Acta Crystallographica Section A
**Foundations of Crystallography**
ISSN 0108-7673

Received 23 June 2000
Accepted 3 October 2000



# The pair-functional method. III. The pairing forces

**A. D. McLachlan**

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

The theory of the pair-functional ensemble is developed to provide estimates of the pairing forces from experimental X-ray intensities. The statistical mechanics of the grand ensemble leads to a diagram expansion for the forces, in terms of the direct correlation function of the fluid ensemble combined with a series of small higher-order corrections. A simpler treatment, based on a biased Gaussian probability distribution, gives approximate formulae, valid for reflections of any type in all space groups. The role of symmetry is analysed. The entropy of an asymmetrical ensemble can always be increased by averaging it over equivalent positions of the atoms in the true space group, with the result that the atoms naturally tend to adopt the highest symmetry compatible with the data. In a cell with different types of atom, the atoms experience a single force function but they interact with a strength proportional to the products of their scattering factors. Numerical estimates are given for typical cases.

## 1. Introduction

The pair-functional theory is the basis of a new direct method for solving the structures of small molecules [McLachlan (1999); McLachlan (2001*a*,*b*) (referred to below as papers I and II)]. It uses a unique statistical ensemble of atoms that interact through a specific set of long-range pairing forces. The structure is solved by searching the ensemble for well paired clusters of atoms and this needs a rather accurate initial estimate of the pairing force. We use the term *pairing force* as the name for the functions $\psi(\mathbf{u})$ in real space or $\psi_E(\mathbf{H})$ in reciprocal space, and *total pair potential* for the function $\Psi_N$, which is the sum of the pair interactions of all the atoms in the cell. The pairing force (see paper I) is a statistical force, defined in terms of the gradients of the ensemble entropy. In this paper, we show how to calculate the normalized Fourier components $\psi_E(\mathbf{H})$ of the pairing forces from the observed X-ray intensities of the measured reflections, $\mathbf{H}$.

There are several possible ways to deduce these forces. The most rigorous is from the many-body statistical mechanics of fluids (Hansen & McDonald, 1986), which uses the grand canonical ensemble with a fluctuating number of atoms (Mayer & Mayer, 1940; Hill, 1956). This theory produces exact infinite-order perturbation expansions for the forces in terms of many-particle diagrams (Morita & Hiroike, 1961; De Dominicis, 1962, 1963). The first term in this expansion is the direct correlation function of the fluid (Ornstein & Zernicke, 1914).

A far simpler approach is to use a general maximum-entropy argument which works with biased probability distributions (Jaynes, 1978; Levine & Tribus, 1978). This exploits the well known Gaussian probability distributions of the structure factors. The argument, based on the concept of unbiased natural probability distributions described in Appendix *C*, agrees with the first level of approximation in the many-body theory, and makes it possible to derive a more general approximate pairing force for molecules which contain several types of atom.

A third purely numerical way to estimate the forces would seek to minimize the dual function of the ensemble (Agmon *et al.*, 1978; Gill *et al.*, 1981; Luenberger, 1984), which is defined in paper I (McLachlan, 2001*a*). Unfortunately, this method would need as much computation as a full solution of the structure.

We also consider the important effects of space-group symmetry on the pairing forces. The correct way to build a pair-functional ensemble in a cell with symmetry is to use interacting systems of symmetry-related atom clusters (Castleden, 1987) in equivalent positions (*International Tables for Crystallography*, 1987). These must be set up with one of the standard allowed origins (Hauptman & Karle, 1956; Hauptman, 1972; Giacovazzo, 1980). However, it is also permissible to map the atoms onto a cell of lower fictitious symmetry, such as *P*1. Thus, in a cell where the measured data indicate that the space group is actually *G*, it is possible to represent the constrained paired-atom ensemble in various alternative fictitious ways, using subgroups of *G* with lower symmetry. We shall show below that the ensemble with the true symmetry *G* always has a higher entropy than any of the others. This means that the pair interaction forces should normally enable a random asymmetric starting set of atoms to adapt itself to the true symmetry during a search for the correct structure. When the symmetry arguments and the biased Gaussian approximation are combined together they

lead to a set of simple rules (Stewart & Karle, 1976; Giacovazzo, 1980) that give the approximate pairing force for any class of reflection.

Higher-order corrections to the pairing forces will often be small. The Gaussian approximation holds accurately for all but the weakest and strongest reflections and the pairing force for a particular reflection **H** is only slightly affected by the intensities of other symmetry-independent reflections **K**. Another kind of correction is sometimes important for tightly packed atoms at short distances, where exclusion effects come into play. The classical phase probability distributions of crystallography (Hauptman & Karle, 1953; Naya *et al.*, 1965; Hauptman, 1975) are based on the properties of independent atoms scattered randomly throughout the cell. Real atoms behave more like hard spheres (Hansen & McDonald, 1986) with a certain collision diameter and exclude one another from the overlap regions. The last topics considered in this paper are numerical estimates of the total pair potentials and the complete ensemble entropies under typical conditions.

The main mathematical proofs and definitions will be found in the Appendices.

## 2. Many-body theory of pairing forces

### 2.1. The grand ensemble

The many-body theory is based on the particle distribution functions of a spatially uniform grand ensemble, as described in Appendix *A*, where the variable number of atoms has an average value equal to $N$, the fixed number of atoms in the target molecule. To set up the ensemble, we first have to convert the observed intensities $I_{\rm obs}(\mathbf{H})$ of the measured reflections **H**, with $\mathbf{H} \neq 0$, into normalized structure intensities $|E_{\rm obs}(\mathbf{H})|^2$ (Blessing *et al.*, 1998). These will be the target intensities for the representative ensemble, so that

$$|T(\mathbf{H})|^2 = |E_{\rm obs}(\mathbf{H})|^2. \tag{1}$$

The grand ensemble distributions are usually expressed in terms of fractional cell coordinates, with a two-particle correlation function $k^{(2)}(\mathbf{u})$ and a normalized pair-correlation function $h^{(2)}(\mathbf{u})$, which both describe pairs of particles at positions **x** and **y** separated by a shift vector **u**. These functions are related by the equation

$$h^{(2)}(\mathbf{u}) = (1/N^2)k^{(2)}(\mathbf{u}) - 1. \tag{2}$$

The pair-correlation function of the ensemble has Fourier components $\hat{h}^{(2)}(\mathbf{H})$ and these components have to match the target intensities.

$$\hat{h}^{(2)}(\mathbf{H}) = (1/N)\{|T(\mathbf{H})|^2 - 1\}. \tag{3}$$

In other words, the originless Patterson function of the ensemble must agree with that deduced from the measured data. The constrained grand ensemble behaves like the Boltzmann distribution of a set of atoms that interact through a unique pairing force $\psi(\mathbf{u})$, as outlined in Appendix *B* and paper I (McLachlan, 2001*a*). The pairing force is determined implicitly as a functional of all the target intensities. In the grand ensemble, $N$ is not fixed, and in the remainder of this section we must understand $N$ to represent the ensemble average $\langle N \rangle$.

### 2.2. The direct correlation function

The statistical theory of fluids provides a well known perturbation analysis for the variations of the particle distribution functions under weak changes of the potentials (Hansen & McDonald, 1986). If the perturbation is a change of the single-particle potential energy, then Yvon's theorem (Yvon, 1958) states that an initially uniform fluid with mean density $N$, subjected to a potential $-\chi(\mathbf{y})$ at a fixed point **y**, undergoes small changes of probability density $\delta\rho^{(1)}(\mathbf{x})$ at other points **x**. In our ensemble, the equivalent equation is

$$\delta\rho^{(1)}(\mathbf{x}) = N\beta\chi(\mathbf{x}) + N^2\beta \int h^{(2)}(\mathbf{x}-\mathbf{y})\chi(\mathbf{y})\,{\rm d}\mathbf{y}, \tag{4}$$

where $\beta = 1/kT = 1$ is the effective inverse temperature. The density changes consist of a local single-particle effect at the point **x** itself, combined with an indirect long-range two-particle effect proportional to the unperturbed pair-correlation function $h^{(2)}(\mathbf{u})$. The signs in the Yvon equation above are both positive because of the positive sign of the Boltzmann factor $\exp(\beta\Psi)$ in the paired-atom ensemble. The more complicated changes of $h^{(2)}(\mathbf{u})$ produced by introducing a weak two-particle potential $\psi(\mathbf{u})$ into an initially perfect gas without interatomic forces are expressed in terms of the Ornstein–Zernicke direct correlation function $c^{(2)}(\mathbf{x}, \mathbf{y}) = c^{(2)}(\mathbf{u})$, which is a function of the separation of the points $\mathbf{u} = \mathbf{x} - \mathbf{y}$ (Ornstein & Zernicke, 1914). This function is defined implicitly by the non-linear convolution relation

$$h^{(2)}(\mathbf{u}) = c^{(2)}(\mathbf{u}) + N \int c^{(2)}(\mathbf{u}-\mathbf{u}_1)h^{(2)}(\mathbf{u}_1)\,{\rm d}\mathbf{u}_1. \tag{5}$$

A perturbation analysis based on Yvon's equation then shows that the two-particle potential required to generate a prescribed function $h^{(2)}(\mathbf{u})$ is

$$\beta\psi(\mathbf{u}) = c^{(2)}(\mathbf{u}) \tag{6}$$

in which $\beta = 1$ and $c^{(2)}(\mathbf{u})$ itself is derived implicitly from $h^{(2)}(\mathbf{u})$ *via* the convolution relation. Introducing the Fourier components $\hat{h}^{(2)}(\mathbf{H})$ and $\hat{c}^{(2)}(\mathbf{H})$ into the convolution, we find

$$\hat{h}^{(2)}(\mathbf{H}) = \hat{c}^{(2)}(\mathbf{H}) + N\hat{c}^{(2)}(\mathbf{H})\hat{h}^{(2)}(\mathbf{H}) \tag{7}$$

or

$$\hat{c}^{(2)}(\mathbf{H}) = \frac{\hat{h}^{(2)}(\mathbf{H})}{1 + N\hat{h}^{(2)}(\mathbf{H})}. \tag{8}$$

We therefore arrive at the important result that the normalized pairing force is

$$\psi_E(\mathbf{H}) = \frac{|T(\mathbf{H})|^2 - 1}{|T(\mathbf{H})|^2}. \tag{9}$$

Any Fourier components of the force must be set to zero if the amplitude $T(\mathbf{H})$ is not measured or when $\mathbf{H} = (0, 0, 0)$. Note that in this approximation each Fourier component $\psi_E(\mathbf{H})$ of the force is independent of the target intensities $|T(\mathbf{K})|^2$ of other reflections **K**. Thus the different reflections behave

nearly independently in the pair ensemble. In real space, the Yvon approximation to the two-particle pair force can be written as a non-linear convolution series in terms of the levelled originless Patterson function or the levelled autocorrelation function. Using the relation

$$\Delta k^{(2)}(\mathbf{u}) = k^{(2)}(\mathbf{u}) - \langle N(N-1)\rangle, \tag{10}$$

we obtain the infinite series

$$\begin{aligned} N^2\psi(\mathbf{u}) = {} & \Delta k^{(2)}(\mathbf{u}) - (1/N)\int \Delta k^{(2)}(\mathbf{u}-\mathbf{u}_1)\Delta k^{(2)}(\mathbf{u}_1)\,\mathrm{d}\mathbf{u}_1 \\ & + (1/N^2)\iint \Delta k^{(2)}(\mathbf{u}-\mathbf{u}_1)\Delta k^{(2)}(\mathbf{u}_1-\mathbf{u}_2) \\ & \times \Delta k^{(2)}(\mathbf{u}_2)\,\mathrm{d}\mathbf{u}_1\,\mathrm{d}\mathbf{u}_2 - \ldots . \end{aligned} \tag{11}$$

The theory of fluids thus yields a simple approximate initial form for the pairing force. It does not, however, give any guide to the accuracy of the perturbation formula. In a later section, we shall see that there are other arguments that confirm the approximation and that these can be developed further to indicate the limits of accuracy of the Yvon equation when applied to crystallographic problems.

### 2.3. Estimates from integral equations

The theory of fluids provides further more accurate estimates of the two-body potentials in terms of the observed correlation functions. These are valid even for strongly interacting sets of atoms, such as hard spheres. Morita & Hiroike (1961) and De Dominicis (1963) each give equivalent infinite-order diagram expansions of the potential in terms of the functions $h^{(2)}(\mathbf{u})$ and $g^{(2)}(\mathbf{u}) = [1 + h^{(2)}(\mathbf{u})]$. Both their theories do, however, assume a complete knowledge of all the Fourier components of the relevant distributions, including the unmeasured intensities. Their results are of the form

$$\psi(\mathbf{u}) = \log[1 + h^{(2)}(\mathbf{u})] - [h^{(2)}(\mathbf{u}) - c^{(2)}(\mathbf{u})] - B^{(2)}(\mathbf{u}). \tag{12}$$

Here $B^{(2)}(\mathbf{u})$ is the sum of the so-called bridge diagrams. Each bridge diagram specifies a pair of fixed atoms and a further set of at least two moveable atoms with five or more bonds between the atoms. The whole pattern of a connected diagram forms an irreducible network according to certain topological rules. The value of any diagram is a power of $N$ multiplied by integrals of products of factors $h^{(2)}(\mathbf{r}_1 - \mathbf{r}_2)$ from all the bonds. Usually, $B^{(2)}$ is neglected, giving rise to the well known *hypernetted chain approximation* (HNC), which can then be expanded as a power series in $h^{(2)}(\mathbf{u})$.

$$\begin{aligned} \psi(\mathbf{u}) &= \log[1 + h^{(2)}(\mathbf{u})] - [h^{(2)}(\mathbf{u}) - c^{(2)}(\mathbf{u})] \\ &= c^{(2)}(\mathbf{u}) - \tfrac{1}{2}[h^{(2)}(\mathbf{u})]^2 + \tfrac{1}{3}[h^{(2)}(\mathbf{u})]^3 - \ldots \quad \text{(HNC)}. \end{aligned} \tag{13}$$

Another valuable approximation, especially for short-range potentials acting on hard spheres, is the *Percus–Yevick equation* (P-Y), which may be rewritten in the form

$$\psi(\mathbf{u}) = \log[1 + h^{(2)}(\mathbf{u})] - \log[1 + h^{(2)}(\mathbf{u}) - c^{(2)}(\mathbf{u})]. \quad \text{(P-Y)}. \tag{14}$$

Before leaving the topic of the many-body pair potentials, we note that the hypernetted chain equation introduces a coupling between different Fourier components. Thus the Fourier transform of the series expansion above gives

$$\hat{\psi}(\mathbf{H}) = \hat{c}^{(2)}(\mathbf{H}) - \tfrac{1}{2}\sum_{\mathbf{K}} \hat{h}^{(2)}(\mathbf{H}-\mathbf{K})\hat{h}^{(2)}(\mathbf{K}) + \ldots \quad \text{(HNC)} \tag{15}$$

with leading terms

$$\begin{aligned} \psi_E(\mathbf{H}) = {} & [|T(\mathbf{H})|^2 - 1]/|T(\mathbf{H})|^2 \\ & - (1/2N)\sum_{\mathbf{K}}[|T(\mathbf{H}-\mathbf{K})|^2 - 1][|T(\mathbf{K})|^2 - 1]. \end{aligned} \tag{16}$$

The correction terms proportional to $1/2N$ show that the measured intensities that belong to triplets of reflections such as $(\mathbf{H}, \mathbf{K}-\mathbf{H}, -\mathbf{K})$ give rise to interference effects on the pairing forces.

## 3. Symmetry and uniqueness

In paper I, we gave a short account of the uniqueness principles for maximum-entropy ensembles. The key conclusion was that any combination of linear constraints on the many-particle distribution functions will generate an effectively unique ensemble (Jaynes, 1978, 1983). We showed that, if there are two or more distributions that satisfy the constraints (McLachlan & Harris, 1961), any mixture of these distributions will have at least as high an entropy. Degenerate solutions may occur in practice, associated with different cell origins or opposite handedness. These general results mean that the pair-functional ensemble generated by any feasible set of experimental data is effectively unique, apart from degeneracy. We can extrapolate this conclusion to conjecture that under normal conditions, when there is a sufficiently complete and accurate set of high-resolution data, the ensemble will also contain an effectively unique structural solution of the phase problem. The Boltzmann distribution of atomic conformations will be as well defined as possible, consistent with the quality of the data. There are certain special sets of different small structures of point atoms that have identical Patterson functions. These homometric structures (Lipson & Cochran, 1966, p. 164) are unlikely to arise in large molecules and would be treated as coexisting alternative solutions within the paired-atom ensemble.

### 3.1. Symmetry groups

Symmetry enters into practical problems in several different ways. First, because the constraints implied by the data refer to an assumed molecular structure of identical particles, the given conditions are symmetric under particle exchange. Thus, the many-body probability $f^{(N)}(\mathbf{r}^N)$ must be a symmetric function of the particle coordinates $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$. Secondly, the observed data are normally consistent with a single known crystallographic symmetry group $\mathbf{G}$. As we mentioned in the *Introduction*, there is generally a choice in practical calculations between constructing a maximum-entropy ensemble in the asymmetric unit of the group $\mathbf{G}$ or in the full cell with symmetry $P1$. We now show that, for given data that is

consistent not only with the full group **G** but also with subgroups of **G** having lower symmetry, the ensemble with the highest entropy also belongs to the subgroup of highest symmetry, that is to **G** itself. Therefore, if an ensemble is set up in $P1$ without any special symmetry and varied to reach its global maximum entropy under the given data conditions, compatible with **G**, it should spontaneously acquire the symmetry of **G**.

To prove the general result, we consider the construction of symmetric probability distributions (Giacovazzo, 1980). Suppose that there are $G$ operations $\mathbf{O}_g$ in the group that act on an atomic cordinate $\mathbf{r}$ through a rotation matrix $\mathbf{R}_g$ and a translation vector $\mathbf{t}_g$, so that

$$\mathbf{O}_g\mathbf{r} = (\mathbf{R}_g\mathbf{r} + \mathbf{t}_g). \tag{17}$$

Apply any symmetry operation to the full coordinates $\mathbf{r}^N$ of all the atoms and the full probability distribution, in which $f^{(N)}(\mathbf{r}^N)$ becomes $f^{(N)}(\mathbf{O}_g\mathbf{r}^N)$, where $\mathbf{O}_g\mathbf{r}^N$ stands for the $N$-particle coordinate vector $(\mathbf{O}_g\mathbf{r}_1, \mathbf{O}_g\mathbf{r}_2, \ldots, \mathbf{O}_g\mathbf{r}_N)$. With this notation, we see that any unsymmetrical distribution that satisfies the given linear constraints can be made symmetrical by averaging over a combination of transformed distributions under the symmetry group. For example, if $\mathbf{O}_g$ is a twofold rotation, then any many-particle distribution $f_A^{(N)}$ can be symmetrized about the rotation axis by constructing the combined distribution

$$f^{(N)}(\mathbf{r}_N) = \tfrac{1}{2}[f_A^{(N)}(\mathbf{r}^N) + f_A^{(N)}(\mathbf{O}_g\mathbf{r}^N)], \tag{18}$$

which has entropy $S_N \geq S_A$. By repeating this process with other group generator symmetry elements, the distribution can be given the full symmetry **G**. The symmetry transformations of the group are uniquely specified by the conventions for the choice of cell origin (Hauptman & Karle, 1956; Giacovazzo, 1980) and the fully symmetrized ensemble for a properly defined group is an equally weighted mixture of all the permissible origins and enantiomorphs. For example, in $P1$, the fully symmetrized ensemble has a continuous distribution of origins and both enantiomorphs are present. In $P2_1$, the origins are continuously distributed along the screw axis, but the axis itself is fixed in the cell.

### 3.2. Equivalent positions

The most direct way to generate a many-body distribution with the correct symmetry of a group **G** is by using the standard equivalent positions of the space group, with a set of $N/G$ reference atoms at positions $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N/G})$. Each reference atom generates a multiplet of $G$ image points denoted

$$\mathbf{x}^{(G)} = (\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gG}), \tag{19}$$

in which $(g_1, g_2, \ldots, g_G)$ are the group operations, with $g_1 = E$ as the identity operation. The image points are distinct unless the reference atom occupies a special position in the cell. This approach has been used in reciprocal space by Castleden (1987) to derive phase probability distribution formulae for a general space group. With the multiplet method, one may set up a distribution function $f^{(N/G)}(\mathbf{x}^{(N/G)})$ in terms of the positions of the reference atoms and assign to it an entropy

$$S_{N/G} = -\int f^{(N/G)}(\mathbf{x}^{(N/G)}) \log[f^{(N/G)}(\mathbf{x}^{(N/G)})(N/G)!]\, \mathrm{d}\mathbf{x}^{(N/G)}. \tag{20}$$

It can then be shown by arguments similar to those used in paper I that the maximum-entropy ensemble for a cell with symmetry **G** and a given pair-correlation function is of the same Boltzmann form as in $P1$ except that the pair potential involves interactions between all the images of all the reference atoms, including pairs of different images of the same reference atom.

$$\Psi_{N/G}(\mathbf{x}^{(N/G)}) = \sum_{ig1<jg2} \psi(\mathbf{O}_{g1}\mathbf{x}_i - \mathbf{O}_{g2}\mathbf{x}_j). \tag{21}$$

This sum is over all $N/G$ values of $i$ and $j$ and all $G$ values of $g1$ and $g2$, without self-interactions, and has $N(N-1)/2$ terms in all. As a simple example, to illustrate the pair potential in the space group $P\bar{1}$, we take the case of just two reference atoms at $\mathbf{x}$ and $\mathbf{y}$. The inversion centre is at the origin. These atoms generate the four images $\pm\mathbf{x}$, $\pm\mathbf{y}$ and then the six image pairs formed by these points have a total pairing potential of

$$\Psi_{N/G} = \psi(2\mathbf{x}) + \psi(2\mathbf{y}) + 2\psi(\mathbf{x}-\mathbf{y}) + 2\psi(\mathbf{x}+\mathbf{y}). \tag{22}$$

The first two terms are functions of the single-particle coordinates $2\mathbf{x}$, $2\mathbf{y}$ and they show how the effect of symmetry is to convert some of the pair interactions into an apparent single-particle field $\chi(\mathbf{r}) = \psi(2\mathbf{r})$. Hence the $P\bar{1}$ pair-functional ensemble will in general have a non-uniform single-particle probability density. This non-uniform spatial density is accounted for in reciprocal space by the fixed signs of the even–even–even Fourier components $2\mathbf{h} = (2h, 2k, 2l)$, which are semi-invariant under changes of origin (Hauptman & Karle, 1956; Hauptman, 1972; Giacovazzo, 1980). This brief discussion of symmetry indicates how the pair-functional theory can be extended, in principle, to any type of crystal symmetry.

## 4. The perturbed Gaussian approximation

### 4.1. The biased ensemble

The probability distributions of the normalized Fourier intensities within the pair-functional ensemble can be understood readily in terms of the concept of natural distributions described in Appendix *C*. The natural distribution for any reflection **H** is well approximated by a Gaussian and the normalized pairing force $\psi_E(\mathbf{H})$ represents an imposed bias. In the space group $P1$, the natural distribution of the complex structure factor $E = (A + iB)$ is

$$f_{\mathrm{nat}}(A, B)\, \mathrm{d}A\, \mathrm{d}B = (1/\pi)\exp(-E^2)\, \mathrm{d}A\, \mathrm{d}B. \tag{23}$$

The calculation in Appendix *D* shows the effect of a biasing term which multiplies the probability by an adjustable factor $\exp(\lambda E^2)$ and leads to a new maximum-entropy distribution

$$f(A, B) = [1/Z(\lambda)] f_{\rm nat}(A, B) \exp(\lambda E^2). \quad (24)$$

The value of $\lambda$ is chosen so that the mean intensity $E^2$ in the revised distribution matches the observed intensity $T^2$, that is

$$\lambda = (T^2 - 1)/T^2. \quad (25)$$

We identify $\lambda$ with the normalized pairing force $\psi_E(\mathbf{H})$ and the result agrees exactly with the one already derived from Yvon's equation. The full distribution changes to

$$f(E^2)\,{\rm d}A\,{\rm d}B = (1/\pi T^2)\exp(-E^2/T^2)\,{\rm d}A\,{\rm d}B, \quad (26)$$

which is simply another Gaussian with a changed width, adjusted to match the target $T^2$. The pairing potential thus has the effect of either inhibiting or amplifying the natural fluctuations of $E^2$ in a random collection of atoms to generate an ensemble with the desired average properties.

The theory for space group $P\bar{1}$ follows similar lines, starting from the natural distribution of the real structure factor

$$f_{\rm nat}(A)\,{\rm d}A = [1/(2\pi)^{1/2}]\exp(-\tfrac{1}{2}A^2)\,{\rm d}A \quad (27)$$

with the result that

$$\lambda = (T^2 - 1)/2T^2. \quad (28)$$

### 4.2. Relaxed potentials

One difficulty with the biased ensemble is that the pairing force $\psi_E(\mathbf{H})$ becomes very large and negative when $|T(\mathbf{H})|$ approaches zero. This can be avoided by introducing a small tolerance $\sigma$ in the fit to the target intensities. The error potential $Y$ (Appendix $D$) provides a finite relaxed pairing force. Alternatively, we can use a simplified empirical form of pairing force function that includes a small constant cut-off intensity $|T_{\rm low}|^2$. The error potential calculation for $P1$ gives a finite pairing force

$$\psi_E(\mathbf{H}) = \lambda = \frac{2(T^2 - 1)}{(T^2 + \sigma^2) + W}, \qquad W^2 = (T^2 - \sigma^2)^2 + 4\sigma^2. \quad (29)$$

The empirical cut-off function leads to a useful simple formula, valid in both $P1$ and $P\bar{1}$. The results for $P1$ are

$$\psi_{\rm acentric}(\mathbf{H}) = \frac{(T^2 - 1)}{(T^2 + T_{\rm low}^2)} \quad (30)$$

$$S_{\rm acentric} = \log\left(\frac{T^2 + T_{\rm low}^2}{1 + T_{\rm low}^2}\right) - \left(\frac{T^2 - 1}{1 + T_{\rm low}^2}\right). \quad (31)$$

In $P\bar{1}$, the pairing force and the entropy are

$$\psi_{\rm centric}(\mathbf{H}) = \tfrac{1}{2}\psi_{\rm acentric}(\mathbf{H}) \quad (32)$$

$$S_{\rm centric}(\mathbf{H}) = \tfrac{1}{2}S_{\rm acentric}(\mathbf{H}) + \tfrac{1}{2}\log 2\pi. \quad (33)$$

In both space groups, the mean value of $E^2$ obtained from the biased distribution does not fit the target exactly, especially for the weak reflections:

$$\langle E^2 \rangle = \frac{(T^2 + T_{\rm low}^2)}{(1 + T_{\rm low}^2)}. \quad (34)$$

### 4.3. Symmetry weights

The results of the Gaussian approximation above are valid for all space groups. This is because the probability distribution of every normalized structure factor depends principally on the symmetry class of the reflection. Acentric reflections have an isotropic two-dimensional distribution in the complex $(A, B)$ plane, while centric reflections have a one-dimensional distribution in $A$ or $B$ only. For each space group, the conversion from structure factors to normalized intensities follows the usual rule (Giacovazzo, 1980; Blessing *et al.*, 1998),

$$|E(\mathbf{H})|^2 = |F(\mathbf{H})|^2/\varepsilon(\mathbf{H})\Sigma_I, \quad (35)$$

where

$$\Sigma_I = \sum_{j=1}^{N} f_j^2(\mathbf{H}) \quad (36)$$

is the sum of the squared atomic scattering factors and $\varepsilon(\mathbf{H})$ is the statistical weight of the reflection. The weight depends in turn on the centring order of the lattice and the multiplicity of the individual reflection $\mathbf{H}$ (Stewart & Karle, 1976). We conclude that the biased Gaussian approach gives a correctly weighted first approximation to the pairing force in every space group.

### 4.4. Limits of the Gaussian approximation

When the observed intensity $|T|^2$ is very large, the biased Gaussian distribution becomes broad and there are large fluctuations of $E^2$ within the pair-functional ensemble. The root-mean-square intensity fluctuation $(\Delta I)^2 = \langle (E^2 - T^2)^2 \rangle$ becomes large, with $\Delta I = |T|^2$, and the mode $\mathbf{H}$ becomes unstable. This instability associated with the strong reflections can be countered by using an additional protective term in the total many-body potential. For example, the pairing force can be mixed with a term proportional to the correlation coefficient between the intensity fluctuations of the model and the target. This term will keep the value of $E^2$ closer to $T^2$.

Another source of error is that the natural probability distributions of the structure factors themselves deviate from the Gaussian limit when $E^2$ is large. A careful analysis of the natural distribution up to the fourth order in the Gram–Charlier series (Cramer, 1951; Klug, 1958; Giacovazzo, 1980) gives

$$f_{\rm nat}(A, B) = (1/\pi)\exp(-E^2)[1 - (1/4N)(E^4 - 4E^2 + 2)] \quad (37)$$

with significant corrections only when $E^2$ approaches $N^{1/2}$. Another analysis by the independent-atom maximum-entropy approximation (Bricogne, 1984) gives an expansion of $\log f_{\rm nat}$ in terms of unitary structure factors $U = E/N^{1/2}$. For small values of $U$, this method gives a power-series expansion

$$f_{\rm nat}(A, B) = (1/\pi)\exp(-E^2)[1 - (1/4N)E^4 - (5/36N^2)E^6]. \quad (38)$$

The independent-atom entropy is derived from an ensemble, with specified phases, which is different from the one considered here, and so its entropy has no direct relation with the entropy of the pair-functional ensemble used in this paper.

Clearly, when there are more than 100 atoms, any corrections to the Gaussian distributions for strong reflections are likely to be important only when $E^2$ approaches ten, and will affect only a very small fraction of the observed reflections. The corrections caused by weak interference effects between different reflections may be more important, since a large number of triplets and quartets can all contribute to a single potential.

In a later paper, we shall describe the many-body theory of the strong coupling limit of the paired-atom ensemble. This gives an accurate expression for the pairing force over the entire range of intensities.

## 5. Several types of atom

If the cell contains two or more types of atom, with different scattering factors $f_a, f_b, \ldots$, the ensemble behaves like a multicomponent fluid mixture (Morita & Hiroike, 1961). The generalized maximum-entropy ensemble that represents a mixture with a given observed originless Patterson function is described in Appendix *E*. The most important conclusion is that atoms of different types interact with potentials of the form $\Psi = f_a f_b \psi(\mathbf{r}_{iA} - \mathbf{r}_{jB})$ which are proportional to their scattering factors. The pairing force $\psi(\mathbf{u})$ is the same for all the atoms and is estimated in the Gaussian approximation, as before. The normalized pairing force $\psi_E(\mathbf{H})$ can still be expressed in terms of the normalized target amplitude $|T(\mathbf{H})|^2 = |F_T(\mathbf{H})|^2/\Sigma_I$, with the result that

$$\psi_E(\mathbf{H}) = \Sigma_I \hat{\psi}(\mathbf{H}) = \frac{|T(\mathbf{H})|^2 - 1}{|T(\mathbf{H})|^2}. \qquad (39)$$

## 6. Exclusion effects

Exclusion effects arise when two atoms cannot occupy the same space in the cell. The simplest case is in the point-atom model, where $N$ point atoms of scattering factor $f$ are sprinkled randomly over a fine grid with $L$ vertices. Here it is easy to show, by Parseval's theorem, that the mean-square structure factor for $\mathbf{h} \neq 0$ is

$$\langle |F(\mathbf{h})|^2 \rangle = Nf^2(1 - x_{\rm fill}), \qquad (40)$$

where $x_{\rm fill} = N/L$ is the filling factor of the grid or the fraction of occupied points. This correction $(1 - x_{\rm fill})$ should be included in the natural Gaussian distribution of any structure factor. A similar correction occurs in the structure factors of a uniform molecular envelope and is related to Babinet's principle. A general further treatment of exclusion effects would be complicated, even for hard-sphere atoms.

## 7. Numerical estimates

It is useful to estimate the expected total pairing potential and entropy of the pair ensemble that matches a typical set of measured X-ray intensities. In the Gaussian approximation, all the $n_R$ measured reflections contribute independently and so the mean expected values can be written

$$\Psi_{\rm expec} = n_R \langle \psi(T) \rangle_{\rm nat}, \quad S_{\rm expec} = n_R \langle S(T) \rangle_{\rm nat}, \qquad (41)$$

where the averages are performed over the natural probability distribution of a typical target amplitude $T$. Thus, for a particular measured reflection $\mathbf{H}$ and a definite value of $T$,

$$\psi(T) = \psi_E(\mathbf{H})\{|T(\mathbf{H})|^2 - 1\} = \lambda(T)(T^2 - 1), \qquad (42)$$

with $\lambda(T)$ given by one of the empirical formulae in §4. Also, $S(T)$ is the estimated value of $(S - S_{\rm flat})$. For example, with the intensity cut-off formula averaged over the Gaussian distributions, we obtain

$$\langle \psi(T) \rangle_{\rm acentric} = \int_0^\infty \frac{(T^2 - 1)^2}{(T^2 + T_{\rm low}^2)} \exp(-T^2) 2T \, {\rm d}T \qquad (43)$$

$$\langle \psi(T) \rangle_{\rm centric} = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^\infty \frac{(T^2 - 1)^2}{2(T^2 + T_{\rm low}^2)} \exp(-\tfrac{1}{2}T^2) \, {\rm d}T \qquad (44)$$

with similar expressions for the entropies. The results are collected in Table 1. For example, in the space group $P1$, using the intensity cut-off formula with $T_{\rm low}^2 = 0.25$, we obtain

$$\Psi_{\rm expec} = 0.8442 n_R, \quad S_{\rm expec} = -0.2679 n_R. \qquad (45)$$

The error potential with $\sigma = 0.25$ gives similar values. Since $\Psi$ and $S$ are both logarithms, respectively of a likelihood and a volume in configuration space, we can estimate that each reflection measured multiplies the likelihood by 2.326 and divides the acceptable volume by 1.307. The measurements restrict the degrees of freedom of each atom in proportion to $n_R/N$, the number of reflections per atom.

**Table 1**
Pair potential and entropy per reflection.

| Cut-off intensity estimate | | | |
|---|---|---|---|
| Quantity | Acentric | Centric | $T_{\rm low}^2$ |
| $\Psi$ | 0.8442 | 0.7344 | 0.25 |
| | 1.3368 | 1.3368 | 0.10 |
| $S$ | −0.2679 | −0.2162 | 0.25 |
| | −0.3825 | −0.3225 | 0.10 |

| Error potential estimate | | | |
|---|---|---|---|
| Quantity | Acentric | Centric | $\sigma$ |
| $\Psi$ | 0.9046 | 0.9590 | 0.25 |
| | 1.5141 | 1.8218 | 0.10 |
| $S$ | −0.3226 | −0.2893 | 0.25 |
| | −0.4378 | −0.3919 | 0.10 |

## 8. Conclusions

In summary, this paper, together with the basic principles described in paper I, completes the formal framework of the pair-functional method, up to the point where it is possible to generate the pairing forces that fit any set of experimental data. Further developments of the theory will analyse the probability distributions of structure factors within the paired-

atom ensemble and look at the limit of strong coupling when the pairing forces are very large. These ideas will lead to the temperature-dependent self-consistent-field approximation as a practical method for solving structures. The main results of the present work on pairing forces are:

(i) The many-body theory shows that the forces associated with different reflections are almost independent of one another and they are proportional to the well known direct correlation function of the analogous fluid.

(ii) Small corrections of order $1/N$ arise from interference effects between different reflections or when strong reflections have $E^2$ of order $N^{1/2}$. But these details may not be critical for solving structures.

(iii) The biased Gaussian method gives a useful correct first-order guide to the results of the many-body theory.

(iv) The ensemble for a cell with higher symmetry is also unique and adopts the highest correct symmetry. There are simple rules for calculating the forces.

(v) The ensemble can handle different types of atom.

(vi) The strong pairing forces that are needed to match very weak reflections should be moderated by using a relaxed fit to these intensities.

Given these principles, the programme for applying the pair-functional method in crystallography needs to follow two further directions. One is to construct specialized ensembles for each experimental application, such as isomorphous replacement, solvent flattening, fragment recognition. The other is to develop computing methods that use the pair potential to solve or refine structures, going beyond the elementary methods described in paper II. The pair potential of a set of atoms, or a positive density map, is not hard to calculate and could be included as an option in many kinds of standard refinement programs.

## APPENDIX *A*
## Grand distributions

Many important theoretical questions are best treated by using a grand ensemble in the cell, where the number of particles is not precisely known but has a definite average value $\langle N\rangle$. A grand ensemble is described by a series of distributions $f(N, \mathbf{r}^N)$ for the spatial arrangement of each possible number of atoms. These functions are normalized so that

$$A = \sum_{N=0}^{\infty} \int f(N, \mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N = 1 \tag{46}$$

$$\langle N\rangle = \sum_{N=0}^{\infty} \int N f(N, \mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N. \tag{47}$$

It may also be useful to write each of these distributions in the form

$$f(N, \mathbf{r}^N) = P(N) f^{(N)}(\mathbf{r}^N) \tag{48}$$

in which $P(N)$ is the probability that the system contains exactly $N$ particles and $f^{(N)}$ is a normalized spatial probability. The single-particle and two-particle distribution functions in the grand ensemble are

$$\rho^{(1)}(\mathbf{x}) = \sum_{N=0}^{\infty} \sum_{i=1}^{N} \int \delta(\mathbf{r}_i - \mathbf{x}) f(N, \mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N \tag{49}$$

$$\rho^{(2)}(\mathbf{x}, \mathbf{y}) = \sum_{N=0}^{\infty} \sum_{i\neq j} \iint \delta(\mathbf{r}_i - \mathbf{x}) \delta(\mathbf{r}_j - \mathbf{y}) f(N, \mathbf{r}^N)\, \mathrm{d}\mathbf{r}^N \tag{50}$$

and they obey the conditions

$$\int \rho^{(1)}(\mathbf{x})\, \mathrm{d}\mathbf{x} = \langle N\rangle \tag{51}$$

$$\iint \rho^{(2)}(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = \langle N(N-1)\rangle. \tag{52}$$

The scaled pair distributions are then defined analogously as

$$g^{(2)}(\mathbf{r}_1, \mathbf{r}_2) = \rho^{(2)}(\mathbf{r}_1, \mathbf{r}_2)/\rho^{(1)}(\mathbf{r}_1)\rho^{(1)}(\mathbf{r}_2) \tag{53}$$

with

$$k^{(2)}(\mathbf{u}) = \int \rho^{(2)}(\mathbf{r}, \mathbf{r}+\mathbf{u})\, \mathrm{d}\mathbf{r} \tag{54}$$

and in a uniform ensemble this gives

$$h^{(2)}(\mathbf{u}) = g^{(2)}(\mathbf{u}) - 1. \tag{55}$$

The entropy of the grand ensemble requires a non-uniform prior weight $m_N = 1/N!$ to be given to the particle number $N$ and the resulting entropy must therefore be defined as

$$S = -\sum_{N=0}^{\infty} \int f(N, \mathbf{r}^N) \log[f(N, \mathbf{r}^N) N!]\, \mathrm{d}\mathbf{r}^N, \tag{56}$$

which can be broken down into parts that depend separately on $P(N)$ and $f^{(N)}(\mathbf{r}^N)$:

$$S = -\sum_N P(N) \log[P(N)N!] + \sum_N P(N) S_N. \tag{57}$$

When the value of $N$ is known exactly, with $N = N_0$ and $P(N_0) = 1$, the grand entropy reduces correctly to $S_{N_0} - \log(N_0!)$. Another important special case is when the mean value of $N$ is known to be $M$ and there are no spatial constraints at all. Now the maximum-entropy grand ensemble is a simple Poisson distribution over $N$ with $f^{(N)}(\mathbf{r}^N) = 1$ for each geometrical factor and

$$P(N) = \exp(-M)(M^N/N!), \quad \langle N\rangle = M. \tag{58}$$

## APPENDIX *B*
## Grand ensemble with given particle distributions

The grand ensemble is used for general theoretical purposes, such as the construction of many-body diagrams. The relevant maximum-entropy distributions are constructed in the usual way, using the normalized distribution functions $f(N, \mathbf{r}^N)$ to reproduce the desired average values of $\langle N\rangle$, $\rho^{(1)}(\mathbf{x})$ and $\rho^{(2)}(\mathbf{x}, \mathbf{y})$. The Lagrange multipliers yield the equations

$$\delta S + \lambda \delta A + \mu \delta\langle N\rangle + \int \chi(\mathbf{x}) \delta\rho^{(1)}(\mathbf{x})\, \mathrm{d}\mathbf{x} + \tfrac{1}{2} \iint \psi(\mathbf{x}, \mathbf{y}) \delta\rho^{(2)}(\mathbf{x}, \mathbf{y})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} = 0 \tag{59}$$

and the solution depends on the classical grand partition function $\Xi$, with $\beta = 1$.

$$f(N, \mathbf{r}^N) = (1/\Xi)[\exp(\beta\mu N)/N!] \exp[\beta\Lambda_N(\mathbf{r}^N)], \quad (60)$$

where

$$\Xi = \sum_{N=0}^{\infty} [\exp(\beta\mu N)/N!] Z_N \exp[\beta\Lambda_N(\mathbf{r}^N)]$$
$$\lambda = -\log \Xi + 1. \quad (61)$$

The resultant total entropy is

$$S(\text{max.}) = \log \Xi - \mu\langle N\rangle - \langle\Lambda\rangle. \quad (62)$$

The probability distribution for the number of particles in the grand sample is a distorted Poisson distribution with biases proportional to $Z_N$:

$$P(N) = (1/\Xi)[\exp(\beta\mu N)/N!] Z_N. \quad (63)$$

## APPENDIX *C*
## Entropy in terms of natural probability distributions

Under some conditions, one may obtain a useful aproximation for the maximum-entropy distributions of a quantity from limited knowledge. Consider a general quantity $X(\mathbf{r}^N)$ which is a function of the fractional cell coordinates of $N$ atoms, $\mathbf{r}^N = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, where $\mathbf{r} = (x, y, z)$. The volume of an element of the $3N$-dimensional configuration space is written as $\mathrm{d}\mathbf{r}^N$ and the complete volume is

$$\int \mathrm{d}\mathbf{r}^N = 1. \quad (64)$$

We define the density of states, or volume of configuration space, that corresponds to a particular value $X'$ to be the volume of all points for which $X(\mathbf{r}^N) = X'$ within a range $\mathrm{d}X'$, through the equation

$$\Gamma(X') = \int \delta[X(\mathbf{r}^N) - X'] \, \mathrm{d}\mathbf{r}^N, \quad (65)$$

where $\delta$ is the Dirac delta function. This is normalized so that

$$\int \Gamma(X') \, \mathrm{d}X' = 1. \quad (66)$$

The natural probability distribution of $X$ within the $N$-particle space $\mathbf{r}^N$ with uniform weights is derived from the uniform distribution $f^{(N)}(\mathbf{r}^N) = 1$, whose entropy is $S_{\text{nat}} = 1$. The natural probability distribution thus defined is

$$f_{\text{nat}}(X) \, \mathrm{d}X = \Gamma(X) \, \mathrm{d}X \quad (67)$$

with the mean value

$$\langle X\rangle_{\text{nat}} = \int X\Gamma(X) \, \mathrm{d}X. \quad (68)$$

Now consider the entropy in $\mathbf{r}^N$ of a special form of probability distribution $f(X) \, \mathrm{d}X$ over $X$, in which $f(X)$ is any function, but the probability corresponding to any particular value $X'$ is smeared evenly over the whole region of $\mathbf{r}^N$ for which $X(\mathbf{r}^N) = X'$. Thus, at any point the generated distribution is

$$f^{(N)}(\mathbf{r}^N) \, \mathrm{d}\mathbf{r}^N = [f(X')/\Gamma(X')] \, \mathrm{d}\mathbf{r}^N. \quad (69)$$

The entropy of the special ensemble is

$$S_N = -\int f^{(N)}(\mathbf{r}^N) \log f^{(N)}(\mathbf{r}^N) \, \mathrm{d}\mathbf{r}^N$$
$$= -\int f(X) \log[f(X)/\Gamma(X)] \, \mathrm{d}X \quad (70)$$

and corresponds to the entropy of a distribution in the reduced space of $X$ with a prior weighting of $\Gamma(X)$ or equivalently $f_{\text{nat}}(X)$. The standard maximum-entropy method may now be used to construct a biased ensemble in the reduced space, which matches some desired non-natural value of $\langle X\rangle$. The result is

$$f(X) = [1/Z(\lambda)]\Gamma(X) \exp(\lambda X)$$
$$= [1/Z(\lambda)] f_{\text{nat}}(X) \exp(\lambda X) \quad (71)$$

with

$$Z(\lambda) = \int f_{\text{nat}}(X) \exp(\lambda X) \, \mathrm{d}X = \int \exp[\lambda X(\mathbf{r}^N)] \, \mathrm{d}\mathbf{r}^N \quad (72)$$

for which the entropy is

$$S = \log Z(\lambda) - \lambda\langle X\rangle \quad (73)$$

and the biased mean value is

$$\langle X\rangle = \partial \log Z(\lambda)/\partial\lambda. \quad (74)$$

Only the natural distribution $f_{\text{nat}}(X)$ is needed for this purpose. The detailed distribution of the values of $X$ within $\mathbf{r}^N$ is irrelevant. The numerical value of $\lambda$ needed to match a given target average $\langle X\rangle = X^T$ may be calulated by minimizing the target potential

$$Q(\lambda) = \log Z(\lambda) - X^T\lambda. \quad (75)$$

## APPENDIX *D*
## The Gaussian distribution of structure factors

An important application of the natural distributions is to the normalized complex structure factor $E = (A + iB)$ in the space group $P1$. Here the natural distribution is well approximated by a Gaussian

$$f_{\text{nat}}(A, B) \, \mathrm{d}A \, \mathrm{d}B = (1/\pi) \exp[-(A^2 + B^2)] \, \mathrm{d}A \, \mathrm{d}B$$
$$= \exp(-E^2) 2E \, \mathrm{d}E. \quad (76)$$

We seek the maximum-entropy ensemble which reproduces a given target value of $E^2$, $\langle E^2\rangle = T^2$. The distribution in the reduced space has the form

$$f(A, B) = [1/Z(\lambda)] f_{\text{nat}}(A, B) \exp(\lambda E^2)$$
$$= (1/\pi)(1 - \lambda) \exp[-(1 - \lambda)E^2], \quad (77)$$

in which

$$Z(\lambda) = 1/(1 - \lambda) \quad (78)$$

and

$$T^2 = \partial(\log Z)/\partial\lambda = 1/(1 - \lambda). \quad (79)$$

Thus the required solution has

$$\lambda = (T^2 - 1)/T^2, \quad S = -(T^2 - 1) + \log T^2. \quad (80)$$

This maximum-entropy ensemble for $E^2$ is an isotropic two-dimensional Gaussian in the complex plane, expanded or

contracted according as to whether $T^2$ is greater or less than the natural average $\langle E^2\rangle_{\rm nat} = 1$. The real and imaginary parts fluctuate independently about zero, with $\langle A^2\rangle = \langle B^2\rangle = \frac{1}{2}T^2$. The same form of solution is valid for correctly scaled acentric reflections in every space group.

The biased Gaussian ensemble breaks down as a useful approximation when $T$ approaches zero. The reason for this is simply that a structure factor of exactly zero is an extremely improbable event in any random collection of atoms (except for reasons of symmetry). Therefore, the pairing force $\lambda$ needed to ensure a small ensemble average value for $T$ becomes very large and negative. Also, the entropy $S$ diverges like $\log T^2$. In practice, it is important not to distort the whole force field by trying to fit small structure factors with high precision and we shall be content to fit the intensities within a small tolerance $\sigma$. The calculation uses the dual function $Q(\lambda)$ and the error potential $Y(\lambda)$ defined in Appendix *C* of paper I, in which

$$Y(\lambda) = Q + \tfrac{1}{2}\sigma^2\lambda^2, \tag{81}$$

$$Q = \log Z(\lambda) - \lambda T^2. \tag{82}$$

The unique minimum of $Y(\lambda)$ as a function of $\lambda$, for given $T^2$ and $\sigma$, is defined by the condition

$$\partial Y/\partial\lambda = 1/(1-\lambda) + \lambda\sigma^2 - T^2 = 0 \tag{83}$$

and $\lambda$ is the root of this quadratic equation.

$$\lambda = 2(T^2-1)/[(T^2+\sigma^2) + W] \tag{84}$$

$$W^2 = (T^2-\sigma^2)^2 + 4\sigma^2. \tag{85}$$

The solution is finite when $T = 0$, where $W^2 = (4\sigma^2 + \sigma^4)$, and if $\sigma$ is small the limit is

$$\lambda \approx -1/\sigma \quad (T = 0). \tag{86}$$

In practice, we use a simpler approximation for the acentric reflections. This is to adjust all the target intensities by adding a small empirical correction $T^2_{\rm low}$ to them. The pairing force and the mean fitted intensity become

$$\lambda_{\rm acentric} = \frac{(T^2-1)}{(T^2+T^2_{\rm low})} \tag{87}$$

$$\langle E^2\rangle = \frac{(T^2+T^2_{\rm low})}{(1+T^2_{\rm low})} \tag{88}$$

and the effective entropy is

$$S_{\rm acentric} = \log\left(\frac{T^2+T^2_{\rm low}}{1+T^2_{\rm low}}\right) - \left(\frac{T^2-1}{1+T^2_{\rm low}}\right). \tag{89}$$

In this empirical potential, $T^2_{\rm low}$ takes the place of $\sigma$.

The theory for centric reflections follows similar lines. Starting from the Gaussian natural distribution of the real amplitude $E = A$ and applying the bias, we obtain

$$f_{\rm nat}(A)\,{\rm d}A = [1/(2\pi)^{1/2}]\exp(-\tfrac{1}{2}A^2)\,{\rm d}A \tag{90}$$

$$f(A) = [1/Z(\lambda)]\exp[-\tfrac{1}{2}(1-2\lambda)A^2] \tag{91}$$

with

$$\log Z(\lambda) = \tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(1-2\lambda). \tag{92}$$

The required force and the entropy of the centric solution are

$$\lambda = (T^2-1)/2T^2, \quad S = \tfrac{1}{2}\log 2\pi + \log T - \tfrac{1}{2}(T^2-1). \tag{93}$$

Except for the constant part $\frac{1}{2}\log 2\pi$, these values are half as large as for the acentric reflections. The error potential estimate leads to a similar condition

$$\partial Y/\partial\lambda = 1/(1-2\lambda) + \lambda\sigma^2 - T^2 = 0. \tag{94}$$

The empirically adjusted force is

$$\lambda_{\rm centric} = \tfrac{1}{2}\frac{(T^2-1)}{(T^2+T^2_{\rm low})} \tag{95}$$

with the entropy

$$S_{\rm centric} = \tfrac{1}{2}\log(2\pi) + \tfrac{1}{2}\log\left(\frac{T^2+T^2_{\rm low}}{1+T^2_{\rm low}}\right) - \tfrac{1}{2}\left(\frac{T^2-1}{1+T^2_{\rm low}}\right). \tag{96}$$

Lastly, the pair-functional ensemble with the highest possible entropy is the featureless random distribution in which $T^2 = 1$ for every reflection. This defines a reference level $S_{\rm flat}$ for the entropy of each reflection, with the values $0, \frac{1}{2}\log 2\pi$ for acentric and centric reflections. The difference between $S$ and $S_{\rm flat}$ measures the degree of constraint imposed on the ensemble by the experimental data.

## APPENDIX *E*
## Cell with two types of atom

Here we derive the pair-functional ensemble for a cell that contains $N_A$ atoms of type $A$ and $N_B$ atoms of type $B$, with different scattering factors $f_a$ and $f_b$. The many-particle probability function is a function of the joint coordinates $(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B})$ and is normalized so that

$$A_N = \int f^{(N)}(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B})\,{\rm d}\mathbf{r}_A^{N_A}\,{\rm d}\mathbf{r}_B^{N_B} = 1. \tag{97}$$

The associated entropy is taken as

$$S_N = -\int f^{(N)}\log f^{(N)}\,{\rm d}\mathbf{r}_A^{N_A}\,{\rm d}\mathbf{r}_B^{N_B} - \log(N_A!N_B!). \tag{98}$$

We suppose that the average phased structure factors and originless intensities are to be matched for a certain set of reflections $\mathbf{H}$, with the target values

$$\langle F(\mathbf{H})\rangle = G(\mathbf{H}), \quad \langle I^{(2)}(\mathbf{H})\rangle = J(\mathbf{H}). \tag{99}$$

The structure factors and intensities are defined in terms of the Fourier transforms of the probability distributions $\eta_A(\mathbf{H})$ and $\eta_B(\mathbf{H})$ for the two types of particle.

$$F(\mathbf{H}) = f_a\eta_A(\mathbf{H}) + f_b\eta_B(\mathbf{H}) \tag{100}$$

$$I^{(2)}(\mathbf{H}) = |F(\mathbf{H})|^2 - \Sigma_I. \tag{101}$$

Here

$$\eta_A(\mathbf{H}) = \sum_{j=1}^{N_A}\exp(2\pi i\mathbf{H}\cdot\mathbf{r}_{jA}),$$
$$\eta_B(\mathbf{H}) = \sum_{j=1}^{N_B}\exp(2\pi i\mathbf{H}\cdot\mathbf{r}_{jB}) \tag{102}$$

and

$$\Sigma_I = N_A f_a^2 + N_B f_b^2. \quad (103)$$

The method of Lagrange multipliers gives the stationary condition

$$\delta S_N + \lambda \delta A_N + \sum_{\mathbf{H}} \hat{\chi}(-\mathbf{H}) \delta G(\mathbf{H}) + \tfrac{1}{2} \sum_{\mathbf{H}} \hat{\psi}(\mathbf{H}) \delta J(\mathbf{H}) = 0. \quad (104)$$

The many-particle equilibrium probability distribution for the system is

$$f^{(N)}(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B}) = (1/Z_N) \exp[\beta \Lambda_N(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B})], \quad (105)$$

where $\beta = 1$ and $\Lambda_N$ is the combined statistical potential

$$\Lambda_N = \sum_{\mathbf{H}} \hat{\chi}(-\mathbf{H})[f_a \eta_A(\mathbf{H}) + f_b \eta_B(\mathbf{H})] + \tfrac{1}{2} \sum_{\mathbf{H}} \hat{\psi}(\mathbf{H}) \{|f_a \eta_A(\mathbf{H}) + f_b \eta_B(\mathbf{H})|^2 - \Sigma_I\}. \quad (106)$$

The potential takes a rather simpler form in real space, in terms of the particle coordinates. Now the single-particle part of $\Lambda_N$ is

$$\mathbf{X}_N = \sum_{j=1}^{N_A} f_a \chi(\mathbf{r}_{jA}) + \sum_{j=1}^{N_B} f_b \chi(\mathbf{r}_{jB}) \quad (107)$$

and the paired particle terms contain $N(N-1)/2$ interactions between atoms of all types.

$$\Psi_N = \sum_{i<j} f_a^2 \psi(\mathbf{r}_{iA} - \mathbf{r}_{jA}) + \sum_{i<j} f_b^2 \psi(\mathbf{r}_{iB} - \mathbf{r}_{jB}) + \sum_{i,j} f_a f_b \psi(\mathbf{r}_{iA} - \mathbf{r}_{jB}). \quad (108)$$

The distributions of both types of atom are specified by the same field functions $\chi(\mathbf{r})$ and $\psi(\mathbf{u})$ but the fields act on the two types with different strengths. Thus, $f_a$ and $f_b$ behave like different effective atomic charges in a system of electrically charged particles.

Up to this point, the analysis is exact. It is rather more difficult to estimate the pair potentials, although in principle we could use the theory of the direct correlation function of a fluid mixture, which is expressed in the form of a matrix (Morita & Hiroike, 1961; Hansen & McDonald, 1986).

Here we use the simpler method of natural probability distributions to derive an approximation for the pair potential $\hat{\psi}(\mathbf{H})$ belonging to a reflection $\mathbf{H}$ for the simplest case where the single-particle electron-density variation $G(\mathbf{H})$ is unconstrained and the target intensity is

$$\langle |F(\mathbf{H})^2| \rangle = J(\mathbf{H}) + \Sigma_I = |F_T(\mathbf{H})|^2. \quad (109)$$

The natural probability distributions for the separate normalized structure factors $E_A(\mathbf{H})$ and $E_B(\mathbf{H})$ of the two atom types are independent two-dimensional Gaussians with variances of unity. The resulting combined natural distribution of $F(\mathbf{H}) = A_{\mathbf{H}} + iB_{\mathbf{H}}$ is a similar Gaussian

$$f_{\rm nat}(A_{\mathbf{H}}, B_{\mathbf{H}}) = (1/\pi\Sigma_I) \exp[-(A_{\mathbf{H}}^2 + B_{\mathbf{H}}^2)/\Sigma_I]. \quad (110)$$

The maximum-entropy ensemble includes a Boltzmann factor for the reflection $\mathbf{H}$ of the form

$$\exp \Psi = \exp\{\hat{\psi}(\mathbf{H})[|F(\mathbf{H})|^2 - \Sigma_I]\} \quad (111)$$

and so, following the standard natural distribution method, the perturbed distribution is

$$f(A_{\mathbf{H}}, B_{\mathbf{H}}) = [1/Z(\psi)](1/\pi\Sigma_I) \exp\{[-(A_{\mathbf{H}}^2 + B_{\mathbf{H}}^2)] \times [1/\Sigma_I - \hat{\psi}(\mathbf{H})]\}. \quad (112)$$

The value of $\hat{\psi}(\mathbf{H})$ must be chosen to match a standard Gaussian form, with the target variance

$$f(A_{\mathbf{H}}, B_{\mathbf{H}}) = [1/\pi|F_T(\mathbf{H})|^2] \exp[-(A_{\mathbf{H}}^2 + B_{\mathbf{H}}^2)/|F_T(\mathbf{H})|^2] \quad (113)$$

and this leads to the estimated potential

$$\hat{\psi}(\mathbf{H}) = \frac{|F_T(\mathbf{H})|^2 - \Sigma_I}{\Sigma_I |F_T(\mathbf{H})|^2}. \quad (114)$$

The Gaussian approximation for the distribution of $|F(\mathbf{H})|$ should be valid for a reasonably wide range of amplitudes, provided that both $N_A$ and $N_B$ are large.

I thank Ian McDonald and Robert Harris for discussions about the statistical theory of fluids.

## 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.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 47–71. Kluwer. Dordrecht, Netherlands.

Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.

Castleden, I. R. (1987). *Acta Cryst.* A**43**, 384–393.

Cramer, H. (1951). *Mathematical Methods of Statistics.* Princeton University Press.

De Dominicis, C. (1962). *J. Math. Phys.* **3**, 983–1002.

De Dominicis, C. (1963). *J. Math. Phys.* **4**, 255–265.

Giacovazzo, C. (1980). *Direct Methods in Crystallography*. London: Academic Press.

Gill, P. E., Murray, W. & Wright, M. H. (1981). *Practical Optimization.* London: Academic Press.

Hansen, J. P. & McDonald, I. R. (1986). *Theory of Simple Liquids*, 2nd ed. London: Academic Press.

Hauptman, H. A. (1972). *Crystal Structure Determination. The Role of the Cosine Seminvariants.* New York: Plenum Press.

Hauptman, H. (1975). *Acta Cryst.* A**31**, 671–679.

Hauptman, H. & Karle, J. (1953). *The Solution of the Phase Problem: I. The Centrosymmetric Crystal. American Crystallographic Association Monograph* No 3. Pittsburgh, PA: Polycrystal Book Service.

Hauptman, H. & Karle, J. (1956). *Acta Cryst.* **9**, 45–55.

Hill, T. L. (1956). *Statistical Mechanics.* New York: McGraw-Hill.

*International Tables for Crystallography* (1987). Vol. A. *Space Group Symmetry*, 2nd ed., edited by T. Hahn. Dordrecht: Kluwer.

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.

Klug, A. (1958). *Acta Cryst.* **11**, 515–543.

Levine, R. D. & Tribus, M. (1978). Editors. *The Maximum Entropy Formalism.* Cambridge, MA: MIT Press.
Lipson, H. & Cochran, W. (1966). *The Crystalline State*, Vol. III. *The Determination of Crystal Structures*. London: G. Bell and Sons.
Luenberger, D. G. (1984). *Linear and Nonlinear Programming*, 2nd ed. Reading, MA: Addison-Wesley.
McLachlan, A. D. (1999). *Acta Cryst.* A**55** Supplement, Abstract P12.BB.007.
McLachlan, A. D. (2001*a*). *Acta Cryst.* A**57**, 125–139.
McLachlan, A. D. (2001*b*). *Acta Cryst.* A**57**, 140–151.
McLachlan, A. D. & Harris, R. A. (1961). *J. Chem. Phys.*, **34**, 1451–1452.
Mayer, J. E. & Mayer, M. G. (1940). *Statistical Mechanics*. New York: John Wiley.
Morita, T. & Hiroike, K. (1961). *Prog. Theor. Phys. (Jpn)*, **25**, 537–578.
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.
Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005–1007.
Yvon, J. (1958). *Nuovo Cim. Suppl.* **9**, 144–151.