*International Tables for Crystallography* (2012). Vol. F, Section 18.4.5.8, pp. 494–495.

structure by applying the appropriate scale to the solvent contribution:

$$F_{\rm total} = F_{\rm prot} + k_m \exp(-B_m s^2/4) F_{\rm mask}, \qquad (18.4.5.2)$$

where $k_m$ and $B_m$ are the scale and temperature factor for the solvent contribution, and $F_{\rm total}$, $F_{\rm prot}$ and $F_{\rm mask}$ are the complex structure factors corresponding to the combined contribution from the protein and solvent region. The scale and $B$ values are usually refined iteratively to find the best match to the observed structure factors, and are implemented in *CNS*, *REFMAC* and *PHENIX.REFINE*.

Nevertheless, there remain severe problems in the modelling of the interface. The border between the two regions is not abrupt, as there is a smooth and continuous change from the region with fully occupied, discrete sites to one which is truly fluid, but this passes through a volume with an increasing level of dynamic disorder and associated partial occupancy. Modelling of this region poses major problems, as described above, and the definition of disordered sites with low occupancy remains difficult even at atomic resolution. At which stage the occupancy and associated ADP can be defined with confidence is not yet an objective decision. In addition, refinement and modelling at this level of detail is very time consuming in terms of human intervention.

### 18.4.5.7. *Metal ions and other ligands in the solvent*

In general, proteins are crystallized from aqueous solutions which contain various additives, such as anions or cations (especially metals), organic solvents, including those used as cryoprotectants, and other ligands. Some of these may bind in specific or indeed non-specific sites in the ordered solvent shell, in addition to any functional binding sites of the protein. To identify such entities at limited resolution is often impossible, as the range of expected ADPs is large and there is very poor discrimination in the appearance of such sites and of water in the electron density. Atomic resolution assists in resolving ambiguities, as the interatomic distances, ADPs and occupancies are all better defined.

For metal ions, two additional criteria can be invoked. Firstly, the coordination geometry, with well defined bond lengths and angles, provides an indication of the identity of the ion, as different metals have different preferred ligand environments (Harding, 1999, 2006). The bond-valence approach is also applicable (Müller *et al.*, 2003). In addition, the value of the refined ADP and/or occupancy is helpful. Secondly, the anomalous signal in the data should reveal the presence of metal and some other non-water sites in the solvent through computation of the anomalous difference synthesis (Dauter & Dauter, 1999). This emphasizes the need to retain the anomalous signal during the collection and reduction of native data. While these approaches can be applied at lower resolution, they both become much more powerful at atomic resolution.

The presence of bound organic ligands has become especially relevant since the advent of cryogenic freezing. Compounds such as ethylene glycol and glycerol possess a number of functional hydrogen-bonding groups that can attach to sites on the protein in a defined way. Indeed, these may often bind in the active sites of enzymes such as glycosyl hydrolases, where they mimic the hydroxyl groups of the sugar substrate. It is most important to identify such moieties properly, particularly if substrate studies are to be planned successfully.

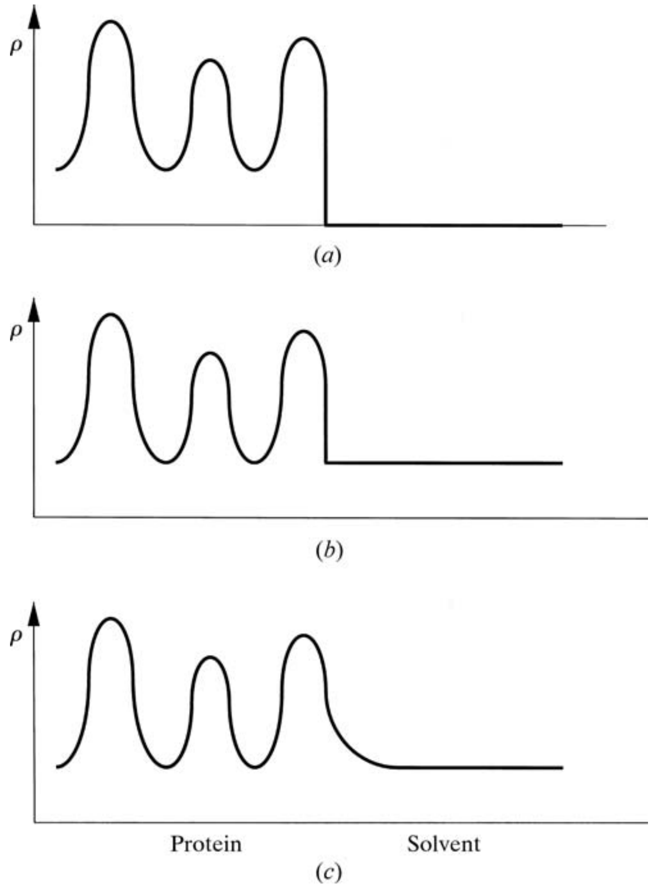

**Figure 18.4.5.2**
Schematic representation of the bulk-solvent models described in the text. (*a*) No bulk-solvent correction, *i.e.* solvent density set to zero. (*b*) Constant level of solvent outside the macromolecule and ordered water envelope. Here, sharp edge effects remain. (*c*) The model as in (*b*), but smoothed at the edge of a macromolecule, equivalent to the application of a *B* value to the solvent model.

### 18.4.5.8. *Deformation density*

X-ray structures are generally modelled using the spherical-atom approximation for the scattering, which ignores the deviation from sphericity of the outer bonding and lone-pair electrons. Extensive studies over a long period have confirmed that the so-called deformation density, representing deviation from this spherical model, can be determined experimentally using data to very high resolution, usually from 0.8 to 0.5 Å. An excellent review of this field has been provided by Coppens (1997). The observed deviations can be compared with those expected from the available theories of chemical bonding and the densities derived therefrom.

The application of atomic resolution analysis to proteins allowed the observation of the deformation density in macromolecules (Lamzin *et al.*, 1999). Data for two proteins were analysed: crambin (molecular weight 6 kDa) at 0.67 Å resolution and a subtilisin (molecular weight 30 kDa) at 0.9 Å resolution. Significant and interpretable deformation density could not be observed for the individual residues. However, on averaging the density over 40 peptide units for crambin and more than 250 for the subtilisin, the deformation density within the peptide unit was clearly visible and could be related to the expected bonding features in these units. This shows the real power of atomic resolution crystallography, which can reveal features containing no more than 0.2 e Å$^{-3}$.

Deformation density studies are now being applied to many polypeptides (Jelsch *et al.*, 1998; Koritsanszky *et al.*, 2002; Pichon-

Pesme *et al.*, 2004; Afonine *et al.*, 2007; Zarychta *et al.*, 2007). It has been observed that details of deformation densities are most clearly revealed when only the highest resolution (so-called high order) terms are included in the refinement and the Fourier maps (Coppens, 1997). This is reported to result from the deconvolution of the effects of the anisotropic ADPs, which can to some extent take up the fine features corresponding to the deformation density (bonding electrons and lone pairs). After proper modelling of the deformation density features, the overall model is refined against all the data.

### 18.4.6. Quality assessment of the model

The refinement of proteins at resolutions lower than atomic depends upon the use of restraints on the geometry and ADPs. The almost exclusively used library of target geometric restraints for refinement and validation of protein structures (Chapter 18.3) is derived from structures of amino acids and peptides in the Cambridge Crystallographic Data Centre's small-molecule crystal structure database (Allen *et al.*, 1979). Stereochemical parameters, such as conformational angles $\varphi$, $\psi$, should ideally not be restrained, as they allow independent validation of the model. As stated in Section 18.4.4.2, these restraints are required even at atomic resolution to maintain the chemical integrity of flexible regions, although their impact will be limited on ordered regions.

Owing to the excess of accurate X-ray observations over parameters at atomic resolution, extensive validation of individual structures should be less challenging than for those at lower resolution. It is hard to achieve an $R$ factor around 10% with an incorrect model. However, considerable attention needs to be given to detail and great care taken to avoid over-interpretation, especially of the flexible regions.

An analysis of eight structures determined at atomic resolution some years ago (EU 3-D Validation Network, 1998) indicated that they follow the expected rules of chemistry more closely than those of lower-resolution analyses in the PDB, confirming that atomic resolution indeed provides more precise coordinates. A subsequent analysis of ten structures at ultra-high resolution, 0.8 Å or better (Jaskolski *et al.*, 2007) confirmed these conclusions but identified a few possible adjustments to some targets. Following this analysis, Karplus *et al.* (2008) proposed that protein stereochemistry is context dependent, *e.g.* it differs in detail between $\alpha$-helices and $\beta$-strands, and that this should be reflected in future target libraries. Thus, the availability of atomic resolution structures will provide a more objective basis for the construction of such libraries.

### 18.4.7. Relation to biological chemistry

A question arises as to what biological issues are addressed by analysis of macromolecular structures at atomic resolution. For any protein, the overall structure of its fold, and hence its homology with other proteins, can already be provided by analyses at low to medium resolution. However, proteins are the active entities of cells and carry out recognition of other macromolecules, ligand binding and catalytic roles that depend upon subtle details of chemistry, for which accurate positioning of the atoms is required. Even at atomic resolution, the accuracy of structural definition is less than what would ideally be required for the changes observed during a chemical reaction. At lower resolutions, structure–function relations require yet further extrapolation of the experimental data.

To understand the function of many macromolecules, such as enzymes, it is not sufficient to determine the structure of a single state. Alongside the native structure, those of various complexes will also be required. The differences between the states provide additional information on the functionality. For an understanding of the chemistry involved, atomic resolution has tremendous advantages in terms of accuracy, as reliable judgments can be based on the experimental data alone.

Advantages of atomic resolution include the following:

(1) The positions of all atoms that possess defined conformations are more accurately defined. This means that all bond lengths and angles in the structure have lower standard uncertainties (EU 3-D Validation Network, 1998). For regions of the molecule where the conformation is representative of the norm, this is of purely quantitative significance, but where the stereochemistry deviates from the expected value this accuracy takes on a special significance, which poses questions to the theoretical chemist. Such deviations from standard geometry often play an important role in biological function.

(2) The better the ADP definition, notably its anisotropy, the greater the insight into the static or thermal flexibility of individual regions of the molecule. Macromolecules are crucially dependent upon flexibility for properties such as induced fit in substrate or ligand recognition, allosteric responses or responses to the biological environment. More detailed definition of the position and mobility of flexible regions may be assisted by atomic resolution analysis.

(3) A few amino-acid side chains play an active role in catalysis. Those that do include histidine, aspartic and glutamic acids and serine, all through protonation–deprotonation events, and hydrogen atoms are crucial to their function. Hydrogen atoms are usually treated as riding on their parent atoms and should be included in the model, even at medium resolution. Unfortunately, those hydrogen atoms that are of interest can only rarely be treated as rigidly bonded at a predictable position. However, atomic resolution allows many hydrogen atoms to be clearly identified in the refined electron density. In addition, the presence or absence of hydrogen may be inferred by accurate estimation of the bond lengths between atoms, *e.g.* within the carboxylate groups.

(4) The relative orientation of reacting moieties is crucial to enzyme catalysis. If chemical hypotheses of mechanism are to be subjected to appropriate Popperian scrutiny (Popper, 1959), then precise definition of atomic coordinates in native and complex structures is necessary.

(5) Enzyme catalysis provides a reduction of the activation energy of the reaction, which can be achieved by distortion of the conformation of the substrate bound to the enzyme (the so-called Michaelis complex) towards the transition state or by the stabilization of the latter by the enzyme. For both, the study of complexes of inhibitors or substrate analogues at a sufficient resolution to clarify the fine detail of the structures is required.

(6) Adaptation of the enzyme to the substrate is postulated by the induced-fit theory of catalysis. The level of adjustment can be very small, and energy calculations again require that this be precisely defined.

(7) In metalloproteins, the ligand field, and hence geometry and bond lengths, around the metal ion are essential indicators of any variation in valence electrons between different states. For example, bond lengths between oxidized and reduced states of metal ions vary by the order of 0.1 Å or less, and

references