Theor Chem Acc (2012) 131:1076
DOI 10.1007/s00214-011-1076-8

REGULAR ARTICLE

Toward ab initio refinement of protein X-ray crystal structures: interpreting and correlating structural fluctuations

Olle Falklöf · Charles A. Collyer · Jeffrey R. Reimers

Received: 4 July 2011 / Accepted: 9 August 2011 / Published online: 11 January 2012

Abstract The refinement of protein crystal structures currently involves the use of empirical restraints and force fields that are known to work well in many situations but nevertheless yield structural models with some features that are inconsistent with detailed chemical analysis and therefore warrant further improvement. Ab initio electronic structure computational methods have now advanced to the point at which they can deliver reliable results for macromolecules in realistic times using linear-scaling algorithms. The replacement of empirical force fields with ab initio methods in a final refinement stage could allow new structural features to be identified in complex structures, reduce errors and remove computational bias from structural models. In contrast to empirical approaches, ab initio refinements can only be performed on models that obey basic qualitative chemical rules, imposing constraints on the parameter space of existing refinements, and this in turn inhibits the inclusion of unlikely structural features. Here, we focus on methods for determining an appropriate ensemble of initial structural models for an ab initio X-ray refinement, modeling as an example the high-resolution single-crystal X-ray diffraction data reported for the structure of lysozyme (PDB entry "2VB1"). The AMBER force field is used in a Monte Carlo calculation to determine an ensemble of 8 structures that together embody all of the partial atomic occupancies noted in the original refinement, correlating these variations into a set of feasible chemical structures while simultaneously retaining consistency with the X-ray diffraction data. Subsequent analysis of these results strongly suggests that the occupancies in the empirically refined model are inconsistent with protein energetic considerations, thus depicting the 2VB1 structure as a deep-lying minimum in its optimized parameter space that actually embodies chemically unreasonable features. Indeed, density-functional theory calculations for one specific nitrate ion with an occupancy of 62% indicate that water replaces this ion 38% of the time, a result confirmed by subsequent crystallographic analysis. It is foreseeable that any subsequent ab initio refinement of the whole structure would need to locate a *globally* improved structure involving significant changes to 2VB1 which correct these identified *local* structural inconsistencies.

Keywords Monte Carlo · Density-functional theory · Protein refinement · Ensemble refinement · Lysozyme

1 Introduction

X-ray crystallography is the most important experimental technique in protein structure determination. In order to

Published as part of the special collection of articles celebrating the 50th anniversary of Theoretical Chemistry Accounts/Theoretica Chimica Acta.

O. Falklöf · J. R. Reimers (✉)
School of Chemistry, The University of Sydney,
Sydney, NSW 2006, Australia
e-mail: jeffrey.reimers@sydney.edu.au

O. Falklöf
Department of Chemistry, The University of Gothenburg,
Gothenburg, Sweden

Present Address:
O. Falklöf
Department of Physics, Chemistry and Biology,
Linköping University, 581 83 Linköping, Sweden

C. A. Collyer
School of Molecular Bioscience, The University of Sydney,
Sydney, NSW 2006, Australia

understand biological functions and chemical mechanisms, the structures of macromolecules need to be obtained with high accuracy. In general, crystal structures of small molecules (~150 atoms) determined at high resolution match closely with models derived from ab initio calculations due to the low complexity of their molecular structure, the overdetermination of the refinement process and the use of unbiased free atom refinement methods. In contrast, despite technological enhancements during the last 30 years, it is difficult to generate accurate atomic-resolution macromolecular structures. For macromolecular systems, the structures are often highly complex and usually the number of observations per atomic parameter is low and therefore empirical information must be included to enable a structure to be refined. Even in a handful of cases, where structures have been subjected to free atom refinement against high-resolution data, the structural models are inadequate in representing the unknown ensemble of conformers which compose the observed average structure of the molecule.

For X-ray refinement, an initial protein model structure is generated and then improved by matching it to observed reflection data while simultaneously maintaining geometrical restraints. Geometrical restraints have a large impact on the structure if the reflection data is poor but are normally weighted down in high-resolution refinements. A common approach is to use the geometrical parameters derived from accurate measurements of bond and angle parameters in small protein structures [1]. These standardized parameters are used in refinement programs such as SHELXL [2] and REFMAC5 [3] that are used widely. For a "normal" protein structure, these values are in general realistic, but for unusual structural features, such as cofactors, this methodology can yield inaccurate results. For example, we recently optimized the structure of photosystem-I by linear-scaling density-functional theory without reference to the original X-ray diffraction data [4]. This introduced only small changes to well-represented features such as main chain conformations but significant changes to the chlorophylls, cofactors, oligomerization features and group orientations. More generally, significant problems with even the determination of the backbone structure have been noted [5–12], and Eyal et al. [13] have shown that structures by the same authors in the PDB are more similar to each other than structures from independent groups.

Restrained refinement was pioneered by Konnert and Hendrickson [14–17], and the function minimized is the weighted sum of the difference of the observed and predicted intensities combined with the differences of the squared ideal and observed interatomic distances together with other types of geometrical restraints. Jack and Levitt [18] introduced a refinement approach where a sum of energy and X-ray terms was minimized. Brunger et al. [19] further developed the methods and combined the X-ray data with a potential from molecular dynamics.

Due to the computational time and problematic scaling properties, high-level density functional calculations have typically not been applicable to protein structure refinement. During the last few years, a number of linear-scaling semi-empirical and density-functional theory (DFT) algorithms have been developed [4, 20–39] (for recent reviews, see e.g., [40–44]). The development of modern density functionals capable of describing dispersion interactions [45–61] allows such methods to properly describe all of the chemical bonding types associated with protein structure. These methods are commonly used to enhance descriptions of protein active sites but have also been used to optimize whole protein structures, but without reference to the original raw diffraction intensities used to generate them [4, 23–25, 27–29, 33]. Recently, dispersion-corrected DFT optimization led to the revision of a protein X-ray structure, reducing the R factor to increase the quality of the interpretation of the diffraction data [62].

For many reasons, it is preferable to employ such methods *within* the refinement stage, however. Pioneering studies employing quantum chemical methods focused on improving low-resolution structures. Ryde et al. [63–66] have developed a method that combined quantum mechanics and molecular mechanics, where only a part of the structure used restraints from quantum mechanical calculations. Also, Merz et al. [67–73] and Stewart [23, 33, 43] made use of another approach using a semi-empirical method for the energy restraints. These approaches have yielded more chemically reasonable aspects of critical structural features than was previously obtained using traditional refinement methods, often also increasing agreement between observed and calculated diffraction intensities.

A major challenge at this time is the demonstration that quantum chemical methods can be used to enhance the refinement of *all* aspects of, in particular, high-resolution X-ray structures. These are the structures that have the most significant features resolved. The advantage that ab initio methods have to offer protein crystallography is that they know most about the highest-resolution features of the protein, precisely the area in which little information is directly available from current experiments. Complementary to this, ab initio calculations know least about the overall protein conformation and other lower-resolution features that are readily determined using crystallography. There is, however, an intermediate regime concerning torsional motions and short-range intermolecular forces in which useful information could be expected from both approaches, and indeed it is important to verify that any used computation method can accurately reproduce such features. All computational methods will suffer from some systematic failures, and perhaps now the time has come where ab initio methods can be utilized to simply add information to crystallographic refinements.

Here, we investigate one of the most basic issues facing the melding of ab initio computations of protein structure with X-ray structure refinement: the presence of multiple conformations in protein structures, a feature that often leads to poor X-ray structural models [7]. A multiple-copy refinement scheme (ensemble refinement) [74–80] is developed, which treats multiple conformations as an ensemble of fully connected structures, each of which obeys basic chemical rules. Modern X-ray structures often detect many atoms with multiple sites. Major correlations between the locations of such atoms, especially those directly related to function, are often noted in structural models, but systematic approaches to determining such correlations are rare, and PDB files are rarely constructed so as to make correlations explicit. As a result, most correlations go unnoticed, yet such features are critical to any ab initio refinement of crystal structures. Ensemble refinement can lead to significant improvements of observed and fitted diffraction intensities [75] and are currently of great interest as a means of improving structural models [11, 81–91], including refinement of NMR data [92].

The structure of lysozyme is used as a sample system to consider the issue of correlated structural fluctuations as lysozyme is one of the most studied proteins. It was discovered by Fleming [93], and its three-dimensional structure was first solved by Blake et al. [94], providing the first enzyme structure to be solved by X-ray crystallography. The number of lysozyme structures contained in the Protein Data Bank (PDB) [95] is large. Even so, the exact catalytic mechanism of lysozyme is still a subject of debate [96, 97], indicating that further structural characterization and computational analysis may be required. From an experimental point of view, lysozyme is, in comparison with other proteins, easy to purify and crystallize, and crystals diffract at high resolution. Moreover, lysozyme crystallizes in different polymorphic space groups depending on the crystallization conditions [98]. We consider only the refined "2VB1" structure from the PDB of triclinic hen egg white lysozyme (HEWL) by Wang et al. [98]. This is currently the third-highest-resolution protein structure deposited in the PDB and is at a resolution of 0.65 Å. This structure was refined in SHELXL by *removing* the restraints for the well-ordered parts of the model (those in a single conformation) and represents a close approximation to the structures resulting from free atom refinements conducted in small-molecule X-ray crystallography.

In Sect. 2, we review features of standard protein X-ray structure refinement that are critical to the development of enhanced methods. These methods focus on providing the best-possible representation of the raw X-ray diffraction data in terms of atomic coordinates, given options for the application of non-experimental features such as constraints and empirical force fields. Our aim is to replace these empirical conditions with ab initio ones, but in order to commence such an operation, many significant issues involving lack of experimental knowledge of required information must first be addressed. Section 3 describes critical data that is either present or absent from the 2VB1 X-ray structure of lysozyme [98]. Section 4 describes a Monte Carlo technique for determining an ensemble of chemically realistic structures that represent the multiple conformations detected in the original refinement. This involves adding additional constraints to the refinement that enforce basic chemical rules, but otherwise this Monte Carlo step uses the AMBER force field to distinguish between a vast number of chemically realistic structures that would each generate the *same* R factor during X-ray refinement. Section 5, however, considers basic chemical features that appear *inconsistent* with the original refinement, indicating that the construction of a chemically realistic model that is also in good agreement with the raw diffraction data is a considerable challenge.

2 Basic relevant features of protein X-ray diffraction analyses

The raw data collected during X-ray diffraction experiments consist of diffraction intensities I (h, k, l) and associated structure-factor amplitudes F_{obs} (h, k, l), with

$$I(h, k, l) \propto F^2_{\mathrm{obs}}(h, k, l) \tag{1}$$

where h, k and l index the observed reflections. The electron density within the crystal at point (x, y, z) may be generated from a Fourier transform of the structure factors,

$$\rho_{\mathrm{obs}}(x, y, z) = \frac{1}{V}\sum_h\sum_k\sum_l F_{\mathrm{obs}}(h, k, l) \times \exp(-2\pi i(hx + ky + lz) + i\alpha_{\mathrm{obs}}(h, k, l))), \tag{2}$$

where V is the volume of the unit cell, if the phases α_{obs} (h, k, l) are known [99]. Phases and structure factors may both be readily determined from an atomic model of the crystal structure, allowing for the construction of the comparable electron density

$$\rho_{\mathrm{calc}}(x, y, z) = \frac{1}{V}\sum_h\sum_k\sum_l F_{\mathrm{calc}}(h, k, l) \times \exp(-2\pi i(hx + ky + lz) + i\alpha_{\mathrm{calc}}(h, k, l))). \tag{3}$$

However, phases are not directly determinable from the diffraction data and so estimated phases $\alpha_{\text{obs}} \approx \alpha_{\text{calc}}$ are utilized in Eq. 2. The aim of crystallographic refinement is to determine a reliable atomic coordinate model and therefore reliable estimates of the phases. This scenario is readily achievable for small-molecule crystallography as realistic initial phases are derived using direct methods, but this approach cannot usually be applied in the case of protein crystallography [99]. While femtosecond nanocrystallography [100] offers in the next decade a possible experimental solution to this problem, the analysis of protein X-ray diffraction data currently usually proceeds through a number of stages, with at each stage the phases deduced from the atomic coordinate model; as better phases are obtained in each cycle, the “observed” electron density ρ_{obs} (x, y, z) reveals more of the atomic structure, leading then to better phases. Use of ab initio calculations in aiding protein structure refinement should be conceived as simply adding another stage to the existing analysis sequence, striving again to improve accuracy and provide a better interpretation of the raw diffraction data.

The final steps of the structure determination of a protein involve the refinement of the structure. In a structure refinement, the model structure is refined against the experimental diffraction data to improve the R factor defined as

$$R = \frac{\sum_{hkl}\left||F_{\text{obs}}| - \tilde{k}|F_{\text{calc}}|\right|}{\sum_{hkl}|F_{\text{obs}}|}, \tag{4}$$

where $\tilde{k}$ is a scale factor. Note that this involves comparing functions of purely observed quantities F_{obs} to those of purely calculated ones F_{calc}; the precise nature of the function that is actually minimized depends on the software used, with the maximum likelihood method [3] being a common choice.

Density difference maps such as

$$\begin{aligned}\Delta\rho(x,y,z) &= \rho_{\text{obs}}(x,y,z) - \rho_{\text{calc}}(x,y,z)\\ &\approx \text{``}F_O - F_C\text{''} = \frac{1}{V}\sum_{hkl}(|F_{\text{obs}}| - |F_{\text{calc}}|)\\ &\quad\times \exp[-2\pi i(hx + ky + lz) + i\alpha_{\text{calc}}]\end{aligned} \tag{5}$$

are usually created in order to enhance refinement but because of the use of α_{calc} in approximating ρ_{obs}, their quality varies spatially within the structure and so they can be difficult to interpret. Of particular concern for the generation of a structural model suitable for ab initio refinement is the feature that typically α_{calc} has a more significant influence on $F_O - F_C$ than does $|F_{\text{obs}}| - |F_{\text{calc}}|$, sometimes allowing the magnitude of $F_O - F_C$ to decrease while R increases [1]. Note too that a commonly reported type is also the combined map with an observed density added to the difference map [99]

$$\text{``}2F_O - F_C\text{''} = \frac{1}{V}\sum_{hkl}(2|F_{\text{obs}}| - |F_{\text{calc}}|)\exp[-2\pi i(hx + ky + lz) + i\alpha_{\text{calc}}|]. \tag{6}$$

The electron density around an atomic nucleus is typically temperature independent, though large-amplitude motions of atoms with frequencies less than $kT/\hbar$ can blur the density. If all of the equivalent molecules in a protein crystal do not adopt the same conformation, then the electron density will be distributed, possibly mimicking the effects of thermal motion. The effects of large thermal motions are not usually included explicitly via ensemble representations in protein structure refinements, though the presence of multiple conformers can sometimes be detected and explicit structures modeled for each conformer [99]. Explicitly identified conformers are ascribed weights indicating the fraction of molecules adopting each particular conformer. Thermal effects and non-explicitly represented conformational effects are usually treated in protein structure refinement implicitly by smearing out the atomic electron density using

$$T_{\text{iso}} = \exp\left(-B\frac{\sin^2\theta}{\lambda^2}\right) = \exp\left(-\frac{B}{4}\left(\frac{2\sin\theta}{\lambda}\right)^2\right), \tag{7}$$

where B is an “atomic displacement parameter” that reflects isotropic thermal motion with mean square displacement, $\overline{u^2}$,

$$B = 8\pi^2\overline{u^2}. \tag{8}$$

If high-resolution experimental data are available, then this isotropic approximation may be inadequate so that an alternate anisotropic temperature factor

$$\begin{aligned}T_{\text{aniso}}(h,k,l) = \exp[-2\pi^2(&U_{11}h^2a^{*2} + U_{22}k^2b^{*2} + U_{33}l^2c^{*2}\\ &+ 2U_{12}hka^*b^* + 2U_{13}hla^*c^*\\ &+ 2U_{23}klb^*c^*)]\end{aligned} \tag{9}$$

is often used, where a^*, b^* and c^* are the lengths of the axes in the reciprocal space. Here, U_{ij} is the element in the ith row and jth column of the displacement tensor with respect to the axes in reciprocal space.

Every observation is associated with some level of noise. Since the aim is to model real features and not noise through over-fitting, the experimental data are usually divided into a working and a test set. Only reflections from the working set are used in the refinement. The indicator of the quality of the model is the R_{free}-factor, which is an R factor calculated considering only the reflections from the test set that was not used during the refinement [101].

Hydrogen atoms give a very weak signal in the diffraction pattern. Usually, only non-hydrogen atoms are modeled in X-ray structures. If the hydrogen atoms are included, they are commonly included as riding groups and follow the movement of their neighboring heavy atom. From a chemical point of view, hydrogen atoms are crucial for the understanding of the stability and mechanisms of the structures and must be considered carefully [99]. Any ab initio calculation requires an accurate description of the locations of the hydrogen atoms.

The observed number of reflections per fitting parameter is usually low, and to enhance refinement, extra-geometrical conditions are usually included. Jack and Levitt proposed a method that takes an energy term into account,

$$Q = (1 - w_x) \cdot E + w_x \cdot \sum_{hkl} w(h,k,l)(|F_{\mathrm{obs}}(h,k,l)| - |F_{\mathrm{calc}}(h,k,l)|)^2 \qquad (10)$$

where the function involving an atomic energy contribution E was minimized, and $w_x \in [0, 1]$ controls the relative contributions of the energy and the X-ray term [18]. It is this function E that could be replaced with an ab initio estimate rather than its estimation using empirical force fields.

Large volumes in the crystal may be occupied primarily by solvent molecules, however; and in these regions, it is typically not adequate to use atomic models to describe the diffraction as the degree of disorder is typically too large. Such regions are typically modeled more realistically as a homogeneous electron gas. It is, however, critical that the best-possible description of these regions be obtained as diffraction contributions from them add to the structure factors and influence the determination of phases. A commonly used correction based on Babinet's principle [102] is

$$F_{\mathrm{calc}} = \left[1 - K \exp\left[-B_{\mathrm{solvent}} \frac{\sin^2\theta}{\lambda^2}\right]\right] \cdot F_{\mathrm{protein}}. \qquad (11)$$

3 Hen egg white lysozyme 2VB1 structure properties

The observed [98] crystal structure of HEWL is triclinic with just one molecule in the asymmetric unit. The unit-cell parameters are $|a| = 27.07$ Å, $|b| = 31.25$ Å, $|c| = 33.76$ Å, $\alpha = 87.98°$, $\beta = 108.00°$ and $\gamma = 112.11°$. Only one protein chain, containing 129 amino acids, is present per cell. Other identified molecules at least partially included in the structure are one acetate ion, nine nitrate ions, three ethylene glycol molecules and 170 water molecules. The structure 2VB1 contains coordinates for all hydrogen atoms except those in the water molecules. Nine atoms are missing in the B conformation of TYR-20. Based on our analysis, 17.5% of the volume of the unit cell is not explicitly represented using atoms, enough volume to accommodate 146 additional water molecules. However, the identified regions of the unit cell are also positively charged and so at least some counter anions are missing, and it is impossible that some of the identified areas in hydrophobic regions contain no atoms at all.

Most significantly, 33% of the atoms are identified with multiple sites, reflecting observed conformational differences between molecules in different unit cells within the crystal. At most three configurations were identified for any one atom, named "A", "B" and "C"; conformation "A" of one atom is not necessarily correlated with conformation "A" of another atom, however. Nevertheless, the set of atoms from LEU-17 to LEU-25 have been identified as giving rise to two distinct conformations of the protein chain, depicting a correlated structural fluctuation. As such, chemical rules indicate that the occupancy of each atom in the set in each configuration must be the same, but despite this, the reported structure optimizes individual weights for *each* atom in *each* conformation. Such a feature is not allowed if ab initio computations are to be used to refine the structure. In other cases, for example, for the anions and ethylene glycols, the same occupancy is specified for all atoms in a chemical group. However, no correlations were presented between the occupancies of the vast majority of atoms with multiple identified sites. As ab initio calculations manipulate explicit knowledge of correlations between atoms, the 2VB1 structure is far from one that is suitable for such a treatment. Yet it is only because 2VB1 is obtained at such high resolution that the very presence of multiple conformations is identified, and so only for structures like this can an ab initio calculation be conceived. As an indication of the severity of the problem faced by an ab initio calculation, only 25% of the residues of the protein are surrounded by a 5-Å region in which the atomic structure is unambiguously defined in 2VB1.

The structure 2VB1 was also refined using anisotropic displacement parameters. This process is desirable in that it improves phases and allows the structure to be indentified in more regions of the crystal. However, these enhancements arise at the cost of increasing the number of parameters per heavy atom in a single conformer from 4 to 9. Just as a chemically meaningful structure suitable for ab initio refinement requires consistent values for the occupancies of each atom in a fluctuating chain conformation, so it also requires the anisotropy values of atoms need to be correlated with each other. Such correlations are *not* built into the 2VB1 analysis, however, and as a result, many parameters are introduced into the anisotropic refinement that have no physical basis but significantly improve accuracy measures such as R: for lysozyme, R decreases from 19.48 for an isotropic analysis to 8.40 for an anisotropic one, but it is not clear what fraction of this decrease can be attributed to the inclusion of real chemical effects and how much can be attributed to over-parameterization. Indeed, 20025 free parameters were fitted to 187165 observed reflections, locating just 1814 non-hydrogen atoms, or 11 parameters per heavy atom on average. It is most likely that ab initio refinement will preclude the use of unconstrained anisotropic temperature factors, however.

4 A Monte Carlo scheme to determine a small set of chemically reasonable structures that represent the original 2VB1 X-ray model

In an ab initio calculation, all atoms must be explicitly represented. Variations in conformers within the protein crystal can be accounted for by performing calculations on an ensemble of possible protein structures, averaging the energies and forces. To represent truly disordered regions such as the 17.5% of the protein structure not so far explicitly represented, a large number of configurations could be necessary but the use of a small number of structures may be feasible [103, 104]. While this task is a central feature in all molecular dynamics simulations of protein function performed outside the scope of X-ray structural refinement, we focus here on other, more basic, ambiguous features of the data analysis and choose to ignore this region completely. In a subsequent improved implicit approach, this region could be represented by a dielectric material [9, 105–107] or by explicit molecules [108, 109].

We focus on the atoms in the 2VB1 structure that are attributed partial or multiple occupancies, and an experimental error bar of ±10% in these occupancies is stated [98]. Given this and the triclinic symmetry of the crystal, the simplest possibly realistic description of conformation variation of the represented atoms would include an ensemble of 8 chemically complete structures, allowing occupancies to be represented to an accuracy of ±6.25%. Converting all of the stated occupancies to their nearest multiple of one-eighth increases the *R* factor from 19.48 to 19.52% for an isotropic analysis and from 8.40 to 8.51% for an anisotropic analysis, as determined using REFMAC5 [3]; these increases are much less than what could be considered as physically meaningful. However, changing these occupancies to enforce the chemical restriction that all atoms in a particular conformation have the same weight has a larger effect, increasing the *R* factors to 19.60 and 8.88%, respectively. The quality of some subsequent ab initio optimization should be compared to these modified *R* factors rather than the originals as these increased values reflect mostly the chemically realistic requirement that occupancies are assigned to conserved chemical groups as whole entities.

An ensemble of 8 chemically feasible structures is then produced that maintains the generated atomic-site occupancies (i.e., this procedure essentially does not alter the *R* factor). This ensemble is represented in two different ways. As presented to the X-ray refinement codes, 8 all-atom structures are defined in which every atom has precisely an occupancy of one-eighth. As presented to a subsequently described energy analysis program, the coordinates are obtained expanding the original unit cell into a 2 × 2 × 2 superlattice. Each site of the superlattice is assigned one of the 8 structures from the ensemble, the entire structure satisfying the new boundary conditions. It is possible to transform between these two representations without difficulty, and using an explicit superlattice representation during X-ray refinement would be inconvenient as the unit-cell parameters, and hence all diffraction indices, would require modification. In either case, refinement of the expanded model using standard means would not be possible as the number of atomic coordinates required to be fitted is increased eightfold, making such an analysis overparameterized. However, if ab initio forces are included during refinement and a chemically based scheme is used to specify ab initio the thermal parameters, then a solution should be possible as our model includes 10,871 heavy atoms in the 2 × 2 × 2 superlattice and 187,165 reflections are observed. It is also feasible to constrain the coordinates of some atomic fragments in the 8 cells to be equivalent, thus reducing overparameterization for chemical units dominated by only one significant structure.

Here, we focus not on such a refinement but rather on the required task of generating a starting structure for the 2 × 2 × 2 superlattice. Fluctuations of conformations of nearby atoms are likely to be highly correlated (e.g., the presence or absence of a nearby nitrate ion can determine the conformation of a cationic residue), yet such correlations are not identified in the existing crystal structure [98]. A realistic configuration of the 2 × 2 × 2 superlattice must take these correlations into account. We consider a vast number of possibilities, rating them using a molecular mechanics energy function. All of the structures considered in this section essentially share the same *R* factors (there are trivial variations between them caused by differing hydrogen locations that have no quantitative effect) and hence the energy function is being used purely to add *additional* information required for a subsequent refinement process.

As previously mentioned, all atoms in the identified [98] nine-residue group LEU-17 to LEU-25 that form a loop region with two distinct conformations were given the same weights and treated as *one* chemical unit. Similarly, the residues [98] ASP-87 to ALA-90, LYS-97 to VAL-99, and ASN-103 and GLY-104 were treated each as single correlated chemical units. All other single residues were identified as chemical units and assigned the same weight in each configuration. A single chemically realistic structure was added for the nine missing atoms in the B conformation of TYR-20, with subsequent structural refinement being assumed to be sufficient to describe the required conformational variations of these atoms.

To determine all other correlations, Monte Carlo simulations were conducted using specifically developed software. If a chemical unit has a weight of say 5/8 for one conformation and 3/8 for a second, then 5 copies of the first configuration and three of the second are added in random order to occupy the 8 sections of the 2 × 2 × 2 superlattice. Alternatively, if say a water molecule had an occupancy of 5/8, then water molecules were added to 5 of the 8 sections with the other 3 sections remaining vacant. The purpose of the Monte Carlo program is to determine the optimum distribution for the included chemical groups among the 8 cells of the 2 × 2 × 2 superlattice, establishing correlations between the occupancies of different chemical sites. Note that the full crystallographic symmetry properties of the superlattice are used in all calculations.

The AMBER force field [110, 111] was used to drive the Monte Carlo simulations. This is made up of a local part (bond stretches, bends, torsions) and a long-range part (dispersion and electrostatic interactions). If the correlation between fluctuations of two neighboring residues needs to be determined, then bond bending and torsional motions will differ between conformations. This effect is small, however, and is also a rare occurrence, and hence we consider only the long-range part of the AMBER force field,

$$E^{\mathrm{LR}} = \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right]. \tag{12}$$

The dielectric constant, ε, was set to unity and $A_{ij} = e_{ij}^{*}(R_{ij}^{*})^{12}$ and $B_{ij} = 2e_{ij}^{*}(R_{ij}^{*})^{6}$, where R_{ij}^{*} and e_{ij}^{*} are the equilibrium bond length and bond energy, respectively, of the Lennard–Jones potential between atoms i and j. For interacting atoms of different types, the Lennard–Jones bond length was taken as a sum of the van der Waals radius of each atom, $R_{ij}^{*} = R_i^{*} + R_j^{*}$, while the well depth was taken as the geometric average of the well depths for each atoms, $e_{ij}^{*} = \sqrt{e_i^{*}e_j^{*}}$. The Lennard–Jones parameters and atomic charges of the residue atoms were taken from the study of Cornell et al. [111]. For acetate, nitrate and ethylene glycol, charges were calculated, after optimization, at the B3LYP/6-31G(d) level of theory [112, 113] by GAUSSIAN 09 [114] including the "prop = (fitcharge, dipole)" and "scrf = cosmo" options; the resulting charges are shown in Fig. 1. Use of a unit dielectric constant is appropriate if all atoms in the crystal are explicitly represented and properly sampled, and its use herein is based on

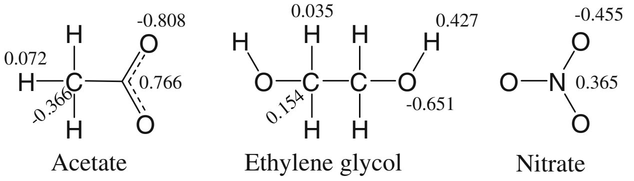


Fig. 1 Atomic charges used for acetate, ethylene glycol and nitrate; AMBER atom types "C", "N", "H" and "O" were also applied

the somewhat crude assumption that most correlations are induced by short-range intermolecular forces rather than by long-range electrostatics. Note that the full crystallographic boundary conditions are used in all calculations.

Hydrogen atoms were added to all of the water molecules in the 2VB1 structure [98]. To describe the gross features of the possible water configurations, a grid of 72 different water orientations per water molecule was set up. These grid points were distributed on the surface of two cones pointing toward each other with an angular displacement of 30° between each point. By using such a grid, the water molecule were able to do physically unlikely movements and jump between possible alternate local minima in a single Monte Carlo step. To make the grid more flexible and address more subtle bonding features, each individual molecule was also allowed to rotate small amounts about its x-, y- and z-axes. The TIP3P water molecule model by Jorgensen et al. [115] was applied with both fixed bond length (0.9572 Å) and bond angle (104.45°).

The Monte Carlo scheme functioned by selecting at random one of five possible operations, making a random move for that operation, and determining the associated change in the total energy. This Monte Carlo move was accepted according to the Metropolis algorithm:

$$\begin{array}{llll} \text{If} & E_{\text{new}} \le E_{\text{old}} & \text{always accept,} & \\ \text{If} & E_{\text{new}} > E_{\text{old}} & \text{accept if} & \xi \le \exp\left(\frac{E_{\text{old}}-E_{\text{new}}}{k_B T}\right), \end{array} \tag{13}$$

where $\xi \in [0,1]$ is a random number. The five possible operations and the relative weighting used in determining which type of operation to make next are described in Table 1. Four of these operations involve simply interchanging the conformations between a selected two of the eight sections of the 2 × 2 × 2 superlattice containing either part of the protein chain or else interchanging present and absent sites for ethylene glycol, nitrate ions or waters, while the fifth operation involves small-angle changes to the configuration of a specific water molecule. For the water molecules, seven random numbers are generated per move: selecting the water to move, selecting a new grid point number, selecting a rotation axis, selecting the

Table 1 Types of Monte Carlo operations

Operation	Weight
Residue conformation (A, B or C)	35 (1/group)
Ethylene glycol (present or absent)	3 (1/group)
Nitrate (present or absent)	9 (1/group)
Water (present or absent)	170 (1/group)
Water configuration	1,700 (10/group)

Table 2 Average fraction of accepted moves after each temperature step used in the Monte Carlo simulation

Cycles	Temperature/K	Acceptance rate
1–20,000,000	4,273	0.522
20,000,001–40,000,000	4,173	0.518
40,000,001–60,000,000	4,073	0.514
60,000,001–80,000,000	3,773	0.506
80,000,001–100,000,000	0	0.405

Table 3 Average fraction of accepted trial moves for each operation

Group type	Acceptance rate
Residue	0.287
Ethylene glycol	0.518
Nitrate	0.158
Water	0.348
Water configuration	0.413

rotation angle and optionally three specifying a small translation of the oxygen coordinates.

The Monte Carlo simulations must be run at some temperature T. This temperature does not reflect that used in the crystallography experiments but instead is adjusted to produce an optimized representative structure at 0 K. Hence, a run is initially performed at a temperature sufficiently high to sample all of the available parameter space. As a typical quantity associated with a configuration change is the making or breaking of a hydrogen bond, likely temperatures fall in the region of 5–10 kcal mol^{-1} energy, or 2,000–4,000 K. We found a temperature of $\sim$4,300 K to be sufficient based on the generally accepted criterion that a Monte Carlo calculation works best if ca. half of the moves are accepted and half rejected. After this, the calculation was slowly quenched to 0 K; Table 2 shows the temperatures used, the total number of Monte Carlo moves made and the fraction of accepted moves. As the five classes of operations have intrinsically different associated energy scales, the fraction of accepted moves varies between each class. Acceptance ratios for each class are shown in Table 3 and indicate acceptable rates, the most difficult operation being exchange of nitrate ions with an acceptance ratio of just 0.158. In total, 10^8 Monte Carlo moves were made per calculation, and 10 separate calculations were performed.

Since the Monte Carlo simulations embody a constant total number for each possibility of each chemical group and hence fixed marginal distributions, Fisher's exact test [116] is a suitable independency test for the configuration of each chemical unit considered. It was calculated under the null hypothesis that the configurations of the different units are independent and 4.7% were found to have a p value below 0.05 in a two-sided test. This two-sided test was performed as described by Freeman and Halton [117–119], where the p value is the sum of all possibilities with a probability less or equal to the observed one. It reveals that some of the variations noted originally in the 2VB1 structure, in particular groups close to each other, are indeed strongly correlated.

In order to compare the ten individual configurations (named "A"–"J") obtained after quenching to 0 K each simulation of the 2 × 2 × 2 superlattice, a similarity value S was evaluated,

$$S_{pq} = \frac{\sum I_{ii}^{pq}}{N} \tag{14}$$

where

$$I_{ii}^{pq} = \begin{cases} 1 & \text{if} \quad I_i^p = I_i^q \\ 0 & \text{if} \quad I_i^p \neq I_i^q \end{cases}, \tag{15}$$

N is the total number of independent chemical units correlated, and I_i^p and I_i^q are the states (1 if present, 0 if absent, etc.) of chemical unit number i in simulation runs p and q, respectively. As the 8 individual structures may be placed inside a 2 × 2 × 2 superlattice in a number of equivalent ways, all possibilities were considered and the largest similarity value chosen. Table 4 shows the deduced similarities between the 10 final structures, and all values are close to 0.7. Since this value is larger than the lowest-possible value of about 0.5, many features are conserved throughout the 10 structure, indicating that essential correlations have indeed been identified by the Monte Carlo procedure. Furthermore, this value is less than unity, indicating that there are some significant differences between the 10 structures and hence many correlations are possibly not of great significance. Such a result is a requirement if just any one configuration of a 2 × 2 × 2 superlattice is to provide a useful description of the actual inhomogeneity evidenced in the original X-ray refinement. Simulation J had the lowest total interaction energy and the relative energies of the other runs are plotted against the similarity to this structure in Fig. 2. It is found that the total energy is not dependent on the similarity, another desired feature. This result shows that the Monte Carlo simulation has sampled and obtained low energy structures in different

Table 4 Comparison of the similarities between the 10 different Monte Carlo runs

	A	B	C	D	E	F	G	H	I	J
A	1.00	0.72	0.72	0.73	0.73	0.73	0.72	0.72	0.72	0.72
B	0.72	1.00	0.73	0.71	0.71	0.73	0.74	0.73	0.72	0.72
C	0.72	0.73	1.00	0.71	0.71	0.72	0.72	0.74	0.72	0.73
D	0.73	0.71	0.71	1.00	0.71	0.71	0.72	0.72	0.72	0.73
E	0.73	0.71	0.71	0.71	1.00	0.70	0.72	0.73	0.71	0.71
F	0.73	0.73	0.72	0.71	0.70	1.00	0.71	0.72	0.71	0.72
G	0.72	0.74	0.72	0.72	0.72	0.71	1.00	0.73	0.72	0.72
H	0.72	0.73	0.74	0.72	0.73	0.72	0.73	1.00	0.71	0.71
I	0.72	0.72	0.72	0.72	0.71	0.71	0.72	0.71	1.00	0.72
J	0.72	0.72	0.73	0.73	0.71	0.72	0.72	0.71	0.72	1.00

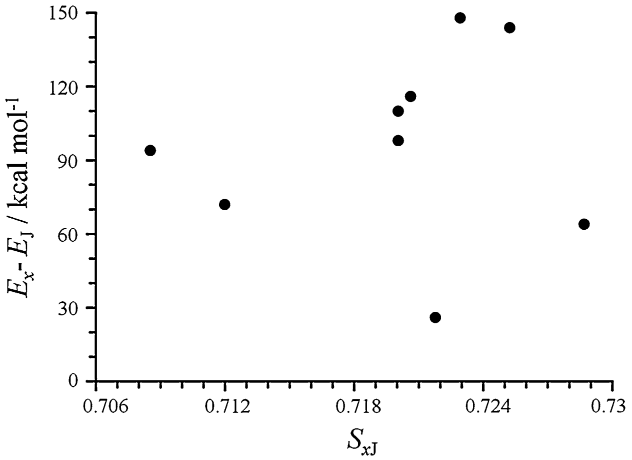


Fig. 2 The energy of quenched optimized 2 × 2 × 2 superlattice structure x, less that of structure J, as a function of the similarity of the two structures. This energy arises from variations in 840 *df*

parts of the configuration space. The energy variation between the 10 structures appears quite large, but there are 840 *df* in the Monte Carlo simulations, and so the energy spread per degree of freedom is just 0.18 kcal mol^{-1}, less than thermal energy at 300 K (0.3 kcal mol^{-1} per degree of freedom). Nevertheless, this result indicates that it is worthwhile maximizing the number of quenches made in calculations of this type.

Further, the calculated wide dispersion in the total energies indicates that the energy landscape is composed of a large number of local minima. All 10 structures have been determined to be local minima in the parameter space by considering all possible single-parameter variations, with the energy increasing in all cases. Figure 3 shows the energy increases determined for the lowest-energy structure, J, as well as for the highest-energy structure, G. These profiles are quite similar, as indeed are also those for the other 8 configurations. The left-most peaks ($\Delta E = 0 - 1\,\mathrm{kcal\,mol}^{-1}$) arise typically from ILE, TRP, PRO and VAL residues, uncharged groups that in specific instances occupy different sites without steric hindrance. Strong correlations manifest through large (say >10 kcal mol^{-1}) to extreme (100 kcal mol^{-1}) costs for a single exchange, mostly due to steric repulsion and ion–ion interactions. In summary, this energy analysis thus also indicates that while the 2 × 2 × 2 superlattice model does capture many (previously unknown) essential correlations, further expansion in its size would yield an improved description. However, such an expansion would undesirably add more parameters to a subsequent X-ray refinement, so careful choice of the superlattice size must be made.

5 Possible changes of the solvent atoms in the X-ray structure

In this section, we consider the chemical feasibility of the partial occupancies attributed in 2VB1 to many of the nitrate ions, ethylene glycol molecules, and water molecules. In earlier models for the triclinic lysozyme, some of the molecules identified in 2VB1 as nitrate ions and ethylene glycol molecules were identified as water molecules instead [120], and it is clear that only very high-quality data and analyses can accurately discriminate between various explanations of observed electron density in these regions. In this section, we consider variations of the proposed atomic structures for the nitrates, ethylene glycols and water molecules, examining them in terms of both the associated energetics and induced changes to the *R* factor.

Figure 4 shows the calculated AMBER energy for inserting water molecules into holes left by vacant nitrate ions, ethylene glycol molecules or water molecules in the optimized 2 × 2 × 2 superlattice structure J representing the 2VB1 optimized coordinates. For a missing nitrate ion, one water molecule is used to fill the resulting cavity, whereas two water molecules are used to fill an ethylene glycol hole.

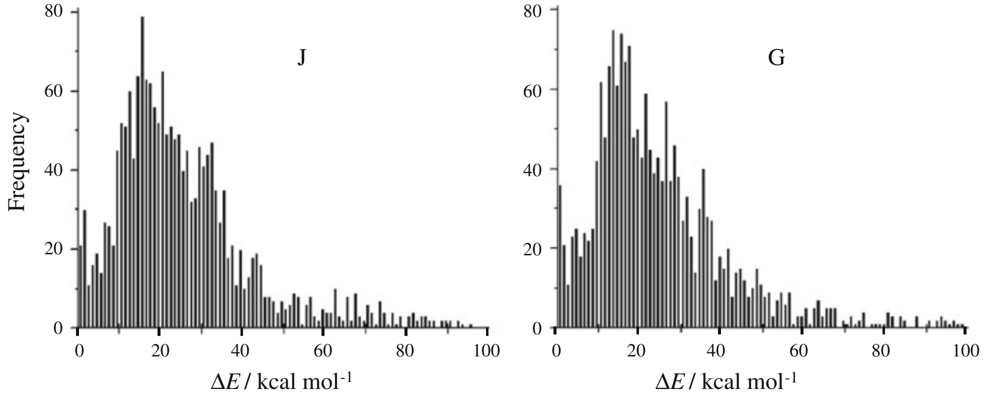


Fig. 3 Frequency of occurrence as a function of energy cost for all possible single flips of elements of the quenched structures obtained from simulations J (lowest energy) and G (highest energy)

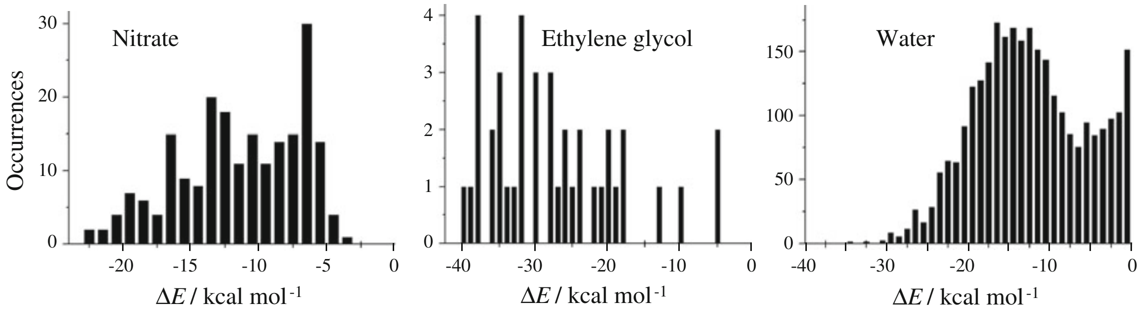


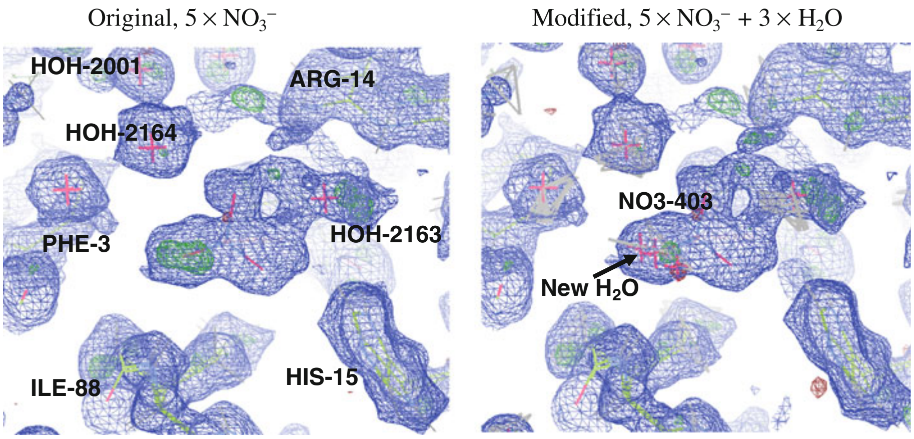
Fig. 4 Distribution of the energy change when water molecules were added to the holes due to missing nitrate ions (1 water), ethylene glycol molecules (2 waters) and water molecules (1 water)

Addition of a water molecule from the gas phase to a nitrate hole was always found to be attractive, but the energy change ΔE varied from -3 to -23 kcal mol^{-1}. However, the calculated energy for liquid water is -9.9 kcal mol^{-1} [115], close to the observed enthalpy change for formation of liquid water of -10.5 kcal mol^{-1} [121], and so, ignoring entropy changes between the liquid and protein environments, only insertions that are more exothermic than this are likely to proceed. The calculations thus indicate that many of the nitrate vacancies in the 2VB1 structure will in fact be filled with water. For ethylene glycol holes, Fig. 4 shows that a similar situation arises with most substitutions resulting in exothermicities of up to 40 kcal mol^{-1}, well in excess of the 20 kcal mol^{-1} required to extract two water molecules from the bulk liquid. However, even larger exothermicities per molecule are also depicted in Fig. 4 for the situation in which a water molecule fills a water hole, up to 35 kcal mol^{-1}. Strong interactions of water in these holes often arise from direct hydrogen-bonding interactions with charged residues such as LYS, ARG, GLU and ASP, and the desolvation of these residues seems highly unlikely.

A limitation of these calculations is that they ignore the 17.5% of the volume of the unit cell for which no atomic structure has been proposed. Some of the sites considered will be near this vacant volume and so may be poorly described, missing perhaps additional local attractive forces. The long-range effect of dielectric media is to mitigate the effects of electrostatic forces, however, and so could significantly reduce some of the calculated water substitution energies. Conversely, other effects such as thermal motion of solvent could easily enhance these energies.

Despite such uncertainties, the emphatic result from these calculations, that water molecules should spontaneously fill many of the holes created by partially occupied water sites, stands in contrast to the current experimental interpretation of the occupancies of the water sites: simply increasing the water occupancies as demanded by the calculations would, by construct, automatically lead to an increase in the refined *R* factor. This implies that the perceived electron density from Eq. 2 in the vicinity of these water sites is underestimated, an effect possibly attributable to poor calculated phases caused by an incorrect description of other either explicitly or implicitly represented regions of the unit cell.

Such a dramatic contradiction of calculated and refined properties is not automatically implied by the prediction that water molecules should fill nitrate or ethylene glycol holes, however. The changes to the assigned compositions of these sites with improved structural refinement [98] indicate that the total electron densities and its distribution in these regions are difficult to determine. Adding water to



Original, $5 \times NO_3^-$

Modified, $5 \times NO_3^- + 3 \times H_2O$

Fig. 5 Electron density maps (*blue*: $2F_O - F_C$, *green*: $F_O - F_C$ positive, *red*: $F_O - F_C$ negative) around NO3-403 as is the 2VB1 refinement, showing significant unmodeled electron density on one side of the ion, and after the vacant, nitrate locations in the $2 \times 2 \times 2$ supercell are occupied by additional water molecules

previously vacant cells could even lead to a decrease in the *R* factor, indicating, by all measures, an improved structure.

In a second set of Monte Carlo simulations, all nitrate ions and ethylene glycol holes were filled with water, setting the occupancies of each species so as to conserve the total number of electrons perceived in the region. Because of the nonlinear nature of crystallographic refinement with calculated phase factors used in generating density maps (Eqs. 5–6), such conservation of net charge is not a requirement, but here it is used in a minimalist approach to modification of the original structural model. A sequence of simulations with reducing temperature was performed as before, finally quenching the structure to 0 K. This structure was verified to be a local minimum in configuration space.

When two water molecules filled an ethylene glycol hole, the optimized structure always located the waters on one side of the hole, producing an electron density map that looks quite different from the original (that coming from assuming either ethylene glycol or else nothing occupied the cavity). For the nitrate holes, the water molecules often located in regions of unaccounted electron density in the original maps, however, suggesting that an improved structure is possible. Specifically, nine nitrate ions are identified in 2VB1, named NO3-401 to NO3-409. Of these, five (NO3-402, NO3-403, NO3-405, NO3-407 and NO3-409) have $F_O - F_C$ maps showing significant observed electron density that is not accounted for by the structural model; these groups have occupancies of 73, 62, 68, 69 and 49%, respectively. Conversely, the other four nitrates (NO3-401, NO3-404, NO3-406 and NO3-408) with occupancies of 73, 52, 85 and 100%, respectively, appear to account for most of the observed electron density. Here, we focus only on the nitrates with unaccounted electron density, specifically NO3-403 whose $2F_O - F_C$ and $F_O - F_C$ maps from structure J are shown in Fig. 5. The appearance of substantial unaccounted electron density near one of the nitrate oxygen atoms is readily apparent.

Nitrate NO3-403, with 62% occupancy (thus present in only 5 of the 8 cells in the superlattice structure J) forms a hydrogen bond to ILE-88, and indeed, it is near this residue that the unaccounted electron density is centered. Other close residues include ARG-14, while LYS-1 is more distant, and many water molecules are also found nearby. Most significant, however, is ASP-87 that is located above the plane of the nitrate molecule and hence not visualized in Fig. 5. This residue has two observed conformations, one that is in van der Waals contact with NO3-403 with 82% occupancy, while the less-prevalent conformer is most distant, but in supercell J the vacant NO3-403 cells all have ASP-87 very close. As additional electron density needs to be accounted for, water molecules were added to the three cells unoccupied by NO3-403 rather than utilizing the simplistic density-conserving scheme used in the previous Monte Carlo simulations. These added molecule interact strongly with ASP-87, ILE-88 and the surrounding water molecules.

The structures of the added water molecules, as well as the surrounding water molecule orientations, were then optimized using our ONIOM-based linear-scaling DFT scheme [4] implemented in GAUSSIAN-09 [114] using the PW91 density functional [122] and the 6-31G* basis set for all atoms except those bearing anions, for which 6-31+G* was used. This scheme breaks a full optimization down into small pieces, and for the three critical units, this included 27, 29 and 60 nearby water molecules and 1213, 1290 and 1915 atoms in total; such significant variation in the size of the critical units occurs owing to unoccupied water sites in structure J as complete local hydrogen-bonded networks are included in all calculations. The energies for water molecules placed into the three holes were thus calculated to be −18, −20 and −21 kcal mol^{-1}, typical of the values found in the previous Monte Carlo simulations using the Amber force field. Single-point energy calculations using an external reaction field to model neglected paths of the structure has no effect on these energy differences, as one would expect given that the calculations involve large regions of explicitly included matter (effective cavity radii 13–15 Å). Energetic considerations thus strongly favor occupation of these three nitrate holes.

Figure 5 shows also the density maps generated using the revised structure with water molecules filling the three NO3-403 holes. It is clear from the figure that this modification results in significant reduction of the unaccounted electron density, but this is not a good measure of the quality of the modified model as the nonlinearities in the refinement procedure can easily result in the *R* factor increasing despite the visualized apparent improvement to the model. However, using an isotropic temperature factor of 8 for the added water molecules, the *R* factors are found to decrease by 0.01%, a value that is significant given that artificially changing the occupancy of NO3-403 from 5/8 to 1 increase the *R* factors by 0.01% and that decreasing the occupancy to 0 increases *R* by 0.08%. Hence, both energetic and crystallographic measures indicate that water molecules fill the site when NO3-403 is not present.

The unaccounted electron density manifest for the original structure in Fig. 5 can readily be accounted for using anisotropic B factor chosen for each atom in the nitrate ion, but such a choice would not be consistent with the actual motions of a nitrate ion. Hence, we see that while the anisotropic X-ray structure refinement resulted in a deep local minimum in its parameter space, some of the deduced structural features may not necessarily be realistic. While anisotropic *B* factors can account for real non-spherical electron density and produce a dramatic decrease in the *R* factor in 2VB1, they may do this simply by overparameterization, eroding the physical meaning of all deduced parameters. We thus recommend that use of anisotropic thermal displacement parameters be minimized during subsequent ab initio refinement.

6 Conclusions

A Monte Carlo procedure utilizing the AMBER force field was developed to find an ensemble of 8 structures represented in a 2 × 2 × 2 superlattice that embodies in a properly correlated fashion all of the structural variations previously reported in the 2VB1 structure of lysozyme crystal. This calculation included explicitly all correlations between atoms identified by high-resolution X-ray refinement methods [98] and deduced many new correlations manifested through large calculated interaction energies. Statistical analyses indicate that an average structure (with only some local multiple conformations) does include many key correlations and suggest that expanded superlattices could yield improved results. Such superlattices are a critical requirement before ab initio electronic structure computational methods can be embedded inside any X-ray refinement packages, but expansion is limited as an undesired feature of their inclusion is an increase in the number of free parameters during refinement. Further extensions of this method are required, however, in order to include the 17.5% of the crystal volume that was not explicitly represented in the original X-ray structure.

The development of such a structure by necessity introduces chemically meaningful constraints, constraints not present in the original X-ray refinement. In this case, these constraints ensured that all atoms appearing in different chain conformations have the same occupancy, an important physical property, yet their introduction by necessity *decreases* the quality of the structural refinement as perceived by increases in *R* and R_{free}. In addition, use of the 2 × 2 × 2 superlattice does decrease the quality of the refinement in real terms by limiting the occupancies to a discrete set of values, multiples of one-eighth for the present example. The quality of ab initio refined structures needs to be compared to the values of *R* and R_{free} determined at this stage, not the original unconstrained values.

Ab initio refinement will, in general, require some modifications to the way X-ray refinements are completed. In particular, *B* factors should be modified to include only the effects of thermal motion and not the effects of multiple conformations as multiple conformations are now being included explicitly. Such isotropic or anisotropic displacement factors could be determined ab initio using say molecular dynamics simulations and hence not appear as free parameters during X-ray structure refinement [81, 86, 123].

The development of a single 2 × 2 × 2 superlattice representing an ensemble of 8 structures to describe conformational variations by default allows for individual coordinates for all atoms in each of the 8 cell copies. While this is demanded for regions showing significant variation, it may not be justified for other regions in which one single structure dominates the X-ray diffraction. In this region, replicated copies of all coordinates could be maintained in some or all of the 8 cells, reducing considerably the number of free parameters in an X-ray refinement. Established methods for treating the *B* factors would be appropriate for such atoms.

Initially, superlattices were developed embodying most the features of the original 2VB1 structure for lysozyme. Each considered superlattice thus generated essentially the same *R* factor and so was equally consistent with the raw X-ray diffraction data. These structures varied considerably in terms of their perceived total energies although the maximum energy difference per degree of freedom was quite small. Many of the geometrical fluctuations apparent in the 2VB1 structure were found to be strongly correlated with each other, while many others were found to be uncorrelated. The lowest-energy structure found was considered as a starting point for future ab initio X-ray refinement. Subsequently, however, the question as to the chemical feasibility of the originally deduced nitrate, ethylene glycol and water occupancies was examined. The calculations predict that it is highly exothermic to fill many of these holes with water molecules taken from bulk solution, indicating that the original occupancies are not actually feasible. While authoritative calculations require an improved treatment of missing solvent atoms than is used herein, one specific example is considered, nitrate NO3-403, for which DFT calculations predict that water fills the nitrate holes while this change actually does result in a decrease in the *R* factor, indicating that indeed an improved structural model is produced.

Such a simple correlation between predicted and actual model improvements will not be universal, however, and it could be that the application of ab initio computational methods to X-ray structure refinement does not easily lead to improved measures of structure quality such as simply a decrease in R_{free}. Because of the nonlinearities inherent in X-ray diffraction modeling, it is feasible that such a decrease could only be obtained by taking the whole optimized structure from its original local minimum in parameter space and transforming it to a quite different local minimum, affecting many of the atomic parameters of the structure, especially perhaps solvent and unassigned regions. This is a difficult task.

Acknowledgments We thank Aaron McGrath for technical advice, Zbigniew Dauter from the Argonne National Laboratory Biosciences Division for providing detailed information regarding the refinement of 2VB1, the Australian Research Council for funding this research, and National Computer Infrastructure (NCI) and Australian Centre for Advanced Computing and Communications (AC3) for computing resources.

References

1. Engh RA, Huber R (1991) Acta Crystallogr A 47:392–400
2. Sheldrick G, Schneider T (1997) SHELXL: high-resolution refinement. Methods Enzymol 277:319–343
3. Murshudov GN, Vagin AA, Dobson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 53:240–255
4. Canfield P, Dahlbom MG, Hush N, Reimers JR (2006) Density-functional geometry optimization of the 150000-atom photosystem-I trimer. J Chem Phys 124:024301
5. Kleywegt GJ (1999) Experimental assessment of differences between related protein crystal structures. Acta Crystallogr D Biol Crystallogr 55:1878–1884
6. Cruickshank DWJ (1999) Remarks about protein structure precision. Acta Crystallogr D Biol Crystallogr 55:583–601
7. DePristo MA, De Bakker PIW, Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. Structure 12:831–838
8. Jaskolski M, Gilski M, Dauter Z, Wlodawer A (2007) Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? Acta Crystallogr D Biol Crystallogr 63:611–620
9. Chen J, Brooks CL (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins Struct Funct Bioinform 67:922–930
10. Karplus PA, Shapovalov MV, Dunbrack RL, Berkholz DS (2008) A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions. Acta Crystallogr D Biol Crystallogr 64:335–336
11. Rashin AA, Rashin AHL, Jernigan RL (2009) Protein flexibility: coordinate uncertainties and interpretation of structural differences. Acta Crystallogr D Biol Crystallogr 65:1140–1161
12. Jaskolski M (2010) From atomic resolution to molecular giants: an overview of crystallographic studies of biological macromolecules with synchrotron radiation. Acta Physica Polonica A 117:257–263
13. Eyal E, Gerzon S, Potapov V, Edelman M, Sobolev V (2005) The limit of accuracy of protein modeling: influence of crystal packing on protein structure. J Mol Biol 351:431–442
14. Konnert JH (1976) A restrained parameter structure-factor least-squares refinement procedure for large asymmetric units. Acta Crystallogr A 32:614–617
15. Hendrickson WA, Konnert JH (1979) Stereochemically restrained crystallographic least-squares refinement of macromolecule structures. In: Srinivasan R (ed) Biomolecular structure, conformation, function, and evolution, vol 1. Pergamon Press, Oxford, pp 43–57
16. Konnert JH, Hendrickson WA (1980) A restrained-parameter thermal-factor refinement procedure. Acta Crystallogr A 36:344–350
17. Hendrickson WA (1985) Stereochemically restrained refinement of macromolecular structures. Methods Enzymol 115:252–270
18. Jack A, Levitt M (1978) Refinement of large structures by simultaneous minimization of energy and R factor. Acta Crystallogr A 34:931–935
19. Brunger AT, Kuriyan J, Karplus M (1987) Crystallographic R factor refinement by molecular dynamics. Science 235:458–460
20. Ohta K, Yoshioka Y, Morokuma K, Kitaura K (1983) The effective fragment potential method. An approximate ab initio mo method for large molecules. Chem Phys Lett 101:12–17
21. Stewart JJP (1996) Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. Int J Quantum Chem 58:133–146
22. White CA, Johnson BG, Gill PMW, Head-Gordon M (1996) Linear scaling density functional calculations via the continuous fast multipole method. Chem Phys Lett 253:268–278
23. Stewart JJP (1997) Calculation of the geometry of a small protein using semiempirical methods. J Mol Struct Theochem 401:195–205
24. Lee TS, Lewis JP, Yang W (1998) Linear-scaling quantum mechanical calculations of biological molecules: the divide-and-conquer approach. Comput Mater Sci 12:259–277
25. Van Alsenoy C, Yu CH, Peeters A, Martin JML, Schäfer L (1998) Ab initio geometry determinations of proteins. 1. Crambin. J Phys Chem A 102:2246–2251
26. Artacho E, Sánchez-Portal D, Ordejón P, García A, Soler JM (1999) Linear-scaling ab initio calculations for large and complex systems. Phys Status Solidi B 215:809–817
27. Sato F, Yoshihiro T, Era M, Kashiwagi H (2001) Calculation of all-electron wavefunction of hemoprotein cytochrome c by density functional theory. Chem Phys Lett 341:645–651
28. Inaba T, Tahara S, Nisikawa N, Kashiwagi H, Sato F (2005) All-electron density functional calculation on insulin with quasi-canonical localized orbitals. J Comput Chem 26:987–993

29. Wada M, Sakurai M (2005) A quantum chemical method for rapid optimization of protein structures. J Comput Chem 26:160–168
30. Li S, Shen J, Li W, Jiang Y (2006) An efficient implementation of the "cluster-in-molecule" approach for local electron correlation calculations. J Chem Phys 125:074109
31. Sale P, Høst S, Thøgersen L, Jørgensen P, Manninen P, Olsen J, Jansik B, Reine S, Pawlowski F, Tellgren E, Helgaker T, Coriani S (2007) Linear-scaling implementation of molecular electronic self-consistent field theory. J Chem Phys 126:114110
32. Cankurtaran BO, Gale JD, Ford MJ (2008) First principles calculations using density matrix divide-and-conquer within the SIESTA methodology. J Phys Condens Matter 20:294208
33. Stewart JJP (2009) Application of the PM6 method to modeling proteins. J Mol Model 15:765–805
34. Gordon MS, Mullin JM, Pruitt SR, Roskop LB, Slipchenko LV, Boatz JA (2009) Accurate methods for large molecular systems. J Phys Chem B 113:9646–9663
35. Fedorov DG, Alexeev Y, Kitaura K (2010) Geometry optimization of the active site of a large system with the fragment molecular orbital method. J Phys Chem Lett 2:282–288
36. Kobayashi M, Kunisada T, Akama T, Sakura D, Nakai H (2010) Reconsidering an analytical gradient expression within a divide-and-conquer self-consistent field approach: exact formula and its approximate treatment. J Chem Phys 134:034105
37. Mayhall NJ, Raghavachari K (2010) Molecules-in-molecules: an extrapolated fragment-based approach for accurate calculations on large molecules and materials. J Chem Theory Comput 7:1336–1343
38. Nagata T, Brorsen K, Fedorov DG, Kitaura K, Gordon MS (2010) Fully analytic energy gradient in the fragment molecular orbital method. J Chem Phys 134:124115
39. Reine S, Krapp A, Iozzi MF, Bakken V, Helgaker T, Pawowski F, Saek P (2010) An efficient density-functional-theory force evaluation for large molecular systems. J Chem Phys 133:044102
40. Bylaska E, Tsemekhman K, Govind N, Valiev M (2011) Large-scale plane-wave-based density-functional theory: formalism, parallelization, and applications. In: Reimers JR (ed) Computational methods for large systems: electronic structure approaches for biotechnology and nanotechnology. Wiley, Hoboken, pp 77–116
41. Gale JD (2011) SIESTA: a linear-scaling method for density functional calculations. In: Reimers JR (ed) Computational methods for large systems: electronic structure approaches for biotechnology and nanotechnology. Wiley, Hoboken, pp 45–74
42. Li W, Hua W, Fang T, Li S (2011) The energy-based fragmentation approach for computing total energies, structures, and molecular properties of large systems at the ab initio levels. In: Reimers JR (ed) Computational methods for large systems: electronic structure approaches for biotechnology and nanotechnology. Wiley, Hoboken, pp 227–258
43. Clark T, Stewart JJP (2011) MNDO-like semiempirical molecular orbital theory and its application to large systems. In: Reimers JR (ed) Computational methods for large systems: electronic structure approaches for biotechnology and nanotechnology. Wiley, Hoboken, pp 259–286
44. Elstner M, Gaus M (2011) The self-consistent-charge density-functional tight-binding (SCC-DFTB) method: an efficient approximation of density functional theory. In: Reimers JR (ed) Computational methods for large systems: electronic structure approaches for biotechnology and nanotechnology. Wiley, Hoboken, pp 287–308
45. Zimmerli U, Parrinello M, Koumoutsakos P (2004) Dispersion corrections to density functionals for water aromatic interactions. J Chem Phys 120:2693–2699
46. Antony J, Grimme S (2006) Density functional theory including dispersion corrections for intermolecular interactions in a large benchmark set of biologically relevant molecules. Phys Chem Chem Phys 8:5287–5293
47. Grimme S, Antony J, Schwabe T, Mück-Lichtenfeld C (2007) Density functional theory with dispersion corrections for supramolecular structures, aggregates, and complexes of (bio)organic molecules. Org Biomol Chem 5:741–758
48. Zhao Y, Truhlar DG (2007) Density functionals for noncovalent interaction energies of biological importance. J Chem Theory Comput 3:289–300
49. Murdachaew G, De Gironcoli S, Scoles G (2008) Toward an accurate and efficient theory of physisorption. I. Development of an augmented density-functional theory model. J Phys Chem A 112:9993–10005
50. DiLabio GA (2008) Accurate treatment of van der Waals interactions using standard density functional theory methods with effective core-type potentials: application to carbon-containing dimers. Chem Phys Lett 455:348–353
51. Gräfenstein J, Cremer D (2009) An efficient algorithm for the density-functional theory treatment of dispersion interactions. J Chem Phys 130:124105
52. Liu Y, Goddard WA (2009) A universal damping function for empirical dispersion correction on density functional theory. Mater Trans 50:1664–1670
53. Sato T, Nakai H (2009) Density functional method including weak interactions: dispersion coefficients based on the local response approximation. J Chem Phys 131:224104
54. Foster ME, Sohlberg K (2010) Empirically corrected DFT and semi-empirical methods for non-bonding interactions. Phys Chem Chem Phys 12:307–322
55. Grimme S, Antony J, Ehrlich S, Krieg H (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. J Chem Phys 132:154104
56. Riley KE, Pitoňák M, Jurečka P, Hobza P (2010) Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories. Chem Rev 110:5023–5063
57. MacKie ID, Dilabio GA (2010) Accurate dispersion interactions from standard density-functional theory methods with small basis sets. Phys Chem Chem Phys 12:6092–6098
58. Goerigk L, Grimme S (2011) A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. Phys Chem Chem Phys 13:6670–6688
59. Grimme S, Ehrlich S, Goerigk L (2011) Effect of the damping function in dispersion corrected density functional theory. J Comput Chem 32:1456–1465
60. Steinmann SN, Corminboeuf C (2011) A density dependent dispersion correction. Chimia 65:240–244
61. Zhao Y, Truhlar DG (2011) Density functional theory for reaction energies: test of meta and hybrid meta functionals, range-separated functionals, and other high-performance functionals. J Chem Theory Comput 7:669–676
62. Brüning J, Alig E, Van De Streek J, Schmidt MU (2011) The use of dispersion-corrected DFT calculations to prevent an incorrect structure determination from powder data: The case of acetolone, C 11H11N3O3. Z Kristallogr 226:476–482
63. Ryde U, Olsen L, Nilsson K (2002) Quantum chemical geometry optimizations in proteins using crystallographic raw data. J Comput Chem 23:1058–1070
64. Ryde U, Nilsson K (2003) Quantum chemistry can locally improve protein crystal structures. J Am Chem Soc 125:14232–14233

65. Ryde U (2007) Accurate metal-site structures in proteins obtained by combining experimental data and quantum chemistry. Dalton Trans 607–625
66. Ryde U, Greco C, De Gioia L (2010) Quantum refinement of [FeFe] hydrogenase indicates a dithiomethylamine ligand. J Am Chem Soc 132:4512–4513
67. Yu N, Yennawar HP, Merz KM Jr (2005) Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics. Acta Crystallogr D Biol Crystallogr 61:322–332
68. Yu N, Li X, Cui G, Hayik SA, Merz KM Jr (2006) Critical assessment of quantum mechanics based energy restraints in protein crystal structure refinement. Protein Sci 15:2773–2784
69. Yu N, Hayik SA, Wang B, Liao N, Reynolds CH, Merz KM Jr (2006) Assigning the protonation states of the key aspartates in beta-secretase using QM/MM X-ray structure refinement. J Chem Theory Comput 2:1057–1069
70. Van Der Vaart A, Suárez D, Merz KM Jr (2000) Critical assessment of the performance of the semiempirical divide and conquer method for single point calculations and geometry optimizations of large chemical systems. J Chem Phys 113:10512–10523
71. Van Der Vaart A, Gogonea V, Dixon SL, Merz KM Jr (2000) Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. J Comput Chem 21:1494–1504
72. Dixon SL, Merz KM Jr (1997) Fast, accurate semiempirical molecular orbital calculations for macromolecules. J Chem Phys 107:879–893
73. Dixon SL, Merz KM Jr (1996) Semiempirical molecular orbital calculations with linear system size scaling. J Chem Phys 104:6643–6649
74. Pellegrini M, Grønbech-Jensen N, Kelly JA, Pfluegl GMU, Yeates TO (1997) Highly constrained multiple-copy refinement of protein crystal structures. Proteins Struct Funct Bioinform 29:426–432
75. Levin EJ, Kondrashov DA, Wesenberg GE, Phillips GN Jr (2007) Ensemble refinement of protein crystal structures: validation and application. Structure 15:1040–1052
76. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Adams PD, Moriarty NW, Zwart P, Read RJ, Turk D, Hung LW (2007) Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. Acta Crystallogr D Biol Crystallogr 63:597–610
77. Stewart KA, Robinson DA, Lapthorn AJ (2008) Type II dehydroquinase: molecular replacement with many copies. Acta Crystallogr D Biol Crystallogr 64:108–118
78. Stewart JJP (2008) Application of the PM6 method to modeling the solid state. J Mol Model 14:499–535
79. Genheden S, Ryde U (2011) A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations. J Comput Chem 32:187–195
80. Genheden S, Diehl C, Akke M, Ryde U (2010) Starting-condition dependence of order parameters derived from molecular dynamics simulations. J Chem Theory Comput 6:2176–2190
81. Delarue M (2007) Dealing with structural variability in molecular replacement and crystallographic refinement through normal-mode analysis. Acta Crystallogr D Biol Crystallogr 64:40–48
82. Knight JL, Zhou Z, Gallicchio E, Himmel DM, Friesner RA, Arnold E, Levy RM (2008) Exploring structural variability in X-ray crystallographic models using protein local optimization by torsion-angle sampling. Acta Crystallogr D Biol Crystallogr 64:383–396
83. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP (2008) Toward better refinement of comparative models: predicting loops in inexact environments. Proteins Struct Funct Genet 72:959–971
84. Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu G, Van Den Bedem H, Latombe JC, Halperin-Landsberg I, Altman RB (2008) Efficient algorithms to explore conformation spaces of flexible protein loops. IEEE/ACM Trans Comput Biol Bioinform 5:534–545
85. Lindorff-Larsen K, Ferkinghoff-Borg J (2009) Similarity measures for protein ensembles. PLoS One 4:e4203
86. Yang L, Song G, Jernigan RL (2009) Comparisons of experimental and computed protein anisotropic temperature factors. Proteins Struct Funct Bioinform 76:164–175
87. Dhanik A, Van Den Bedem H, Deacon A, Latombe JC (2010) Modeling structural heterogeneity in proteins from X-ray data. Springer Tracts Adv Robot 57:551–566
88. Schwander P, Fung R, Phillips GN Jr, Ourmazd A (2010) Mapping the conformations of biological assemblies. New J Phys 12:035007
89. Lang PT, Ng HL, Fraser JS, Corn JE, Echols N, Sales M, Holton JM, Alber T (2010) Automated electron-density sampling reveals widespread conformational polymorphism in proteins. Protein Sci 19:1420–1431
90. Kohn JE, Afonine PV, Ruscio JZ, Adams PD, Head-Gordon T (2010) Evidence of functional protein dynamics from X-ray crystallographic ensembles. PLoS Comput Biol 6:e1000911
91. Tyka MD, Keedy DA, André I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D (2011) Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 405:607–618
92. Ramelot TA, Raman S, Kuzin AP, Xiao R, Ma L-C, Acton TB, Hunt JF, Montelione GT, Baker D, Kennedy MA (2009) Improving NMR protein structure quality by Rosetta refinement: a molecular replacement study. Proteins Struct Funct Bioinform 75:147–167
93. Fleming A (1922) On a remarkable bacteriolytic element found in tissues and secretions. Proc R Soc Ser B 93:306–317
94. Blake CCF, Fenn RH, North ACT, Phillips DC, Poljak RJ (1962) Structure of lysozyme. Nature 196:1173–1176
95. Berman HM, Henrick K, Nakamura H (2003) Announcing the world wide protein data bank. Nat Struct Biol 10:980
96. Vocadlo DJ, Davies GJ, Laine R, Withers SG (2001) Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. Nature 412:835–838
97. Bottoni A, Miscione GP, De Vivo M (2005) A theoretical DFT investigation of the lysozyme mechanism: computational evidence for a covalent intermediate pathway. Proteins Struct Funct Genet 59:118–130
98. Wang J, Dauter M, Alkire R, Joachimiak A, Dauter Z (2007) Triclinic lysozyme at 0.65 a resolution. Acta Crystallogr D Biol Crystallogr 63:1254–1268
99. Blundell TL, Johnson LN (1976) Protein crystallography. Academic Press, London
100. Chapman HN, Fromme P, Barty A, White TA, Kirian RA, Aquila A, Hunter MS, Schulz J, DePonte DP, Weierstall U, Doak RB, Maia FRNC, Martin AV, Schlichting I, Lomb L, Coppola N, Shoeman RL, Epp SW, Hartmann R, Rolles D, Rudenko A, Foucar L, Kimmel N, Weidenspointner G, Holl P, Liang M, Barthelmess M, Caleman C, Boutet S, Bogan MJ, Krzywinski J, Bostedt C, Bajt S, Gumprecht L, Rudek B, Erk B, Schmidt C, Homke A, Reich C, Pietschner D, Struder L, Hauser G, Gorke H, Ullrich J, Herrmann S, Schaller G, Schopper F, Soltau H, Kuhnel K-U, Messerschmidt M, Bozek JD, Hau-Riege SP, Frank M, Hampton CY, Sierra RG, Starodub D, Williams GJ, Hajdu J, Timneanu N, Seibert MM, Andreasson J, Rocker A, Jonsson O, Svenda M, Stern S, Nass K, Andritschke R, Schroter C-D, Krasniqi F, Bott M, Schmidt KE, Wang X, Grotjohann I,

Holton JM, Barends TRM, Neutze R, Marchesini S, Fromme R, Schorb S, Rupp D, Adolph M, Gorkhover T, Andersson I, Hirsemann H, Potdevin G, Graafsma H, Nilsson B, Spence JCH (2011) Femtosecond X-ray protein nanocrystallography. Nature 470:73–77

101. Brunger AT (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355:472–475
102. Badger J (1997) Modeling and refinement of water molecules and disordered solvent. Methods Enzymol 277:344–352
103. Podjarny AD, Howard EI, Urzhumtsev A, Grigera JR (1997) A multicopy modeling of the water distribution in macromolecular crystals. Proteins Struct Funct Bioinform 28:303–312
104. Colominas C, Luque FJ, Orozco M (1999) Monte Carlo–MST: new strategy for representation of solvent configurational space in solution. J Comput Chem 20:665–678
105. Liu Y, Beveridge DL (2002) Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. Proteins Struct Funct Bioinform 46:128–146
106. Das B, Meirovitch H (2003) Solvation parameters for predicting the structure of surface loops in proteins: transferability and entropic effects. Proteins Struct Funct Bioinform 51:470–483
107. Hassan SA, Mehler EL, Zhang D, Weinstein H (2003) Molecular dynamics simulations of peptides and proteins with a continuum electrostatic model based on screened coulomb potentials. Proteins Struct Funct Bioinform 51:109–125
108. Dechene M, Wink G, Smith M, Swartz P, Mattos C (2009) Multiple solvent crystal structures of ribonuclease A: an assessment of the method. Proteins Struct Funct Bioinform 76:861–881
109. Kannan S, Zacharias M (2010) Application of biasing-potential replica-exchange simulations for loop modeling and refinement of proteins in explicit solvent. Proteins Struct Funct Bioinform 78:2809–2819
110. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta SJ, Weiner P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106:765–784
111. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM Jr, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197
112. Becke AD (1993) Density-functional thermochemistry. III. The role of exact exchange. J Chem Phys 98:5648–5652
113. Hehre WJ, Ditchfield R, Pople JA (1972) Self-consistent molecular orbital methods. XII. Further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules. J Chem Phys 56:2257–2261
114. Frisch MJ, Trucks GW, Schlegel HB et al (2009) Gaussian 09, revision A.02. Gaussian, Inc., Pittsburgh
115. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
116. Fischer RA (1935) The logic of inductive inference. J R Stat Soc A 98:39–54
117. Freeman GH, Halton JH (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. Biometrika 38:141–149
118. Agresti A (1990) Categorical data analysis. Wiley, New York
119. Bartoszyński R, Niewiadomska-Bugaj M (1996) Probability and statistical inference. Wiley, New York
120. Walsh MA, Schneider TR, Sieker LC, Dauter Z, Lamzin VS, Wilson KS (1998) Refinement of triclinic hen egg-white lysozyme at atomic resolution. Acta Crystallogr D Biol Crystallogr 54:522–546
121. Lide DR (ed) (2005) CRC handbook of chemistry and physics, 86th edn. CRC Press, Boca Raton
122. Perdew JP, Wang Y (1992) Accurate and simple analytic representation of the electron-gas correlation energy. Phys Rev B 45:13244–13249
123. Vitkup D, Ringe D, Karplus M, Petsko GA (2002) Why protein R-factors are so large: a self-consistent analysis. Proteins Struct Funct Genet 46:345–354