainties from two-dimensional detectors (Yang *et al.*, 2014), to test a PDF with accurate uncertainty estimates.

All trials were performed on a desktop computer with a 3 GHz Intel Core2Duo processor. Execution time for a single ParSCAPE trial depends on PDF complexity, the number of points over which sampling is performed, and especially on the PDF uncertainty. In our tests these ranged from a few seconds to several minutes, with 15–30 s most typical.

The peak function used is the Gaussian over r,

$$\frac{|a|}{r(\frac{\pi}{4\log 2}w^2)^{1/2}}\exp\left[\frac{-4\log 2}{w^2}(r-r_0)^2\right], \tag{9}$$

where

$$w^2 = \tfrac{1}{2}(\sin w' + 1)w_{\rm max}^2, \tag{10}$$

and (in terms of the underlying Gaussian) a is the area, w is full width at half-maximum (FWHM) and r_0 is the peak position. Since Levenburg–Marquardt is an unconstrained optimization method, we implicitly enforce mild restrictions with this parameterization. The absolute value of a ensures only positive peaks are found in the physical region (see §8.3). The parameter w' limits peak FWHM to the interval $[0, w_{\rm max}]$, and reduces the likelihood of extracting unphysically wide peaks in regions of high overlap. In our trials $w_{\rm max} = 0.8$ Å for the zinc sample, and 0.7 Å for all others. (Typical FWHM of a single peak due to thermal broadening is 0.2–0.4 Å.)

Termination ripples are applied by evaluating the fast Fourier transform of equation (9), zeroing the Fourier components corresponding to $Q > Q_{\rm max}$, and taking the inverse transform. The target grid for peak function evaluation can vary over a ParSCAPE trial, both in extent and sampling rate, so the impact of edge effects, discretization error and effectively varying Q resolution can lead to numerically inconsistent termination ripples despite identical peak parameters. To limit these to negligible levels ($\ll \delta$G) the transform is performed on a grid sampled at five times the desired sampling rate and extended by about four additional ripples ($8\pi/Q_{\rm max}$) at both ends.

We estimate the linear baseline for crystal systems in a semi-automated fashion by fitting the large, well separated peaks at very low r starting from an approximate slope determined by inspection. Unphysically large δg is used to avoid fitting anything but the baseline and most important peaks. The C_{60} baseline is estimated by subtraction of the interparticle correlations fit by an analytical RDF of hollow spheres in an f.c.c. (face-centered cubic) lattice (Heiney *et al.*, 1991). The PbTe nanoparticle baseline is an *ad hoc* fit using the characteristic function of a sphere (Guinier *et al.*, 1955), with no corrections for interparticle correlations. See the supporting information for details on these baselines.

Multimodel selection on a population of models generated by 500 ParSCAPE trials was performed by the method described in §6 for each sample. We empirically observed tolerance parameter $t = 0.1$ to be sufficient for this analysis. (Using $t = 0.05$ and $t = 0.2$, for example, results in slightly different classifications, but does not change the models selected as best nor otherwise affect our conclusions.) Further details of testing differ for PDFs with unknown and known δG, described in §§7.3 and 7.4 below.

7.3. Unknown δG

The commonality of X-ray PDFs obtained from integrating detectors using software tools that do not report an accurate δG compels consideration of mitigation strategies. One might be tempted to use $\rm AIC_c$ to estimate uncertainties while comparing models, but as ParSCAPE generates these models assuming known uncertainty this clearly begs the question. Instead, to obtain a population of models with a wide variety of complexity, 500 ParSCAPE trials were performed with δg equal to 0.5–5% the maximum value of $G(r)$ in the extracted region, which are typical values of δG for a high-quality PDF based on the results of structure modeling. Multimodel selection analysis was then performed 500 times, treating each of these plausible uncertainties as the true δG in turn. Although δG can be estimated by other means (*e.g.* bootstrapping, residual analysis of the 17 tested PDFs with known structures) we perform the analysis without this information to show that it remains effective, to guide investigators in similar situations, to test the degree ParSCAPE results are consistent with their assumptions, and to show approximately how multimodeling results depend on PDF quality.

We define two sets to summarize results of these 500 multimodel analyses. Let $C_{\rm best}$ be the set of similarity classes with maximum Akaike weight for at least one assumed value of δG, and $M_{\rm best}$ be the set of best models from those classes. Although the actual likelihood of these models cannot be compared without knowing δG, we interpret these as the most plausible models in the absence of this knowledge. The contrast between the $M_{\rm best}$ models and those of standard AIC multimodeling, where the best models are those with significant Akaike weight determined only at the true experimental δG, should not be overlooked.

For samples with a well characterized structure (all but the PbTe nanoparticle) the quality of extracted peaks can be

Table 1
Summary of peak extraction and AIC-based multimodeling results from 500 ParSCAPE trials for each of 18 X-ray PDFs with unknown δG.

R_w is the residuum of Nyquist-sampled PDF structure refinement. Cl is the number of classes. The $M_{\rm best}$ column reports the number of these models, the number consistent with plausibility by the $\chi^2_{\rm red}$ criterion [equation (11)], and the number resulting in correct Liga solution (including chemical species assignment). The Peaks column gives the number of peaks extracted in the various $M_{\rm best}$ models, and in brackets the number calculated from the reference structure with precision 0.01 Å. Details of the forward and backward cases of consistency checks are given in the main text.

					$M_{\rm best}$ models				Mean $\delta g - \delta G$ (%)	
$\mathrm{Sample}^{\mathrm{ref.\ structure}}$	Atoms	$Q_{\rm max}$	R_w	Cl	No.	χ^2	Liga	Peaks	Forward	Backward
Ag[1]	4	35	0.095	13	3	2	3	11-16 [11]	−1.00 (87)	0.02 (94)
$BaTiO_3$[2]	5	26	0.123	63	5	2	5	13-23 [89]	−0.76 (57)	−0.71 (62)
C graphite[1]	4	22	0.266	358	6	2	1	19-26 [52]	−1.8 (10)	−1.7 (11)
C_{60}[3]	60	21.3		63	3	2	2	12-14 [21]	−0.86 (210)	0.53 (62)
$CaTiO_3$[4]	20	26	0.100	218	9	5	0	19-28 [312]	−0.78 (44)	−0.78 (53)
CdSe[1]	4	29	0.149	111	7	5	1	9-21 [26]	−0.45 (68)	−0.40 (37)
CeO_2[1]	12	27	0.119	35	4	4	2	11-18 [19]	−0.21 (120)	−0.24 (97)
NaCl[5]	8	19	0.161	178	6	5	6	11-17 [11]	−1.50 (79)	−0.90 (65)
Ni[1]	4	27	0.110	30	4	4	4	15-20 [15]	−0.22 (92)	−1.1 (11)
PbS[6]	8	28	0.086	15	5	4	5	10-14 [10]	−1.40 (84)	−1.40 (78)
PbTe[1]	8	26	0.073	71	4	4	4	8-13 [8]	−1.4 (11)	−1.1 (8)
PbTe NP[7]	$\gg$ 100	28		426	10	7		27-53 [–]	−0.88 (46)	−0.50 (46)
Si[1]	8	27	0.202	52	6	4	6	12-20 [13]	−0.87 (57)	−0.75 (57)
$SrTiO_3$[8]	5	26	0.167	70	5	4	2	16-22 [23]	−1.10 (57)	−1.00 (64)
TiO_2 rutile[9]	6	26	0.164	104	7	5	0	15-25 [100]	−1.1 (8)	−0.40 (74)
Zn[1]	2	32	0.105	34	6	4	6	12-21[30]	−0.66 (59)	−0.77 (50)
ZnS sphalerite[10]	8	24	0.103	40	7	5	7	12-19 [13]	−0.41 (89)	−0.60 (52)
ZnS wurtzite[11]	4	26.5	0.196	20	5	4	5	12-18 [35]	−0.80 (58)	−0.49 (38)

[1] Wyckoff (1963). [2] Megaw (1962). [3] Truncated icosahedron with nearest-neighbor distance 1.44 Å. [4] Sasaki *et al.* (1987). [5] Jurgens *et al.* (2000). [6] Ramsdell (1925). [7] NP = nanoparticle. Precise structure unknown. For comparison, bulk PbTe has 32 distinct peaks within 20 Å. [8] Mitchell *et al.* (2000). [9] Meagher & Lager (1979). [10] Skinner (1961). [11] Wyckoff (1963). PDF refined as a mixture of wurtzite and sphalerite phases.

determined directly. We also consider whether each $M_{\rm best}$ model is selected assuming δG's where that model is statistically implausible according to the reduced chi-square statistic $\chi^2_{\rm red} = \chi^2/K$ for K degrees of freedom. Reduced chi-square has expectation value 1 and variance $2/K$, and as K increases it approaches the normal distribution and the regime where the familiar rule of thumb that $\chi^2_{\rm red} \sim 1$ is reasonable. Rather than ponder the subtle issues of reduced chi-square in model analysis (*e.g.* the distribution assumes the 'true' model, fitting alters the chi-square distribution's assumption that the residuals are normally distributed, the effective degrees of freedom for nonlinear models are non-obvious) we take a qualitative approach. Namely, the test statistic χ^2 calculated from a plausible model with k parameters fit to the Nyquist-sampled PDF with n data points should, over repeated experiments, be distributed approximately as the chi-square distribution with $K = n - k$ degrees of freedom. If the observed test statistic occurs at the fringes of the distribution the assumption of plausibility is likely false, and otherwise the statistic is consistent with (but not necessarily indicative of) a plausible model. The goal is to identify implausible models that nevertheless have significant Akaike weight due to the relative nature of AIC, so plausible models should be rejected by random chance infrequently. Define a region consistent with plausibility

$$\frac{\Phi^{-1}(0.00135)}{n-k} \le \chi^2_{\rm red} \le \frac{\Phi^{-1}(1-0.00135)}{n-k}, \qquad (11)$$

where Φ^{-1} is the inverse cumulative distribution function of the chi-square distribution. This region includes 99.73% of the distribution and rejects a plausible model by chance once every ~370 tests (*cf.* 500 ParSCAPE trials), and in the asymptotic limit defines a region of approximately ±3 standard deviations about the mean. We consider a model selected by AIC implausible if the test statistic falls outside this range, and examine whether or not the $M_{\rm best}$ models are implausible for the δG where they are selected as best. Helpfully, if δG is unknown but at least some models are plausible, this test can narrow the region of uncertainties reasonably considered physical.

We also attempt structure solution for all $M_{\rm best}$ models of C_{60} and the 16 crystal samples with ten trials of the Liga algorithm, which performs geometric build-up from pair distances (Juhás *et al.*, 2006, 2008). Peak extraction for these purposes is considered successful if the correct structure is found for at least one Liga trial. Structure solution is considered correct if, with respect to a reference structure obtained from the Crystallography Open Database (Gražulis *et al.*, 2009), it has the same nearest-neighbor coordination and no position offset by more than 0.3 Å. Liga trials for the crystal structures replicate the method of Juhás *et al.* (2010), which attempts structure solution within the simple [111] cell assuming known lattice parameters and stoichiometry, followed by a downhill method which assigns chemical species to each atom in the structure geometry determined by Liga. That study used the same crystal data, but peak extraction was performed by a very early precursor to ParSCAPE that lacked robust multimodeling capabilities and careful consideration of PDF uncertainty and sampling. Furthermore, termination ripples were not modeled, so even clearly spurious peaks required manual removal before running Liga.

Finally, consistency of ParSCAPE and multimodeling results with respect to δg, the uncertainty assumed when running ParSCAPE, and δG, the uncertainty assumed during multimodel analysis, are tested for two cases. In the *forward* case fix δg and for the resultant model calculate δG at which it attains its maximum Akaike weight. In the *backward* case fix δG and find the δg which produced the model with greatest Akaike weight. The forward case therefore measures how consistent the results of individual ParSCAPE trials are with their assumptions, while the backward case measures how consistent the best models are with the ParSCAPE trials that produce them.

Table 1 summarizes results from all 18 PDFs with unknown δG. For the purposes of discussion we categorize the PDFs into those with qualitatively low, moderate and high overlap. 'Correctness' of models in this section is with respect to results of structure solution to a modest tolerance, and not the intrinsic peaks.

The six samples with least peak overlap have f.c.c. (Ag, Ni), rock-salt (NaCl, PbS, PbTe) and diamond (Si) crystal structure. No hidden peaks are present in Ag, Ni, PbS and PbTe, while the others have only one or two. In all these cases at least one of the $M_{\rm best}$ models is correct, with the exception of Si due to a hidden peak near 9.4 Å displaced towards the mean of the combined feature by ~0.2 Å. The correct position of this peak is found for trials assuming small ($<1\%$) uncertainties, but these models have extraneous peaks and insignificant Akaike weight. The $M_{\rm best}$ model at greatest δG omits this peak entirely but is otherwise correct. The incorrect position is found in the remaining five $M_{\rm best}$ models, but in these cases the position is correct until the resampling step of ParSCAPE, indicating the Nyquist-sampled PDF and our peak function insufficiently constrain this aspect of the model. The width of the incorrect peak is somewhat greater than normal for this PDF, and reducing the peak function's maximum allowed width nearer to that of the other extracted peaks recovers the correct structure. Liga solution was considerably tolerant of spurious peaks for these six PDFs, obtaining the correct structure from every $M_{\rm best}$ model.

Nearly every $M_{\rm best}$ model was consistent with plausibility according to our $\chi^2_{\rm red}$ criterion. Three models selected at low δG, one for PbS and two for Si, are well outside the limits, with multiple spurious peaks yet $\chi^2_{\rm red}$ indicative of severe underfitting. Here $\chi^2_{\rm red}$ is complementary to AIC, allowing us to reject these models despite ignorance of the experimental δG. Curiously, the correct model for Ag was selected where $\chi^2_{\rm red}$ indicates overfitting, although the peaks are so distinct no simpler model could reasonably exist. However, the Akaike weight for this model is >0.2 for the vast majority of δG's where $\chi^2_{\rm red}$ indicates plausibility, so the correct model would almost certainly not be rejected by the AIC-based multimodeling approach.

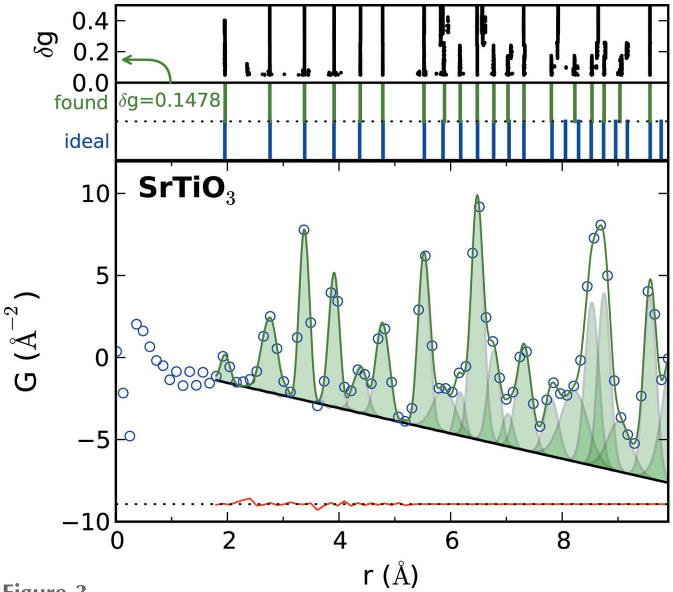


Figure 3
Experimental Nyquist-sampled SrTiO$_3$ X-ray PDF with $Q_{\rm max} = 26$ Å^{-1} showing extracted peaks of the $M_{\rm best}$ model from δG = 0.1961–0.2497 Å^{-2} (maximum Akaike weight ~0.1) generated assuming $\delta g = 0.1478$ Å^{-2}. The top inset shows extracted peak positions from every ParSCAPE trial as a function of δg. The inset of vertical lines compares extracted to ideal peak positions. The bottom line is the difference between the observed PDF and this model (offset for clarity). ParSCAPE ably handles the difficult feature between 8 and 9 Å, though several peaks can only be discerned in combination with a neighbor. Liga obtains the correct structure from these extracted peaks. The methods of §7.3 underestimate the range of physically plausible δG for this sample, which is shown in §7.4 to be about 0.588 Å^{-2}.

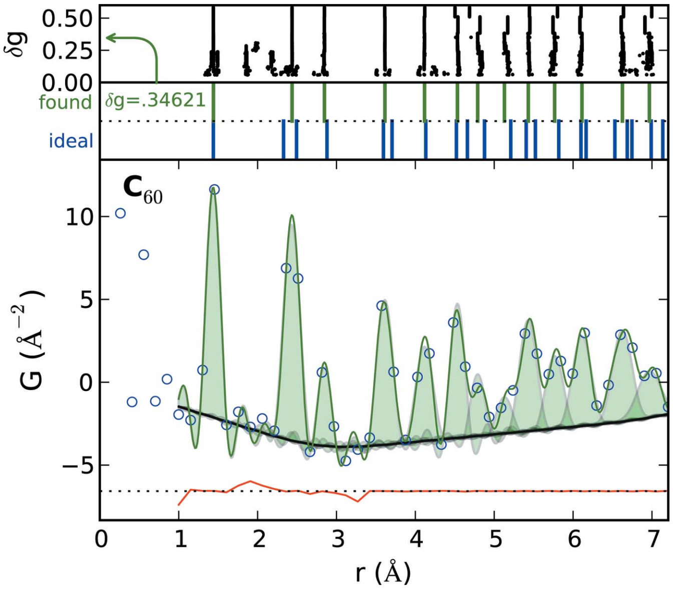


Figure 4
Experimental Nyquist-sampled C$_{60}$ X-ray PDF, $Q_{\rm max} = 21.3$ Å^{-1}, showing extracted peaks of the $M_{\rm best}$ model from δG = 0.26950–0.42188 Å^{-2} (maximum Akaike weight ~0.2, see class 57 in Fig. 5) generated assuming $\delta g = 0.34621$ Å^{-2}. The top inset shows extracted peak positions from every ParSCAPE trial as a function of δg. The inset of vertical lines compares extracted to ideal peak positions. The bottom line is the difference between the observed PDF and this model (offset for clarity). Although no hidden peaks are resolved, many pairs of peaks are so close together that fitting them separately is likely not justified, statistically or from resolution considerations. The prominent termination ripples are not extracted as peaks, but their mediocre fit compared to the rest of the data suggests systematic error, likely from baseline misspecification. Liga obtains the correct structure from these extracted peaks.

Structures with moderate peak overlap, up to a few dozen obscured peaks, include C (graphite), C_{60}, CdSe, CeO_2, $SrTiO_3$, Zn, ZnS (sphalerite) and ZnS (wurtzite). In all these structures some structural peaks remain unresolved, and as δg increases there is a trade-off between removing spurious peaks and resolving or retaining real ones. Many extracted peaks will have the combined contribution of multiple intrinsic peaks, and the degree to which these can be resolved depends on PDF quality but also the structural details of the sample. In fact, the correct peak model may require more parameters than Nyquist-sampled data points, so cannot be found by ParSCAPE even in principle.

Fig. 3 summarizes peak extraction from $SrTiO_3$, an undistorted (cubic unit cell) perovskite, and is an example where multiple hidden peaks can be resolved from high-quality PDF with moderate overlap. The topmost inset shows the clear (and very typical) progression of features which are removed as δg increases, with neighboring peaks shifting to compensate for the loss. In this case spurious peaks can be removed while resolving hidden peaks, particularly in the very difficult feature from 8 to 9 Å. The latter has a single local maximum with six intrinsic peaks, of which two are resolved directly, and the others as two pairs of underlying peaks. This sample is tested with accurate uncertainties in §7.4.

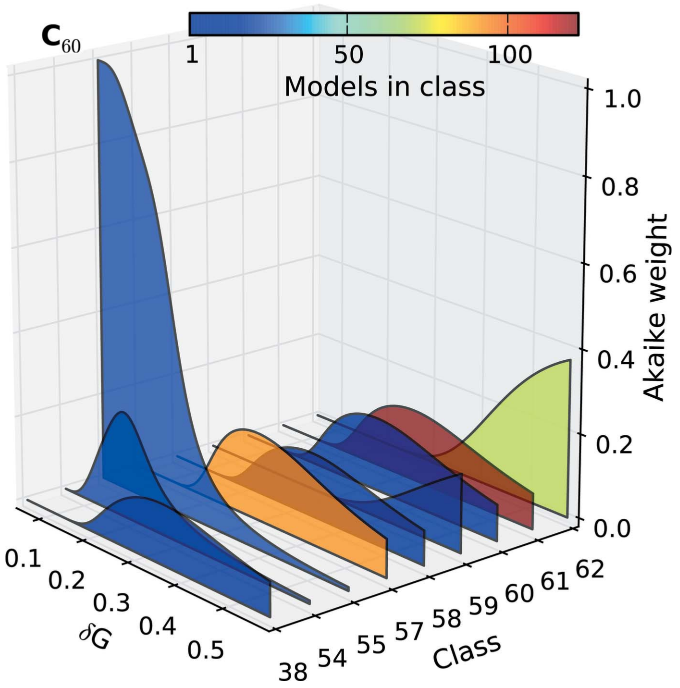


Figure 5
The Akaike weights for the best nine classes extracted from C_{60}, representing 329 of 500 ParSCAPE trials. The other 54 classes have total weight < 1% for all δG. The bell shape of weight *versus* δG reflects the AIC-enforced trade-off between chi-square error and the number of parameters given a fixed population of models. In this case the classes with 14 peaks reach maximum Akaike weight at low δG, those with 13 peaks at mid δG and those with 12 peaks at large δG. The basic results of multimodeling with ParSCAPE can be quickly determined by this graph, including the number of statistically plausible models and relative model strength. This particular plot is typical of PDFs with little evident overlap, as very few classes contribute nearly the total Akaike weight.

Fig. 4 summarizes peak extraction from C_{60}, in which hidden peaks cannot be resolved because the intrinsic peaks are so close and the prominent termination ripples are easily misidentified as peaks (roughly half of all trials). Nevertheless, of the nine classes of models with appreciable Akaike weight only two include these spurious peaks, comprising less than 10% of all trials, although these dominate the Akaike weights for low δG (Fig. 5). Five of these classes have 13 peaks, distinguished by various shifts in the peaks between 4.5 and 6.5 Å (several visible in the top inset of Fig. 4).

In other cases, tolerating a few spurious peaks to resolve possible hidden peaks may be an acceptable compromise, especially if the spurious peaks can be identified from physical considerations. This is the case for ZnS (wurtzite), for example, where one model has three small spurious peaks, but resolves three hidden peaks otherwise missed. The latter are suspect, of course, if the model has appreciable Akaike weight only for very small δG or if $\chi^2_{\rm red}$ suggests the model is misspecified. Examination of $\chi^2_{\rm red}$ shows that models selected at small δG were implausible for C (graphite), CdSe, $SrTiO_3$ and ZnS (sphalerite). The most interesting of these is the borderline high-overlap graphite, which exhibits poorly separated peaks above 5 Å. All its $M_{\rm best}$ models are likely

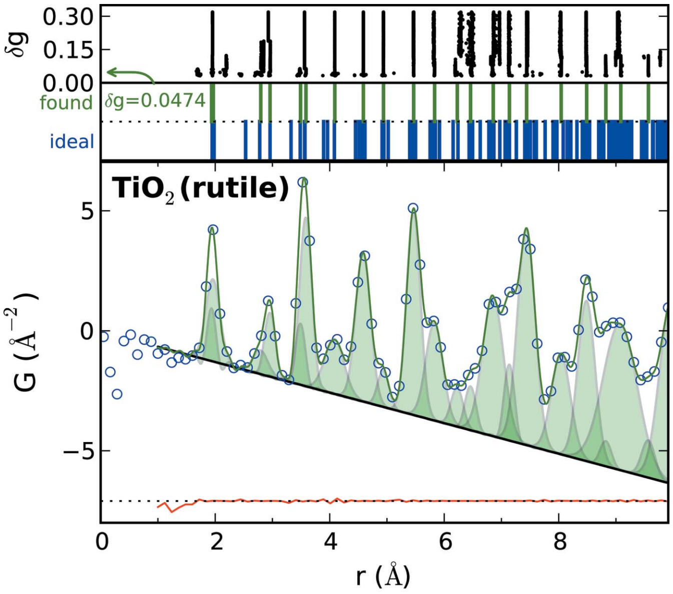


Figure 6
Experimental Nyquist-sampled TiO_2 X-ray PDF with $Q_{\rm max}$ = 26 Å^{-1} showing extracted peaks of $M_{\rm best}$ model from δG = 0.051–0.149 Å^{-2} (maximum Akaike weight ~0.38) generated assuming δg = 0.0474 Å^{-2}. Insets of vertical lines compare extracted to ideal peak positions. The bottom line is the difference between the observed PDF and sum of extracted peaks (offset for clarity). The top inset shows extracted peak positions of every ParSCAPE trial as a function of δg. Although this PDF is of high quality, and the extracted peaks appear to fit well, it has deceptively difficult aspects. The substantial peak overlap is not obvious by inspection, which may hinder the correct interpretation of extracted peaks. Fitting a single feature in the PDF with two nearly coincident peaks is a common ParSCAPE trait that may be accurate (as in this case where two ideal peaks 0.04 Å apart contribute to the observed nearest-neighbor peak), but can also be characteristic of inaccurate baseline estimation, indistinct termination ripples, or another effect causing apparent broadening. ParSCAPE does not find sufficient peaks for Liga solution, which is straightforward from ideal TiO_2 interatomic distances.

underfit, and the two with nearly implausible $\chi^2_{\rm red}$ exhibit clear unphysical attributes. The marginal quality of the sample even when fitting the correct structure (PDF fit residuum $R_w = 0.266$) suggests ParSCAPE cannot draw strong conclusions from this PDF. Despite these signs of misspecification, the model selected assuming δG $\sim$ 3% (Akaike weight $\sim$0.13) retained enough detail to yield graphite structure in seven of ten Liga trials. Both Zn and ZnS (wurtzite) favored implausibly overfit models assuming large δG, but as earlier these were either the simplest models found or had significant Akaike weight for δG where $\chi^2_{\rm red}$ is not implausible. Liga build-up was successful for all structures with moderate peak overlap.

The rutile TiO_2 (Fig. 6), PbTe nanoparticle (Fig. 7), and distorted perovskites $BaTiO_3$ and $CaTiO_3$ are examples of structures with substantial peak overlap. The degree of overlap may not be obvious, as for TiO_2, where a single extracted peak may fit many structural peaks, perhaps spanning tenths of Å's.

An inconsistency between ParSCAPE's assumptions and results after multimodeling is observed from the negative values of δg $-$ δG, typically about $-$0.9% in the forward case and $-$0.6% in the backward case, with standard deviations of comparable magnitude (Table 1). The sign of this trend indicates that individual ParSCAPE trials tend to miss statistically salutary features when δg = δG, and that the $M_{\rm best}$ models are usually drawn from these more complex models. This is not surprising because determining whether a feature exists or not is very difficult given preliminary peaks on very localized data, precisely the conditions early in a ParSCAPE trial, while multimodeling draws on a large population of fully specified alternate models. A typical example is shown in Fig. 8, where the decrease in complexity of the $M_{\rm best}$ models as uncertainty δG increases lags the decrease for individual trials as δg increases. Clearly an individual ParSCAPE trial, even at δg = δG, does not necessarily generate a model that compares favorably with the best models from a more diverse set. Therefore, we feel this study's approach of generating a diverse population of models (treating δg with some latitude) is a much stronger procedure. Combined with the other benefits of AIC-based multimodeling we believe this represents the best practice for ParSCAPE when δG is unknown.

7.4. Known δG

Fortunately, it has recently become possible to test the above procedure at experimentally determined δG by using *SrXplanar* to obtain the one-dimensional diffraction pattern. More importantly, the peak models best supported by the data can be directly assessed without considering the entire range of physically plausible δG. Finally, it permits investigating ParSCAPE's performance as different values of $Q_{\rm max}$ alter the resolution *versus* uncertainty trade-off.

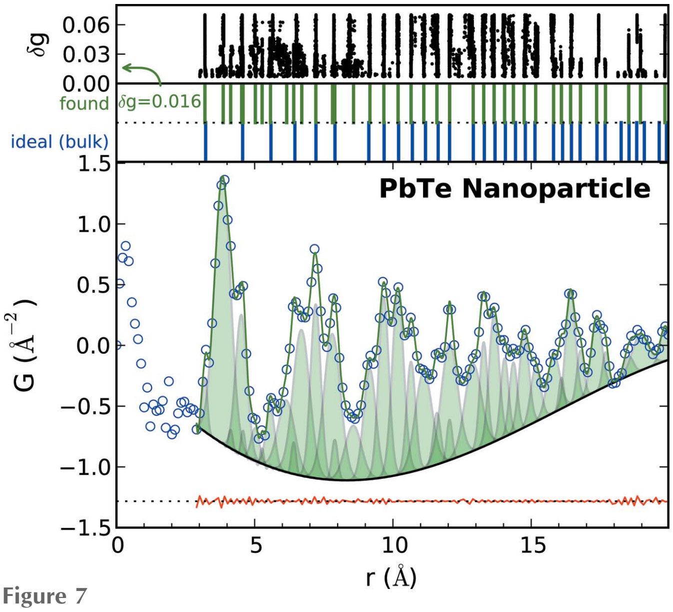


Figure 7
Experimental Nyquist-sampled PbTe nanoparticle X-ray PDF, $Q_{\rm max}$ = 28 Å^{-1}, showing extracted peaks of $M_{\rm best}$ model (δg = 0.01598 Å^{-2}). Its maximum Akaike weight is about 0.13, near δG = 0.028 Å^{-2}. Inset of vertical lines shows extracted peak positions compared to ideal positions of the bulk. The bottom line is the difference between observed PDF and sum of extracted peaks (offset for clarity). The top inset shows the peak positions of every model as a function of the δg which generated it. Although the bulk PbTe distances are identified in this challenging system, other large features suggest the presence of an additional phase. Very few of the 500 trials result in identical models, with over 400 classes identified. About 20 classes have non-negligible Akaike weight for some δG.

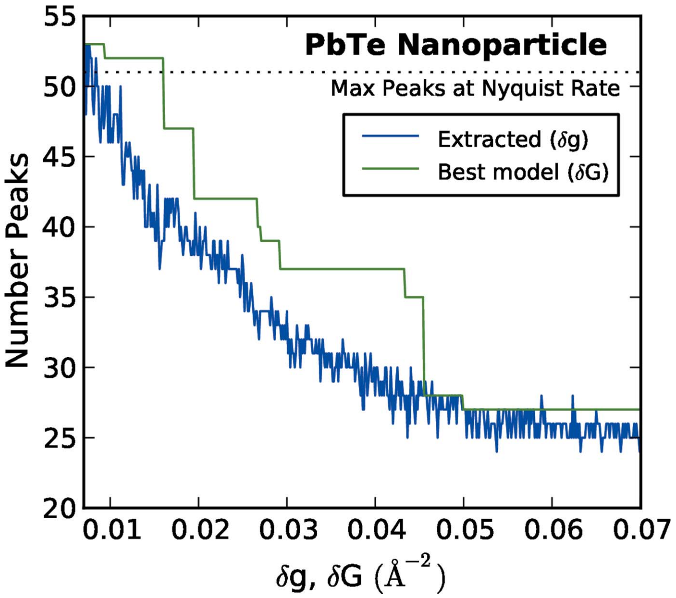


Figure 8
Comparison for PbTe nanoparticle of the number of peaks extracted by ParSCAPE assuming uncertainty δg to that of the best (greatest Akaike weight) model found by multimodel selection given δG. Both decrease as uncertainty increases, but clearly the best model for given δG tends to come from a ParSCAPE trial of lesser δg. The dotted line indicates the maximum number of peaks which can be fit given Nyquist-sampled data, one parameter per point. For small δg simultaneous pruning and downsampling stalls when removing peaks leads to no improvement in AIC before Nyquist sampling is achieved. In that case the model is fit using the least possible oversampling, $\sim$5% in this example, and the model should be considered more questionable than its AIC might indicate.

We generated a $SrTiO_3$ PDF with experimental δG (Q_{max} = 26 Å^{-1}) starting from the raw images used to obtain the PDF in the previous section, replicating the original data reduction procedure as closely as possible. These data are a 65 s exposure measured on a Marresearch Mar345 image-plate detector using 87 keV synchrotron X-rays at the 6ID-D beamline of the Advanced Photon Source, Argonne National Laboratory, USA. See Juhás *et al.* (2010) for further experimental details. PDFs were also obtained at six additional values of Q_{max} between 16 and 28 Å^{-1}. For each of these, 500 ParSCAPE trials were run using δg ranging from 0.070 to 0.921 Å^{-2}, approximately half the mean δG at Q_{max} = 16 Å^{-1} to 1.25 times mean δG at Q_{max} = 28 Å^{-1}, respectively. For given Q_{max} the δG in the range of extraction are reasonably constant, within 2% of the mean. Fig. 9 shows for all seven PDFs, over the region of significant overlap, peak positions from the models with greatest Akaike weight (evaluated using the procedure in §7.2) at the experimentally determined δG.

For Q_{max} = 26 Å^{-1}, mean δG = 0.588 Å^{-2}, about 20% above the maximum δg tested in §7.3. Consequently, it is unsurprising that the model with greatest Akaike weight ($\simeq$ 0.33) is less complex than the one in Fig. 3, which had greatest weight for δG near 0.25 Å^{-2}. From an AIC-based perspective, one should be reluctant to assign much value to the specific bump in, for example, the large feature near 8.75 Å. Is this a 'bad' fit for missing physical features? We don't have a truly independent way to answer that question, but the models favored if δG had actually been lower do capture additional physical features, suggesting that the AIC-based approach would capture them given more data collection. In addition, among information criteria of similar form to the AIC, we are aware of none with a smaller penalty term. If an information criterion approach is viable, as we believe, no alternate information criterion will consistently suggest these data can support even more complex models. In short, there is no replacement for more data.

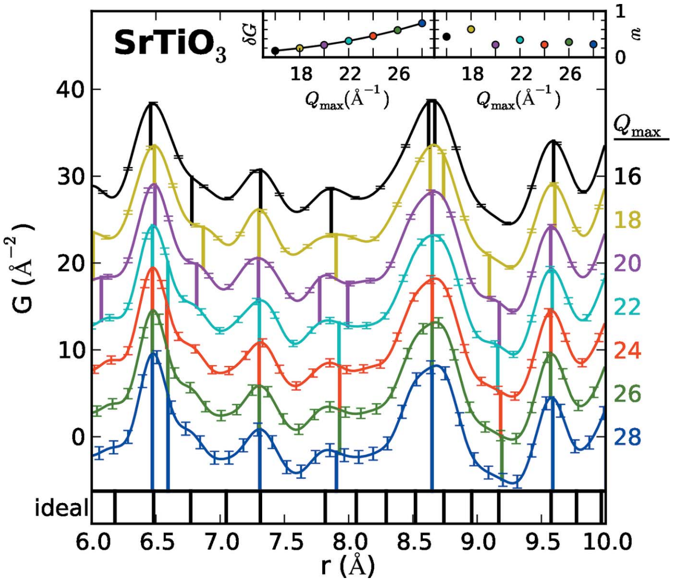


Figure 9
Detail of $SrTiO_3$ X-ray PDFs for various values of Q_{max} (offset vertically) with minimally correlated uncertainties propagated from the same raw diffraction data as the PDF in Fig. 3. The error bars appear at Nyquist-sampled positions. Vertical lines show the peak positions of the model, selected from 500 ParSCAPE trials, with greatest Akaike weight w (top-right inset) at the experimentally determined mean δG for the PDF of corresponding color. The bottom inset shows ideal peak positions. The top-center inset shows mean δG as a function of Q_{max}, which is well approximated by an empirical power law (plus a small constant) with exponent $\simeq$ 3.1. The models selected for Q_{max} = 16, 18 Å^{-1} have more free parameters than Nyquist-sampled data points, so those data do not truly support models of that complexity. For larger values of Q_{max} there are sufficient Nyquist-sampled data points for the models shown, but the uncertainties are too large to justify the models favored assuming smaller δG which, in fact, better match the ideal peak positions. For the models shown the trade-off between resolution and uncertainty appears best managed at Q_{max} = 20 Å^{-1}, which accurately identifies the most peaks.

As such, a reasonable question is how much additional data taking would be required to recover a model similar to the one in Fig. 3. The uncertainty of a Q_{max} = 26 Å^{-1} PDF created from the image which captured a 30 s subset of the data is 0.788 Å^{-2}. Increasing the exposure from 30 to 65 s thus reduced the uncertainty by a factor of roughly 0.75, close to the factor of 0.68 one might suspect if uncertainties were precisely Gaussian. (This scaling was independent of Q_{max}.) If additional data taking follows the same scaling properties, a factor of ten increase in collection time would be sufficient to reduce δG to 0.25 Å^{-2}, certainly within the time budget at a user facility for a sample with similarly strong scattering.

The impact of Q_{max} is also visible. At Q_{max} = 16, 18 Å^{-1} the uncertainties are small, but the only models sufficiently complex given these uncertainties require more parameters than Nyquist-sampled points. This is evidence that greater resolution is required for best results, despite the enviably low uncertainties. For large Q_{max}, physical features like the peak near 6.15 Å are lost because the peak does not, to the AIC, appear to add sufficient value. Without this peak, invariably the ones near 6.5 and 6.75 Å broaden, while the latter shifts (see the inset in Fig. 3). The best showing is Q_{max} = 20 Å^{-1}, which correctly identifies the most physical features. For this specific measurement it has sufficient resolution to support its model, yet small enough uncertainty to retain features which larger Q_{max} models lose. While investigation with more samples is required, these preliminary results suggest ParSCAPE reacts to Q_{max} in the anticipated fashion.

8. Key issues

The principal goal of ParSCAPE is unbiased peak extraction enabling structure determination without a prior structure model, yet the best model or models in the experimental PDF are rarely manifest. Searching for them is a challenging endeavor due to peak overlap, but also due to features inherent to the PDF measurement, principally the baseline and termination ripples. Consequently, evaluating alternate models is critical for prioritizing the initial conditions passed to a structure determination method. In addition, PDF science is in transition to a new degree of quantitative maturity, and the design of ParSCAPE anticipates features such as experimentally determined δG which are not yet ubiquitous. These topics are explored below.

8.1. The PDF baseline

Peaks extracted with ParSCAPE are conditioned on the estimate of the PDF baseline. Estimating the linear baseline of bulk crystal PDFs is generally straightforward, but nanoparticles pose significant difficulties. Since the baseline arises from unmeasured scattering below $Q_{\rm min}$ (Farrow & Billinge, 2009) a natural solution is direct determination with small-angle scattering experiments. At present, however, these data are not widely acquired in the nanoparticle PDF community. In their absence the best structural knowledge available should be used. The nanoparticle baseline can be computed analytically for simple shapes (Rayleigh, 1914; Glatter & Kratky, 1982; Müller *et al.*, 1996; Gilbert, 2008; Lei *et al.*, 2009) and *via* integral equations for more complex cases (Kodama *et al.*, 2006). Recent *ad hoc* techniques have also proven successful (Korsunskiy *et al.*, 2007). Direct calculation of $\gamma_0(r)$ from a structure model is a self-consistent approach in structural modeling, but usually not applicable to peak extraction.

Several simple baseline functions are included with *SrMise*, as well as arbitrary polynomials and interpolation from specified points, but the tools for estimating the baseline before peak extraction are very rudimentary. For crystals, a successful strategy appears to be fitting the first several peaks with unphysically large δg. Such trials take a few seconds to perform and capture the gross behavior of the (typically) most distinct peaks. These baselines can then be examined using more realistic δg to see if unlikely features appear. For example, if exactly one atomic pair is known to contribute to the nearest-neighbor peak, but multiple peaks are consistently found for that feature, this may be a sign the baseline should be adjusted slightly. On the other hand, if the investigator has reason to believe the nearest-neighbor peak is also strongly anharmonic, an extra peak may reflect that rather than a poor baseline.

At present the investigator should expect to craft a nanoparticle baseline manually. One approach will start from simple analytical models and estimates of nanoparticle properties (particularly size) and see if, for example, major peaks from a corresponding bulk structure are identified. Integrating both analytical and small-angle scattering nanoparticle baselines with robust estimation procedures is a major goal for a future version of *SrMise*.

8.2. Termination ripples

ParSCAPE aids the exercise of sound judgment in deciding whether a peak is real or a ripple. Techniques for handling termination ripples are a venerable topic in the PDF community (Lovell *et al.*, 1979; Warren & Mozzi, 1975; Warren, 1990), although current-generation beamlines with good, high-Q counting statistics reduce the need for methods to deal with low counting statistics. One common technique applies a smooth window function in reciprocal space near $Q_{\rm max}$. This damps the oscillation, but induces peak broadening, as from increased Debye–Waller factor. Another calculates the PDF for many values of $Q'_{\rm max} \le Q_{\rm max}$ and observes how features change. Modern structural refinement usually models termination ripples directly, and ParSCAPE adopts this approach.

ParSCAPE's clustering is inherently local, and suited for unimodal peaks, but poorer when satellite peaks are evident. Attempts to include termination ripples in descriptive peaks from the beginning of ParSCAPE were ineffective, since these introduce oscillations that may themselves be fit during recursive search. Furthermore, even if a physical peak is identified clearly, when the cluster grows to contain the ripples these features may be wrongly identified as the central peak of a separate feature also generating termination ripples, compounding the problem. In regions with sufficient peak overlap (usually at large r) the ripples will cancel to some degree, but introducing oscillations in the hopes they cancel later is misguided. Finally, using termination ripples from the start negatively affected both speed and model convergence.

Applying termination ripples to existing peaks followed by pruning was more effective in our tests, as this asks each peak whether it is a termination ripple, and gives its neighbors a chance to fit any vacated features in the data as evidence. Fig. 10 demonstrates this process for C_{60}. Simultaneously down-

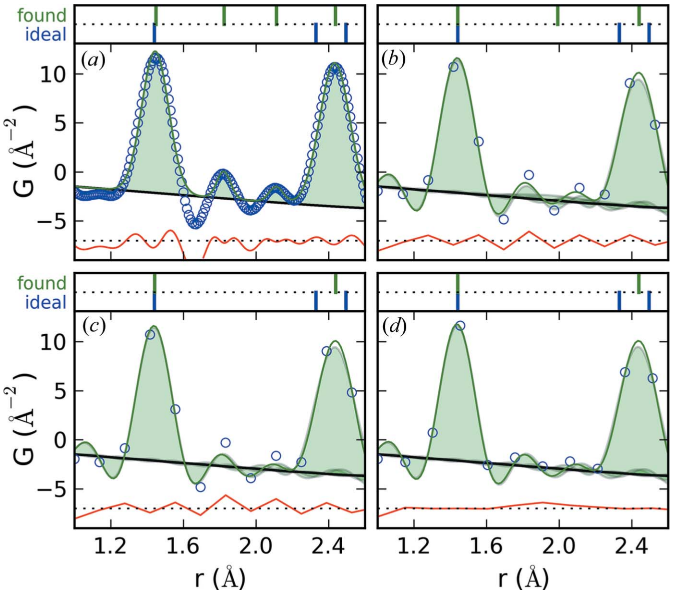


Figure 10
An example of termination ripple removal during the pruning/resampling step on C_{60} PDF, with comparison of distances and offset residual as in previous figures. (*a*) Immediately before the pruning/resampling step. The peak function does not yet model termination effects, and the ripples observed at approximately 1.8 and 2.1 Å are fit by distinct peaks. (*b*) After the first round of pruning. Termination effects have been introduced and the sampling rate is decreased. The data are still oversampled because the model in part (*a*) has more parameters than the Nyquist-sampled data permit. One extraneous peak has been removed, while the other shifts and broadens. (*c*) After the second round of pruning. The second extraneous peak has been removed, but the sampling rate is unchanged. The residual is increased compared to part (*b*), but not enough to offset the improvement in AIC due to three fewer parameters. (*d*) The final fit. Additional pruning removed no further peaks, and the PDF is now sampled at the Nyquist rate.

sampling the PDF during pruning yielded the greatest improvement. The principal reason is that some features are discarded as the number of points treated as independent decreases, causing the AIC to favor fewer parameters. Furthermore, model parameters do not need to 'escape' their previously converged values in order to fit slightly different features, as downsampling the PDF perturbs the optimal values so that the old values behave as a good initial guess. Evidence for the latter is order of magnitude reduction in termination ripple convergence problems (with or without pruning) compared to applying Nyquist sampling afterward. (Note that Nyquist sampling from the start of ParSCAPE retains these convergence problems once termination ripples are applied, further justifying the choice to oversample initially.)

Pruning is not always successful, but other aspects of ParSCAPE can provide insight as well. One way is examining how certain features are fit as δg changes. For example, if a feature is fit with its own peak at low δg, but well fit as the termination ripple of a nearby peak at larger δg, questioning the low-δg result is entirely justifiable. Similarly, the models found when varying $Q_{\max}$ may provide a more rigorous analog to the traditional method of visually inspecting changes to the peak profile. Multimodeling is flexible with regard to the investigator's judgment. If a model appears unphysical simply exclude it and recalculate the Akaike weights.

8.3. Negative peaks

The neutron PDF of systems containing elements with neutron structure factors of opposite sign will contain negative peaks. Although positive structure factors are more common, examples of negative structure factors include hydrogen and titanium. Even moderate peak overlap will cancel significant positive contributions to the PDF. Consequently, a model with an overlapping positive and negative peak becomes indistinguishable at that position from one with neither.

The usefulness of ParSCAPE in this context depends on the investigator's goals. Isolated peaks, both negative and positive, are trivial to extract. Furthermore, peak fitting starting from an existing model is straightforward. However, the principal goal of ParSCAPE, namely peak extraction over an extended range of the PDF starting from no model whatsoever, becomes daunting. Apart from the exceptions just noted, ParSCAPE assumes all peaks are positive, and there are no plans to address this issue algorithmically.

8.4. PDF sampling and uncertainty

The sampling rate and uncertainties in $G(r)$ are critical determinants of ParSCAPE's result because the AIC tolerates more parameters as the former increases, but fewer as the latter increases (Fig. 8). $Q_{\max}$, as discussed in §4.2, determines their balance. Approximate conditions on $Q_{\max}$ and the observation time necessary to optimize the uncertainties of $G(r)$, assuming each observation of $S(Q)$ has equal weight, were calculated by Thijsse (1984). ParSCAPE or another AIC-based technique may permit an independent check on this balance for an individual PDF with peak extraction over a range of maximum Q values. For a sufficiently complex PDF the best statistics presumably occur when the number of peaks that can be justified from the data is maximized. If the region were wide it might guide the investigator in generating a PDF with $Q_{\max}$ that emphasizes fine detail or lower uncertainties as desired. The results of §7.4 are promising in this regard, but we defer conclusions to a future study. In particular, even if ParSCAPE reliably identifies an 'optimal' $Q_{\max}$, it is not clear adopting that value should consistently and measurably benefit structure modeling.

An alternate application for future investigation is dynamic determination of sample exposure time during PDF acquisition *via* rapid multimodeling until a feature of interest can be justified among, or over, alternate models. Potential examples include extraction of a supposed hidden peak, a nearest-neighbor peak with shape indicative of nonharmonic interactions, or a given number of extracted peaks. This could be of particular value in time-intensive experiments, such as neutron scattering or temperature studies, where insufficient statistics are not easily rectified.

8.5. Applicability of the AIC

Careful practitioners of the AIC emphasize that models under consideration should be plausible *a priori* rather than arise from *post hoc* considerations, that multiple such models should be retained rather than choosing a single 'best' model, and that inference from *post hoc* models is meaningless (Burnham & Anderson, 2002). With its stepwise approach culminating in a single model, ParSCAPE might seem to violate these best practices, even 'data dredge'. These are legitimate concerns, arising primarily from confirmatory modeling in the biological sciences, but we feel they are largely mitigated due to the usual purpose and context of PDF peak extraction. First, the peak function (and frequently baseline) is based on well understood theory. Second, the total number of peaks is restricted to a small interval by the physically motivated clustering method and the experimentally estimated uncertainties in the data. Third, peak extraction without a structural model often has an exploratory rather than confirmatory character, with all due interpretive caution. Fourth, fitting individual features is frequently driven by strong *a priori* considerations, often in conjunction with a structural model. Finally, ParSCAPE lends itself to multimodel investigations of many plausible models, which we strongly encourage. If the AIC were applied to structural modeling, greater consideration of these cautions would be required.

An additional concern is that 'model' as defined in the theory considers only its functional form, not the parameters' refined values. From this perspective there is only one model with a given number of peaks (assuming the same peak function), not many with different refined values. This is a more subtle manifestation of model redundancy than that discussed in §6. The calculated AIC and its interpretation remain valid, as the approximations made in its derivation hold even for refined models with likelihoods 'close' to that of the maximum likelihood estimate (Burnham & Anderson, 2002). The Akaike weights, however, assume all models have equal prior likelihood of being the K–L best model, which is not true if models are redundant. Burnham & Anderson (2002), in a confirmatory paradigm, suggest using non-equal *a priori* weights to deal with this kind of redundancy if it is not removed outright. Whether such an approach is appropriate or necessary for ParSCAPE is under investigation.

9. Conclusion

While peak extraction and peak fitting of the PDF can provide useful information about individual peaks, the purpose of ParSCAPE is peak extraction over an extended range to aid methods of *ab initio* structure determination based on peak positions. ParSCAPE substantially reduces the user intervention required for this task and makes a statistically defensible estimate of sets of peaks that explain the measured PDF over a wide range of r in the absence of any structural model. This is expected to find use beyond structure solution from PDF.

To obtain an unbiased estimate of the best peak models the algorithm utilizes a multimodeling procedure based on the information-theoretic Akaike information criterion to balance goodness of fit and model complexity in a statistically justified fashion. In the spirit of the AIC, which is at heart a method for comparing competing models, the algorithm constructs classes of distinct models and scores them, allowing the user to select the best model, or any number of different preferred models, for further study, where by model we mean a set of Gaussian peak positions, widths and intensities.

ParSCAPE has been implemented in a software program, *SrMise*, available at http://www.diffpy.org. A scripting interface provides full access to *SrMise*, while single trials of peak extraction may be rapidly performed from the command line. The release includes examples demonstrating peak extraction and multimodeling with both crystals and nanoparticles. Nearly every parameter can, in principle, be determined or estimated from experiment, the notable exceptions being the PDF baseline in most cases (*i.e.* absent small-angle scattering data).

The PDF baseline must be specified before peak extraction, and results are conditioned upon it. Reasonable crystal baselines can be quickly estimated by trial and error, but nanoparticle baselines remain challenging. At this time *SrMise* supports parameterized nanoparticle baselines for a few simple shapes and arbitrary numerical baselines. Robust procedures for automated baseline modeling and estimation are a major goal for future versions.

The experimental uncertainty δG is, unfortunately, not commonly propagated to the PDF by popular data reduction software. Although new tools are changing this, 'legacy' PDFs will likely remain in use for some time. When δG is known, multimodeling analysis is straightforward. When it is not, all conclusions are premised upon an assumed δG, and the investigator's work increases. However, even in this situation, ParSCAPE can help the experimenter to rank competing models by estimating them systematically for different assumed uncertainty levels. Peak parameters found by ParSCAPE over a small range of assumed data uncertainties are largely observed to be stable to within the maximum likelihood estimate of their uncertainties. Finally, *ab initio* structure determination of 15 structures from peaks extracted without the use of δG demonstrates ParSCAPE can be a pragmatic tool even when statistical rigor is not possible.

We have also demonstrated ParSCAPE on a $SrTiO_3$ PDF with known δG. By our AIC-based method, the data best support a simpler model than those for which Liga construction was successful. A straightforward argument suggests such models would likely be favored (in this case) given an order of magnitude increase in exposure time. ParSCAPE also provides a new way to examine the resolution *versus* uncertainty trade-off imposed by $Q_{\max}$, one that may lead to future applications.

We believe that ParSCAPE can become a powerful tool for unbiased, model-free quantitative analysis of PDFs in the absence of a structural model and it has already been demonstrated in this regard (Terban *et al.*, 2015). It is easy to use and its success in providing sets of peaks that result in successful structure solutions using Liga demonstrates that it is robust.

APPENDIX *A*
Integrating detector uncertainties

Statistical uncertainties in PDF data come from various origins which depend on the measurement system. For many detectors such as photon counting detectors this is Poissonian, with initial uncertainty square-root the number of counts. Software commonly used during data reduction often assumes the latter, but this is invalid for integrating detectors such as CCDs, in which case the uncertainties determined this way are incorrect. Moreover, obtaining a one-dimensional powder diffraction pattern from area detectors requires integrating around Debye–Scherrer rings, the details of which affect the statistical correlations between neighboring bins in the one-dimensional pattern (Yang *et al.*, 2014). Data reduction and modeling software currently available do not typically utilize the full variance–covariance (VC) matrix, so area integration should be carried out so as to minimize statistical correlations between points. Nevertheless, the most common integrating software use a 'pixel-splitting' algorithm which introduces rather than suppresses correlations. For details see Yang *et al.* (2014).

Consequently, statistical uncertainties have been only rarely determined and propagated in powder diffraction from area detectors. For structural modeling, incorrect uncertainty magnitudes invalidate uncertainty estimates on model parameters, but have minimal impact on the refined values themselves. PDF reduction tools currently in development propagate the full VC matrix to $G(r)$, which in conjunction with *SrXplanar* will obviate these issues.

All data reduction in this paper was originally performed using *PDFgetX2* (Qiu *et al.*, 2004), which correctly propagates uncertainties (though not the VC matrix) given an input one-dimensional diffraction pattern with accurate uncertainties. When *SrXplanar* made the latter possible, we continued to use *PDFgetX2*, despite the availability of *PDFgetX3* (Juhás *et al.*, 2013), to best match the PDFs already tested by ParSCAPE.

Acknowledgements

We thank Dr Chris Farrow and Dr Pavol Juhás for productive conversations, particularly regarding the PDF baseline and uncertainties. We gratefully acknowledge the preceding as well as Dr Saurabh Gujarathi, Dr Peng Tian, Max Terban and Soham Banerjee for help testing *SrMise*. We appreciate the efforts of Dr Emil Božin, Dr Ahmad Masadeh and Dr Douglas Robinson for assistance with the X-ray measurements at the Advanced Photon Source, Argonne National Laboratory (APS, ANL). We thank Dr Christos Malliakas for providing the NaCl, PbTe nanoparticle and ZnS wurtzite samples, as well as Professor Harry Dorn for the C_{60} sample. Work at MSU was supported by the Center of Research Excellence in Complex Materials at Michigan State University. Work in the Billinge group was supported by the US Department of Energy, Office of Basic Energy Sciences (DOE–BES), as the complex modeling laboratory directed research and development grant at Brookhaven National Laboratory through contract DE-AC02-98CH10886. The Advanced Photon Source is supported by the US DOE–BES under contract No. W-31-109-Eng-38.

References

Akaike, H. (1973). *Second International Symposium on Information Theory*, pp. 267–281. Budapest: Academiai Kiado.

Anderson, D. R., Burnham, K. P. & White, G. C. (1994). *Ecology*, **75**, 1780–1793.

Andrae, R., Schulze-Hartung, T. & Melchior, P. (2010). arXiv: 1012.3754.

Arnold, T. W. (2010). *J. Wildl. Manag.* **74**, 1175–1178.

Billinge, S. J. L. & Levin, I. (2007). *Science*, **316**, 561–565.

Bock, H.-H. (1996). *Clustering and Classification*, edited by P. Arabie, L. J. Hubert & G. De Soete, pp. 377–453. Singapore: World Scientific Publishing Company.

Bozdogan, H. (1987). *Psychometrika*, **52**, 345–370.

Bozdogan, H. (2000). *J. Math. Psychol.* **44**, 62–91.

Božin, E. S., Malliakas, C. D., Souvatzis, P., Proffen, T., Spaldin, N. A., Kanatzidis, M. G. & Billinge, S. J. L. (2010). *Science*, **330**, 1660.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference.* New York, NY: Springer.

Claeskens, G. & Hjort, N. L. (2003). *J. Am. Stat. Assoc.* **98**, 900–916.

David, W. I. F. & Shankland, K. (2008). *Acta Cryst.* A**64**, 52–64.

David, W. I. F., Shankland, K., McCusker, L. B. & Baerlocher, C. (2002). *Structure Determination from Powder Diffraction Data.* Oxford: Oxford University Press.

Dimitrov, D. A., Röder, H. & Bishop, A. R. (2001). *Phys. Rev. B*, **64**, 14303.

Egami, T. & Billinge, S. J. L. (2012). *Underneath the Bragg Peaks: Structural Analysis of Complex Materials*, 2nd ed. Amsterdam: Elsevier.

Farrow, C. L. & Billinge, S. J. L. (2009). *Acta Cryst.* A**65**, 232–239.

Farrow, C. L., Shaw, M., Kim, H.-J., Juhás, P. & Billinge, S. J. L. (2011). *Phys. Rev. B*, **84**, 134105.

Gan, G., Ma, C. & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications* (ASA-SIAM Series on Statistics and Applied Probability). Philadelphia, Pennsylvania: SIAM and Alexandria, Virginia: ASA.

Gilbert, B. (2008). *J. Appl. Cryst.* **41**, 554–562.

Glatter, O. & Kratky, O. (1982). *Small-angle X-ray Scattering*, 1st ed. London: Academic Press Inc.

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 726–729.

Guinier, A., Fournet, G., Walker, C. & Yudowitch, K. (1955). *Small-angle Scattering of X-rays.* New York: John Wiley and Sons, Inc.

Heiney, P., Fischer, J., McGhie, A., Romanow, W., Denenstein, A., McCauley Jr, J., Smith, A. & Cox, D. (1991). *Phys. Rev. Lett.* **66**, 2911–2914.

Hurvich, C. M. & Tsai, C.-L. (1989). *Biometrika*, **76**, 297–307.

Juhás, P., Cherba, D. M., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. (2006). *Nature (London)*, **440**, 655–658.

Juhás, P., Davis, T., Farrow, C. L. & Billinge, S. J. L. (2013). *J. Appl. Cryst.* **46**, 560–566.

Juhás, P., Granlund, L., Duxbury, P. M., Punch, W. F. & Billinge, S. J. L. (2008). *Acta Cryst.* A**64**, 631–640.

Juhás, P., Granlund, L., Gujarathi, S. R., Duxbury, P. M. & Billinge, S. J. L. (2010). *J. Appl. Cryst.* **43**, 623–629.

Jurgens, B., Irran, E., Schneider, J. & Schnick, W. (2000). *Inorg. Chem.* **39**, 665–670.

Kodama, K., Iikubo, S., Taguchi, T. & Shamoto, S. (2006). *Acta Cryst.* A**62**, 444–453.

Korsunskiy, V. I., Neder, R. B., Hofmann, A., Dembski, S., Graf, C. & Rühl, E. (2007). *J. Appl. Cryst.* **40**, 975–985.

Kotz, S. & Johnson, N. L. (1992). *Breakthroughs in Statistics.* New York: Springer.

Kullback, S. & Leibler, R. A. (1951). *Ann. Math. Stat.* **22**, 79–86.

Le Bail, A., Duroy, H. & Fourquet, J. L. (1987). *Mater. Res. Bull.* **23**, 447–452.

Lebreton, J.-D., Burnham, K. P., Clobert, J. & Anderson, D. R. (1992). *Ecol. Monogr.* **62**, 67.

Lei, M., de Graff, A. M. R., Thorpe, M. F., Wells, S. A. & Sartbaeva, A. (2009). *Phys. Rev. B*, **80**, 024118.

Levashov, V. A., Billinge, S. J. L. & Thorpe, M. F. (2007). *J. Comput. Chem.* **28**, 1865–1882.

Liddle, A. (2007). *Mon. Not. Royal Astron. Soc. Lett.* **377**, L74–L78.

Lovell, R., Mitchell, G. R. & Windle, A. H. (1979). *Acta Cryst.* A**35**, 598–603.

Ma, D., Stoica, A. D. & Wang, X. (2009). *Nature Mater.* **8**, 1–5.

McCusker, L. B., Von Dreele, R. B., Cox, D. E., Louër, D. & Scardi, P. (1999). *J. Appl. Cryst.* **32**, 36–50.

McQuarrie, A. D. R. (1998). *Regression and Time Series Model Selection.* Singapore: World Scientific Publishing Company.

Meagher, E. P. & Lager, G. A. (1979). *Can. Mineral.* **17**, 77–85.

Megaw, H. D. (1962). *Acta Cryst.* **15**, 972–973.

Mitchell, R. H., Chakhmouradian, A. R. & Woodward, P. M. (2000). *Phys. Chem. Miner.* **27**, 583–589.

Morohashi, M., Shimizu, K., Ohashi, Y., Abe, J., Mori, H., Tomita, M. & Soga, T. (2007). *J. Chromatogr. A*, **1159**, 142–148.

Mullen, K. & Levin, I. (2011). *J. Appl. Cryst.* **44**, 788–797.

Müller, J. J., Hansen, S. & Pürschel, H.-V. (1996). *J. Appl. Cryst.* **29**, 547–554.

Pawley, G. S. (1981). *J. Appl. Cryst.* **14**, 357–361.

Petkov, V., Trikalitis, P. N., Božin, E. S., Billinge, S. J. L., Vogt, T. & Kanatzidis, M. G. (2002). *J. Am. Chem. Soc.* **124**, 10157.

Qiu, X., Thompson, J. W. & Billinge, S. J. L. (2004). *J. Appl. Cryst.* **37**, 678.

Ramsdell, L. S. (1925). *Am. Mineral.* **10**, 281–304.

Rayleigh, L. (1914). *Proc. R. Soc. London Ser. A*, **90**, 219.

Rodriguez, A. & Laio, A. (2014). *Science*, **344**, 1492–1496.

Sasaki, S., Prewitt, C. T., Bass, J. D. & Schulze, W. A. (1987). *Acta Cryst.* C**43**, 1668–1674.

Schwarz, G. (1978). *Ann. Stat.* **6**, 461–464.

Shannon, C. E. (1949). *Proc. IRE*, **37**, 10–21.

Shao, Q. & Wu, Y. (2005). *J. Stat. Plan. Inference*, **135**, 461–476.

Skinner, B. J. (1961). *Am. Mineral.* **46**, 1399–1411.

Spiegelhalter, D. J. (2002). *J. R. Stat. Soc. B*, **93**, 120–639.

Stoica, P. & Sel, Y. (2004). *IEEE Signal Process. Mag.* **21**, 36–47.

Stone, M. (1977). *J. R. Stat. Soc. B*, **39**, 44–47.

Sugiura, N. (1978). *Commun. Stat. Theory Methods*, **7**, 13–26.

Takeuchi, K. (1976). *Suri Kagaku*, **153**, 12–18.

Terban, M. W., Johnson, M., Di Michiel, M. & Billinge, S. J. L. (2015). *Nanoscale*, **7**, 5480–5487. doi:10.1039/C4NR06486K.

Thijsse, B. J. (1984). *J. Appl. Cryst.* **17**, 61–76.

Thorpe, M. F., Levashov, V. A., Lei, M. & Billinge, S. J. L. (2002). *From Semiconductors to Proteins: Beyond the Average Structure*, edited by S. J. L. Billinge & M. F. Thorpe, pp. 105–128. New York: Kluwer/Plenum.

Tibshirani, R. (2001). *J. R. Stat. Soc. B*, **63**, 411–423.

Toby, B. H. & Billinge, S. J. L. (2004). *Acta Cryst.* A**60**, 315–317.

Transtrum, M. K., Machta, B. B. & Sethna, J. P. (2010). *Phys. Rev. Lett.* **104**, 060201.

Warren, B. E. (1934). *J. Phys. Chem.* **2**, 551.

Warren, B. E. (1990). *X-ray Diffraction.* New York: Dover.

Warren, B. E. & Mozzi, R. L. (1975). *J. Appl. Cryst.* **8**, 674–677.

Wei, H. (2010). *J. Cosmol. Astropart. Phys.* **08**, 020.

Wright, A. C. (1998). *Glass Phys. Chem.* **24**, 148–179.

Wyckoff, R. W. G. (1963). *Crystal Structures*, Vol. 1, 2nd ed. New York: Wiley,

Yang, X., Juhás, P. & Billinge, S. J. L. (2014). *J. Appl. Cryst.* **47**, 1273–1283.

Zhao, Z., Zhang, Y. & Liao, H. (2008). *Eng. Appl. Artif. Intell.* **21**, 1182–1188.