Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 23 June 2000
Accepted 3 October 2000

The pair-functional method for direct solution of molecular structures. I. Statistical principles

A. D. McLachlan

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England. Correspondence e-mail: admcl@mrc-lmb.cam.ac.uk

The pair-functional principle shows how to construct a unique statistical ensemble of strongly interacting atoms that corresponds to any feasible measured set of X-ray intensities. The ensemble and all its distribution functions are strictly periodic in the crystal lattice, so that each unit cell has exactly the same arrangement of atoms at all times. The mean particle density in the cell is uniform because the ensemble has undefined phases and the origin is not fixed. The atoms in this maximum-entropy ensemble interact through pairwise additive periodic statistical forces within the unit cell. The ensemble average pair-correlation function is matched to the observed originless Patterson function of the crystal. The derived pairing force then becomes approximately proportional to the Ornstein–Zernicke direct correlation function of the ensemble. The atoms have a many-body Boltzmann distribution and the logarithm of the likelihood of any particular conformation is related to its total pairing potential. The pairing potential of a group of atoms acts like a local field in the cell. This property is used in the pair-functional method. Molecular structures can be solved by a direct search in real space for clusters of atoms with high pair potentials. During a successful search, the atoms move from their original random positions to form larger and larger clusters of correctly formed fragments. Finally, every atom belongs to a single cluster, which is the correct solution.

1. Introduction

1.1. A unique ensemble

The pair-functional method described in this paper is a new approach to the direct solution of molecular structures which is based on a uniqueness principle from statistical mechanics. The principle applies to a maximum-entropy statistical ensemble of atomic structures that together reproduce the observed magnitudes of the X-ray intensity data. This is a many-body ensemble of sets of N interacting atoms rather than either a collection of independent atoms or a single definite structure and the interactions take the form of fixed pairwise additive statistical forces that act simultaneously between all the pairs of atoms. The forces are uniquely defined by the X-ray intensities and they come out as anisotropic direction-dependent functions of the atom–atom separation vectors. They provide the essential link between the experimental observations and the statistical distribution of atomic positions in the crystal unit cell. Hence, in principle, once the ensemble has been set up, a structure may be solved by searching within the many-body space for the most probable sets of atomic positions that satisfy the experimental constraints with reasonable accuracy (McLachlan, 1999).

1.2. Background: statistical mechanics

Equilibrium ensembles for both classical and quantum statistical mechanics were fully developed and understood in the 1930's, and used to deduce the properties of many forms of matter from known interatomic forces (Fowler, 1936; Mayer & Mayer, 1940; Hill, 1956; Landau & Lifshitz, 1958). The concept of entropy was then completely explained in terms of the density of states of the equilibrium system in quantum mechanics (Mayer & Mayer, 1940) or the corresponding volume in phase space in classical mechanics. More formally, the entropy could also be derived in terms of the partition function and the Darwin–Fowler method of steepest descents (Fowler, 1936). The statistical theory of fluids, which we draw on in this paper, became logically complete with Mayer's cluster integral theory of the grand ensemble and the succeeding many-body methods based on Mayer diagrams (Mayer & Mayer, 1940; Morita & Hiroike, 1961; De Dominicis, 1962, 1963). These methods led to approximate integral equations that explain the main features of the pair distribution functions in dense fluids, notably the Percus–Yevick equation (Percus & Yevick, 1958). An excellent account of these theories is given in the book by Hansen & McDonald (1986).

The pair-functional ensemble that we use in this paper is a maximum-entropy ensemble that is identical in form with an equilibrium Boltzmann distribution for an atomic fluid with interparticle forces. Robert Harris and I originally derived it in 1961 for use in statistical mechanics and proved its uniqueness. An important conceptual difference between conventional statistical mechanics and the present work is that we are trying to solve the inverse of the usual scientific problem. That is, given the known pair distribution functions of the atoms in a crystal unit cell, we seek a suitable ensemble and a suitable corresponding set of fictitious statistical forces that will reproduce the observed distribution.

In statistical physics, the interaction energies between atoms are real and known quantum-mechanical two-body and three-body forces that are deduced from physical laws. The aim of the theory is to deduce the behaviour of the fluid from known forces.

In contrast, the information theory of inverse inference seeks to represent observed properties in terms of fictitious forces that do not even pretend to be real. Their main purpose is to give a provisional fit to the observations. For example, given a dense fluid, where three-body forces are important but only the pair distribution function is observable, one might derive a unique two-body force that fits the distribution and represents some of the measurable effects of the three-body force.

This point of view is similar in spirit to the inverse Monte Carlo theory used in chemical physics (Allen & Tildesley, 1987; McGreevy & Howe, 1991; Hummer *et al.*, 1998; DaSilva *et al.*, 1998).

The pair-functional method also has connections with Klug's (1958) analysis of the probability distributions of structure factors. The partition function used in the present paper is identical with Klug's moment-generating function for the intensities, but not the amplitudes, of structure factors. Consequently, the partition function with imaginary statistical forces is identified with the characteristic function (Cramer, 1951) and therefore represents the Fourier transform of the intensity probability distribution.

1.3. Background: maximum entropy

Jaynes (1957*a*,*b*, 1978) developed maximum-entropy methods as a general framework for solving inverse problems, initially to provide a description of non-equilibrium phenomena in quantum statistical mechanics, and then later as a self-sufficient general-purpose system of logical inference. Jaynes regarded the entropy itself as a fundamental entity, justified by axiomatic principles from information theory and Bayesian statistics, and he showed how to harness the methods in a wide variety of applications (Levine & Tribus, 1978; Jaynes, 1983).

One special form of maximum-entropy ensemble, extensively used for noise reduction and image processing (Gull & Daniell, 1978; Bryan & Skilling, 1980; Gull & Skilling, 1984), is an independent-atom ensemble in real space in which the mean probability density has prescribed Fourier amplitudes and specified phases (Bricogne, 1984; Navaza, 1985, 1986). The ensemble is analogous to the Boltzmann distribution of a collection of strictly independent atoms in an applied external field which varies throughout space. This type of ensemble has been tried in crystallography as a phase-determination method (Collins, 1982; Wilkins, 1983; Gull *et al.*, 1987; Prince *et al.*, 1988), but now finds its most useful application as a starting point for other Bayesian and tree-based phasing methods (Bricogne, 1988; Bricogne & Gilmore, 1990; Gilmore *et al.*, 1990).

The pair-functional ensemble described in the present paper is completely different from this simple type of independent-atom ensemble, since it consists of sets of strongly interacting atoms in real space, with given Fourier intensities, but no explicitly specified phases. The strong long- and short-range statistical forces derived from a high-resolution data set are sufficiently specific to localize all the atoms in a definite structure.

We often use the term *statistical force* in this paper. In statistical mechanics and information theory, a statistical force is analogous to the thermodynamic forces used in equilibrium and non-equilibrium thermodynamics, and is defined in terms of constrained entropy variations. Thus, if C is the mean value of any statistical variable and S is the entropy of the ensemble, the statistical force λ corresponding to C is defined as $\lambda = -\partial S/\partial C$, with a negative sign. Appendix B gives an elementary example.

1.4. Two classes of direct methods

The majority of existing successful direct methods in crystallography are either Patterson methods or phase-determination methods (Giacovazzo, 1980; Fortier, 1998). In Patterson search methods, there is no statistical element and clusters of atoms are constructed geometrically from the peak-to-peak vector maps (Lipson & Cochran, 1966; Sheldrick, 1985). The search is thus limited to a fairly small number of atoms. Phase-determination methods divide into two main subclasses.

The macromolecular subclass includes important and highly successful phasing methods that make use of extra phase information from physical sources. Examples are heavy atoms, anomalous dispersion, solvent envelopes or the presence of identical subunits within one cell. The macromolecular methods work mainly with density maps and often only assign positions to atoms during the final stages.

The subclass of unaided statistical phasing methods that solve structures at atomic resolution relies on sets of trial phases or on phase probability distributions, derived from the expected properties of a random atomic model that fits the data. The basic principles are embodied in Sayre's equation (Sayre, 1952), the tangent formula (Karle & Hauptman, 1956), and the joint phase probability distributions of triplets and quartets (Hauptman & Karle, 1953; Naya *et al.*, 1965; Hauptman, 1975*a*,*b*). In earlier years, structures were often solved by refining a carefully chosen starting set of phases (Germain & Woolfson, 1968; Main *et al.*, 1980). In the newest

methods, the starting point is usually a random set of atoms. Refinement of phases in reciprocal space then alternates with a refinement and reselection of trial atomic electron-density peaks in the real space of the cell (Sheldrick, 1990; Hauptman, 1991; DeTitta *et al.*, 1994; Weeks *et al.*, 1994).

1.5. Pair-functional theory as a new class of method

The pair-functional theory leads to a new class of solution methods. This class is firmly founded on a thorough statistical analysis of random structures, but it uses atomic positions and atomic probability distributions as its working material and makes no explicit use of phases. Phases have no place in the basic method, except for their indirect role in auxiliary Fourier transform calculations. Although the method works with atoms it need not be restricted to structures with data measured to atomic resolution. For example, a low-resolution molecular envelope could be modelled as a limited number of large globular pseudo-atoms and treated by the method.

The aim of the theory is to develop a useful new series of tools for crystallographers that complement or improve on existing methods. This programme involves three distinct activities:

(i) The theoretical specification of specialized many-body ensembles that embody the constraints imposed by different kinds of experimental data. These are usually constraints on the single-particle and two-particle probability densities.

(ii) To develop methods of inference for deducing the underlying molecular structure from an analysis of each type of ensemble.

(iii) The design of computer algorithms to carry out the solution process.

The pair-functional theory allows a wide and flexible choice of unique ensembles, constructed from various types of linear constraints in the many-dimensional space of $3N$ atomic coordinates for N atoms. The basic uniform ensemble for identical atoms in a cell with $P1$ symmetry is only one starting point. There are also restricted ensembles with higher symmetry for each type of crystal space group [see paper III of this series (McLachlan, 2001)]. Further special-purpose ensembles can also be constructed to include heavy-atom data, known structural fragments, and light atoms of different types. Each of these ensembles uses uniquely determined statistical forces that are matched to a well defined physically feasible set of constraint values.

1.6. Advantages of a unique ensemble

At this point, a reader may ask what is gained by setting up a complicated statistical ensemble of atoms instead of simply looking for one actual molecular structure that fits the data. One answer to this question is that in practice the actual structure itself is often in thermal motion, with large temperature factors for some atoms, and there may be regions of the model which are disordered or fluctuating. Random errors in the atomic coordinates lead to uncertainties in the predicted structure factors (Read, 1990). The pair-functional ensemble can easily be generalized to allow for many of these effects. However, this first paper will only treat crystals of fixed identical atoms.

More fundamentally, however, the ensemble has at least two advantages. The first is that the power of uniqueness theorems is well established in many branches of science – particularly in electromagnetism, statistical mechanics and quantum mechanics. Such theorems not only provide objective tests of quality and completeness but often they also lead to reliable variational methods for solving practical problems. For example, the density-functional method in quantum chemistry (Hohenberg & Kohn, 1964; Parr & Yang, 1989) has led to exceedingly accurate ways of estimating chemical bond energies. Single-particle density-functional theories have also been used successfully in the theory of non-uniform fluids (Mermin, 1964; Evans, 1979). The second advantage concerns the search for trial structures. The paired-atom ensemble is a collection of structures that is heavily biased towards those models that satisfy the experimental conditions. An efficient search of the ensemble, guided by the pairing statistics of the atoms, can arrive at the correct structure much faster than many simpler unguided methods.

In terms of phases, each member of the paired-atom ensemble has a certain set of phase angles. The mean probability distribution of the phases in the ensemble as a whole acquires a definite character, corresponding to the actual molecular solution. However, this result is not achieved by explicit phase bias, but through a process of enrichment, whereby atomic structures with the correct Fourier intensities are strongly concentrated in the ensemble.

1.7. Plan of the paper

The main topic of the present paper is the mathematical foundations of the theory. We shall begin by stating the pair-functional principle itself and showing how it can be derived within the framework of the statistical mechanics of fluids and the related maximum-entropy ensembles. The foundations also yield a simple approximation for the pair potential in the basic ensemble. On the basis of these fundamentals, it becomes possible to discuss the scope of the theory and the precise nature of the unique solutions that the ensembles provide.

In the statistical mechanics of large systems, there is also a well known property: the average features of the distribution are strongly concentrated in the neighbourhood of the most probable part of configuration space. This property leads, in the pair-functional theory, to a working rule, the *most-probable pairing hypothesis*. This proposes that the ensemble contains many members that are close to the actual molecular structure and have the predicted pairing potential, while they also closely match the observed experimental data. The most-probable pairing hypothesis has two useful applications: first it enables a correct solution to be identified without reasonable doubt by checking that it does indeed have a nearly maximal pairing potential. Secondly, it suggests a number of practical computational methods for discovering the solution, by making an efficient search for well paired clusters of atoms.

§2 defines the essential statistical variables for a unit cell in a strictly periodic crystal. §3 introduces the paired-atom ensemble and the fictitious two-body pairing force. §4 discusses the uniqueness of the ensemble. §5 sums up these results in the form of the *pair-functional principle*.

The remaining parts of the paper develop ideas needed to search for solutions. These constitute the *pair-functional method*. §6 introduces the direct correlation function as an approximate pairing force and §7 describes the local pairing field produced at a point in the cell by all the surrounding atoms.

1.8. Mathematical appendices

In the remainder of this paper, the main mathematical derivations have been confined to a series of separate short Appendices. These should be readily understood by readers with a knowledge of the statistical mechanics of fluids or of maximum-entropy theory and they draw on well known results. There are, however, important differences in detail between a classical fluid and a crystal unit cell, so that careful attention must be given to the altered definitions of the statistical quantities that are used.

It is hoped that readers whose main interest is in the application to crystallography rather than the theory will be able to follow the outline of argument in the next sections. Our first task will be to describe the connection between standard crystallographic variables and the corresponding probability variables of a statistical ensemble. This involves a conversion from physical units such as electron densities to dimensionless probability densities.

2. Statistical variables for the cell

The fundamental quantities that describe a crystal cell that contains N atoms in the volume V are the electron density $\rho_e(\mathbf{r})$, the structure factors $F(\mathbf{h})$ and the Fourier intensities $I(\mathbf{h})$:

$$\rho_e(\mathbf{r}) = (1/V)\sum_{j=1}^{N} f_j \delta(\mathbf{r} - \mathbf{x}_j) \tag{1}$$

$$F(\mathbf{h}) = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j) \tag{2}$$

$$I(\mathbf{h}) = |F(\mathbf{h})|^2, \tag{3}$$

where f_j is the scattering factor of an atom at $\mathbf{x}_j$ and δ stands for the Dirac delta function. All coordinates, such as $\mathbf{r}$, are vectors measured in fractional cell coordinates, where the components (x, y, z) range from 0 to 1. The expected mean value of $I(\mathbf{h})$ for a random arrangement of atoms is

$$\Sigma_I = \sum_{j=1}^{N} f_j^2. \tag{4}$$

The Patterson function for a displacement vector $\mathbf{u}$ is

$$P(\mathbf{u}) = (1/V)\sum_{\mathbf{h}} I(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}) \tag{5}$$

measured in units of electrons^2 Å^{-3}. The origin peak of the Patterson is not wanted, hence we use the originless Patterson multiplied by V, which contains information about the distribution of the $N(N-1)$ pairs of atoms.

$$\begin{aligned} P_N^{(2)}(\mathbf{u}) &= \sum_{i\neq j} f_i f_j \delta(\mathbf{u} - \mathbf{x}_i + \mathbf{x}_j) \\ &= \sum_{\mathbf{h}} I_N^{(2)}(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}), \end{aligned} \tag{6}$$

where $I_N^{(2)}(\mathbf{h})$ is the reduced intensity from N atoms

$$I_N^{(2)}(\mathbf{h}) = I(\mathbf{h}) - \Sigma_I. \tag{7}$$

$P_N^{(2)}(\mathbf{u})$ and $I_N^{(2)}(\mathbf{h})$ are both measured in electrons^2 cell^{-1}. Sometimes it is useful to work with a levelled originless Patterson function, whose mean value over the unit cell is zero. This is defined as

$$\Delta P_N^{(2)}(\mathbf{u}) = P_N^{(2)}(\mathbf{u}) - \sum_{i\neq j} f_i f_j. \tag{8}$$

The choice of notation for the rest of this paper involves some clashes between the well established systems of symbols used in X-ray crystallography and in the statistical mechanics of fluids. To change either scheme could cause confusion but it is necessary to bring together results from both disciplines and a few clashes are inevitable. For example, f is used to define an atomic scattering factor or a probability distibution in different contexts. To minimize problems of notation, most of the statistical mathematics and its detailed derivation is placed in self-contained Appendices, which we refer to when needed.

In order to set up statistical ensembles, we have to define suitable atomic probability densities. The model for the full crystal is a strictly periodic repeating lattice with identical atomic positions in each unit cell. This means that the entire conformation of the crystal is specified by the coordinates of just N atoms in one representative central cell. Pairing forces therefore operate between any atom in the central cell and all the repeated atom images in this and other cells. In our description, using cell fractional coordinates, the volume of the cell is taken to be unity.

For simplicity, we now consider only the case of N equal atoms with a constant scattering factor f. Many quantities are labelled here with a subscript N to indicate that they refer to an exact number of atoms, in a canonical ensemble, rather than a fluctuating number of atoms in a grand ensemble. First we consider a set of fixed atoms, without averaging.

For a cell with atoms at fixed positions, the particle number density measured in atoms per cell and its Fourier transform are

$$p_N(\mathbf{r}) = \sum_{j=1}^{N} \delta(\mathbf{r} - \mathbf{x}_j) = V\rho_e(\mathbf{r})/f \tag{9}$$

$$\eta_N(\mathbf{h}) = \sum_{j=1}^{N} \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j) = F(\mathbf{h})/f. \tag{10}$$

The rescaled originless Patterson function becomes a statistical joint two-particle density $K_N^{(2)}(\mathbf{u})$ and the reduced intensity becomes $\zeta_N(\mathbf{h})$:

$$\begin{aligned} K_N^{(2)}(\mathbf{u}) &= \sum_{i\neq j} \delta(\mathbf{u}-\mathbf{x}_i+\mathbf{x}_j) \\ &= P_N^{(2)}(\mathbf{u})/f^2 \\ &= \sum_{\mathbf{h}} \zeta_N(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}), \end{aligned} \tag{11}$$

where

$$\zeta_N(\mathbf{h}) = |\eta_N(\mathbf{h})|^2 - N = I_N^{(2)}(\mathbf{h})/f^2. \tag{12}$$

The mean value of $K_N^{(2)}(\mathbf{u})$ over the whole cell is $N(N-1)$.

In the special case of equal atoms, the statistical structure factors are almost the same as the normalized structure factors and intensities:

$$E_N(\mathbf{h}) = \eta_N(\mathbf{h})/N^{1/2} \tag{13}$$

$$\varepsilon_N(\mathbf{h}) = |E_N(\mathbf{h})|^2 - 1 = \zeta_N(\mathbf{h})/N. \tag{14}$$

In general, however, with atoms of different types, there is no simple relationship between $E_N(\mathbf{h})$ and the $\eta_N(\mathbf{h})$ variables.

We now turn to the corresponding quantities for the cell atomic fluid ensemble, which are averages over distributions of atoms. These are defined in Appendix *D*. The single-particle distribution function $\rho_N^{(1)}(\mathbf{r})$ has a Fourier transform $\hat{\rho}_N^{(1)}(\mathbf{h})$ with

$$\rho_N^{(1)}(\mathbf{r}) = N + \sum_{\mathbf{h}\neq 0} \hat{\rho}_N^{(1)}(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{r}). \tag{15}$$

These single-particle quantities are the ensemble averages of the atomic probability density $p_N(\mathbf{r})$ and its transform $\eta_N(\mathbf{h})$.

$$\rho_N^{(1)}(\mathbf{r}) = \langle p_N(\mathbf{r})\rangle, \quad \hat{\rho}_N^{(1)}(\mathbf{h}) = \langle \eta_N(\mathbf{h})\rangle. \tag{16}$$

The fluid autocorrelation function $k_N^{(2)}(\mathbf{u})$ and its transform $\hat{k}_N^{(2)}(\mathbf{h})$ are related to the two-particle fluid distribution function $\rho_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)$ by the equations

$$k_N^{(2)}(\mathbf{u}) = \int \rho_N^{(2)}(\mathbf{r},\mathbf{r}+\mathbf{u})\,\mathrm{d}\mathbf{r} \tag{17}$$

$$k_N^{(2)}(\mathbf{u}) = N(N-1) + \sum_{\mathbf{h}\neq 0} \hat{k}_N^{(2)}(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}). \tag{18}$$

These quantities are respectively the averages of the scaled originless Patterson function and of the reduced intensity.

$$k_N^{(2)}(\mathbf{u}) = \langle K_N^{(2)}(\mathbf{u})\rangle, \quad \hat{k}_N^{(2)}(\mathbf{h}) = \langle \zeta_N(\mathbf{h})\rangle. \tag{19}$$

Sometimes it is useful to work with a levelled originless Patterson function, whose mean value over the unit-cell volume is zero. The corresponding levelled statistical autocorrelation function is

$$\Delta K_N^{(2)}(\mathbf{u}) = \sum_{i\neq j} \delta(\mathbf{u}-\mathbf{x}_i+\mathbf{x}_j) - N(N-1) \tag{20}$$

and its mean value is denoted by $\Delta k_N^{(2)}(\mathbf{u})$. The Fourier transform of $\Delta k_N^{(2)}(\mathbf{u})$ is the same as $\hat{k}_N^{(2)}(\mathbf{h})$, except that $\Delta\hat{k}_N^{(2)}(\mathbf{h})$ vanishes for $\mathbf{h}=0$. The pair-correlation function of a uniform fluid [not to be confused with the above autocorrelation function $k_N^{(2)}(\mathbf{u})$] is properly defined here as

$$h_N^{(2)}(\mathbf{u}) = [1/N(N-1)]k_N^{(2)}(\mathbf{u}) - 1, \tag{21}$$

but in the physics of real fluids, where N is extremely large, the factor $N(N-1)$ is normally replaced by N^2.

These correspondences can be used to set up the pair-functional maximum-entropy ensemble that matches the observations of certain X-ray intensities on a definite, but unknown, structure of equal atoms in space group $P1$. The ensemble is constructed with a uniform mean density of N but a non-uniform specified mean autocorrelation function $k_N^{(2)}(\mathbf{u})$. Note that the mean density must be chosen to be uniform, for otherwise the proposed density would specify a choice of the unknown X-ray phases and constitute an implicit bias on the unknown structure. Initially, we shall assume for simplicity that every Fourier intensity is measured and known exactly. This ideal situation implies that $k_N^{(2)}(\mathbf{u})$ is known for every particle separation $\mathbf{u}$ and has the specified target value

$$k_N^{(2)}(\mathbf{u}) = N(N-1) + N\sum_{\mathbf{h}\neq 0} \{|E_{\rm obs}(\mathbf{h})|^2 - 1\} \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}). \tag{22}$$

Later we shall relax these ideal conditions to allow for a limited number of measured reflections, which may also be accompanied by estimates of their experimental errors.

3. The paired-atom ensemble

Our objective is to find the constrained statistical distribution that maximizes the entropy of exactly N atoms in a crystal cell and matches a given originless Patterson function, $P_N^{(2)}(\mathbf{u})$. We shall use a many-body ensemble of correlated atoms, in which the mean value of the Patterson function is expressed as a linear constraint.

The standard maximum-entropy theory developed by Jaynes (1978) starts from the concept of a discrete distribution of probabilities p_j over a set of states j with prior statistical weights m_j, as outlined in Appendix *A*. The theory postulates that the most appropriate inferred distribution that satisfies constraints on its mean values is the one that maximizes the entropy

$$S = -\sum_j p_j \log(p_j/m_j) \tag{23}$$

subject to these constraints. Here the prior will be a uniform distribution over the cell volume multiplied by a factor of $1/N!$ for the number of identical particles. The standard theory shows that, under a feasible linear constraint

$$\sum_j c_j p_j = C \tag{24}$$

with weights c_j and a target value C, the normalized probabilities that maximize the entropy are

$$p_j = (1/Z) m_j \exp(\lambda c_j), \quad Z = \sum_j m_j \exp(\lambda c_j), \tag{25}$$

where λ is a Lagrangian multiplier and Z is the probability partition function. Thus, λ is the statistical force associated with the mean value of C and is the negative gradient of the

entropy $\lambda = -\partial S/\partial C$. Appendix *B* summarizes the theory for a general set of many linear constraints. The solution of the equations also needs a general procedure for calculating λ. This can best be performed by minimizing the dual function defined in Appendix *C*.

The pair-functional ensemble is constructed by maximizing the entropy of a continuous probability distribution in many-dimensional space. As described in Appendix *D*, the probability distribution for a cell with exactly N atoms is a function of all $3N$ fractional cell coordinates for the atoms, written

$$f^{(N)}(\mathbf{r}^N) = f^{(N)}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \tag{26}$$

and normalized to unity over the cell, so that

$$\int f^{(N)}(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N = 1. \tag{27}$$

The geometrical part of the entropy, omitting the $\log N!$ correction, is defined in terms of $f^{(N)}$ as

$$S_N = -\int f^{(N)}(\mathbf{r}^N) \log f^{(N)}(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N. \tag{28}$$

The required ensemble for a continuous fluid confined in a cell is the one that maximizes S_N subject to an infinite number of constraints, which are the values of the distance autocorrelation function $k_N^{(2)}(\mathbf{u})$ for every spatial separation $\mathbf{u}$ in the cell. This requires a continuous distribution of Lagrange multipliers $\psi(\mathbf{u})$.

In order to establish the correspondence between the crystal cell theory and the established statistical mechanics of fluids, it is helpful first to consider the ensemble for a general non-uniform fluid, with given non-uniform single-particle and two-particle distributions $\rho_N^{(1)}(\mathbf{x})$ and $\rho_N^{(2)}(\mathbf{x}, \mathbf{y})$. Here there are two kinds of Lagrangian multipliers $\chi(\mathbf{x})$ and $\psi(\mathbf{x}, \mathbf{y})$ that behave like negative effective single-particle and pair-interaction potential energies, and the maximum-entropy ensemble is the Boltzmann distribution in this energy field (see Appendix *E*). This general type of ensemble will be used in later work to derive pair-functional theories that include single-particle constraints derived from information about solvent envelopes, heavy-atom derivatives and known structural fragments in the cell. The grand canonical form of the non-uniform ensemble is also used for many-body treatments of the problem, including the estimation of the Lagrangian multipliers from the theory of fluids, which is treated in a later paper.

Our limited objective of fitting the distance autocorrelation function can be achieved with a spatially uniform ensemble and a simple periodic distance- and direction-dependent pairing force (Appendix *F*). This was the result proved by McLachlan & Harris (1961). Note that the constraints represented by the Patterson function of a crystal are highly anisotropic, unlike the spherically symmetric radial distribution function of a fluid. Therefore, the pairing force is strongly directional. When the entropy is maximized, varying $f^{(N)}$ so as to keep the autocorrelation function fixed, the probability of an arbitrary atomic conformation $\mathbf{r}^N$ is found to be

$$f^{(N)}(\mathbf{r}^N) = (1/Z_N) \exp[\Psi_N(\mathbf{r}^N)], \tag{29}$$

where the many-body pair potential

$$\Psi_N(\mathbf{r}^N) = \sum_{i<j} \psi(\mathbf{r}_i - \mathbf{r}_j) \tag{30}$$

is a sum of two-particle interactions, constructed from the unique pairing force function $\psi(\mathbf{u})$, which is in turn determined implicitly by the autocorrelation function $k_N^{(2)}(\mathbf{u})$ given as data. Z_N is the partition function of the system, while the resulting maximum value of the entropy depends on Z_N and the average value of Ψ_N in the probability distribution $f^{(N)}$.

$$Z_N = \int \exp[\Psi_N(\mathbf{r}^N)]\, \mathrm{d}\mathbf{r}^N \tag{31}$$

$$S_N(\text{max.}) = \log Z_N - \langle \Psi_N \rangle. \tag{32}$$

These general relations take a more useful form when expressed in terms of Fourier variables. Suppose first that all the atoms occupy definite positions in the cell, so that their probability density $p_N(\mathbf{r})$ has definite Fourier coefficients $\eta_N(\mathbf{h})$ and reduced intensities $\zeta_N(\mathbf{h})$. Also let us take the Fourier transform of $\psi(\mathbf{u})$,

$$\psi(\mathbf{u}) = \sum_{\mathbf{h}\neq 0} \hat{\psi}(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}), \tag{33}$$

where $\hat{\psi}(\mathbf{h}) = \hat{\psi}(-\mathbf{h})$ and $\hat{\psi}(\mathbf{h})$ is therefore real. [We note that addition of a uniform constant to $\psi(\mathbf{u})$ makes no difference to the probability distribution $f^{(N)}(\mathbf{r}^N)$ in the canonical ensemble, and so it is always possible to choose $\hat{\psi}(\mathbf{h}) = 0$ for $\mathbf{h} = 0$.] The total pair potential of any atomic conformation $\mathbf{r}^N$ now becomes

$$\begin{aligned}\Psi_N(\mathbf{r}^N) &= \tfrac{1}{2} \sum_{\mathbf{h}\neq 0} \hat{\psi}(-\mathbf{h}) \zeta_N(\mathbf{h}) \\ &= \sum_{\mathbf{h}>0} \psi_E(\mathbf{h}) \{|E_N(\mathbf{h})|^2 - 1\},\end{aligned} \tag{34}$$

where the second sum is over the half-space of $\mathbf{h}$. Here,

$$\psi_E(\mathbf{h}) = N\hat{\psi}(\mathbf{h}) \tag{35}$$

is the normalized Fourier coefficient of the pair potential (see Appendix *G*). The half-space in the reciprocal lattice may be defined for space group $P1$ as the set of (h, k, l) indices with $h \geq 0$, excluding the origin and giving half weight to the indices $(0, k, l)$.

Although the above ensemble for the completely specified Patterson function is a correct direct analogy of the distribution functions used in the statistical mechanics of fluids, it is not yet suitable for use in crystallography. The experimental conditions are different, since the entire Patterson function is never known. Instead, a certain number of X-ray intensities are measured for a limited set of reflections, with reciprocal-lattice vectors denoted by $\mathbf{H}$. The realistic constraints on the ensemble are to maximize S_N, given the sparse conditions

$$\hat{k}_N^{(2)}(\mathbf{H}) = \langle \zeta_N(\mathbf{H}) \rangle \tag{36}$$

for the measured reflections $\mathbf{H}$ only, while the unmeasured components $\hat{k}_N^{(2)}(\mathbf{h})$ are unknown. These conditions lead to a solution in terms of the Fourier components $\hat{\psi}(\mathbf{H})$ of the pair potential for the measured reflections, while $\hat{\psi}(\mathbf{h}) = 0$ for all the unmeasured ones. See Appendix *G*. Under these sparse constraints, the total many-body pair potential of any parti-

cular atomic conformation within the solution ensemble takes the form

$$\Psi_N(\mathbf{r}^N) = \sum_{\mathbf{H}>0} \psi_E(\mathbf{H})\{|E_N(\mathbf{H})|^2 - 1\}. \quad (37)$$

Here the sum only includes the measured reflections and $E_N(\mathbf{H})$ are the normalized structure factors for N equal atoms. It is convenient to subtract the origin peak from the spatial potential $\psi(\mathbf{u})$ to give an originless function $\psi^0(\mathbf{u})$ as described in Appendix G, so that the pair potential of the atoms is reduced to its simplest form

$$\Psi_N(\mathbf{r}^N) = \sum_{\mathbf{H}>0} \psi_E^0(\mathbf{H})|E_N(\mathbf{H})|^2. \quad (38)$$

Now the sum is over the half-space of measured reflections $\mathbf{H}$.

4. Linear constraints and uniqueness

4.1. Mixed distributions

The uniqueness properties of maximum-entropy ensembles are general theorems that only hold under certain specified conditions (Jaynes, 1978). The most important condition is that all the constraints refer to average values of quantities that are linear functions of the underlying probability distributions f.

Suppose that f_A and f_B are probabilities of an event, taken from two distributions that both satisfy the same set of linear constraints. Clearly any linear mixture

$$f = (1 - w)f_A + wf_B \quad (39)$$

also satisfies the constraints. Furthermore, the logarithms satisfy the inequality

$$-f \log f \geq -(1 - w)f_A \log f_A - wf_B \log f_B. \quad (40)$$

Thus, if a many-body distribution is a mixture of two possible solutions to a given set of linear constraints:

$$f^{(N)}(\mathbf{r}^N) = (1 - w)f_A^{(N)}(\mathbf{r}^N) + wf_B^{(N)}(\mathbf{r}^N), \quad (41)$$

it follows immediately that

$$S_N \geq (1 - w)S_{NA} + wS_{NB}. \quad (42)$$

In consequence, either the maximum-entropy solution is unique, with a single distribution $f^{(N)}$, or there are a number of possible solutions, each of which have the same entropy (McLachlan & Harris, 1961). Any mixture of these degenerate solutions is equally valid. Examples of non-unique statistical ensembles occur in mechanics when there is insufficient information to specify the physical situation unambiguously. For example, an isolated system of particles with a given energy may have an unknown centre of mass or unknown orientation. The inequality above may become an equality in practice if the two trial many-body distributions f_A and f_B have negligible overlap with one another, for example if they represent two enantiomorphs of a molecule in a cell.

4.2. Linear many-body constraints

An important class of linear constraints in a many-body ensemble of N particles with a distribution function $f^{(N)}(\mathbf{r}^N)$ is that which specifies the values of any reduced distribution function for a smaller number of particles, of the form $\rho_N^{(n)}(\mathbf{r}^n) = \rho_N^{(n)}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$, or values for the Fourier components of such distributions. The reduced distributions are themselves linear averages over the full distribution $f^N(\mathbf{r}^N)$. Any combination of these linear constraints may therefore be used to construct a unique maximum-entropy ensemble in the N-body space. The pair-functional ensemble is derived from linear conditions on $\rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$ alone and consequently it is unique in the full space $\mathbf{r}^N$.

We note here that it would be possible to generate unique ensembles that also set limits on the fluctuations of two-particle quantities. For example, two-, three- and four-atom potentials could be combined to match any mean square structure factor to an observed value $\langle E^2 \rangle = T^2$ and at the same time assign a chosen value to the fluctuations about the mean defined as $\langle (E^2 - T^2)^2 \rangle = U^2$. These more complicated ensembles are less smooth than the bare pair-functional ensemble and have a lower entropy, so they may not be as useful for computational purposes.

As a counter-example to the uniqueness theorems above, consider the well known independent-atom maximum-entropy ensemble consisting of a distribution $\rho^{(1)}(\mathbf{r}_1)$, normalized to a total of N atoms (Gull & Daniell, 1978; Gull & Skilling, 1984). Here the derived pair-correlation function is not linear but is a quadratic function of products $\rho^{(1)}(\mathbf{r})\rho^{(1)}(\mathbf{r} + \mathbf{u})$ and cannot be used as a constraint to generate a unique maximum-entropy distribution (Lemaréchal & Navaza, 1991).

These two examples demonstrate that the multidimensional N-body ensemble with its linear particle distribution functions and constraints has important advantages when compared with an independent-atom ensemble.

5. The pair-functional principle

We now state the *pair-functional principle* in its simplest basic form, as proved above:

Every feasible observed set of normalized structure-factor intensities from a definite arrangement of N equal atoms in a crystal cell generates a unique maximum-entropy statistical ensemble. Here atoms at positions $\mathbf{r}_i$ and $\mathbf{r}_j$ all interact in pairs through a common long-range distance- and direction-dependent potential $\psi(\mathbf{r}_i - \mathbf{r}_j)$. We call $\psi(\mathbf{u})$ for separation $\mathbf{u}$ the *pairing force function*. This is a functional of all the observed intensities (or equally of the entire originless Patterson function of the crystal), assumed here to be error-free. It is not a mechanical force but an entropy gradient. In reciprocal space, the pairing force is expressed in terms of its normalized Fourier components $\psi_E(\mathbf{H})$ for the measured reflections, while the statistical forces associated with the unmeasured reflections all vanish.

The pair-functional principle can immediately be generalized in several ways, which will be explored later:

(i) The target intensities and model intensities may equally well be generated from any valid statistical distribution of N-atom structures (*e.g.* a disordered crystal or a model with random positional errors rather than one definite set of coordinates).

(ii) Crystals where some information about the single-particle density is given (*e.g.* partial phase information) are described by adding in a further unique single-particle potential $\chi(\mathbf{r})$.

(iii) In any space group of higher symmetry, the unique ensemble has the highest possible symmetry compatible with the space group.

(iv) In a cell with two or more types of atom, having scattering factors $f_a, f_b, \ldots$, the pairing force between atoms is proportional to $f_a f_b \psi(\mathbf{r}_{ai} - \mathbf{r}_{bj})$, with a single pairing force function $\psi(\mathbf{u})$ operating on all atom types.

(v) Supplementary linear constraints could be applied to generate further unique ensembles in which three- or four-particle averages of $\rho_N^{(n)}(\mathbf{r}^n)$ have specified values.

The quantity $\Psi_N(\mathbf{r}^N)$ for any atomic structure with coordinates $\mathbf{r}^N$ will be called the *total pair potential* of the conformation. The significance of Ψ can be expressed in Bayesian terms, where we consider the structure-solution process as a chain of statistical inferences that lead from the observed intensities to the true atomic structure (Bricogne, 1991, 1993). In the pair-functional framework,

$$f^{(N)}(\mathbf{r}^N)\,\mathrm{d}\mathbf{r}^N = (1/Z_N)\exp[\Psi_N(\mathbf{r}^N)]\,\mathrm{d}\mathbf{r}^N \tag{43}$$

is the exact normalized conditional probability that the structure $\mathbf{r}^N$ with ranges $\mathrm{d}\mathbf{r}^N$ is present in the paired ensemble. Thus, $\Psi_N(\mathbf{r}^N)$ is the logarithm of the likelihood of this event. The ensemble itself is not perfect as it contains samples of erroneous structures with intensities that deviate randomly from the observed values. Therefore, Ψ only gives an approximation to the likelihood that the atomic coordinates are correct; the paired atoms must also be correctly placed so as to reproduce the observed intensities.

Because the ensemble that describes the data is unique, the pair-functional principle offers a new way to solve crystallographic problems. A search for correct sets of trial phases is replaced by a phaseless search for correctly placed clusters of atoms from within the specified ensemble, working directly with atoms and physically feasible positive-density maps at all times.

6. An estimate of the pairing force

Before any search begins, it is first necessary to know the pairing force. The maximum-entropy method provides a complete prescription for calculating the unique paired-atom ensemble that represents a given experimental set of X-ray intensities. But we also need a simple practical way to estimate a good approximation for the pairing force $\psi(\mathbf{u})$ or its normalized Fourier components $\psi_E(\mathbf{H})$. This calculation really requires the grand ensemble and will be discussed fully in a later paper. An excellent starting point is to use the Ornstein–Zernicke direct correlation function of the atoms, considered as a fluid (Ornstein & Zernicke, 1914; Hansen & McDonald, 1986).

We assume that the observed intensities $I_{\rm obs}(\mathbf{H})$ of the measured reflections $\mathbf{H}$ have been suitably scaled to give target values $T(\mathbf{H})$ for the normalized structure amplitudes with $\mathbf{H} \neq 0$,

$$|T(\mathbf{H})|^2 = |E_{\rm obs}(\mathbf{H})|^2, \tag{44}$$

so that the autocorrelation function or pair-correlation function to be matched in the grand ensemble will have the Fourier coefficients

$$\hat{k}^{(2)}(\mathbf{H}) = N\{|T(\mathbf{H})|^2 - 1\} \tag{45}$$

$$\hat{h}^{(2)}(\mathbf{H}) = (1/N)\{|T(\mathbf{H})|^2 - 1\}. \tag{46}$$

Suppose that $c^{(2)}(\mathbf{u})$ is the direct correlation function of the paired-atom fluid and its Fourier components are $\hat{c}^{(2)}(\mathbf{H})$. Then it can be shown by perturbation theory that for weak forces

$$\psi(\mathbf{u}) = c^{(2)}(\mathbf{u}) \tag{47}$$

while the exact Ornstein–Zernicke relation states that

$$\hat{c}^{(2)}(\mathbf{H}) = \frac{\hat{h}^{(2)}(\mathbf{H})}{1 + N\hat{h}^{(2)}(\mathbf{H})}. \tag{48}$$

This leads immediately to the result that the normalized pairing force is

$$\psi_E(\mathbf{H}) = \frac{|T(\mathbf{H})|^2 - 1}{|T(\mathbf{H})|^2}. \tag{49}$$

A fuller discussion follows in paper III of this series. Any other Fourier components of the pairing force $\psi_E(\mathbf{k})$ must be set to zero if the amplitude $|T(\mathbf{k})|$ is not measured or if $\mathbf{k} = 0$. Note that in this approximation each Fourier component $\psi_E(\mathbf{H})$ of the potential is independent of the target intensities $|T(\mathbf{k})|^2$ of other reflections $\mathbf{k}$. Thus the different reflections behave nearly independently in the many-body ensemble. The approximation above breaks down for very weak intensities, where it would produce a very large negative force. A large negative statistical force implies that atomic arrangements with a large Fourier intensity for reflection $\mathbf{H}$ have extremely small probabilities. Thus these terms in Ψ act like strongly repulsive mechanical forces in a real fluid. In practice, it is useful to make an empirical correction, using the formula

$$\psi_E(\mathbf{H}) = \frac{|T(\mathbf{H})|^2 - 1}{|T(\mathbf{H})|^2 + T_{\rm low}^2}, \tag{50}$$

where $T_{\rm low}$ is a small cut-off amplitude. This approximate force function has an interesting series expansion, valid for typical reflections with $|T(\mathbf{H})|^2$ close to 1:

$$\psi_E = \gamma(T\,T^* - 1) - \gamma^2(T\,T^* - 1)^2 + \gamma^3(T\,T^* - 1)^3 - \ldots, \tag{51}$$

in which $\gamma = (1 + T_{\rm low}^2)^{-1}$. Here the first term is the transform of the levelled originless Patterson and the remaining correction terms correspond in real space to repeated self-

convolutions of $\Delta P_N^{(2)}(\mathbf{u})$. Thus, the force $\psi(\mathbf{u})$ is a generalization of the Patterson function and shares many of its properties.

7. Atom searches and the local field

Having selected a suitable force, $\psi(\mathbf{u})$, the next stage in a structure solution is to search for the most probable conformations of the interacting atoms in the ensemble. These conformations will be selected from among those that have the highest values for the total pair potential Ψ. The conformations must also be constrained to yield Fourier intensities close to the observed ones. The pairing force is a long-range wavelike potential, extending across the whole cell, and good clusters of interacting atoms need not be localized in any small region.

The physical nature of the ensemble depends on the number of measured X-ray intensities. If there are very few, the ensemble is like an interacting fluid with large fluctuations and no fixed atomic positions. If there are many measurements at atomic resolution, the ensemble represents a condensed solid phase of fixed atoms in a single structure.

The ensemble is analogous to a Boltzmann distribution of interacting atoms in a fluid and so a wide variety of search methods might be used, ranging from simple iterations or gradient optimizations of Ψ to Monte Carlo thermal simulations. The computational cost is dominated by the frequent Fourier transforms needed to calculate Ψ.

Any trial arrangement of atoms generates a local *pairing field* $V(\mathbf{x})$ at every point in the cell, which is the gradient of Ψ for variations of the probability density, in the sense that

$$\mathrm{d}\Psi_N = \int V(\mathbf{x})\,\mathrm{d}p_N(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{52}$$

Therefore, $V(\mathbf{x})$ is a scalar variable and is defined in terms of the originless pairing force $\psi^0(\mathbf{u})$. For an arrangement of point atoms at definite positions, the particle number density takes the form

$$p_N(\mathbf{r}) = \sum_j p_j \delta(\mathbf{r} - \mathbf{r}_j), \tag{53}$$

where p_j is the occupancy of the site at $\mathbf{r}_j$. In particular, for N equally occupied sites, the total pair potential is

$$\Psi_N(\mathbf{r}^N) = \tfrac{1}{2}\sum_{i,j} \psi^0(\mathbf{r}_i - \mathbf{r}_j), \tag{54}$$

and the field is

$$V(\mathbf{x}) = \sum_{j=1}^{N} \psi^0(\mathbf{x} - \mathbf{r}_j). \tag{55}$$

If a new atom is introduced at a vacant point $\mathbf{x}$ then Ψ is increased by $V(\mathbf{x})$. Similarly, if an existing atom is removed from the point $\mathbf{r}_i$ the change is $-V(\mathbf{r}_i)$. Remember that the originless force $\psi^0(\mathbf{u})$ has no spurious self-energy terms between an atom and itself. A trial structure can be improved by moving atoms from improbable filled points of low potential to more probable vacant points of higher potential. The total pair potential of all the atoms is also expressed in terms of $V(\mathbf{x})$ and the positive atomic probability distribution $p_N(\mathbf{x})$ by the equation

$$\Psi_{\rm trial} = \tfrac{1}{2}\int V(\mathbf{x}) p_N(\mathbf{x})\,\mathrm{d}\mathbf{x} = \tfrac{1}{2}\sum_{\mathbf{H}} \hat{V}(-\mathbf{H})\eta_N(\mathbf{H}). \tag{56}$$

Since $V(\mathbf{x})$ is periodic in the crystal it has a Fourier expansion

$$V(\mathbf{x}) = \sum_{\mathbf{h}\neq 0} \hat{V}(\mathbf{h}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{x}), \tag{57}$$

where the components vanish unless $\mathbf{H}$ is a measured reflection:

$$\hat{V}(\mathbf{H}) = N^{-1/2}\psi_E^0(\mathbf{H})E(\mathbf{H}), \tag{58}$$

so that

$$\Psi_{\rm trial} = \sum_{\mathbf{H}>0} \psi_E^0(\mathbf{H})|E(\mathbf{H})|^2. \tag{59}$$

Suppose now that an actual target structure exists that exactly matches the observed data. Then the target value of the normalized structure amplitude is $|T(\mathbf{H})|$ for each measured reflection, with

$$|T(\mathbf{H})|^2 = |E_{\rm obs}(\mathbf{H})|^2. \tag{60}$$

The total pair potential of the target is therefore known in advance and has the value

$$\Psi_{\rm targ} = \sum_{\mathbf{H}>0} \psi_E^0(\mathbf{H})|T(\mathbf{H})|^2. \tag{61}$$

The structure search is governed by the requirement that $\Psi_{\rm trial}$ should be matched to $\Psi_{\rm targ}$ at the same time as each $|E(\mathbf{H})|$ is matched to $|T(\mathbf{H})|$. This outline of a typical search process brings out several points:

(i) *Acceptability*. Any point-atom structure $\mathbf{r}^N$ or indeed any valid positive probability density $p_N(\mathbf{x})$ with the correct amplitudes $|T(\mathbf{H})|$ has $\Psi = \Psi_{\rm targ}$.

(ii) *Non-physical maps*. Any density map constructed from $|T(\mathbf{H})|$ with arbitrary phases also has $\Psi = \Psi_{\rm targ}$. But this is a physically meaningless Ψ value because the maps generally contain false negative densities that contribute spuriously to Ψ. The pair-functional principle applies only to positive probabilities and any proposed negative density must be filtered out.

(iii) *Pair-potential as a search criterion*. The utility of Ψ is twofold, as a figure of merit (relative to $\Psi_{\rm targ}$), and as a target for optimization, which guides atoms towards a solution more or less efficiently.

(iv) *Most-probable pairing hypothesis*. The unique ensemble generated by the experimental data is defined to have a mean total potential of $\langle\Psi\rangle = \Psi_{\rm targ}$. It must contain a scatter of values both higher and lower than the mean. In large statistical ensembles, it is usually found that the most probable value coincides with the mean value. Therefore we postulate that $\Psi_{\rm targ}$ will normally be the most probable value in the ensemble. Consequently, a search of the $3N$-dimensional space of $\mathbf{r}^N$ is likely to yield many structures that match $\Psi_{\rm targ}$. Some among these will also nearly match the observed Fourier intensities and be useful trial solutions.

Tests with a variety of search methods that use the pairing force field show that the progress of the atomic model through successive iterations towards a correct solution often mimics the condensation of a fluid into an ordered solid state. Initially there is a rapid rise in the total pair potential, followed by a long run of small fluctuations during which the system hardly appears to change. Small clusters of correctly placed atoms form and dissolve in turn. Eventually a larger stable nucleating cluster may form that appears to act as a template. Then the whole process of condensation to a complete solution is suddenly completed in a few steps.

This behaviour has the characteristics of a highly cooperative dynamic phase transition. It recalls the similar behaviour of the steps in *SHELX* or *Shake-and-Bake* searches (Sheldrick, 1990; DeTitta *et al.*, 1994; Weeks *et al.*, 1994).

There are some suggestive similarities between the construction of the pair force field $V(\mathbf{r})$ for a trial set of atoms, as described above, and the use of multiple Patterson peaks to solve structures with the help of the symmetry minimum function (Simpson *et al.*, 1965) or other multiple vector search methods (Beevers & Robertson, 1950; Buerger, 1951, 1959; Jacobson & Beckman, 1979). All of these methods use the image-forming properties of the Patterson function to pick out probable atomic positions and then take special precautions to weight down overlapping peaks (Luger & Fuchs, 1986; Pavelčík, 1986; Terwilliger *et al.*, 1987). Such methods are particularly suitable for solving heavy-atom structures (Sheldrick *et al.*, 1993; Pavelčík, 1994).

The pairing force $\psi(\mathbf{u})$ also has strong imaging properties in a good trial structure at high resolution, since the peaks of $V(\mathbf{r})$ pick out the correct atomic positions. There are, however, important differences. The pairing force is a statistical log-likelihood function rather than a geometrical construct, so its values must be added and not multiplied. Also the formula for $\psi_E(\mathbf{H})$ weights the strong and weak reflections very differently from the Patterson, taking large negative values for the weakest reflections.

8. Conclusions

This paper completes the outline of the statistical principles of the pair-functional method and aims to show that it is a logically consistent and well founded theory. We have only touched on the derivation of the pairing force function from the direct correlation function and this will be completed elsewhere. The strong- and weak-coupling limits of the theory are also important future topics, since they show the connections with other theories, such as the temperature-dependent self-consistent field and the independent-atom maximum-entropy method. Another important connection will be established between the well known phase probability distributions of triplets and quartets and a new set of corresponding averages in the paired-atom ensemble.

The actual solution of real structures and a full description of the computational methods employed will be given in a second paper. The choice of search algorithm is partly governed by the costs of Fourier transforms and peak-searching methods carried out on the trial set of atoms. A Gibbs ensemble search at a controlled temperature would be a theoretical ideal, but has been postponed for future use in favour of two simple faster procedures. One is a variant of normal tangent formula refinement in which a large set of trial peaks from a density map is pruned by selecting a smaller subset of well paired atoms. The atoms are chosen to have the best combined pairing potential in the presence of all their selected neighbours. The other procedure is a self-consistent molecular field search, at a specified fictitious temperature, in which every atom is represented by a temperature-dependent probability peak in the map. The peak occupancy for a given atom is an equilibrium Boltzmann distribution generated by the mean long-range pairing field from all the other atomic densities in the cell.

APPENDIX *A*
Entropy of a probability distribution

In probability theory, entropy is used as a measure of the spread, or amount of uncertainty, in a probability distribution (Jaynes, 1978). For example, in a discrete distribution over n possible states, with probabilities $p_1, p_2, \ldots, p_n$, the entropy is defined as

$$S = -\sum_j p_j \log p_j. \tag{62}$$

The smoothest possible distribution, with all p_j equal to $1/n$, has the maximum possible entropy $S(\text{max.}) = \log n$. This definition assumes implicitly that all the states have equal statistical weights. More generally, if the n possibilities are not single states but groups or clusters of states with statistical weights m_j, adding up to a total weight of M, the entropy becomes

$$S = -\sum_j p_j \log(p_j/m_j) \tag{63}$$

and the smoothest distribution has $p_j = m_j/M$. In Bayesian statistics, the quantities m_j often represent assumed prior probabilities.

There are a variety of logical paths and statistical arguments that lead to these initial definitions, and the foundations have been discussed extensively by Shannon (Shannon, 1948*a*,*b*; Shannon & Weaver, 1949) and Jaynes (1983). One of the simplest approaches is to consider an experiment of repeated random trials. Suppose that a single trial has n possible outcomes, all equally probable, with known probabilities of $1/n$ for each one. We attempt to verify these probabilities by carrying out a large number, ν, of repeated trials in succession. Consider the result of one particular series experiment, in which the various outcomes are seen to occur respectively $\nu_1, \nu_2, \ldots, \nu_n$ times. The experimenter would deduce that the measured probabilities for the outcomes are

$$p_j = \nu_j/\nu. \tag{64}$$

These will not in general be exactly the same as the actual underlying probabilities $1/n$. According to the binomial

probability distribution, it can be predicted that the number of different ways of obtaining the observed counts $\nu_1, \nu_2, \ldots, \nu_n$ in any one series of ν trials is exactly

$$W = \frac{\nu!}{\nu_1!\nu_2!\ldots\nu_n!}. \tag{65}$$

In the limit of large ν, keeping the ratios p_j fixed, W can be estimated accurately by Stirling's formula ($\log \nu = \nu \log \nu - \nu$) with the result that

$$\log W = \nu S \quad \text{and} \quad S = -\sum_j p_j \log p_j. \tag{66}$$

Thus the entropy of the inferred distribution p_j is a measure of how likely it is to be observed when the actual probabilities are all equal.

In the more general case, where a single trial has unequal probabilities, with weights m_j, a similar argument gives the entropy derived from $\log W$ as

$$S = -\sum_j p_j \log(p_j/m_j) - \log M. \tag{67}$$

The constant term $\log M$ is usually omitted.

APPENDIX *B*
The most probable distribution with constraints

One often needs to deduce the form of the most probable, or smoothest, non-uniform probability distribution that is compatible with given average values for a number of quantities of interest. Let there be l quantities $c_1, c_2, \ldots, c_l$ in a system with n discrete states, and suppose that the quantity c_α takes the value $c_{\alpha j}$ in state j. We seek the maximum-entropy distribution of p_j that yields given average values C_α for each of the quantities

$$\sum_j c_{\alpha j} p_j = C_\alpha, \quad \text{where} \quad A = \sum_j p_j = 1. \tag{68}$$

The solution is well known from statistical mechanics and uses Lagrangian multipliers. The Lagrangian equation for variations of the probabilities is

$$\delta S + \lambda_0 \delta A + \sum_\alpha \delta C_\alpha = 0 \tag{69}$$

and leads to the condition

$$\sum_j \left\{ -[1 + \log(p_j/m_j)] + \lambda_0 + \sum_\alpha \lambda_\alpha c_{\alpha j} \right\} \delta p_j = 0. \tag{70}$$

The multiplier λ_0 is eliminated and the normalized probabilities are

$$p_j = \frac{1}{Z(\lambda_1, \lambda_2, \ldots, \lambda_l)} m_j \exp\left(\sum_\alpha \lambda_\alpha c_{\alpha j} \right), \tag{71}$$

where Z is the partition function,

$$Z = \sum_j m_j \exp\left(\sum_\alpha \lambda_\alpha c_{\alpha j} \right). \tag{72}$$

The value of the entropy for this distribution is an implicit function of the average values

$$S(C_1, C_2, \ldots, C_l) = S(\text{max.}) = \log Z - \sum_\alpha \lambda_\alpha C_\alpha. \tag{73}$$

The multipliers λ_α and the mean values C_α are obtained as partial derivatives of S and $\log Z$, respectively,

$$\frac{\partial S}{\partial C_\alpha} = -\lambda_\alpha, \quad \frac{\partial \log Z}{\partial \lambda_\alpha} = C_\alpha. \tag{74}$$

Because λ_α are the negative gradients of the entropy, they are often described as the *statistical forces* associated with the constraints, by analogy with mechanical forces which are derivatives of the energy and the thermodynamic entropic forces which act in irreversible processes. This maximum-entropy distribution is generally a unique true maximum, with negative definite second derivatives, provided that the chosen values of C_α permit a feasible solution.

APPENDIX *C*
The dual function

When Z is considered as a function of the variables λ_α, each choice of the multipliers generates a possible maximum-entropy probability distribution p_j with certain varying associated averages $C_\alpha(\lambda_1, \lambda_2, \ldots, \lambda_l)$. It then becomes necessary to determine the correct values of λ_α which reproduce a desired set of given fixed numerical target values $C_\alpha = C_\alpha^T$. A convenient way to do this (Agmon *et al.*, 1978) is through the dual target function $Q(\lambda_1, \lambda_2, \ldots, \lambda_l)$, defined as

$$Q = \log Z(\lambda) - \sum_\alpha C_\alpha^T \lambda_\alpha. \tag{75}$$

Q has a unique minimum which is defined by the conditions

$$\frac{\partial Q}{\partial \lambda_\alpha} = C_\alpha(\lambda) - C_\alpha^T = 0 \tag{76}$$

and therefore ensures that $C_\alpha(\lambda)$ has the correct value. For an account of dual functions, see Gill *et al.* (1981) and Luenberger (1984). An elementary example of a minimal function with a dual is seen in thermodynamics, where the Helmholtz free energy A is the dual of the negative Gibbs free energy $-G$, where the relationship exchanges the roles of P and V, the pressure and volume.

Often the requirement is to fit the distribution to a set of averages that are only known with limited accuracy. For example, when the target values C_α^T have standard deviations σ_α, our objective may be to maximize the quantity

$$U = S - \tfrac{1}{2} \sum_\alpha (C_\alpha - C_\alpha^T)^2/\sigma_\alpha^2 \tag{77}$$

by solving the equations

$$\frac{\partial U}{\partial C_\alpha} = -\lambda_\alpha - (C_\alpha - C_\alpha^T)/\sigma_\alpha^2 = 0. \tag{78}$$

This can be converted into an equivalent dual problem (McLachlan, 1989) to minimize the function $Y(\lambda_1, \lambda_2, \ldots, \lambda_l)$,

$$Y = Q + \tfrac{1}{2} \sum_\alpha \sigma_\alpha^2 \lambda_\alpha^2, \tag{79}$$

in which Q is supplemented by penalty terms that limit the size of the multipliers λ_α. The condition for a minimum of Y is then the same as for the maximum of U:

$$\frac{\partial Y}{\partial \lambda_\alpha} = (C_\alpha - C_\alpha^T) + \sigma_\alpha^2 \lambda_\alpha = 0. \tag{80}$$

APPENDIX *D*
Cell distribution functions for *N* atoms

The classical theory of statistical mechanics of fluids deals with a very large number N of interacting atoms in motion, enclosed in a large box of volume V. Usually conditions are chosen so that the number density N/V of the fluid tends to a limiting value ρ.

The ensembles that we use for crystallographic purposes are different. Every ensemble now represents a finite number N of atoms at rest in a single unit cell of the crystal, whose volume V, measured in standard cell fraction coordinates $\mathbf{r} = (x, y, z)$, is equal to unity. The crystal is assumed to be strictly periodic and every other unit cell is an exact copy of the first, with all N atoms in the same positions.

When a cell contains exactly N atoms, these are described by an N-particle probability distribution function $f^{(N)}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ of the $3N$ cell coordinates, which is written in abbreviated form as $f^{(N)}(\mathbf{r}^N)$. The distribution is normalized to 1 over the cell volume, with

$$A_N = \int f^{(N)}(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N = 1. \tag{81}$$

All positions in space are equally probable so that the statistical weight of an element of $\mathbf{r}^N$ is equal to its volume $\mathrm{d}\mathbf{r}^N$. The reduced single-particle and two-particle distribution functions of the ensemble are then defined as the averages

$$\rho_N^{(1)}(\mathbf{x}) = \sum_i \int \delta(\mathbf{r}_i - \mathbf{x}) f^{(N)}(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N \tag{82}$$

$$\rho_N^{(2)}(\mathbf{x}, \mathbf{y}) = \sum_{i,j} \iint \delta(\mathbf{r}_i - \mathbf{x}) \delta(\mathbf{r}_j - \mathbf{y}) f^{(N)}(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N \tag{83}$$

and satisfy the normalization conditions

$$\int \rho_N^{(1)}(\mathbf{x})\, \mathrm{d}\mathbf{x} = N \tag{84}$$

$$\iint \rho_N^{(2)}(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = N(N-1). \tag{85}$$

In a uniform ensemble, such as the ensemble for a cell in space group $P1$ without any defined origin, $\rho_N^{(1)}(\mathbf{x})$ has the constant value N. Thus, $\rho_N^{(1)}$ is measured in units of atoms per cell.

The statistical theory of infinite fluids normally uses a scaled distribution function $g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$, which tends to a limit of 1 at well separated points. This is defined as

$$g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \rho_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) / \rho_N^{(1)}(\mathbf{r}_1) \rho_N^{(1)}(\mathbf{r}_2). \tag{86}$$

In a crystal ensemble, this long-range limit no longer holds but in a uniform cell $g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2)$ is a periodic function of the separation between particles. Thus, when $\mathbf{r}_1 - \mathbf{r}_2 = \mathbf{u}$,

$$g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = g_N^{(2)}(\mathbf{u}). \tag{87}$$

Two other useful two-particle functions are the distance autocorrelation function $k_N^{(2)}(\mathbf{u})$ and the pair-correlation function $h_N^{(2)}(\mathbf{u})$, which are

$$k_N^{(2)}(\mathbf{u}) = \int \rho_N^{(2)}(\mathbf{r}, \mathbf{r} + \mathbf{u})\, \mathrm{d}\mathbf{r} \tag{88}$$

$$h_N^{(2)}(\mathbf{u}) = g_N^{(2)}(\mathbf{u}) - 1. \tag{89}$$

In a spatially uniform ensemble, $h_N^{(2)}(\mathbf{u})$ and $k_N^{(2)}(\mathbf{u})$ are closely related, with

$$h_N^{(2)}(\mathbf{u}) = (1/N^2) k_N^{(2)}(\mathbf{u}) - 1. \tag{90}$$

The definition of the entropy of N identical particles in classical statistical mechanics is taken from the classical limit of quantum theory and takes the form

$$S = S_N - \log N!, \tag{91}$$

where S_N is the classical geometrical entropy of the distribution function $f^{(N)}$ and $-\log N!$ is a correction for the indistinguishability of identical particles. Here,

$$S_N = -\int f^{(N)}(\mathbf{r}^N) \log f^{(N)}(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N. \tag{92}$$

The most probable distribution is the uniform one, $f^{(N)}(\mathbf{r}^N) = 1$.

The distribution function of the atoms in a unit cell, which we use here, $f_{\mathrm{cell}}^{(N)}(\mathbf{r}^N)$ is normalized to a cell of unit volume in cell coordinates, unlike the conventional function $f_{\mathrm{gas}}^{(N)}(\mathbf{X}^N)$ for a classical fluid, normalized for Cartesian coordinates in a volume V. This accounts for a difference in the two forms of entropy for the same physical situation

$$S_{\mathrm{gas}} = S_{\mathrm{cell}} + N \log V. \tag{93}$$

APPENDIX *E*
Cell with *N* atoms and given particle distributions

Suppose that a crystal cell is known to contain exactly N atoms with given physically feasible single-particle and two-particle distributions $\rho_N^{(1)}(\mathbf{x})$ and $\rho_N^{(2)}(\mathbf{x}, \mathbf{y})$. We seek the unique maximum-entropy distribution $f^{(N)}(\mathbf{r}^N)$, correctly normalized to $A_N = 1$, which achieves these average values. This distribution can be derived with Lagrange's multipliers, using a multiplier λ for A_N and a continuous distribution of multipliers $\chi(\mathbf{x})$ and $\psi(\mathbf{x}, \mathbf{y})$ for $\rho_N^{(1)}$ and $\rho_N^{(2)}$, respectively. The conditions satisfied by variations $\delta f^{(N)}(\mathbf{r}^N)$ are

$$\delta S_N + \lambda \delta A_N + \int \chi(\mathbf{x}) \delta \rho_N^{(1)}(\mathbf{x})\, \mathrm{d}\mathbf{x} + \tfrac{1}{2} \iint \psi(\mathbf{x}, \mathbf{y}) \delta \rho_N^{(2)}(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = 0. \tag{94}$$

The standard partition-function method of Appendix B gives a form of Boltzmann distribution, having an effective temperature factor $\beta = 1/kT = 1$ and a positive sign to the exponential.

$$f^{(N)}(\mathbf{r}^N) = (1/Z_N) \exp[\beta \Lambda_N(\mathbf{r}^N)]. \tag{95}$$

The partition function is

$$Z_N = \int \exp[\beta \Lambda_N(\mathbf{r}^N)]\, \mathrm{d}\mathbf{r}^N \tag{96}$$

in which $\Lambda_N(\mathbf{r}^N)$ is a sum of effective single-particle and two-particle energies

$$\begin{aligned}\Lambda_N(\mathbf{r}_N) &= \mathbf{X}_N(\mathbf{r}^N) + \Psi_N(\mathbf{r}^N) \\ &= \sum_i \chi(\mathbf{r}_i) + \sum_{i<j} \psi(\mathbf{r}_i, \mathbf{r}_j). \end{aligned} \tag{97}$$

The value of the entropy is

$$\begin{aligned}S_N(\text{max.}) &= \log Z_N - \int \Lambda_N(\mathbf{r}^N) f^N(\mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N \\ &= \log Z_N - \beta \frac{\partial}{\partial \beta} \log Z_N. \end{aligned} \tag{98}$$

The effective potentials are the functional derivatives of the entropy with respect to the constraints

$$\delta S_N / \delta \rho_N^{(1)}(\mathbf{x}) = -\chi(\mathbf{x}) \tag{99}$$

$$\delta S_N / \delta \rho_N^{(2)}(\mathbf{x}, \mathbf{y}) = -\tfrac{1}{2} \psi(\mathbf{x}, \mathbf{y}). \tag{100}$$

Notice that this distribution is identical in form to the statistical-mechanical Boltzmann distribution of a classical fluid of atoms in a known non-uniform external field $\chi(\mathbf{x})$ with a given pair-interaction potential $\psi(\mathbf{x}, \mathbf{y})$, which is a function of both positions $\mathbf{x}$ and $\mathbf{y}$. An important difference is that the maximum-entropy ensemble represents the inverse of the usual situation in physics. Here the distributions $\rho^{(1)}$ and $\rho^{(2)}$ are given, but the statistical potentials are unknown and must be deduced from the constraints. Under these general conditions, it is not usually possible to choose $\psi(\mathbf{x}, \mathbf{y})$ to be a simple function of the particle separation $\mathbf{u} = \mathbf{y} - \mathbf{x}$.

The ensemble for a crystal cell without additional symmetry, in space group $P1$, with no defined origin, necessarily has a uniform single-particle distribution and only the pair distribution is known. Under these conditions, the pseudo-potential $\chi(\mathbf{x})$ vanishes and the ensemble reduces to

$$f^{(N)}(\mathbf{r}^N) = (1/Z_N) \exp[\beta \Psi_N(\mathbf{r}^N)]. \tag{101}$$

In all these ensembles, the unique appropriate statistical potentials $\chi(\mathbf{x})$ and $\psi(\mathbf{x}, \mathbf{y})$ will exist for any well posed feasible set of constraints. The problem of finding these functions in practice is more difficult. In principle, however, the solution is easily obtained by minimizing the corresponding dual function Q_N of Appendix *C* for specified target particle distributions $\rho_N^{(1)T}(\mathbf{x})$ and $\rho_N^{(2)T}(\mathbf{x}, \mathbf{y})$. The dual function Q_N is derived from $\log Z_N$ and is therefore a functional of the proposed fields $\chi(\mathbf{x})$ and $\psi(\mathbf{x}, \mathbf{y})$ with the form

$$\begin{aligned}Q_N = \log Z_N(\chi, \psi) &- \int \rho_N^{(1)T}(\mathbf{x}) \chi(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &- \tfrac{1}{2} \int\int \rho_N^{(2)T}(\mathbf{x}, \mathbf{y}) \psi(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}. \end{aligned} \tag{102}$$

To use this approach successfully, one needs a suitable many-body approximation for Z_N as a function of the variables (χ, ψ).

APPENDIX *F*
Cell with *N* atoms and given autocorrelation function

The statistical quantity in the theory of fluids that corresponds to the crystallographic originless Patterson function is the particle pair probability autocorrelation function for a vector separation $\mathbf{u}$,

$$k_N^{(2)}(\mathbf{u}) = \int \rho_N^{(2)}(\mathbf{r}, \mathbf{r} + \mathbf{u})\, \mathrm{d}\mathbf{r}, \tag{103}$$

which has a mean value of $N(N-1)$ over all separations $\mathbf{u}$. For a cell with non-overlapping hard-sphere atoms, there is no origin peak and $k_N^{(2)}(0) = 0$.

The maximum-entropy ensemble for a spatially uniform cell ensemble in which the form of $k_N^{(2)}(\mathbf{u})$ is specified can be deduced immediately from the more general conditions in Appendix *E*. Thus, the Lagrangian multipliers give

$$\delta S_N + \lambda \delta A_N + \tfrac{1}{2} \int \psi(\mathbf{u}) \delta k_N^{(2)}(\mathbf{u})\, \mathrm{d}\mathbf{u} = 0 \tag{104}$$

in which $\chi(\mathbf{x})$ is unused and $\psi(\mathbf{u})$ is now a centrosymmetric two-particle potential. It is a periodic function of the interparticle separation vector $\mathbf{u} = \mathbf{x} - \mathbf{y}$. The form of the probability distribution is again

$$f^{(N)}(\mathbf{r}^N) = (1/Z_N) \exp[\beta \Psi_N(\mathbf{r}^N)] \tag{105}$$

but with the difference that $\Psi_N(\mathbf{r}^N)$ depends on the separation-dependent potential $\psi(\mathbf{u})$ instead of the more general two-point potential $\psi(\mathbf{r}_1, \mathbf{r}_2)$.

$$\Psi^{(N)}(\mathbf{r}^N) = \sum_{i<j} \psi(\mathbf{r}_i - \mathbf{r}_j). \tag{106}$$

The entropy for this solution is

$$S_N(\text{max.}) = \log Z_N - \langle \Psi_N \rangle \tag{107}$$

and the functional derivative of the entropy with respect to the autocorrelation function becomes

$$\delta S_N / \delta k_N^{(2)}(\mathbf{u}) = -\tfrac{1}{2} \psi(\mathbf{u}). \tag{108}$$

APPENDIX *G*
Cell with *N* atoms and given Fourier components of the distribution functions

Next we suppose that the ensemble in the cell is not necessarily uniform but has known average values for certain Fourier components $\mathbf{H} \neq 0$ of the real single-particle distribution function and the real two-particle correlation function. (A slightly more general situation would allow for two independent selections of the reflections $\mathbf{H}'$ and $\mathbf{H}''$, which have given one- and two-particle components.) Thus,

$$\langle \eta_N(\mathbf{H}) \rangle = \hat{\rho}_N^{(1)}(\mathbf{H}) \quad \text{and} \quad \langle \zeta_N(\mathbf{H}) \rangle = \hat{k}_N^{(2)}(\mathbf{H}) \tag{109}$$

for each of the selected reflections. The Lagrangian multipliers for this system are now taken to be $\hat{\chi}(\mathbf{H})$ and $\hat{\psi}(\mathbf{H})$, with the maximum-entropy condition

$$\delta S_N + \lambda \delta A_N + \sum_{\mathbf{H}} \hat{\chi}(-\mathbf{H}) \delta \hat{\rho}_N^{(1)}(\mathbf{H}) + \tfrac{1}{2} \sum_{\mathbf{H}} \hat{\psi}(-\mathbf{H}) \delta \hat{k}_N^{(2)}(\mathbf{H}) = 0. \tag{110}$$

The sums are taken over both positive and negative values of $\mathbf{H}$. Notice that, by definition, $\hat{\chi}(\mathbf{h}) = \hat{\psi}(\mathbf{h}) = 0$ for the unmeasured reflections, $\mathbf{h}$, so that the values of $\hat{\rho}_N^{(1)}(\mathbf{h})$ and

$\hat{k}_N^{(2)}(\mathbf{h})$ are unconstrained. The N-particle probability distribution is again

$$f^{(N)}(\mathbf{r}^N) = (1/Z_N)\exp[\beta\Lambda_N(\mathbf{r}^N)], \tag{111}$$

where $\Lambda_N(\mathbf{r}^N)$ contains both single-particle and two-particle Fourier components.

$$\begin{aligned}\Lambda_N(\mathbf{r}_N) &= \mathbf{X}_N(\mathbf{r}^N) + \Psi_N(\mathbf{r}^N) \\ &= \sum_{\mathbf{H}} \hat{\chi}(-\mathbf{H})\eta_N(\mathbf{H}) + \tfrac{1}{2}\sum_{\mathbf{H}} \hat{\psi}(-\mathbf{H})\zeta_N(\mathbf{H}).\end{aligned} \tag{112}$$

Although this solution appears superficially to be the same as the one derived in Appendix E, it differs because the Fourier components of the statistical forces for the unmeasured reflections all vanish. The entropy for this solution is

$$S_N(\text{max.}) = Z_N - \sum_{\mathbf{H}} \hat{\chi}(-\mathbf{H})\hat{\rho}_N^{(1)}(\mathbf{H}) - \tfrac{1}{2}\sum_{\mathbf{H}} \hat{\psi}(-\mathbf{H})\hat{k}_N^{(2)}(\mathbf{H}), \tag{113}$$

and the solution is unique, provided that these linear constraints are physically feasible. The gradients of the entropy and partition function are respectively given by

$$\mathrm{d}S_N = -\sum_{\mathbf{H}} \hat{\chi}(-\mathbf{H})\,\mathrm{d}\hat{\rho}_N^{(1)}(\mathbf{H}) - \tfrac{1}{2}\sum_{\mathbf{H}} \hat{\psi}(-\mathbf{H})\,\mathrm{d}\hat{k}_N^{(2)}(\mathbf{H}) \tag{114}$$

$$\mathrm{d}Z_N = \sum_{\mathbf{H}} \hat{\rho}_N^{(1)}(-\mathbf{H})\,\mathrm{d}\hat{\chi}(\mathbf{H}) + \tfrac{1}{2}\sum_{\mathbf{H}} \hat{k}_N^{(2)}(-\mathbf{H})\,\mathrm{d}\hat{\psi}(\mathbf{H}). \tag{115}$$

It is useful to rescale $\hat{\psi}(\mathbf{H})$ into a normalized form that is independent of the number of atoms by making the transformation

$$\psi_E(\mathbf{H}) = N\hat{\psi}(\mathbf{H}). \tag{116}$$

With this convention, the total pair potential for a set of equal atoms becomes

$$\Psi_N(\mathbf{r}^N) = \sum_{\mathbf{H}>0} \psi_E(\mathbf{H})\{|E_N(\mathbf{H})|^2 - 1\}, \tag{117}$$

summed over a hemisphere of reciprocal space. A further reduction is often convenient. This is to remove the origin peak from the spatial pairing force $\psi(\mathbf{u})$ and use a set of originless Fourier potentials $\psi_E^0(\mathbf{H})$. Suppose that there are n_R measured reflections and that ψ_{Em} is the mean of their pair potentials. Then we define

$$\begin{aligned}\psi_E^0(\mathbf{H}) &= \psi_E(\mathbf{H}) - \psi_{Em} \\ \psi_{Em} &= (1/n_R)\sum_{\mathbf{H}\neq 0} \psi_E(\mathbf{H})\end{aligned} \tag{118}$$

with $\psi_E^0(\mathbf{h}) = 0$ for all unmeasured reflections. The originless spatial potential becomes

$$\psi^0(\mathbf{u}) = (1/N)\sum_{\mathbf{H}\neq 0} \psi_E^0(\mathbf{H})\exp(-2\pi i\mathbf{H}\cdot\mathbf{u}). \tag{119}$$

This transformation reduces the pair potential for equal atoms to its simplest form

$$\Psi_N(\mathbf{r}^N) = \sum_{\mathbf{H}>0} \psi_E^0(\mathbf{H})|E_N(\mathbf{H})|^2. \tag{120}$$

The sum is now over the half-space of measured reflections $\mathbf{H}$.

The correct values of the statistical forces $\hat{\chi}(\mathbf{H})$ and $\hat{\psi}(\mathbf{H})$ may in principle be determined by setting up the dual minimal function Q_N, as described in Appendix E, with the desired target values $\hat{\rho}_N^{(1)T}(\mathbf{H})$ and $\hat{k}_N^{(2)T}(\mathbf{H})$. Alternatively, if the data are only to be fitted to a certain accuracy, the error dual function Y_N can be used. The error function, unlike Q_N, will always yield a solution with finite values for the statistical forces.

It is a pleasure to thank Ian McDonald, Lawrence Pratt and Robert Harris for advice on the statistical mechanics of fluids; John Skilling and Steve Gull for discussions on Bayesian statistics; David Eisenberg, Tony Crowther and Phil Evans for guidance on crystallographic matters. I am also indebted to Gerard Bricogne for his early lectures on maximum entropy and direct methods. I thank two referees for useful suggestions.

References

Agmon, N., Alhassid, Y. & Levine, R. D. (1978). *The Maximum Entropy Formalism*, edited by R. D. Levine & M. Tribus, pp. 207–208. Cambridge, MA: MIT Press.
Allen, M. P. & Tildesley, D. J. (1987). *Computer Simulation of Liquids.* Oxford University Press.
Beevers, C. A. & Robertson, J. H. (1950). *Acta Cryst.* **3**, 164.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.
Bricogne, G. (1991). *Acta Cryst.* A**47**, 803–829.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. & Gilmore, C. J. (1990). *Acta Cryst.* A**46**, 284–297.
Bryan, R. K. & Skilling, J. (1980). *Mon. Not. R. Astronom. Soc.* **191**, 69–79.
Buerger, M. J. (1951). *Acta Cryst.* **4**, 531–544.
Buerger, M. J. (1959). *Vector Space and its Application in Crystal Structure Investigation.* New York: John Wiley.
Collins, D. M. (1982). *Nature (London)*, **298**, 49–51.
Cramer, H. (1951). *Mathematical Methods of Statistics.* Princeton University Press.
DaSilva, F. L. B., Svensson, B., Akesson, T. & Jonsson, B. (1998). *J. Chem. Phys.* **109**, 2624–2629.
De Dominicis, C. (1962). *J. Math. Phys.* **3**, 983–1002.
De Dominicis, C. (1963). *J. Math. Phys.* **4**, 255–265.
De Titta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.
Evans, R. (1979). *Adv. Phys.* **28**, 143–200.
Fortier, S. (1998). Editor. *Direct Methods for Solving Macromolecular Structures. NATO Adv. Sci. Inst. Ser. C. Mathematical and Physical Sciences*, Vol. 507. Dordrecht: Kluwer.
Fowler, R. H. (1936). *Statistical Mechanics*, 3rd ed. Cambridge University Press.
Germain, G. & Woolfson, M. M. (1968). *Acta Cryst.* B**24**, 91–96.
Giacovazzo, C. (1980). *Direct Methods in Crystallography*. London: Academic Press.
Gill, P. E., Murray, W. & Wright, M. H. (1981). *Practical Optimization.* London: Academic Press.
Gilmore, C. J., Bricogne, G. & Bannister, C. (1990). *Acta Cryst.* A**46**, 297–308.
Gull, S. F. & Daniell, G. J. (1978). *Nature (London)*, **272**, 686–690.
Gull, S. F., Livesey, A. K. & Sivia, D. S. (1987). *Acta Cryst.* A**43**, 112–117.
Gull, S. F. & Skilling, J. (1984). *IEEE Proc.* **131**, 646–659.
Hansen, J. P. & McDonald, I. R. (1986). *Theory of Simple Liquids*, 2nd ed. London: Academic Press.
Hauptman, H. (1975*a*). *Acta Cryst.* A**31**, 671–679.
Hauptman, H. (1975*b*). *Acta Cryst.* A**31**, 680–687.

Hauptman, H. (1991). *Crystallographic Computing 5. From Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. Oxford University Press/IUCr.
Hauptman, H. & Karle, J. (1953). *The Solution of the Phase Problem: I. The Centrosymmetric Crystal. American Crystallographic Association Monograph* No 3. Pittsburgh: Polycrystal Book Service.
Hill, T. L. (1956). *Statistical Mechanics.* New York: McGraw-Hill.
Hohenberg, P. & Kohn, W. (1964). *Phys. Rev.* **136**, B864–B871.
Hummer, G., Garde, S., Garcia, A. C., Paulaitis, M. E. & Pratt, L. R. (1998). *J. Phys. Chem.* B**102**, 10469–10482.
Jacobson, R. A. & Beckman, D. E. (1979). *Acta Cryst.* A**35**, 339–340.
Jaynes, E. T. (1957*a*). *Phys. Rev.* **106**, 620–630.
Jaynes, E. T. (1957*b*). *Phys. Rev.* **108**, 171–190.
Jaynes, E. T. (1978). *The Maximum Entropy Formalism*, edited by R. D. Levine & M. Tribus, pp. 15–118. Cambridge, MA: MIT Press.
Jaynes, E. T. (1983). *Papers on Probability, Statistics and Statistical Physics, Brandeis Lectures*, edited by R. D. Rosencrantz, pp. 39–76. Dordrecht: Reidel.
Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.
Klug, A. (1958). *Acta Cryst.* **11**, 515–543.
Landau, L. D. & Lifshitz, E. M. (1958). *Statistical Physics.* London: Pergamon Press.
Lemaréchal, C. & Navaza, J. (1991). *Acta Cryst.* A**47**, 631–632.
Levine, R. D. & Tribus, M. (1978). *The Maximum Entropy Formalism.* Cambridge, MA: MIT Press.
Lipson, H. & Cochran, W. (1966). *The Crystalline State*, Vol. III. *The Determination of Crystal Structures.* London: Bell and Sons.
Luenberger, D. G. (1984). *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley.
Luger, P. & Fuchs, J. (1986). *Acta Cryst.* A**42**, 380–386.
McGreevy, R. L. & Howe, M. A. (1991). *Phys. Chem. Liquids*, **24**, 1–12.
McLachlan, A. D. (1989). *Entropy and Bayesian Methods*, edited by J. Skilling, pp. 241–249. Dordrecht: Kluwer.
McLachlan, A. D. (1999). *Acta Cryst.* A**55** Supplement, Abstract P12.BB.007.
McLachlan, A. D. (2001). *Acta Cryst.*, A**57**, 152–162.
McLachlan, A. D. & Harris, R. A. (1961). *J. Chem. Phys.* **34**, 1451–1452.
Main, P., Fiske, S. J., Hull, S. E., Lessinger, L., Germain, G., Declercq, J. P. & Woolfson, M. M. (1980). *MULTAN80. A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Universities of York, England, and Louvain, Belgium.
Mayer, J. E. & Mayer, M. G. (1940). *Statistical Mechanics.* New York: John Wiley.
Mermin, N. D. (1964). *Phys. Rev.* **137**, A1441–A1443.
Morita, T. & Hiroike, K. (1961). *Prog. Theor. Phys. Jpn*, **25**, 537–578.
Navaza, J. (1985). *Acta Cryst.* A**41**, 232–244.
Navaza, J. (1986). *Acta Cryst.* A**42**, 212–223.
Naya, S., Nitta, I. & Oda, T. (1965). *Acta Cryst.* **19**, 734–747.
Ornstein, L. S. & Zernicke, F. (1914). *Proc. Akad. Sci. (Amsterdam)*, **17**, 793–806.
Parr, R. G. & Yang, W. (1989). *Density-Functional Theory of Atoms and Molecules.* Oxford University Press.
Pavelčík, F. (1986). *J. Appl. Cryst.* **19**, 488–491.
Pavelčík, F. (1994). *Acta Cryst.* A**50**, 467–474.
Percus, J. K. & Yevick, G. J. (1958). *Phys. Rev.* **110**, 1–13.
Prince, E., Sjolin, L. & Alenljung, R. (1988). *Acta Cryst.* A**44**, 216–222.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Shannon, C. E. (1948*a*). *Bell System Tech. J.* **27**, 379–423.
Shannon, C. E. (1948*b*). *Bell System Tech. J.* **27**, 623–656.
Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois, IL, USA.
Sheldrick, G. M. (1985). *SHELXS86. Program for the Solution of Crystal Structures.* University of Göttingen, Germany.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Simpson, P. G., Dobrott, R. D. & Lipscomb, W. N. (1965). *Acta Cryst.* **18**, 169–179.
Terwilliger, T. C., Kim, S. H. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 1–5.
Weeks, C. M., De Titta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Wilkins, S. W. (1983). *Acta Cryst.* A**39**, 896–898.