Z. Kristallogr. **219** (2004) 847–856

A better FOX: using flexible modelling and maximum likelihood to improve direct-space *ab initio* structure determination from powder diffraction

Vincent Favre-Nicolin*,[I] and Radovan Černý[II]

[I] Université Joseph Fourier and CEA, DRFMC/SP2M/NRS, 17, rue des Martyrs, F-38054 Grenoble Cedex 9, France
[II] Laboratoire de Cristallographie, Université de Genève, 24, quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland

Received June 1, 2004; accepted August 6, 2004

Powder diffraction / Ab initio structure anaylsis / Maximum likelihood / Restraints

Abstract. *Ab initio* structure determination using direct-space methods, although relying on an essentially brute-force approach, can be greatly improved through smarter algorithms. The most basic improvement involves the use of prior information to reduce the number of configurations evaluated to find the structure solution. It is however vitally important that the parametrization used to incorporate this prior information does not reduce the efficiency with which the configuration space is explored. We will show that this can be achieved by defining molecules and polyhedra through a set of restraints associated to dedicated random changes, allowing to solve structures up to three times as fast as with the 'standard' approach where atomic positions are parametrized *directly* from bond lengths, bond angles and dihedral angles.

To further enhance the efficiency of the algorithm, it is also possible to 'tune' the convergence criterion used to compare the structural model to the observed diffraction data (usually χ^2 or R_{wp}). By using Maximum Likelihood principles, it is shown that incorporating the fact that the model is *approximate* in the χ^2 evaluation can improve the algorithm convergence towards the structure solution.

Introduction

Structure determination for small structures (<100 independent atoms) can be seen as a simple task in 2004 Crystallography. Indeed, direct methods (combined with density modification and Fourier recycling) have proven strong enough to solve structures from single crystal data for more than 1000 independent atoms, requiring relatively little computer time.

However this is still not applicable to powder diffraction: indeed, direct methods rely on the precise measurement of individual reflexion intensities, which is generally not possible for powders due to the projection of the 3D reciprocal space on a single dimension. Worse, it often happens that samples for which no single crystal of sufficient size (>10 μm) can be grown also diffract weakly at medium and high resolutions, making the extraction of structure factors with reasonable uncertainties difficult. The development of micro-diffraction (beam size <10 μm) with synchrotron radiation is also showing its limits as compounds may suffer from radiation damage.

All these elements show that despite the considerable algorithmic and technical developments of the last 20 years, direct-space methods are and will still be required to solve crystal structures. The relatively slow speed is compensated by the robustness inherent to the ergodic search.

Global optimization in direct space

A basic direct-space algorithm can be described by very few principles:

(1) **Parametrization:** the structure must be described using continuous or discrete parameters
(2) **Ergodic algorithm:** the algorithm shall vary parameters to eventually explore the entire configuration space, ensuring going through or near the true structure.
(3) **Cost function:** each configuration can be evaluated by using one or several criteria (comparison between calculated and observed diffraction pattern with χ^2 or R_{wp}, energetic evaluation, ...), yielding the "cost" (or "fitness", "penalty", "score", ...) of the configuration, and enabling to decide which configurations are better.

The above principles describe a pure "brute-force" approach where *all* possible configurations are tested, until a global minimum is found. While such an algorithm guarantees finding the structure solution, it will only be adequate for the *very* patient crystallographer, since the number of trials required varies exponentially as a function of the total number of parameters. Positive results were nevertheless reported using grid-search techniques (Masciocchi *et al.* (1994); Chernyshev and Shenk (1998)) For a faster convergence, a few principles should be added to smarten up the algorithm:

* Correspondence author
(e-mail: Vincent.Favre-Nicolin@ujf-grenoble.fr))

(4) **Biased algorithm:** the algorithm, while remaining ergodic, can be biased so that more time is spent near configurations with low costs, and rejects higher costs.

This has long been used for Structure Determination from Powder Diffraction (SDPD) with Monte-Carlo and Simulated Annealing (SA) (Newsam, Deem Freeman, (1992); Harris *et al*., (1994); Andreev, MacGlashan Bruce, (1997); Andreev, Lightfoot and Bruce (1997)), where the probability of a given configuration is proportional to $e^{\frac{-\text{cost}}{T}}$, following a Boltzmann-type distribution with temperature T. The temperature is gradually reduced during the optimization, to focus on the best configuration. When several criteria (R_{wp}, energy) are used, it was also proposed to minimize the two cost functions independently (Brodski, Peschar and Shenk (2003)), rather than use a sum.

In order to avoid being trapped in a local minimum, multiple SA runs can be used, or parallel optimizations using different temperatures in the Parallel Tempering (PT) method (Falcioni and Deem (1999)). Fox (Favre-Nicolin and Černý (2002)) uses PT with an automatic adjustment of the Monte-Cralo temperature.

Another major approach uses Genetic Algorithms (GA) (Shankland, David and Csoka (1997); Kariuki *et al*. (1997); Harris, Johnston and Kariuki (1998)). In this method there is not a single configuration but rather of *population* of configuration tested together, and the biasing comes from (i) the selective survival of the best fitting trial configuration, and (ii) the generation of new trial configurations using both random changes and mating between existing configurations.

(5) **Reducing parameter space:** To accelerate the structure solution the structure description must use the smallest number of parameters by grouping atoms, using *a priori* chemical information.

This can be done using rigid bodies (Dinnebier (1999)), as well as by defining groups of atoms where atomic positions are deduced from previous atoms using bond lengths, bond angles and dihedral angles (Andreev, Lightfoot and Bruce (1997)), generally using a *Z*-matrix approach (Fig. 1).

Principles (1)–(5) were already used for numerous SDPD programs: *PowderSolve* (Engel *et al*., (1999)), *Espoir* (Le Bail, (2001))[1], *Topas* (Bruker AXS, (2000)), *Endeavour* (Putz, Schön and Jansen, (1999)), *Dash* (ex-Druid) (David, Shankland and Shankland, (1998)), *GAP* (Shankland, David and Csoka, (1997)), *GAPSS* (Kariuki *et al*., (1997)), PSSP (Pagola *et al.* (2000)).

In the first version of FOX (Favre-Nicolin and Černý (2002)), several new features were introduced:

– the use of a Dynamical Occupancy Correction, which allows to automatically take into account atoms overlapping with a symmetrical atom (special position) or from a different buiding units, without any *a priori* information about the overlap or special position. This feature is especially important for inorganic and intermetallic compounds, where the structure is generally built from stacked polyhedra, and often presents a high symmetry.
– a modular design, so that all "objects" involved (crystal structures, diffraction datasets) could each supply their criterion to evaluate the model. The description of the crystal structures uses a basic description of groups of atoms (Scatterers) so that the parametrization could be modified independently of the scattering calculations. The global optimization algorithm was also built as a general algorithm, and not specific to crystal structure determination.

Details about general features of the FOX program can be found in the first article (Favre-Nicolin and Černý (2002)); tutorials for inorganic and organic compounds are available in the FOX package, as well as on the website (http://objcryst.sourceforge.net/Fox/). In this article we will focus on improvements to the structure determination algorithms using two added principles:

(6) **Flexible parametrization:** the structure shall be described in a way that reduces parameter space exploration, *without restricting the possible movements between configurations*. A corollary is that the convergence shall not be sensitive to the details of the parametrization.
(7) **Maximum likelihood:** the correct evaluation of each configuration cannot rely solely on the calculation of R_{wp} or χ^2 to compare a model to the experimental diffraction data, but should use Maximum Likelihood principles for a correct evaluation of the approximate models.

Improving the modelling of organic molecules for direct-space solution

The *Z*-matrix description

To describe a group of atoms (a molecule or polyhedron), the simplest approach is to use a *Z*-matrix: all atoms are entered into an ordered list, and their position is defined using one bond length, one bond angle and one dihedral angle relatively to previous atoms in the list (see Fig. 1). The absolute position and orientation of the entire group can be defined by three translation parameters and three Euler angles[2].

This description is the most natural and direct description of any group of atoms, and is very efficient to reduce the number of Degrees of Freedom (DOF) for the structure solution: as bond lengths and angles are generally well known or predictible (within a few 0.01 Å for bond lengths and degrees for angles), the only remaining DOF are the dihedral angles for free torsion bonds, plus the six translation-orientation parameters, all aforementionned parameters being linearly independant.

However this description presents a few pitfalls:

(a) Some structures cannot be adequately described using a *Z*-matrix, when there is at least one cycle.

[1] *Espoir* is free, open-source software which can be downloaded from http://sdpd.univ-lemans.fr/sdpd/espoir/

[2] Note that while we focus on the *Z*-matrix approach, the limitations outlined here apply to *any* parametrization of atomic positions directly from geometrical descriptors.

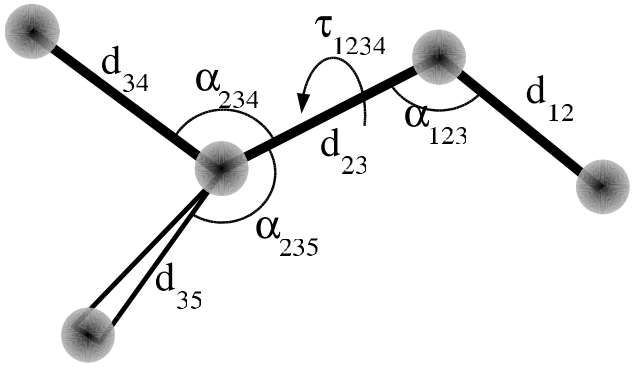

Fig. 1. In a *Z*-matrix description, atomic positions are deduced in order from the position of the atoms preceding them, using bond lengths (d_{ij}), bond angles (α_{ijk}) and torsion (dihedral) angles (τ_{ijkl}). As bond length and angles are generally predictable, this means that conformations are essentially described by a unique set of dihedral (torsion) angles $\{\tau_i\}$. The drawback of this modelling is that any conformation change must be parametrized as a combination of torsion angle changes, which makes more complex the exploration of all possible configurations.

Indeed an atom can only be defined from one end of the cycle, so that the only option not to break the cycle is to make it rigid. This prevents the description of flexible cycles.

(b) The description of the conformation directly from dihedral angles implies that a given molecule can be described by a number of different *Z*-matrices, which are not equivalent (different order of atoms, different choice for torsion angles).

(c) Complex conformation changes (mirroring the structure, twisting an internal chain while keeping the rest of the structure) are impossible or complex to describe directly in terms of dihedral angles, even though they can be 'natural' modifications that an intelligent algorithm could try to find the global minimum.

An extreme consequence of (b) is that since any change in the beginning of the *Z*-matrix will modify the position of all subsequent atoms in the structure, it is therefore required to find first the position of the atoms at the beginning of the structure. This is in practice not entirely true, as the stochastic nature of the random configuration changes (small modifications simultaneously on all DOF) makes it possible to change the structure at the "beginning" (in the *Z*-matrix description) of the molecule, while keeping the "end" unchanged. But this will only happen stochastically, and cannot be driven by an intelligent algorithm. This was pointed out by Favre-Nicolin and Černý (2002).

It was practically demonstrated (Shankland *et al.* (2002)) that using different orders for the atoms in a *Z*-matrix description led to different convergence speeds, which is due to (b) and (c), as the type of conformation changes which are easily possible will depend on the *Z*-matrix chosen. It is possible to work around these effects by a carefully chosen *Z*-matrix, e.g. by describing flexible structures from the center rather than from one end, but this is not trivial and therefore not very practical for the person who only occasionaly solves structures.

(c) should *a priori* not hinder excessively the convergence of the algorithm: the use of random displacements, jointly on all torsion angles, guarantees that *all* possible conformations will be explored. However in practice it can impose more complex changes in parameter space to go from one configuration to another closely related. Worse, there is no guarantee that the χ^2 will remain at reasonable values along the pathway to the new configuration. The *Z*-matrix approach, with its very constrained method, results in artificial barriers on the Hypersurface $\chi^2 = f(parameters)$ which hinders its exploration. It is then necessary to improve the flexibility of the modelling, which was already pointed out by Andreev, MacGlashan and Bruce (1997).

A pure restraints-based approach

In order to keep the advantages of the *Z*-matrix approach (reducing the parameter space) while keeping a complete freedom to the conformation changes that can be directly used, it is possible to use a restraints-based system: all atomic positions are defined directly by their three positional parameters (x, y, z), and the conformation of the molecule is *statistically* imposed through a set of restraints on bond lengths, bond angles and dihedral angles.

These so-called "soft restraints" have been used for protein structure determination (e.g. see Herzberg and Sussman (1982); Chapman (1995)), as well as for small molecules and inorganic structures, including powder diffraction in the GSAS software package (Dinnebier (1999); Larson and Von Dreele (2000); Von Dreele *et al.* (2000); Von Dreele (2001)). We are extending this use to *ab initio* structure determination in direct space. Generally the restraints only use an expected value, and a σ which defines how quickly the cost (or χ^2) rises when the value departs from the ideal value. As during a global optimization the expected values can have a wide range (e.g. inorganic crystals when the valence of metal atoms is *a priori* unknown), we also use an inner range $\pm\delta$ without any penalty: e.g. for a bond d of expected length d_0, we have:

$$\text{if} \quad d \in [d_0 - \delta; d_0 + \delta], \qquad \chi^2_{\text{bond}} = 0\,,$$

$$\text{if} \quad d \leq d_0 - \delta, \qquad \chi^2_{\text{bond}} = \left(\frac{d - (d_0 - \delta)}{\sigma}\right)^2,$$

$$\text{if} \quad d \geq d_0 + \delta, \qquad \chi^2_{\text{bond}} = \left(\frac{d - (d_0 + \delta)}{\sigma}\right)^2.$$

The same is used for bond angles and dihedral angle restraints, and more could be added (imposing the planeity of a group of atoms, etc.). Default values for σ (resp. δ) are 0.01 Å (0.02 Å) for bond lengths, and 0.6° (1.2°) for bond and dihedral angles.

In order to correctly restrain the conformation of the group of atoms, it is necessary to keep the restraints very selective. This cannot be done simply by adding the $\chi^2_{\text{restraints}}$ to the χ^2 from the diffraction data, for two reasons:

(i) the ergodic nature of global optimization algorithms implies that we must explore the entire space, including very improbable configurations.

What we want to achieve is to explore the entire space, with loose restraints with regards to the absolute atomic positions, but tight conformational restraints.

(ii) the range of χ^2 values for the diffraction data is unpredictable, with several order of magnitude differences possible, depending on the background level and on the sample crystallinity. Scaling the $\chi^2_{\text{restraints}}$ with $\chi^2_{\text{diffraction data}}$ is therefore difficult.

To avoid these issues we are using a dual restraints approach: the $\chi^2_{\text{restraints}}$ is evaluated by the Molecule object immediately after a new conformation has been generated: it is then accepted with a Boltzmann-type probability, $\exp\left(\frac{-\chi^2_{\text{restraints}}}{T}\right)$; a new conformation is generated when a trial is rejected. The restraint temperature T[3] is dynamically adjusted to accept 70% of new configurations. The validation of new configuration is therefore made before calculating the diffraction pattern associated with the new structure, which also saves computing time.

Intelligent random moves

The random generation of new configurations for molecules cannot simply be random displacements of individual atoms: otherwise, the imposed 70% acceptance of new configurations would lead to a high restraint temperature and unacceptable configurations, or would require the use of too small atomic displacements to be useful. We use a combination of random configuration changes:

- **rotation around free torsion bonds:** this is similar to the random moves obtained by a *Z*-matrix approach, except that it is possible to rotate *any* bond, and does not depend on the order of either the atoms or restraints. Also, the rotation only changes the side of the bond which has the smallest number of atoms, so that if the rest of the molecule is in a correct position, it will remain in place.
- **rotating single chains:** when one atom is connected to at least three other atoms, it is also possible to rotate a single chain of atoms around the chosen torsion bond, instead of rotating all the chains on a given side of the torsion bond.
- **exchanging atom groups:** it is also possible to exchange the positions of two chains connected to a given atom. This is especially useful when similar chains are connected to a given atom, to 'flip' from a false to the global minimum. This also allows changing the absolute configuration of an asymmetric atom. This is only tried in 5% of trials.
- **random atomic displacements** all atoms are randomly moved with a maximum translation of 0.02 Å along each coordinate.

All these moves are automatically generated at the beginning of a global optimization, and checked with respect to the set of restraints to make sure that they are all allowed (e.g. a dihedral angle restraint around a bond will effectively exclude the bond from all free rotations around it). The user also has an option to label explicitly the free rotation bonds, as well as the option to mark the entire group of atoms as a rigid body.

In addition to these internal changes, the molecule is randomly translated and rotated, the rotation being defined by a quaternion[4] which allows (i) to avoid the 'gimbal lock' when Euler angles fall into special positions, and (ii) more generally allows to sample the orientation space in an isotropic manner.

It is important to note that the given list of random moves for the conformation of a Molecule is not exhaustive: the {individual atoms + restraints} approach allows *any* configuration change (in practice a combination of random changes, *linearly independent or not*) to be trivially implemented at no computing cost, so that new moves can be tuned for different types of atom groups:

- for long chains of atoms it could be possible to use a helicoïdal or sinusoidal change on all or part of the chain
- for large flexible cycles it would be possible to "twist" part of the cycle while keeping the rest fixed
- etc. ...

Finally, most of these configuration changes (except the flipping of atom groups) are continuous and reversible so that it is possible to compute the derivative of the atomic positions (and therefore of the diffraction data) with respect to all the individual moves. This can be useful for a steepest-descent algorithm, molecular-dynamics and the recently proposed Hybrid Monte-Carlo algorithm (Johnston, David, Markvardsen and Shankland (2002)).

In the case of Genetic Algorithms, it is necessary to associate an *absolute* set of the torsion angle values to each configuration, in order to be able to combine these torsions to mate and produce new configurations. In a restraints approach, only torsion angles *variations* are recorded but not their absolute value: it would still be possible to cumulate the torsion changes from a reference structure to obtain the same effect.

Results of the new modeling

We have compared the *Z*-matrix and restraints-based approach on the structure determination of Cimetidine from powder diffraction: this structure is well-known as a test-case for SDPD (Cernik *et al*. (1991)), and the presence of 8 free torsion angles makes it an interesting case of a 'flexible' molecule which should benefit from the new approach.

Figure 2 shows the convergence with both models, for 20 independent runs. The calculation speed is approximately the same for both models (4700 trials/s on a

[3] This temperature is independent from the Simulated Annealing/ Parallel Tempering temperatures used for the global evaluation of the new configuration using other criteria (diffraction data, anti-bump, ...).

[4] http://en.wikipedia.org/wiki/Quaternion.
Quaternions were first used for global optimization of crystal structures from powder diffraction by David, Shankland and Shankland (1998).

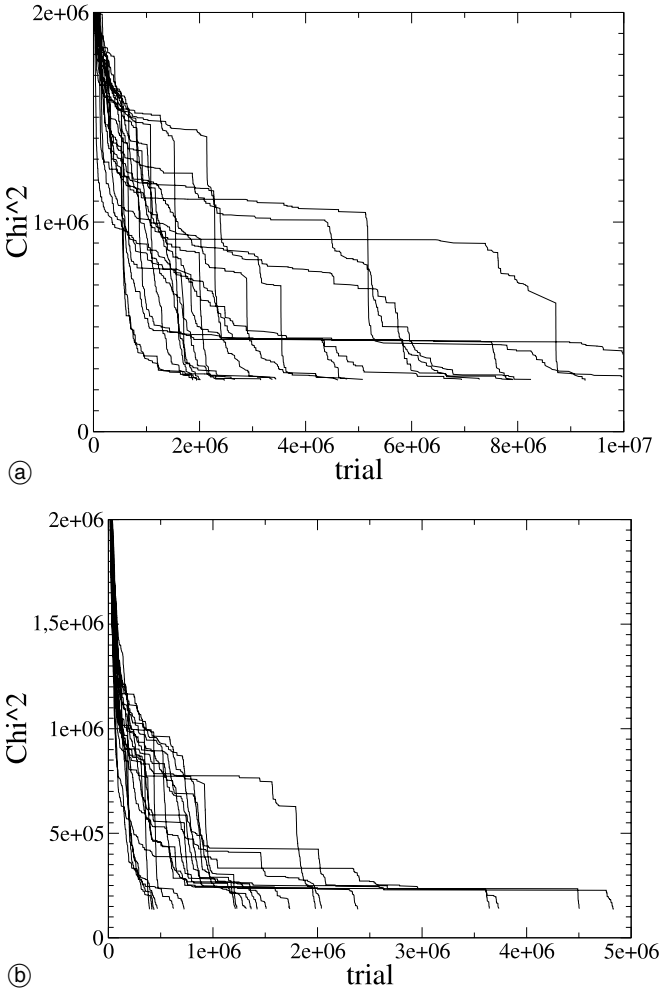

Fig. 2. Convergence of the global optimization $\chi^2 = f(trial)$ for 20 independent runs, using (**a**) a *Z*-matrix description and (**b**) a restraints-based description. The latter allows to use more intelligent random configuration changes while still restraining the exploration of parameter space to chemically sound trial structures, and yields a faster convergence.

2.16 GHz Athlon XP running Linux (kernel 2.6.3)). The average number of trials required to solve the structure[5] is 4.8 million (17 mn) with a *Z*-matrix approach and 1.6 million (5 mn 40 s) for the restraints-based description, i.e. three times as fast. Tests on other molecular structures follow the same trend, with a convergence speed two to three times faster. Inorganic structures, being mostly built of quasi-rigid-bodies, benefit less from the new modelling.

Using maximum likelihood to improve the exploration of the HyperSurface

In order to determine whether a structural model is correct, it is necessary to evaluate the *Likelihood* of that model given the diffraction data, and maximize that likelihood as a function of the model parameters. In practice, it is $-\log$ (Likelihood) which is maximized. Disregarding normalization terms, this means maximizing:

$$-\log\,(\text{Likelihood}) = \chi^2 = \sum \frac{(y_i^{\text{calc}} - y_i^{\text{obs}})^2}{(\sigma_i^{\text{obs}})^2}.$$

This analysis, *where the only sources of error are the experimental uncertainties* σ_i^{obs}, implies that we are able to provide a *perfect structural model and formulas describing the diffraction experiment:* the fit should eventually be perfect if the diffraction data is. This approach is correct for a least-squares refinement, as the objective is indeed to obtain a near-perfect description of the structure. All intermediate structures evaluated are sufficiently near that 'perfect' model to converge using the χ^2 criterion. For a global optimization, the problem is quite different, as we start 'far' from the solution, and we still need to obtain a criterion which tells whether the model is getting better or not. Obviously when far from the solution, the sources of error when comparing the model to the diffraction data is not only coming from the experimental errors, but also from the approximate description of the structure.

Several approaches can be used to help the algorithm move towards the global minimum: begin the optimization with more weight for the low-resolution part of the diffraction data, and progressively evolve to weights obtained from counting statistics. This intuitive approach can be difficult to use as it requires tuning the weight versus the convergence of the algorithm, which will depend largely on the type and quality of the data, and on the complexity of the structure. It is also possible to restrict the space explored by using low-resolution electron density envelopes (Brenner, McCusker and Baerlocher (1997); Brenner, McCusker and Baerlocher (2002)) in which the atoms can evolve, but this will only work for structures where clear envelopes can be generated (e.g. organic structures surrounded by disordered solvant, . . .).

A rigorous approach is provided by Maximum Likelihood (ML) (see Sivia (1996) for an introduction), which has been used in macromolecular crystallography for some time (see Read (1990); Pannu and Read (1996); Murshudov, Vagin and Dodson (1997); Read (2001)). ML allows to implement in the algorithm the fact that we are working with an *approximate structural model:* the atomic positions are approximations of the true structure, associated with a positional error σ_{ML}, and therefore leading to a distribution of calculated patterns rather than a single one.

This ML approach was proposed for powder diffraction refinements (Antoniadus, Berruyer and Filhol (1990)), and has already been used for SDPD by Markvardsen *et al.* (2003) to take into account completely "missing" fragments in a structure, allowing to find a partial structure. We will now show that this can be used in a more general way to improve the convergence of a global optimization algorithm[6].

[5] The algorithm is stopped after falling below a predefined χ^2, and all structures are checked versus the known correct configuration.

[6] There also exist numerour uses of Maximum Likelihood in Crystallography, mostly for phasing and direct methods, which is beyond the scope of this article. For an application to phasing from powder diffraction data, see Bricogne (1991).

Structure factor for an approximate structural model

To describe an *approximate* structural model, we follow Luzzati's (1952) approach and associate every atom at position $\vec{\mathbf{r}}_j^0$ with a translationnal error $\overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j$. The most likely value for the structure factor (which we will assume is only real, for the sake of simplicity) can then be written as:

$$\langle F^{\text{calc}} \rangle = \left\langle \sum_j f_j \cos\left(2\pi\vec{\mathbf{k}} \cdot (\vec{\mathbf{r}}_j^0 + \overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j)\right) \right\rangle$$

$$\langle F^{\text{calc}} \rangle = \sum_j f_j \cos\left(2\pi\vec{\mathbf{k}} \cdot \vec{\mathbf{r}}_j^0\right) D_j$$

where f_j is the atomic scattering factor of atom j, and $\vec{\mathbf{k}}$ the scattering vector. $D_j = \langle \cos\left(2\pi\vec{\mathbf{k}} \cdot \overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j\right) \rangle$ is the Luzzati factor associated with atom j.

Assuming a 3D isotropic gaussian for the positional error $\overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j$, for a sufficiently small $\overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j$, we can write:

$$D_j(\vec{\mathbf{k}}) = \mathrm{e}^{\frac{-2\pi^2}{3}\langle|\overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j|^2\rangle\,|\vec{\mathbf{k}}|^2}$$

and with $\langle|\overrightarrow{\boldsymbol{\delta}\mathbf{r}}_j|^2\rangle = 3\sigma_j^2$, where σ_j^2 is the mean-square displacement of atom j, this can be written as:

$$D_j(\vec{\mathbf{k}}) = \mathrm{e}^{-2\pi^2\sigma_j^2|\vec{\mathbf{k}}|^2} = \mathrm{e}^{-8\pi^2\sigma_j^2\frac{\sin^2\theta}{\lambda^2}}$$

This is exactly the same expression as for an atomic displacement factor, and corresponds to the standard description of thermic disorder in the structure. The important addition in the ML approach is that the calculated structure factor also has a variance, and Luzzati (1952) showed that this variance can be written as[7]:

$$\sigma^2_{F^{\text{calc}}} = \sum_j f_j^2 (1 - D_j^2)$$

The calculated structure factor is then represented by a probability distribution:

$$P(F; F^{\text{calc}}) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{\frac{-(F-\langle F^{\text{calc}}\rangle)^2}{2\sigma^2}}$$

If we assume that all atoms have the same error distribution σ_{ML}, then we have:

$$\sigma^2_{F^{\text{calc}}} = (1 - D^2)\sum_j f_j^2 \quad \text{and} \quad \langle F \rangle = DF^{\text{calc}}$$

which leads to:

$$P(F; F^{\text{calc}}) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{\frac{-(F-DF^{\text{calc}})^2}{2\sigma^2}}$$

We see that an approximate model therefore yields a structure factor damped by the Luzzati factor D, which is equivalent to an atomic displacement factor: the difference between the two is that D represents a lack of information, rather than a disorder. Henceforward it is associated with a variance proportional to $1 - D^2$, i.e. increasing as a first approximation like $\dfrac{\sin^2\theta}{\lambda^2}$. Qualitatively, we can see that this will give increased variances (decreased weights) at high angle.

Applying the ML principles to SDPD

The experimental data points are all independent measurements associated with a gaussian distribution. To derive a computationaly agreeable joint probability $P(model|data)$, it is necessary to derive a gaussian probability associated with each calculated diffraction intensity. For SDPD, this is straightforward when structure factors modulus have been extracted from the powder pattern. However, not all diffraction patterns are suitable for the extraction of structure factors. More importantly, it is precisely those average or low-quality diffraction data that absolutely require global optimization methods to find the structure. For those reasons it was chosen in Fox to perform the optimization directly on the full powder pattern, to be able to work with low-quality, multiple phases, etc.[8]...

The gaussian approximation on the intensities is acceptable since the initial gaussian distribution of atomic displacements is chosen mostly out of mathematical convenience; the approximation is nevertheless quite wrong for $I \ll \sigma(I)$.

$$\langle I^{\text{calc}} \rangle = m \times LP(|\langle F^{\text{calc}}\rangle|^2 + 2\sigma^2_{F^{\text{calc}}})\,,$$

$$\sigma^2_{I^{\text{calc}}} = 2(m \times LP)^2\,\sigma^2_{F^{\text{calc}}}(2\langle F^{\text{calc}}\rangle^2 + \sigma^2_{F^{\text{calc}}})\,,$$

where m is the multiplicity of the reflection, and LP the Lorentz-polarization correction.

Finally, the probability distribution (again gaussian for computational convenience) has to be derived for the full profile: theoretically, this should be done by taking into account the full correlation between the calculated values along one profile, and calculating the joint probability according to this correlation. However in Fox we are not using full but rather *integrated profiles* (Favre-Nicolin and Černý (2002)), so that the quantities compared are the integrated intensities around the expected positions for all reflections. An acceptable approximation consists in imposing that the integrated variance on a reflection profile be equal to the variance on the integrated intensity. The gaussian distribution on successive integration ranges can then be considered reasonably independent.

The likelihood of the model given the diffraction data can then be written as:

$$-\log\,(\text{Likelihood}) = \chi^2_{\text{likelihood}}$$

$$= \sum_i \left[\frac{1}{2}\log\left(2\pi\left(\sigma_i^{\text{obs}\,2} + s^2\sigma_i^{\text{calc}\,2}\right)\right) + \frac{\left(Y_i^{\text{obs}} - s\langle Y\rangle_i^{\text{calc}}\right)^2}{\sigma_i^{\text{obs}\,2} + s^2\sigma_i^{\text{calc}\,2}} \right]$$

where i varies over all the integration segments of the full pattern, σ_i^{obs} and σ_i^{calc} are the observed and calculated un-

[7] This is only true in the absence of symmetry. It is more generally necessary to multiply σ^2 by the *expected intensity factor* ε, which is the number of times a given reflection is left unchanged by a symmetry element (e.g. see Stewart and Karle (1976)).

[8] It is true that extracting the structure factors while keeping the correlation matrix allows to keep the complete information from the raw powder pattern. But this extraction can be difficult for average and low quality data, or patterns with multiple cristalline phases, and the equivalence between the raw powder pattern and the {extracted structure factors; correlation matrix} is only valid if the profiles used describe the diffraction pattern perfectly.

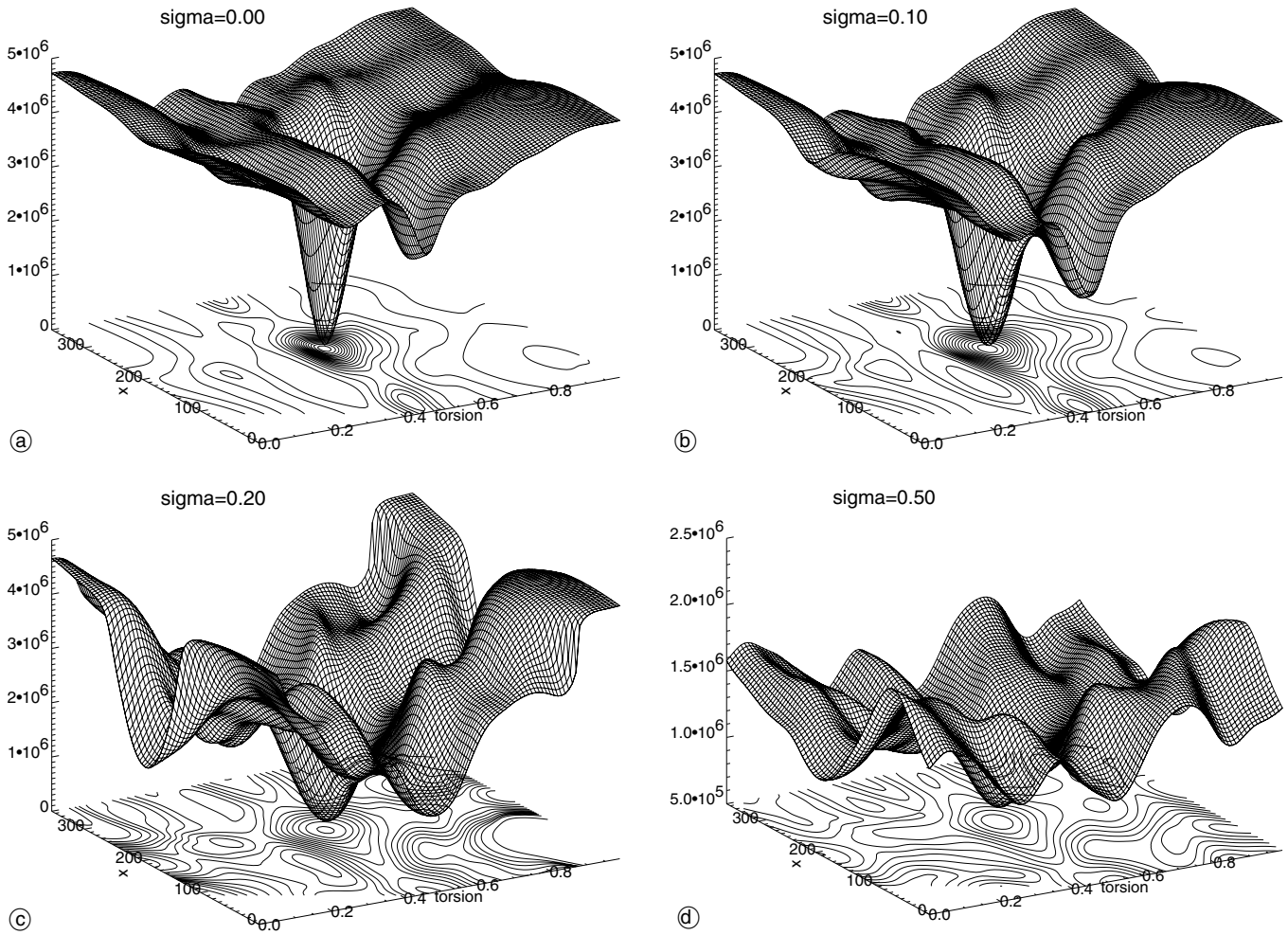

Fig. 3. Modification of the HyperSurface $\chi^2 = f(parameters)$ when introducing Maximum Likelihood model errors for a Cimetidine diffraction data. The 2D cut represents the evolution of χ^2 near the global minimum, when modifying the x coordinate of the molecule atom, and the $N_7 - C_8$ torsion angle. The 2D cut represents for $[0; 2\pi]$ along the torsion angle and $[0; {}^{1}/_{2}]$ for the x translation. (**a**) without atomic positional error. (**b**) $\sigma_{ML} = 0.1$Å (**c**) $\sigma_{ML} = 0.2$Å (**d**) $\sigma_{ML} = 0.5$Å. Taking into account a statistical positional error of all atoms allows to enlarge the global minimum, effectively increasing the radius of convergence of the algorithm. However for large σ_{ML}, the algorithm begins to have difficulties discriminating configurations which differ only by misplacing one or two atoms, due to the large flat global minimum.

certainties, Y_i^{obs} (resp. $\langle Y \rangle_i$) are the integrated observed intensity (resp. the most likely calculated integrated intensity), and s[9] is the scale factor between observed and calculated patterns. The first term corresponds to the normalization of the probability; it is usually forgotten but is required here, lest the maximum likelihood be reached for infinite σ_{ML}.

Effect of maximum likelihood

We have first evaluated the influence of the introduction of a ML positional error on the aspect of the HyperSurface, using the cimetidine powder pattern (Cernik *et al.* (1991)). We have plotted χ^2 as a function of two parameters (the x translation of the molecule and the torsion angle of the $N_7 - C_8$ bond). The 2D maps are represented in Fig. 3 for different values of the atomic position uncertainty σ_{ML}.

The HyperSurface shows a clear widening of the global minimum, which increases with σ_{ML}. This implies that the algorithm should find this minimum more easily. However for larger σ_{ML} values, the minimum becomes so large that the algorithm would not be able to discriminate trial structures that correspond to similar electronic densities, due to the 'blurring' introduced by the ML error.

We have also tested the effect of a ML error on the convergence of the algorithm, using the Cimetidine structure (Cernik *et al.* (1991)). Due to the complex calculations, the speed (2.16 GHz Athlon XP running Linux 2.6.3) decreases from 4700 to 3800 trials/s. The convergence behaviour is displayed in Fig. 4, and the average number of trials required for the solution is listed in Table 1. For each σ_{ML} value, a different value for $\chi^2_{likelihood}$ was used to test to decide that the solution was found, since the value at the global minimum changes[10]. For larger

[9] Note that as the scale factor appears both in the numerator and denominator, an iterative algorithm must be used to find the best scale.

[10] The global minimum should normally be found for $\sigma_{ML} = 0$. The fact that it is actually found for $\sigma_{ML} \approx 0.1$ Å indicates that the model is incomplete, probably because we are only optimizing positional parameters, and that displacement parameters are only set to 'expected' values for that structure, or it could be due to the absence of the hydrogen atoms in the structure. Practically, the σ_{ML} value for which the minimum is found is a indicator on how "complete" the model is.

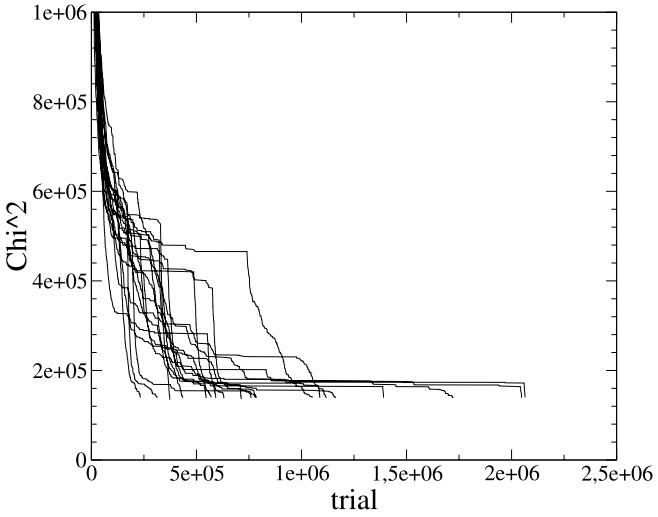

Fig. 4. Convergence of the Global Optimization of the Cimetidine structure, shown on 20 independent runs, when using a Maximum Likelihood error $\sigma = 0.15$ Å. The average convergence is faster (see Fig. 2b.) due to the larger radius of the global minimum (Fig. 3), and also reduces the number of long runs (those requiring more than twice the average length for a successfull run).

σ_{ML} values, the discrimination between the true solution and configurations where a single atom is misplaced becomes difficult: for $\sigma_{\mathrm{ML}} = 0$, $\chi^2_{\mathrm{likelihood}}$ values below 200000 ($\approx 3 \times (\chi^2_{\mathrm{likelihood}})_{\mathrm{min}}$) are in the final minimum, whereas for $\sigma_{\mathrm{ML}} = 0.3$, $\chi^2_{\mathrm{likelihood}}$ must be below 320000 ($\approx 1.07 \times (\chi^2_{\mathrm{likelihood}})_{\mathrm{min}}$) to guarantee that it corresponds to the global minimum.

Tests have also been conducted on a potassium tratrate powder pattern recorded on a laboratory diffractometer, and the conclusions differ, with the average number of trials required increasing from 340000 ($\sigma_{\mathrm{ML}} = 0.0$ Å) to 440000 ($\sigma_{\mathrm{ML}} = 0.1$ and 0.2 Å). Improvement is however obtained when the potassium atom is fixed. This different behaviour can either be due to the presence of the heavy atom (which may require a different error) or to the in-house diffraction data, for which the higher σ^{obs} does not permit the model errors to play the same role as for the cimetidine data set.

Table 1. Average number of trials required to solve the Cimetidine structure from its powder pattern, as a function of the ML error. The numbers were calculated on 20 independent runs, all successful after less than 5 million trials. The Third column indicates the minimum χ^2 value at the global minimum. Note that the large χ^2 values are due to the fact that (i) these are *un-normalized* χ^2 (i.e. not the Crystallographic Goodness of Fit), (ii) it includes the normalization term which increases with the sigma values, and (iii) it includes the Molecule restraints likelihood.

σ_{ML}(Å)	$\langle trials \rangle$ ($\times 10^6$)	$\left(\chi^2_{\mathrm{likelihood}}\right)_{\mathrm{min}}$
0.00	1.6	70000
0.10	1.01	58000
0.15	0.90	94000
0.20	0.85	156000
0.30	0.85	300000

Future uses of ML for direct-space SDPD algorithms

The results obtained on the Cimetidine sample confirm the applicability of ML for real-space SDPD: besides the ability to take into account a non-localized fragment using an infinite σ_{ML} for a few atoms (Markvardsen *et al.* (2003)), it allows to increase the radius of convergence of the algorithm by enlarging the global minimum. This should be especially useful for high-quality diffraction data where the contribution of model errors will be largest compared to the experimental (diffraction measurements) errors. Several other uses can be suggested:

- use a progressive decrease of the ML position error σ_{ML} during the algorithm convergence. It could be done both for SA and PT, although the comparison of $\chi^2_{\mathrm{likelihood}}$ requires the use of the same σ_{ML}, as it alters the HyperSurface.
- the enlargement of the global minimum suggests that introducing ML error could improve minimization algorithms that use partial derivatives: steepest descent, the recently proposed Hybrid Monte-Carlo (Johnston *et al.* (2002)), as well as Least-Squares. It could improve the efficiency of algorithms combining random structure generation and local minimization (Turner *et al.* (2000)).
- the ML position error σ_{ML} could be optimized as any other parameter, since the 'true' minimization of the model versus the diffraction data should incorporate this parameter describing the incompleteness of the model.
- separate σ_{ML} parameters should be used for different atoms, mostly to distinguish heavy and light atoms.

Other FOX highlights

Recent structures

Here are a few complex structures which were recently solved by FOX:

- the $Mg_{1+x}Ir_{1-x}$ ($x = 0$, 0.037 and 0.054) structure was solved (Černý *et al.* (2004)) with 25 independent atoms (75 DOF), largely thanks to the Dynamical Occupancy Correction (Favre-Nicolin and Černý (2002)).
- the structure of β'-PSP (1,3-di-n-hexadecanoyl-2-n-octadecanoylglycerol), $C_{53}H_{102}O_6$, was solved with FOX by increasing gradually the degrees of freedom [bond distances and angles around the glycerol moiety followed by all (56) non-H torsion angles], requiring two months of calculation time (De Ridder *et al.* (2004)).

New features

Brian Toby recently added the ability to create FOX xml files from EXPGUI (Toby (2001), http://www.ncnr.nist.gov/programs/crystallography/software/

expgui/expgui.html), therefore allowing to display crystal structures refined by GSAS (Larson and Von Dreele (2000)) using FOX. Brian Toby and Michael Polyakov also contributed code to display electron (or neutron) density mesh in 3D.

The ability to export the 3D crystal view to POV-Ray was also added (see examples at http://objcryst.sourceforge.net/Fox/screenshot.html), allowing the production of high quality drawings.

Conclusion

In this article we have presented how fundamental a 'flexible' approach is to modern direct-space SDPD algorithms:

- first in the modelisation of groups of atoms, where the coupling of individual atoms, restraints and a set of 'smart' random moves largely improves the convergence, compared to simpler modelling which restrict the exploration of all possible structures.
- secondly by introducing *explicitly* the *approximate* nature of the structural models evaluated during a global optimization using a Maximum Likelihood approach, it is possible to validate more efficiently models while they are still far from the structure solution.

FOX (Free Objects for Xtallography) is a free, open-source project (http://www.gnu.org/philosophy/philosophy.html). It can be downloaded and redistributed from http://objcryst.sourceforge.net under the terms of the GNU General Public License (http://http://www.gnu.org/licenses/gpl.html). It is developed under Linux, precompiled binaries are also available for windows (98 and above), with preliminary support for Mac OS X.

Acknowledgments. The authors would like to thank Anders Markvardsen for several discussions on the use of Maximum Likelihood. We would also like to thank Brian Toby and Michael Polyakov for their contribution to improve the 3D view of crystal structures.

References

Andreev, Y. G.; MacGlashan, G. S.; Bruce, P. G.: Ab initio solution of a complex crystal structure from powder-diffraction data using simulated-annealing method and a high degree of molecular flexibility. Phys. Rev: Cond. Matter **B55** (1997) 12011–12017.

Andreev, Yu. G.; Lightfoot, P.; Bruce, P. G.: A General Monte Carlo Approach to Structure Solution from Powder Diffraction Data: Application to $Poly(ethyleneoxide)_3 : LiN(SO_3CF_3)_2$. J. Appl. Cryst. **30** (1997) 294–305.

Antoniadus, A.; Berruyer, J.; Filhol, A.: Maximum-Likelihood Methods in Powder Diffraction Refinements. Acta Cryst. **A46** (1990) 692–711.

Brenner, L. B.; McCusker L. B.; Baerlocher, C.: Using a structure envelope to facilitate structure solution from powder diffraction data. J. Appl. Cryst. **30** (1997) 1167–1172.

Brenner, L. B.; McCusker L. B.; Baerlocher, C.: The application of structure envelopes in structure determination from powder diffraction data. J. Appl. Cryst. **35** (2002) 243–252.

Bricogne, G.: A multisolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data. Acta Cryst. **A47** (1991) 803–829.

Bruker AXS GmbH: TOPAS User's Manual (2000).

Cernik, R. J.; Cheetham, A. K.; Prout, C. K.; Watkin, D. J.; Wilkinson, A. P.; Willis, B. T. M.: The Structure of Cimetidine ($C_{10}H_{16}N_6S$) Solved from Synchrotron-Radiation X-ray Powder Diffraction Data. J. Appl. Cryst. **24** (1991) 222–226.

Černý, R.; Renaudin, G.; Favre-Nicolin, V.; Hlukhyy, R.; Pöttgen, R.: $Mg_{1+x}Ir_{1-x}$ ($x = 0$, 0.037 and 0.054), a binary intermetallic compound with a new orthorhombic structure type determined from powder and single-crystal X-ray diffraction. Acta Cryst. **B60** (2004) 272–281.

Chapman, M. S.: Restrained Real-Space Macromolecular Atomic Refinement using a New Resolution-Dependent Electron-Density Function. Acta Cryst. **A51** (1995) 69–80.

Chernyshev, V. V.; Schenk, H.: A grid search procedure of positioning a known molecule in an unknown crystal structure with the use of powder diffraction data. Z. Kristallogr. **213** (1998) 1–3.

David, W. I. F.; Shankland, K.; Shankland, N: Routine determination of molecular crystal structures from powder diffraction data. Chem. Commun. **8** (1998) 931–932.

De Ridder, D. J. A.; Goubitz, K.; Pop, M. M.; Driessen, R. A. J.; Peschar, R; Schenk, H: in preparation for Chem. Phys Chem. (2004).

Dinnebier, R. E.: Rigid Bodies in Powder Diffraction. Powder Diffraction **14** (1999) 84–92.

Engel, G. E.; Wilke, S.; Knig, O.; Harris, K. D. M.; Leusen, F. J. J.: PowderSolve – a complete package for crystal structure solution from powder diffraction patterns. J. Appl. Cryst. **32** (1999) 1169–1179.

Falcioni, M.; Deem, M. W.: A biased Monte Carlo scheme for zeolite structure solution. Chem. Phys. **110** (1999) 1754–1766.

Favre-Nicolin, V.; Černý, R.: FOX, 'free objects for crystallography': a modular approach to ab initio structure determination from powder diffraction. J. Appl. Cryst **35** (2002) 734–743.

Harris, K. D. M.; Tremyane, M.; Lightfoot, P.; Bruce, P. G.: Crystal Structure Determination from Powder Diffraction Data by Monte Carlo Methods. J. Am. Chem. Soc. **116** (1994) 3543–3547.

Harris, K. D. M.; Johnston, R. L.; Kariuki, B. M.: The genetic algorithm: foundations and applications in structure solution from powder diffraction data. Acta Cryst. **A54** (1998) 632–645.

Herzberg, O.; Sussman, J. L.: Protein model building by the use of a constrained-restrained least-squares procedure. J. Appl. Cryst. **16** (1983) 144–150.

Johnston, J. C.; David, W. I. F.; Markvardsen, A.; Shankland, K.: A hybrid Monte Carlo method for crystal structure determination from powder diffraction data. Acta Cryst. **A58** (2002) 441–447.

Kariuki, B. M.; Serrano-González, H.; Johnston, R. L.; Harris, K. D. M.: The application of a genetic algorithm for solving crystal structures from powder diffraction data. Chem. Phys. Lett. **280** (1997) 189–195.

Larson, A. C.; Von Dreele, R. B.: General Structure Analysis System (GSAS), Los Alamos National Laboratory Report LAUR 86–748 (2000).

Le Bail, A.: ESPOIR: A Program for Solving Structures by Monte Carlo Analysis of Powder Diffraction Data. Materials Science Forum **378–381** (2001) 65–70.

Luzzati, V.: Traitement statistique des erreurs dans la détermination des structures cristallines. Acta Cryst. **A5** (1952) 802–810.

Markvardsen, A.; David, W. I. F.; Shankland, K.: A maximum-likelihood method for global-optimization-based structure determination from powder diffraction data. Acta Cryst. **A58** (2002) 316–326.

Masciocchi, N.; Bianchi, R.; Cairati, P.; Mezza, G.; Pilati, T.; Sironi, A.: P-RISCON: a real-space scavenger for crystal structure determination from powder diffraction data. J. Appl. Cryst. **27** (1994) 426–429.

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

Murshudov, G. N.; Vagin, A. A.; Dodson, E. J.: Refinement of Macromolecular Structures by the Maximum-Likelihood Method. Acta Cryst. **D53** (1997) 240–255.

Newsam, J. M.; Deem, M. W.; Freeman, C. M.: Accuracy in Powder Diffraction II. NIST Special Publ. No. **846** (1992) 80–91.

Nowell, H.; Attfield, J. P.; Cole, J. C.: The use of restraints in Rietveld refinement of molecular compounds; a case study using the crystal structure determination of tryptamine free base. Acta Cryst. **B58** (2002) 835–840.

Pagola, S.; Stephens, P. W.; Bohle, D. S.; Kosar, A. D.; Madsen, S. K.: The structure of malaria pigment β-haematin. Nature **404** (2000) 307–310.

Pannu, N. S.; Read, R. J.: Improved Structure Refinement Through Maximum Likelihood. Acta Cryst. **A52** (1996) 659–668.

Putz, H.; Schön, J. C.; Jansen, M.: Combined method for ab initio structure solution from powder diffraction data. J. Appl. Cryst. **32** (1999) 864–870.

Read, R. J.: Structure-Factor Probabilities for Related Structures. Acta Cryst. **A46** (1990) 900–912.

Read, R. J.: Pushing the boundaries of molecular replacement with maximum likelihood. Acta Cryst. **D57** (2001) 1373–1382.

Shankland, K.; McBride, L.; David, W. I. F.; Shankland, N.; Steele, G.: Molecular, crystallographic and algorithmic factors in structure determination from powder diffraction data by simulated annealing. J. Appl. Cryst. **35** (2002) 443–454.

Shankland, K.; David, W. I. F.; Csoka, T.: Crystal structure determination from powder diffraction data by the application of a genetic algorithm. Z. Kristallogr. **212** (1997) 550–552.

Sivia, D. S.: Data Analysis: A Bayesian Tutorial. Oxford University Press (1996).

Stewart, J. M.; Karle, J.: The Calculation of ε with Normalized Structure Factors, E. Acta Cryst. **A32** (1976) 1005.

Toby, B. H.: EXPGUI, a graphical user interface for GSAS. J. Appl. Cryst. **34** (2001) 210–213.

Turner, G. W.; Tedesco, E.; Harris, K. D. M.; Johnston, R. L.; Kariuki, B. M.: Implementation of Lamarckian concepts in a Genetic Algorithm for structure solution from powder diffraction data. Chem. Phys. Lett. **321** (2000) 183–190.

Von Dreele, R. B.; Stephens, P. W.; Smith, G. D.; Blessing, R. H.: The first protein crystal structure determined from high-resolution X-ray powder diffraction data: a variant of T3R3 human insulin-zinc complex produced by grinding. Acta Cryst. **D56** (2000) 1549–1553.

Von Dreele, R. B.: Binding of N-acetylglucosamine to chicken egg lysozyme: a powder diffraction study. Acta Cryst. **D57** (2001) 1836–1842.