Chapter 6

Extracting Structural Information from the PDF

Chapter Outline

6.1. INTRODUCTION

When the full corrections are carried out, the experimentally derived PDF is an absolute function: the data have been properly normalized and the absolute, rather than the relative, intensities of the peaks are meaningful. Even with *ad hoc* corrections (Chapter 5), the peak positions, widths, and relative intensities are correct. A great deal of structural information can therefore be deduced directly from the data without resorting to models. Ultimately, the most information is obtained from the PDF by structural modeling: calculating the PDF from structural models and comparing them to the data. The structural origin of different features of the PDF function can be understood by careful consideration of the equations which define the PDF (Section 6.3.1). This is explicitly described in Section 6.2. Different approaches for modeling the data using regression techniques are described in Section 6.3. Finally, in Section 6.4, we describe recent developments in the use of PDF data for *ab initio* structure solution.

Pergamon Materials Series, Vol. 16. http://dx.doi.org/10.1016/B978-0-08-097133-9.00006-X

6.2. DIRECT INFORMATION

6.2.1. The PDF from a Structure

The relationship between an atomic structure (i.e., an arrangement of atoms in space) and the PDF was discussed in Chapter 3. We reiterate this argument here since it is highly pertinent to the discussion of obtaining structural information from the PDF. First, assume we have a model for our sample that consists of a set of N atoms at positions $\mathbf{r}_n$ with respect to some origin. Intuitively, we obtain the PDF in the following way. We first choose an atom at random and place the origin of our space at the position of that atom. We then systematically find every other atom in the sample and measure the distance from the origin atom to that atom. Each time we find an atom, we place a unit of intensity at the position $r_m = |\mathbf{r}_m|$ on the axis of our function $R(r)$. We continue this until we find every atom in the sample. We then move the origin of our space to another atom and repeat the process adding intensity to the same $R(r)$ function. This is systematically repeated until every atom in the sample has had its turn at the origin. To keep $R(r)$ as an intrinsic function (independent of sample size), it is divided by N to normalize it. If all the atoms in the sample are of the same chemical species, this is the end of the story. To take into account the different scattering powers of different chemical species, we multiply the unit of intensity for each atom pair by $b_m b_n / \langle b \rangle^2$ where b_i is the scattering length of the ith atom. Mathematically, this can be expressed as in Eq. (3.5):

$$R(r) = \frac{1}{N} \sum_{mn} \frac{b_m b_n}{\langle b \rangle^2} \delta(r - (r_n - r_m)). \tag{6.1}$$

To understand how this works in practice, it is useful to consider a simple example. Let us consider a crystalline single-element material; for example, the neutron scatterers' favorite, nickel. A typical neutron sample is a fine powder which can be as much as 10 g. Such a sample contains $\sim 10^{23}$ atoms. To properly calculate $R(r)$, we would therefore need to carry out a double sum over these many atoms, which is clearly impractical. Two things help us in practice. First, we are generally only interested in calculating $R(r)$ over a relatively narrow range or r, say 20 Å. Thus, we still need to put the origin on each of the 10^{23} atoms in turn, but the second sum need only be taken over atoms that lie within 20 Å of the origin atom. Second, the material in question is crystalline. In this case, the total sample is made up of many equivalent unit cells which are periodically repeated in space. In this favorable situation, we need only to place the origin of each atom in the unit cell since the equivalent atom in all the other unit cells has exactly the same atomic environment. This is now a computationally tractable problem: a double sum where the first sum has taken over the atoms in the unit cell (< 100 typically) and the second sum over all atoms within $r_{\max}$ of the origin atom where $r_{\max}$ is the maximum extent over which the PDF is to be calculated.

6.2.2. Direct Information from the PDF

6.2.2.1. Atom-Pair Separation from Peak Positions

It is clear from this description that the PDF is a heavily averaged representation of the structure. First, directional information is lost since we consider only $r_n - r_m$ and not $\mathbf{r}_n - \mathbf{r}_m$. Second, it is a linear superposition of the local environments of many atoms, more than 10^{20} in fact. How can such a function contain any useful information at all? The reason is that, especially on very short-length scales, the possible environments of particular atoms are very limited. In Ni, for example, all the atoms have the same nearest-neighbor distance, r_{nn}, if we neglect thermal vibration. There will be no intensity in $R(r)$ for $r < r_{nn}$ and a sharp peak at r_{nn}. This behavior is very general and true even in atomically disordered systems such as glasses, liquids, and gasses. The second neighbor distance is generally less well defined, the PDF peak will be broader, but will still be apparent even in disordered materials. In crystals, because of the long-range order of the structure, all neighbors at all lengths are well defined and give rise to sharp PDF peaks. The position of these peaks gives the separations of pairs of atoms in the structure directly.

6.2.2.2. Coordination Number from Peak Integrated Intensity

It is clear from Eq. (3.5) that if a well-defined PDF peak can be observed, we can determine the coordination number of the origin atom by integrating the intensity under that peak. The correlation function which yields the coordination number directly by integration is $R(r) = 4\pi r^2 \rho(r)$. This is related to $G(r)$, the function obtained directly from the Fourier transform of the data, by $R(r) = r[G(r) + 4\pi r \rho_0]$. In the case of crystalline Ni, there are four Ni atoms in the unit cell (fcc structure). Each nickel ion has 12 neighbors at 2.49 Å (Wyckoff, 1963). When we construct our PDF, we will therefore place 48 units of intensity at position $r = 2.49$ Å (the weighting factor, $b_m b_n / \langle b \rangle^2$, is unity since there is only one kind of scatterer) and divide by $N = 4$ since we put four atoms, respectively, at the origin. Thus, integrating the first peak will yield 12 which is the coordination number of Ni. The same information can be obtained from multielement samples if the chemical origin of the PDF peak, and therefore the weighting factor, is known. If, as is often the case, PDF peaks from different origins overlap, this process is complicated. Information can be extracted by measuring the chemical-specific differential or partial PDFs directly (see Chapter 3) or with less certainty by fitting the peaks with a series of Gaussian functions. As we discuss later, full-scale structural modeling largely overcomes the problem of PDF peak overlap, at least in crystalline materials.

6.2.2.3. Atom-Pair Probability Distribution from the Peak Shape

The PDF is made up of a sum of well-defined delta functions. In the real material (and therefore in the experimentally derived PDF), the sums are

taken over the entire sample. Atomic disorder in the form of thermal and zero-point motion of atoms and any static displacements of atoms away from ideal lattice sites give rise to a distribution of atom–atom distances. The PDF peaks are therefore broadened resulting in Gaussian-shaped peaks.[1]

The width, and shape, of the PDF peaks therefore contains information about the real atomic probability distribution. In the case where a probability distribution is non-Gaussian, for example, if the atomic potential is multi-welled and atoms in the solid occupy each well in some disordered fashion, the PDF will reflect this and, in principle, this information can also be extracted from the PDF by an analysis of the PDF peak shape.

To summarize, there are three independent pieces of information which are contained in a PDF peak: its position gives the average separation of the pair of atoms in question, its integrated intensity yields the coordination number of that pair of atoms, and the width and shape of the peak gives the underlying atomic probability distribution.

6.2.3. Examples

6.2.3.1. Peak Position

The peak position yields bond lengths directly. One clear example where this has proved to be useful is in understanding the local atomic structure of semiconductor alloys, as introduced in Section 1.2.1. In $In_{1-x}Ga_xAs$, the metal-ion sublattice contains a random solid solution of indium and gallium atoms which have very different covalent radii. The material on the whole forms a long-range-ordered crystal structure. However, the In/Ga bond length obtained from the average crystal structure bears little resemblance to the real, local, bond lengths in the material: locally, In–As and Ga–As pairs have bonds that are close to their length in the end-member compounds. The two distinct bond lengths (In–As and Ga–As) are different by $\sim$0.14 Å and can be resolved in a high-resolution PDF measurement. Then, by fitting two Gaussian functions (suitably convoluted with a Sinc function to account for termination effects in the Fourier transform), the evolution of the local bond lengths with alloy composition could be directly elucidated (Petkov *et al.*, 1999). This is shown in the so-called *Z*-plot in Fig. 1.3. The inset shows the nearest-neighbor PDF peak which is clearly resolved into two features. The Gaussian fits are also shown. The main panel shows the evolution of the In–As and Ga–As bond lengths with alloy composition obtained from these peak fits. Also shown are earlier results of EXAFS (Mikkelson and Boyce, 1982) who were the first to observe this phenomenon experimentally. There is excellent agreement between the EXAFS and the PDF data. Beyond the *Z*-plot, the PDF data have allowed different structural models for the local

1. The PDF peaks, even from Gaussian atomic probability distributions, are actually slightly non-Gaussian. However, the deviations from Gaussian are negligibly small as discussed in Section 6.4.

structure of the semiconductor alloy to be differentiated (Jeong *et al.*, 2001; Peterson *et al.*, 2001).

Another elegant example of the use of the PDF to measure local bond lengths directly was demonstrated by Dove *et al.* (1997) in crystalline phases of silica and introduced in Section 1.2.5. The low-temperature crystalline form of silica is α-quartz. It is made up of corner-shared SiO_4 tetrahedra with a physically reasonable Si–O bond length of 1.61 Å and a Si–O–Si bond angle of 144°. As temperature increases, the Si–O–Si bond angle increases and the Si–O bond length decreases. The sample then transforms first into *β*-quartz, HP-tridymite, and then *β*-cristobalite on further heating (Keen, 1998). The latter two phases both have Si–O–Si bond angles of 180°, known to be chemically unfavorable, and an unphysical Si–O bond length of 1.54 Å. Large thermal factors are also seen in the crystal structure refinements. An analysis of the PDF from these materials immediately shows (Fig. 1.11) that the local Si—O and O—O bonds are significantly longer than those obtained from the average crystal structure. Furthermore, as a function of temperature, the Si–O bond length determined from the PDF increases very slightly even though the average value obtained crystallographically is smoothly decreasing as the phase transition to *β*-cristobalite is approached (Tucker *et al.*, 2000). As in the semiconductor alloys, in silica the O ions do not lie on lattice sites of the crystallographic model but are displaced away from them in a disordered way. Similar effects are seen in other materials, for example, AlF_3 (Chupas *et al.*, 2004). These examples illustrate convincingly, and in a model-independent way, that the PDF contains additional information beyond the average structure in crystalline materials and that this information can be reliably recovered experimentally. Detailed modeling, as we describe later in the chapter, reveals a great deal of additional information beyond this model-independent analysis.

6.2.3.2. *Integrated Intensity in PDF Peaks*

The integrated intensity under a PDF peak yields the coordination number of that atom–atom correlation (the number of neighbors at a specific distance). This type of analysis is widely used in studies of glasses (Waseda, 1980); however, we will give two examples from partially crystalline samples. The first example is from disordered nanoporous carbon. Some of the earliest ever PDF studies were carried out on carbon blacks by Warren (1934). There is currently renewed interest in disordered carbons because of their potential uses for storing lithium ion battery applications and as conducting nanoporous electrode materials in fuel cells, among many other interesting applications. A number of studies use the PDF method to study the local structure of disordered carbons (Claye and Fischer, 1999; Petkov *et al.*, 1999). In the latter study, nanoporous disordered carbon is made by pyrolyzing poly furfuryl alcohol by heating it in an inert atmosphere. This produces a carbonaceous

product which is more or less disordered depending on the temperature of pyrolysis. A PDF study showed that a 1200 °C treatment produced almost perfect graphene sheets; an 800 °C treatment introduced significant disorder into the sheets in the form of greater sheet fragmentation and non-6-membered rings. A 400 °C treatment resulted in highly distorted carbon planes bearing little resemblance to graphitic material and more resembling the alcohol starting material. The data PDFs are shown in Fig. 6.1. Two key pieces of evidence revealed the nature of the disorder in the 800 °C sample. First, the integrated intensity of the first C—C peak fell from having 3.0 neighbors in the 1200 °C sample to having 2.6 neighbors in the 800 °C sample. Second, the third PDF peak, coming from the C—C bond diametrically across the six-member ring in graphite, became broader and lost intensity in the 800 °C sample. This suggests the loss of, and an increased distribution of, third-neighbor correlations, at this position diagonally across the ring, a telltale indication for the presence of higher-membered rings in the structure (Fig. 6.1). The graphene sheets are thus becoming fragmented and higher neighbor rings are being introduced. In a TEM image, the sheets indeed appear more curved in this sample, a by-product of the presence of non-6-member rings. The near-neighbor C—C peak in the 400 °C sample when integrated indicated 1.9 neighbors. This number is somewhat uncertain because this peak is not fully resolved in the structure; however, it is a clear indication that the graphene sheets are highly fragmented and even that the structure is closer to the polymeric starting material (two carbon–carbon near neighbors) than the graphitic product. Again, additional information was

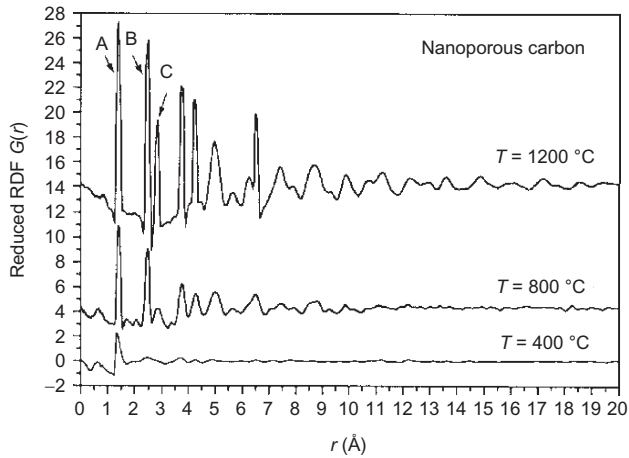

FIGURE 6.1 PDFs from disordered carbon made by pyrolyzing poly furfuryl alcohol at different temperatures (Petkov *et al.*, 1999). These PDFs were created by Fourier transforming the $S(Q)$ data shown in Fig. 9.16b.

elucidated from modeling, but significant knowledge of the structure could be deduced directly from the PDF.

6.2.3.3. PDF Peak Width

The width of PDF peaks reveals information about the static and dynamic disorder of atoms involved in the pair. Measuring the peak width as a function of temperature gives information about the Debye temperature of a bond; the width as a function of atomic separation yields information about correlated atom dynamics which, in turn, reveals information about the underlying atomic potential; the width as a function of doping gives information about doping-induced disorder; and so on.

The Debye model is a simple way to explain the PDF peak width with a one (or usually two) parameter model (Debye, 1912). This is the same Debye model used in undergraduate physics and chemistry textbooks for explaining the low-r behavior of the specific heat (Kittel, 1996). Thermal energy is absorbed into the material in the form of atomic vibrations by exciting phonons. There is a phonon spectrum (allowed energies of different modes) which results in a density of states that gives the number of available phonon modes at each energy. These modes are then populated statistically governed by a Boltzmann relation. If the phonon spectrum is known, and any anharmonic effects are unimportant, then the dynamics of the atoms and the PDF peak widths can be calculated explicitly. The phonon spectrum can be determined from an inelastic neutron scattering measurement. However, it is rare that this is done before the powder diffraction and PDF work is carried out, if at all. From a practical perspective, there are virtually no studies where PDFs are calculated using the whole known lattice dynamics. The Debye model provides a useful first step in this situation. A simple form is assumed for the phonon density of states (quadratically increasing up to a cutoff that depends on the Debye temperature), and the Debye–Waller factors and uncorrelated PDF peak widths, which are appropriate for the high-r region of the PDF, are determined at any given temperature (Willis and Pryor, 1975; Billinge *et al*., 1991). This model shows how, as a function of temperature, the high-r PDF peaks should broaden. The behavior is shown in Fig. 6.2 (Levashov *et al*., 2007) (the Debye model curve is the "QM Harmonic" line in the figure). In the high-temperature region, the PDF peaks broaden linearly with temperature. This is the behavior that is expected for a classical harmonic solid and this line should extrapolate to zero at $T=0$ where the atoms would become stationary in a classical picture ("CM Harmonic" curve in Fig. 6.2). However, at low temperatures, the quantum effects become important and the PDF peaks widths do not go to zero; they saturate at a finite width corresponding to the quantum zero-point motion. Figure 6.2 also shows how this simple behavior is modified if the interatomic potential is not strictly harmonic, for example, a Morse potential, $U_{\mathrm{M}}(x) = -U_0[1 - \mathrm{e}^{-\alpha x}]^2$ (Morse, 1929). The Debye model is very simplified, yet works surprisingly well in most

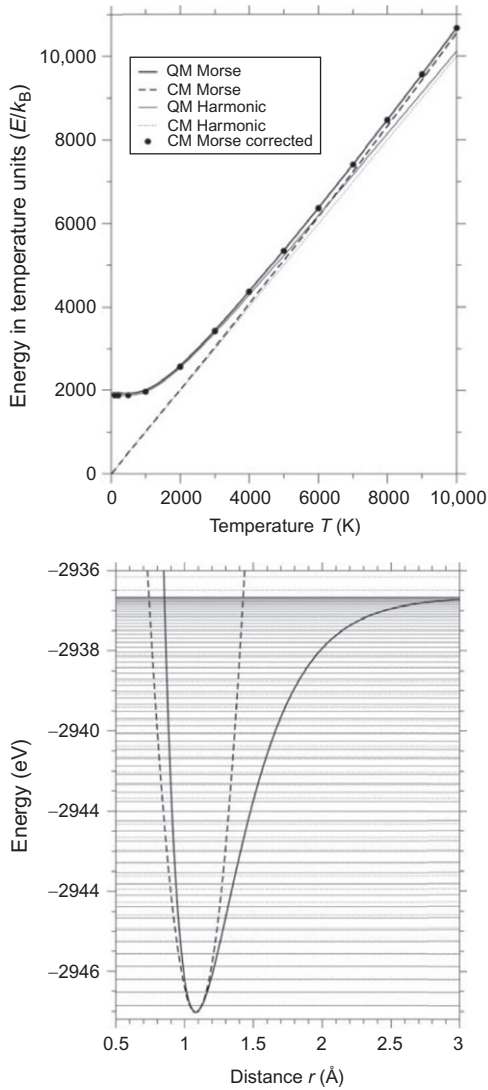

FIGURE 6.2 (Left) Illustration of the temperature dependence of the PDF peak width. The *y*-axis is proportional to the mean-square atomic displacement, which is directly related to the PDF peak widths. The classical harmonic behavior is shown by the thin solid line. It is linear over the entire temperature range intercepting zero at $T=0$. The quantum mechanical harmonic behavior is a thicker solid line. It lies on top of the classical result at high temperature but crosses over to the curve that intercepts the *y*-axis near 2000 due to the QZPM. A more realistic interatomic pair potential is the Morse potential (the harmonic and Morse potentials are shown in the right-hand panel). The anharmonicity results in a high-temperature result that is not linear but curves upward, giving the impression that the potential is becoming softer with increasing temperature, which is indeed the case (Levashov *et al*., 2007).

cases. One reason may be that the phonon density of states assumed by the Debye model is not a million miles away from the actual densities of states of materials, as illustrated for Ni in Fig. 6.3 (Jeong *et al.*, 2003).

In the manganite system $La_{1-x}Ca_xMnO_3$, a sharp deviation from canonical Debye behavior of the PDF peak width with temperature was observed (Fig. 1.10) (Billinge *et al.*, 1996). This deviation correlated with the metal–insulator transition temperature in this material, which gave strong evidence for the appearance of lattice polarons and carrier localization at this phase transition. In the figure, the peak width is not plotted directly but the PDF peak height is. The width can be extracted by fitting Gaussian functions or more commonly Gaussians convoluted with a Sinc function to account for termination effects. However, because the number of neighbors is constant, the integrated area under the peak is invariant and the peak height, extracted directly from the data, gives the inverse peak width. This can often give a more accurate determination of the peak width, especially when plotting trends such as temperature dependence, than carrying out Gaussian fitting.

The PDF peak width as a function of atomic separation, r, has been studied in a metallic and a semiconducting system (Jeong *et al.*, 1999). Peak widths were extracted by fitting Gaussians convoluted with Sinc functions. The peak width increases with r following a curve with a $(1-1/r^2)$ dependence (Fig. 6.4). The sharpening of the low-r peaks arises because in a solid the neighboring atoms tend to move in a correlated fashion because they are directly bonded to each other. The strain field around a misfitting impurity in an elastic continuum is expected to fall-off continuously as $1/r^2$ which explains why the correlations also die off somewhat continuously as they do. The r-dependence of the peak width comes from motional correlations which are discussed in Chapter 7. They are the real-space manifestation of the reciprocal-space phonon modes. The Debye model has been extended to account in a simplified way for the motional correlations (Beni and Platzman,

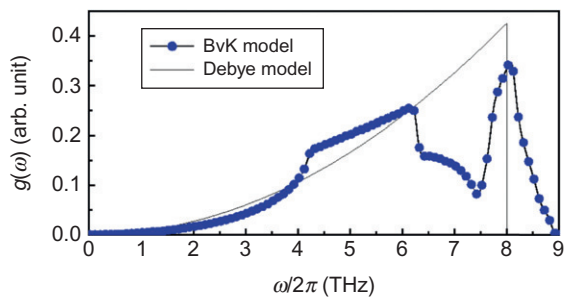

FIGURE 6.3 Comparison of a realistic phonon density of states (PDOS) and the PDOS assumed in the Debye model. Shown here is the case for Ni (Jeong *et al.*, 2003).

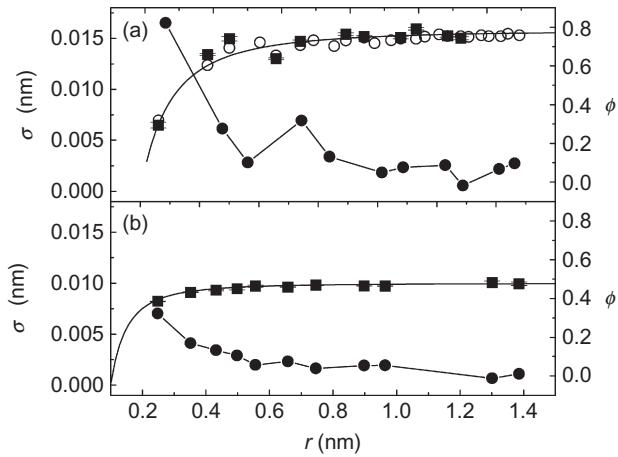

FIGURE 6.4 *r*-Dependence of the PDF peak width. Peaks at low-*r* become sharpened (σ becomes smaller) due to correlated atomic motion. (a) In–As and (b) Nickel were measured using high-energy X-rays at CHESS. Also shown is a correlation parameter (see Jeong *et al.*, 1999 for the definition) that is a measure of how correlated is the atomic motion. For near-neighbor atoms in In–As, the motions are 80% correlated.

1976; Jeong *et al.*, 2003). This treatment shows that the motional correlation results in a PDF peak width that has an r_{ij}-dependence according to $1/r_{ij}^2$ at low temperature (strictly, when $T \ll T_\theta$, where T_θ is the Debye temperature) and a $1/r_{ij}$ dependence at high temperature (when $T \gg T_\theta$) (Jeong *et al.*, 2003).

Interestingly, the peak widths deviate from the smooth $1/r_{ij}^2$ curve and potential-based modeling shows that these deviations are real. Certain directions in the solid (in the semiconductor alloys, it is the $\langle 110 \rangle$ directions) are more stiffly bonded and displacement correlations extend further in these directions leading, for example, to the anomalously sharp fifth peak in the In–As PDF. These deviations in peak width from the canonical Debye behavior are small, as shown in Fig. 6.5. However, these deviations (e.g., the upward fluctuations of the circles for the PDF peaks at 3.8, 5.5, and 7 Å in the figure) contain the detailed information about the real phonon spectrum and the nature of the phonon modes. The information is clearly present, but there is very little of it, limiting the value of the PDF for determining phonons directly, as discussed more in Chapter 7.

6.2.3.4. *Local Atomic Density*

There is one more piece of information the PDF can provide, in principle. It is the atomic density, ρ_0. Because $g(r) = 0$ at distances much shorter than the nearest-neighbor distance, $G(r) = -\rho_0 r$ at such distances, and the slope of $G(r)$ gives the value of ρ_0. This, however, assumes that one can get a perfect $G(r)$. Unfortunately, in reality, various errors, such as the normalization error,

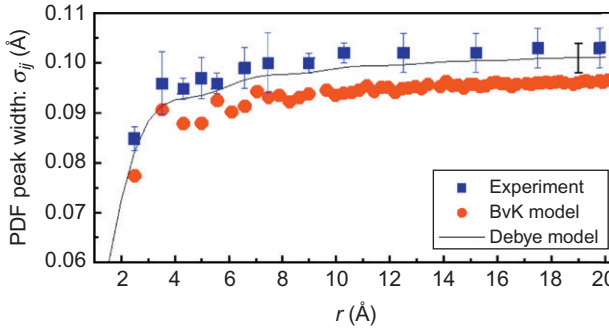

FIGURE 6.5 PDF peak widths as a function of interatomic separation, r for Ni. The solid line is the behavior of the correlated Debye model, fit to the data. The solid circles are obtained from a Born von Karman model, which gives a good account of the phonons in Nickel. The solid squares are measured peak widths (Jeong *et al*., 2003).

tend to accumulate at short distances, and it is difficult to determine the slope reliably. But in some cases, we can use this equality to estimate the atomic density. Here, it is important that the atomic density thus determined can be different from the macroscopic density which includes the contribution from macroscopic voids.

An example is the case of graphite and nanoporous carbon shown in Fig. 6.6 (Dmowski *et al*., 2012). It is possible to estimate the atomic density within 20% of accuracy in spite of noise. Here, the estimated density of nanoporous carbon, 0.088 atoms/Å^3, is lower than that of graphite, 0.110 atoms/Å^3, but is significantly higher than the macroscopic density, 0.037 atoms/Å^3, reflecting the effect of large voids. In modeling this material, it is important to use the local density determined by the PDF, rather than the macroscopic density.

6.3. MODELING THE PDF

In the previous section, we described how structural information can be obtained in a model-independent way from the PDF. In practice, the most information, and the most quantitatively reliable information, is obtained by fitting structural models to the data. The method for calculating a PDF from a structural model was discussed in detail in Section 6.2. Once the model PDF has been calculated, it can be compared to a PDF derived from data to assess how good the structural model is. At this point, it is typical to quantify the "goodness of fit" with a suitably defined residual function. Two different models can then be compared with each other by comparing their goodness of fit. The model with the lower goodness-of-fit parameter is the better model; the difference between the models is significant if the difference in the goodness of fit exceeds the uncertainty in this parameter coming from the random

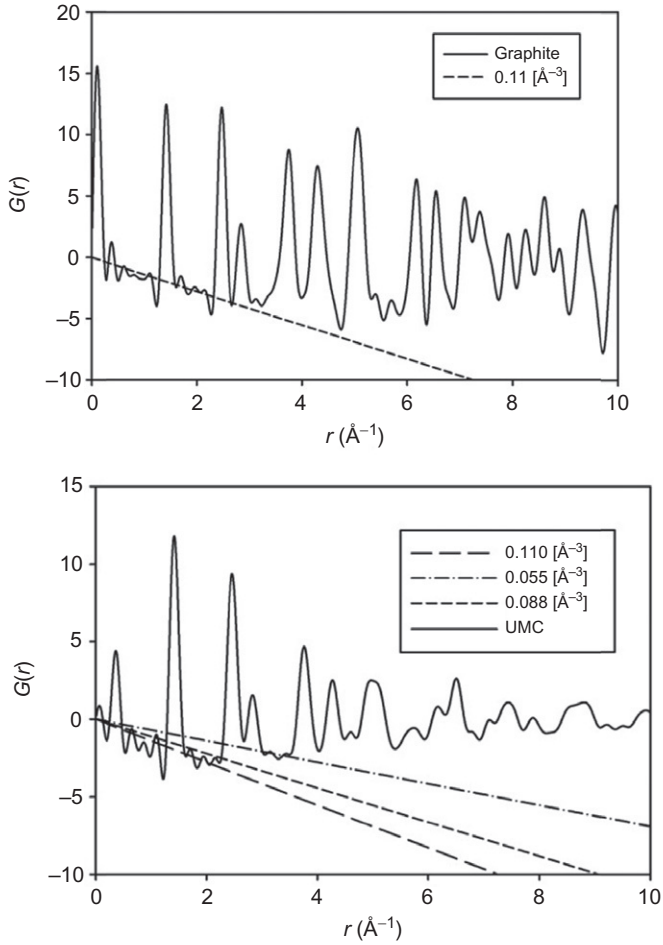

FIGURE 6.6 (a) $G(r)$ for graphite with the slope corresponding to atomic density 0.11 atoms/Å^3. (b) $G(r)$ for the nanoporous carbon with three different slopes. It is clear that the local atomic density is 80–100% of atomic density in graphite, significantly higher than the macroscopic density which is only 1/3 of the density in graphite (Dmowski *et al.*, 2012).

errors in the data. This trial-and-error approach to finding a good model can then be automated using some regression technique where model parameters are allowed to vary, subject to certain user-defined constraints, and these variations continue until a minimum is found for the goodness-of-fit parameter.

Roughly speaking, there are two different philosophies used for calculating models, "Big Box" and "Small Box" modeling, and there are a wide range of different approaches used for the minimization procedure. The Small Box approach for calculating the model is to specify the smallest possible unit cell that explains the structure and to calculate the PDF from this unit cell.

Distributions in the atomic distances are accounted for by convoluting the resulting calculated PDF with Gaussian functions. The basic unit cell is periodically repeated in space as far as is necessary to calculate the PDF to the required r_{max}. The "Big Box" approach is to model the sample as a large box of atoms which is much larger than the range over which the PDF is to be calculated. In this approach, the atomic distributions due to thermal motion are either explicitly included in the distribution of atom positions in the box or if the box does not contain enough atoms to give smooth distributions, the convolution approach can also be used. Often, both approaches of accounting for the atomic disorder are used together. As time goes on, the distinction between these modeling approaches is becoming somewhat blurred as we discuss below.

It should be pointed out that once a structural model has been defined, both the scattering in Q-space and the PDF can easily be calculated. Since the PDF is simply the Fourier transform of the total scattering structure factor, $S(Q)$, the model could be fit to the data in real or reciprocal space. Since the data are the same, each approach should yield the same information; but there is often debate about which approach is superior. Even though it is true that $S(Q)$ and the PDF contain the same information, there are some pertinent differences between the two representations and it turns out that, from a practical point of view, the two representations lend themselves to being modeled in different ways. Depending on the question being asked, the answer is often more accurately determined in one representation or the other. Even though in principle you could, you would not attempt to change the oil on your car from under the hood or change the spark plugs by lying under the car. In the same way, the total scattering yields its various pieces of information more readily either in real- or reciprocal-space depending on which piece of information is being sought. In the following sections, we describe different modeling approaches in more detail and this discussion will be made more concrete at this point.

6.3.1. Small Box Modeling

The Rietveld full-profile fitting method (Rietveld, 1969) for extracting structural information from crystalline powder diffraction patterns has revolutionized the use of powder diffraction for structure studies (Young, 1993). While powder diffraction is not the method of choice for determining new structures (although it can be, and is, used successfully in the absence of single-crystal data; David, 2002), its simplicity and the ease of making samples, collecting and analyzing the data, make it by far the most widely used method for refining accurate structural parameters for a given sample under varying conditions. Furthermore, because the problem of extinction (a dynamical scattering effect) is much smaller in powder than single-crystal diffraction, once the gross features of a structure have been determined from single crystals,

the quantitative fine details are often determined by crushing up the single crystal and doing a powder measurement.

The Rietveld method (Rietveld, 1969) was motivated by the fact that in a powder measurement, many of the Bragg peaks are overlapped especially in the high-Q (small d-spacing) region of the pattern. Apart from the inherent loss of directional information in a powder measurement, this fact severely limited the set of data points that could be obtained by integrating the intensity of separated Bragg peaks to obtain the crystallographic structure factors needed for crystal structure solution. The Rietveld approach was to calculate the complete set of crystallographic structure factors, and therefore the Bragg peak intensities and positions, for a trial structure. These were then convoluted with profile functions to account for the instrument resolution and sample-dependent peak-broadening effects. The intensities were modulated with experimental effects such as absorption, extinction, incident flux, background intensities, and so on and sample-dependent effects such as Debye–Waller factors. In this way, the entire experimental dataset was simulated and compared to the measured one. All the above-mentioned effects are parameterized and the parameters are allowed to vary until a best fit is obtained, traditionally, using a least-squares approach. The sample-dependent parameters thus derived include the unit cell parameters (unit cell lengths and angles), atomic positions in the unit cell expressed in fractional coordinates, anisotropic thermal ellipsoids for each atom, and the average atomic occupancy of each site.

This approach has been applied, in exact analogy, to the PDF (Dmowski *et al*., 1988; Billinge, 1998; Proffen and Billinge, 1999; Farrow *et al*., 2007) in the program PDFFIT (Proffen and Billinge, 1999), and more recently PDFgui (Farrow *et al*., 2007). We highlight here the similarities and differences with conventional Rietveld. The main similarity is that the model is defined in a small unit cell with atom positions specified in terms of fractional coordinates. The refined structural parameters are exactly the same as those obtained from Rietveld. The main difference from conventional Rietveld is that the local structure is being fit which contains information about short-range atomic correlations. There is additional information in the data, which is not present in the average structure, about disordered and short-range ordered atomic displacements. To successfully model these displacements, it is often necessary to utilize a "unit cell" which is larger than the crystallographic one. It is also a common strategy to introduce disorder in an average sense without increasing the unit cell. For example, in the case where an atom is sitting in one of two displaced minima in the atomic potential, its probability of being in either well is random and can be modeled as a split atomic position with 50% occupancy in each well. This is not a perfect, but a very good, approximation of the real situation and is very useful as a first-order attempt at modeling the data.

This "Real-space Rietveld" approach is proving to be very useful and an important first step in analyzing PDFs from crystalline materials. This is

because of two main reasons. First, its similarity with traditional Rietveld means that a traditional Rietveld-derived structure can be compared quantitatively with the results of the PDF modeling. This is an important first step in determining whether there is significant evidence for local distortions beyond the average structure. The Rietveld model is refined to the PDF without relaxing any of the symmetry constraints of the crystallographic model. The resulting fit can be assessed to look for significant deficiencies. In addition, the refined parameters can be compared directly with the results from Rietveld to make sure they are consistent. If evidence exists to suggest that local structural distortions beyond the average structure are present, these can then be incorporated in the PDF model. The second strength of the real-space Rietveld approach is the simplicity of the structural models making it quick and straightforward to construct the structural models and making physical interpretations from the models similarly quick and straightforward.

Why might one want to analyze a well-ordered crystalline material in real space? In this case, there is no diffuse scattering in the diffraction pattern and there seems little advantage in going for the PDF. In fact, it can still be a worthwhile exercise for the reason illustrated with the oil-change analogy in Section 6.2. Below we describe three reasons for carrying out a real-space Rietveld refinement on a well-ordered material, beyond the obvious advantages that are apparent if disordered local atomic displacements are present.

Local structural parameters, such as thermal factors and anharmonicities in the crystal potential, may be more accurately determined in real space than reciprocal space. On the other hand, average structure parameters, such as lattice parameters, are much more accurately obtained in Q-space. As a concrete example, let us consider thermal (or more precisely displacement) factors. The atomic thermal parameters are measured by considering the monotonic Gaussian fall-off in intensity of Bragg peaks: the Debye–Waller factor. For a sample with a high Debye temperature, this fall-off can be quite slow. For example, a $U_{ii}=0.0025$ Å^{-2} results in a Debye–Waller envelope with a standard deviation (half-width) of $\sigma=14$ Å^{-1}. Many Rietveld refinements are not carried out over a range beyond 14 Å^{-1} ($d=0.45$ Å) because of the problem of Bragg peak overlap at high-Q. On the other hand, PDFs are routinely measured taking a range of data up to 30 Å^{-1} and beyond. Furthermore, the arbitrary background function used in most Rietveld refinements is extrapolated into the high-Q region where it cannot be separated explicitly from the data. A small error in the background at high-Q can have a significant effect on the refined thermal factor because of the low intensity in the Bragg peaks themselves. In a PDF analysis, the background is extracted explicitly. Similar arguments apply to other slowly Q-varying experimental effects such as absorption and multiple scattering. It is, therefore, to be expected that more reliable thermal factors can be obtained from a PDF analysis than from a Rietveld refinement; though it is clearly true that if Rietveld was carried out on a corrected $S(Q)$ function (as will increasingly happen in the future) over the

same range of Q as the PDF analysis, the resulting thermal factors will have similar accuracy. To date, no quantitative comparison of the relative accuracy of Rietveld and PDF-derived thermal parameters has been carried out. Where the values have been compared, they are in rather good agreement with the PDF-derived parameters showing less of a tendency to yield unphysical values (e.g., becoming negative) (Proffen *et al.*, 1999; Gutmann *et al.*, 2000). However, when a series of datasets collected as a function of temperature are compared, more scatter is apparent in the PDF-derived values than the Rietveld. The reason is not clear though it may result from inadequate (simple Gaussian!) profile functions being used in the PDF refinements and the situation may improve as the sophistication of the PDF modeling approaches that of Rietveld, the more mature technique.

Another possible reason to carry out a real-space Rietveld analysis on a well-ordered crystal is that the structural parameters are differently correlated in real- and reciprocal-space refinements because the equations to calculate the signal from the model are different. In any refinement, different refinement variables can become significantly correlated with each other and each one can take on an unphysical value which, nonetheless, when applied together reproduces the data well. This is true in real- and reciprocal-space refinements. However, since the various variables will be differently correlated with each other in the two cases, a joint refinement in real and reciprocal space can remove the correlations. At present, there is no joint real- and reciprocal-space refinement code in existence, though this will undoubtedly be remedied in the future.

Finally, more precise information about the atomic potential can be gleaned from a real-space analysis in a well-ordered material. As discussed in Chapter 7, directly bonded, nearest-neighbor atoms tend to move in phase with each other and their motion is positively correlated. If the bond is very stiff, the relative motion of these neighbors with respect to each other can be quite small (in covalent semiconductors the nearest-neighbor motion is $>80\%$ correlated!). The motion of far neighbors is uncorrelated. The thermal factor measured in a conventional crystallographic refinement is the uncorrelated one. This often reflects the softest bonds in the material. For example, in the semiconductor alloys, it reflects the weak bond-bending forces. The stiff bond-stretching force can be measured from the correlated motion of the near neighbors which is obtained from the width of near-neighbor peaks in the PDF.

When local atomic displacements away from the average structure exist, and therefore significant diffuse scattering intensity is present in the scattering, a conventional Rietveld refinement becomes inadequate. At this point, a model is required that can fit both Bragg and diffuse intensities and which includes short-range ordered structures that are different from the average structure. This is much more readily done in real space where there is a natural separation in r, where it is now common to fit different models to explain the low-r region and the high-r region.

This approach to studying the structure of complex materials has grown rapidly, in part because it is a powerful way of obtaining practical quantitative information about the structure of complex materials, in part because the availability of PDF data is increasing with the developments of the RAPDF and ePDF data collection coupled with rapid data reduction methods, and finally, facilitated by the availability of a rather easy-to-use graphical user interface (GUI)-driven program PDFgui (Farrow *et al.*, 2007). The approach was developed to study local structure in crystalline materials, but has found natural extension to nanoparticles, clusters, and nanoporous systems.

6.3.1.1. Example of Small Box Modeling: PDFgui

The most sophisticated and robust real-space Rietveld code available at the time of writing is the PDFgui code (Farrow *et al.*, 2007). This code allows multiple datasets to be refined and can handle multiple phases. It has a comprehensive and easy-to-use GUI but also has a scripting interface to the underlying refinement engine (called PDFFIT2) allowing great flexibility and user control. Arbitrary constraints can be introduced between the parameters as required, for example, to reproduce the average symmetry of the sample or to allow rigid rotations or translations of sets of atoms. Helper functions are present in the program for automatically expanding the unit cell (important as local structure often has lower symmetry than the average structure implying a supercell is needed to model it) and for creating the proper constraint equations so that the refinements respect some desired space-group symmetry. The program also incorporates scripting interpreter allowing mathematical formulae to be evaluated dynamically as the program executes. It can be controlled using macros allowing the refinement to be automated, and the most useful scripts, such as temperature-dependent refinements on multiple datasets, are incorporated into the GUI. The program uses a least-squares minimization procedure and yields parameters with estimated standard deviations associated with them, though this implies that the uncertainties on the data are accurately known. This is often highly problematic as discussed in Chapter 5. As is commonly acknowledged in Rietveld refinements, the estimated errors coming from the software tend to be an underestimate of the real absolute errors because they do not account for systematic errors and statistical correlations between points in the data (e.g., see Tian and Billinge, 2011). However, they provide a good baseline for estimating the significance of structural parameters and should be a good estimate of the relative changes of parameters between similar datasets (e.g., temperature dependence of parameters).

The residual functions used in PDFfit are equivalent to the Rietveld R-values. The weighted agreement factor, R_w, is defined as

$$R_{\mathrm{w}} = \sqrt{\frac{\sum_{i=1}^{N} w(r_i)\left[G_{\mathrm{obs}}(r_i) - G_{\mathrm{calc}}(r_i)\right]^2}{\sum_{i=1}^{N} w(r_i) G_{\mathrm{obs}}^2(r_i)}}, \tag{6.2}$$

where G_{obs} and G_{calc} are the observed and calculated PDFs (in the form of $G(r)$) and w is the weighting factor, $w(r_i) = 1/\sigma^2(r_i)$, where σ is the estimated standard deviation on the ith data point at position r_i. Note that PDF residual functions are not strictly statistically significant quantities since neighboring points in the PDF are not statistically independent. This is discussed in more detail in Appendix 5.3. For example, a reliable χ^2 cannot be determined from $R_{\mathrm{wp}}^{\mathrm{PDF}}$ as it can from a Rietveld refinement by dividing R_{wp} by $(N-P)$ where N are the numbers of points and P the numbers of parameters in the fit. Nonetheless $R_{\mathrm{wp}}^{\mathrm{PDF}}$ is a quantitative measure of goodness of fit and can be used to compare models and minimized to optimize a model. Points in the PDF separated by $\Delta r \sim \pi/Q_{\max}$ are approximately statistically independent and so a reasonable estimate of χ^2 is possible by taking (Billinge, 1992):

$$\frac{R_{\mathrm{wp}}^{\mathrm{PDF}}}{\left(\frac{(r_{\max} - r_{\min})Q_{\max}}{\pi}\right) - P}. \tag{6.3}$$

6.3.2. Big Box Modeling

The experimental PDF comes from a sample with many atoms in it, $>10^{20}$. Big Box modeling takes the same approach. A model is created with many tens of thousands of atoms in a box to simulate the solid. The PDF is then calculated from this model and compared to the measured PDF (of course the comparison can be made also in reciprocal space). As we discussed, most PDF measurements yield the instantaneous structure, like photographs of the structure taken with a very fast shutter speed. A photograph is taken for each recorded scattering event and the final PDF is the average of all the instantaneous snapshots. Because of ergodicity (Huang, 1987), the temporal and ensemble averages give the same result. This means that if the model in the box contains enough atoms, the average PDF of those atoms can be used to simulate both static and dynamic disorder. Large boxes and/or multiple configurations of the model are needed to get sufficient statistics on the calculated PDF to give a good agreement to the measured one.

Direct simulations of solids may also be done this way. For example, given an interatomic pair potential that gives the energy of pairs of atoms as a function of their separation, it is possible to put a large number of atoms in a box and to calculate the energy of the configuration. It is useful to then have a method to find the configuration that yields the minimum energy. This

is a classic global optimization problem and there are various algorithms for minimizing the energy. One of the most widely used is the Metropolis simulated annealing algorithm (Metropolis *et al*., 1953). This is a hard problem because the energy landscape may be quite complicated and it is not straightforward to find the minimum. In the Metropolis algorithm, atom positions are allowed to vary in some random way. After each change, the energy of the system is calculated using the specified potential energy function. If a change reduces the energy of the system, it is accepted. If it raises the energy of the system, then it can be accepted or rejected. This decision is made randomly according to an underlying probability. It is this "gambling" aspect of the technique that got its name: Monte-Carlo is the famous gambling center for the rich and famous on the French Riviera in the sovereign principality of Monaco. In simulated annealing, the underlying probability is given by

$$P = \mathrm{e}^{-\Delta E/kT}, \tag{6.4}$$

where ΔE is the change in energy, k is the Boltzmann constant, and T is the "temperature" of the system which is initially set by the experimenter. When T is higher, more "bad" moves are accepted and more of the energy landscape is probed by the simulation (it is easier to get out of local minima). The temperature can then be systematically lowered to guide the simulation into the global energy minimum. This is a widely used and powerful method in statistical physics and the computational technology is well developed (Binder and Heerman, 1992).

Clearly, Monte-Carlo simulated annealing is a powerful approach to minimization of any parameterized function and is particularly applicable if there are a large number of parameters and a complicated parameter space. From this perspective, it lends itself to being applied to solving complicated structural problems, where the function to be minimized is not an energy but a residual function: that is, the agreement between structural data and a calculated scattering pattern from a model. In this case, there is no obvious reason why the underlying probability should be a Boltzmann distribution as in simulated annealing. However, in the absence of rigorous arguments justifying the adoption of another form for the probability distribution, and the extensive understanding that exists of the behavior of simulated annealing with the Boltzmann distribution (Binder and Heerman, 1992), this was the natural place to start and this has been developed with the name Reverse Monte-Carlo or RMC following the pioneering work of McGreevy and Puszati (1988), although it has a longer history.

The approach was first used as long ago as the 1960s by the MIT group (Kaplow *et al*., 1968; Renninger and Averbach, 1973). Limited computing power really hindered progress at that time and the widespread use of this approach has really occurred since the late 1980s. McGreevy and Puszati (1988) initially applied the technique to study the structure of liquid argon and coined the name Reverse Monte-Carlo to differentiate the technique from

conventional simulated annealing. Shortly afterward, Dmowski *et al*. (1988) and Toby *et al*. (1990) also used simulated annealing to study correlated local structural distortions in crystalline materials using a more "Small Box" approach but using a simulated annealing algorithm to do the regression. The implementations of the technique were quite different and both have been applied with some success. The strength of the McGreevy and Puszati (1988) implementation is also a potential weakness. In this case, a large box of atoms is used as the structural model. These are allowed to arrange themselves, with the minimum of constraints, in such a way as to give good agreement with the diffraction data. In the most unconstrained case, the hard-sphere repulsions, which prevent atoms overlapping, are the only constraints. Originally, the data to be fit were in the form of a PDF (Keen *et al*., 1990), but more recently total scattering data (Montfrooij *et al*., 1996) and even single-crystal data (Nield *et al*., 1995) have been used. Clearly, the strength of this approach is that it is unbiased. The resulting structural model gives a solution that is unprejudiced but consistent with the data.

The use of simulated annealing to minimize a residual function constitutes solving the famous diffraction "inverse problem." The direct problem is that given a 3D arrangement of atoms, what is the diffraction pattern (or in this case the PDF). The inverse problem is that, given a PDF, what is the 3D structure that gave rise to it? The PDF data are the input to the inverse problem; the structural model is then fit to the data by adjusting atom positions to minimize the residual function. The problem is that, in general, the model is underconstrained in which case there is not enough information in the data to yield a unique solution. This is the case of an ill-posed inverse problem. The most difficult question to answer when carrying out Big Box modeling often is "that's a nice model, but is it correct?" meaning, "is it unique?" The problem is that there is currently no definitive way of counting the constraints and the degrees of freedom in the problem. This is illustrated with an intuitive example. Let us say that your model consists of 10,000 free atoms. The degrees of freedom are then 29,997, that is, three spatial coordinates for each atom minus 3 because the origin is not specified (we neglect the momenta of the atoms in this modeling). This means you would need at least 29,997 independent data points to constrain this model. Now, let us apply some constraints. First, we will make the atoms nonoverlapping hard spheres, then we will confine them to a box. What is the number of degrees of freedom now? We can imagine shrinking the box until the atomic density inside was exactly the theoretical maximum density for close packing of the spheres. In this case, if the spheres are identical, there is only one way to pack them and there is only one solution to the problem, in fact, in this case you do not even need any data to constrain the solution, it is completely determined just by the constraints. If the atoms are different from each other, there are the permutations of the different atom types over the available sites. Some of these will give equivalent PDFs and some will not. If the box is infinitely

large, then the number of degrees of freedom asymptotically approaches 29,997 again, but how does this happen? How does the number of constraints increase when we slowly increase the size of the box away from the theoretical maximum density? It is evident that this problem is very complex and, currently, generally ignored.

Despite these difficulties, it is still of value to model in this way. Any approach that provides physical insight and helps you to interpret your data is valid, as long as the conclusions drawn from the study are not overreached, and even if the uniqueness of the solution is not established, one additional "constraint" is to pick solutions that answer the question "is it physically reasonable" in the affirmative. An argument that has been made in the RMC literature is that *any* structural motifs that emerge in the model, such as local tetrahedral atomic arrangements in network glasses, are probably real. The reason is that the simulated annealing, by its very nature, will find the most probable, and therefore most disordered, structural solution that is consistent with the data. Any atomic correlations (ordered atomic arrangements) that emerge in the model must be more or less uniquely specified in the data themselves. This is a slightly dangerous line or argument though as it does not address the uniqueness issue. There may be other structural solutions that give the same signal in the PDF and there are a few examples where this has been clearly demonstrated (Soper, 2007, 2011). The problem is especially acute in disordered systems, where the structural information is limited and there is an enormous degeneracy in possible structural solutions consistent with the same dataset. If one does not constrain the structural model in physically reasonable ways, your structure solution is quite likely to be wrong and often unphysical (Keen, 1998). Clearly, care must be taken when interpreting the results of this kind of modeling. At this point, the ways forward are to get additional data and to constrain the model and this is discussed in Section 6.3.3.

When faced with crystalline data, a great deal is known about the structure from a conventional crystallographic analysis. It is clearly inefficient, and most probably disastrous, to specify the starting model by placing atoms randomly in the box and hoping that the correct crystal structure is found by random mutations of the model.[2] In this case, a more conventional refinement is preferred where a well-defined starting model is specified and this is allowed to distort randomly to fit the data. When a Monte-Carlo algorithm is used for the regression, the term RMC refinement has been coined (Keen, 1998),

2. This actually points to the inefficiency of the Monte Carlo algorithm. It is one of the easiest regression algorithms to implement, is versatile and widely used, but is not efficient and often will not find the real global minimum on attainable time-scales (given infinite time it will find it!). Thus, another problem with RMC modeling is that it is never clear if the "best-fit" model is actually the best fit. When the problem is ill-posed it is rather easy to find good solutions but they are not unique, when there is a unique solution it is often difficult for Monte Carlo to find it, rather it finds a more easily accessible local minimum.

distinguishing the approach from RMC modeling as described in the previous paragraph. As RMC modeling becomes more constrained (Tucker *et al*., 2007), the distinction becomes somewhat cloudy; however, it is useful to distinguish the two approaches since in many people's eyes, RMC equates to unconstrained RMC modeling with all its shortcomings.

For modeling local structure in crystalline systems, a number of powerful RMC refinement codes are emerging, mostly around the RMCprofile program of Tucker *et al*. (2007) (http://www.rmcprofile.org). This is the best possible scenario for Big Box modeling. First, because the material under study tends to be more highly ordered, there is more information in the diffraction patterns. Also, RMCprofile does a good job of incorporating additional constraints into the modeling. This is discussed in more detail in Section 6.3.3.

A number of other RMC modeling and refinement codes are also freely available. Used mostly for disordered materials, RMC++ is still under active development (http://www.szfki.hu/~nphys/rmc++/opening.html) (Evrard and Pusztai, 2005; Gereben *et al*., 2007). As well as its special features for modeling nanoparticles using evolutionary algorithms, the DISCUS program (http://discus.sourceforge.net/) also has an implementation of RMC (Neder and Proffen, 2008). A number of examples of data analyses using these programs will be presented in later chapters.

6.3.3. Complex Modeling

The discussion above of Big Box modeling introduced the idea of an inverse problem and whether it is well posed or ill posed. The issue is the following. For the forward problem, given a 3D structure, the diffraction pattern can be calculated directly. However, the reverse is not true because of the phase problem. Diffraction intensities contain information about the amplitude but not the phase of the Fourier components giving rise to the intensities. There is an inherent loss in information meaning that there is no direct mathematical transform that can take you from the diffracted intensities to the 3D density distribution: the structure. This does not mean that it is impossible to determine the structure from the diffraction. If sufficient or redundant information is measured, it is possible to find algorithms that will determine the structure by solving the inverse problem, usually by highly automated trial-and-error methods. For example, if you have a unit cell containing 5 symmetry-unrelated atoms, roughly speaking you need 18 independent constraints to solve it (6 to define the unit cell and 15 for the atoms, minus 3 as the origin is unspecified). Even with the loss of the phase information, if enough Bragg peaks are measured, there is enough amplitude information to find a solution. Typically, the position and intensity of many more than 18 Bragg peaks will have to be measured as they do not contain exactly orthogonal information. If your diffraction information is coming from powder data, there is an additional loss of directional information due to powder averaging. However,

again, by oversampling (obtaining redundant information) this 3D information can be reconstructed and a unique 3D structure can be obtained from the 1D data. The ability to solve crystal structures from powder data is now well established and used (David, 2002). As discussed below, it has also been demonstrated that, under favorable circumstances, unique 3D structures can be reconstructed from PDF data alone (Juhas *et al*., 2006, 2008, 2010).

Things go bad when the complexity of the problem increases. For example, the number of atoms in the asymmetric unit may increase and more Bragg peaks are needed for a solution. Also, the symmetry of the solution may decrease, also increasing the number of atoms that must be located. Happily, the increase in unit cell size and the decrease in symmetry result in there being many more Bragg peaks available to measure, and this increase in the available information matches the increase in the complexity of the solution so that structure solutions of mind-bogglingly complicated structures, including large proteins and even viruses and chromosomes, are possible. This kind of reasoning has been extended to the case of diffuse scattering from nonperiodic objects (Sayre, 1952) where the data can be oversampled if the object to be imaged is sitting on a null-scattering (or very weakly scattering) support that is at least twice as large as the object itself. This was experimentally verified (Miao *et al*., 1999), and great progress has been made to image at the atomic scale (Zuo *et al*., 2003) and ultimately to solve nanostructure in this way (see e.g., Scott *et al*., 2012).

However, in the context of powder diffraction, the increase in the density of diffraction peaks presents an experimental challenge due to peak overlap when the pattern is orientationally averaged. This leads to a *loss* in information, especially in the high-Q region where the Bragg peak density diverges, as the peaks merge and overlap, and the loss is worse for low-symmetry structures. This is mitigated by making a very high-resolution measurement which is allowing even protein structures to be refined (not solved) from powder data (von Dreele *et al*., 2000). However, for nanomaterials, the intrinsic width of the diffraction features due to finite size effects (Scherrer broadening) means that this is not possible even in principle. Thus, for total scattering studies aimed at elucidating nanostructure, the information in the data is disappearing just as the complexity of the structure aiming to be solved is increasing. There is certainly a point where there will be insufficient information in the diffraction pattern alone to constrain a unique solution to the inverse problem. An information gap opens up and the problem becomes ill posed. Systems with strong disorder, such as liquids and glasses, fall in this class. From a theoretical perspective, this can only be solved by closing the gap: regularization of the problem. This means either increasing the constraints coming from data or decreasing the degrees of freedom in the model, also by introducing theoretical constraints. We refer to the practical implementation of this as "complex modeling," where "complex" implies the joining, or complexing, of multiple information sources in a coherent global optimization scheme as illustrated in Fig. 6.7.

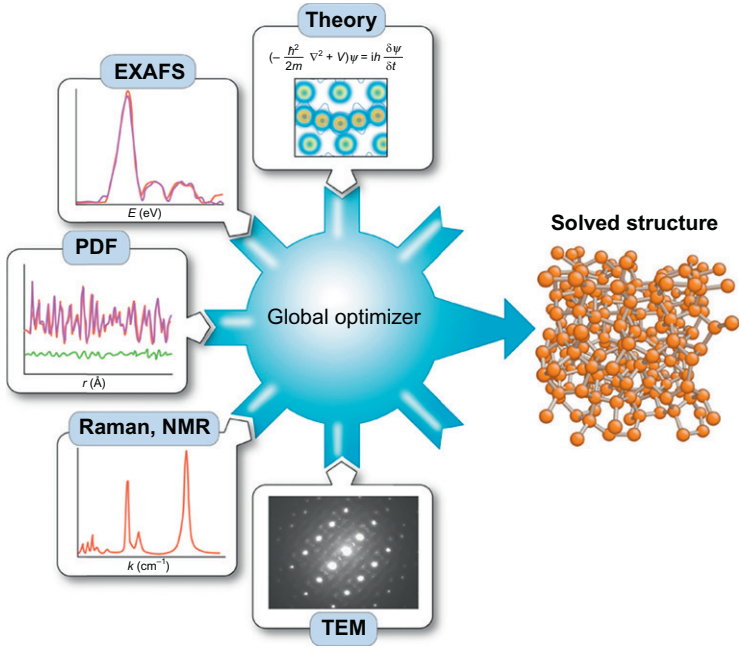

FIGURE 6.7 Schematic illustration of complex modeling. Multiple different heterogeneous information sources are joined, or complexed, together to constrain a global optimization of the inverse problem, in this case the diffraction inverse problem to find a complicated 3D structure (Billinge, 2010).

In some sense, complex modeling has been around about as long as modeling. It is actually rare not to apply constraints such as symmetry and hard spheres for atoms and multiple datasets such as different resolution tof neutron banks, for example, all of which could correctly be described as forms of complex modeling. Nonetheless, we think it is helpful to couch the problem in these terms in order to move forward in a more systematic way. A number of existing modeling efforts can be couched in these terms and a roadmap for future modeling efforts becomes apparent.

6.3.3.1. Example of Complex Modeling: Empirical Potential Structure Refinement

Empirical Potential Structure Refinement (EPSR) is a method developed by Alan Soper at ISIS in the United Kingdom (Soper, 1996, 2001) originally for extracting structural information about solvation states of ionic and molecular complexes in solution. This is very important question (most of life on earth takes place in water) but extremely hard to solve because of the low information content of the data and the complexity of the models. The data

that are used here to increase the information content are complexes of multiple neutron liquid structure factors and PDFs with different kinds of deuteration. As discussed in Chapter 3, by using isotopic substitution it is possible to get differential- and possibly partial PDFs. An excellent strategy for studying hydrogenated liquids and hydrogenated solute species is to preferentially deuterate different parts of the system and collect complementary datasets. For example, it is possible to have a hydrogenated molecular complex in water and heavy water and then to selectively deuterate different parts of the solute molecule. The problem of collecting data from liquid systems containing hydrogen is also large, but largely solved, with the specially designed SANDALS and NIMROD diffractometers at ISIS. Despite the extra datasets, the problem still needs to be modeled to extract the desired information and this is done using EPSR.[3] This approach uses a big box in the sense that a big box is created containing a large number of atoms. However, it is not a Big Box modeling method in the sense that only a small number of parameters are refined: the parameters that define an interatomic potential between the atoms. A scheme is then used to refine the values of the potential parameters to give a good fit to all the data.

One of the first attempts to extract pair potential information directly from measured PDF data was by Kaplow and coworkers (Kaplow *et al.*, 1964; Lagneborg and Kaplow, 1967). In simple metals such as solid lead and cobalt, they extracted the mean force potential, $U(r)$ (i.e., a weighted sum of all the pair potentials of atoms surrounding the atom at the origin, e.g., see Hansen and McDonald, 1986) from the expression

$$\rho(r) = \rho_0 \exp\left(\frac{-U(r)}{k_B T}\right), \tag{6.5}$$

where k_B is Boltzmann's constant (Kaplow *et al.*, 1964). The derivative of $U(r)$ with respect to distance gives the net force on the atom at the origin (due to all the neighbors at various distances). Since $\rho(r)$ was explicitly measured (it is the experimental PDF!), $U(r)$ can be obtained directly by inverting Eq. (6.5).

In EPSR, a model is set up with initial values for the pair potential parameters and the mean force potential, $U(r)$, calculated for each pair of atoms. Inverting Eq. (6.5), we get

3. This is an example of so-called inverse Monte-Carlo (Gerold and Kern, 1987; Livet, 1987) which should be distinguished from "reverse Monte-Carlo" discussed in Section 6.3.2. The principle difference is that in an inverse Monte-Carlo refinement, potential parameters are modified and an energy is minimized using a Monte-Carlo algorithm whereas in reverse Monte-Carlo, the χ^2 of the fit is used directly as an effective energy that is minimized using Monte-Carlo simulated annealing.

$$U(r) = -k_B T \ln\left(\frac{\rho(r)}{\rho_0}\right) = -k_B T \ln(g(r)), \tag{6.6}$$

where $g(r)$ is the pair distribution function defined in Section 3.1.3.1. We therefore have a reference $U^m(r)$ from the model and a measured $U^D(r)$. We would like to think of a way to modify $U^m(r)$ to bring it closer to $U^D(r)$. To do this, a perturbation is added to the original reference potential, $U_0^m(r)$, that is, the new potential,

$$U_1^m(r) = U_0^m(r) + kT \ln\left(\frac{g^m(r)}{g^D(r)}\right). \tag{6.7}$$

The box of atoms is then relaxed to minimize the total energy, the new potential, resulting in a new calculated $g(r)$. The process is then iterated until it reaches convergence. This approach has proved to be very powerful in the study of complex liquids (Soper, 1996, 2000; Landron *et al.*, 2001). It is now being applied to more complex systems such as micelles in solution (Hargreaves *et al.*, 2011).

6.3.3.2. *Example of Complex Modeling: Constrained RMC*

As we discussed above, RMC is moving in a direction toward more constrained fits. The RMCprofile is a case in point (Tucker *et al.*, 2001, 2007). It already utilizes information from the Bragg scattering and the total scattering by ensuring that the model, when periodically averaged into a unit cell, reproduces the Bragg scattering. This should probably be thought of as an additional theory constraint (a periodicity constraint) rather an additional data constraint since the same total scattering data are used as fit for the local structure. Additional data that may be included are differential and partial PDFs, X-ray, and neutron data. Addition of EXAFS data has also been added recently (Krayzman *et al.*, 2008, 2009). Additional constraints are also being added to the model, such as limiting bond angle distortions to within physical ranges (Keen, 1998). Recently, a bond valence sum restraint was added (Norberg *et al.*, 2009). Often, something is known about the network connectivity of a structure, and constraints can be introduced that maintain the network. Two such examples are Geometric Modeling (Wells *et al.*, 2002, 2004; Sartbaeva *et al.*, 2007) and a similarity constraint that ensures that every atom has the same local environment as every other atom (Cliffe *et al.*, 2010).

These developments appear to be quite promising and this approach to modeling looks set to grow in the coming years. It is still difficult to know whether the resulting model is unique and the inverse problem well posed, but more constraints definitely take one in the right direction, at the expense of possibly introducing bias into the modeling. It is also important not to introduce a constraint and then use that constraint again as a result, or to validate a result. For example, introducing a constraint that the atoms in a

tetrahedral network glass are fourfold coordinated may be a good idea, but when the resulting model is nicely tetrahedral one cannot say that the model is now "more physical because it is tetrahedral" and then extrapolate to study the statistics of higher member rings from the model. Since they are not subject to the constraint, they may or may not be correct in the model; it is impossible to know. One can only know that the resulting model is more physical in the sense in which it was constrained to be so. Sometimes this can be subtle and can catch out even experts in the art. A tetrahedral model of a network glass was obtained and proclaimed by Cliffe *et al*. (2010) using just the similarity constraint mentioned above and without specifying a fourfold coordination *a priori*. However, the similarity constraint coupled with PDF data which have an average nearest-neighbor coordination of 4 is effectively the same as applying a fourfold constraint: the constraint was there in the model even though it was not explicitly set. This is still a valid approach to studying glassy structures to gain some insight, but extrapolating with claims that it is an *ab initio* solution of glass structure (Cliffe *et al*., 2010) are unwarranted and may be dangerous. The resulting model is fourfold coordinated because that was an (effective) fourfold constraint, but this does not necessarily make the problem well posed and the rest of the solution reliable. On the other hand, the same approach applied to data from C_{60} did result in a correct structure solution (Cliffe *et al*., 2010). In this case, we know that this problem is well posed because the C_{60} structure was previously solved from just PDF data as we describe below (Juhas *et al*., 2006). Here, adding the similarity constraint allowed the relatively inefficient simulated annealing algorithm to find the solution. This in itself is an exciting result. It shows how constraints can also greatly reduce search spaces during global optimization and aid convergence, as well as doing their job of helping to regularize problems.

Used with the proper care, these constrained Big Box approaches are very powerful and are providing new insights into complex materials. Because of the ease of implementing constraints with the RMC algorithm, this will continue to be a rich test bed of complex modeling ideas and capabilities. Both RMCprofile and DISCUS, the two main programs used for Big Box modeling of crystalline and nanocrystalline materials, have excellent, well-documented public releases and their use by the community is increasing with time.

6.3.3.3. *Complex Modeling Example: Nanoparticle Modeling*

Determining the structures of nanoparticles is currently one of the most interesting problems in complex materials. This is a canonical example of a "nanostructure problem" (Billinge, 2010). Whereas we have powerful methods to solve the structure of crystals (crystallography), these methods break down for structure that exists only on the nanoscale and for nanostructure we do not have robust structure solution methods, the topic of this whole book basically! As mentioned above, if you can get "single nanocrystal" data

(or data from a single crystal of nanocrystals; Jadzinsky *et al*., 2007), there are algorithms available that can reconstruct the structure (Sayre, 1952), and the challenge there is to get the nonorientationally averaged data from a single nanoparticle. We concern ourselves here with the problem of nanostructure solution from powder data.

One thing that makes nanostructure solution from powder data so interesting is that the inverse problem is marginally posed. As discussed below, we know that in the case of C_{60} it is well posed. We also know that in more general cases there is not enough information in a single PDF dataset to constrain a unique solution. Still, a great deal can be learned about the structure of nanoparticles using structure refinement methods, as discussed at length in Chapter 10. In that case, the Small Box modeling approach yields abundant useful quantitative information about the system (Masadeh *et al*., 2007). However, this approach treats nanoparticles like small pieces of attenuated bulk material and becomes challenged when nanoparticles are not homogeneous (although it is actually quite remarkable how far the approach can be taken, even in this regard, as discussed in Section 9.2). A different approach to this problem has been taken in the work of the group of Neder and incorporated into the latest versions of the DISCUS modeling program. In the absence of *ab initio* methods, one of the great challenges in complicated systems such as nanoparticles is actually building the initial trial models in a refinement. DISCUS has ingenious and powerful tools for doing this, as illustrated schematically in Fig. 6.8. For example, for heterogeneous nanoparticles such as core-shell nanoparticles, the shell can be grown by seeding it at a point on the surface of the core and growing the shell crystal with some specified orientation, then removing from the shell model any atoms that are inside the volume defined by the core. Having built the nanoparticle, it is necessary to carry out a refinement by making changes to the model and then accepting or rejecting the changes based on some regression algorithm. For nanoparticles, a key issue is the presence of stacking faults in the structure. These are extended defects where the change in the model involves the translation of a whole stack of atomic planes rather than a small shift of a single atom, which is a typical update step in a Monte-Carlo refinement. In this case, a differential evolution regression algorithm (a form of genetic algorithm) was found to be more robust and work better than simulated annealing.

We might ask whether this modeling is Big Box or Small Box? It is "Small Box" in the sense that the model consists of an individual nanoparticle which may have only a few hundred to a few thousand atoms. However, it is "Big Box" in the sense that to calculate the PDF with good statistics, it is necessary to make an ensemble of many such nanoparticles and calculate the average PDF from the whole ensemble. This is because, for example, a particular nanoparticle may have a stacking fault in the third layer and another have it in the fourth layer, and yet another may have two stacking faults and a fourth have no stacking faults. This ensemble of particles has, on average,

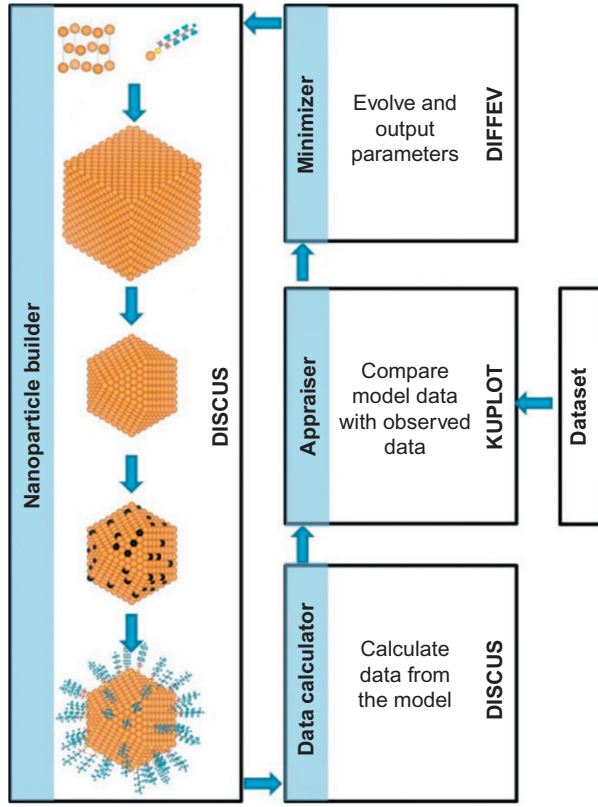

FIGURE 6.8 Schematic of how the DISCUS program can be used to build realistic nanoparticle models and how parameters in the model can be refined using the DIFFEV part of DISCUS (Page *et al.*, 2011).

one stacking fault per particle, and the measured PDF from the real sample will resemble that of the average of the ensemble more than the PDF of any of the individual particles in the ensemble. To build up enough statistics, it is necessary to have hundreds or thousands of nanoparticles in the ensemble, all generated stochastically but with the same set of parameters (such as stacking fault density). In actual fact, it can be argued that this is not Small Box or Big Box *per se*, but an example of complex modeling where prior knowledge in the form of the fact that the nanoparticles are faulted is used to constrain the possible solutions and refinements are done on a small number of parameters such as the stacking fault density. Needless to say, with this approach, as with RMC, it is straightforward to add more datasets and more constraints into the complex. The main drawback of this method is that the regression is slow and that some parameters such as stacking fault density may be determined more quickly by other methods (such as fitting two-phase models of

wurtzite and zinc blende to the data using PDFgui, which if you fit up to a range of four atomic layers should give an accurate value from the proportion of each phase that is refined, though this needs to be validated against PDF data generated by DISCUS). It is also early in its development and its power is expected to grow in the future.

6.3.3.4. Complex Modeling Framework: SrFit

The examples of complex modeling described above suggest a number of things. First, that complex modeling is required for solving many nanostructure problems from powder data: it is not going away. Second, that complex modeling needs to be flexible and extensible to get the most power from it. Third, that the most complicated complexes will require high performance computing to work. These things suggest that a modular, flexible, software framework based on object-oriented programming principles could be very helpful. This is the idea behind the SrFit complex modeling framework. The initial implementation is working and undergoing testing in the Billinge group at the time of writing.

The complex modeling process was first abstracted into its component parts and a design sought that modularized the components.

Complex modeling components:

1. Structure model representation
2. Structure model builder
3. Function calculator
4. Model definition (target function and constraint handler)
5. Minimizers.

In global optimization, "the model" is actually the function to be minimized, or cost function, plus any constraints. For us, the cost function will be some kind of agreement factor between calculated and measured data plus any hand-built restraints which add cost when they are deviated from by too much (e.g., bounds on bond lengths and angles, bond valence sums, and so on). The hard constraints can be expressed mathematically, such as a symmetry constraint that says that $a=b=c$ for a cubic material, for example. The inverse problem "model" should not to be confused with the structure model, which is really what we are after scientifically. We would like to be able to build up the model definition on the fly by adding different contributions to the cost function as desired at runtime, and to have tools for flexibly defining constraints, also on the fly. We would also like, to the extent possible, to be able to use different minimizers on an as-needed basis, for example, as a check for convergence.

The structure model has a *representation*, for example, atom coordinates in a Cartesian box, fractional coordinates in a nonorthonormal crystallographic asymmetric unit, plus symmetry information to build the full

structure, a molecule represented as a Z-matrix, and so on. We would like to be able to choose which representation to use given our particular experimental situation. Indeed, we may want to even mix different representations in the same fit, for example, as done using SrFit to study the growth of SnO_2 nanoparticles out of a solvent containing tin(IV) complexes in solution in a hydrothermal synthesis, as described in Section 9.3.2 (Jensen *et al.*, 2012) where the reaction precursors were modeled as individual molecules and the nanoparticles as fragments of crystal in the same fit. Next, we want to be able to define different function calculators for each structure model depending on what measurements we have, for example, neutron PDF, X-ray PDF, small-angle scattering, EXAFS, and so on. The software should also be as easy as possible to extend, so that new function calculators can be written by users and, if possible, made available to other users.

The basic framework for SrFit was developed as part of the NSF-funded DANSE project. It exists and is undergoing testing, though at the time of writing it does not have an easy-to-use interface, which will be developed later in the project. It is written in Python and C++. It is open source and available to Python compliant users who want to try it out by contacting Prof. Billinge at Columbia University.

6.4. *AB INITIO* NANOSTRUCTURE SOLUTION FROM PDF DATA

Outside the world of glasses and liquids, virtually all of the PDF modeling that has been done to date is refinement of a known (or guessed) starting model. However, crystal structures may be *solved* from powder diffraction data. Is it possible to solve, rather than refine, nanoparticle structures from PDF data? There are a number of questions that must be addressed. First, is the problem well posed? that is, is there enough information in the PDF to yield a unique 3D structure model? Second, how can we find the solution? As discussed above, the simulated annealing algorithm is a global optimizer and is widely used, but it is an inefficient algorithm. It works well for ill-posed models, for example, of amorphous materials where there are a multitude of solutions and the pathway to the minimum is relatively smooth, but it had trouble reconstructing clusters larger than a few tens of atoms from idealized PDF data because it tends to be trapped in local minima and will not reach the global minimum. For most more well-posed problems, it is actually used as a local (or probably more correctly a regional) minimizer rather than as a global optimizer in the RMC refinement programs such as RMCprofile, described above.

Without an efficient algorithm for solving the problem, it is not even possible to answer the uniqueness question. There is no uniqueness theorem for the nanostructure problem but a practical answer may be obtained by trial and error: by testing whether multiple optimizations that converge to the global minimum

give the same structure or not. This can only be done once an algorithm can be found that can access the global minimum in reasonable time.

A canonical system to test is C_{60}, a moderate sized, unique, rigid cluster whose structure is known and for which it is straightforward to get good data. Furthermore, a genetic algorithm had been proposed that could find the structure of C_{60} that minimized the total energy of a cluster of 60 atoms given the interatomic potential (Deaven and Ho, 1995). Could this same algorithm be adapted to find the solution when it was not an energy being minimized but χ^2, the fit of the model PDF to the measured PDF, a kind of "reverse genetic algorithm" method? Unfortunately, it proved difficult to make it work (it was finally made to work, but proved less efficient and less robust than the alternative LIGA algorithm described here (Juhas *et al*., 2006)) and a new algorithm was sought. The basic idea behind the new LIGA algorithm was that the PDF function is an *r*-dependent function. It might be possible to find candidate clusters that fit the low-*r* region well and then build out such promising candidates to larger size. There will still be many clusters, and many of them incorrect subclusters that nonetheless are consistent with the low-*r* part of the PDF, so a backtracking method is required so that when promising but wrong clusters are built out, they can be taken apart again and rebuilt differently. The algorithm is described in detail in Juhas *et al*. (2006), with all the implementational details described in Juhas *et al*. (2008). The resulting algorithm worked and gave the first *ab initio* structure solution from PDF data (Juhas *et al*., 2006). The structural solutions are shown in Fig. 6.9. Figure 6.10 shows some intermediate clusters on the way to a final solution for the 88-atom Lennard-Jones cluster. The Lennard-Jones clusters can be found online and are often used to benchmark minimization algorithms, where it is normally energy that is being minimized. In this case, the atomic pair distances were extracted from the clusters and the list of distances used as a target for LIGA to reconstruct the structure.

The algorithm needs methods for growing and shrinking clusters and for scoring clusters. The information in the list of interatomic distances from the PDF is actually used to build up the clusters since it is known that only atom pairs with these distances are allowed. In this way, when building a cluster by adding a new atom, three distances can be selected from the available distances in the list, and three atoms selected at random on the cluster. The new atom is then added to the cluster at the point of intersection of the three distances that emanate from the three atoms on the cluster. This more or less uniquely places the atom (it is not quite unique in the early stages of growth when the cluster is planar). A number of clusters are grown in this way to different sizes. The clusters are then scored by looking at all the interatomic distances on the cluster and comparing them to the allowed distances coming from the PDF. The further a distance is away from its nearest allowed distance in the table, the more error it contributes. In this way, both the total error contained in a cluster and marginal error contributed by each atom are

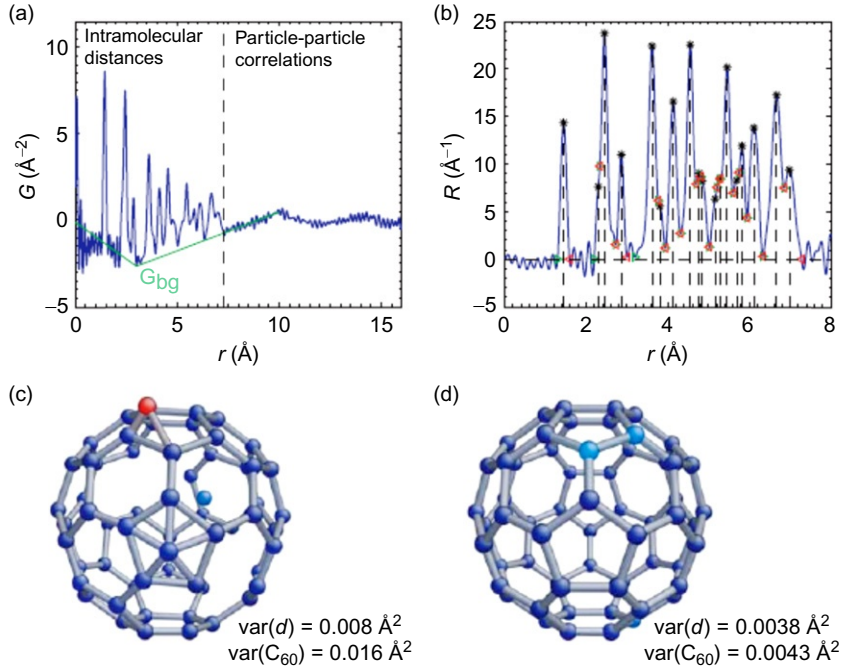

FIGURE 6.9 (a) The PDF data from C_{60} that were used to solve the C_{60} structure. The PDF of the molecule sits on top of a background that was removed to get the RDF shown in (b). Peak positions and heights were then determined and this was the information that was fed to the algorithm. The data contain noise and other aberrations. When the model was forced to use the absolute peak heights from the data, the best solution contained defects as shown in (c). However, when the peak positions were enforced but their height was not, the correct solution was found as shown in (d) (Juhas *et al.*, 2006).

determined. Clusters of the same size compete with each other and the clusters with the lowest total error are considered the best clusters and "promoted" to the next level by adding an atom. Clusters with high error are "relegated" to a lower division by removing an atom. The atom that is chosen for removal is done stochastically, but weighted by the error that the atom contributes to the cluster. The similarity to the promotion/relegation cycle in European soccer leagues was the inspiration for the name, LIGA, of the algorithm. Well-performing teams get promoted to the next division up and, playing in front of larger crowds, can afford to buy new players. Poorly performing clusters face relegation. When they go down, they sell one of their worst players. The analogy to soccer is only cute, but what makes the algorithm powerful in this context is that it uses the available information very well to limit the search space. The algorithm could find the solution to C_{60} from ideal data in less than a second and to noisy data in a few minutes, on a laptop computer (Juhas *et al.*, 2006, 2008).

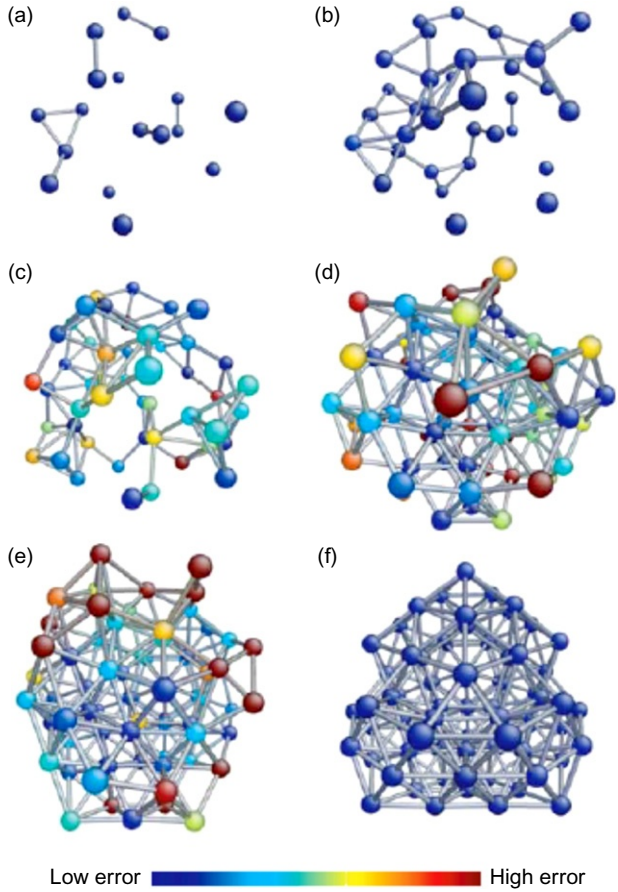

FIGURE 6.10 Example of intermediate clusters, and the final successful cluster, obtained during testing of the algorithm. The target was to find the lowest energy cluster from 88 atoms interacting through a Lennard-Jones potential, so-called LJ-88. The Lennard-Jones clusters can be found online and are often used to benchmark minimization algorithms where it is energy that is being minimized (Juhas *et al*., 2006).

The algorithm works particularly well for elemental systems where the integrated intensity of the PDF peak is actually the multiplicity of that interatomic distance. For compounds, the PDF peak height is weighted by the scattering power of the atoms involved, turning an integer algorithm into a floating point one and increasing the complexity of the problem because not only the geometry but also the combinatorics of coloring (correctly placing different "color" atoms on the atomic sites) must be solved. It turns out that, in the spirit of complex modeling, chemically reasonable constraints such as ensuring a different atom on a nearest-neighbor site (a dissimilarity constraint)

is the best way to do the coloring. In this way, LIGA was used to solve a number of known crystal structures of compounds (Juhas *et al*., 2010).

More recently, the addition of one more constraint to the C_{60} problem, the similarity constraint ensuring that every atom has the same environment, allowed the structure to be solved *ab initio* using simulated annealing (Cliffe *et al*., 2010). This is an easier problem to solve than the original problem where the structure was recreated from the data alone: it has an added symmetry (in fact, the symmetry of C_{60} is so high that if you know the full symmetry of the molecule, you actually only need to specify two parameters to solve the C_{60} structure; David *et al*., 1993). However, it shows how creative use of constraints can really help solving nanostructure problems and gives hope that complex modeling will be successful in practice as well as just sensible in theory.

REFERENCES

Beni, G. & Platzman, P.M. (1976) *Phys. Rev. B*, **14**, 1514.

Billinge, S.J.L. (1992) PhD Thesis, University of Pennsylvania.

Billinge, S.J.L. (1998) S.J.L. Billinge & M.F. Thorpe (Eds.), *Local Structure from Diffraction* (p. 137). New York: Plenum Press.

Billinge, S.J.L. (2010) *Physics*, **3**, 25.

Billinge, S.J.L., Davies, P.K., Egami, T. & Catlow, C.R.A. (1991) *Phys. Rev. B*, **43**, 10340.

Billinge, S.J.L., DiFrancesco, R.G., Kwei, G.H., Neumeier, J.J. & Thompson, J.D. (1996) *Phys. Rev. Lett*., **77**, 715.

Binder, K. & Heerman, D.W. (1992) *Monte Carlo Simulation in Statistical Physics* (2nd Edition). Berlin: Springer.

Chupas, P.J., Chaudhuri, S., Hanson, J.C., Qiu, X.Y., Lee, P.L., Shastri, S.D., Billinge, S.J.L. & Grey, C.P. (2004) *J. Am. Chem. Soc*., **126**, 4756.

Claye, A. & Fischer, J.E. (1999) *Electrochim. Acta*, **45**, 107.

Cliffe, M.J., Dove, M.T., Drabold, D.A. & Goodwin, A.L. (2010) *Phys. Rev. Lett*., **104**, 4.

David, W.I.F. (2002) *Structure Determination from Powder Diffraction Data*. Oxford: Oxford University Press.

David, W.I.F., Ibberson, R.M. & Matsuo, T. (1993) *Proc. R. Soc. Lond. A*, **442**, 129.

Deaven, D.M. & Ho, K.M. (1995) *Phys. Rev. Lett*., **75**, 288.

Debye, P. (1912) *Annalen Der Physik*, **39**, 789.

Dmowski, W., Contescu, C.I., Llobet, A., Gallego, N.C. & Egami, T. (2012) *J. Phys. Chem. C*, **116**, 2946.

Dmowski, W., Toby, B.H., Egami, T., Subramanian, M.A., Gopalakrishnan, J. & Sleight, A.W. (1988) *Phys. Rev. Lett*., **61**, 2608.

Dove, M.T., Keen, D.A., Hannon, A.C. & Swainson, I.P. (1997) *Phys. Chem. Miner*., **24**, 311.

Evrard, G. & Pusztai, L. (2005) *J. Phys. Condens. Matter*, **17**, S1.

Farrow, C.L., Juhas, P., Liu, J.W., Bryndin, D., Bozin, E.S., Bloch, J., Proffen, T. & Billinge, S.J.L. (2007) *J. Phys. Condens. Matter*, **19**, 335219.

Gereben, O., Jovari, P., Temleitner, L. & Pusztai, L. (2007) *J. Optoelectron. Adv. Mater*., **9**, 3021.

Gerold, V. & Kern, J. (1987) *Acta Metall*., **35**, 393.

Gutmann, M.S., Billinge, S.J.L., Brosha, E.L. & Kwei, G.H. (2000) *Phys. Rev. B*, **61**, 11762.

Hansen, J.P. & McDonald, I.R. (1986) *Theory of Simple Liquids*. London: Academic Press.
Hargreaves, R., Bowron, D.T. & Edler, K. (2011) *J. Am. Chem. Soc*., **133**, 16524.
Huang, K. (1987) *Statistical Mechanics* (2nd Edition). New York: Wiley.
Jadzinsky, P.D., Calero, G., Ackerson, C.J., Bushnell, D.A. & Kornberg, R.D. (2007) *Science*, **318**, 430.
Jensen, K.M.O., Christensen, M., Juhas, P., Tyrsted, C., Bojesen, E.D., Lock, N., Billinge, S.J.L. & Iversen, B.B. (2012) *J. Am. Chem. Soc*., **134**, 6785.
Jeong, I.-K., Heffner, R.H., Graf, M.J. & Billinge, S.J.L. (2003) *Phys. Rev. B*, **67**, 104301.
Jeong, I.-K., Proffen, T., Mohiuddin-Jacobs, F. & Billinge, S.J.L. (1999) *J. Phys. Chem. A*, **103**, 921.
Jeong, I.-K., Mohiuddin-Jacobs, F., Petkov, V., Billinge, S.J.L. & Kycia, S. (2001) *Phys. Rev. B*, **63**, 205202.
Juhas, P., Cherba, D.M., Duxbury, P.M., Punch, W.F. & Billinge, S.J.L. (2006) *Nature*, **440**, 655.
Juhas, P., Granlund, L., Duxbury, P.M., Punch, W.F. & Billinge, S.J.L. (2008) *Acta Crystallogr. A*, **64**, 631.
Juhas, P., Granlund, L., Gujarathi, R., Duxbury, P.M. & Billinge, S.J.L. (2010) *J. Appl. Crystallogr*., **43**, 623.
Kaplow, R., Averbach, B.L. & Strong, S.L. (1964) *J. Phys. Chem. Solids*, **25**, 1195.
Kaplow, R., Rowe, T.A. & Averbach, B.L. (1968) *Phys. Rev*., **168**, 1068.
Keen, D.A. (1998) S.J.L. Billinge & M.F. Thorpe (Eds.), *Local Structure from Diffraction*. New York: Plenum Press.
Keen, D.A., Hayes, W. & McGreevy, R.L. (1990) *J. Phys. Condens. Matter*, **2**, 2773.
Kittel, C. (1996) *Introduction to Solid State Physics* (7th Edition). New York: Wiley.
Krayzman, V., Levin, I. & Tucker, M.G. (2008) *J. Appl. Crystallogr*., **41**, 705.
Krayzman, V., Levin, I., Woicik, J.C., Proffen, T., Vanderah, T.A. & Tucker, M.G. (2009) *J. Appl. Crystallogr*., **42**, 867.
Lagneborg, R. & Kaplow, R. (1967) *Acta Metall*., **15**, 13.
Landron, C., Hennet, L., Jenkins, T.E., Greaves, G.N., Coutures, J.P. & Soper, A.K. (2001) *Phys. Rev. Lett*., **86**, 4839.
Levashov, V.A., Billinge, S.J.L. & Thorpe, M.F. (2007) *J. Comput. Chem*., **28**, 1865.
Livet, F. (1987) *Acta Metall*., **35**, 2915.
Masadeh, A.S., Bozin, E.S., Farrow, C.L., Paglia, G., Juhás, P., Karkamkar, A., Kanatzidis, M.G. & Billinge, S.J.L. (2007) *Phys. Rev. B*, **76**, 115413.
McGreevy, R.L. & Puszati, L. (1988) *Mol. Simul*., **1**, 359.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) *J. Chem. Phys*., **21**, 1087.
Miao, J.W., Charalambous, P., Kirz, J. & Sayre, D. (1999) *Nature*, **400**, 342.
Mikkelson, J.C. & Boyce, J.B. (1982) *Phys. Rev. Lett*., **49**, 1412.
Montfrooij, W., McGreevy, R.L., Hadfield, R.A. & Andersen, N.-H. (1996) *J. Appl. Crystallogr*., **29**, 285.
Morse, P.M. (1929) *Phys. Rev*., **33**, 0932.
Neder, R.B. & Proffen, T. (2008) *Diffuse Scattering and Defect Structure Simulations: A Cook Book Using the Program DISCUS*. Oxford: Oxford University Press.
Nield, V.M., Keen, D.A. & McGreevy, R.L. (1995) *Acta Crystallogr. A*, **51**, 763.
Norberg, S.T., Tucker, M.G. & Hull, S. (2009) *J. Appl. Crystallogr*., **42**, 179.
Page, K., Hood, T.C., Proffen, T. & Neder, R.B. (2011) *J. Appl. Crystallogr*., **44**, 327.
Peterson, P.F., Proffen, T., Jeong, I.-K., Billinge, S.J.L., Choi, K.-S., Kanatzidis, M.G. & Radaelli, P.G. (2001) *Phys. Rev. B*, **63**, 165211.

Petkov, V., DiFrancesco, R.G., Billinge, S.J.L. & Acharya Foley, H.C. (1999) *Philos. Mag. B*, **79**, 1519.
Proffen, T. & Billinge, S.J.L. (1999) *J. Appl. Crystallogr.*, **32**, 572.
Proffen, T., DiFrancesco, R.G., Billinge, S.J.L., Brosha, E.L. & Kwei, G.H. (1999) *Phys. Rev. B*, **60**, 9973.
Renninger, A.L. & Averbach, B.L. (1973) *Acta Crystallogr. B*, **29**, 1583.
Rietveld, H.M. (1969) *J. Appl. Crystallogr.*, **2**, 65.
Sartbaeva, A., Wells, S.A., Thorpe, M.F., Bozin, E.S. & Billinge, S.J.L. (2007) *Phys. Rev. Lett.*, **99**, 4.
Sayre, D. (1952) *Acta Crystallogr.*, **5**, 843.
Scott, M.C., Chen, C.C., Mecklenburg, M., Zhu, C., Xu, R., Ercius, P., Dahmen, U., Regan, B.C. & Miao, J.W. (2012) *Nature*, **483**, 444.
Soper, A.K. (1996) *Chem. Phys.*, **202**, 295–306.
Soper, A.K. (2000) *Chem. Phys.*, **258**, 121.
Soper, A.K. (2001) *Mol. Phys.*, **99**(17), 1503–1516.
Soper, A.K. (2007) *J. Phys. Condens. Matter*, **19**, 15.
Soper, A.K. (2011) *J. Phys. Condens. Matter*, **23**, 8.
Tian, P. & Billinge, S.J.L. (2011) *Z. Kristallogr.*, **226**, 898.
Toby, B.H., Egami, T., Jorgensen, J.D. & Subramanian, M.A. (1990) *Phys. Rev. Lett.*, **64**, 2414.
Tucker, M.G., Dove, M.T. & Keen, D.A. (2000) *J. Phys. Condens. Matter*, **12**, L425.
Tucker, M.G., Dove, M.T. & Keen, D.A. (2001) *J. Appl. Crystallogr.*, **34**, 630.
Tucker, M.G., Keen, D.A., Dove, M.T., Goodwin, A.L. & Hui, Q. (2007) *J. Phys. Condens. Matter*, **19**, 16.
Von Dreele, R.B., Stephens, P.W., Smith, G.D. & Blessing, R.H. (2000) *Acta Crystallogr. D*, **56**, 1549.
Warren, B.E. (1934) *J. Chem. Phys.*, **2**, 551.
Waseda, Y. (1980) *The Structure of Non-Crystalline Materials*. New York: McGraw-Hill.
Wells, S.A., Dove, M.T. & Tucker, M.G. (2002) *J. Phys. Condens. Matter*, **14**, 4567.
Wells, S., Dove, M. & Tucker, M. (2004) *J. Appl. Crystallogr.*, **37**, 536.
Willis, B.T.M. & Pryor, A.W. (1975) *Thermal Vibrations in Crystallography*. London: Cambridge University Press.
Wyckoff, R.M.G. (1963) *Crystal Structures* (2nd Edition). New York: Interscience.
Young, R.A. (1993) *The Rietveld Method International Union of Crystallography Monograph*. Oxford: Oxford Science Publications.
Zuo, J.M., Vartanyants, I., Gao, M., Zhang, R. & Nagahara, L.A. (2003) *Science*, **300**, 1419.