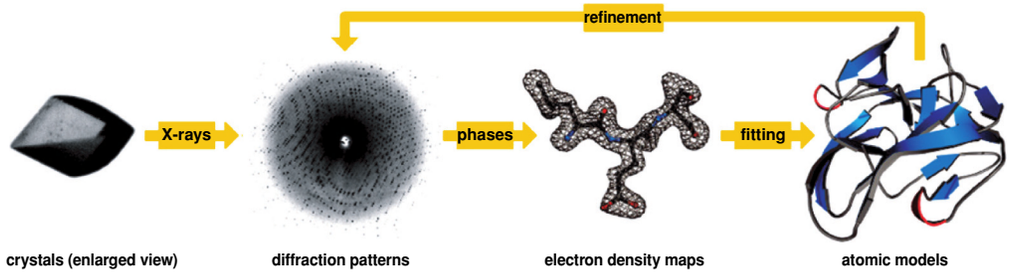

Figure 2.9. The process of refinement is an iterative one, in which small changes are made to the atomic model, a new diffraction pattern is calculated from that model, and that pattern is compared to the measured diffraction pattern. The discrepancy between the two is expressed as the *R* factor. The lower the *R* factor, the better the agreement between the two. Taken from G. A. Petsko and D. Ringe, in *Protein Structure and Function*, New Science Press Ltd, London, 2004.

guide how far a molecule may deviate from ideality in attempting to fit an electron density feature. The limits are determined from what is known about the structures of small molecules that are representative of substructures of the protein (Figure 2.9).[11–14]

The measure of success for this fitting procedure is the *R* factor, a measure of the disagreement between model and experiment. Although the fitting of the model to the electron density, and the agreement between the two, is seen at the level of the electron density map, this measure of agreement can be calculated only at the level of the measurements made to obtain that map (i.e., the structure amplitudes). The calculation determines the difference between the calculated reflection amplitudes derived from the model and the measured ones derived from the x-ray experiment. Thus, *R* factors reflect not only the quality of the fit of the model but also the quality and resolution of the data. For protein macromolecules, *R* factors are usually within the range of 15% to the low 20% for data around 1.8–2.5Å resolution, which is a common resolution range for protein/ligand complexes. These numbers mean that roughly 80% of the measured scattering from the crystal has been accounted for by the model. In many cases, the unaccounted-for scattering will include not only errors of measurement but also the failure to model the disordered solvent in the channels within the crystal lattice (Figure 2.10).

Because of the overwhelming influence of the phases (calculated) over the reflection amplitudes (observed data) in determining the final electron density, the *R* factor can be manipulated. Consequently, it is now common practice to calculate an *R* factor from data that have not been used in the refinement process and therefore not been biased by calculated phases. This measure is called the R_{free} and usually uses 5–10% of the data, randomly chosen and excluded from all refinement steps, to calculate the measure of disagreement.[16] Because of the incompleteness of the data used, and the lack of phase bias, the R_{free} is always higher than the *R* factor by approximately 3–5% in the case of well-refined structures. In the early stages of refinement it may be 10% higher.

The molecules of protein in the crystal are packed in such a way that solvent channels are found between them. These are filled with the solution from which the protein was crystallized, and that solution may contain other ions and molecules that are associated with the protein. An estimate of the volume of the crystal that is attributable to solvent comes from the Matthews coefficient, calculated from the crystal data.[17] The most important component of the solvent is the bound water molecules that are found

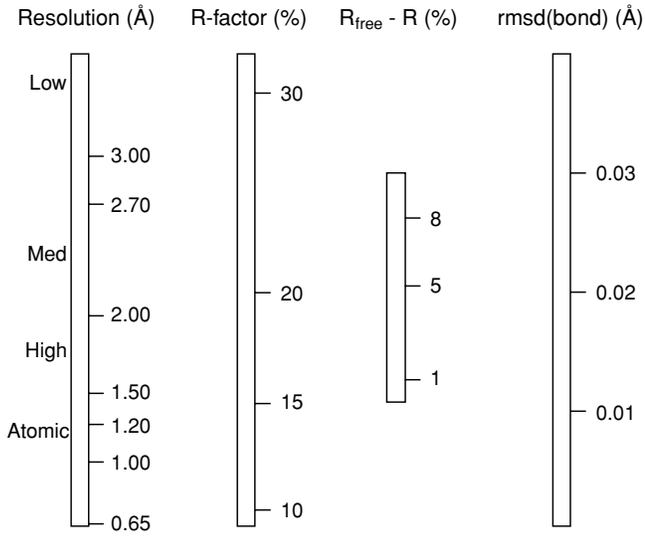

Figure 2.10. Approximate correlations between resolution of the data, the expected *R* factors for a refined structure, the expected differences between the *R* factor and R_{free} of the final refined structure, and the approximate precision of bond distances ascertainable from a structure.[7] These values can be used as criteria to assess the quality of a crystallographic model. For instance, a large difference between *R* and R_{free} could indicate possible overinterpretation of the data; if the difference is very low, it could mean that the test data set used to calculate R_{free} is not in fact "free." rmsd (root-mean-squared deviation) indicates the deviation of protein geometry from ideality. For instance, high rmsd(bonds) indicates model error. If it is too low, the refinement may have been dominated by strict adherence to geometry rather than refinement to experimental diffraction data. It should be noted that "ideal" bond lengths may have errors of approximately 0.02Å, so the expectation that a model is better than that is unreasonable. Redrawn from data in Wlodawer et al., 2008.[7]