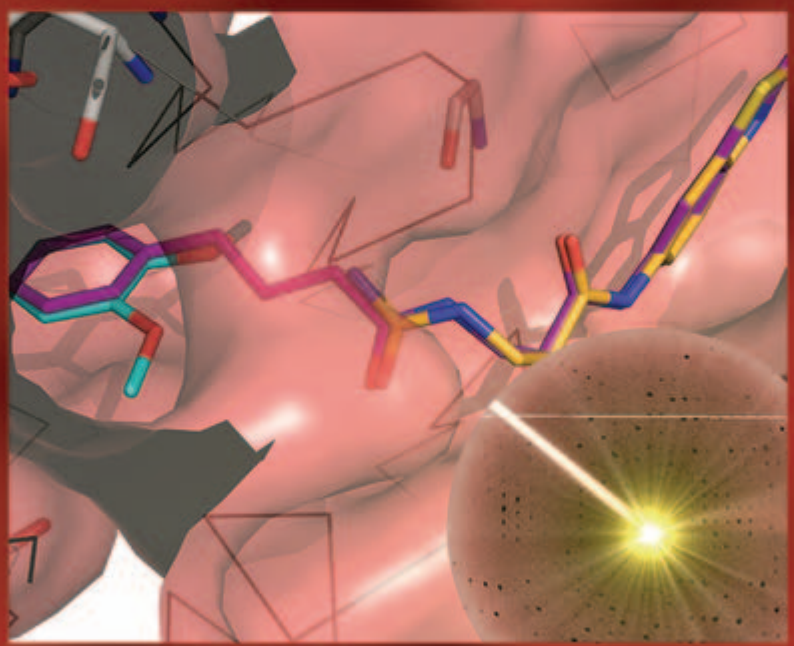
IUCr Texts on Crystallography · 20

Phasing in Crystallography

A Modern Perspective

CARMELO GIACOVAZZO



INTERNATIONAL UNION OF CRYSTALLOGRAPHY

OXFORD SCIENCE PUBLICATIONS

INTERNATIONAL UNION OF CRYSTALLOGRAPHY
BOOK SERIES

IUCr Monographs on Crystallography

1 *Accurate molecular structures*
A. Domenicano, I. Hargittai, editors
2 *P.P. Ewald and his dynamical theory of X-ray diffraction*
D.W.J. Cruickshank, H.J. Juretschke, N. Kato, editors
3 *Electron diffraction techniques*, *Vol. 1*
J.M. Cowley, editor
4 *Electron diffraction techniques*, *Vol. 2*
J.M. Cowley, editor
5 *The Rietveld method*
R.A. Young, editor
6 *Introduction to crystallographic statistics*
U. Shmueli, G.H. Weiss
7 *Crystallographic instrumentation*
L.A. Aslanov, G.V. Fetisov, J.A.K. Howard
8 *Direct phasing in crystallography*
C. Giacovazzo
9 *The weak hydrogen bond*
G.R. Desiraju, T. Steiner
10 *Defect and microstructure analysis by diffraction*
R.L. Snyder, J. Fiala, H.J. Bunge
11 *Dynamical theory of X-ray diffraction*
A. Authier
12 *The chemical bond in inorganic chemistry*
I.D. Brown
13 *Structure determination from powder diffraction data*
W.I.F. David, K. Shankland, L.B. McCusker, Ch. Baerlocher, editors
14 *Polymorphism in molecular crystals*
J. Bernstein
15 *Crystallography of modular materials*
G. Ferraris, E. Makovicky, S. Merlino
16 *Diffuse X-ray scattering and models of disorder*
T.R. Welberry
17 *Crystallography of the polymethylene chain*: *an inquiry into the structure of waxes*
D.L. Dorset

18 *Crystalline molecular complexes and compounds*: *structure and principles*
F.H. Herbstein
19 *Molecular aggregation*: *structure analysis and molecular simulation of crystals and liquids*
A. Gavezzotti
20 *Aperiodic crystals*: *from modulated phases to quasicrystals*
T. Janssen, G. Chapuis, M. de Boissieu
21 *Incommensurate crystallography*
S. van Smaalen
22 *Structural crystallography of inorganic oxysalts*
S.V. Krivovichev
23 *The nature of the hydrogen bond*: *outline of a comprehensive hydrogen bond theory*
G. Gilli, P. Gilli
24 *Macromolecular crystallization and crystal perfection*
N.E. Chayen, J.R. Helliwell, E.H. Snell
25 *Neutron protein crystallography*: *hydrogen*, *protons*, *and hydration in bio-macromolecules*
N. Niimura, A. Podjarny

IUCr Texts on Crystallography

1 *The solid state*
A. Guinier, R. Julien
4 *X-ray charge densities and chemical bonding*
P. Coppens
8 *Crystal structure refinement*: *a crystallographer's guide to SHELXL*
P. Müller, editor
9 *Theories and techniques of crystal structure determination*
U. Shmueli
10 *Advanced structural inorganic chemistry*
Wai-Kee Li, Gong-Du Zhou, Thomas Mak
11 *Diffuse scattering and defect structure simulations*: *a cook book using the program DISCUS*
R.B. Neder, T. Proffen
12 *The basics of crystallography and diffraction*, *third edition*
C. Hammond
13 *Crystal structure analysis*: *principles and practice*, *second edition*
W. Clegg, editor
14 *Crystal structure analysis*: *a primer, third edition*
J.P. Glusker, K.N. Trueblood
15 *Fundamentals of crystallography*, *third edition*
C. Giacovazzo, editor
16 *Electron crystallography*: *electron microscopy and electron diffraction*
X. Zou, S. Hovmöller, P. Oleynikov
17 *Symmetry in crystallography*: *understanding the International Tables*
P.G. Radaelli
18 *Symmetry relationships between crystal structures*: *applications of crystallographic group theory in crystal chemistry*
U. Müller
19 *Small angle X-ray and neutron scattering from biomacromolecular solutions*
D.I. Svergun, M.H.J. Koch, P.A. Timmins, R.P. May
20 *Phasing in crystallography*: *a modern perspective*
C. Giacovazzo

Phasing in Crystallography

A Modern Perspective

CARMELO GIACOVAZZO

Professor of Crystallography,
University of Bari, Italy
Institute of Crystallography, CNR, Bari, Italy

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2014
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2013943731

ISBN 978–0–19–968699–5

Printed in Great Britain by
Clays Ltd, St Ives plc

Dedication

To my mother,
to my wife Angela,
my sons Giuseppe and Stefania,
to my grandchildren Agostino, Stefano and Andrea Morris

Acknowledgements

I acknowledge the following colleagues and friends for their generous help:

Caterina Chiarella, for general secretarial management of the book and for her assistance with the drawings;

Angela Altomare, Benedetta Carrozzini, Corrado Cuocci, Giovanni Luca Cascarano, Annamaria Mazzone, Anna Grazia Moliterni, and Rosanna Rizzi for their kind support, helpful discussions, and critical reading of the manuscript. Corrado Cuocci also took care of the cover figure.

Facilities provided by the Istituto di Cristallografia, CNR, Bari, are gratefully acknowledged.

Preface

A short analysis of the historical evolution of phasing methods may be a useful introduction to this book because it will allow us to better understand efforts and results, the birth and death of scientific paradigms, and it will also explain the general organization of this volume. This analysis is very personal, and arises through the author's direct interactions with colleagues active in the field; readers interested in such aspects may find a more extensive exposition in *Rend. Fis. Acc. Lincei* (2013), **24**(1), pp. 71–76.

In a historical sense, crystallographic phasing methods may be subdivided into two main streams: the small and medium-sized molecule stream, and the macro-molecule stream; these were substantially independent from each other up until the 1990s. Let us briefly consider their achievements and the results of their subsequent confluence.

Small and medium-sized molecule stream

The *Patterson* (1934) *function* was the first general phasing tool, particularly effective for heavy-atom structures (e.g. this property met the requirements of the earth sciences, the first users of early crystallography). Even though subsequently computerized, it was soon relegated to a niche by *direct methods*, since these were also able to solve light-atom structures (a relevant property towards the development of organic chemistry).

Direct methods were introduced, in their modern probabilistic guise, by Hauptman and Karle (1953) and Cochran (1955); corresponding phasing procedures were automated by Woolfson and co-workers, making the crystal structure solution of small molecules more straightforward. Efforts were carried out exclusively in reciprocal space (*first paradigm of direct methods*); the paradigm was systematized by the neighbourhood (Hauptman, 1975) and representation theories (Giacovazzo, 1977, 1980). Structures up to 150 non-hydrogen (non-H) atoms in the asymmetric unit were routinely able to be solved.

The complete success of this stream may be deduced from the huge numbers of structures deposited in appropriate data banks. Consequently, western national research agencies no longer supported any further research in the small to medium-sized molecule area (the work was done!); research groups working on methods moved instead to powder crystallography, electron crystallography, or to proteins, all areas of technological interest for which phasing was still a challenge. Direct space approaches were soon developed, which enhanced our capacity to solve structures, even from low quality diffraction data.

The macromolecule stream

Since the 1950s, efforts were confined to *isomorphous replacement* (SIR, MIR; Green et al., 1954), *molecular replacement* (MR; Rossmann and Blow, 1962), and *anomalous dispersion techniques* (SAD-MAD; Okaya and Pepinsky, 1956; Hoppe and Jakubowski, 1975). *Ab initio approaches*, the main techniques of interest for the small and medium-sized molecule streams, were neglected as being unrealistic; indeed, they are less demanding in terms of prior information but are very demanding in terms of data resolution.

The popularity of protein phasing techniques changed dramatically over the years. At the very beginning, SIR-MIR was the most popular method, but soon MR started to play a more major role as good structural models became progressively more readily available. About 75% of structures today are solved using MR. The simultaneous technological progress in synchrotron radiation and its wide availability have increased the appeal of SAD-MAD techniques.

The achievements obtained within the macromolecular stream have been impressive. A huge number of protein structures has been deposited in the Protein Data Bank, and the solution of protein structures is no longer confined to just an elite group of scientists, it is performed in many laboratories spread over four continents, often by young scientists. Crucial to this has been the role of the *CCP4* project, for the coordination of new methods and new computer programs.

The synergy of the two streams

It is the opinion of the author that synergy between the two streams originated due to a common interest in *EDM* (*electron density modification*) techniques. This approach, first proposed by Hoppe and Gassman (1968) for small molecules, was later extensively modified to be useful for both streams. Confluence of the two streams began in the 1990s (even if contacts were begun in the 1980s), when EDM techniques were used to improve the efficiency of direct methods. That was the beautiful innovation of s*hake and bake* (Weeks et al., 1994); both direct and reciprocal space were explored to increase phasing efficiency (this was the *second paradigm of direct methods*). It was soon possible to solve ab initio structures with up to 2000 non-hydrogen atoms in the asymmetric unit, provided data at atomic or quasi-atomic resolution are available. As a consequence, the ab initio approach for proteins started to attract greater attention. A secondary effect of the EDM procedures was the recent discovery of new ab initio techniques, such as *charge flipping* and *VLD* (*vive la difference*), and the newly formulated Patterson techniques.

The real revolution in the macromolecular area occurred when probabilistic methods, already widely used in small and medium-sized molecules, erupted into the protein field. Joint probability distributions and maximum likelihood approaches were tailored to deal with large structures, imperfect isomorphism, and errors in experimental data; and they were applied to SAD-MAD, MR, and SIR-MIR cases. For example, protein substructures with around 200 atoms in the asymmetric unit, an impossible challenge for traditional techniques, could easily be solved by the new approaches.

High-throughput crystallography is now a reality: protein structures, 50 years ago solvable only over months or years, can now be solved in hours or days; also due to technological advances in computer sciences.

The above considerations have been the basic reason for reconsidering the material and the general guidelines given in my textbook *Direct Phasing in Crystallography*, originally published in 1998. This was essentially a description of the mathematical bases of direct methods and of their historical evolution, with some references to applicative aspects and ancillary techniques.

The above described explosion in new phasing techniques and the improved efficiency of the revisited old methods made impellent the need for a new textbook, mainly addressing the phasing approaches which are alive today, that is those which are applicable to today's routine work. On the other hand, the wide variety of new methods and their intricate relationship with the old methods requires a new rational classification: methods similar regarding the type of prior information exploited, mathematical technique, or simply their mission, are didactically correlated, in such a way as to offer an organized overview of the current and of the old approaches. This is the main aim of this volume, which should not therefore simply be considered as the second edition of *Direct Phasing in Crystallography*, but as a new book with different guidelines, different treated material, and a different purpose.

Attention will be focused on both the theoretical and the applicative aspects, in order to provide a friendly companion for our daily work. To emphasize the new design the title has been changed to *Phasing in Crystallography,* with the subtitle, *A Modern Perspective*. In order to make the volume more useful, historical developments of phasing approaches that are not in use today, are simply skipped, and readers interested in these are referred to *Direct Phasing in Crystallography*.

This volume also aims at being a tool to inspire new approaches. On the one hand, we have tried to give, in the main text, descriptions of the various methods that are as simple as possible, so that undergraduate and graduate students may understand their general purpose and their applicative aspects. On the other hand, we did not shrink from providing the interested reader with mathematical details and/or demonstrations (these are necessary for any book dealing specifically with methods). These are confined in suitable appendices to the various chapters, and aimed at the trained crystallographer. At the end of the book, we have collected together mathematical appendices of a general character, appendices denoted by the letter M for mathematics and devoted to the bases of the methods (e.g. probability theory, basic crystallography, concepts of analysis and linear algebra, specific mathematical techniques, etc.), thus offering material of interest for professional crystallographers.

A necessary condition for an understanding of the content of the book is a knowledge of the fundamentals of crystallography. Thus, in Chapter 1 we have synthesized the essential elements of the general crystallography and we have also formulated the *basic postulate of structural crystallography*; the entire book is based on its validity.

In Chapter 2, the statistics of structure factors is described simply: it will be the elementary basis of most of the methods described throughout the volume.

Chapter 3 is a simplified description of the concepts of structure invariant and seminvariant, and of the related origin problem.

In Chapter 4, we have synthesized the methods of joint probability distributions and neighbourhoods–representation theories. The application of these methods to three-phase and four-phase structure invariants are described in Chapter 5. The probabilistic estimation of structure seminvariants has been skipped owing to their marginal role in modern phasing techniques. In Chapter 6, we discuss direct methods and the most traditional phasing approaches.

Chapter 7 is dedicated to joint probability distribution functions when a model is available, with specific attention to two- and to three-phase invariants. The most popular Fourier syntheses are described in the same chapter and their potential discussed in relation with the above probability distributions.

Chapter 8 is dedicated to phase improvement and extension via electron density modification techniques, Chapter 9 to two new phasing approaches, *charge flipping* and *VLD* (*vive la difference*), and Chapter 10, to *Patterson techniques*. Their recent revision has made them one of the most powerful techniques for ab initio phasing and particularly useful for proteins.

X-rays are not always the most suitable radiation for performing a diffraction experiment. Indeed, neutron diffraction may provide information complementary to that provided by X-ray data, electron diffraction becoming necessary when only nanocrystals are available. In Chapter 11 phasing procedures useful for this new scenario are described.

Often single crystals of sufficient size and quality are not available, but microcrystals can be grown. In this case powder data are collected; diffraction techniques imply a loss of experimental information, and therefore phasing via such data requires significant modifications to the standard methods. These are described in Chapter 12.

Chapters 13 to 15 are dedicated to the most effective and popular methods used in macromolecular crystallography: the non-ab initio methods, *Molecular Replacement* (*MR*), *Isomorphous Replacement* (*SIR-MIR*), and *Anomalous Dispersion* (*SAD-MAD*) techniques.

The reader should not think that the book has been partitioned into two parts, the first devoted to small and medium-sized molecules, the second to macromolecules. Indeed in the first twelve chapters, most of the mathematical tools necessary to face the challenges of macromolecular crystallography are described, together with the main algorithms used in this area and the fundamentals of the probabilistic approaches employed in macromolecular phasing. This design allows us to provide, in the last three chapters, simpler descriptions of MR, SIR-MIR, and SAD-MAD approaches.

Contents

Symbols and notation

The following symbols and conventions will be used throughout the full text.

The **bold** character is used for denoting vectors and matrices.

$\mathbf{h}\cdot\mathbf{r}$	the dot indicates the scalar product of the two vectors **h** and **r**
$\mathbf{a}\wedge\mathbf{b}$	cross-product of the two vectors **a** and **b**
$\bar{\mathbf{A}}$	the bar indicates the transpose of the matrix **A**
s.f.	structure factor
n.s.f.	normalized structure factor
s.i.	structure invariant
s.s.	structure seminvariant
cs.	centrosymmetric
n.cs.	non-centrosymmetric
RES	experimental data resolution (in Å)
CORR	correlation between the electron density map of the target structure (the one we want to solve) and that of a model map
	$R_{cryst} = \frac{\sum_{\mathbf{h}} \lvert\lvert F_{obs}\rvert - \lvert F_{calc}\rvert\rvert}{\sum_{\mathbf{h}} \lvert F_{obs}\rvert}$ crystallographic residual
SIR-MIR	single–multiple isomorphous replacement
SAD-MAD	single–multiple anomalous dispersion
MR	molecular replacement

Fundamentals of crystallography

1

1.1 Introduction

In this chapter we summarize the basic concepts, formulas and tables which constitute the essence of general crystallography. In Sections 1.2 to 1.5 we recall, without examples, definitions for unit cells, lattices, crystals, space groups, diffraction conditions, etc. and their main properties: reading these may constitute a useful reminder and support for daily work. In Section 1.6 we establish and discuss the *basic postulate of structural crystallography*: this was never formulated, but during any practical phasing process it is simply assumed to be true by default. We will also consider the consequences of such a postulate and the caution necessary in its use.

1.2 Crystals and crystallographic symmetry in direct space

We recall the main concepts and definitions concerning crystals and crystallographic symmetry.

Crystal. This is the periodic repetition of a motif (e.g. a collection of molecules, see Fig. 1.1). An equivalent mathematical definition is: the crystal is the convolution between a lattice and the unit cell content (for this definition see (1.4) below in this section).

Unit cell. This is the parallelepiped containing the motif periodically repeated in the crystal. It is defined by the unit vectors **a**, **b**, **c**, or, by the six scalar parameters $a, b, c, \alpha, \beta, \gamma$ (see Fig. 1.1). The generic point into the unit cell is defined by the vector

$$\mathbf{r} = x\,\mathbf{a} + y\,\mathbf{b} + z\,\mathbf{c},$$

where x, y, z are fractional coordinates (dimensionless and lying between 0 and 1). The volume of the unit cell is given by (see Fig. 1.2)

$$V = \mathbf{a} \wedge \mathbf{b} \cdot \mathbf{c} = \mathbf{b} \wedge \mathbf{c} \cdot \mathbf{a} = \mathbf{c} \wedge \mathbf{a} \cdot \mathbf{b}. \tag{1.1}$$

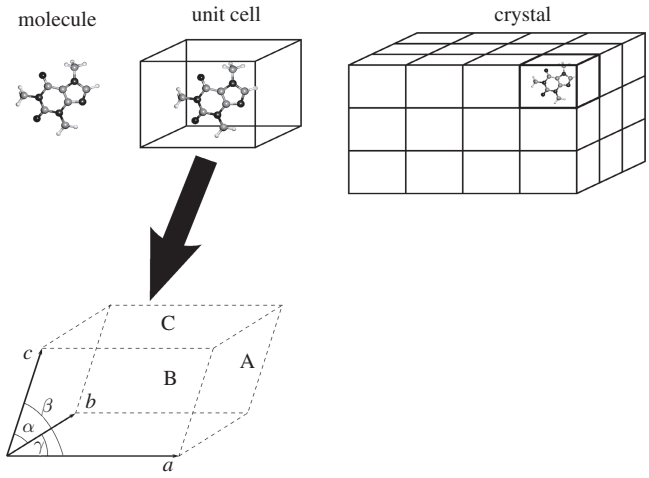


Fig. 1.1
The motif, the unit cell, the crystal.

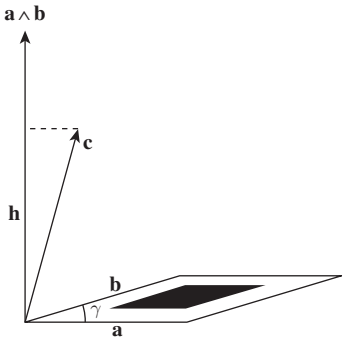


Fig. 1.2
The vector $\mathbf{a} \wedge \mathbf{b}$ is perpendicular to the plane $(\mathbf{a}, \mathbf{b})$: its modulus $|ab \sin\gamma|$ is equal to the shaded area on the base. The volume of the unit cell is the product of the base area and **h**, the projection of **c** over the direction perpendicular to the plane $(\mathbf{a}, \mathbf{b})$. Accordingly, $V = (\mathbf{a} \wedge \mathbf{b}) \cdot \mathbf{c}$.

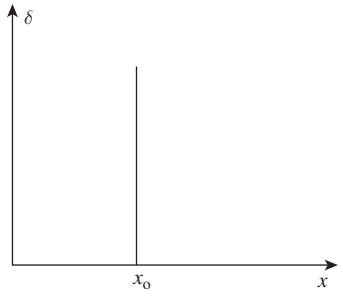


Fig. 1.3
Schematic representation of the Dirac function $\delta(x - x_o)$.

Dirac delta function. In a three-dimensional space the Dirac delta function $\delta(\mathbf{r} - \mathbf{r}_0)$ is defined by the following properties:

$$\delta = 0 \quad \text{for} \quad (\mathbf{r} \neq \mathbf{r}_0), \quad \delta = \infty \quad \text{for} \quad (\mathbf{r} = \mathbf{r}_0), \int_S \delta(\mathbf{r} - \mathbf{r}_0) d\mathbf{r} = 1,$$

where S is the full $\mathbf{r}$ space. The function δ is highly discontinuous and is qualitatively represented in Fig. 1.3 as a straight line.

Crystal lattice. This describes the repetition geometry of the unit cell (see Fig. 1.4). An equivalent mathematical definition is the following: a crystal lattice is represented by the lattice function $L(\mathbf{r})$, where

$$L(\mathbf{r}) = \sum_{u,v,w=-\infty}^{+\infty} \partial(\mathbf{r} - \mathbf{r}_{u,v,w}); \tag{1.2}$$

where $\partial(\mathbf{r} - \mathbf{r}_{u,v,w})$ is the Dirac delta function centred on $\mathbf{r}_{u,v,w} = u\mathbf{a} + v\mathbf{b} + w\mathbf{c}$ and u,v,w are integer numbers.

Convolution. The convolution of two functions $\rho(\mathbf{r})$ and $g(\mathbf{r})$ (this will be denoted as $\rho(\mathbf{r}) \otimes g(\mathbf{r})$) is the integral

$$C(\mathbf{u}) = \rho(\mathbf{r}) \otimes g(\mathbf{r}) = \int_S \rho(\mathbf{r}) g(\mathbf{u} - \mathbf{r}) d\mathbf{r}. \tag{1.3}$$

The reader will notice that the function g is translated by the vector $\mathbf{u}$ and inverted before being integrated.

The convolution of the function $\rho(\mathbf{r})$, describing the unit cell content, with a lattice function centred in $\mathbf{r}_0$, is equivalent to shifting $\rho(\mathbf{r})$ by the vector $\mathbf{r}_0$. Indeed

$$\delta(\mathbf{r} - \mathbf{r}_0) \otimes \rho(\mathbf{r}) = \rho(\mathbf{r} - \mathbf{r}_0).$$

Accordingly, the convolution of $\rho(\mathbf{r})$ with the lattice function $L(\mathbf{r})$ describes the periodic repetition of the unit cell content, and therefore describes the crystal (see Fig. 1.5):

$$L(\mathbf{r}) \otimes \rho(\mathbf{r}) = \sum_{u,v,w=-\infty}^{+\infty} \partial(\mathbf{r} - \mathbf{r}_{u,v,w}) \otimes \rho(\mathbf{r}) = \sum_{u,v,w=-\infty}^{+\infty} \rho(\mathbf{r} - \mathbf{r}_{u,v,w}). \tag{1.4}$$

Primitive and centred cells. A cell is primitive if it contains only one lattice point and centered if it contains more lattice points. The cells useful in crystallography are listed in Table 1.1: for each cell the multiplicity, that is the number of lattice points belonging to the unit cell, and their positions are emphasized.

Symmetry operators. These relate symmetry equivalent positions. Two positions $\mathbf{r}$ and $\mathbf{r}'$ are symmetry equivalent if they are related by the symmetry operator $\mathbf{C} = (\mathbf{R}, \mathbf{T})$, where $\mathbf{R}$ is the rotational component and $\mathbf{T}$ the translational component. More explicitly,

$$\begin{vmatrix} x' \\ y' \\ z' \end{vmatrix} = \begin{vmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{vmatrix} \begin{vmatrix} x \\ y \\ z \end{vmatrix} + \begin{vmatrix} \mathrm{T}_1 \\ \mathrm{T}_2 \\ \mathrm{T}_3 \end{vmatrix}, \tag{1.5}$$

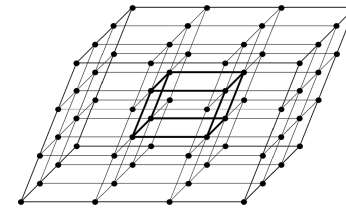


Fig. 1.4
The unit cell (bold lines) and the corresponding lattice.

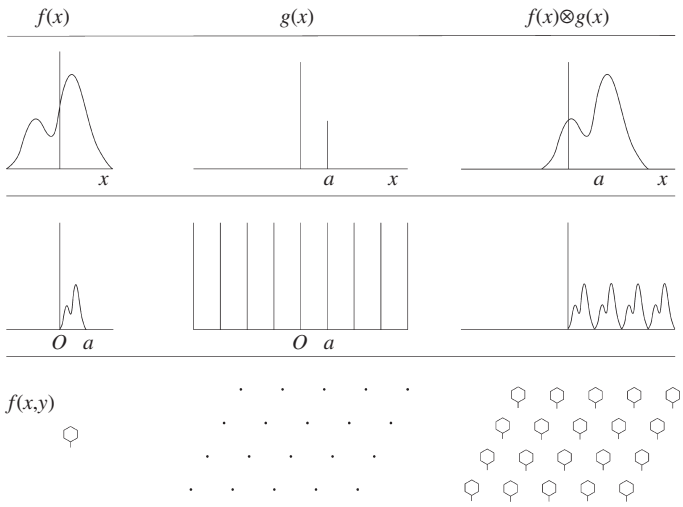


Fig. 1.5
The convolution of the motif f with a delta function is represented in the first line. In the second line f is still the motif, g is a one-dimensional lattice, $f(x) \otimes g(x)$ is a one-dimensional crystal. In the third line, a two-dimensional motif and lattice are used.

Table 1.1 The conventional types of unit cell and corresponding lattice multiplicity

Symbol	Type	Positions of additional lattice points	Number of lattice points per cell
P	Primitive	—	1
I	body-centred	(1/2, 1/2, 1,2)	2
A	A-face centred	(0, 1/2, 1/2)	2
B	B-face centred	(1/2, 0, 1/2)	2
C	C-face centred	(1/2, 1/2, 0)	2
F	All faces centred	(1/2, 1/2, 0), (1/2, 0, 1/2) (0, 1/2, 1/2)	4
R	Rhombohedrally centred (description with 'hexagonal axes')	(1/3, 2/3, 2/3), (2/3, 1/3, 1/3)	3

where (x',y',z') and (x,y,z) are the coordinates of $\mathbf{r}'$ and $\mathbf{r}$ respectively. In a vectorial form,

$$\mathbf{r}' = \mathbf{R}\mathbf{r} + \mathbf{T}.$$

If the determinant $|\mathbf{R}| = 1$ the symmetry operator is proper and refers to objects directly congruent; if $|\mathbf{R}| = -1$ the symmetry operator is improper and refers to enantiomorph objects. The type of symmetry operator may be identified according to Table 1.2:

Table 1.2 Trace and determinant of the rotation matrix for crystallographic symmetry operators

Element	1	2	3	4	6	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{6}$
trace	3	$\bar{1}$	0	1	2	$\bar{3}$	1	0	$\bar{1}$	$\bar{2}$
determinant	1	1	1	1	1	$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$	$\bar{1}$

Point group symmetry. This is a compatible combination of symmetry operators, proper or improper, without translational components, and intersecting at one point. The number of crystallographic point groups is 32 and their symbols are shown in Table 1.3. Most of the physical properties depend on the point group symmetry of the crystal (they show a symmetry equal to or larger than the point group symmetry: *Neumann principle*).

Crystal systems. Crystals belonging to point groups with common features can be described by unit cells of the same type. For example, crystals with only three twofold axes, no matter if proper or improper, can be described by an orthogonal cell. These crystals then belong to the same crystal system, the orthorhombic system. The relations between crystal system-point groups are shown in Table 1.4. For each system the allowed Bravais lattices, the characterizing symmetry, and the type of unit cell parameters are reported.

Table 1.3 List of the 32 crystal point groups, Laue groups, and lattice point groups

Crystal systems	Point groups			Laue classes	Lattice point groups
	Non-centrosymmetric		Centrosymmetric		
Triclinic	1		$\bar{1}$	$\bar{1}$	$\bar{1}$
Monoclinic	2	m	$2/m$	$2/m$	$2/m$
Orthorhombic	222	$mm2$	mmm	mmm	mmm
Tetragonal	$\left[\begin{matrix}4\\422\end{matrix}\right.$	$\bar{4}$ $4mm, \bar{4}2m$	$4/m$ $4/mmm$	$4/m$ $4/mmm$	$\left.\right] 4/mmm$
Trigonal	$\left[\begin{matrix}3\\32\end{matrix}\right.$	 $3m$	$\bar{3}$ $\bar{3}m$	$\bar{3}$ $\bar{3}m$	$\left.\right] \bar{3}m$
Hexagonal	$\left[\begin{matrix}6\\622\end{matrix}\right.$	$\bar{6}$ $6mm, \bar{6}2m$	$6/m$ $6/mmm$	$6/m$ $6/mmm$	$\left.\right] 6/mmm$
Cubic	$\left[\begin{matrix}23\\432\end{matrix}\right.$	 $\bar{4}3m$	$m\bar{3}$ $m\bar{3}m$	$m\bar{3}$ $m\bar{3}m$	$\left.\right] m\bar{3}m$

Table 1.4 Crystal systems, characterizing symmetry and unit cell parameters

Crystal system	Bravais type(s)	Characterizing symmetry	Unit cell properties
Triclinic	P	None	a, b, c, α, β, γ
Monoclinic	P, C	Only one 2-fold axis	a, b, c, 90°, β, 90°
Orthorhombic	P, I, F	Only three perpendicular 2-fold axes	a, b, c, 90°, 90°, 90°
Tetragonal	P, I	Only one 4-fold axis	a, a, c, 90°, 90°, 90°
Trigonal	P, R	Only one 3-fold axis	a, a, c, 90°, 90°, 120°
Hexagonal	P	Only one 6-fold axis	a, a, c, 90°, 90°, 120°
Cubic	P, F, I	Four 3-fold axes	a, a, a, 90°, 90°, 90°

Space groups. Three-dimensional crystals show a symmetry belonging to one of the 230 space groups reported in Table 1.5. The space group is a set of symmetry operators which take a three- dimensional periodic object (say a crystal) into itself. In other words, the crystal is invariant under the symmetry operators of the space group.

The space group symmetry defines the *asymmetric unit*: this is the smallest part of the unit cell applying to which the symmetry operators, the full content of the unit cell, and then the full crystal, are obtained. This last statement implies that the space group also contains the information on the repetition geometry (this is the first letter in the space group symbol, and describes the type of unit cell).

1.3 The reciprocal space

We recall the main concepts and definitions concerning crystal reciprocal space.

Reciprocal space. In a scattering experiment, the amplitude of the wave (say $F(\mathbf{r}^*)$, in Thomson units) scattered by an object represented by the function $\rho(\mathbf{r})$, is the Fourier transform of $\rho(\mathbf{r})$:

$$F(\mathbf{r}^*) = T[\rho(\mathbf{r})] = \int_S \rho(\mathbf{r}) \exp(2\pi i \mathbf{r}^* \cdot \mathbf{r}) d\mathbf{r}, \tag{1.6}$$

where T is the symbol of the Fourier transform, S is the full space where the scattering object is immersed, $\mathbf{r}^* = \mathbf{s} - \mathbf{s}_0$ is the difference between the unit vector $\mathbf{s}$, oriented along the direction in which we observe the radiation, and the unit vector $\mathbf{s}_0$ along which the incident radiation comes (see Fig. 1.6). We recall that $|\mathbf{r}^*| = 2 \sin\theta/\lambda$, where 2θ is the angle between the direction of incident radiation and the direction along which the scattered radiation is observed, and λ is the wavelength. We will refer to $\mathbf{r}^*$ as to the generic point of the reciprocal space S^*, the space of the Fourier transform.

$F(\mathbf{r}^*)$ is a complex function, say $F(\mathbf{r}^*) = A(\mathbf{r}^*) + iB(\mathbf{r}^*)$. It may be shown that, for two enantiomorphous objects, the corresponding $F(\mathbf{r}^*)$ are the complex conjugates of each other: they therefore have the same modulus $|F(\mathbf{r}^*)|$. As a consequence, for a centrosymmetrical object, $F(\mathbf{r}^*)$ is real.

Table 1.5 The 230 three-dimensional space groups arranged by crystal systems and point groups. Point groups not containing improper symmetry operators are in a square box (the corresponding space groups are the only ones in which proteins may crystallize). Space groups (and enantiomorphous pairs) that are uniquely determinable from the symmetry of the diffraction pattern and from systematic absences (see Section 1.5) are shown in bold type

Crystal system	Point group	Space groups
Triclinic	$\boxed{1}$	P1
	$\bar{1}$	P$\bar{1}$
Monoclinic	$\boxed{2}$	P2, P2_1, C2
	m	Pm, P*c*, C*m*, C*c*
	2/*m*	P2/*m*, P2_1/*m*, C2/*m*, P2/*c*, **P2_1/*c***, C2/*c*
Orthorhombic	$\boxed{222}$	P222, **P222_1**, **P2_12_12**, **P$2_12_12_1$**, **C222_1**, C222, F222, I222, I$2_12_12_1$
	*mm*2	P*mm*2, P*mc*2_1, P*cc*2, P*ma*2_1, P*ca*2_1, P*nc*2_1, P*mn*2_1, P*ba*2, P*na*2_1, P*nn*2, C*mm*2, C*mc*2_1, C*cc*2, A*mm*2, A*bm*2, A*ma*2, A*ba*2, F*mm*2, ***Fdd2***, I*mm*2, I*ba*2, I*ma*2
	mmm	P*mmm*, **P*nnn***, P*ccm*, **P*ban***, P*mma*, **P*nna***, P*mna*, **P*cca***, P*bam*, **P*ccn***, P*bcm*, P*nnm*, P*mmn*, **P*bcn***, **P*bca***, P*nma*, C*mcm*, C*mca*, C*mmm*, C*ccm*, C*mma*, C*cca*, F*mmm*, **F*ddd***, I*mmm*, *Ibam*, **I*bca***, I*mma*
Tetragonal	$\boxed{4}$	P4, **P4_1**, P4_2, **P4_3**, I4, **I4_1**
	$\bar{4}$	P$\bar{4}$, I$\bar{4}$
	4/*m*	P4/*m*, P4_2/*m*, **P4/*n***, **P4_2/*n***, I4/*m*, **I4_1/*a***
	$\boxed{422}$	P422, P42_12, P4_122, **P4_12_12**, **P4_222**, **P4_22_12**, **P4_322**, **P4_32_12**, I422, **I4_122**
	4*mm*	P4*mm*, P4*bm*, P4_2*cm*, P4_2*nm*, P4*cc*, P4*nc*, P4_2*mc*, P4_2*bc*, I4*mm*, I4*cm*, I4_1*md*, **I4_1*cd***
	$\bar{4}$*m*	P$\bar{4}2m$, P$\bar{4}2c$, P$\bar{4}2_1m$, **P$\bar{4}2_1$*c***, P$\bar{4}m2$, P$\bar{4}c2$, P$\bar{4}b2$, P$\bar{4}n2$, I$\bar{4}m2$, I$\bar{4}c2$, I$\bar{4}2m$, I$\bar{4}2d$
	4/*mmm*	P4/*mmm*, P4/*mcc*, **P4/*nbm***, **P4/*nnc***, P4/*mbm*, P4/*mnc*, **P4/*nmm***, **P4/*ncc***, P4_2/*mmc*, P4_2/*mcm*, **P4_2/*nbc***, **P4_2/*nnm***, P4_2/*mbc*, P4_2*mnm*, **P4_2/*nmc***, **P4_2/*ncm***, I4/*mmm*, I4/*mcm*, **I4_1/*amd***, **I4_1/*acd***
Trigonal–hexagonal	$\boxed{3}$	P3, **P3_1**, **P3_2**, R3
	$\bar{3}$	P$\bar{3}$, R$\bar{3}$
	$\boxed{32}$	P312, P321, **P3_112**, **P3_121**, **P3_212**, **P3_221**, R32
	3*m*	P3*m*1, P31*m*, P3*c*1, P31*c*, R3*m*, R3*c*
	$\bar{3}$*m*	P$\bar{3}1m$, P$\bar{3}1c$, P$\bar{3}m1$, P$\bar{3}c1$, R$\bar{3}m$, R$\bar{3}c$
	$\boxed{6}$	P6, **P6_1**, **P6_5**, P6_3, **P6_2**, **P6_4**
	$\bar{6}$	P$\bar{6}$
	6/*m*	P6/*m*, P6_3/*m*
	$\boxed{622}$	P622, **P6_122**, **P6_522**, **P6_222**, **P6_422**, **P6_322**
	6*mm*	P6*mm*, P6*cc*, P6_3*cm*, P6_3*mc*
	$\bar{6}$*m*	P$\bar{6}m2$, P$\bar{6}c2$, P$\bar{6}2m$, P$\bar{6}2c$
	6/*mmm*	P6/*mmm*, P6/*mcc*, P6_3/*mcm*, P6_3/*mmc*
Cubic	$\boxed{23}$	P23, F23, I23, **P2_13**, I2_13
	m$\bar{3}$	P*m*$\bar{3}$, **P*n*$\bar{3}$**, F*m*$\bar{3}$, **F*d*$\bar{3}$**, I*m*$\bar{3}$, **P*a*$\bar{3}$**, **I*a*$\bar{3}$**
	$\boxed{432}$	P432, **P4_232**, F432, **F4_132**, I432, **P4_332**, **P4_132**, **I4_132**
	$\bar{4}$3*m*	P$\bar{4}3m$, F43*m*, I$\bar{4}3m$, P43*n*, F$\bar{4}3c$, **I$\bar{4}$3*d***
	m$\bar{3}$*m*	P*m*$\bar{3}m$, **P*n*$\bar{3}$*n***, P*m*$\bar{3}n$, **P*n*$\bar{3}$*m***, F*m*$\bar{3}m$, F*m*$\bar{3}c$, **F*d*$\bar{3}$*m***, **F*d*$\bar{3}$*c***, I*m*$\bar{3}m$, **I*a*$\bar{3}$*d***

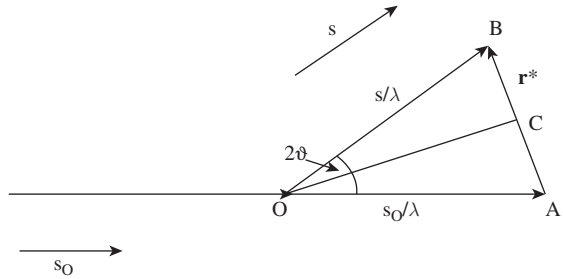


Fig. 1.6
The scatterer is at O, $\mathbf{s}_o$ and $\mathbf{s}$ are unit vectors, the first along the incident X-ray radiation, the second along the direction in which the scattered intensity is observed. To calculate $|\mathbf{r}^*|$ it is sufficient to notice that the triangle AOB is isosceles and that point C divides AB into two equal parts.

$\rho(\mathbf{r})$ may be recovered via the inverse Fourier transform of $F(\mathbf{r}^*)$:

$$\rho(\mathbf{r}) = T^{-1}[F(\mathbf{r}^*)] = \int_{S^*} F(\mathbf{r}^*) \exp(-2\pi i \mathbf{r}^* \cdot \mathbf{r}) d\mathbf{r}^*. \tag{1.7}$$

The reciprocal lattice. It is usual in crystallography to take, as a reference system for the reciprocal space, the reciprocal vectors $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$, defined below. Given a direct lattice, with unit vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, its reciprocal lattice is identified by the vectors $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ satisfying the following two conditions:

1. $\mathbf{a}^* \wedge \mathbf{b} = \mathbf{a}^* \wedge \mathbf{c} = \mathbf{b}^* \wedge \mathbf{a} = \mathbf{b}^* \wedge \mathbf{c} = \mathbf{c}^* \wedge \mathbf{a} = \mathbf{c}^* \wedge \mathbf{b} = \mathbf{0}$
2. $\mathbf{a}^* \cdot \mathbf{a} = \mathbf{b}^* \cdot \mathbf{b} = \mathbf{c}^* \cdot \mathbf{c} = 1$

Condition 1 defines the orientation of the reciprocal basis vectors (e.g. $\mathbf{a}^*$ is perpendicular to $\mathbf{b}$ and $\mathbf{c}$, etc.), whereas condition 2 fixes their modulus. From the above conditions the following relations arise:

(i) $$\mathbf{a}^* = \frac{1}{V}\mathbf{b} \wedge \mathbf{c},\ \mathbf{b}^* = \frac{1}{V}\mathbf{c} \wedge \mathbf{a},\ \mathbf{c}^* = \frac{1}{V}\mathbf{a} \wedge \mathbf{b},\ V^* = V^{-1}, \tag{1.8}$$

(ii) the scalar product of the two vectors $\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$ and $\mathbf{r}^* = x^*\mathbf{a}^* + y^*\mathbf{b}^* + z^*\mathbf{c}^*$, one defined in direct and the other in reciprocal space, reduces to the sum of the products of the corresponding coordinates:

$$\mathbf{r} \cdot \mathbf{r}^* = x^*x + y^*y + z^*z = \bar{\mathbf{X}}^*\mathbf{X} = \left|x^*y^*z^*\right| \left|\begin{matrix} x \\ y \\ z \end{matrix}\right|; \tag{1.9}$$

(iii) the generic reciprocal lattice point is defined by the vector $\mathbf{r}^*_{hkl} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, with integer values of h, k, l. We will also denote it by $\mathbf{r}^*_{\mathbf{H}}$ or $\mathbf{r}^*_{\mathbf{h}}$, where $\mathbf{H}$ or $\mathbf{h}$ represent the triple h,k,l.

(iv) $\mathbf{r}^*_{hkl}$ represents the family (in direct space) of lattice planes with Miller indices (hkl). Indeed $\mathbf{r}^*_{hkl}$ is perpendicular to the planes of the family (hkl) and its modulus is equal to the spacing of the planes (hkl): i.e.

$$\mathbf{r}^*_{hkl} \perp (hkl), \quad \text{and} \quad |\mathbf{r}^*_{hkl}| = 1/d_{hkl}. \tag{1.10}$$

(v) the reciprocal lattice may be represented by the reciprocal lattice function

$$L(\mathbf{r}^*) = \frac{1}{V} \sum\nolimits_{h,k,l=-\infty}^{+\infty} \partial\left(\mathbf{r}^* - \mathbf{r}^*_{\mathbf{H}}\right); \tag{1.11}$$

$L(\mathbf{r}^*)$ is the Fourier transform of the direct lattice:

$$L(\mathbf{r}^*) = T[L(\mathbf{r})] = T\left[\sum_{u,v,w=-\infty}^{+\infty} \partial(\mathbf{r} - \mathbf{r}_{u,v,w})\right] = \frac{1}{V}\sum_{\mathbf{H}} \partial(\mathbf{r}^* - \mathbf{r}_{\mathbf{H}}^*). \quad (1.12)$$

Atomic scattering factor $f(\mathbf{r}^*)$. This is the amplitude, in Thomson units, of the wave scattered by the atom and observed at the reciprocal space point $\mathbf{r}^*$. $f(\mathbf{r}^*)$ is the Fourier transform of the atomic electron density ρ_a:

$$f(\mathbf{r}^*) = T[\rho_a(\mathbf{r})] = \int_S \rho_a(\mathbf{r}) \exp(2\pi i \mathbf{r}^* \cdot \mathbf{r}) d\mathbf{r}. \quad (1.13)$$

Usually $\rho_a(\mathbf{r})$ includes thermal displacement: accordingly, under the isotropic scattering assumption,

$$f(r^*) = f_0(r^*) \exp\left(-Br^{*2}/4\right) = f_0(r^*) \exp(-B \sin^2\theta/\lambda^2), \quad (1.14)$$

where $f_0(r^*)$ is the scattering factor of the atom at rest, and B is the isotropic temperature factor. At $r^* = 0$, $f(r^*)$ is maximum (then $f(r^*) = Z$, where Z is the atomic number). The decay with r^* is sharper for high B values (see Fig. 1.7).

Molecular scattering factor $F_M(\mathbf{r}^*)$. This is the amplitude, in Thomson units, of the wave scattered by a molecule, observed at the reciprocal space point $\mathbf{r}^*$. It is the Fourier transform of the electron density of the molecule:

$$\begin{aligned} F_M(\mathbf{r}^*) = T[\rho_M(\mathbf{r})] &= \int_S \sum_{j=1}^{N} \rho_{aj}(\mathbf{r} - \mathbf{r}_j) \exp(2\pi i \mathbf{r}^* \cdot \mathbf{r}) d\mathbf{r} \\ &= \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{r}^* \cdot \mathbf{r}_j), \end{aligned} \quad (1.15)$$

where $\rho_M(\mathbf{r})$ is the electron density of the molecule and N is the corresponding number of atoms. $F_M(\mathbf{r}^*)$ is a continuous function of $\mathbf{r}^*$.

Structure factor $F_M(\mathbf{r}^*)$ *of a unit cell*. This is the amplitude, in Thomson units, of the wave scattered by all the molecules contained in the unit cell and observed at the reciprocal space point $\mathbf{r}^*$. $F_M(\mathbf{r}^*)$ is the Fourier transform of the electron density of the unit cell:

$$\begin{aligned} F_M(\mathbf{r}^*) = T[\rho_M(\mathbf{r})] &= \int_S \sum_{j=1}^{N} \rho_{aj}(\mathbf{r} - \mathbf{r}_j) \exp\left(2\pi i \mathbf{r}^* \cdot \mathbf{r}\right) d\mathbf{r} \\ &= \sum_{j=1}^{N} f_j \exp\left(2\pi i \mathbf{r}^* \cdot \mathbf{r}_j\right). \end{aligned} \quad (1.16)$$

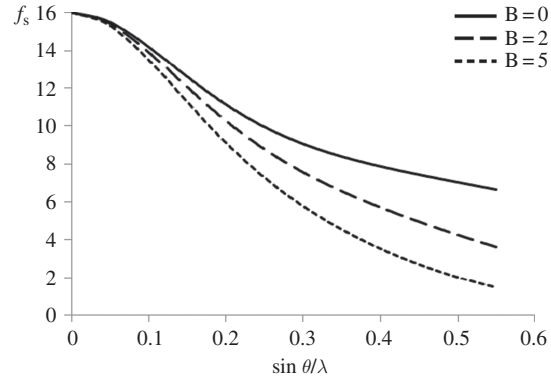


Fig. 1.7
Scattering factor of sulphur for different values of the temperature factor.

$\rho_M(\mathbf{r})$ is now the electron density in the unit cell, N is the corresponding number of atoms, and $F_M(\mathbf{r}^*)$ is a continuous function of $\mathbf{r}^*$. The reader will certainly have noted that we have used for the unit cell the same notation employed for describing the scattering from a molecule: indeed, from a physical point of view, the unit cell content may be considered to be a collection of molecules.

Structure factor $F(\mathbf{r}^*)$ for a crystal. This is the amplitude, in Thomson units, of the wave scattered by the crystal as observed at the reciprocal space point $\mathbf{r}^*$. It is the Fourier transform of the electron density of the crystal. In accordance with equation 1.4

$$F(\mathbf{r}^*) = T[\rho_{cr}(\mathbf{r})] = T[\rho_M(\mathbf{r}) \otimes L(\mathbf{r})]$$

and, owing to the convolution theorem,

$$F(\mathbf{r}^*) = T[\rho_M(\mathbf{r})] \cdot T[L(\mathbf{r})] = F_M(\mathbf{r}^*) \cdot \frac{1}{V} \sum\nolimits_{\mathbf{H}} \partial(\mathbf{r}^* - \mathbf{r}^*_{\mathbf{H}}). \qquad (1.17)$$

$F(\mathbf{r}^*)$ is now a highly discontinuous function which is different from zero only at the reciprocal lattice points defined by the vectors $\mathbf{r}^*_{\mathbf{H}}$. From now on, $F_M(\mathbf{r}^*_{\mathbf{H}})$ will be written as $F_{\mathbf{H}}$ and will simply be called the *structure factor*. The study of $F_{\mathbf{H}}$ and of its statistical properties is basic for phasing methods.

Limits of a diffraction experiment. Diffraction occurs when $\mathbf{r}^*_{\mathbf{H}}$ meet the Ewald sphere (see Fig. 1.8). A diffraction experiment only allows measurement of reflections with $\mathbf{r}^*_{\mathbf{H}}$ contained within the *limiting sphere* (again, see Fig. 1.8). Data resolution is usually described in terms of the maximum measurable value of $|\mathbf{r}^*_{\mathbf{H}}|$ (say $|\mathbf{r}^*_{\mathbf{H}}|_{\max}$): in this case the resolution is expressed in Å^{-1}. More frequently, because of equation (1.10), in terms of the minimum measurable value of $d_{\mathbf{H}}$ (say $(d_{\mathbf{H}})_{\min}$): in this case data resolution is expressed in Å. Accordingly, stating that data resolution is 2 Å is equivalent to saying that only reflections with $d_{\mathbf{H}} > 2$ Å were measured. Severe resolution limits are frequent for proteins: often reflections inside and close to the limiting sphere cannot be measured because of the poor quality of the crystal. Usually, better

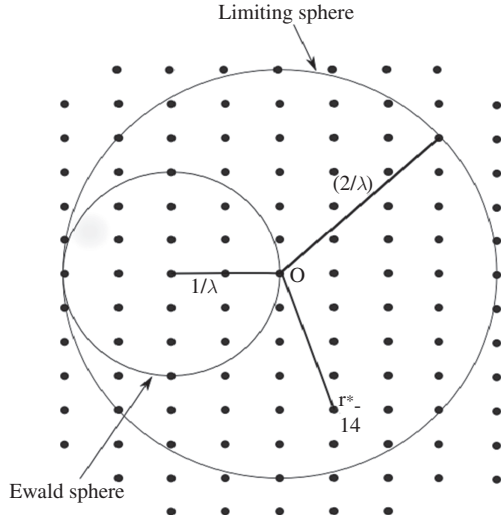


Fig. 1.8
Ewald and limiting spheres.

data can be collected, not by diminishing λ, but by performing the experiment in cryo-conditions, to fight decay of the scattering factor due to thermal displacement.

Electron density calculations. According to equation (1.17), the electron density in a point $\boldsymbol{r}$ having fractional coordinates (x,y,z) may be estimated via the Fourier series

$$\begin{aligned}\rho(\mathbf{r}) &= \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} F_{hkl}\exp[-2\pi i(hx+ky+lz)] \\ &= \frac{2}{V}\sum_{h=0}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} |F_{hkl}|\cos[\phi_{hkl}-2\pi(hx+ky+lz)].\end{aligned} \tag{1.18}$$

The last term is obtained by applying the Friedel law, and shows that the electron density is a real function. As previously recalled, there are limitations to the number of measurable reflections: accordingly, series (1.18) will show truncation effects which are more and more severe as soon as the resolution becomes worse (see Fig. 1.9 and Section 7.3.1).

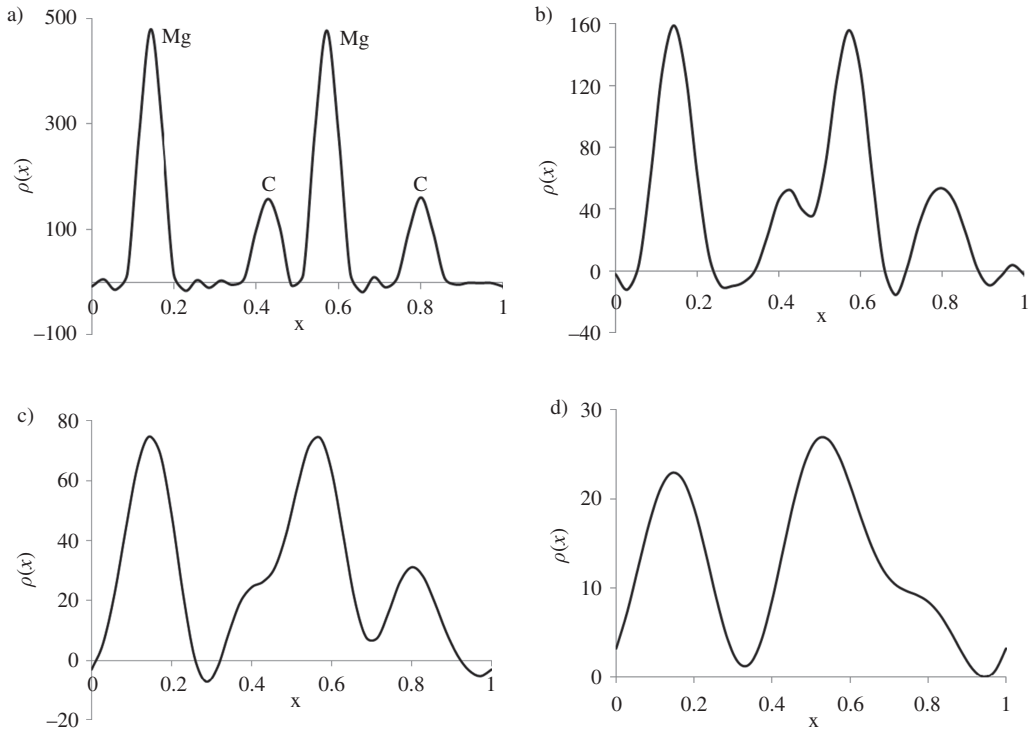


Fig. 1.9
Electron density maps of a (non-realistic) four-atom one-dimensional structure. Data up to: (a) 0.9 Å; (b) 1.5 Å; (c) 2 Å; (d) 3 Å. In all cases true phases have been used: the differences between the maps are only due to truncation effects. Changes in peak intensity and positions are clearly visible.

1.4 The structure factor

The structure factor $F_{\mathbf{h}}$ plays a central role in phasing methods: its simple geometrical interpretation is therefore mandatory. Let N be the number of atoms in the unit cell, f_j the scattering factor of the jth atom, and x_j, y_j, z_j its fractional coordinates: then

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp\left(2\pi i\, \mathbf{h} \cdot \mathbf{r}_j\right) = \sum_{j=1}^{N} f_j \exp\left[2\pi i(hx_j + ky_j + lz_j)\right]. \quad (1.19)$$

f_j includes the thermal displacement and must be calculated at the $\sin\theta/\lambda$ corresponding to the reflection $\mathbf{h}$: to do that, firstly, the modulus of the vector $\mathbf{r}^*_{hkl} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$ should be calculated and then, by using the equation $|\mathbf{r}^*| = 2\sin\theta/\lambda$, the searched f value may be obtained.

Let us rewrite (1.19) in the form

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(i\alpha_j) = |F_{\mathbf{h}}| \exp(i\phi_{\mathbf{h}}) = A_{\mathbf{h}} + iB_{\mathbf{h}}, \quad (1.20)$$

where

$$\alpha_j = 2\pi\, \mathbf{h} \cdot \mathbf{r}_j, \quad A_{\mathbf{h}} = \sum_{j=1}^{N} f_j \cos(2\pi\, \mathbf{h} \cdot \mathbf{r}_j),$$

$$B_{\mathbf{h}} = \sum_{j=1}^{N} f_j \sin(2\pi\, \mathbf{h} \cdot \mathbf{r}_j).$$

On representing $F_{\mathbf{h}}$ in an Argand diagram (Fig. 1.10), we obtain a vectorial diagram with N vectors each characterized by a modulus f_j and an angle α_j with the real axis: the value

$$\phi_{\mathbf{h}} = \tan^{-1}(B_{\mathbf{h}}/A_{\mathbf{h}}) \quad (1.21)$$

depends on the moduli and on the mutual orientation of the vectors $\mathbf{f}_j$ and is said to be the *phase* of $F_{\mathbf{h}}$.

In a space group with symmetry higher than P1, with point group symmetry of order m, for each atomic position $\mathbf{r}_j$, located in the asymmetric unit, there are m symmetry equivalent positions

$$\mathbf{r}_{js} = \mathbf{R}_s \mathbf{r}_j + \mathbf{T}_s.$$

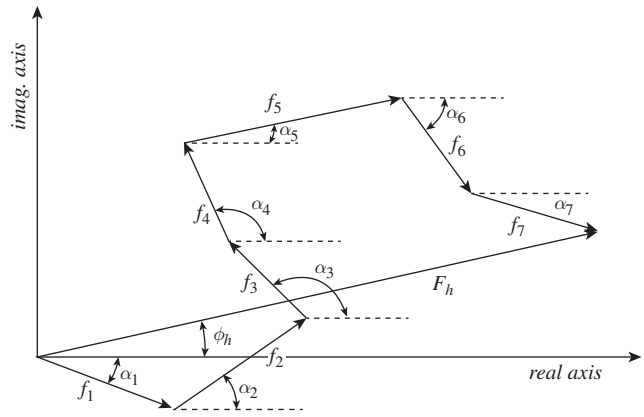


Fig. 1.10
The structure factor $F_{\mathbf{h}}$ is represented in the Argand plane as the sum of $N = 7$ $\mathbf{f}_j$ vectors, with modulus f_j and phase angle α_j.

Then the structure factor takes the form

$$F_{\mathbf{h}} = \sum_{j=1}^{t} f_j \sum_{s=1}^{m} \exp 2\pi i\mathbf{h}(\mathbf{R}_s\mathbf{r}_j + \mathbf{T}_s)$$

where t is the number of atoms in the asymmetric unit.

1.5 Symmetry in reciprocal space

A diffraction experiment allows us to *see* the reciprocal space: it is then very important to understand which symmetry relations can be discovered there as a consequence of the symmetry present in direct space. Here we summarize the main effects.

1.5.1 Friedel law

In accordance with equation (1.20) we write $F_{\boldsymbol{h}} = A_{\boldsymbol{h}} + \mathrm{i}B_{\boldsymbol{h}}$. Then it will follow that $F_{-\boldsymbol{h}} = A_{\boldsymbol{h}} - \mathrm{i}B_{\boldsymbol{h}}$, and consequently

$$\phi_{-\boldsymbol{h}} = -\phi_{\boldsymbol{h}}. \tag{1.22}$$

The value of $\phi_{-\boldsymbol{h}}$ is the opposite of the value of $\phi_{\boldsymbol{h}}$, see Fig. 1.11. Since

$$I_{\boldsymbol{h}} = (A_{\boldsymbol{h}} - \mathrm{i}B_{\boldsymbol{h}})(A_{\boldsymbol{h}} + \mathrm{i}B_{\boldsymbol{h}}) = A_{\boldsymbol{h}}^2 + B_{\boldsymbol{h}}^2,$$

$$I_{-\boldsymbol{h}} = (A_{\boldsymbol{h}} + \mathrm{i}B_{\boldsymbol{h}})(A_{\boldsymbol{h}} - \mathrm{i}B_{\boldsymbol{h}}) = A_{\boldsymbol{h}}^2 + B_{\boldsymbol{h}}^2,$$

we deduce the Friedel law, according to which the diffraction intensities associated with the vectors $\boldsymbol{h}$ and $-\boldsymbol{h}$ of reciprocal space are equal. Since these intensities appear to be related by a centre of symmetry, usually, although imperfectly, it is said that the diffraction by itself introduces a centre of symmetry.

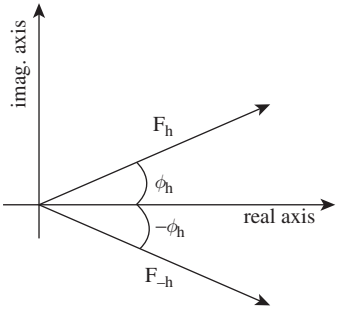


Fig. 1.11
The Friedel law.

1.5.2 Effects of symmetry operators in reciprocal space

Let us suppose that the symmetry operator $\boldsymbol{C} = (\mathbf{R}, \mathbf{T})$ exists in direct space. We wonder what kind of relationships the presence of the operator $\boldsymbol{C}$ brings in reciprocal space.

Since

$$\begin{aligned} F_{\bar{\boldsymbol{h}}\mathbf{R}} \exp\left(2\pi\,\mathrm{i}\bar{\boldsymbol{h}}\mathrm{T}\right) &= \sum_{j=1}^{N} f_j \exp\left(2\pi\,\mathrm{i}\bar{\boldsymbol{h}}\mathbf{R}\mathbf{X}_j\right) \cdot \exp\left(2\pi\,\mathrm{i}\bar{\boldsymbol{h}}\mathbf{T}\right) \\ &= \sum_{j=1}^{N} f_j \exp\left[2\pi\,\mathrm{i}\bar{\boldsymbol{h}}(\mathbf{R}\mathbf{X}_j + \mathbf{T})\right] = F_{\boldsymbol{h}}, \end{aligned}$$

we can write

$$F_{\bar{\boldsymbol{h}}\mathbf{R}} = F_{\boldsymbol{h}} \exp(-2\pi\,\mathrm{i}\bar{\boldsymbol{h}}\mathbf{T}). \tag{1.23}$$

Sometimes it is convenient to split equation (1.23) into two relations, the first involving moduli and the second the phases

$$|F_{\bar{\boldsymbol{h}}\mathbf{R}}| = |F_{\boldsymbol{h}}|, \tag{1.24}$$

$$\phi_{\bar{\boldsymbol{h}}\mathbf{R}} = \phi_{\mathbf{h}} - 2\pi\bar{\boldsymbol{h}}\mathbf{T}. \tag{1.25}$$

From (1.23) it is concluded that intensities $I_{\boldsymbol{h}}$ and $I_{\bar{\boldsymbol{h}}\mathbf{R}}$ are equal, while their phases are related by equation (1.25).

Reflections related by (1.24) and by the Friedel law are said to be *symmetry equivalent reflections.* For example, in P2 the set of symmetry equivalent reflections is

$$|F_{hkl}| = |F_{\bar{h}k\bar{l}}| = |F_{\bar{h}\bar{k}l}| = |F_{h\bar{k}l}|. \tag{1.26}$$

The reader will easily verify that space groups P4, P$\bar{4}$, and P4/m show the following set of symmetry equivalent reflections:

$$|F_{hkl}| = |F_{\bar{h}\bar{k}l}| = |F_{\bar{k}hl}| = |F_{k\bar{h}l}| = |F_{\bar{h}\bar{k}\bar{l}}| = |F_{hk\bar{l}}|, = |F_{kh\bar{l}}| = |F_{\bar{k}h\bar{l}}|.$$

1.5.3 Determination of reflections with restricted phase values

Let us suppose that for a given set of reflections the relationship $\bar{\boldsymbol{h}}\mathbf{R} = -\bar{\boldsymbol{h}}$ is satisfied. If we apply (1.25) to this set we will obtain $2\phi_{\boldsymbol{h}} = 2\pi\bar{\boldsymbol{h}}\mathbf{T} + 2n\pi$, from which

$$\phi_{\boldsymbol{h}} = \pi\bar{\boldsymbol{h}}\mathbf{T} + n\pi. \tag{1.27}$$

Equation (1.27) restricts the phase $\phi_{\boldsymbol{h}}$ to two values, $\pi\bar{\boldsymbol{h}}\mathbf{T}$ or $\pi(\bar{\boldsymbol{h}}\mathbf{T} + 1)$. These reflections are called reflections with restricted phase values, or less correctly, 'centrosymmetric'.

If the space group is centrosymmetric (cs.) the inversion operator

$$\mathbf{R} = \begin{vmatrix} \bar{1} & 0 & 0 \\ 0 & \bar{1} & 0 \\ 0 & 0 & \bar{1} \end{vmatrix}, \quad \mathbf{T} = \begin{vmatrix} T_1 \\ T_2 \\ T_3 \end{vmatrix}$$

will exist. In this case every reflection is a restricted phase reflection and will assume the values $\pi\bar{\boldsymbol{h}}\mathbf{T}$ or $\pi(\bar{\boldsymbol{h}}\mathbf{T} + \mathbf{1})$. If the origin is assumed to be the centre of symmetry then $\mathbf{T} = 0$ and the permitted phase values are 0 and π. Then $F_{\boldsymbol{h}}$ will be a real positive number when $\phi_{\boldsymbol{h}}$ is equal to 0, and a negative one when $\phi_{\boldsymbol{h}}$ is equal to π. For this reason we usually talk in cs. space groups about the sign of the structure factor rather than about the phase.

In Fig. 1.12, $F_{\boldsymbol{h}}$ is represented as an Argand diagram for a centrosymmetric structure of six atoms. Since for each atom at $\boldsymbol{r}_j$ another symmetry equivalent atom exists at $-\boldsymbol{r}_j$, the contribution of every couple to $F_{\boldsymbol{h}}$. will have to be real.

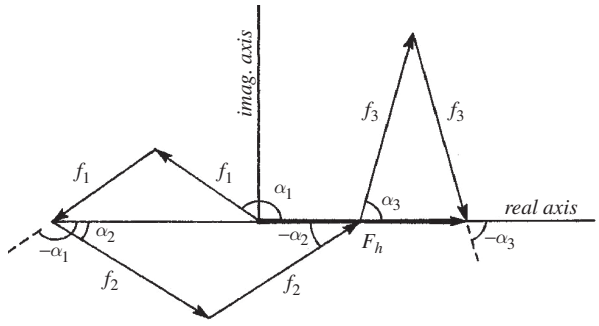


Fig. 1.12
$F_{\mathbf{h}}$ is represented in the Argand plane for a cs. crystal structure with $N = 6$. It is $\alpha_j = 2\pi\bar{\mathbf{H}}\mathbf{X}_j$.

Table 1.6 Restricted phase reflections for the 32 point groups

Point group	Sets of restricted phase reflections
1	None
$\bar{1}$	All
m	(0, *k*, 0)
2	(*h*, 0, *l*)
2/*m*	All
*mm*2	[(*h*, *k*, 0) masks (*h*, 0, 0), (0, *k*, 0)]
222	Three principal zones only
mmm	All
4	(*h*, *k*, 0)
$\bar{4}$	(*h*, *k*, 0); (0, 0, *l*)
4/*m*	All
422	(*h*, *k*, 0); {*h*, 0, *l*}; {*h, h, l*}
$\bar{4}$2*m*	[(*h*, *k*, 0), {*h*, *h*, 0}]; [{*h, 0, l*}, (0, 0, *l*)]
4*mm*	[(*h*, *k*, 0), {*h*, 0, 0}, {*h, h,* 0}]
4/*mmm*	All
3	None
$\bar{3}$	All
3*m*	{*h*, 0, $\bar{h}$, 0}
32	{*h*, 0, $\bar{h}$, *l*}
$\bar{3}$*m*	All
6	(*h*, *k*, 0)
$\bar{6}$	(0, 0, *l*)
6/*m*	All
$\bar{6}$*m*2	[{*h*, *h*, *l*}, {*h*, *h*, 0}, (0, 0, *l*)]
6*mm*	[(*h*, *k*, 0), {*h*, *h*, 0}, {*h*, 0, 0}]
62	(*h*, *k*, 0); (*h*, 0, *l*); (*h*, *h*, *l*)
6/*mmm*	All
23	{*h*, *k*, 0}
m$\bar{3}$	All
$\bar{4}$3*m*	[{*h*, *k*, 0}, {*h*, *h,* 0}]
432	{*h*, *k*, 0}; {*h*, *h*, *l*}
m$\bar{3}$*m*	All

Table 1.7 If $\bar{\boldsymbol{h}}\mathbf{R} = -\mathbf{h}$ the allowed phase values ϕ_a of $F_{\boldsymbol{h}}$ are $\pi\bar{\boldsymbol{h}}\mathbf{T}$ and $\pi\bar{\boldsymbol{h}}\mathbf{T} + \pi$. Allowed phases are multiples of 15° and are associated, in direct methods programs, with a symmetry code (SCODE). For general reflections SCODE = 1

$\phi_a^{(0)}$	SCODE
Any	1
(30,210)	3
(45,225)	4
(60,240)	5
(90,270)	7
(120,300)	9
(135,315)	10
(150,330)	11
(180,360)	13

As an example of a non-centrosymmetric (n.cs.) space group let us examine $P2_12_12_1$, $\left[(x, y, z), \left(\frac{1}{2} - x, \bar{y}, \frac{1}{2} + z\right), \left(\frac{1}{2} + x, \frac{1}{2} - y, \bar{z}\right), \left(\bar{x}, \frac{1}{2} + y, \frac{1}{2} - z\right)\right]$, where the reflections $(hk0)$, $(0kl)$, $(h0l)$ satisfy the relation $\bar{\boldsymbol{h}}\mathbf{R} = -\boldsymbol{h}$ for $\mathbf{R} = \mathbf{R}_2, \mathbf{R}_3, \mathbf{R}_4$ respectively. By introducing $\mathbf{T} = \mathbf{T}_2$ in equation (1.27) we obtain $\phi_{hk0} = (\pi h/2) + n\pi$. Thus ϕ_{hk0} will have phase 0 or π if h is even and phase $\pm\pi/2$ if h is odd. By introducing $\mathbf{T} = \mathbf{T}_3$ in equation (1.27) we obtain $\phi_{0kl} = (\pi k/2) + n\pi$: i.e. ϕ_{0kl} will have phase 0 or π if k is even and $\pm\pi/2$ if k is odd. In the same way, by introducing $\mathbf{T} = \mathbf{T}_4$ in equation (1.27) we obtain $\phi_{h0l} = (\pi l/2) + n\pi$: i.e. ϕ_{h0l} will have phase 0 or π if l is even and $\pm\pi/2$ if l is odd. In Table 1.6 the sets of restricted phase reflections are given for the 32 point groups.

The allowed values of restricted phases depend on the translational component of the symmetry element and on its location with respect to the cell origin. For conventional three-dimensional space groups the allowed phase values are multiples of 15°. In Table 1.7 the different types of phase restriction are shown: in the second column the characteristic

codes associated in direct methods programs with the various restrictions are quoted. It should not be forgotten that symmetry equivalent reflections can have different allowed phase values. For example, in the space group $P4_12_12\left[(x,\ y,\ z);\left(\bar{x},\ \bar{y},\ z+\frac{1}{2}\right);\left(\bar{y}+\frac{1}{2},\ x+\frac{1}{2},\ z+\frac{1}{4}\right);\left(y+\frac{1}{2},\ \bar{x}+\frac{1}{2},\ z+\frac{3}{4}\right);\left(\bar{x}+\frac{1}{2},y+\frac{1}{2},\bar{z}+\frac{1}{4}\right);\left(x+\frac{1}{2},\bar{y}+\frac{1}{2},\bar{z}+\frac{3}{4}\right);(y,x,\bar{z});\left(\bar{y},\bar{x},\bar{z}+\frac{1}{2}\right)\right]$, the reflection (061) has phase values restricted to $(-(\pi/4),\ 3\pi/4)$. Its equivalent reflections are also symmetry restricted, but the allowed phase values may be different from $(-(\pi/4),\ 3\pi/4)$. On the assumption that $\phi_{061} = 3\pi/4$, the reader will find for the equivalent reflections the phase restrictions shown in Fig. 1.13.

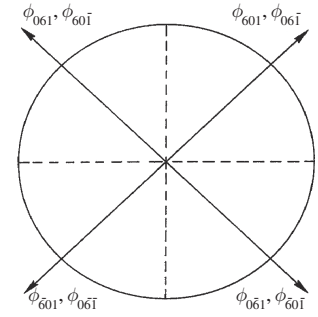


Fig. 1.13
Phase restrictions for the reflection (061) and its symmetry equivalents.

1.5.4 Systematic absences

Let us look for the class of reflections for which $\bar{\boldsymbol{h}}\boldsymbol{R} = \bar{\boldsymbol{h}}$ and apply equation (1.23) to them. This relation would be violated for those reflections for which $\bar{\boldsymbol{h}}\mathbf{T}$ is not an integer number unless $|\mathbf{F}_{\boldsymbol{h}}| = 0$. From this fact the rule follows: reflections for which $\bar{\boldsymbol{h}}\mathbf{R} = \bar{\boldsymbol{h}}$ and $\bar{\boldsymbol{h}}\mathbf{T}$ is not an integer will have diffraction intensity zero or, as is usually stated, will be systematically absent or extinct. Let us give a few examples.

In the space group $P2_1\left[(x,y,z),\left(\bar{x},y+\frac{1}{2},\bar{z}\right)\right]$, the reflections $(0k0)$ satisfy the condition $\bar{\boldsymbol{h}}\mathbf{R}_2 = \bar{\boldsymbol{h}}$. If k is odd, $\bar{\boldsymbol{h}}\boldsymbol{T}_2$ is semi-integer. Thus, the reflections $(0k0)$ with $k \neq 2n$ are systematically absent.

In the space group $P4_1\left[(x,y,z),\left(\bar{x},\bar{y},\frac{1}{2}+z\right),\left(\bar{y},x,\frac{1}{4}+z\right),\left(y,\bar{x},\frac{3}{4}+z\right)\right]$, only the reflections $(00l)$ satisfy the condition $\boldsymbol{h}\mathbf{R}_j = \boldsymbol{h}$ for $j = 2,3,4$. Since $\bar{\boldsymbol{h}}\mathbf{T}_2 = l/2$, $\bar{\boldsymbol{h}}\mathbf{T}_3 = l/4$, $\bar{\boldsymbol{h}}\mathbf{T}_4 = 3l/4$, the only condition for systematic absence is $l \neq 4n$, with n integer.

In the space group $Pc\left[(x,y,z),\left(x,\bar{y},z+\frac{1}{2}\right)\right]$, the reflections $(h0l)$ satisfy the condition $\bar{\boldsymbol{h}}\mathbf{R}_2 = \bar{\boldsymbol{h}}$. Since $\bar{\boldsymbol{h}}\mathbf{R}_2 = l/2$, the reflections $(h0l)$ with $l \neq 2n$ will be systematically absent.

Note that the presence of a slide plane imposes conditions for systematic absences to bidimensional reflections. In particular, slide planes opposite to $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ impose conditions to classes $(0kl)$, $(h0l)$, and $(hk0)$ respectively. The condition will be $h = 2n$, $k = 2n$, $l = 2n$ for the slide planes of type a, b, or c respectively.

Let us now apply the same considerations to the symmetry operators centring the cell. If the cell is of type A, B, C, I, symmetry operators will exist whose rotational matrix is always the identity, while the translational matrices are

$$\mathbf{T}_{\mathrm{A}} = \begin{bmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} \quad \mathbf{T}_{\mathrm{B}} = \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix} \quad \mathbf{T}_{\mathrm{C}} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix} \quad \mathbf{T}_{\mathrm{I}} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

respectively. If we use these operators in equation (1.24), we find that (1) the relation $\bar{\boldsymbol{h}}\mathbf{R} = \bar{\boldsymbol{h}}$ is satisfied for any reflection and (2) the systematic absences,

of three-dimensional type, are $k + l = 2n$, $h + l = 2n$, $h + k = 2n$, $h + k + l = 2n$, respectively.

A cell of type F is simultaneously A-, B-, and C-centred, so the respective conditions for systematic absences must be simultaneously valid. Consequently, only the reflections for which h, k, and l are all even or all odd will be present.

The same criteria lead us to establish the conditions for systematic absences for rhombohedral lattices ($-h + k + l \neq 3n$ for obverse setting and $h - k + l \neq 3n$ for reverse setting). The list of systematic absences for any symmetry element is given in Table 1.8.

Table 1.8 Systematic absences

Symmetry elements		Set of reflections	Conditions
Lattice	P	*hkl*	None
	I		$h + k + l = 2n$
	C		$h + k = 2n$
	A		$k + l = 2n$
	B		$h + l = 2n$
	F		$\begin{cases} h + k = 2n \\ k + l = 2n \\ h + l = 2n \end{cases}$
	R_{obv}		$-h + k + l = 3n$
	R_{rev}		$h - k + l = 3n$
Glide plane ∥ (001)	*a*	*hk*0	$h = 2n$
	b		$k = 2n$
	n		$h + k = 2n$
	d		$h + k = 4n$
Glide plane ∥ (100)	*b*	0*kl*	$k = 2n$
	c		$l = 2n$
	n		$k + l = 2n$
	d		$k + l = 4n$
Glide plane ∥ (010)	*a*	*h*0*l*	$h = 2n$
	c		$l = 2n$
	n		$h + l = 2n$
	d		$h + l = 4n$
Glide plane ∥ (110)	*c*	*hhl*	$l = 2n$
	b		$h = 2n$
	n		$h + l = 2n$
	d		$2h + l = 4n$
Screw axis ∥ *c*	$2_1, 4_2, 6_3$	00*l*	$l = 2n$
	$3_1, 3_2, 6_2, 6_4$		$l = 3n$
	$4_1, 4_3$		$l = 4n$
	$6_1, 6_5$		$l = 6n$
Screw axis ∥ *a*	$2_1, 4_2$	*h*00	$h = 2n$
	$4_1, 4_3$		$h = 4n$
Screw axis ∥ *b*	$2_1, 4_2$	0*k*0	$k = 2n$
	$4_1, 4_3$		$k = 4n$
Screw axis ∥ [110]	2_1	*hh*0	$h = 2n$

1.6 The basic postulate of structural crystallography

In the preceding paragraphs we have summarized the basic relations of general crystallography: these can be found in more extended forms in any standard textbook. The reader is now ready to learn about the topic of phasing, one of the most intriguing problems in the history of crystallography. We will start by illustrating its logical aspects (rather than its mathematics) via a short list of questions.

Given a model structure, can we calculate the corresponding set (*say* $\{|F_{\mathbf{h}}|\}$) *of structure factor moduli?* The answer is trivial; indeed we have only to introduce the atomic positions and the corresponding scattering factors (including temperature displacements) into equation 1.19. As a result of these calculations, moduli and phases of the structure factors can be obtained. It may therefore be concluded that there is no logical or mathematical obstacle to the symbolic operation

$$\rho(\mathbf{r}) \Rightarrow \{|F_{\mathbf{h}}|\}\,.$$

A second question is: *given only the structure factor moduli*, *can we entertain the hope of recovering the crystal structure*, *or*, *on the contrary*, *is there some logical impediment to this* (*for example*, *an irrecoverable loss of information*)? In symbols, this question deals with the operation

$$\{|F_{\mathbf{h}}|\} \Rightarrow \rho(\mathbf{r}). \tag{1.28}$$

As an example, let us suppose that the diffraction experiment has provided 30 000 structure factor moduli and lost 30 000 phases. Can we recover the 30 000 phases given the moduli, and consequently determine the structure, or are the phases irretrievably lost?

A first superficial answer may be provided by our daily experience. To give a simple example, if we are looking for a friend in New York but we have lost his address, it would be very difficult to find him. This allegorical example is appropriate as in New York there are millions of addresses, similarly, millions of structural models may be conceived that are compatible with the experimental unit cell. The search for our friend would be much easier if some valuable information were still in our hands: e.g. he lives in a flat on the 130th floor. In this case we could discard most of the houses in New York. But where, in the diffraction experiment, is the information hidden which can allow us to discard millions of structural models and recover the full structure?

A considered answer to the problem of phase recovery should follow reference to modern structural databases (see Section 1.7). In Fig. 1.14 statistics are shown from the *Cambridge Structural Database*, where the growth in numbers of deposited structures per year is shown. Hundreds of thousands of crystal structures have been deposited, the large majority of these having been solved starting from the diffraction moduli only. In Fig. 1.15, similar statistics are shown for the *Protein Data Bank* (*PDB*).

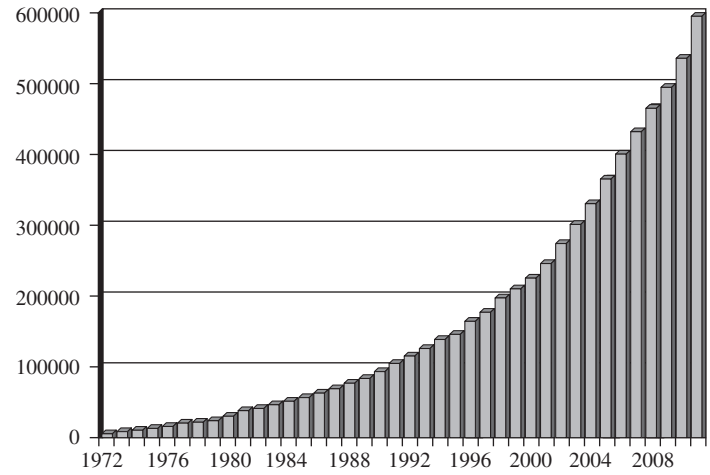


Fig. 1.14
Cumulative growth per year of the structures deposited in the *Cambridge Structural Database* (*CSD*).

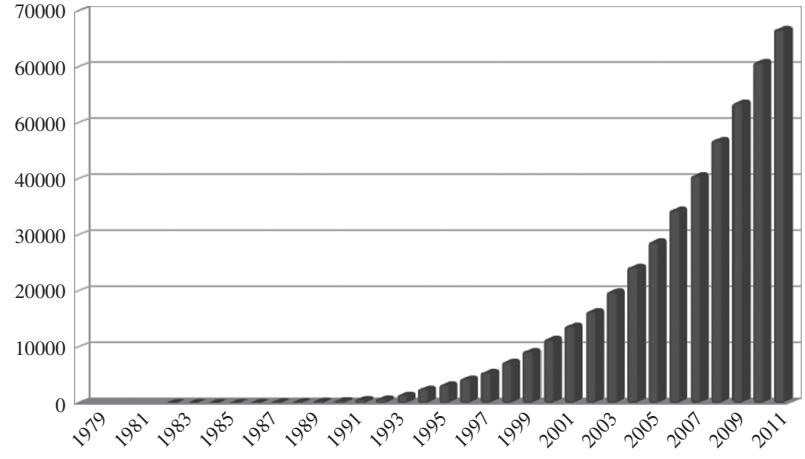


Fig. 1.15
Growth per year of the structures deposited in the *Protein Data Bank* (*PDB*).

Such huge numbers of structures could not have been solved without valuable information provided by experiment and since X-ray experiments only provide diffraction amplitudes we have to conclude that the phase information is hidden in the amplitudes. But at the moment we do not know how this is codified.

Before dealing with the code problem, we should answer a preliminary question: *how can we decide* (*and accept*) *that such huge numbers of crystal structures are really* (*and correctly*) *solved*? Each deposited structure is usually accompanied by a *cif* file, where the main experimental conditions, the list of the collected experimental data, their treatment by crystallographic programs, and the structural model are all described. Usually residuals such as (Booth, 1945)

$$R_{cryst} = \frac{\sum_{\mathbf{h}} ||F_{obs} - |F_{calc}||}{\sum_{\mathbf{h}} |F_{obs}|} \tag{1.29}$$

are mentioned as mathematical proof of the correctness of a model: if R_{cryst} is smaller than a given threshold and no crystal chemical rule is violated by the proposed model, then the model is assumed to be correct. This assumption is universally accepted, and is the basic guideline for any structural crystallographer, even though it is not explicitly formulated and not demonstrated mathematically. But, how can we exclude two or more crystal structures which may exist, which do not violate well-established chemical rules, and fit the same experimental data? A postulate should therefore be evoked and legitimized, in order to allow us to accept that a crystal structure is definitively solved: this is what we call the basic postulate of structural crystallography.

The basic postulate of structural crystallography: *only one chemically sound crystal structure exists that is compatible with the experimental diffraction data*.

Before legitimizing such a postulate mathematically a premise is necessary: the postulate is valid for crystal structures, that is, for structures for which chemical (i.e. the basic chemical rules) and physical constraints hold. Among physical constraints we will mention atomicity (the electrons are not dispersed in the unit cell, but lie around the nucleus) and positivity (i.e. the electron density is non-negative everywhere). The latter two conditions are satisfied if X-ray data and, by extension, electron data (electrons are sensible to the potential field) are collected: the positivity condition does not hold for neutron diffraction, but we will see that the postulate may also be applied to neutron data.

Let us now check the postulate by using the non-realistic four-atom one-dimensional structure shown in Fig. 1.9a: we will suppose that the chosen interatomic distances comply with the chemistry (it is then a *feasible model*). In Fig. 1.16a–c three electron densities are shown at 0.9 Å resolution, obtained by using, as coefficients of the Fourier series (1.18), the amplitudes of the true structure combined with random phases. All three models, by construction, have the same diffraction amplitudes ($R_{cryst} = 0$ for such models), but only one, that shown in Fig. 1.9a, satisfies chemistry and positivity–atomicity postulates. All of the random models show positive peaks (say potential atoms) in random positions, there are always a number of negative peaks present, and the number of positive peaks may not coincide with the original structure. Any attempt to obtain other feasible models by changing the phases in a random way will not succeed: this agrees well with the postulate.

A more realistic example is the following (structure code *Teoh*, space group *I-4*, $C_{42}\ H_{40}\ O_6\ Sn_2$). Let us suppose that the crystallographer has requested his phasing program to stop when a model structure is found for which $R_{cryst} < 0.18$ and that the program stops, providing the model depicted in Fig. 1.17a, for which $R_{cryst} = 0.16$. This model, even if it is further refinable up to smaller values of R_{cryst}, has to be rejected because it is chemically invalid, even if the crystallographic residual is sufficiently small. If the crystallographer asks the phasing program to stop only when a model is found for which $R_{cryst} < 0.10$, then the model shown in Fig. 1.17b is obtained, for which $R_{cryst} = 0.09$. This new model satisfies basic crystal chemical rules and may be further refined.

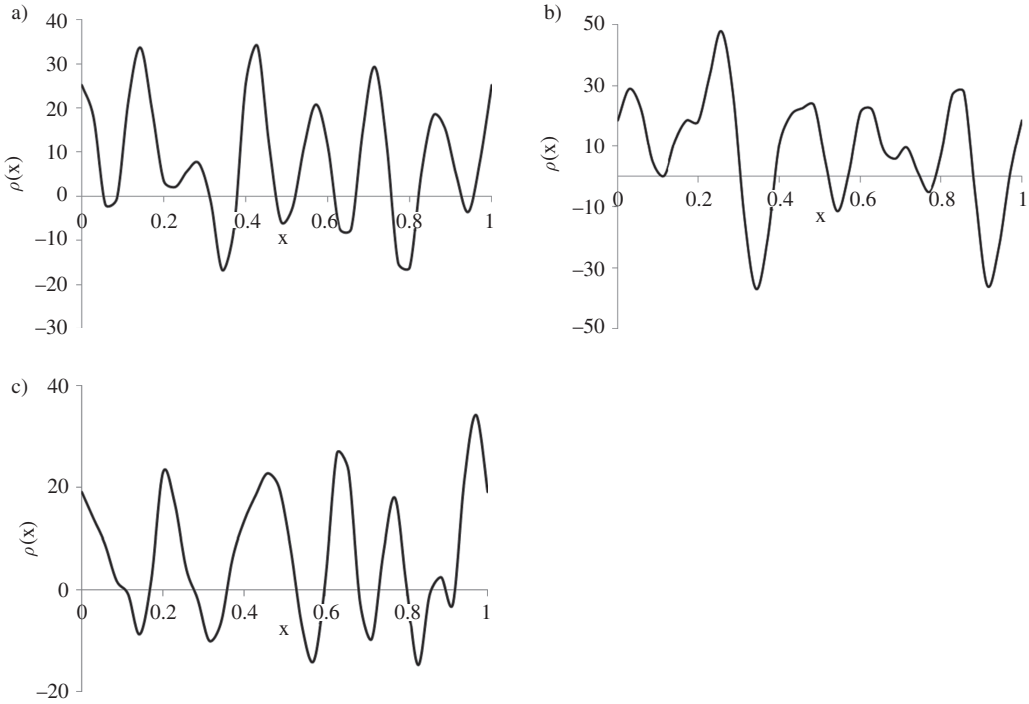


Fig. 1.16
For the four-atom one-dimensional structure shown in Fig. 1.9a, three models, obtained using random phases, are shown. Data resolution: 0.9 Å.

The above results lead to a practical consequence: even if experimental data are of high quality, and even if there is very good agreement between experiment and model (i.e. a small value of R_{cryst}), *structure validation* (i.e. the control that the basic crystal chemical rules are satisfied by the model) *is the necessary final check of the structure determination process*. Indeed it is an obligatory step in modern crystallography, a tool for a posteriori confirmation of the basic postulate of crystallography.

The basic postulate may be extended to neutron data, but now the positivity condition does not hold: it has to be replaced by the chemical control and validation of the model, but again, there should not exist two chemically sound crystal structures which both fit high quality experimental data.

In order to legitimize the basic postulate of structural crystallography mathematically, we now describe how the phase information is codified in the diffraction amplitudes. We observe that the modulus square of the structure factor, say

$$|F_{\mathbf{h}}|^2 = F_{\mathbf{h}} \cdot F_{-\mathbf{h}} = \sum_{j=1}^{N} f_j \exp\left(2\pi i\, \mathbf{h} \cdot \mathbf{r}_j\right) \cdot \sum_{j=1}^{N} f_j \exp\left(-2\pi i\, \mathbf{h} \cdot \mathbf{r}_j\right)$$

$$= \sum_{j\,1,j\,2=1}^{N} f_{j1} f_{j\,2} \exp\left[2\pi i\, \mathbf{h} \cdot (\mathbf{r}_{j1} - \mathbf{r}_{j\,2})\right] \tag{1.30}$$

depends on the interatomic distances: inversely, the set of interatomic distances defines the diffraction moduli. If one assumes that only a crystal structure exists with the given set of interatomic distances, the obvious conclusion should be that only one structure exists (except for the enantiomorph structure) which is compatible with the set of experimental data, and vice versa, only one set of diffraction data is compatible with a given structure. In symbols

$$\text{crystal structure} \Leftrightarrow \left\{\mathbf{r}_i - \mathbf{r}_j\right\} \Leftrightarrow \{|F_{\mathbf{h}}|\}\,. \tag{1.31}$$

This coincides exactly with the previously defined basic postulate.

The conclusion (1.31), however, must be combined with structure validation, as stated in the basic postulate. Indeed Pauling and Shapell (1930) made the observation that for the mineral bixbyite there are two different solutions, not chemically equivalent, with the same set of interatomic vectors. Chemistry (i.e. structure validation) was invoked to define the correct structure. Patterson (1939, 1944) defined these kinds of structure as *homometric* and investigated the likelihood of their occurrence. Hosemann and Bagchi (1954) gave formal definitions of different types of homometric structures. Further contributions were made by Buerger (1959, pp. 41–50), Bullough (1961, 1964), and Hoppe (1962a,b). In spite of the above considerations it is common practice for crystallographers to postulate, for structures of normal complexity, a biunique correspondence between the set of interatomic vectors and atomic arrangement. Indeed for almost the entire range of the published structures, two different *feasible* (this property being essential) structures with the same set of observed moduli has never been found.

Some care, however, is necessary when the diffraction data are not of high quality and/or some pseudosymmetry is present. Typical examples of structural ambiguity are:

(a) The low quality of the crystal (e.g. high mosaicity), or the disordered nature of the structure. In this case the quality of the diffraction data is depleted, and therefore the precision of the proposed model may be lower.
(b) The structure shows a symmetry higher than the real one. For example, the structure is very close to being centric but it is really acentric, or it shows a strong pseudo tetragonal symmetry but it is really orthorhombic. Deciding between the two alternatives may not be easy, particularly when the pseudosymmetry is very close to crystal symmetry and data quality is poor.
(c) Strong pseudotranslational symmetry is present. This occurs when a high percentage of electron density satisfies a translational vector $\boldsymbol{u}$ smaller than that allowed by the crystal periodicity: for example, if $\boldsymbol{u} = \boldsymbol{a}/3$ and 90% of the electron density is invariant under the pseudotranslation. In this case reflections with h = 3n are very strong, the others are very weak. If only substructure reflections are measured, the substructure only is defined (probably with a quite good R_{cryst} value), but the fine detail of the structure is lost.

In all of the cases a–c the final decision depends on the chemistry and on the fit between model and observations. To give a general view of what the fit means

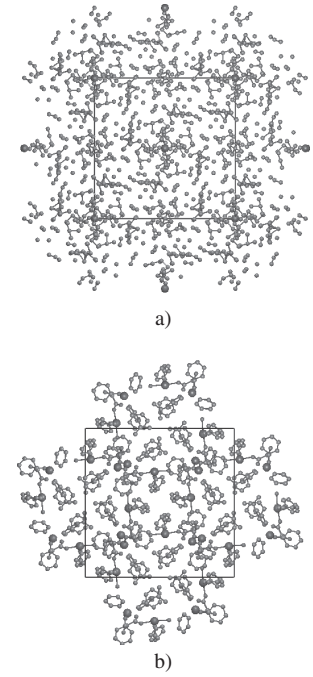


Fig. 1.17
Teoh: (a) false structural model with $R_{cryst} = 0.16$; (b) correct structural model with $R_{cryst} = 0.08$.

Table 1.9 Statistics on R_{cryst} for structures deposited in the *Cambridge Structural Database* up to 1 January 2012. For each range ΔR_{cryst} in which R_{cryst} lies, *Nstr* and % are the corresponding number of structures and percentage, respectively

ΔR_{cryst}	*Nstr*	%
0.01–0.03	62 774	10.5
0.03–0.04	122 706	20.6
0.04–0.05	135 525	22.7
0.05–0.07	163 269	27.4
0.07–0.09	60 651	10.2
0.09–0.10	13 370	2.2
0.10–0.15	18 353	3.1
0.15–. . . .	3835	0.6

numerically today, we report in Table 1.9 statistics on the crystallographic residual R_{cryst} performed over the structures deposited in the *Cambridge Structural Database* up to January 2012. We see that the precision of the structural determination may vary over a wide range: indeed R_{cryst} values are found between 0.01 and more than 0.1, and this wide range is often due to the different quality of the crystals. For the large majority of structures, even those with a relatively high value of R_{cryst}, the structure is uniquely fixed in all details, eventually with limited precision in unit cell regions where structural disorder is present. These details, however, do not destroy the general validity of the basic postulate.

The basic postulate of structural crystallography should be considered by any rational crystallographer before initiating their daily structural work. This may be further summarized as follows: *in a diffraction experiment the phase information is not lost, it is only hidden within the diffraction amplitudes. Accordingly, any phasing approach is nothing else but a method for recovering the hidden phases from the set of diffraction amplitudes.*

Let us now suppose that the basic postulate is consciously considered by our young crystallographer. A further problem then arises: is the amount of information stored in the diffraction amplitudes sufficient to define the structure? For example, in the case of a low resolution diffraction experiment the crystallographic data may not be sufficient to define the short interatomic distances, making it impossible, therefore, to uniquely define the structure. This is a crucial problem for structural crystallography, since the crystal structure solution may depend on the amount of information provided by the diffraction experiment. *What then are the resolution limits for a useful diffraction experiment*?

Suppose we have a crystal with *P1* symmetry: let N be the number of non-hydrogen atoms in the unit cell, and $N_{sp} = 4N$ the number of structural parameters necessary for defining the structure (four parameters per atom, say *x, y, z* and the corresponding isotropic thermal factor). For a small- or medium-sized molecule, $V = k\ N$, where k is usually between 15.5 and 18.5; for a protein, owing to the presence of the solvent, k may be significantly larger, up to or even exceeding 40. According to equation (1.8), $V^* = V^{-1} = (kN)^{-1}$.

Let us suppose that a diffraction experiment provides data up to $r^*_{\max}$, or, equivalently, up to $d_{\min}$. The number of measurable reflections (say N_{ref}) may be calculated as follows. The reciprocal space measured volume may be parameterized as

$$\Phi^*_{meas} = \frac{4}{3}\pi \left(r^*_{\max}\right)^3 = \frac{4\pi}{3d^3_{\min}},$$

and

$$N_{ref} = \frac{\Phi^*_{meas}}{V^*} = \frac{4\pi}{3d^3_{\min}} kN.$$

Let us now estimate the index,

$$R_{\inf} = \textit{ratio between the experimental information and the structural complexity}. \tag{1.32}$$

When no prior supplemental information is available besides experimental data, R_{inf} may be qualitatively approximated as follows:

$$R_{\text{inf}} = \textit{number of measured symmetry independent reflections/number of structural parameters} \tag{1.33}$$

To compute R_{inf}, the Friedel law should be taken into account: thus we divide N_{ref} by 2 and then write the resulting expression for R_{inf} :

$$R_{\text{inf}} \approx \frac{\pi}{6d_{\text{min}}^3}k.$$

The numerical values of R_{inf} for specific values of k and d_{min} are shown in Table 1.10: larger values of R_{inf} correspond with cases in which the structure is overdetermined by the observations, while small values of R_{inf} do not uniquely fix the structure. Let us suppose, just as a rule of thumb, that a structure may be solved, from diffraction data only, if $R_{\text{inf}} \geq 3$: Table 1.10 suggests that $d_{\text{min}} \approx 1.4$ Å is the resolution threshold below which a small molecule structure cannot be solved ab initio. The threshold moves to ≈ 1.6 Å for a protein with a small percentage of solvent, and to ≈ 1.8 Å for a protein with a larger solvent percentage.

Table 1.10 R_{inf} in *P1* is shown for some values of d_{min} and k. $k = 17$ is representative of the small- to medium-sized structures, $k = 25$, 35, of the proteins

d_{min}	$k = 17$	$k = 25$	$k = 35$
0.4	139	204	286
0.6	41	60.6	84.7
0.8	17.4	25.6	35.7
1.0	8.9	13.1	18.3
1.4	3.2	4.8	13.7
1.8	1.5	2.2	3.1
2.2	0.8	1.2	1.7
2.5	0.6	0.8	1.2
3.0	0.3	0.5	0.7
3.5	0.2	0.3	0.4
4.0	0.1	0.2	0.3

The conclusion is that the solvent is a valuable source of information: the larger the solvent, the higher the threshold for the ab initio crystal structure solution (modern solvent flattening techniques are able to efficiently exploit this information). A special case occurs when one is interested in solving a substructure, for example the heavy-atom substructure in SIR-MIR cases and the anomalous scatterer substructure in SAD-MAD cases. If it is supposed that the structure factor amplitudes of such substructures are estimated with reasonable approximation, then the atoms belonging to the substructure are dispersed in a big empty space (i.e. the unit cell of the structure). In this case the estimated structure factor amplitudes of the substructure overdetermine it, and the substructure could be solved even at very low resolution (worse than 3.5 Å).

The above conclusions do not change significantly if the space group has symmetry higher than triclinic. Indeed in this case R_{inf} is the ratio between the number of unique reflections and the number of structural parameters corresponding to the symmetry independent atoms.

Additional difficulties with the phasing process arise when experimental data quality is poor. If there are errors in the diffraction amplitudes, since information on the phases is hidden within the amplitudes, such errors will inevitably cause a deterioration in the efficiency of any phasing procedure. This is particularly important in the case of powder data (see Chapter 12) and also electron data (see Chapter 11), but it is also important for proteins, because the presence of the solvent implies disordered regions in the unit cell and therefore limited data resolution.

So far we have answered the question: under what conditions is a structure univocally fixed from its diffraction data? We have skipped cases where some previous additional information is available; here, the number of measured symmetry independent reflections in the numerator of R_{inf} is only part of the total information available and therefore the conclusions drawn from

Table 1.10 must be corrected. In this book we will consider four cases in which additional information is present:

1. *Non-crystallographic symmetry*. This is an important source of information which permits a reduction in the number of structural parameters in equation 1.33. It occurs when there are more identical molecules in the asymmetric unit: in this case they may be defined in terms of one molecule by applying the local symmetry operators. *Non-crystallographic symmetry* allows the structural solution of large biological assemblies such as viruses.
2. *Molecular replacement*. A model molecule, similar geometrically to that under investigation, is available.
3. *Isomorphous derivatives*. Diffraction data for the target and one or more isomorphous structures are measured.
4. *Anomalous dispersion data*. Diffraction data with anomalous dispersion effects are collected (we will see that this case is similar to case 3).

Because of the additional experimental information available, the value of R_{inf} increases substantially which allows structure solution even at data resolutions larger than 4 Å.

1.7 The legacy of crystallography

Human beings periodically visit museums to enjoy the artistic masterpieces exhibited in witness of human sensitivity to beauty. Historical and technical museums are often consulted in relation to their acquaintance with the evolution of human civilization andwith man′s capacity for improving human life through technical innovations. But, where can the products of crystallography be consulted, in witness of its immense legacy to chemistry, physics, mineralogy, and biology?

Over a period of about one century crystallographic phasing methods have solved a huge number of crystal structures, so enriching our understanding of the mineral world, of organic, metallorganic, and inorganic chemistry, and of the bio-molecules. This enormous mine of information is stored in dedicated databases, among which are the following.

1. The *Cambridge Structural Database* (*CSD*), <http://www.ccdc.cam.ac.uk/products/csd/>, where chemical and crystallographic information for organic molecules and metal–organic compounds determined by X-ray or neutron diffraction: powder diffraction studies are deposited.
2. *Inorganic Crystal Structure Database* (*ICSD*), <http://www.fiz-karlsruhe.de/icsd_content.html>, where structural data of pure elements, metals, minerals and intermetallic compounds are deposited. By January 2012 it contained more than 150 000 entries, 75.6% of them with a structure having been assigned.
3. *CRYSTMET*, <http://www.tothcanada.com/>, where structural information on metals and alloys are stored.

Table 1.11 *CSD* entries on 1 January 2012: Nr is the number of entries, % the corresponding percentage over the total

	Nr	%
Total number of structures	596 810	100
Number of compounds	544 565	
Organic compounds	254 475	42.6
Transition metal present	319 188	53.5
Neutron studies	1534	0.3
Powder diffraction studies	2354	0.4

4. *Protein Data Bank* (*PDB*), <http://www.rcsb.org/pdb/>, with about 75 056 entries up to January 2012.
5. *Nucleic Acid Database* (*NDB*), <http://ndbserver.rutgers.edu/>, oligonucleotide structures deposited up to April 2012.

For each structure deposited, an archive typically contains details of the structure solution, citation information, the list of atoms and their coordinates; the structure can be visualized and displayed on the user's computer. In this section we report some statistics on the entries in two of these databases, the *CSD* for small molecules and the *PDB* for macromolecules, in order to provide the reader with some essential information on some of the parameters to which phasing methods are sensitive. For example, which type of radiation is more useful in standard conditions, which are the most frequent space groups or crystal systems, how data resolution is distributed among the deposited structures, etc. This type of information is shown in Tables 1.11 to 1.16. It should be noted that:

(a) Tables 1.11 and 1.14 provide the numbers of deposited structures for small molecules and macromolecules, respectively. Also given is information on the type of radiation used for their solution. The tables justify the special attention we are giving to X-ray diffraction.
(b) Tables 1.13 to 1.16 suggest which are the most frequent space groups, for both small and large molecules. The reader should remember that these are expected to be very different for the two categories: indeed centric space groups and, in general, groups with inversion axes, are not allowed for proteins.

Table 1.12 *CSD*: crystallographic system statistics

System	%
Triclinic	24.7
Monoclinic	52.5
Orthorhombic	18.0
Tetragonal	2.2
Trigonal	1.7
Hexagonal	0.5
Cubic	0.5

Table 1.13 *CSD*: the 10 most frequent space groups

Space group	%
$P2_1/c$	34.9
P-1	23.8
$C2/c$	8.2
$P2_12_12_1$	7.6
$P2_1$	5.3
$Pbca$	3.5
$Pna2_1$	1.6
$Pnma$	1.2
Cc	1.1
P1	1.0

Table 1.14 *PDB*: entries for proteins, nucleic acids, and protein/NA complexes, according to experimental technique

	Proteins	Nucleic acids	Protein/NA complexes
Total	75056	2360	3609
X-ray	66381	1352	3298
NMR	8206	979	186
Electron microscopy	285	22	118

Table 1.15 *PDB*: distribution of data resolution for the deposited structures, in Å

Å	Nr	%
0.5–1	485	0.68
1.0–1.5	6134	8.63
1.5–2.0	27 385	38.53
2.0–2.5	22 144	31.16
2.5–3.0	11 210	15.77
3.0–3.5	2930	4.12
3.5–4.0	602	0.85