possible and the desired level of structural studies (*e.g.* complete) may be specified.

Many more searches can be carried out in the complete Search Dialog. Particularly useful are searches on atomic environments and interatomic distances. The atomic environment is defined by the coordination number, the geometry of the coordination polyhedron and the identities of the central and peripheral atoms. Searches on the number of different atomic environment types in the same structure can also be carried out. Specifying a pair of elements makes it possible to select a range of interatomic distances to be included in the search. These histograms are also useful in assessing the reasonableness of a particular distance in a Rietveld refinement.

3.7.5.8. *Particular software features*

All searches use the 'Perpetual Restraining' feature, which updates the selection set in real time as a new query is introduced, so that the progress of the search scheme can easily be monitored. The complete Search Dialog offers a large variety of features that make the retrieval and presentation of information extremely flexible.

The Chemical System Matrix View makes it easy to locate phases in binary, ternary, quaternary and pseudo-quaternary systems. The Phases List View collects a selection set into its 'distinct' phases. From the individual database entry it is particularly easy to find all database entries with the same prototype structure and plot the unit-cell volume as a function of selected atomic radii. The standard display of an entry includes a short summary about the phase, structural, bibliographic, experimental and editorial data, as well as a structure drawing, a powder pattern and a table of interatomic distances.

The software for producing structure drawings offers the visualization of atomic environments (coordination polyhedra), the statistics of interatomic distances and the calculation of selected distances and angles. Four different models are available (ball and stick, wires, sticks and space-filling), with on-the-fly rotation controlled by the mouse. The nearest-neighbour histogram of a selected atom is compared with a statistical plot containing all distances in the database involving the same chemical elements, and the atomic environments can be instantaneously modified by clicking on the nearest-neighbour histogram.

Powder-diffraction patterns can be computed for any user-defined wavelength and the visualization includes a tool for zoom-in/out tracking. Patterns based on published lists of interplanar spacings can also be visualized. It is further possible to export database entries as CIF files, tables (*e.g.* powder-diffraction pattern, distances and angles) or graphics (*e.g.* structure drawings; BMP, GIF, JPG, PNG, TIFF or Diamond documents), and individually tailored dossiers can be designed and printed.

3.7.6. Metals data file (CRYSTMET)

CRYSTMET (White *et al.*, 2002) began as a database of critically evaluated crystallographic data for metals, including alloys, intermetallics and minerals, and has grown to include inorganic compounds in general. It was started in 1960 by Cromer and Larson at Los Alamos National Laboratory, and its development was continued by the National Research Council of Canada. In 1996, the production and dissemination was transferred to Toth Information Systems.

CRYSTMET contains chemical, crystallographic and bibliographic data, together with comments regarding experimental details for each study. Using these data, a number of associated data files are generated, with the major one being a file of calculated powder patterns. Entry into CRYSTMET is *via* a number of search screens, including chemistry, bibliographic information, unit cell and reduced cell, powder patterns (using the positions of the strongest peaks as input), formula, structure type, Pearson symbol and space group. The results of queries reside in sets, which can be further manipulated using logical operations.

The results are displayed as a series of screens, which include crystallographic data, distances and angles, and the powder pattern. There is some ability to customize the calculation of the powder pattern of an entry; the calculation is performed for Debye–Scherrer geometry. Included on the Results tabs is a direct interface to the *MISSYM* program (Le Page, 1987, 1988), which searches the reported structure for additional symmetry elements. This is a very useful tool for detecting missed symmetry.

3.7.7. Protein Data Bank (PDB)

The Protein Data Bank is described in Chapter 24.1 of *International Tables for Crystallography* Volume F (Berman *et al.*, 2011). Current information is available on the web at https://www.wwpdb.org/.

3.7.7.1. *Powder diffraction by proteins*

Although powder-diffraction techniques had been applied to proteins as long ago as 1936 (Wyckoff & Corey, 1936; Corey & Wyckoff, 1936), and proof-of-principle experiments had been carried out (Rotella *et al.*, 1998, 2000), real progress in protein powder crystallography began with the work of Von Dreele (Von Dreele, 1998, 1999, 2003; Von Dreele *et al.*, 2000).

Progress in powder crystallography on macromolecules has been reviewed by Margiolaki & Wright (2008) and is also discussed in Chapter 7.1 of this volume. Notable studies include the characterization of the binding of *N*-acetylglucosamine oligosaccharides to hen egg-white lysozyme (Von Dreele, 2007*a*) and determination of the second SH3 domain of ponsin (Margiolaki *et al.*, 2007).

As with all powder diffraction, peak overlap ultimately limits the information available. Multi-pattern strategies to overcome the overlap problem have been investigated by Von Dreele (2007*b*). Multiple-pattern resonant-diffraction experiments have enabled study of the binding of $PtBr_6^{2-}$ ions to lysozyme (Helliwell *et al.*, 2010). A bootstrap approach has been used to determine the structure of bacteriorhodopsin to 7 Å resolution (Dilanian *et al.*, 2011). Parametric resonant-scattering experiments have been used to determine the secondary structures of lysozyme derivatives (Basso *et al.*, 2010). Powder-diffraction experiments have also been used to gain insight into the general features of a nonstructural protein 3 (nsp3) macro domain (Papageorgiou *et al.*, 2010).

The structure of a five-residue peptide has been determined *ab initio* using laboratory powder data (Fujii *et al.*, 2011). We can expect further useful results at this interface between small-molecule and protein powder crystallography.

As is typical in other areas of science, powder diffraction has proven to be useful in more practical features of protein processing. It has been used to identify insulin (Norrman *et al.*, 2006) and GB1 (Frericks Schmidt *et al.*, 2007) polymorphs and

Copyright © 2018 International Union of Crystallography

lot-to-lot variations in lyophilized protein formulations (Hirakura *et al.*, 2007), and has been explored for use in structure-based generic assays (Allaire *et al.*, 2009).

3.7.7.2. *Calculation of protein powder patterns (with Kenny Ståhl)*

The Powder Diffraction File contains a few experimental powder patterns of proteins. These include silk fibroin protein (00-054-1394), tubulin (00-036-1547 and 00-036-1548), insulin (00-060-1360 through 00-060-1368), tomato bushy stunt virus (00-003-0001) and tobacco mosaic virus (00-003-0003 and 00-003-0004). Patterns have not yet been calculated from the structures in the Protein Data Bank because the calculated intensities generally fit poorly to those in experimental patterns.

Protein structures in the PDB do not generally contain H-atom positions, and the contributions from the disordered solvent in the solvent channels (which is the major source of the discrepancy) is not described (Hartmann *et al.*, 2010). The conventional Lorentz factor tends to infinity when approaching $2\theta = 0^\circ$. Differences in data-collection temperatures and solvent content between powder and single-crystal specimens often mean that the lattice parameters differ. The relatively poor scattering from the protein and the large scattering from the mother liquor and sample holder result in significant background contributions to experimental powder patterns.

Optimization of the lattice parameters is generally straightforward and is important because most protein crystal structures are determined at low temperatures, while powder data are collected under ambient conditions. Protein crystals contain 30–80% disordered solvent. The solvent contribution to the diffraction pattern is most important for the low-angle powder data. In conventional protein crystallography several correction models have been developed (Moews & Kretsinger, 1975; Phillips, 1980; Jiang & Brünger, 1994), but the flat bulk-solvent model is the simplest one which yields a realistic correction (Jiang & Brünger, 1994; Hartmann *et al.*, 2010). This model includes two parameters: $k_{\rm sol}$, which defines the level of electron density in the solvent region, and $B_{\rm sol}$, which defines the steepness of the border between the solvent and macromolecular regions. These parameters are typically refined in contemporary software and cluster around $k_{\rm sol} = 0.35$ e Å^{-3} and $B_{\rm sol} = 46$ Å^2 (Fokine & Urzhumtsev, 2002).

The flat bulk-solvent correction can be applied using *phenix.pdbtools* (Adams *et al.*, 2010), which requires a PDB coordinate file and values of $k_{\rm sol}$ and $B_{\rm sol}$ as input. Average values can be used, but refined values or values from the Electron Density Server (EDS; Kleywegt *et al.*, 2004) can improve the results. The bulk-solvent correction is highly anisotropic, and both parameters affect the anisotropy.

The ideal H-atom positions can be calculated using *phenix.pdbtools*. The solvent and hydrogen contributions to the pattern can be significant (Fig. 3.7.13).

The Lorentz factor L describes the fraction of a reflection that is in the diffracting condition. For Bragg–Brentano and Debye–Scherrer geometries it is given by

$$L = \frac{1}{\sin 2\theta}\frac{1}{\sin\theta}. \qquad (3.7.3)$$

This equation assumes ideal crystals, resulting in infinitesimally small reciprocal-lattice points. The true size of the lattice points depends on the crystallite size and imperfections (strain). This smearing needs to be included in the Lorentz factor at low angles. A revised Lorentz factor for protein powder diffraction has been derived (Hartmann *et al.*, 2010),

$$L_{\rm rev} = \frac{1}{\sin 2\theta}\frac{1}{\sin\theta}\frac{\sin^2\theta}{(\sin^2\theta + \lambda^2\eta^2/12)}, \qquad (3.7.4)$$

in which η reflects the distribution of scattering-vector amplitudes. For Guinier geometry these equations become more complex (Hartmann *et al.*, 2010). Fig. 3.7.14 shows that the Lorentz factor has a smaller effect than the solvent and H atoms, but that it is still significant. By applying these corrections it should be possible for the ICDD editorial staff to calculate useful powder patterns from PDB entries that could be included in the Powder Diffraction File.

Separating the background from the diffraction pattern is not straightforward (Frankaer *et al.*, 2011). Estimation of the background is greatly assisted by a correct calculated pattern. The calculated pattern can be scaled to the experimental data using *PROTPOW* (http://www.kemi.dtu.dk/english/Research/PhysicalChemistry/Protein_og_roentgenkrystallografi/Protpow).

Ståhl *et al.* (2013) have demonstrated that existing search/match procedures can be used to identify proteins using their powder patterns, and that powder patterns calculated from Protein Data Bank coordinates with proper care can be added to a database and included in the search/match procedure. Several problems can be foreseen when including large amounts of protein data into the Powder Diffraction File. It may be worthwhile including powder patterns with several levels of solvent correction, rather than just an average value. Asymmetry from instrumental effects and specimen transparency, which can affect the peak positions, needs to be taken into account. The use of an average thermal expansion coefficient may be sufficient to account for the differences in lattice parameters between low-temperature single-crystal structures and powder patterns measured under ambient conditions.

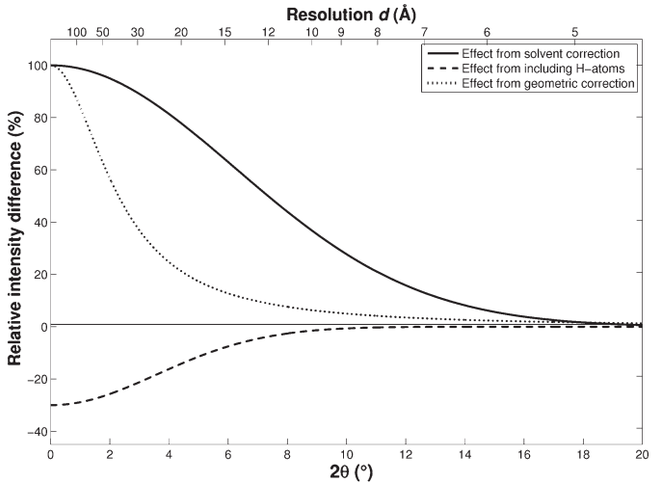

Figure 3.7.13
Overview of the trends from the different corrections. The effects are shown as the relative intensity difference $(I_{\rm non\text{-}corr} - I_{\rm corr})/I_{\rm non\text{-}corr}$ plotted as functions of the scattering angle 2θ (using Cu $K\alpha_1$) and resolution $d = \lambda/(2\sin\theta)$. The curves are based on average corrections of lysozyme and insulin data. $I_{\rm non\text{-}corr}$ is the raw intensity from a calculated pattern which has only been Lorentz corrected. The geometric correction curve was calculated using $\eta = 0.045$ Å^{-1}. From Hartmann *et al.* (2010).

references