Journal of
Applied
Crystallography
ISSN 0021-8898

Received 16 July 2013
Accepted 12 September 2013

Analyzing diffuse scattering with supercomputers

T. M. Michels-Clark,[a,b]* V. E. Lynch,[b] C. M. Hoffmann,[b] J. Hauser,[c] T. Weber,[d] R. Harrison[e] and H. B. Bürgi[c,f]

[a]University of Tennessee, Department of Chemistry, 552 Beuhler Hall, 1420 Circle Drive, Knoxville, TN 37996-1600, USA, [b]Oak Ridge National Laboratory, PO Box 2008, Oak Ridge, TN 37831, USA, [c]Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, CH-3012 Bern, Switzerland, [d]ETH Zürich Laboratorium für Kristallographie, Wolfgang-Pauli-Strasse 10, 8093 Zürich, Switzerland, [e]Stony Brook University, Department of Applied Mathematics and Statistics, Stony Brook, NY 11794-3600, USA, and [f]Institute of Organic Chemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland. Correspondence e-mail: tmichels-clark@ion.chem.utk.edu

Two new approaches to quantitatively analyze diffuse diffraction intensities from faulted layer stacking are reported. The parameters of a probability-based growth model are determined with two iterative global optimization methods: a genetic algorithm (GA) and particle swarm optimization (PSO). The results are compared with those from a third global optimization method, a differential evolution (DE) algorithm [Storn & Price (1997). *J. Global Optim.* **11**, 341–359]. The algorithm efficiencies in the early and late stages of iteration are compared. The accuracy of the optimized parameters improves with increasing size of the simulated crystal volume. The wall clock time for computing quite large crystal volumes can be kept within reasonable limits by the parallel calculation of many crystals (clones) generated for each model parameter set on a super- or grid computer. The faulted layer stacking in single crystals of trigonal three-pointed-star-shaped tris(bicylco[2.1.1]hexeno)benzene molecules serves as an example for the numerical computations. Based on numerical values of seven model parameters (reference parameters), nearly noise-free reference intensities of 14 diffuse streaks were simulated from 1280 clones, each consisting of 96 000 layers (reference crystal). The parameters derived from the reference intensities with GA, PSO and DE were compared with the original reference parameters as a function of the simulated total crystal volume. The statistical distribution of structural motifs in the simulated crystals is in good agreement with that in the reference crystal. The results found with the growth model for layer stacking disorder are applicable to other disorder types and modeling techniques, Monte Carlo in particular.

© 2013 International Union of Crystallography
Printed in Singapore – all rights reserved

source: https://doi.org/10.7892/boris.58136 | downloaded: 12.10.2024

1. Introduction

Interesting and exploitable macroscopic material properties of functional single crystals are often related to microscopic local deviations from a periodic average structure. These deviations may take the form of static or dynamic disorder and manifest themselves as diffuse scattering (DS) in one, two or three dimensions (rods, layers, clouds) concomitant with Bragg scattering (Welberry, 2004). Phonon-related scattering resulting in thermal diffuse scattering (TDS) (Willis & Pryor, 1975) is not the focus of this investigation. Here we focus on static disorder associated with local structural variations. Qualitative DS analysis can show general aspects of disorder, but only a quantitative analysis can reveal details of the deviations from the average (or Bragg) structure and provide a basis for explaining the origin of the functional properties (Welberry & Goossens, 2008; Aebischer *et al.*, 2006). Only the disordered atoms or molecules within the overall structure contribute to diffuse scattering, which, being distributed over extensive volumes of reciprocal space, is usually orders of magnitude weaker per unit volume of scattering space than Bragg diffraction. This means that measuring the diffraction of structurally disordered materials requires a careful experimental setup at powerful neutron and synchrotron sources and careful discrimination of the experimental noise from scattering introduced by the sample environment and not by the crystal itself. Better radiation sources, detectors and data reduction routines make the acquisition of reliable diffuse scattering data increasingly tractable.

Usually, preliminary knowledge of the structural disorder is vague and consists mainly of chemical or geometric rules that are violated in the average structure. The measured, quantitative information requires quantitative modeling, which necessitates iterative optimization of empirical disorder parameters. Growth or Monte Carlo (MC) models are most

Table 1
The transition matrix **T** of probabilities for extending the left-hand column of four-layer motifs into the top row of new four-layer motifs.

Once a new motif is formed by adding a new fifth layer, the first layer and thus the first motif are dropped. The meaning of the symbols is described in the text.

	To								
From	$\ldots eb_{\rm L}$	$\ldots eb_{\rm R}$	$\ldots ee$	$\ldots b_{\rm L}b_{\rm R}$	$\ldots b_{\rm L}b_{\rm L}$	$\ldots b_{\rm L}e$	$\ldots b_{\rm R}b_{\rm L}$	$\ldots b_{\rm R}b_{\rm R}$	$\ldots b_{\rm R}e$
$eb_{\rm L}$	0	0	0	$t-\Delta/2$	$c-\Delta/2$	$e_1+\Delta$	0	0	0
$eb_{\rm R}$	0	0	0	0	0	0	$t-\Delta/2$	$c-\Delta/2$	$e_1+\Delta$
ee	$(1-e_2)/2$	$(1-e_2)/2$	e_2	0	0	0	0	0	0
$b_{\rm L}b_{\rm R}$	0	0	0	0	0	0	$t-\Delta$	$c+\Delta$	e_1
$b_{\rm L}b_{\rm L}$	0	0	0	$t+\Delta$	$c-\Delta$	e_1	0	0	0
$b_{\rm L}e$	$(1-e_2)/2-\Delta$	$(1-e_2)/2+\Delta$	e_2	0	0	0	0	0	0
$b_{\rm R}b_{\rm L}$	0	0	0	$t-\Delta$	$c+\Delta$	e_1	0	0	0
$b_{\rm R}b_{\rm R}$	0	0	0	0	0	0	$t+\Delta$	$c-\Delta$	e_1
$b_{\rm R}e$	$(1-e_2)/2+\Delta$	$(1-e_2)/2-\Delta$	e_2	0	0	0	0	0	0

effective for estimating an initial disorder model and equilibrating it (Proffen & Welberry, 1998). Lattice energy minimizations have also been used to qualitatively verify the local structural disorder, and the energy-minimized structures were found to qualitatively reproduce the observed diffuse diffraction pattern well (Schmidt & Glinnemann, 2012). Intrinsic issues with these techniques are twofold: Firstly, the size of the constructed model crystals needs to be sufficiently large to encompass the observed disorder (short-range-ordered) motifs, but is generally still small compared to the scattering volume of the actual sample. Secondly, a global optimization procedure is needed to find the best empirical parameters for describing the disorder and their numerical values. The substantial computational resources needed to resolve these issues are becoming increasingly available with advances in computer technology.

In this work three aspects of computationally modeling stacking disorder are investigated with the help of growth models (Wilson, 1962). (1) We compare the efficiency of different algorithms for global optimization of model parameters, namely, differential evolution (Storn & Price, 1997), a genetic algorithm (Holland, 1975) and particle swarm optimization (Kennedy & Eberhart, 1995). (2) We analyze the speckle-type intensity variations inherent in all procedures for modeling disorder using the concept of 'clones', *i.e.* model crystals that are independently built from a single set of model parameters. (3) The concept of clones lends itself to parallelization on super- or grid computers. Here we report on the scalability of such parallelization.

In §2 the chemical model system is described and growth modeling of stacking disorder is sketched. §3 describes three global optimization methods for the parameters of the growth model and ways to parallelize the computations. §4 compares the performance of the optimization methods and of the parallelization. This section also summarizes the structural results obtained. Conclusions are given in §5.

2. Chemical model system, growth modeling and reference data

Tris(bicyclo[2.1.1]hexeno)benzene (TBHB) crystallizes in multiple stacking variants of threefold symmetric layers consisting of coplanar three-pointed-star-shaped molecules (Fig. 1). The polymorph with space group $P6_3/mmc$ and Bragg lattice parameters $a' = b' = 5.2145\,(5)$ Å and $c' = 8.9429\,(8)$ Å shows diffuse streaks of scattering intensity at non-integral values of $(-h'+k')/3$ (with $h', k' =$ integer) (Birkedal *et al.*, 2003). The streaks indicate faulted layer stacking (Bürgi, Hostettler *et al.*, 2005). The unit cell of a single layer is $\mathbf{a} = \mathbf{a}' - \mathbf{b}'$, $\mathbf{b} = \mathbf{a}' + 2\mathbf{b}'$, with lengths $a = b = 3^{1/2}a'$. Correspondingly the diffuse streaks are indexed as hkL, with $-h+k \neq 3$ and L the continuous variable along the streaks. The unit along L was chosen as $c = 2c'$.

The disorder has been described with growth modeling, a procedure in which a new layer is added onto the preceding layers of a crystal. Addition in different positions is associated with different probabilities. The probability of each added layer depends on the arrangement of the preceding layers (Bürgi, Hostettler *et al.*, 2005), four of them in the present case. Selected growth sequences labeled with a shorthand and the symbols of the associated stacking probabilities are shown in Fig. 2.

The symbols of the shorthand refer to three layers: the symbol e (for eclipsed) implies that layer $n+2$ sits exactly on top of layer n; $b_{\rm L}$ ($b_{\rm R}$) means that layers n (lowest), $n+1$ and $n+2$ (highest) spiral in a clockwise (anticlockwise) fashion when looking onto the growing crystal face. Fig. 2 uses this nomenclature to describe transitions from four- to five-layer

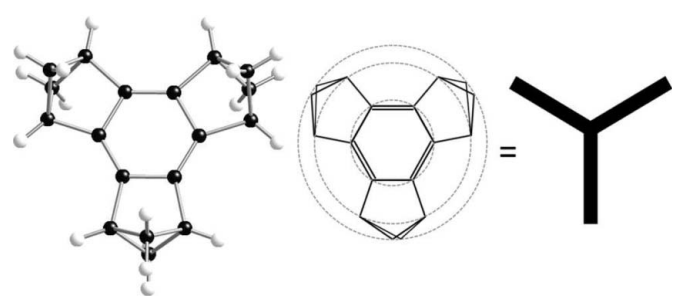

Figure 1
(Left) Tris(bicyclo[2.1.1]hexeno)benzene molecule (TBHB): black atoms are carbon, white atoms are hydrogen. (Center) Skeletal formula representation of TBHB, where all atoms on a given dotted circle were assigned the same isotropic atomic displacement parameter (U_1, U_2, U_3). (Right) Schematic tristar representation of TBHB (used in Fig. 2).

Table 2
Comparison of the reference parameters defining the reference data with mean model parameters and their standard deviations obtained by optimizations with a genetic algorithm (GA), differential evolution (DE) and particle swarm optimization (PSO) using 40 individuals.

The parameter values at the start of the optimizations are randomly generated for each member in the population within the range of values listed in the third column for each parameter. The minimum (min.) and maximum (max.) values are the world size or absolute limits of the parameter values during optimization.

Parameter	Reference	Min., max.	GA	DE	PSO
c	0.48877	0, 0.5	0.49697 (3)	0.4883 (5)	0.488 (2)
Δ†	0.49336	−0.5, 0.5	0.45 (1)	0.482 (7)	0.486 (7)
e_2	0.006748	0, 1.0	0.7372 (6)	0.008 (4)	0.02 (3)
Tilt‡	2.2723	−5.0, 5.0	2.251 (8)	2.27 (2)	2.3 (1)
U_1§	2.6284	0, 5.0	2.67 (1)	2.65 (2)	2.7 (1)
U_2§	2.2734	0, 5.0	2.026 (5)	2.27 (4)	2.2 (2)
U_3§	3.0005	0, 5.0	2.982 (9)	2.97 (6)	3.0 (2)
t†	0.48877	0, 0.5	0.49697 (3)	0.4883 (5)	0.488 (2)
e_1†	0.022452	0, 1.0	0.0061 (7)	0.0235 (7)	0.02 (3)
R¶	–	–	0.0338 (3)	0.0091 (5)	0.02 (1)

† Constrained parameters: $c = t$, $(c \pm \Delta) + (t \mp \Delta) + e_1 = 1$. ‡ Units of tilt in degrees. § Units of U_i in 10^{-2} Å^2. ¶ R factor measures fitness after 150 generations.

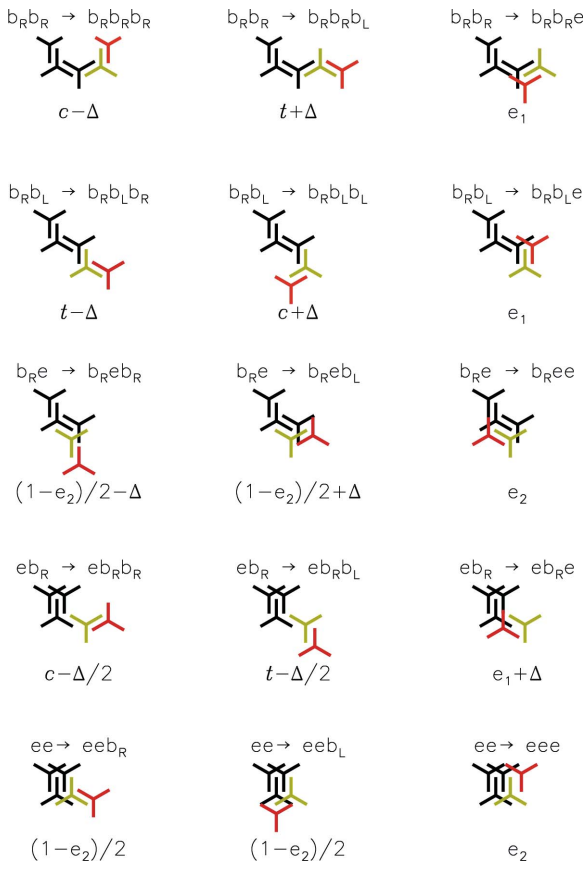

Figure 2
Examples of unique five-layer stackings and associated motifs: layers 1, 2, 3 in solid black, layer 4 in light gray (green in the electronic version of the journal), and layer 5 in dark gray (red); respective transition probabilities from four- to five-layer stackings are shown at the bottom of the motif. The symbols above the motif describe the four- and five-layer stacks. The symbols b_L, b_R, e (bent left, bent right, eclipsed) characterize the three possible three-layer stacks. A four-layer stack is defined by two symbols, and a five-layer stack by three symbols.

stacking sequences. The full Markov matrix of transition probabilities is given in Table 1. Adding a new layer on the right-hand side of the column vector 'from' generates the sequence in the top row vector 'to' of the transition matrix; in short $\mathbf{p}' = \mathbf{pT}$, where $\mathbf{p}$ and $\mathbf{p}'$ are row vectors describing the probabilities of finding a given four-layer sequence before and a five-layer sequence after adding a new layer, respectively; $\mathbf{T}$ is the matrix of transition probabilities. For the sequences b_R and b_L, the molecules in layer $n+1$ are tilted out of the trigonal plane, but not for the sequence e. Chemically equivalent atoms are assigned the same isotropic mean-square displacement parameter (Fig. 1).

Nearly noise-free diffuse intensity data were obtained from 1280 clones, generated with a disorder model derived from an experimental study of a crystal of TBHB (Hauser *et al.*, 2009). The parameters of the model used in this work are the stacking probabilities, a molecular tilt angle and isotropic atomic displacement parameters. A total of 1280 virtual model crystals (clones), each consisting of 0.96×10^5 layers, were grown from a single set of parameters that best fit the 14 experimentally determined hkL lines (Hauser *et al.*, 2009) (Table 2). Each clone was divided into 1600 randomly chosen lots (Proffen & Welberry, 1998; Welberry & Butler, 1994) encompassing 60 layers. Diffuse intensities were calculated by Fourier transformation of each lot. The calculated 2.048×10^6 sets of intensities were then incoherently averaged to create a reference data set of 14 hkL lines including $0kL$ ($k = 1, 2, 4, 5, 7$), $1kL$ ($k = -3, -4, 5, 6$), $2kL$ ($k = 3, 4, 6$), $3kL$ ($k = 4, 5$) and $0 < L < 5$ for all lines. As an example the reference and optimized model intensities of the $01L$ line are compared in Fig. 3.

3. Computations

3.1. Methods for optimizing model parameters

In general, the initial values of the parameters chosen for modeling disorder are 'educated guesses' at best, usually far

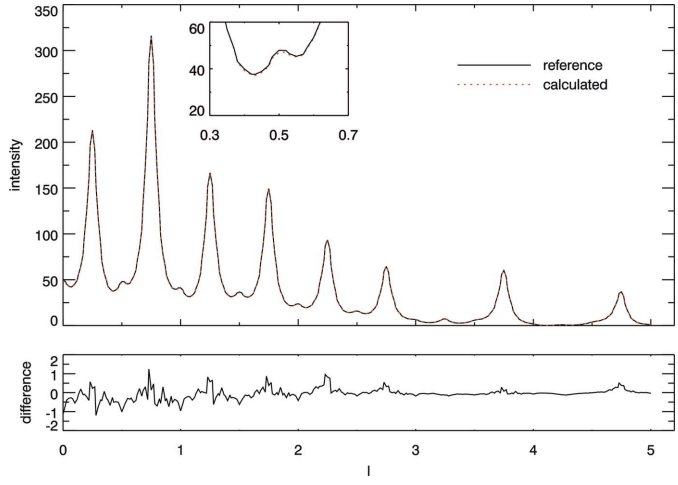

Figure 3
Reference intensity and calculated intensity (averaged over 40 clones) for the $01L$ line are shown overlaid. The difference (calculated − reference) is shown in the plot below (see Table 4, $\overline{R} = 0.0077$).

from their real value. Therefore, a global optimization technique that is not based on sophisticated *a priori* knowledge but is able to optimize sets of random initial model parameters is needed. Population-based metaheuristic algorithms are well suited for the purpose of optimization without preliminary assumptions of the solution.

We selected three representative algorithms for numerical optimization: a genetic algorithm (GA), differential evolution (DE) and particle swarm optimization (PSO). GA and DE are population-based search algorithms that implement principles of genetics. Each individual gene set k in a population is evaluated according to its fitness R_k (high fitness = low R_k).

$$R_k = \left\{ \frac{\sum_i \left[\sum_j (I^j_{i,k}/J) - I_{i,\mathrm{ref}} \right]^2 w_i}{\sum_i (I_{i,\mathrm{ref}})^2 w_i} \right\}^{1/2}, \qquad (1)$$

where the sum over i includes all I data points $I^j_{i,k}$ from 14 diffuse lines ($I = 301 \times 14 = 4214$; weight $w_i = 1$), and the sum over j includes all J clones. All intensities are given unit weight in the calculations. The resulting population of R values, R_k (corresponding to K model parameter sets), is characterized by its mean and standard deviation:

$$\overline{R} = \sum_k \frac{R_k}{K}, \quad s = \left[\frac{\sum_k (R_k - \overline{R})^2}{K-1} \right]^{1/2}. \qquad (2)$$

The individuals yielding the lower R_k values in a comparison between parents and children survive and form the parents for the next generation. This process is repeated until a stopping criterion is reached, in our case a set number of generations. Population convergence to a solution is signalled by a low overall $\overline{R}$ accompanied by a low s value, which essentially remain constant over many generations.

Like GA and DE, PSO is a population-based stochastic search technique; however, it does not use genetic operators. Instead the position and movement of each particle in a swarm is adjusted with respect to the overall trend of velocity and direction for the swarm. Convergence to a solution is achieved when all particles have a small displacement and have clustered together.

3.1.1. Genetic algorithm. GA is a widely used evolutionary algorithm and is described in detail elsewhere (Holland, 1975; Goldberg, 1994; Lucasius & Kateman, 1993; Srinivas & Patnaik, 1994; Gallagher & Sambridge, 1994; Forrest, 1993). The values for the initial generation are randomly generated within a set range for each parameter (Table 2). Cycling through crossover, mutation and selection sequentially creates the subsequent generations. The control parameters of the algorithm are crossover rate and mutation rate (Table 3). During a GA optimization, new individuals are generated from two randomly chosen individuals (genotypes) of a generation, crossover is applied by recombining the parameter vectors at a random point, and then a mutation is applied by randomly selecting and changing parameters. Since GA uses a bit-wise representation of the parameters (genes) during numerical optimization, so-called Hamming cliffs occur when flipping a randomly chosen bit in the binary representation.

Table 3
The control and run parameters for DE, GA and PSO.

Algorithm control parameters.

GA	DE	PSO
Mutation rate = 0.005	$f_\mathrm{m} = 0.70$	$w = 0.95$
Crossover = 0.95	$f_\mathrm{r} = 0.50$	$c_1 = 1.0$
		$c_2 = 1.0$
		$V_\mathrm{max} = 0.25 \times$ (limits†)

† See Table 2 min. and max.

Run parameters.

No. of generations†	150
Population size‡	40
No. of clones§	10–40

† User specified. ‡ Individuals (genotypes) included in the population or swarm. § Copies of each genotype included in the calculation.

This may change parameter values drastically. In order to avoid Hamming cliffs, the genes are represented by so-called 'Gray codes' (Gray, 1953). Once the genetic procedure for the population is complete, the fitness of the individuals in the new generation and the parent generation are compared with the objective function R_k. The individual with the better fitness, either the parent or the new candidate, survives to serve as parent for the next generation. This process is repeated until a stopping criterion is reached, in our case a maximum number of generations (Table 3).

3.1.2. Differential evolution. DE is a vector-based method which has been successfully used for numerical optimization problems and problems that are parameterized with real numbers (Storn & Price, 1997). Application of DE to disorder modeling and the interpretation of diffuse scattering has been described in detail by Weber & Bürgi (2002). DE forms a child, an individual of the subsequent generation, by picking genes from a target individual (t) with gene vector $\mathbf{d}_t$ in the parent generation and from a mate $\mathbf{d}'_c$ created from three randomly chosen parent individuals a, b and c. The three vectors $\mathbf{d}_a$, $\mathbf{d}_b$ and $\mathbf{d}_c$ are combined to create $\mathbf{d}'_c$, according to $\mathbf{d}'_c = \mathbf{d}_c + f_\mathrm{m}(\mathbf{d}_a - \mathbf{d}_b)$, where f_m is a scalar mutation constant, a control parameter of the algorithm. If any gene of a mate $\mathbf{d}'_c$ is outside the set search range, the mate is rejected and a new mate calculated. To create the child, one randomly selected parameter (gene) in t is replaced by the corresponding gene from $\mathbf{d}'_c$, and the remaining genes of the child are inherited from $\mathbf{d}'_c$ with a probability given by the crossover constant f_r, another control parameter of the algorithm. The control parameters used for DE are listed in Table 3. The survival of either the target individual or the child to the next generation is determined by which of the two has the higher fitness. DE is repeated until a stopping criterion is reached, here a maximum number of generations (Table 3).

3.1.3. Particle swarm optimization. PSO is modeled after the behavior of swarms such as birds or insects in nature (Kennedy & Eberhart, 1995). A vector $\mathbf{x}_i$ of parameters defines each individual or particle i within the swarm. Each particle is guided to the optimal solution by the best solution it

has seen plus the best solution seen by the population. The initial population for PSO is randomly generated as for the GA and DE algorithms. Unlike GA and DE, PSO does not operate on the principles of genetics; instead, each particle or candidate solution $\mathbf{x}_i$ is assigned a displacement per unit of time (t), generally referred to as velocity $\mathbf{v}_i$, by which the particle travels the search space. Each variable in $\mathbf{x}_i$ is updated from one generation to the next with $\mathbf{v}_i'$ according to $\mathbf{x}_i' = \mathbf{x}_i + \mathbf{v}_i' t$. The velocity $\mathbf{v}_i'$ modifying the current parameter vector $\mathbf{x}_i$ is influenced by the best solutions seen by the particle, $\mathbf{b}_i$, and the population, $\mathbf{b}_{\rm P}$. It is updated according to

$$\mathbf{v}_i' = w\mathbf{v}_i + c_1 r_1(\mathbf{b}_{\rm P} - \mathbf{x}_i) + c_2 r_2(\mathbf{b}_i - \mathbf{x}_i). \qquad (3)$$

The algorithm depends on the following control parameters:

(a) inertia parameter w (generally <1);

(b) acceleration constants c_1 and c_2 (indicating how much the particle vector $\mathbf{x}_i$ is directed toward the best solution that is seen by the swarm, $\mathbf{b}_{\rm P}$, and the particle, $\mathbf{b}_i$, respectively);

(c) random numbers r_1 and r_2 generated within the range $(-V_{\max}, V_{\max})$.

The control parameters for PSO used in this work are listed in Table 3. Any parameter of $\mathbf{x}_i$ that is outside of the search space is reset to its limits.

3.2. Clones and parallelization

In previous DE optimizations it was observed that the fitness of certain individuals (intensities) was so high that their genes survived many generations. For reasons of computational efficiency, the disordered crystals, their intensities and the fitness of such individuals were not recalculated in subsequent generations. On calculating many individuals and their fitness with the same gene set, it was found that the fitness values covered a distribution of R values. The fitness of individuals surviving many generations was invariably found at the high-fitness end of such distributions. Conversely, the fitness of individuals whose gene sets were frequently replaced was often found at the low-fitness end, even though the average fitness of the distribution might have been quite reasonable. These two phenomena lead to the conclusion that a reliable estimate of fitness requires multiple intensity calculations with the same gene set, subsequent averaging of the intensities, and calculation of a fitness from the averaged intensities (Bürgi, Hauser *et al.*, 2005). Crystals originating from the same parameter or gene set are referred to as clones.

The phenomenon of ultra-stable individuals may be understood in terms of the physical background of a diffuse scattering experiment. Coherent scattering of an object lacking translational symmetry results in a speckle pattern. Locally the scattering intensity in such a pattern may change rapidly. Small differences in the object produce slightly different speckle patterns. The coherence length of X-rays used in diffuse scattering experiments on disordered crystals is typically smaller than the sample size. The experimentally observed signal is thus an incoherent superposition of different speckle patterns originating from slightly offset regions within the sample. Apart from experimental noise, such signals usually look quite smooth. The phenomenon of incoherent superposition is simulated in our disorder modeling with clones, albeit at a much smaller scale. While the real sample may contain of the order of 10^{18} (slightly differing) unit cells, a typical clone consists of a mere 10^4–10^6 unit cells, which are divided up into lots of dimensions that are chosen to match the correlation length defining the short-range order (Welberry & Goossens, 2008; Proffen & Welberry, 1997). The lots are Fourier transformed and averaged incoherently. If the variations of the resulting averaged simulated pattern, also called MC noise, are of the same order as the experimental noise, MC noise may or may not match the experimental noise, thus giving the false impression of unusually high or unusually low fitness of the model. Simulating diffuse scattering patterns with model crystals of inadequate size and insufficient numbers of lots hence runs the risk of mistaking noise in the experimental pattern as being the result of disorder, thus explaining the phenomenon of the unjustified survival of some of the model crystals as described above. To reduce this risk, the volume of simulated crystals must be large enough and thus the calculated diffuse pattern must be smooth enough to minimize bias due to MC noise.

Random number bias is another problem. Disorder modeling usually starts from a randomly seeded crystal that is grown into a full-sized model crystal or equilibrated by using an MC process. Such crystals may be biased by the starting configuration. Building several crystals, each starting from a different random seed, minimizes the risk of random number bias and reduces the probability of ultra-stable individuals.

A disadvantage of clones is the increase of computational cost and a corresponding slowdown of the structure determination process. The latter can be compensated for effectively by parallel computation of the clones, one per processor of a supercomputer or a grid computing facility, as will be discussed in §4.2. In the present case of nearly noise-free data, the use of clones serves to analyze and control the inherent dispersion of results characteristic for crystal growth and MC models (§4.3).

There is an additional dimension to parallelization. The global optimization methods discussed above explore parameter space by calculating individuals and their clones with many different gene sets or swarm particles. Thus, the computation of J clones for K gene sets is easily distributed over JK compute nodes. The efficiency of parallelization is limited only by the amount of communication necessary between the nodes. In the present case this corresponds essentially to the transfer and averaging of the J clone intensities for each of the K individuals. Unless specifically mentioned, optimizations were performed with 40 gene sets, and fitness was calculated from averaging over 20 clones.

Computing resources were provided for the project by the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory (ORNL) and the US National Science Foundation's TeraGrid cyber infrastructure project. The Oak Ridge Institutional Cluster (OIC) at ORNL is a combined 3136 core shared cluster with a 29 teraflop peak performance. This work was run on the SNS data analysis share of the OIC, consisting

of unrestricted parallel use of up to 192 cores, grouped in clusters of eight cores per node. The Extreme Science and Engineering Digital Environment (XSEDE), previously TeraGrid, is a national shared cluster for US NSF users. It encompasses over 20 different computational resources with over 2686 teraflops of combined performance. An allocation of 195 000 core hours with access to five different supercomputers within TeraGrid was granted and used for this work.

4. Results and discussion

4.1. Comparison of global optimization methods

Comparison of global parameter optimization by GA, DE and PSO was performed using 20 clones for each calculation. All three methods show a rapid decrease of $\overline{R}$, within the first 60 generations. The further decrease of $\overline{R}$ in following generations is gradual. GA, DE and PSO start with an average fitness $\overline{R}$ of 0.75, 0.75 and 0.69, respectively, in generation zero. $\overline{R}$ of PSO drops fastest, followed by that of GA. While PSO and GA are seemingly leveled, $\overline{R}$ of DE keeps reducing. At generation 63, GA is surpassed by DE, which shows an $\overline{R}$ (s) value of 0.03 (1). Finally, DE outperforms PSO in generation 82, with an $\overline{R}$ (s) value of 0.019 (4). The log scale in Fig. 4 emphasizes the differences in convergence. The uncertainty $s(\overline{R})$ also decreases. The distributions of $\overline{R}$ become narrower by a factor of ~2 between generations 64 and 82. DE converges to a population with $\overline{R} = 0.0091$ (5), compared to the starting range of $\overline{R} = 0.76$ (23). The distribution of uncertainties of $\overline{R}$ values [0.0338 (6) for GA and 0.02 (1) for PSO] is more than an order of magnitude greater for PSO than for GA.

Both PSO and DE converged to essentially the same parameter values, but for PSO the uncertainties in the final generation are larger than for DE and accompanied by inferior overall population fitness. GA converged to similar parameter values to DE and PSO with the exception of e_2 (Table 2). We hypothesized that GA converged into a local rather than the global minimum, which is supported by a higher $\overline{R}$ value in the final generation. This hypothesis was tested by changing the pathological parameter in small steps while keeping the remaining parameters fixed at their refined values. The $\overline{R}$ value was expected to cross a fitness barrier to arrive at the real solution; however, $\overline{R}$ decreased continuously without going through a maximum. Thus, GA possesses similar global optimization power to DE and PSO, but the local optimization power of GA seems inherently weaker than that of PSO or DE.

As illustrated in Figs. 5(a)–5(g) for DE optimization, the convergence behavior of the different model parameters varies considerably. Figs. 5(a)–5(h) show the maximum, average and minimum parameter values (filled circles) and the standard deviation of the population (vertical lines) for every parameter in every generation. The average $\overline{R}$ value is cut in half after five generations, then again after seven, nine and ~20 generations, showing the decrease in efficiency as the optimization progresses (Fig. 5h). Parameter c converges at the same rate as $\overline{R}$ (Fig. 5a), while the other parameters trail behind. Parameters c and t (constrained to be numerically equal) represent the probability of bent stacking, which is energetically favorable compared to eclipsed stacking (b_R, b_L *versus* e), the former being prevalent in the structure. Thus parameters c and t contribute more than the other model parameters toward modeling the reference intensity and the overall fit.

In contrast to c, e_2 is slower to converge (Fig. 5c). The e_2 parameter is defined as the probability of continuing an

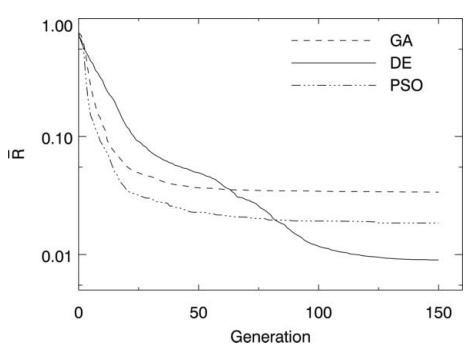

Figure 4
Evolution of population fitness $\overline{R}$ for three global optimization algorithms: general genetic algorithm (GA), differential evolution (DE) and particle swarm (PSO) as a function of generation number.

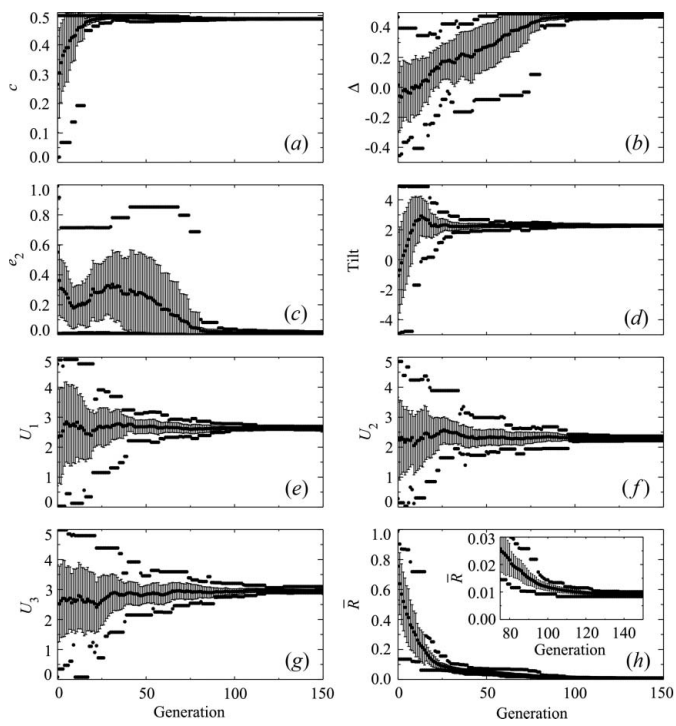

Figure 5
Behavior of the model parameters c, Δ, e_2, tilt, U_1, U_2 and U_3 (a)–(g) and fitness $\overline{R}$ (h) during DE optimization with 40 individuals and 20 clones each. The population mean is indicated by the middle points, and the population standard deviation by the vertical lines. The dots above and below the lines represent the largest and the smallest parameter values in the population.

Table 4
Summary of results after 150 generations of DE optimization with 40 gene sets that have different numbers of layers, lots per crystal and numbers of clones.

The number of cores used, wall clock time and the CPU hours are also reported for each calculation.

Size\layers		96 000	96 000	96 000	96 000	6000	600	600	600
No. of lots		1600	1600	1600	1600	100	10	10	10
No. of clones	Reference	40	20	10	1	16	160	40	10
c	0.48877	0.4884 (3)	0.4883 (4)	0.4883 (5)	0.487 (1)	0.488 (2)	0.487 (2)	0.488 (2)	0.487 (8)
Δ†	0.49336	0.486 (6)	0.482 (7)	0.479 (8)	0.45 (2)	0.452 (22)	0.444 (23)	0.41 (5)	0.25 (11)
e_2	0.006748	0.005 (4)	0.008 (4)	0.014 (10)	0.003 (21)	0.036 (4)	0.0343 (25)	0.057 (7)	0.2 (2)
Tilt	2.2723	2.27 (1)	2.27 (2)	2.27 (3)	2.28 (13)	2.29 (7)	2.27 (8)	2.27 (8)	2.28 (40)
U_1	2.6284	2.651 (13)	2.65 (2)	2.63 (4)	2.61 (13)	2.65 (10)	2.63 (10)	2.6 (2)	2.8 (5)
U_2	2.2734	2.27 (2)	2.27 (4)	2.27 (6)	2.34 (14)	2.31 (20)	2.28 (13)	2.3 (3)	2.5 (6)
U_3	3.0005	2.97 (3)	2.97 (6)	2.99 (8)	3.0 (3)	3.02 (19)	3.01 (22)	3.0 (4)	3.1 (8)
t†	0.48877	0.4884 (3)	0.4883 (4)	0.4883 (5)	0.487 (1)	0.488 (2)	0.487 (2)	0.488 (2)	0.487 (8)
e_1†	0.022452	0.0231 (6)	0.0235 (7)	0.023 (1)	0.026 (2)	0.025 (3)	0.024 (3)	0.024 (4)	0.03 (2)
R‡	–	0.0077 (3)	0.0091 (5)	0.011 (9)	0.029 (3)	0.0297 (3)	0.0288 (3)	0.051 (6)	0.12 (3)
Cores	–	160	160	160	40	160	160	160	160
Wall clock time (h)	–	136	63	27	8	6	57	30	9
CPU hours = cores × wall clock time	–	21 760	10 080	4320	320	960	9120	4800	1440

† Constrained parameters: $c = t$, $(c \pm \Delta) + (t \mp \Delta) + e_1 = 1$. ‡ R factor measures fitness.

energetically unfavorable eclipsed arrangement. Since this stacking option has a low probability, it is infrequently present in the local structure and therefore contributes modestly toward fitting the reference intensity. The Δ parameter (Fig. 5*b*), which distinguishes between the layer stacking either continuing with the same helicity ($+\Delta$) or changing helicity ($-\Delta$), begins to converge approximately at generation 60 and continues to converge within a small uncertainty in the final generation (Table 2, column 4). Since Δ determines the details of bent motifs, it is associated with c and represents a significant determinant of the model intensities.

The tilt parameter, which defines the degree of molecular out-of-plane tilt allowed in layer $n+1$ of a $b_{\rm R}$ or $b_{\rm L}$ (but not an e) motif, refines to its optimal value in approximately 80 generations (Fig. 5*d* and Table 2, column 4). The U_i values, representing the atomic displacement parameters of TBHB, all converge at the same rate, settling to an optimal value in approximately 100 generations (Figs. 5*e*–5*g* and Table 2, column 4).

Generally, small standard deviations of the parameters indicate that the originally quite different 40 gene sets have converged to a single solution. The ratio between the standard deviations and the mean parameter values after 150 generations of DE optimization is in the range of 0.003% for c and 2% for U_3 (Table 2), indicating convergence of the optimization to a single solution.

To summarize this section, we tentatively conclude that PSO can initially navigate the search space most efficiently since all of the variables change simultaneously toward the best solution seen by $\mathbf{b}_i$ and $\mathbf{b}_{\rm P}$: PSO outperforms DE and GA in the first 20 cycles of optimization (Fig. 5). However, towards the end of the optimization process, the collective fitness of the models obtained by DE is better than that from GA and PSO. Initializing an optimization with several generations of PSO followed by DE will most likely make the best use of computational resources. When the parameters are fairly well clustered, convergence tends to slow down; it may then be advantageous to conclude the optimization by a numerical least-squares calculation (Welberry, 2004). In the present case, least-squares optimization could start after 50 generations when considering the rapidly converging parameters, or after 100 generations with regard to the slowly converging parameters.

4.2. Influence of clones

The dependence of the model fitness on speckle-type intensity variations has been tested for DE only. The results after 150 generations of DE optimization with different crystal sizes and different numbers of clones are reported in Table 4. As expected $\overline{R}$ decreases on either increasing the crystal size or increasing the number of clones. The decrease in $\overline{R}$ shows a linear trend with the square root of the reciprocal product of the number of clones and the number of lots per clone, *i.e.* the total number of lots included in the calculation (Fig. 6).

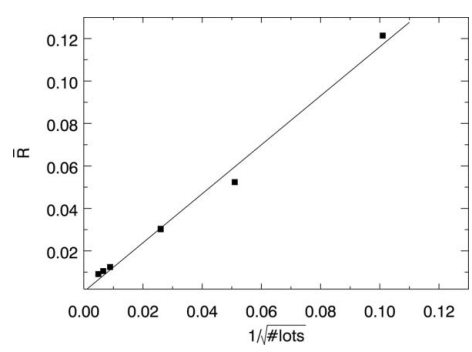

Figure 6
Decrease of $\overline{R}$ with increasing number of lots. $\overline{R}$ depends linearly on the reciprocal square root of the number of lots (#lots).

This behavior indicates that the speckled nature associated with the intensity of individual lots is analogous to noise in an experiment (Sutton *et al.*, 1991). The incorporation of clones reduces the statistical noise inherent in the MC process. On average the standard deviations of the simulated parameters also decrease roughly with the inverse square root of the number of clones (sample size), as expected if the mean parameter values are normally distributed (Table 4). The averages of the parameters over the different gene sets are mostly within one standard deviation of the reference values used to construct the data set (see §2 and Table 4).

Two additional conclusions considering computational efficiency are worth mentioning. Firstly, in columns 6–8 of Table 4, $\overline{R}$ remains approximately constant, showing that decreasing the crystal size can be compensated for by increasing the number of clones, thereby allowing a higher degree of parallelization. Secondly, 40 clones for each individual consisting of 96 000 layers achieved a good, but not perfect, agreement with the reference data, with $\overline{R} = 0.0077$ (3) in the 150th generation (Table 4, column 3). In contrast, the $\overline{R}$ values that can be achieved from state-of-the-art experiments by state-of-the-art structure refinement of disorder models against diffuse scattering data are about 0.1 in favorable cases. As columns 9–10 of Table 4 indicate, a trustworthy estimate of the model parameters may be obtained from crystals of modest size, albeit with relatively large uncertainties.

The most important consequence of replacing large crystals by clones is the possibility to parallelize computations. The calculation time for 40 individuals (without clones) sums up to about 8 h on 40 cores for 150 generations. This corresponds to a total CPU time of $40 \times 8 = 320$ h. As more clones are added, the total CPU time increases from 320 to 4320 to 10 080 to 21 760 h for 1, 10, 20 and 40 clones, respectively (Table 4, columns 3–6). However, if enough cores are available to calculate the fitness of one clone per core, the wall clock time stays approximately the same, as the calculations are performed in parallel. In our case, the predicted scaled numbers, found by dividing the wall clock time by the total number of gene sets and their clones per processor, expand slightly from 8 to 10.8 to 12.6 and 13.6 h. This corresponds to an increase in simulation time by a factor of 1.7 for a calculation that is 40 times larger. The modest expansion of wall clock time is due to increased communication between nodes required for averaging the intensities of the clones. The wall clock times in columns 8–10 of Table 4 follow the same trend as columns 3–6 as more clones are added.

4.3. Motif statistics and comparison with reference model

The purpose of modeling diffuse scattering is to gain insight into the structural motifs composing the crystal, in our case, the types of stacking sequences of TBHB and their lengths. Owing to the probabilistic growth or MC procedures used for building and equilibrating crystals, a model is not expected to be a one-to-one image of the sample investigated. However, it must show the same statistical distribution of structural motifs.

Table 5
Motif statistics of the reference crystals compared with those from the Markov steady-state distribution and those of the best individual in the 150th (final) generation of a DE optimization averaged over 20 and 300 clones (reported as percent of the full crystal).

b without a subscript indicates that either b_L or b_R stacking is allowed.

	Motif statistics (%)			
Motif types	Reference	Markov†	20 clones	300 clones
Eclipsed (*ee*)	0.0284 (63)	0.033	0.032 (7)	0.033 (7)
Bent ($b_L b$)	45.75 (18)	45.594	45.61 (21)	45.61 (19)
Bent ($b_R b$)	45.74 (18)	45.594	45.52 (18)	45.68 (20)
Mixed eclipsed and bent (*eb*)	4.24 (11)	4.388	4.42 (9)	4.39 (10)
Mixed bent and eclipsed (*be*)	4.24 (11)	4.388	4.42 (9)	4.39 (10)

† Steady-state distribution from Markov transition matrix (Table 1).

To test this, the occurrence of the nine structural motifs eb_L, eb_R, ee, $b_L b_R$, $b_L b_L$, $b_L e$, $b_R b_L$, $b_R b_R$ and $b_R e$ in the 1280 virtual reference crystals is compared with that found for the best individual in the 150th (final) generation of a DE optimization and with that calculated from the transition matrix $\mathbf{T}$.

The limiting values of the structural motif probabilities may be found from $\mathbf{p}^n = \mathbf{p}\mathbf{T}^n$, as the number n of layers added approaches infinity. It corresponds to the steady-state distribution $\overline{\mathbf{p}}$, where

$$\begin{aligned}\overline{\mathbf{p}} &= [\overline{p}(eb_L) \quad \overline{p}(eb_R) \quad \overline{p}(ee) \quad \overline{p}(b_L b_R) \quad \overline{p}(b_L b_L) \\ &\quad \overline{p}(b_L e) \quad \overline{p}(b_R b_L) \quad \overline{p}(b_R b_R) \quad \overline{p}(b_R e)] \\ &= [0.02194 \quad 0.02194 \quad 0.00033 \quad 0.22797 \quad 0.22797 \\ &\quad 0.02194 \quad 0.22797 \quad 0.22797 \quad 0.02194]\,. \qquad (4)\end{aligned}$$

The numerical values of the components of $\overline{\mathbf{p}}$ are obtained from the normalized eigenvector of $\mathbf{T}$ with unit eigenvalue. Table 5 gives the four-layer bent left ($b_L b$), bent right ($b_R b$), mixed eclipsed–bent (eb, be) and eclipsed (ee) motif counts in the crystal as percentages (see §2). b without a subscript stands for b_L or b_R.

The statistical distribution of the four-layer motifs in the modeled crystal was also obtained by counting, after 150 generations of DE optimization, the motifs in the best individual averaged over 20 and 300 clones. The motif uncertainties were obtained by calculating the standard deviation of each motif count among its clones. Comparing the results for 20 and 300 clones, it is evident that 20 clones are sufficient to obtain reliable values.

The four-layer motifs containing only bent arrangements were divided into two categories, $b_L b$ and $b_R b$. For symmetry reasons, the frequency of the two motifs should be the same. For the same reason, the frequency of the motifs *eb* and *be* should also be the same. This is found (Table 5). The eclipsed-only motifs, eclipsed in both the second and the third layer, accounted for only 0.033% of the total crystal.

The counts from the growth models agree to within standard errors with the counts from the reference crystals and the limiting values from the Markov model (Table 5, column 3). This shows that the optimized models truly represent the

structural motifs in the crystals from which the reference data set was obtained.

Knowledge of the lengths of repeating motifs (correlation length) is important to ensure that the lot size was chosen appropriately (Welberry & Butler, 1994). The correlation length for the eclipsed case is

$$\langle n_{\rm e}\rangle = \frac{1}{1-\overline{p}(ee)} = \frac{1}{(1-0.00033)} = 1.00033 \qquad (5)$$

layers; for the bent case $\langle n_{\rm b}\rangle$ is given by

$$\langle n_{\rm b}\rangle = \frac{1}{1-[\overline{p}(b_{\rm L}b_{\rm R})+\overline{p}(b_{\rm L}b_{\rm L})+\overline{p}(b_{\rm R}b_{\rm L})+\overline{p}(b_{\rm R}b_{\rm R})]} = \frac{1}{(1-0.9119)} \cong 11.35 \qquad (6)$$

layers (Bürgi, Hostettler *et al.*, 2005).

Continuing an eclipsed stack is of low probability, with a calculated value of 0.008 ± 0.004 (Table 2, column 4). The low probability, e_2 only 2σ above 0, is responsible for the very short average correlation length $\langle n_{\rm e}\rangle$. In contrast, the average length $\langle n_{\rm b}\rangle$ is about 11.35 layers. Both values indicate that a lot size of 60 layers is sufficient to represent the short-range order in crystals of TBHB.

5. Conclusions

In this work we have presented new ways to quantitatively analyze diffuse scattering and have applied them to the one-dimensional stacking disorder described earlier for the organic compound TBHB (Birkedal *et al.*, 2003). Three global optimization algorithms were tested: differential evolution, a general genetic algorithm and particle swarm optimization. All three algorithms converged to similar parameter values, except for one parameter in the genetic algorithm calculations. Particle swarm optimization was found to be most efficient in the initial stages of optimization. After 150 generations of optimization, the parameter values from differential evolution showed the narrowest range and the best agreement between model and reference diffuse intensities. The testing of many sets of parameters is required in all three global optimization algorithms but can be significantly accelerated by parallelization: for each parameter set, model crystals are calculated on separate compute nodes.

In order to reduce the speckle-type intensity variations inherent in the modeling process, intensities may be calculated either from a single large crystal subdivided into many lots or from several smaller crystals consisting of fewer lots but constructed from a single set of modeling parameters (clones). Clones are preferred over a large crystal as they allow further parallelization of the calculation. The use of clones also minimizes any bias that might be associated with the random initial layer configuration that seeds the growth of the model crystal in the modeling process. The dependence of the fitness $\overline{R}$ on the reciprocal square root of the total number of lots (= number of clones times number of lots per clone) was found to be linear, indicating that the behavior of speckle-type intensity variations is analogous to that of experimental noise.

Finally, it was shown that the statistical distribution of four-layer stacking motifs found in the computer simulations was the same within statistical error as that in the reference crystal. These results were verified theoretically using the steady-state probability distribution resulting from the four- to five-layer Markov transition matrix. The chosen lot size of 60 layers was shown to be sufficient, as the largest correlation length was approximately 11.35 layers for a bent arrangement.

The quantitative analysis protocols reported here for analyzing one-dimensional diffuse scattering are applicable not only to crystals with stacking disorder but also to two- and three-dimensional types of structural disorder. A more complex disorder might require more model parameters and thus more individuals as well as larger model crystals that encompass the full range of local structure correlations. However, the parallelized global optimization techniques described in this work will also make such structure determinations feasible if the necessary computing resources are available. The calculation of the diffuse intensities could be further optimized by using GPU processors to Fourier transform the scattering density of the disordered crystals as shown by Gutmann (2010). In addition, they may be combined with other modeling techniques, MC modeling (Proffen & Welberry, 1998) and three-dimensional PDF techniques (Weber & Simonov, 2012) in particular. The resulting local structure variations may then provide a basis for explaining structure–property relationships of disordered materials.

We acknowledge John Cobb and Steve Miller for granting us access to the TeraGrid and OIC computing resources and Phillip Bentley from ESS (European Spallation Source) for the use of PSO and GA algorithms implemented in the optimization suite *Vop*. Many thanks go to Anthony Linden for helpful suggestions and editing and to Michal Chodkiewicz for many fruitful discussions. A special acknowledgment is extended to Andrei Savici for his help in figure and manuscript preparation. We also thank Peter Peterson for his suggestions of computing methods and use of resources. This research was partly funded by a Sinergia project of the Swiss National Science Foundation. This manuscript has been co-authored by the Oak Ridge National Laboratory, managed by UT-Battelle LLC under contract No. DE-AC05-00OR22725 with the US Department of Energy.

References

Aebischer, A., Hostettler, M., Hauser, J., Kramer, K., Weber, T., Güdel, H. U. & Bürgi, H. B. (2006). *Angew. Chem. Int. Ed.* **45**, 2802–2806.

Birkedal, H., Bürgi, H. B., Komatsu, K. & Schwarzenbach, D. (2003). *J. Mol. Struct.* **647**, 233–242.

Bürgi, H. B., Hauser, J., Weber, T. & Neder, R. B. (2005). *Cryst. Growth Des.* **5**, 2073–2083.

Bürgi, H. B., Hostettler, M., Birkedal, H. & Schwarzenbach, D. (2005). *Z. Kristallogr.* **220**, 1066–1075.

Forrest, S. (1993). *Science*, **261**, 872–878.

Gallagher, K. & Sambridge, M. (1994). *Comput. Geosci.* **20**, 1229–1236.

Goldberg, D. E. (1994). *Commun. ACM*, **37**, 113–119.

Gray, F. (1953). US Patent No. 2632058A.
Gutmann, M. J. (2010). *J. Appl. Cryst.* **43**, 250–255.
Hauser, J., Weber, T. & Bürgi, H. B. (2009). Personal communication.
Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Oxford: The University of Michigan Press.
Kennedy, J. & Eberhart, R. (1995). Proceedings of the IEEE International Conference on Neural Networks: Particle Swarm Optimization, 27 November–1 December, Perth, WA, Australia.
Lucasius, C. B. & Kateman, G. (1993). *Chemometr. Intell. Lab. Syst.* **19**, 1–33.
Proffen, Th. & Welberry, T. R. (1997). *Acta Cryst.* A**53**, 202–216.
Proffen, Th. & Welberry, T. R. (1998). *Phase Transitions*, **67**, 373–397.
Schmidt, M. U. & Glinnemann, J. (2012). *Z. Kristallogr.* **227**, 805–817.
Srinivas, M. & Patnaik, L. M. (1994). *Computer*, **27**, 17–26.
Storn, R. & Price, K. (1997). *J. Global Optim.* **11**, 341–359.
Sutton, M., Mochrie, S. G. J., Greytak, T., Nagler, S. E., Berman, L. E., Held, G. A. & Stephenson, G. B. (1991). *Nature*, **352**, 608–610.
Weber, T. & Bürgi, H.-B. (2002). *Acta Cryst.* A**58**, 526–540.
Weber, T. & Simonov, A. (2012). *Z. Kristallogr.* **227**, 238–247.
Welberry, T. R. (2004). *Diffuse X-ray Scattering and Models of Disorder*. New York: Oxford University Press.
Welberry, T. R. & Butler, B. D. (1994). *J. Appl. Cryst.* **27**, 205–231.
Welberry, T. R. & Goossens, D. J. (2008). *Acta Cryst.* A**64**, 23–32.
Willis, B. M. T. & Pryor, A. W. (1975). *Thermal Vibrations in Crystallography*. Cambridge University Press.
Wilson, A. J. C. (1962). *X-ray Optics: the Diffraction of X-rays by Finite and Imperfect Crystals*, 2nd ed., Methuen's Monographs on Physical Subjects. London: Methuen.