IUCrJ
ISSN 2052-2525
BIOLOGY | MEDICINE

Internal protein motions in molecular-dynamics simulations of Bragg and diffuse X-ray scattering

Michael E. Wall*

Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87505, USA. *Correspondence e-mail: mewall@lanl.gov

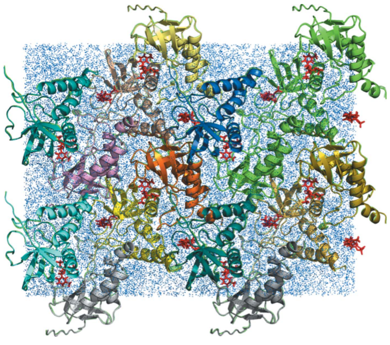
Received 26 September 2017
Accepted 8 January 2018

Edited by E. E. Lattman, University at Buffalo, USA

Keywords: diffuse scattering; protein crystallography; X-ray diffraction; molecular-dynamics simulation; protein conformational ensemble; staphylococcal nuclease; X-ray crystallography; molecular crystals; molecular simulation.

Supporting information: this article has supporting information at www.iucrj.org

Molecular-dynamics (MD) simulations of Bragg and diffuse X-ray scattering provide a means of obtaining experimentally validated models of protein conformational ensembles. This paper shows that compared with a single periodic unit-cell model, the accuracy of simulating diffuse scattering is increased when the crystal is modeled as a periodic supercell consisting of a 2 × 2 × 2 layout of eight unit cells. The MD simulations capture the general dependence of correlations on the separation of atoms. There is substantial agreement between the simulated Bragg reflections and the crystal structure; there are local deviations, however, indicating both the limitation of using a single structure to model disordered regions of the protein and local deviations of the average structure away from the crystal structure. Although it was anticipated that a simulation of longer duration might be required to achieve maximal agreement of the diffuse scattering calculation with the data using the supercell model, only a microsecond is required, the same as for the unit cell. Rigid protein motions only account for a minority fraction of the variation in atom positions from the simulation. The results indicate that protein crystal dynamics may be dominated by internal motions rather than packing interactions, and that MD simulations can be combined with Bragg and diffuse X-ray scattering to model the protein conformational ensemble.



OPEN ACCESS

1. Introduction

In X-ray diffraction from protein crystals, the sharp Bragg peaks are accompanied by diffuse scattering: streaks, cloudy features and other patterns between and under the peaks. Diffuse scattering comes from imperfections in the crystal such as diverse protein conformations. Unlike the Bragg diffraction, which is only sensitive to the mean charge density, diffuse scattering is sensitive to the spatial correlations in charge-density variations. Diffuse scattering therefore provides unique data for modeling protein conformational ensembles.

There is a longstanding interest in using diffuse scattering to validate molecular-dynamics (MD) simulations of protein crystals (Clarage *et al.*, 1995; Faure *et al.*, 1994; Héry *et al.*, 1998; Meinhold & Smith, 2005*a*,*b*, 2007; Wall *et al.*, 2014). Recent advances in computing now enable microsecond duration simulations of diffuse scattering (Wall *et al.*, 2014) and Bragg diffraction (Janowski *et al.*, 2013, 2016) that can overcome the limitations seen using MD trajectories of 10 ns or shorter (Clarage *et al.*, 1995; Meinhold & Smith, 2005*b*). In a microsecond simulation of a single staphylococcal nuclease unit cell (Wall *et al.*, 2014), much of the agreement between the MD simulation and diffuse data is owing to the isotropic component, a small-angle scattering-like pattern seen for all protein crystals. Agreement with this component is significant as it consists of roughly equal contributions from solvent and protein (Meinhold & Smith, 2005*a*; Wall *et al.*, 2014). The anisotropic component, which is about tenfold weaker, agrees less well with the simulation (linear correlation of 0.35–0.43). This gap in accuracy between the isotropic and anisotropic components must be closed because the anisotropic component is richly structured and can come almost entirely from the protein, creating possibilities for the validation of detailed models of protein motions. Accurate modeling of the anisotropic component is the key to unlocking the potential of diffuse scattering and MD simulations for modeling the conformational ensemble.

Wall *et al.* (2014) noted that the simulation of a single unit cell might limit the accuracy of MD models of diffuse scattering, and suggested that simulating a larger section of the crystal with several unit cells might improve the accuracy. Here, this idea is tested by constructing a periodic model of a 2 × 2 × 2 supercell of staphylococcal nuclease and performing an MD simulation of 5.1 µs duration. The linear correlation of the anisotropic component of diffuse intensity computed from this simulation with the data is 0.68, indicating that the supercell simulation greatly increases the accuracy of the model. Analysis using Patterson methods suggests that the distance dependence of the correlations is captured well. The mean structure factors from the simulation largely agree with the crystal structure; however, there are local deviations, suggesting a path to improve the MD model. The *B* factors from the simulation agree well with the crystal structure and improve on a TLS model. Similar to the unit-cell simulation, the agreement of the supercell model with the data reaches a maximum within a microsecond. This suggests the possibility that the simulation duration required for diffuse scattering calculations might become independent of the system size as it is increased beyond the length scale of the correlations. Finally, rigid-body motions explain only a minority component of the dynamics, indicating that internal motions may be more important than packing dynamics in MD simulations of protein crystals.

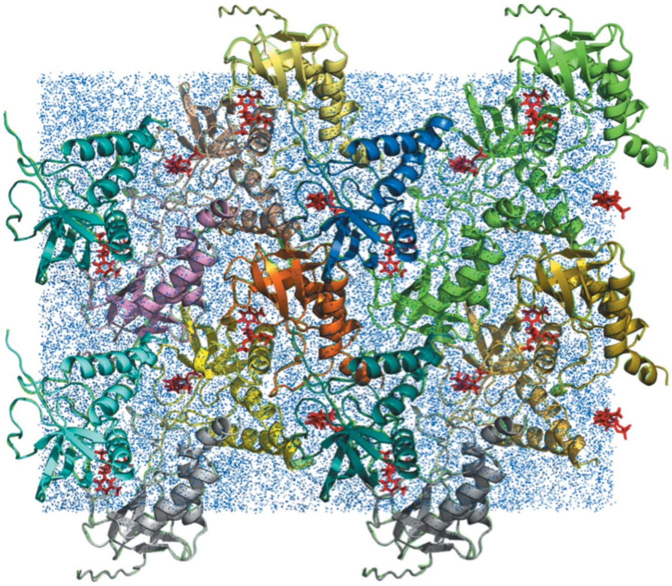


Figure 1
Staphylococcal nuclease supercell model. Eight unit cells containing four protein molecules each are arranged in a 2 × 2 × 2 layout. Protein chains are rendered as cartoons. The pdTP ligand is rendered as red sticks. Water atoms are indicated using speck-like blue spheres. The image was rendered using *PyMOL* (https://pymol.org/).

2. Methods

2.1. Molecular-dynamics simulation

A solvated crystalline model was created using PDB entry 1snc (Loll & Lattman, 1989). After stripping the waters, *UCSF Chimera* (https://www.cgl.ucsf.edu/chimera/) was used to add residues that were absent in the crystal structure. Using the context of the crystal structure as a guide, six missing residues at the N-terminus were modeled as a β-strand and eight missing residues at the C-terminus were modeled as an α-helix. A *P*1 unit cell of the protein and a thymidine 3′,5′-bisphosphate (pdTp) ligand was built in *UCSF Chimera* using space group $P4_1$ (four copies per cell). The unit-cell parameters were $a = b = 48.499$, $c = 63.430$ Å, $\alpha = \beta = \gamma = 90°$. The system was extended to a 2 × 2 × 2 supercell in a 96.998 × 96.998 × 126.860 Å right rectangular box using *PropPDB* from *AmberTools* (http://ambermd.org) (Fig. 1).

Molecular-dynamics simulations were performed using *GROMACS* (Berendsen *et al.*, 1995) v.5.0.2 (setup and first 4.1 µs) and v.5.1.4 (extension to 5.1 µs). The protein topology was defined using *gmx grompp* with CHARMM27 parameters (MacKerell *et al.*, 1998, 2004). The His-protonation states from *grompp* were used without modification. 15 440 TIP3P water molecules were added using *gmx solvate*. To neutralize the system, 192 waters were replaced by Cl^- ions using *gmx genion*. CHARMM27 compatible parameters for the pdTp ligand were obtained using the *SwissParam* server (http://www.swissparam.ch/; Zoete *et al.*, 2011).

Simulations were performed using a constant NVT ensemble. NVT simulations are desired for crystalline simulations to enable comparisons of any calculated densities and structure factors to the crystal structure while avoiding difficulties introduced by drift of the unit-cell parameters during the course of the simulation. The model after *gmx genion* showed large negative pressures when simulated *via* NVT. The standard approach for solvated systems of initially equilibrating the pressure using NPT simulations cannot be used as this would change the box size and therefore the unit cell. The present approach is to iteratively perform energy minimization, NVT simulation adding harmonic restraints, and solvation until a pressure near 100 kPa is obtained. After several iterations, the number of water molecules was increased by 1890 to 17 138. The mean pressure computed from the first 110 ns of the trajectory was 1.8 MPa with a standard deviation of 13 MPa, indicating that the procedure was successful.

The final system consisted of a total of 129 462 atoms. There were 32 copies of the protein, 32 copies of the pdTp ligand, 32 Ca^{2+} ions, 17 138 water molecules and 192 Cl^- counterions.

For the production simulations, a time step of 2 fs was used with LINCS holonomic constraints on all bonds and no harmonic restraints. Neighbor searching was performed every ten steps. The particle mesh Ewald (PME) algorithm was used for electrostatic interactions, with a cutoff of 1.4 nm. A reciprocal grid of 64 × 64 was used with fourth-order B-spline interpolation. A single cutoff of 1.4 nm was used for van der Waals interactions. Temperature coupling was performed with the *v-rescale* algorithm. The protein–ligand complex was treated as a separate temperature group from the rest of the atoms. Periodic boundary conditions were used. Trajectory snapshots were obtained every 2 ps in *GROMACS* `.xtc` format. The first 110 ns of the trajectory was dedicated to initial setup and equilibration. An initial production run then extended the duration to 1.1 µs. Subsequent extensions were performed to 5.1 µs in 1 µs increments (as mentioned above, the final microsecond used a later version of *GROMACS*). The size of the equilibration trajectory was 27 GB and the rest were 245 GB each. Each microsecond of simulation took about 2–4 weeks to complete on LANL Institutional Computing machines, depending on the availability of cycles.

2.2. Simulated diffuse intensity

The diffuse intensity was calculated for 100 ns sections of the MD trajectory. Each section was divided into 200 chunks, which were processed in parallel across ten nodes of an Intel Xeon E5-2660_v3 cluster. Prior to performing the calculation, each snapshot of the trajectory was aligned with the crystal structure using the *GROMACS* `.tpr` structure file. To perform this, the `.tpr` file was converted to a multimodel `.pdb` file using *gmx editconf*. The `.pdb` file was processed to ensure that the coordinates reflected the connectivity of the molecules (`gmx trjconv -pbc mol`). The alignment was performed using the processed `.pdb` file as the reference structure (`gmx trjconv -fit translation -pbc nojump`).

Each chunk of sampled structures was processed using the previously described Python script *get_diffuse_from_md.py* (Wall *et al.*, 2014) to calculate the diffuse intensity to 1.6 Å resolution. The calculation of the diffuse intensity $D_{\rm md}(hkl)$ uses Guinier's equation (Guinier, 1963),

$$D_{\rm md}(hkl) = \langle |f_n(hkl)|^2 \rangle_n - |\langle f_n(hkl) \rangle_n|^2. \quad (1)$$

In the script, the structure factor, $f_n(hkl)$, for each sample n is calculated at Miller indices hkl using the *iotbx* package in the *Computational Crystallography Toolbox* (*cctbx*; Grosse-Kunstleve *et al.*, 2002). The script was modified to accept input of an externally supplied unit-cell specification using the PDB CRYST1 format. Specifying the $P1$ unit cell from the crystal structure yields the diffuse intensity $D_{{\rm md},1\times}(hkl)$ sampled on the Bragg lattice only at integer hkl values. Specifying the $P1$ supercell dimensions in the CRYST1 record yields $D_{{\rm md},2\times}(hkl)$, which is sampled twice as finely at hkl values that are multiples of 1/2. Averages for longer sections of the trajectory were accumulated from averages of the smaller chunks.

To decompose the diffuse intensity into isotropic and anisotropic components, reciprocal space was subdivided into concentric spherical shells, each with a thickness equal to the voxel diagonal. The discretely sampled isotropic intensity $D_{\rm md}(\mathbf{s}_n)$ was calculated as the mean intensity at scattering vector $\mathbf{s}_n$ at the midpoint of each shell n. The anisotropic intensity $D'_{\rm md}(hkl)$ was then calculated at each lattice point hkl by subtracting the isotropic intensity $D_{\rm md}(\mathbf{s}_{hkl})$ from the original signal $D_{\rm md}(hkl)$. The value of $D_{\rm md}(\mathbf{s}_{hkl})$ at scattering vector $\mathbf{s}_{hkl}$ in the range $\mathbf{s}_n$, $\mathbf{s}_{n+1}$ was obtained by cubic B-spline interpolation of $D_{\rm md}(\mathbf{s}_n)$ [previous anisotropic intensity calculations made use of linear interpolation (Wall *et al.*, 2014); results using either interpolation method were similar in the present case, although the spline is generally preferred for increased accuracy]. The same method was used to obtain isotropic [$D_{\rm o}(\mathbf{s}_n)$] and anisotropic [$D'_{\rm o}(hkl)$] components of the experimentally observed diffuse intensity.

Because the experimental diffuse intensity shows symmetry consistent with the $P4_1$ symmetry of the unit cell, the $P4/m$ Laue symmetry (corresponding to the $P4_1$ unit-cell symmetry) was enforced by replacing each $D_{\rm md}(hkl)$ value with the average over all symmetry-equivalent hkl positions in the map.

2.3. Simulated structure factors and average structure

Averages of $f_n(hkl)$ were computed along with the diffuse intensity. To obtain structure factors for comparison with the crystal structure, the $P4_1$ unit-cell CRYST1 record was used in lieu of the $P1$ unit-cell or supercell record. The real-space correlation coefficient (RSCC) was evaluated using the *MolProbity* validation tool in *PHENIX* (Adams *et al.*, 2010), using PDB entry 4wor (Wall, Ealick *et al.*, 1997) and the intensities $I_{\rm MD}(hkl)$ computed as the square of the mean $f_n(hkl)$. The errors in the intensities were calculated as the square root of the intensities. Prior to calculating the RSCC, a molecular-replacement search using *MOLREP* in *CCP*4 was used to determine the placement of the protein in the unit cell. The average structure from the simulation was computed by using *phenix.refine* to refine the crystal structure against the $I_{\rm MD}(hkl)$. The *phenix.refine* option `apply_overall_isotropic_scale_to_adp=false` was used to remove the bulk-solvent scaling component from the B factors. For comparison, the B factors from a TLS model were obtained by refinement against the experimental Bragg data using `refine.adp.tls="chain A" strategy=tls` in *phenix.refine*.

2.4. Image processing and diffuse data integration

Experimental diffuse scattering data from Wall, Ealick *et al.* (1997) were used for validation of the simulations. These data were collected on a custom CCD detector configured in an anti-blooming mode in which charge was drained away from overflow pixels (Wall, 1996). The data were processed using the *Lunus* software for diffuse scattering (Wall, 2009; https://github.com/mewall/lunus). Indexing was performed using image numbers 1, 20 and 40 from the rotation series. Rather than limiting the observations $D_{\rm o}(hkl)$ to integer values of the Miller indices hkl, as was performed for the single-unit-cell MD simulations (Wall *et al.*, 2014), data were sampled twice as finely: both at Miller indices and at the midpoints between.

The twofold sampling yields data that correspond precisely to the reciprocal lattice of the 2 × 2 × 2 supercell from the MD model. To place the model and data in an equivalent orientation, the data were reindexed by applying a 180° rotation about the h axis. Prior to integration, images were mode-filtered to reject the Bragg peak signal (Wall, 1996, 2009). The kernel for the mode filter was a 15 × 15 pixel square, with frequency statistics evaluated in 1 ADU bins.

The accuracy of *Lunus* was improved by shifting from integer to floating-point arithmetic for the polarization correction and solid-angle normalization, which were combined into a single step. The increase in accuracy is small for strong diffraction images such as those used here, but is substantial for weaker diffraction images, for example including pixels with fewer than 10–100 photon counts. In addition, as mentioned above (diffuse scattering computation), a cubic B-spline method was implemented in *Lunus* to improve calculation of the anisotropic intensity.

A collection of helper scripts was added to *Lunus* to enable diffraction images to be processed and integrated in parallel. The helper scripts produce shell scripts with *Lunus* workflows to perform image processing, integration and merging of the data and a *cctbx* (Grosse-Kunstleve *et al.*, 2002) workflow to index the data and obtain a transformation from map pixels in diffraction images to fractional *hkl* values in reciprocal space. The scripts were executed in parallel on 12 nodes of a 32-core Intel Haswell cluster; the 96 1024 × 1024 images of diffraction from staphylococcal nuclease could be processed in 1 min. Real-time parallel processing of diffuse scattering from single-crystal synchrotron data sets is therefore now possible using *Lunus*.

All *Lunus* revisions, including helper scripts for parallel processing, have been committed to the github repository https://github.com/mewall/lunus.

As for the calculated $D_{\rm md}(hkl)$, the $P4/m$ Patterson symmetry was enforced for the experimental $D_{\rm o}(hkl)$. The isotropic and anisotropic components $D_{\rm o}({\bf s}_n)$ and $D'_{\rm o}(hkl)$ were calculated from $D_{\rm o}(hkl)$ as for $D_{\rm md}({\bf s}_n)$ and $D'_{\rm md}(hkl)$ (diffuse scattering computation). The correlation coefficient $r_{\rm oc}$ was used to compare the total calculated diffuse scattering $D_{\rm md}(hkl)$ with the experimental data $D_{\rm o}(hkl)$, and the correlation coefficient $r'_{\rm oc}$ was used to compare the anisotropic component $D'_{\rm md}(hkl)$ with the experimental data $D'_{\rm o}(hkl)$.

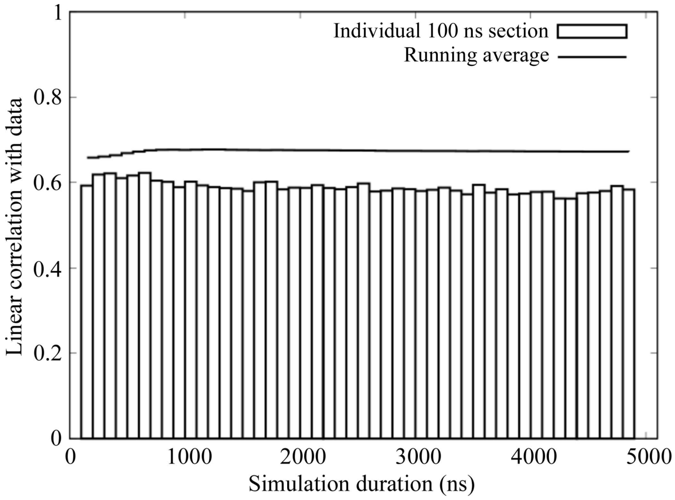


Figure 2
Linear correlation between the simulated and experimental diffuse intensity evaluated for sequential 100 ns sections of the MD trajectory (dashed boxes) and as a running average (solid line). The running average reaches a maximum at a value of 0.68 within the first microsecond.

2.5. Simulated diffraction images

Diffraction images were simulated using methods similar to those used for the data integration, except that instead of the data sets being compiled from the pixel values, the pixel values were obtained from three-dimensional data sets. An indexing solution was obtained as in the data integration, and a template image was provided to determine the crystal orientation. Each pixel was mapped to a fractional Miller index, and the pixel value was calculated as a sum of intensities at the eight nearest grid points in the data set, in proportion to the distance to the center of the grid point along each axis in the space of Miller indices. The method was implemented in a *Lunus* Python script, `simulate_diffraction_image.py`, using *cctbx* methods.

2.6. Patterson maps

Diffuse Patterson maps were created by Fourier transforming diffuse intensities. Symmetrized anisotropic diffuse intensities were output in *hklI* text format using *lat2hkl* in *Lunus*, and were converted to `.mtz` format using *phenix.reflection_file_converter*. Fourier transforms were computed using the Patterson FFT method (Ten Eyck, 1973) in the *CCP*4 suite (Winn *et al.*, 2011). A Patterson map of the anisotropic component of the Bragg reflections was computed for comparison with the diffuse Patterson maps. The anisotropic component of Bragg reflections was calculated using *anisolt* in *Lunus* after conversion of the reflections from `.mtz` format to *hklI* text using *mtz2various* in *CCP*4. The intensities were then converted back to `.mtz` format using *phenix.reflection_file_converter* and the Patterson was obtained using FFT, as mentioned above.

2.7. Rigid-body rotation analysis

Snapshots of C$^{\alpha}$ positions from the supercell were obtained every 40 ps and were translationally aligned with the structure in the `.tpr` file used for the 110–1100 ns simulation. Rotation-matrix analysis was performed using 32 runs of *gmx rotmat*, using the snapshots for each of the 32 copies of the protein as inputs. The *gmx_rotmat.c* source code was edited to add outputs of the root-mean-square deviation (r.m.s.d.) of coordinates between the snapshot and the reference structure both before and after alignment, in addition to the usual output, the elements of the rotation matrix. A custom Python script was written to compute standard deviations of Euler angles from the `.xvg` rotation-matrix element output of *gmx rotmat* and to accumulate trajectory-wide r.m.s.d.s by calculating the square root of the average squared r.m.s.d. for individual snapshots.

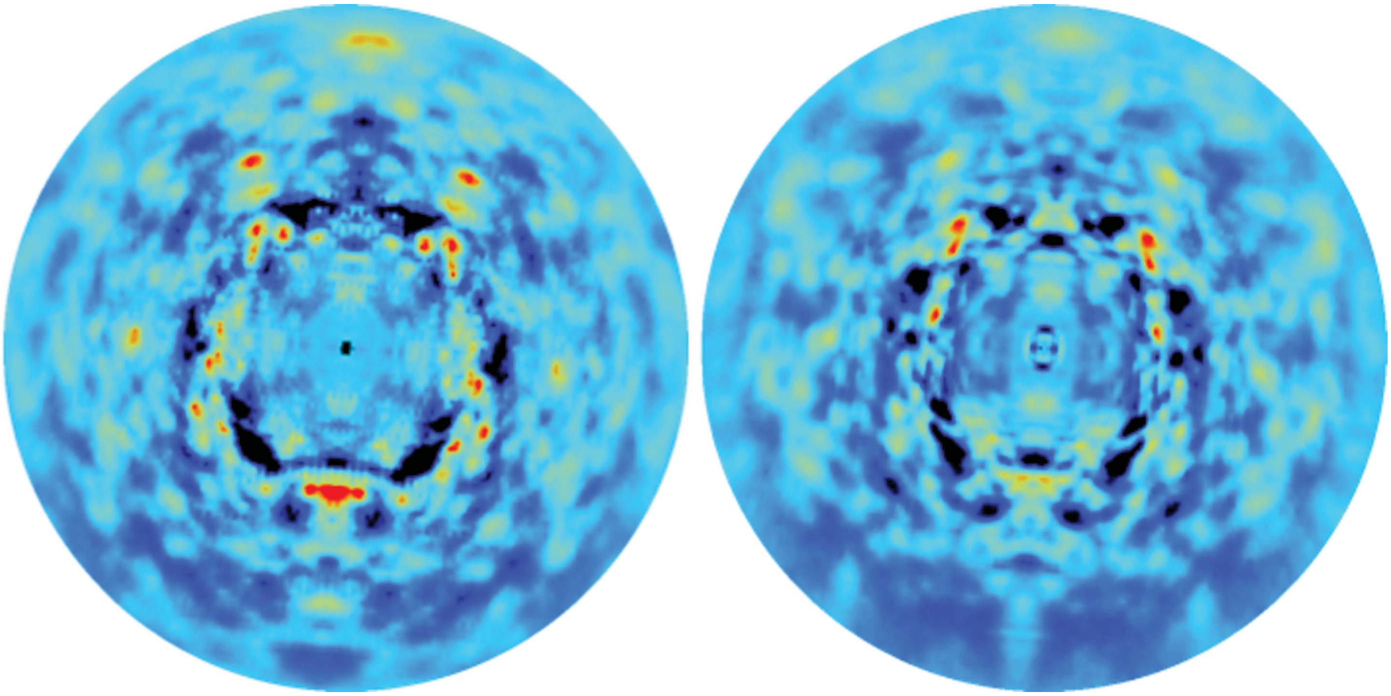


Figure 3
Comparison of simulated diffraction images calculated from the simulated (left panels) and experimental (right panels) three-dimensional diffuse intensities. The crystal orientation corresponds to the first diffraction image in the rotation series. The display is truncated at 1.6 Å. The mean pixel value at each scattering vector was subtracted, and then the minimum value in the image was subtracted, prior to visualization. The images were displayed using the rainbow color map in *Adxv* (Arvai, 2012), with a pixel value range arbitrarily chosen to highlight the similarities.

3. Results

MD simulations were performed using a solvated supercell model of crystalline staphylococcal nuclease consisting of eight unit cells in a 2 × 2 × 2 layout (Fig. 1, §2). The total simulation duration was 5.1 µs. Trajectories were obtained for the following segments, sampled every 2 ps: the initial equilibration (0–110 ns), 110–1100, 1100–2100, 2100–3100, 3100–4100 and 4100–5100 ns.

Diffuse intensities were calculated from the trajectories in 100 ns sections, and the agreement with the data was assessed using the anisotropic component (§2). In the 100 ns immediately following the equilibration, the linear correlation between the simulation and the data is 0.62. Correlations range between 0.59 and 0.62 for subsequent sections through 1100 ns (Fig. 2, boxes). Beyond 1100 ns, the agreement with the data decreases somewhat: the correlation for 100 ns sections between 1100 and 5100 ns ranges from 0.56 to 0.60.

The correlation of the running average of the diffuse intensity (the mean diffuse intensity calculated from sequential 100 ns sections) with the data increases steadily from 0.62 to 0.68 from 110 to 700 ns; it remains at 0.68 through 1100 ns and decreases slowly to 0.67 thereafter (Fig. 2, solid line). The average within the 110–1100 ns range was used as the simulated diffuse intensity for subsequent analyses and visualizations.

Simulated and experimentally derived diffraction images look similar (Fig. 3). There is good correspondence between the shapes of cloudy features in the simulation and the experimental data at all but the lowest resolutions. There are some large differences in the strengths of the features; for example, the large, intense red feature in the bottom half of the simulation image (Fig. 3, left panel) is weaker than the corresponding feature in the experimentally derived image (Fig. 3, right panel).

The correlation between the simulation and the data is substantial over a wide resolution range (Fig. 4, solid steps), consistent with the range over which the diffuse features look similar in Fig. 3. Above 10 Å resolution, the correlation is at

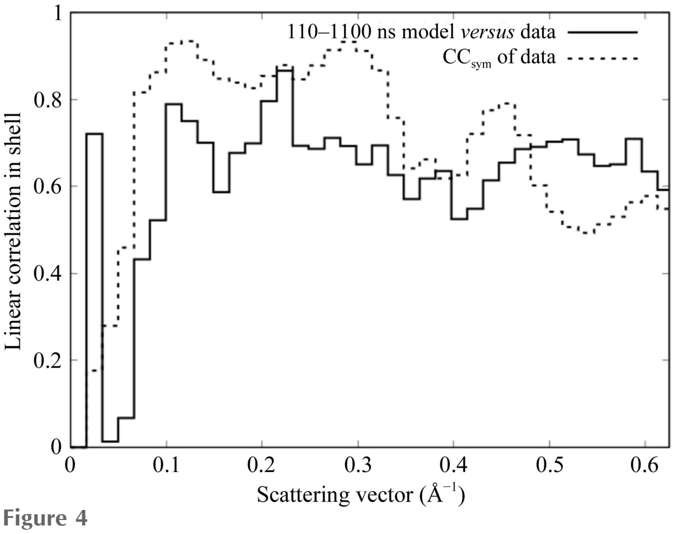


Figure 4
Resolution-dependent agreement between simulation and data (solid line) compared with the self-consistency of the data, as assessed using the expected *P*4/*m* symmetry (dotted line).

least 0.52 in all of the 32 resolution shells, and is above 0.6 in all but three. Below 3 Å resolution the agreement between the simulation and data is substantially less than the agreement between symmetrized and unsymmetrized data sets, indicating that there is room for improvement in modeling the data in this resolution range. Above 2 Å resolution the agreement between the data and simulation is higher than the degree to which the symmetry is obeyed, suggesting that the symmetry averaging might have eliminated some systematic error in the data. Below 10 Å resolution the agreement becomes very small (with the exception of a low-resolution outlier), and the CC_{sym} values also become very small, suggesting that the data were not accurately measured in the neighborhood of the beamstop.

Real-space comparisons of the Patterson function of the charge-density variations (§2) show that both the simulation and the data exhibit similar modulation with distance (Fig. 5). The amplitude of the variations is especially similar in the $x = 0$ section (Figs. 5*a* and 5*b*). In the $z = 0$ section the simulation has higher amplitude variations than the data at longer distances (Figs. 5*d* and 5*e*), indicating that the correlations within this plane are stronger in the simulation than in the data. The linear correlation between the Pattersons computed from the simulation and the data is 0.70. The Patterson computed from the Bragg data (Figs. 5*c* and 5*f*) shows much higher amplitude features at long distances, indicating a longer length scale of correlations for the mean than for the variations in charge density.

The RSCC computed using the crystal structure and the calculated Bragg reflection amplitudes from the simulation (§2) is high in most regions (Fig. 6, purple line); however, there are especially large dips (<0.6) for residues 6–8 at the N-terminus and for residues 46–52. The average RSCC for all residues is 0.80. Regions of the crystal structure with high *B*

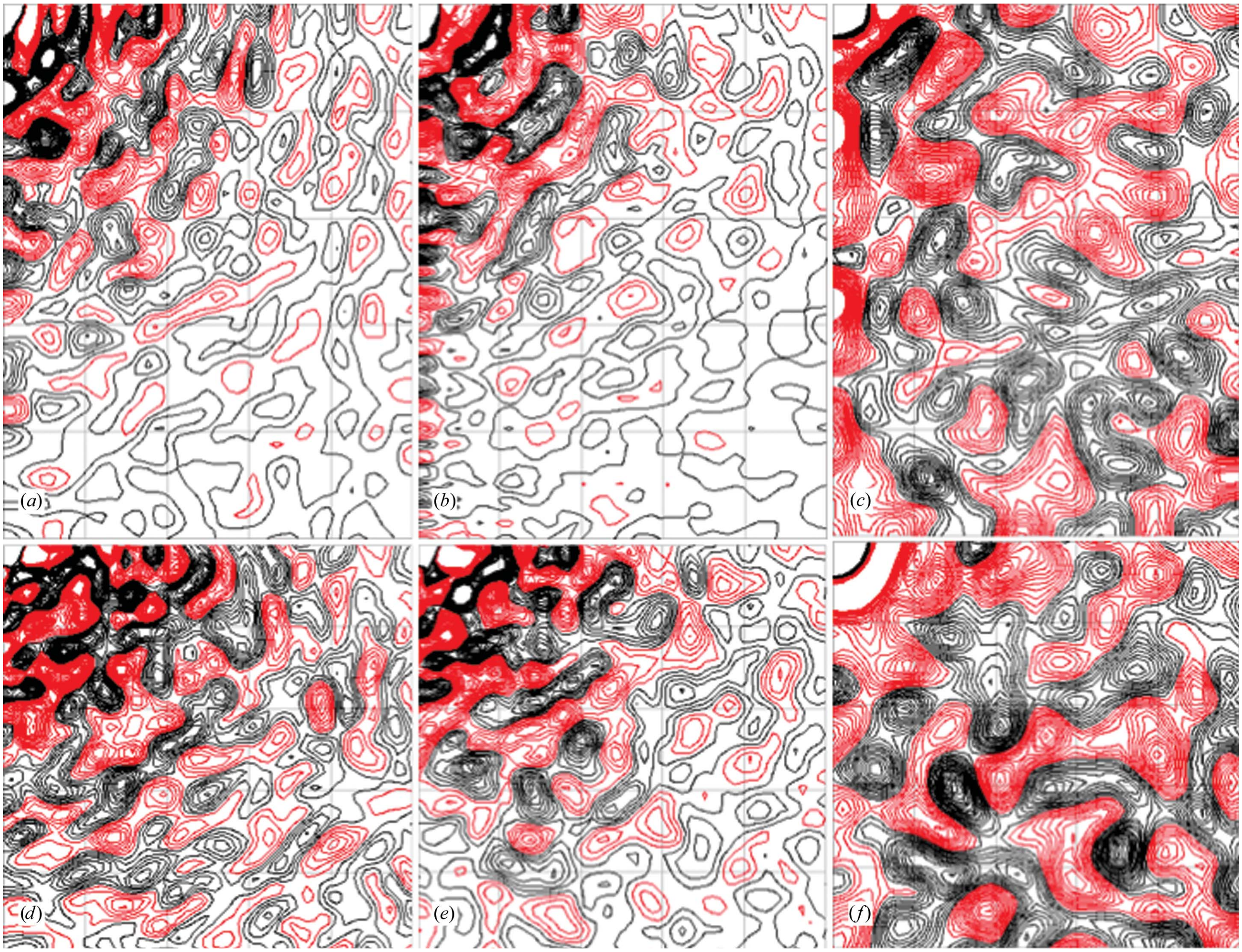


Figure 5
Comparison of the Fourier transform of the simulated diffuse intensity (left panels) with that of the experimental data (middle panels) and the Bragg data (right panels). Only the anisotropic component is used to calculate the transforms. Positive contours are in black and negative contours are in red. Contours are at every 0.5σ between 0 and 10σ in the diffuse data and are adjusted to be equivalent with respect to σ/I_{max} in the other panels. (*a*) $x = 0$ section, simulation; (*b*) $x = 0$ section, diffuse data; (*c*) $x = 0$ section, Bragg data; (*d*) $z = 0$ section, simulation; (*e*) $z = 0$ section, diffuse data; (*f*) $z = 0$ section, Bragg data. The plots were produced using *mapslicer* in *CCP*4 (Winn *et al.*, 2011).

factors (Fig. 6, blue line) include both the N-terminus and a previously observed disordered loop in the crystal structure at residues 44–50 (Loll & Lattman, 1989). The overall r.m.s.d. of atom positions between the simulation average structure (§2) and the crystal structure is 0.7 Å, with high deviations concentrated in local regions of the protein (Fig. 6, yellow line). The residue-wise composite *B* factors from the crystal structure and the simulation average structure are very similar (Fig. 7); the linear correlation between the two is 0.94. By comparison, the residue-wise composite *B* factors from a TLS model of the crystal structure underestimate the disorder in the N-terminus and in the flexible loop (Fig. 7); the linear correlation between the *B* factors derived from the individual ADP *versus* TLS model is 0.89.

4. Discussion

The 0.68 correlation with the anisotropic component of the diffuse data is much higher than has been previously achieved using MD simulations. The present simulation used a supercell model, whereas previous simulations used a unit-cell model of the crystalline protein. In addition, compared with the previous simulation, the present simulation included residues theoretically modeled at the N- and C-termini. To determine the role of including the extra residues in achieving the increased correlation, a 1.1 µs MD simulation of a single unit cell was performed using the extended model (unpublished work). The correlation between the simulated and experimental diffuse intensity within the first microsecond was 0.42, compared with the previous correlation of 0.35–0.43 (Wall *et al.*, 2014). The use of the supercell model therefore accounts for the increased accuracy of the simulated diffuse scattering.

One possible explanation for the increased accuracy of the supercell model is simply the increased size of the ensemble compared with a unit-cell model. The number of intracellular atom pairs is eightfold higher for a 2 × 2 × 2 unit cell than for an equivalent duration of a unit-cell trajectory. However, the agreement with the data for each 100 ns section of the supercell simulation (correlation in the range 0.59–0.62) is much higher than the agreement for the first microsecond of the unit-cell simulation (correlation of 0.42), for which the ensemble is slightly larger. The improvement of the supercell model therefore cannot be attributed to a larger ensemble. Another possible explanation for the improvement is the finer sampling of reciprocal space using the supercell model. The supercell model yields two predictions per Miller index along three directions, or eightfold more predictions of diffuse intensity. These predictions are compared with a diffuse data set that is similarly sampled, including measurements at half-integer Miller indices, where the Bragg peak signal is minimal. In contrast, the unit-cell model yields predictions only at integer Miller indices, where the rejection of the Bragg peak signal in the data is more challenging. To test this explanation, the diffuse intensity was calculated from the supercell model using a *P*1 unit cell (§2), producing a grid sampled at integer Miller indices. The accuracy of this calculation was assessed using the same data set as was used to evaluate the unit-cell simulations. The cumulative correlation over a microsecond was 0.66, which is comparable to the value of 0.68 between the *P*1 supercell calculation and the more finely sampled data set. The improvement therefore cannot be explained by the finer sampling of the diffuse signal.

The most reasonable explanation for the improvement of the supercell model is that it more realistically describes the dynamics, especially the interactions across unit-cell boundaries. This explanation is consistent with normal-mode analysis studies, in which the inclusion of crystal contacts (Kundu *et al.*, 2002; Riccardi *et al.*, 2009) and extension to supercells using Born–von Kármán boundary conditions (Riccardi *et al.*, 2009, 2010) improves the fit to crystallographic data. One rationale for the improvement in the MD is that

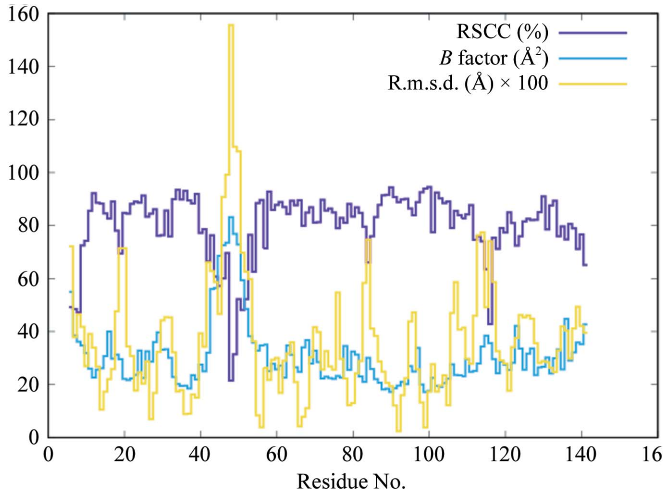


Figure 6
Residue-wise comparisons between the simulated structure factors and the crystal structure. The real-space correlation coefficient (purple line) is computed using the crystal structure and the simulated Bragg reflections. The isotropic *B* factors (blue line) are taken from the crystal structure. The r.m.s.d. of heavy-atom atomic coordinates (yellow line) is computed between the average structure from the simulation and the crystal structure.

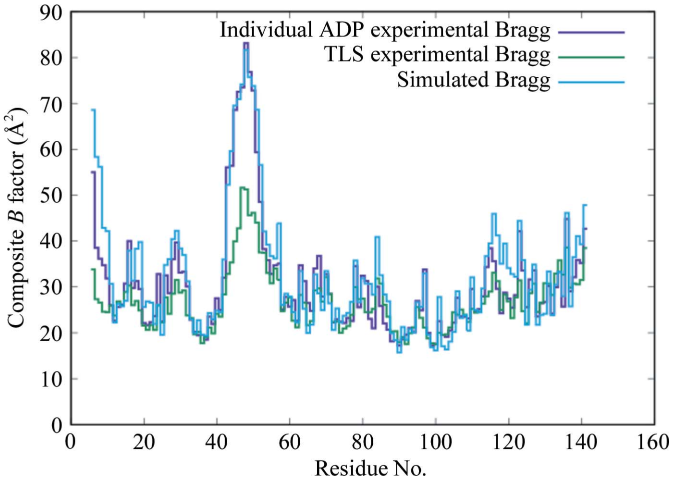


Figure 7
Residue-wise comparison between the *B* factors from the crystal structure (purple), the MD-simulated average structure (cyan) and a TLS model of the crystal structure (green).

using the less accurate unit-cell model displacement correlations between remote atoms can be artificially increased by the periodic boundary conditions, as the atoms can become neighbors across the boundaries of the simulation box. The effect of this artifact on diffuse intensity values sampled at integer Miller indices should be diminished using a 2 × 2 × 2 periodic supercell, as no two atoms in the same unit cell can be neighbors across the simulation box. Because atoms near opposite edges of the simulation box still can be nearby in the supercell model, however, even larger supercell models might further increase the accuracy of the simulations. Larger supercells also can yield detailed models of large-scale motions that have been studied using simpler models of protein diffuse scattering, such as acoustic crystal vibrations in ribosome crystals (Polikanov & Moore, 2015), coupled rigid-body motions in lysozyme crystals (Doucet & Benoit, 1987), and liquid-like motions with long correlation lengths in crystals of lysozyme (Clarage *et al.*, 1992) and calmodulin (Wall, Clarage *et al.*, 1997).

As in the supercell simulation, the residue-wise B factors of the average structure from the single-unit-cell simulation are similar to the crystal structure (Supplementary Fig. S1*a*); the linear correlation between the two is 0.95, compared with a correlation of 0.94 for the supercell simulation. A similar high agreement with crystallographic B factors was seen in MD simulations of a 3 × 2 × 2 supercell of crystalline $P1$ hen egg-white lysozyme (Janowski *et al.*, 2016). The RSCC values computed between the crystal structure and the unit-cell simulation also are similar to those for the supercell simulation (Supplementary Fig. S1*b*); the average RSCC for all residues is 0.82 for the unit-cell simulation, compared with 0.80 for the supercell simulation. Compared with the unit-cell simulation, the supercell simulation therefore specifically improves the model of structure variations, but not the model of the average structure.

The maximum agreement with the data was achieved within the first 1100 ns of the simulation. The time required to reach the maximum is similar to what was seen for the previously published unit-cell simulation (Wall *et al.*, 2014). Because the diffuse intensity is only sensitive to the two-point correlations in the variations, this result suggests that the motions that account for most of the agreement with the experimental data are correlated on a length scale shorter than the unit cell. The short length scale of the correlations is supported by the comparison between the Patterson computed from the Bragg and diffuse data, which reveals that the diffuse Patterson is attenuated at long distances (Fig. 5). Prior to this study, the expectation was that supercell simulations would require a longer duration than unit-cell simulations, as larger systems involve motions on longer length scales, which are generally slower. It is important to note that there is much room for improvement in the accuracy of the MD model, and that to achieve a higher correlation with the data might require longer simulation durations, with stricter requirements for convergence to the thermodynamic ensemble. The relative durations required for diffuse scattering calculations from supercell *versus* unit-cell simulations also might vary for different protein crystals. Nevertheless, this example suggests the possibility that for sufficiently large supercells, the simulation durations required for accurate diffuse scattering calculations might be independent of system size.

There are strong similarities between the Patterson maps calculated from the simulated and experimental diffuse intensities (Fig. 5). The overall modulation of the diffuse Patterson is especially similar between the simulation and the data, indicating that the distance dependence of the correlations is captured well by the MD simulation. The attenuation of the Patterson along the **a** and **b** lattice vectors is more pronounced in the data than in the simulation, however, indicating a longer length scale of correlations in the simulation than in the data within the $z = 0$ plane (Figs. 5*c* and 5*d*).

If the simulation perfectly described the experimental system, the agreement would be expected either to increase or to plateau at long times, depending on how accurately the data were measured. The agreement beyond 1100 ns decreased somewhat, however, indicating a drift of the simulation away from the data. While it is possible that the decrease is transient and that running the present simulation for longer would eventually lead to an increase in the agreement, the simplest explanation is that the MD model is deviating from the experimental behavior at long times.

Dips in the residue-wise RSCC plot (Fig. 6, purple line) indicate regions where the simulated charge density locally deviates from the crystal structure. Dips at the N-terminus and at residues 46–52 correspond to high B-factor regions (Fig. 6, blue line) and might reflect the intrinsic difficulty of capturing discrete conformational variability using B factors (García *et al.*, 1997). Dips in the RSCC also correspond to regions of high r.m.s.d. between the simulation average structure and the crystal structure (Fig. 6, yellow line). These include not only the high B-factor regions, but also many regions with lower B factors. The low B-factor regions with high r.m.s.d.s indicate where the atom positions from the simulation locally deviate from the crystal structure. The discrepancies in these regions would be especially good targets for improving the MD model.

There are a number of specific routes to improving the MD model. Missing residues at the N- and C-termini could be modeled more accurately using more of the context from the crystal structure. The 2 × 2 × 2 supercell could be extended to an even larger supercell. Additional compounds found in the mother liquor could be added to the model (*e.g.* 23% 2-methyl-2,4-pentanediol) and the ionic strength of the solvent could be more accurately modeled; the current model only includes water and neutralizing counterions. There might be inaccuracies in the MD force fields; importantly, the accuracy of the MD models should now be high enough to enable the improvement of force fields using crystallographic data, just as validation using NMR data (Chatfield *et al.*, 1998) has led to improvement of MD force fields (Lindorff-Larsen *et al.*, 2010, 2012; Showalter & Brüschweiler, 2007). Time-averaged ensemble refinement produces models that are closer to the crystal structure (Burnley *et al.*, 2012) and might be used in combination with diffuse data to generate more accurate conformational ensembles. Higher quality data also might be needed to substantially improve the model.

Recent solid-state NMR (ssNMR) experiments combined with crystalline protein simulations (Kurauskas *et al.*, 2017; Ma *et al.*, 2015; Mollica *et al.*, 2012) create opportunities for joint validation of MD simulations using crystallography and NMR. An ssNMR + MD study of the protein GB1 (Mollica *et al.*, 2012) showed reasonable agreement between 200 ns MD simulations and the data for longitudinal relaxation rates and chemical shifts, with lower agreement for the transverse relaxation rates. An ssNMR + MD study of ubiquitin (Kurauskas *et al.*, 2017; Ma *et al.*, 2015) attributed the transverse relaxation rates to rigid-body rotations of whole proteins in the crystal lattice, with amplitudes of 3–5° extracted from the simulations. To assess the importance of rigid-body rotations in the present staphylococcal nuclease simulation, snapshots of each of the 32 copies of the protein were rotationally aligned with a reference structure (§2). Standard deviations of Euler angles were mostly in the 1–2° range, with individual values of as low as 0.8° and as high as 2.3° (Fig. 8*a*). The r.m.s.d. of coordinates between the snapshots and the reference structure decreased after the alignment, but only by 10–20% for most copies of the protein, with a minimum of 8% and a maximum of 28% (Fig. 8*b*). Therefore, rigid-body rotations are not a substantial component of the dynamics in the present simulations.

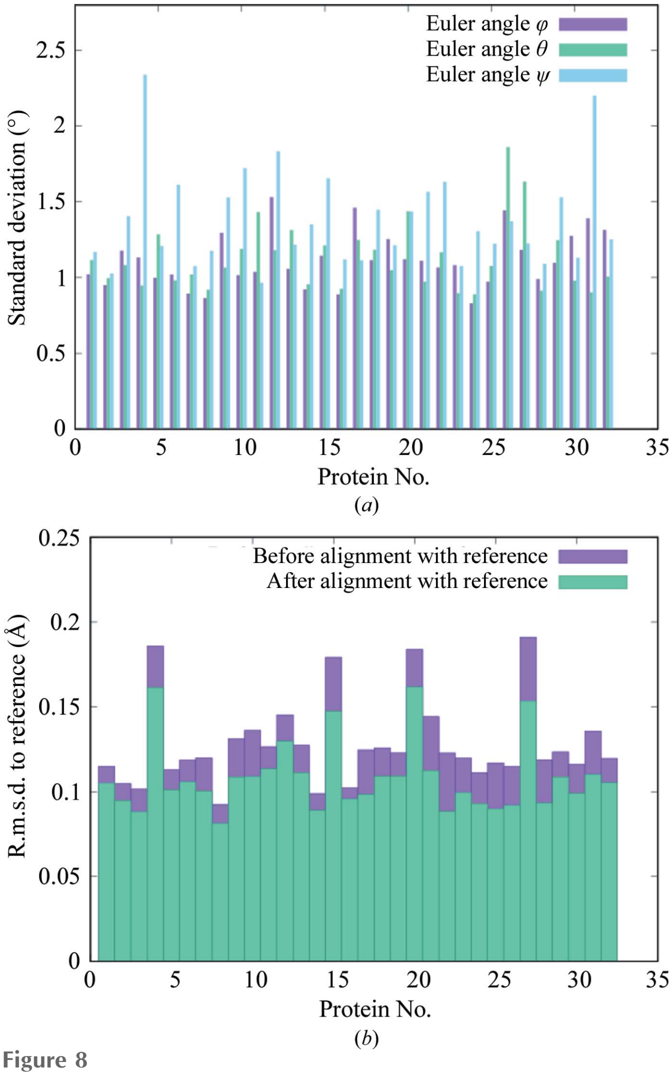


Figure 8
Analysis of the MD trajectory in terms of rigid-body motions of whole proteins. (*a*) Standard deviations of Euler angles that optimally align protein snapshots with the reference structure. (*b*) R.m.s.d.s of coordinates computed before and after aligning protein snapshots with the reference structure.

Further investigation of the rotational matrix fit for protein numbers 4 and 31, which have ψ-angle standard deviations of 2.3 and 2.2°, respectively, revealed a pitfall in rotational analysis. Visual inspection of the trajectories for these protein numbers revealed a conformational change in the flexible loop around residues 42–54 during the first microsecond (Fig. 9). When the tip of the loop was removed (residues 46–52; rendered using sticks in Fig. 9), the ψ-angle standard deviation decreased by 0.7° for these protein numbers. This means that the rotational matrix fit does not solely report on rigid-body motions, as is commonly assumed; therefore, caution is warranted when using rigid-body motions models to interpret MD simulations. It would be interesting to perform a similar analysis of crystalline ubiquitin MD trajectories (Kurauskas *et al.*, 2017; Ma *et al.*, 2015) to see whether the variations in rotational fit correspond to rigid-body rotations, as assumed, or whether they instead might reflect internal motions.

The modeling of the *B* factors is improved using the MD model compared with a TLS model (Fig. 7). This is consistent with the expectation that TLS models might underestimate the disorder in the most mobile regions of the protein (and, in turn, that the *B* factors themselves might underestimate the underlying disorder; García *et al.*, 1997). Overall, the analysis here highlights the importance of internal motions and suggests a more minor role for independent rigid-body translations (Ayyer *et al.*, 2016) or rotations (Pérez *et al.*, 1996)

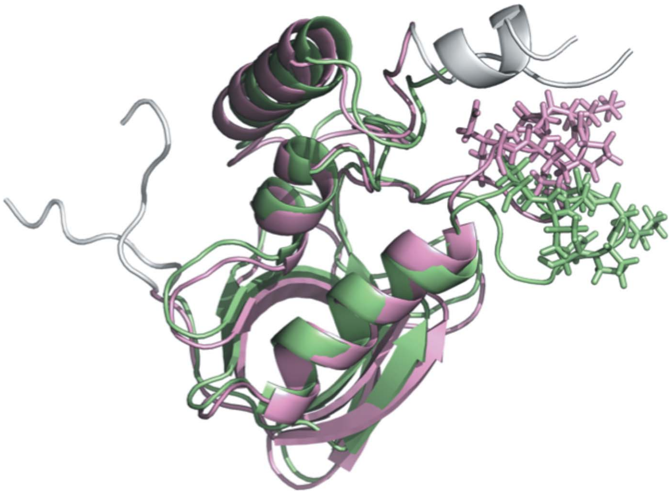


Figure 9
Snapshots of protein 31 at 110 ns (green) and 1001 ns (pink). The seven residues at the N-terminus (left) and C-terminus (upper right), ignored in the rotational fit, are colored white. The SD of the Euler angle ψ of the rotational fit decreases by 0.7° when the tip of the flexible loop (residues 46–52, indicated using sticks) is removed. The image was rendered using *PyMOL* (https://pymol.org/).

in protein diffuse scattering. It will be important to determine whether rigid-body motions are important for other proteins, especially proteins that are stiffer than staphylococcal nuclease. Studies that combine crystallography, ssNMR and MD simulations to develop accurate models of crystalline protein dynamics with multiple points of validation are strongly motivated to reveal the mechanisms of variation that really occur in protein crystals.

In Bragg analysis, a good molecular-replacement solution yields a linear correlation with the Bragg data of about 0.80. The individual atom positions and *B* factors can then be refined to determine a crystal structure that is specific to the crystallographic experiment. Because each Bragg reflection is determined by the entire crystal structure, local atomic details only become resolved once the entire structure is modeled with sufficient accuracy. Similarly, accurate models of diffuse data might only reveal the atomic details of molecular motions when the entire conformational ensemble is modeled with sufficient accuracy.

The present correlation of 0.68, although a significant advance, probably only reflects a global agreement of the model with the data and not a validation of the details of the simulation. As for the Bragg data, once the correlation of models with the anisotropic diffuse data is sufficiently high, diffuse scattering figures of merit such as correlation coefficients or *R* factors might become more sensitive indicators of whether the MD motions are real. If this can be achieved, then crystallography and MD simulations will become a powerful tool for obtaining experimentally validated models of biomolecular mechanisms in crystalline proteins.

Acknowledgements

Many thanks to James S. Fraser for suggesting the rigid-body motion analysis and for leading the 'Macromolecular movements by simulation and diffuse scatter' Laboratory Fees Research Project, which provided additional funding from the University of California. The simulations were performed using Institutional Computing machines at Los Alamos National Laboratory, supported by the US Department of Energy under Contract DE-AC52-06NA25396. Thanks also to Peter B. Moore for reading the manuscript and suggesting the addition of the *B*-factor comparison including the TLS refinement results. This is Los Alamos National Laboratory technical release LA-UR-17-27716.

Funding information

This work was supported by the US Department of Energy *via* the Exascale Computing Project. Additional support was provided by the University of California *via* the Laboratory Fees Research Program.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.

Arvai, A. (2012). *ADXV – A Program to Display X-ray Diffraction Images.* http://www.scripps.edu/tainer/arvai/adxv.html.

Ayyer, K. *et al.* (2016). *Nature (London)*, **530**, 202–206.

Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. (1995). *Comput. Phys. Commun.* **91**, 43–56.

Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. (2012). *Elife*, **1**, e00311.

Chatfield, D. C., Szabo, A. & Brooks, B. R. (1998). *J. Am. Chem. Soc.* **120**, 5301–5311.

Clarage, J. B., Clarage, M. S., Phillips, W. C., Sweet, R. M. & Caspar, D. L. (1992). *Proteins*, **12**, 145–157.

Clarage, J. B., Romo, T., Andrews, B. K., Pettitt, B. M. & Phillips, G. N. Jr (1995). *Proc. Natl Acad. Sci. USA*, **92**, 3288–3292.

Doucet, J. & Benoit, J.-P. (1987). *Nature (London)*, **325**, 643–646.

Faure, P., Micu, A., Pérahia, D., Doucet, J., Smith, J. C. & Benoit, J.-P. (1994). *Nat. Struct. Biol.* **1**, 124–128.

García, A. E., Krumhansl, J. A. & Frauenfelder, H. (1997). *Proteins*, **29**, 153–160.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**, 126–136.

Guinier, A. (1963). *X-ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies.* San Francisco: W. H. Freeman & Co.

Héry, S., Genest, D. & Smith, J. C. (1998). *J. Mol. Biol.* **279**, 303–319.

Janowski, P. A., Cerutti, D. S., Holton, J. & Case, D. A. (2013). *J. Am. Chem. Soc.* **135**, 7938–7948.

Janowski, P. A., Liu, C., Deckman, J. & Case, D. A. (2016). *Protein Sci.* **25**, 87–102.

Kundu, S., Melton, J. S., Sorensen, D. C. & Phillips, G. N. Jr (2002). *Biophys. J.* **83**, 723–732.

Kurauskas, V., Izmailov, S. A., Rogacheva, O. N., Hessel, A., Ayala, I., Woodhouse, J., Shilova, A., Xue, Y., Yuwen, T., Coquelle, N., Colletier, J. P., Skrynnikov, N. R. & Schanda, P. (2017). *Nat. Commun.* **8**, 145.

Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. (2012). *PLoS One*, **7**, e32131.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O. & Shaw, D. E. (2010). *Proteins*, **78**, 1950–1958.

Loll, P. J. & Lattman, E. E. (1989). *Proteins*, **5**, 183–201.

Ma, P., Xue, Y., Coquelle, N., Haller, J. D., Yuwen, T., Ayala, I., Mikhailovskii, O., Willbold, D., Colletier, J.-P., Skrynnikov, N. R. & Schanda, P. (2015). *Nat. Commun.* **6**, 8361.

MacKerell, A. D. Jr *et al.* (1998). *J. Phys. Chem. B*, **102**, 3586–3616.

MacKerell, A. D. Jr, Feig, M. & Brooks, C. L. (2004). *J. Comput. Chem.* **25**, 1400–1415.

Meinhold, L. & Smith, J. C. (2005*a*). *Biophys. J.* **88**, 2554–2563.

Meinhold, L. & Smith, J. C. (2005*b*). *Phys. Rev. Lett.* **95**, 218103.

Meinhold, L. & Smith, J. C. (2007). *Proteins*, **66**, 941–953.

Mollica, L., Baias, M., Lewandowski, J. R., Wylie, B. J., Sperling, L. J., Rienstra, C. M., Emsley, L. & Blackledge, M. (2012). *J. Phys. Chem. Lett.* **3**, 3657–3662.

Pérez, J., Faure, P. & Benoit, J.-P. (1996). *Acta Cryst.* D**52**, 722–729.

Polikanov, Y. S. & Moore, P. B. (2015). *Acta Cryst.* D**71**, 2021–2031.

Riccardi, D., Cui, Q. & Phillips, G. N. Jr (2009). *Biophys. J.* **96**, 464–475.

Riccardi, D., Cui, Q. & Phillips, G. N. Jr (2010). *Biophys. J.* **99**, 2616–2625.

Showalter, S. A. & Brüschweiler, R. (2007). *J. Chem. Theory Comput.* **3**, 961–975.

Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

Wall, M. E. (1996). PhD thesis. Princeton University.

Wall, M. E. (2009). *Methods Mol. Biol.* **544**, 269–279.

Wall, M. E., Clarage, J. B. & Phillips, G. N. (1997). *Structure*, **5**, 1599–1612.

Wall, M. E., Ealick, S. E. & Gruner, S. M. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 6180–6184.

Wall, M. E., Van Benschoten, A. H., Sauter, N. K., Adams, P. D., Fraser, J. S. & Terwilliger, T. C. (2014). *Proc. Natl Acad. Sci. USA*, **111**, 17887–17892.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.

Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. (2011). *J. Comput. Chem.* **32**, 2359–2368.