The Phase Problem of X-ray Crystallography

H.A. Hauptman

Hauptman-Woodward Medical Research Institute, Inc.
73 High Street
Buffalo, NY, USA
hauptman@hwi.buffalo.edu

ABSTRACT. The intensities of a sufficient number of X-ray diffraction maxima determine the structure of a crystal, that is, the positions of the atoms in the unit cell of the crystal. The available intensities usually exceed the number of parameters needed to describe the structure. From these intensities a set of numbers $|E_{\mathbf{H}}|$ can be derived, one corresponding to each intensity. However, the elucidation of the crystal structure also requires a knowledge of the complex numbers $E_{\mathbf{H}} = |E_{\mathbf{H}}| \exp(i\varphi_{\mathbf{H}})$, the normalized structure factors, of which only the magnitudes $|E_{\mathbf{H}}|$ can be determined from experiment. Thus, a "phase" $\varphi_{\mathbf{H}}$, unobtainable from the diffraction experiment, must be assigned to each $|E_{\mathbf{H}}|$, and the problem of determining the phases when only the magnitudes $|E_{\mathbf{H}}|$ are known is called the "phase problem". Owing to the known atomicity of crystal structures and the redundancy of observed magnitudes $|E_{\mathbf{H}}|$, the phase problem is solvable in principle.

Probabilistic methods have traditionally played a key role in the solution of this problem. They have led, in particular, to the so-called tangent formula which, in turn, has played the central role in the development of methods for the solution of the phase problem.

Finally, the phase problem may be formulated as one in constrained global optimization. A method for avoiding the countless local minima in order to arrive at the constrained global minimum leads to the *Shake-and-Bake* algorithm, a completely automatic solution of the phase problem for structures containing as many as 1000 atoms when data are available to atomic resolution.

In the case that single wavelength anomalous scattering (SAS) data are available, the probabilistic machinery leads to estimates of special linear combinations of the phases, the so-called structure invariants. A method of going from estimates of the structure invariants to the values of the individual phases is described.

1. Introduction

When a crystal is irradiated with a beam of X-rays the resulting interference effect gives rise to the so-called diffraction pattern which is uniquely determined by the crystal structure. Only the intensities of the scattered rays can be measured; the phases, which are also needed in order to work backwards, from diffraction pattern to the atomic positions, are lost in the diffraction experiment. However, owing to the known atomicity of real structures and the large number of observable intensities, the lost phase information is in fact contained in the measured intensities. The problem of recovering the missing phases, when only the intensities are available, is known as the phase problem. Alternatively, since the magnitudes $|E|$ of the normalized structure factors $E = |E|\exp(i\varphi)$ are readily determined from the measured diffraction intensities, the phase problem may be defined as the problem of determining the phases φ when the magnitudes $|E|$ are given.

Due to the redundancy of known magnitudes $|E|$, the phase problem is an over determined one and is therefore solvable in principle. This over determination implies the existence of relationships among the Es and, therefore, since the magnitudes $|E|$ are presumed to be known, the existence of identities among the phases φ alone, dependent on the known magnitudes $|E|$, which must of necessity be satisfied. The so-called direct methods are those which exploit these relationships in order to go directly from known magnitudes $|E|$ to desired phases φ. They do not depend on the presence of heavy atoms or atoms having other special scattering properties, for example anomalous scatterers, or prior structural knowledge.

The techniques of modern probability theory lead to the joint probability distributions of arbitrary collections of normalized structure factors from which the conditional probability distributions of selected sets of phases, given the values of suitably chosen magnitudes $|E|$, may be inferred. These distributions, dependent on known magnitudes $|E|$, constitute the foundation on which direct methods are based. They have provided the unifying thread from the beginning, circa 1950, until the present time. In particular, they have led to the recent formulation of the phase problem as one of constrained global optimization [1].

In the case that the structure consists of N identical atoms in the unit cell the relationship between diffraction pattern and crystal structure is given by the pair of equations

$$E_{\mathbf{H}} = |E_{\mathbf{H}}|\exp(i\varphi_{\mathbf{H}}) = N^{-1/2}\sum_{j=1}^{N}\exp(2\pi i\mathbf{H}\cdot r_j) \tag{1.1}$$

$$\langle E_{\mathbf{H}}\exp(-2\pi i\mathbf{H}\cdot r)\rangle_{\mathbf{H}} \approx \begin{cases} N^{-1/2} & \text{if } r = r_j \\ 0 & \text{if } r \neq r_j \end{cases} \tag{1.2}$$

where r_j is the position vector of the atom labeled j and $|E_{\mathbf{H}}|$ is obtained from the intensity of the scattered beam in the direction labeled by the reciprocal lattice vector $E_{\mathbf{H}}$. Clearly if one is to determine the crystal structure from Eq.(1.2) it is necessary to know not only the magnitudes $|E_{\mathbf{H}}|$, obtainable from the diffraction experiment, but also the phases $\varphi_{\mathbf{H}}$, lost in the diffraction experiment.

2. Method

2.1. *The Structure Invariants*

Equation (1.2) implies that the normalized structure factors $E_{\mathbf{H}}$ determine the crystal structure. However, Eq.(1.1) does not imply that, conversely, the crystal structure determines the values of the normalized structure factors $E_{\mathbf{H}}$ since the position vectors r_j depend not only on the structure but on the choice of origin as well. It turns out, nevertheless, that the magnitudes $|E_{\mathbf{H}}|$ of the normalized structure factors are in fact uniquely determined by the crystal structure and are independent of the choice of origin, but that the values of the phases $\varphi_{\mathbf{H}}$ depend also on the choice of origin. Although the values of the individual phases depend on the structure and the choice of origin, there exist certain linear combinations of the phases, the so-called structure invariants, whose values are determined by the structure alone and are independent of the choice of origin. The most important class of structure invariants, and the only one to be considered here, consists of the three-phase structure invariants (triplets),

$$\varphi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \tag{2.1}$$

where $\mathbf{H}$ and $\mathbf{K}$ are arbitrary reciprocal lattice vectors.

2.2. *The Probabilistic Background*

It is assumed that the atomic position vectors r_j are the primitive random variables, uniformly and independently distributed in the unit cell. Then the normalized structure factors $E_{\mathbf{H}}$, as functions of the r_j's, are themselves random variables. The structure invariants $\varphi_{\mathbf{HK}}$ in turn, as functions of the individual phases φ (Eq.(2.1)), are therefore also random variables.

2.3. *The Conditional Probability Distribution of $\varphi_{\mathbf{HK}}$, Given $|E_{\mathbf{H}}|$,$|E_{\mathbf{K}}|$,$|E_{\mathbf{H+K}}|$*

Under the conditions set forth in § 2.2 the conditional probability distribution of the triplet $\varphi_{\mathbf{HK}}$ (Eq.(2.1)), given the presumed known values of $|E_{\mathbf{H}}|$,$|E_{\mathbf{K}}|$,$|E_{\mathbf{H+K}}|$, is known to be

$$P(\Phi|A_{\mathbf{HK}}) = [2\pi I_0(A_{\mathbf{HK}})]^{-1}\exp(A_{\mathbf{HK}}\cos\Phi) \tag{2.2}$$

where Φ represents the triplets $\varphi_{\mathbf{HK}}$, the parameter $A_{\mathbf{HK}}$ is defined by

$$A_{\mathbf{HK}} = 2N^{-1/2}|E_{\mathbf{H}}E_{\mathbf{K}}E_{\mathbf{H+K}}|, \tag{2.3}$$

and I_0 is the Modified Bessel Function. Equation (2.3) implies that the mode of $\varphi_{\mathbf{HK}}$ is zero, and the conditional expected value (or average) of $\cos\varphi_{\mathbf{HK}}$, given $A_{\mathbf{HK}}$, is

$$\varepsilon(\cos\varphi_{\mathbf{HK}}|A_{\mathbf{HK}}) = I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}}) > 0, \tag{2.4}$$

where I_1 is the Modified Bessel Function. It is also readily confirmed that the larger the value of $A_{\mathbf{HK}}$ the smaller is the conditional variance of $\cos\varphi_{\mathbf{HK}}$, given $A_{\mathbf{HK}}$. It is to be stressed that the conditional expected value of the cosine, Eq.(2.4), is always positive since $A_{\mathbf{HK}} > 0$.

2.4. *The Minimal Principle*

In view of Eq.(2.4) one obtains the following estimate of $\cos\varphi_{\mathbf{HK}}$:

$$\cos\varphi_{\mathbf{HK}} \approx I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}}) \tag{2.5}$$

and expects that the smaller the variance, that is the larger $A_{\mathbf{HK}}$, the more reliable this estimate will be. Hence one is led to construct the function (the so-called minimal function), determined by the known magnitudes $|E|$,

$$R = R(\varphi) = \frac{1}{\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}} \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \Big(\cos\varphi_{\mathbf{HK}} - \frac{I_1(A_{\mathbf{HK}})}{I_0(A_{\mathbf{HK}})} \Big)^2 \tag{2.6}$$

which, in view of Eq.(2.1), is seen to be a function of phases φ alone. Equation (2.4) then implies that the global minimum of the minimal function $R(\varphi)$, where the phases are constrained to satisfy the identities known to exist (§ 1), yields the desired phases (the minimal principle). Thus the phase problem is formulated as a problem in constrained global minimization [2]. There remains only the problem of avoiding the myriad local minima of $R(\varphi)$ in order to arrive at the constrained global minimum. The next section shows how this minimum is reached *via* the computer program *Shake-and-Bake*.

2.5. *The Computer Program "Shake-and-Bake"*

[3]

The six-part *Shake-and-Bake* phase determination procedure, shown by the flow diagram in Figure 1, combines minimal-function phase refinement and real-space filtering. It is an iterative process that is repeated until a solution is achieved or a designated number of cycles have been performed. With reference to Figure 1, the major steps of the algorithm are described next.

2.5.1. *Generate invariants.* Normalized structure-factor magnitudes ($|E|$'s) are generated by standard scaling methods and the triplet invariants that involve the largest corresponding $|E|$'s are generated. Parameter choices that must be made at this stage include the numbers of phases and triplets to be used. The total number of invariants is ordinarily chosen to be at least 100 times the number of atoms whose positions are to be determined.

2.5.2. *Generate trial structure.* A trial structure or model is generated that is comprised of a number of randomly positioned atoms equal to the number of atoms in the unit cell. The starting coordinate sets are subject to the restrictions that no two atoms are closer than a specified distance (normally 1.2$\mathring{A}$) and that no atom is within bonding distance of more than four atoms.

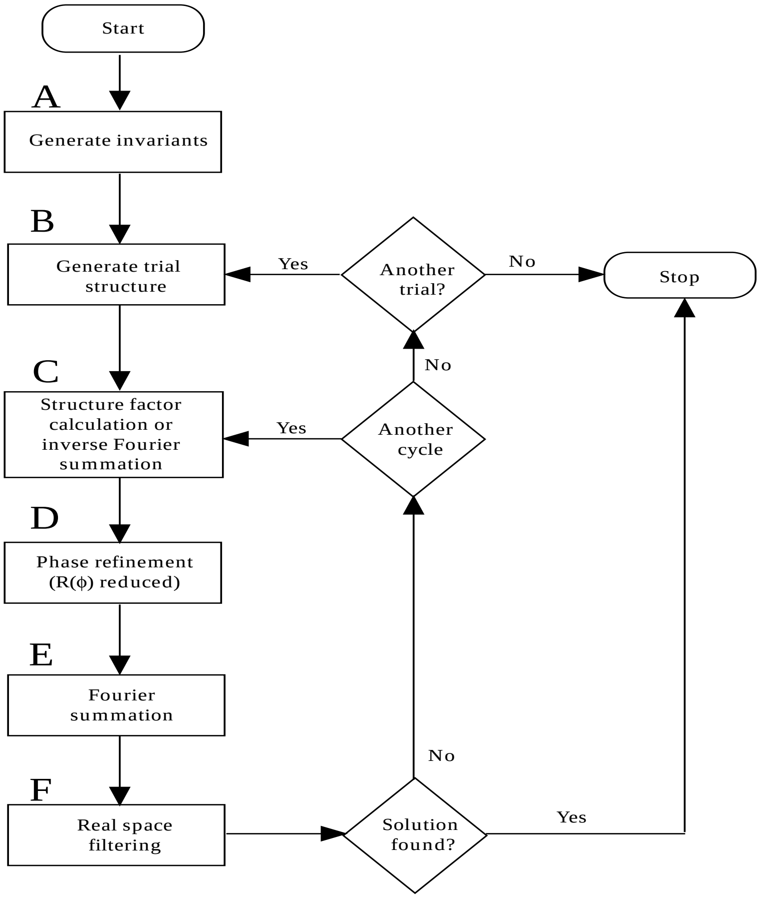

FIGURE 1. Flow chart for *Shake-and-Bake*, the minimal-function phase refinement and real-space filtering procedure.

2.5.3. *Structure-factor calculation.* A normalized structure-factor calculation (see Eq.(1.1)) based on the trial coordinates is used to compute initial values for all the desired phases simultaneously. In subsequent cycles, peaks selected from the most recent Fourier series are used as atoms to generate new phase values.

2.5.4. *Phase refinement.* The values of the phases are perturbed by a *parameter-shift* method in which $R(\varphi)$, which measures the mean-square difference between estimated and calculated structure invariants, is reduced in value. $R(\varphi)$ is initially computed on the basis of the set of phase values obtained from the structure-factor calculation in step C. The phase set is ordered in decreasing magnitude of the associated $|E|$'s. The value of the first phase is incremented by a preset amount and $R(\varphi)$ is recalculated. If the new calculated value of $R(\varphi)$ is lower than the previous one, the value of the first phase is incremented again by the preset amount. This is continued until $R(\varphi)$ no longer decreases or until a predetermined number of increments has been applied to the first phase. A completely analogous course is taken if, on the initial incrementation, $R(\varphi)$ increases, except that the value of the first phase is decremented until $R(\varphi)$ no longer decreases or until the predetermined number of decrements has been applied. The remaining phase values are varied in sequence as just described. Note that, when the ith phase value is varied, the new values determined for the previous $i-1$ phases are used immediately in the calculation of $R(\varphi)$. The step size and number of steps are variables whose values must be chosen.

2.5.5. *Fourier summation.* Fourier summation is used to transform phase information into an electron-density map (Refer to Eq.(1.2)).

2.5.6. *Real-space filtering (Identities among phases imposed).* Image enhancement is accomplished by a discrete electron-density modification consisting of the selection of a specified number of the largest peaks on the Fourier map for use in the next structure-factor calculation. The simple choice, in each cycle, of a number of the largest peaks corresponding to the number of expected atoms has given satisfactory results. No minimum-interpeak-distance criterion is applied at this stage.

Steps C, D, E, and F are repeated until a pre-assigned number of cycles has been completed or until the process converges. The smallest of the final values of the minimal function (one for each trial) reveals the constrained global minimum of the minimal function $R(\varphi)$ and the true values of the phases.

3. Single Wavelength Anomalous Scattering (SAS) Data are Available

3.1. *Introduction*

In this case the normalized structure factors $E_{\mathbf{H}}$ (compare Eq.(1.1)) are defined by

$$E_{\mathbf{H}} = \frac{1}{\alpha_2} \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{H} \cdot r_j) \tag{3.1}$$

$$\alpha_2 = \sum_{j=1}^{N} |f_j|^2 \tag{3.2}$$

where N is the number of atoms in the unit cell, r_j is the position vector of the atom labeled j and f_j is the (complex-valued) atomic scattering factor, presumed to be known.

3.2. *The Probabilistic Background*

With the assumption that SAS diffraction data are available, the conditional probability distribution $P(\Phi)$ of the triplet

$$\varphi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \tag{3.3}$$

given the six magnitudes

$$|E_{\mathbf{H}}|, |E_{-\mathbf{H}}|, |E_{\mathbf{K}}|, |E_{-\mathbf{K}}|, |E_{\mathbf{H}+\mathbf{K}}|, |E_{-\mathbf{H}-\mathbf{K}}| \tag{3.4}$$

is known to be [4]

$$P(\Phi) = [2\pi I_0(A_{\mathbf{HK}})]^{-1} \exp\{A_{\mathbf{HK}} \cos(\Phi - \omega_{\mathbf{HK}})\} \tag{3.5}$$

in which I_0 is the Modified Bessel Function and $A_{\mathbf{HK}}$ and $\omega_{\mathbf{HK}}$ are expressed in terms of the six magnitudes (3.4) and the (presumed known) complex-valued atomic scattering factors f. Compare Eq.(3.5) with Eq.(2.2) and note that the $A_{\mathbf{HK}}$ of Eq.(3.5) is no longer defined by Eq.(2.3) but is instead a much more complicated function of the six magnitudes (3.4) and the atomic scattering factors f. Hence, $A_{\mathbf{HK}}(> 0)$ and $\omega_{\mathbf{HK}}$ are here assumed to be known for every pair $(\mathbf{H}, \mathbf{K})$. Note that, owing to the anomalous scattering, the six magnitudes (3.4) are, in general, distinct in contrast to the normal case when $|E_{-\mathbf{H}}| = |E_{\mathbf{H}}|$, etc.

In view of (3.5), the most probable value of $\varphi_{\mathbf{HK}}$ is $\omega_{\mathbf{HK}}$, and the larger the value of $A_{\mathbf{HK}}$ the better is this estimate of $\varphi_{\mathbf{HK}}$:

$$\varphi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} \approx \omega_{\mathbf{HK}} \tag{3.6}$$

3.3. *The System of SAS Tangent Equations*

Fix the reciprocal lattice vector $\mathbf{H}$. From Eq.(3.6)

$$\varphi_{\mathbf{H}} \approx \omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}} \tag{3.7}$$

$$\sin \varphi_{\mathbf{H}} \approx \sin(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}}) \tag{3.8}$$

which has approximate validity for each fixed value of $\mathbf{K}$. Averaging the right-hand side of (3.8) over $\mathbf{K}$, naturally using weights $A_{\mathbf{HK}}$, one obtains

$$\sin \varphi_{\mathbf{H}} \approx \frac{1}{\sum_{\mathbf{K}} A_{\mathbf{HK}}} \sum_{\mathbf{K}} A_{\mathbf{HK}} \sin(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}}) \tag{3.9}$$

Similarly

$$\cos \varphi_{\mathbf{H}} \approx \frac{1}{\sum_{\mathbf{K}} A_{\mathbf{HK}}} \sum_{\mathbf{K}} A_{\mathbf{HK}} \cos(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}}) \tag{3.10}$$

Eqs.(3.9) and (3.10) imply

$$\tan \varphi_{\mathbf{H}} \approx \frac{\sum_{\mathbf{K}} A_{\mathbf{HK}} \sin(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} A_{\mathbf{HK}} \cos(\omega_{\mathbf{HK}} - \varphi_{\mathbf{K}} - \varphi_{-\mathbf{H}-\mathbf{K}})} \tag{3.11}$$

the system of SAS tangent equations. For each fixed $\mathbf{H}$, Eq.(3.11) yields two values for $\varphi_{\mathbf{H}}$ differing by π. Eqs.(3.9) and (3.10) serve to fix the quadrant.

3.4. *The Maximal Function* $M(\varphi)$

One defines the maximal function $M(\varphi)$, a function of the phases φ, by means of

$$M(\varphi) = \frac{1}{\sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}} \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \cos(\varphi_{\mathbf{HK}} - \omega_{\mathbf{HK}}) \tag{3.12}$$

and infers that $M(\varphi)$ has a global maximum when all the phases appearing in Eq.(3.12) are set equal to their true values.

3.5. *The Maximal Property of the System of SAS Tangent Equations*

Fundamental maximal property. Fix $\mathbf{H}$. Assume that the values of all phases other than $\varphi_{\mathbf{H}}$ are specified arbitrarily. Then the maximal function $M(\varphi)$ becomes a function, $M(\varphi_{\mathbf{H}}|\varphi)$, of the single phase $\varphi_{\mathbf{H}}$. As a function of $\varphi_{\mathbf{H}}$, $M(\varphi_{\mathbf{H}}|\varphi)$ has a unique maximum in the whole interval $(0, 2\pi)$ and the value of $\varphi_{\mathbf{H}}$ that maximizes $M(\varphi_{\mathbf{H}}|\varphi)$ is given by the SAS tangent equation (3.11).

3.6. *Solving the System of SAS Tangent Equations*

Specify arbitrarily initial values for all the phases φ. Fix $\mathbf{H}$. Calculate a new value for the phase $\varphi_{\mathbf{H}}$ by means of the SAS tangent equations (3.9) to (3.11), in this way, in view of § 3.5, increasing the initial value of the maximal function $M(\varphi)$. Fix $\mathbf{H}' \neq \mathbf{H}$. Calculate a new value for $\varphi_{\mathbf{H}'}$, again using (3.9) to (3.11), the new value for $\varphi_{\mathbf{H}}$, and initial values for the remaining phases, thus increasing still further the value of $M(\varphi)$. Continue in this way to obtain new values for all the phases, thus completing the first iteration and, in the process, continuously increasing the value of $M(\varphi)$. Complete as many iterations as necessary in order to secure convergence. Convergence is assured since the iterative process yields a monotonically increasing sequence of numbers, the values of $M(\varphi)$, bounded above by unity. Evidently also, the process leads to a local maximum of $M(\varphi)$ and a corresponding set of values for all the phases φ which depends on the values of the phases chosen initially. By choosing different starting values for the phases one obtains different solutions for the system of SAS tangent equations and different local maxima of $M(\varphi)$. That solution yielding the global maximum of $M(\varphi)$ is the one we seek.

3.7. *The Linear Congruence Connection*

The problem of going from the estimated values $\omega_{\mathbf{HK}}$ of the three-phase structure invariants $\varphi_{\mathbf{HK}}$ to the values of the individual phases φ may be formulated as the problem of solving the redundant system of linear congruences

$$\varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}} \equiv \omega_{\mathbf{HK}} \text{ (modulo } 2\pi) \tag{3.13}$$

each with weight $A_{\mathbf{HK}}$. Our solution of the system of SAS tangent equations also yields the solution of the redundant system of linear congruences (3.13).

This research is supported by National Institutes of Health Grant No. GM-46733.

References

[1] Hauptman, H.A. (1991). *Crystallographic Computing 5: From Chemistry to Biology*, edited by D. Moras, A.D. Podnarny & J.C. Thierry, 324–332. IUCr/Oxford University Press.

[2] DeTitta, G.T., Weeks, C.M., Thuman, P., Miller, R. & Hauptman, H.A. (1994). *Structure Solution by Minimal Function Phase Refinement and Fourier Filtering: I. Theoretical Basis*. Acta Cryst. **A50**, 203–210.

[3] Weeks, C.M., DeTitta, G.T., Hauptman, H.A., Thuman, P. & Miller, R. (1994). *Structure Solution by Minimal Function Phase Refinement and Fourier Filtering: II. Implementation and Applications*. Acta Cryst. **A50**, 210–220.

[4] Hauptman, H. A. (1982) *On Integrating the Techniques of Direct Methods with Anomalous Dispersion: I. The Theoretical Basis*. Acta Cryst. **A38**, 632–641.