Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse X-ray scattering

Michael E. Wall[a,1], Andrew H. Van Benschoten[b], Nicholas K. Sauter[c], Paul D. Adams[c,d], James S. Fraser[b], and Thomas C. Terwilliger[e]

[a]Computer, Computational, and Statistical Sciences Division and [e]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545; [b]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158; [c]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; and [d]Department of Bioengineering, University of California, Berkeley, CA 94720

Edited by Peter B. Moore, Yale University, New Haven, CT, and approved November 4, 2014 (received for review September 1, 2014)

X-ray diffraction from protein crystals includes both sharply peaked Bragg reflections and diffuse intensity between the peaks. The information in Bragg scattering is limited to what is available in the mean electron density. The diffuse scattering arises from correlations in the electron density variations and therefore contains information about collective motions in proteins. Previous studies using molecular-dynamics (MD) simulations to model diffuse scattering have been hindered by insufficient sampling of the conformational ensemble. To overcome this issue, we have performed a 1.1-μs MD simulation of crystalline staphylococcal nuclease, providing 100-fold more sampling than previous studies. This simulation enables reproducible calculations of the diffuse intensity and predicts functionally important motions, including transitions among at least eight metastable states with different active-site geometries. The total diffuse intensity calculated using the MD model is highly correlated with the experimental data. In particular, there is excellent agreement for the isotropic component of the diffuse intensity, and substantial but weaker agreement for the anisotropic component. Decomposition of the MD model into protein and solvent components indicates that protein–solvent interactions contribute substantially to the overall diffuse intensity. We conclude that diffuse scattering can be used to validate predictions from MD simulations and can provide information to improve MD models of protein motions.

diffuse scattering | protein crystallography | molecular-dynamics simulation | protein dynamics | staphylococcal nuclease

Proteins explore many conformations while carrying out their functions in biological systems (1–3). X-ray crystallography is the dominant source of information about protein structure; however, crystal structure models usually consist of just a single major conformation and at most a small portion of the model as alternate conformations. Crystal structures therefore are missing many details about the underlying conformational ensemble (4).

Proteins assembled in crystalline arrays, like proteins in solution, exhibit rich conformational diversity (4) and often can perform their native functions (5). Many methods have emerged for using Bragg data to model conformational diversity in protein crystals (6–17). The development of these methods has been important as conformational diversity can lead to inaccuracies in protein structure models (9, 18–20). A key limitation of using the Bragg data, however, is that different models of conformational diversity can yield the same mean electron density.

Whereas the Bragg scattering only contains information about the mean electron density, diffuse scattering (diffraction resulting in intensity between the Bragg peaks) is sensitive to spatial correlations in electron density variations (21–28) and therefore contains information about the way that atomic positions vary together in protein crystals. Because models that yield the same mean electron density can yield different correlations in electron density variations, diffuse scattering provides a means to increase the accuracy of crystallography for determining protein conformational variations (29). Peter Moore (30) and Mark Wilson (31) have argued that diffuse scattering should be used to test models of conformational diversity in X-ray crystallography.

Several pioneering studies used diffuse scattering to reveal insights into correlated motions in proteins (17, 30, 32–49). Some of these studies used diffuse scattering to experimentally validate predictions of correlated motions from molecular-dynamics (MD) simulations (35–37, 40, 42–44). These studies revealed important insights but were limited by inadequate sampling of the conformational ensemble, leading to lack of convergence of the diffuse scattering calculations (35). Microsecond-scale simulations of staphylococcal nuclease were predicted to be adequate for convergence of diffuse scattering calculations (42). Modern simulation algorithms and computer hardware now enable microsecond or longer MD simulations of protein crystals (50).

Here, we present calculations of diffuse X-ray scattering using a 1.1-μs MD simulation of crystalline staphylococcal nuclease. The results demonstrate that we have overcome the past limitation of inadequate sampling. We chose staphylococcal nuclease because the experiments of Wall et al. (49) still represent the only complete, high-quality, 3D diffuse scattering data set from a protein crystal. The calculated diffuse intensity is very similar using two independent halves of the trajectory; the results therefore are reproducible and can be meaningfully compared with the experimental data. The MD simulation provides a rich

Significance

A major challenge of protein crystallography is to accurately describe protein structure variations. Whereas Bragg peaks only yield a picture in which the variations are superimposed, diffuse X-ray scattering (diffraction between the peaks) reports on the way different atoms move together and can be used to increase the accuracy of models. We have performed a molecular-dynamics (MD) simulation of crystalline staphylococcal nuclease, yielding a rich picture of functionally important motions. This simulation is 100-fold longer than previous studies and yields reproducible calculations of diffuse intensity, overcoming a major challenge in the field. Experimental data support the MD models and indicate that diffuse scattering can be used both to validate and to improve MD simulations.

Author contributions: M.E.W. designed research; M.E.W. performed research; M.E.W., N.K.S., and T.C.T. contributed new reagents/analytic tools; M.E.W., A.H.V.B., and T.C.T. analyzed data; and M.E.W., A.H.V.B., N.K.S., P.D.A., J.S.F., and T.C.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: mewall@lanl.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1416744111/-/DCSupplemental.

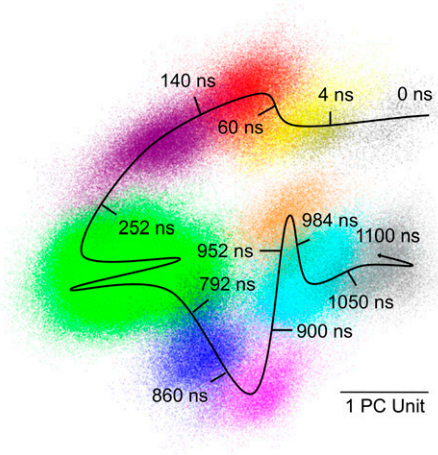

Fig. 1. Scatter plot of structures extracted from the MD trajectory projected on the first two principal components of the α-carbon position covariance matrix. The first component corresponds to the *x* axis, and the second corresponds to the *y* axis. Gray regions correspond to the first 10 ns and last 50 ns of the trajectory; colored regions correspond to metastable states (the yellow region overlaps the first gray region). The curved line indicates the rough trajectory of the system; tick marks indicate approximate times at which state transitions occur.

picture of conformational diversity in the energy landscape of a protein crystal, consisting of at least eight metastable states. Like previous MD studies of crystalline staphylococcal nuclease (42–44), the agreement of the simulation with the total experimental diffuse intensity is excellent, supporting the use of MD simulations to model diffuse scattering data. Unlike previous MD studies, we separately compared the more finely structured, anisotropic component of the diffuse intensity with experimental data. The agreement is substantial but weaker than for the isotropic component, indicating there are inaccuracies in the MD models. Our results therefore point toward using diffuse scattering to improve MD models of protein motions.

Results

Functionally Important Motions of Crystalline Staphylococcal Nuclease. We performed a 1.1-μs MD simulation of a staphylococcal nuclease unit cell (Fig. S1 and *Methods*), revealing a rich conformational landscape. To analyze the trajectory, we performed principal component analysis (*Methods*) and gradually added sequential time points to a scatter plot in the space of the two dominant principal components. The points clustered visually into ellipsoidal shaped regions (Fig. 1).

We identified at least eight metastable states by noting when sharp transitions occurred between the regions (Fig. 1, curved black line with tick marks). The time between the transitions varies considerably from 32 ns (Fig. 1, orange basin) to 540 ns (Fig. 1, green basin); notably, these times are longer than the 10-ns duration of the previous MD simulation of a staphylococcal nuclease unit cell (42) (Fig. 1, gray region). The cyan basin is visited twice: the first time for 52 ns, and the second time for 66 ns. The system spends nearly half the duration of the simulation in the green basin (540 ns); visualization in three dimensions revealed that this basin has a fine structure of substates that, compared with the separations in Fig. 1, lie close together in the space of the dominant principal components.

To gain insight into the functional significance of the states in Fig. 1, we created movies to visualize the motions of the two dominant principal components (Movies S1 and S2). Both components are dominated by internal motions rather than overall rotations and translations of the protein. The residues in the active site (51) cluster into localized regions whose motions are highly correlated (Fig. 2): (A) Glu43, at the beginning of the omega loop (Fig. 2, red sticks); (B) Arg35, Tyr85, and Arg87, at the top of the binding pocket (Fig. 2, Arg35 and Arg87 in orange sticks, and Tyr85 in blue sticks); and (C) Tyr113 and Tyr115, at the bottom of the binding pocket (Fig. 2, cyan sticks). Both of the principal components involve substantial motions of regions A and C with little motion in region B. Especially prominent are a clamping motion of region C against region B, which would likely modulate binding interactions (Fig. 2, blue double-headed arrow), and an opening of region A away from region B, modulating the environment of a water molecule putatively involved in the hydrolysis of the 5′-phosphodiester bond (Fig. 2, red double-headed arrow) (52). The motions vary for the four copies of the protein in the unit cell (Fig. 2, *Inset*): the region C motion is most pronounced for the pink chain; the region A motion is most pronounced for the blue and yellow chains; and the region A and C motions are smallest in the green chain. The states therefore correspond to structures with different active site geometries, resulting in identifiable functional consequences for binding and catalysis.

MD Model of Diffuse Scattering. Wall et al. (49) obtained 3D experimental diffuse scattering data by averaging measurements of the continuous intensity $D_o(\mathbf{s})$ at scattering vectors $\mathbf{s}$ in the neighborhood of each Bragg peak. This procedure yielded a single diffuse intensity $D_o(hkl)$ at integer reciprocal lattice points hkl (*SI Text*); the diffuse lattice showed the expected P4/*m* Patterson symmetry, which was used to create an eightfold redundant expanded lattice with symmetry averaged observations (49). We similarly computed the diffuse intensity $D_{md}(\mathbf{s})$ at each reciprocal lattice point and expanded the lattice using the P4/*m* Patterson symmetry (*Methods*). To assess the reproducibility of diffuse intensity calculations, we compared $D_{md}(hkl)$ calculated using either the first or second half of the following subsections extracted from the full 1.1-μs trajectory: the first 10 ns; the first

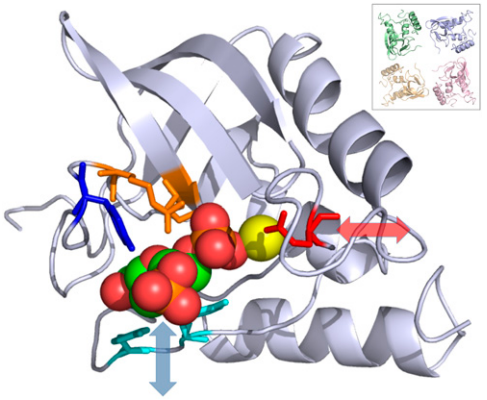

Fig. 2. Active-site conformational dynamics in crystalline staphylococcal nuclease. The backbone is rendered using a ribbon, and the $P4_1$ unit cell packing along the *c* axis is shown in the *Inset* (the screw axis translation is into the page, with the green copy closest and the orange copy farthest away). The residues are shown using sticks, proceeding counterclockwise: Glu43 (red), Arg35 (orange on the β-sheet), Arg87 (orange on the loop), Tyr85 (blue), Tyr115 (cyan), and Tyr113 (cyan). The rest of the protein is rendered as a gray cartoon. The inhibitor thymidine 3′,5′-bisphosphate is shown using spheres and the calcium ion using a yellow sphere. Arrows indicate the direction of motion in the two dominant principal components of the microsecond MD simulation. The loop containing Glu43 (red) moves in the approximate direction indicated by the transparent red double-headed arrow. The loop containing Tyr113 and Tyr115 (cyan) moves in the approximate direction indicated by the transparent blue double-headed arrow. The region containing Tyr85 (blue) and Arg35 and Arg87 (orange) moves much less by comparison. The image was created using PyMOL (66).

100 ns; and the last 1,000 ns. Correlation coefficients using all of the $D_{md}(hkl)$ lattice points (~138,500 in total, of which ~17,300 are independent, due to symmetry) were calculated between the results obtained using either the first or second half (*Methods*). Correlations of the total intensity, r_{12}, were excellent for each of the three subsections of the trajectory: $r_{12} = 0.994$ for 10 ns, 0.995 for 100 ns, and 0.998 for 1,000 ns.

We performed a more detailed assessment of the reproducibility by first decomposing the diffuse intensity into components that are distributed either isotropically or anisotropically about the origin (*Methods*). We then calculated correlations for just the anisotropic component, $D'_{md}(hkl)$, which corresponds to the striking 3D features visible in the experimental data (see, e.g., figure 3 of ref. 49). The correlations for the anisotropic intensity, r'_{12}, were lower than for the total intensity: $r'_{12} = 0.646$ for 10 ns, 0.654 for 100 ns, and 0.832 for 1,000 ns. The first and second halves of the 1,000-ns trajectory sample a similar range in the dominant principal component (Fig. 1), which is consistent with the high correlation for that trajectory. The difference between the values of r_{12} and r'_{12} indicates that the total intensity correlation is dominated by the large isotropic component of the signal that varies gradually with resolution. The increase in the anisotropic intensity correlation with simulation time indicates that diffuse scattering calculations become more reproducible for longer simulations, as expected. The fact that $r'_{12} = 0.832$ for the 1,000-ns trajectory demonstrates that sampling is not a limitation for our validation of MD models of diffuse X-ray scattering.

Meinhold and Smith (42) analyzed diffuse scattering using an approximation in which the variations in atomic positions are normally distributed. In this case, the diffuse intensity in Eq. **1** can be computed using the covariance matrix of atomic displacements $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ between all atom pairs (k, k') (equation 2 in ref. 42). The reproducibility of diffuse scattering calculations then depends on the reproducibility of these matrix element calculations. Similar to Meinhold and Smith (42), we assessed the reproducibility of the matrix elements by just examining α-carbon coordinates. The covariance matrices were calculated from either the early or late halves of the 1,000-ns trajectory (*Methods*), and were compared by calculating the Pearson correlation coefficient between all elements. The resulting value of 0.517 is much lower than the values of r_{12} (0.9975) or r'_{12} (0.832) computed from the diffuse intensity, indicating that calculations of the covariance matrix were not as reproducible as diffuse scattering calculations. A possible explanation is that the diffuse intensity results from the statistically accumulated signal of many atom pairs, whereas the covariance matrix has an element for each individual atom pair.

Experimental Validation of the MD Model. We compared the MD model to the 64,335 experimental observations collected by Wall et al. (49). The observations were placed on a symmetry expanded diffuse intensity lattice containing 120,845 points (*Methods*). For each simulation, we multiplied all of the $D_{md}(hkl)$ by a single scalar weight w and optimized the value of w to minimize the root-mean-squared deviation (RMSD) of the calculated $D_{md}(hkl)$ with respect to the experimental $D_o(hkl)$ values. Comparisons were performed by calculating correlation coefficients using the lattice points where both calculations and observations were available (~118,500, of which ~14,800 are independent due to symmetry). Correlations were computed for four trajectories, either with or without using hydrogen atoms: 0–10, 0–100, 0–1,100, and 100–1,100 ns (Table S1). The correlation r_{oc} of the total intensity ranged from 0.90 to 0.94 for the all-atom calculations, and from 0.85 to 0.92 for the heavy-atom calculations. The correlations for heavy-atom models were in each case lower than for all-atom models (difference of 0.02–0.05), indicating that adding hydrogens somewhat improves the model of total intensity. The correlations increase as the trajectories increase in duration. Excellent agreement was achieved for the 0- to 1,100-ns and 100- to 1,100-ns all-atom simulations, where $r_{oc} = 0.94$.

We also calculated Meinhold and Smith's R-factor–like agreement statistic (equation 3 in ref. 42) between $D_o(hkl)$ and the total $D_{md}(hkl)$ computed from the 100- to 1,100-ns simulation. This involved a minimization with respect to an overall weighting factor and baseline shifting of the $D_{md}(hkl)$. We found a value of 0.029, which is much lower than the minimum value of 0.081 found by Meinhold and Smith (table 2 in ref. 42). The agreement factor therefore is substantially improved for our microsecond simulation compared with their 10-ns simulation.

The isotropically averaged diffuse intensity from the MD model and experimental data were very similar (Fig. 3*A*), which is consistent with the global correlation results. To better understand the origin of the isotropic intensity, we decomposed $D_{md}(hkl)$ for the 1,000-ns trajectory into protein ($D_{md,p}$), solvent ($D_{md,s}$), and protein–solvent ($D_{md,x}$) terms (*Methods*). (Note that the contribution of $D_{md,x}$ to the sum is negative.) Each component contributes about equally to the total signal (Fig. 3*B*).

The protein and solvent curves are similar to those calculated by Meinhold and Smith (42). The cross term, $D_{md,x}$, starts high at small scattering vectors, dips below zero at about 0.3 $Å^{-1}$, and settles to zero at ~0.35 $Å^{-1}$. To gain further insight into the nature of the correlations implied by this behavior, we fit the cross term to a model of correlations that decay exponentially in real space over a length γ (*Methods*). A reasonable fit was obtained using $\gamma = 0.89 \pm 0.08$ Å (Fig. S2), suggesting that the cross term is dominated by short-range correlations between the fluctuations of the protein and solvent (*Discussion*).

As for the model D'_{md}, we calculated the anisotropic component of the experimentally observed diffuse intensity, D'_o, by subtracting the isotropic component (*Methods*). For both the experimental data and the MD model, at each resolution the SD of the anisotropic intensity is less than 10% of the isotropic intensity (Fig. 3*A*, error bars). The anisotropic component therefore is weaker than the isotropic component. The anisotropic intensity due the individual protein, solvent, and protein–solvent cross terms show differences in magnitude (Fig. 3*B*). At scattering vectors greater than 0.1 $Å^{-1}$, the protein has the strongest anisotropic intensity (largest green error bars). The protein–solvent cross term is weaker (small cyan error bars), and the solvent is weakest (very small magenta error bars). Below 0.1 $Å^{-1}$, each component has substantial anisotropic intensity; however, different components nearly cancel, summing to a very small total anisotropic intensity (blue bars).

To assess the agreement of the anisotropic component of the MD model (D'_{md}) with experimental data (D'_o), we visualized

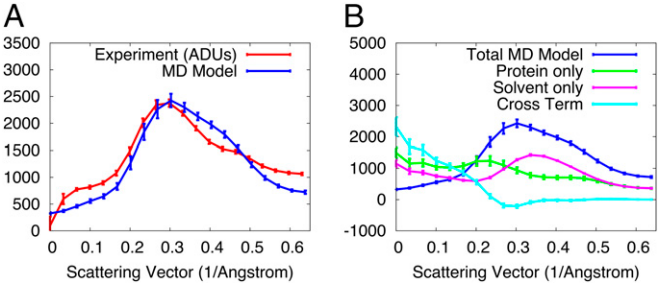

Fig. 3. Analysis of isotropic diffuse intensity. (*A*) Comparison of isotropic diffuse intensity for data (red) to the MD model (blue). (*B*) Decomposition of the total simulated isotropic intensity (blue) into contributions from protein (green), solvent (magenta), and the protein–solvent cross term (cyan). The total intensity is equal to the protein term plus the solvent term minus the cross term (*Methods*). In both *A* and *B*, the SD of the anisotropic diffuse intensity is indicated using error bars.

level surfaces of D'_o (Fig. 4*A*), D'_{md} (Fig. 4*B*), and $D'_o - D'_{md}$ (Fig. 4*C*). The comparison revealed places where the MD model intensity was either higher (Fig. 4*C*, green) or lower (Fig. 4*C*, red) than the experimental intensity. As for the total intensity, to quantify the agreement we calculated the anisotropic correlation coefficient r'_{oc} (*Methods*) between D'_{md} and D'_o for each of the trajectories, either with or without hydrogens (Table S1). The correlation r'_{oc} of the anisotropic intensity ranged from 0.35 to 0.43 for the all-atom calculations, and from 0.34 to 0.40 for the heavy-atom calculations. The highest anisotropic correlation ($r'_{oc} = 0.43$) is strongly positive, but is lower than the lowest total intensity correlation ($r_{oc} = 0.85$), indicating that the MD model more accurately describes the isotropic than the anisotropic diffuse intensity. The best all-atom correlation is for the 0- to 10-ns trajectory ($r'_{oc} = 0.43$), and the worst is for the 0- to 100-ns trajectory ($r'_{oc} = 0.35$). The 0- to 1,000-ns and 0- to 1,100-ns trajectories yield intermediate correlations ($r'_{oc} = 0.40$).

Correlations for heavy-atom models were lower than for all-atom models by 0.01 or less (Table S1), indicating that adding hydrogens did not substantially improve the model of anisotropic intensity. The correlation for the protein component of the MD model ($D_{md,p}$) was $r'_{oc} = 0.38$ (not listed in Table S1), which is very similar to that for the entire MD model. This supports the finding that the protein is the dominant contributor to the anisotropic intensity (Fig. 3*A*, error bars).

We also calculated the anisotropic correlation within resolution shells, and found the highest value was 0.58 in the 0.27 Å^{-1} resolution shell. Visualization of the experimental and MD model intensity in this shell shows arcs of diffuse intensity extending between the two poles, and rich large-scale features near the equator (Fig. 4*D*). A liquid-like motions model showed similar patterns (figure 5 in ref. 49), suggesting that they can be explained largely by the protein structure factor (equation 4 in ref. 49).

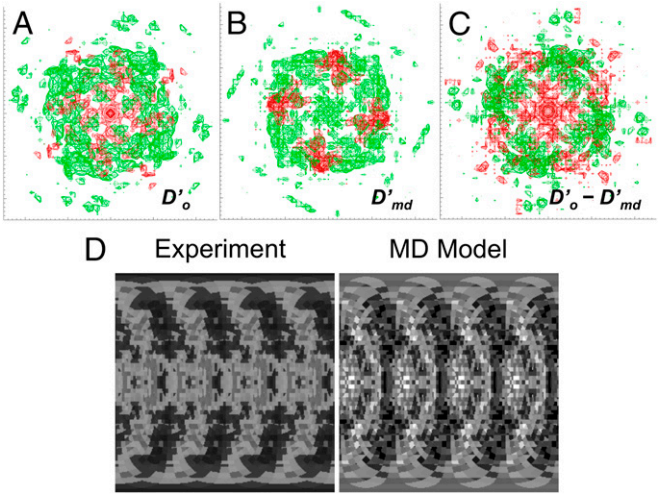

Fig. 4. Comparison of anisotropic diffuse intensity between models and experimental data. (*A*–*C*) Isosurfaces in the (*A*) experimental (D'_o), (*B*) scaled MD model (D'_{md}), and (*C*) difference ($D'_o - D'_{md}$) intensity maps. Positive intensity is shown in green, and negative intensity in red. All isosurfaces including the difference map are displayed at an intensity level equal to the SD of D'_o in the solvent ring. The values of D'_{md} were multiplied by a uniform scale factor to yield the same SD in the solvent ring as D'_o. The *x* direction in each panel corresponds to the *a** axis, varying from −0.5 Å^{-1} at the left to 0.5 Å^{-1} at the right; the *y* direction corresponds to the *b** axis, varying from −0.5 Å^{-1} at the bottom to 0.5 Å^{-1} at the top. (*D*) Visualizations of the experimental (*Left*) and MD model (*Right*) anisotropic diffuse intensity in the 0.27-Å^{-1} resolution shell, for which the agreement is best. The images were constructed as in figure 3 of ref. 49, using the shimlt and seesh routines in LUNUS (63, 72). The *y* direction corresponds to the polar angle in the shell as measured from the *c** axis, varying from 0 at the top to π at the bottom of each image. The *x* direction corresponds to the azimuthal angle as measured from the *a** axis in a right-handed sense, varying from −π at the left to π at the right. Pixel values are displayed as the deviation from the mean on a linear gray scale, with −500 corresponding to black, and 500 corresponding to white.

Discussion

The present 1.1-μs simulation, which is 100-fold longer than the previous 10-ns simulation of a staphylococcal nuclease unit cell (42), revealed eight metastable states with lifetimes that are longer than the duration of the shorter simulation. Analysis of the motions revealed that the basins correspond to structures with different active-site geometries, resulting in identifiable functional consequences for binding and catalysis.

Diffuse scattering can independently validate the predictions of MD simulations. Conversely, interpretation of the diffuse intensity in terms of a multistate conformational ensemble is now possible (Fig. 1). Although there is room for improvement in modeling the isotropic intensity at resolutions exceeding 0.3 Å^{-1} (Fig. 3*A*), the agreement of the MD model with the total experimental diffuse intensity is remarkable considering that no free parameters were adjusted. Moreover, this agreement is robust to sampling (Table S1) and to a model perturbation that includes a change of force field (*SI Text*).

Compared with the total intensity, the MD model had a lower, but still strongly positive, correlation with the anisotropic diffuse intensity. The lower correlation of the anisotropic intensity indicates there are inaccuracies in the MD models. Maximizing the correlation of the anisotropic intensity is therefore a potential strategy for increasing the accuracy of the MD models. For example, both the simulation of a single unit cell with periodic boundary conditions and the assumption of independent unit cells inherent in Eq. **1** might limit the accuracy of the MD model. One way to test this possibility is to simulate larger sections of the crystal with several unit cells (50), as this will enable correlations between unit cells to be included in the model.

Decomposition of the diffuse intensity into protein and solvent components revealed that the protein component $D_{md,p}$ dominates the overall magnitude of the anisotropic intensity (Fig. 3*A*, error bars). In addition, the MD model of anisotropic intensity using the protein alone ($r'_{oc} = 0.38$) was almost as accurate as using the protein and solvent ($r'_{oc} = 0.40$). These results support the hypothesis that the anisotropic intensity reports largely on variations in the protein structure (49).

The decomposition also revealed that the protein–solvent interactions contribute a strong negative component to the sum ($D_{md,x}$ is mostly positive in Fig. 3*B* and enters with a minus sign). Combined with the subatomic correlation length derived from the fit to Eq. **2**, these results suggest that the cross term might be mainly due to negative correlations in the density fluctuations that occur in the region between the protein and solvent atoms. One possible explanation for this is that, when atoms move together, the density shifts, resulting in a positive fluctuation ($+\Delta\rho$) for the leading edge density of one atom (e.g., a protein atom) coupled a negative fluctuation ($-\Delta\rho$) for the trailing edge density of a neighboring atom (e.g., a solvent atom) (Fig. S2, *Inset*).

Increasing the duration of the MD trajectory yielded more reproducible calculations of diffuse intensity and better models of the total intensity. Calculations of the covariance matrix of α-carbon displacements, however, were less reproducible than the calculations of diffuse intensity. In addition, only the cyan basin in Fig. 1 is visited more than once. Thus, even though we have achieved reproducible diffuse scattering calculations using an MD model, the sampling of the conformational ensemble is still incomplete. Millisecond all-atom simulations (53), advanced sampling methods such as parallel tempering (54), and acceleration schemes such as Markov state models (55) and parallel

replica MD (56, 57) can be used to pursue increased sampling and improved convergence of the calculations.

Our comparisons point to opportunities for improving models of protein motions using diffuse scattering. For example, if the accuracy of the model of anisotropic diffuse intensity can be improved, it might become possible to use diffuse scattering to improve MD force fields. Indeed, use of NMR data for force field improvement provides an example of the potential impact of diffuse scattering. Early comparisons of MD to NMR data for staphylococcal nuclease revealed differences in the residue-by-residue backbone flexibility (58), and more recent force fields now do a much better job of reproducing the NMR data (59–62). Another possibility is to use diffuse scattering for validating results obtained using schemes for increasing sampling (54–57).

There is also room for improvement in integration of experimental diffuse scattering data from diffraction images. For example, our current methods aim to extract the large-scale features (due to short-range correlations within the unit cell) and to ignore the small-scale, streaked features (due to long-range correlations across unit cells) (63). However, these methods can be improved by accounting for better separation of the large-scale from small-scale features, or a finer sampling of all diffuse features (29). Increased diffuse scattering data collection efforts and advances in X-ray detector technology will further this goal (29).

Methods

MD Simulations. The 48.5 Å × 48.5 Å × 63.5-Å P1 unit cell model was prepared from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (64) (www.rcsb.org) entry 1STN (65) using the UnitCell code in AmberTools (ambermd.org/#AmberTools). The RMSD of the atomic coordinates [calculated using PyMOL (66)] between 1STN and PDB entry 4WOR (49) (the crystal structure refined against the Bragg data) is 0.70 Å, with backbone deviations concentrated in loops near the ligand binding site (Fig. S3). Waters and counterions were added using genbox and genion in GROMACS (67), resulting in a model with 8,904 protein atoms (four proteins arranged in a $P4_1$ configuration), 6,477 TIP3P water atoms, and 40 Cl^- ions (Fig. S1). The OPLS-AA force field (68, 69) was selected. An initial model was prepared using energy minimization followed by 100-ps equilibration in a constant-particle number, -volume, and -temperature (NVT) ensemble at 300 K, using harmonic restraints for the protein atom coordinates. The modified Berendsen thermostat with a 0.1-ps time constant was used for temperature equilibration.

The production 1.1-μs simulation was performed in GROMACS (67) on a 128 core 2.00 GHz Intel Xeon X6550 machine using a 2-fs time step with a constant-particle number, -pressure, and -temperature (NPT) ensemble at 300 K and 1 bar. The Parrinello–Rahman barostat with a 2.0-ps time constant was used for pressure equilibration. At a rate of 44.6-ns system time per day of wall clock time, it took roughly 30 d to complete, including periods of down time. Snapshots were recorded every 2 ps in compressed (.xtc) format, yielding a total of 5.5×10^8 frames in a 31.1-GB file.

Upon the switch from the restrained NVT to the unrestrained NPT simulation, the unit cell dimensions began fluctuating (all scaled simultaneously with the volume fluctuations) and the mean values rapidly shrank by 1.3% of their initial values. The structure also drifted away from the crystal structure: the RMSD between the mean coordinates from the simulation and PDB entry 4WOR was 1.8 Å (1.3 Å) for heavy atoms (α-carbons). These values are similar to those obtained by Meinhold and Smith (42) for their 10-ns simulation: they reported deviations of less than 0.8% in unit cell dimensions, and a RMSD between the mean simulation coordinates and PDB entry 2SNS (52) (an experimentally determined crystal structure available at the time) of 1.7 Å (1.3 Å) for heavy atoms (α-carbons).

MD Model Diffuse Scattering Calculations and Reproducibility. To calculate the diffuse intensity for a given section of the trajectory, ensembles of atomic coordinates were extracted from the full trajectory in 5-ns chunks. This procedure yielded protein and solvent coordinates of the P1 unit cell in a series of separate files (*SI Text*).

Each of the files was then processed to calculate the diffuse intensity. The structure factor, $f_n(hkl)$, for each unit cell structure n was calculated to 1.6-Å resolution at Miller indices hkl using the iotbx package in the computational crystallography toolbox (CCTBX) (70). The average structure factor, $\langle f_n(hkl)\rangle_n$, and the average squared structure factor, $\langle |f_n(hkl)|^2\rangle_n$, were calculated using a modified version of the *Phenix* (71) Python script get_struct_fact_from_md.py, which we called get_diffuse_from_md.py. Averages for longer sections of the trajectory were accumulated from averages of the 5-ns chunks. The intensity of the diffuse scattering for the first $[D_{md1}(hkl)]$ and second $[D_{md2}(hkl)]$ parts of the trajectory were calculated using the following equation:

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

Eq. **1** appears in the classic text of Guinier (24) and assumes independent unit cell fluctuations. It states that the diffuse intensity is equal to the variance of the unit cell structure factor, and therefore contains information that is distinct from the Bragg peak intensities, which are determined by the square of the mean structure factor (second term of Eq. **1**). Eq. **1** is an established method for calculating diffuse scattering from protein MD simulations (35, 42–44).

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 reciprocal unit cell diagonal ($\Delta s = 0.0336$ Å^{-1}). The isotropic intensity $D_{md}(s_n)$ was calculated as the mean intensity at scattering vector s_n at the midpoint of each shell n. The anisotropic intensity $D'_{md}(hkl)$ was then calculated at each lattice point hkl by subtracting the isotropic intensity $D_{md}(s_{hkl})$ from the original signal $D_{md}(hkl)$. The value of $D_{md}(s_{hkl})$ at scattering vector s_{hkl} in the range (s_n, s_{n+1}) was obtained by linear interpolation of $D_{md}(s_n)$. The same method was used to obtain isotropic $[D_o(s_n)]$ and anisotropic $[D'_o(hkl)]$ components of the experimentally observed diffuse intensity.

Because the experimental diffuse intensity shows symmetry consistent with the $P4_1$ symmetry of the unit cell (49), we enforced the P4/*m* Patterson symmetry (corresponding to the $P4_1$ unit cell symmetry) by replacing each $D_{md}(hkl)$ value with the average over all symmetry equivalent hkl positions in the map. The resulting symmetry-expanded map had ∼17,300 independent values on ∼138,500 lattice points with eightfold redundancy. The values were very similar to the original map (Pearson correlation coefficient $r_{sym} = 0.996$ for total intensity and $r'_{sym} = 0.789$ for anisotropic intensity).

To assess the reproducibility of diffuse scattering calculations in the MD model, we divided the extracted trajectory into first and second halves of equal duration. We then calculated the simulated diffuse scattering from the first and second parts independently, and compared the results quantitatively. A global comparison of $D_{md}(hkl)$ and $D'_{md}(hkl)$ was made using the Pearson correlation coefficients r_{12} and r'_{12}, respectively.

Covariance Matrices and Principal Components. To calculate covariance matrices and to perform principal component analysis, the trajectory was adjusted to remove discontinuous jumps of atomic positions to symmetry-related positions during the course of the simulation (trjconv –pbc nojump). Next, the coordinates were adjusted to remove translational drift (trjconv –fit translation) and to preserve the covalent bonding structure of the protein (trjconv –pbc mol). Finally the α-carbon coordinates were selected and used to calculate and diagonalize the covariance matrix (g_covar). Comparisons of covariance matrices and projections of trajectories onto principal components were performed using the tool g_anaeig. Visualization of the energy landscape was performed by displaying the projected coordinates in a 2D scatter plot.

Decomposition into Protein and Solvent Components. We used Eq. **1** and the equation $D_{md} = D_{md,p} + D_{md,s} - D_{md,x}$ to decompose D_{md} into protein $(D_{md,p} = \langle |f_{prot}|^2 \rangle - |\langle f_{prot} \rangle|^2)$, solvent $(D_{md,s} = \langle |f_{solv}|^2 \rangle - |\langle f_{solv} \rangle|^2)$ and protein–solvent $(D_{md,x} = D_{md,p} + D_{md,s} - D_{md})$ terms, where f_{prot} and f_{solv} are the structure factors computed from just the protein or solvent component, respectively. Note that positive values of $D_{md,x}$ contribute negatively to D_{md}.

We modeled $D_{md,x}$ using the following function (34, 49):

$$\mathrm{FT}[e^{-\gamma r}] = \frac{8\pi\gamma^3}{\left[1 + (2\pi s\gamma)^2\right]^2}, \quad [2]$$

where s is the scattering vector, γ is the correlation length of atomic displacements, and $\mathrm{FT}[e^{-\gamma r}]$ indicates the 3D Fourier transform of the real-space function $e^{-\gamma r}$, which decays exponentially with distance r. The fitting of $D_{md,x}$ to Eq. **2** was performed using the nonlinear-least-squares fitting feature of gnuplot (www.gnuplot.info).

ACKNOWLEDGMENTS. This work was supported by the US Department of Energy under Contract DE-AC52-06NA25396 through the Laboratory-Directed Research and Development Program at Los Alamos National

Laboratory. N.K.S. is supported by NIH Grant GM095887. P.D.A. and T.C.T. are supported by NIH Grant 1P01GM063210. J.S.F. is a Searle Scholar and a Pew Scholar, and is supported by NIH Grants OD009180 and GM110580, and National Science Foundation Grant STC-1231306.

1. Austin RH, et al. (1973) Dynamics of carbon monoxide binding by heme proteins. *Science* 181(4099):541–543.
2. Weber G (1972) Ligand binding and internal equilibria in proteins. *Biochemistry* 11(5): 864–878.
3. Frauenfelder H, Parak F, Young RD (1988) Conformational substates in proteins. *Annu Rev Biophys Biophys Chem* 17:451–479.
4. Frauenfelder H, Petsko GA, Tsernoglou D (1979) Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature* 280(5723):558–563.
5. Doscher MS, Richards FM (1963) The activity of an enzyme in the crystalline state: Ribonuclease S. *J Biol Chem* 238(7):2393–2398.
6. Burnley BT, Afonine PV, Adams PD, Gros P (2012) Modelling dynamics in protein crystal structures by ensemble refinement. *Elife* 1:e00311.
7. Chaudhry C, Horwich AL, Brunger AT, Adams PD (2004) Exploring the structural dynamics of the *E. coli* chaperonin GroEL using translation-libration-screw crystallographic refinement of intermediate states. *J Mol Biol* 342(1):229–245.
8. Chen Z, Chapman MS (2001) Conformational disorder of proteins assessed by real-space molecular dynamics refinement. *Biophys J* 80(3):1466–1472.
9. DePristo MA, de Bakker PI, Blundell TL (2004) Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure* 12(5):831–838.
10. Fraser JS, et al. (2009) Hidden alternative structures of proline isomerase essential for catalysis. *Nature* 462(7273):669–673.
11. Fraser JS, et al. (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography. *Proc Natl Acad Sci USA* 108(39):16247–16252.
12. Kuriyan J, et al. (1991) Exploration of disorder in protein structures by X-ray restrained molecular dynamics. *Proteins* 10(4):340–358.
13. Kuriyan J, Weis WI (1991) Rigid protein motion as a model for crystallographic temperature factors. *Proc Natl Acad Sci USA* 88(7):2773–2777.
14. Pellegrini M, Grønbech-Jensen N, Kelly JA, Pfluegl GM, Yeates TO (1997) Highly constrained multiple-copy refinement of protein crystal structures. *Proteins* 29(4): 426–432.
15. Sternberg MJ, Grace DE, Phillips DC (1979) Dynamic information from protein crystallography. An analysis of temperature factors from refinement of the hen egg-white lysozyme structure. *J Mol Biol* 130(3):231–252.
16. van den Bedem H, Bhabha G, Yang K, Wright PE, Fraser JS (2013) Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat Methods* 10(9):896–902.
17. Wall ME, Clarage JB, Phillips GN (1997) Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering. *Structure* 5(12):1599–1612.
18. García AE, Krumhansl JA, Frauenfelder H (1997) Variations on a theme by Debye and Waller: From simple crystals to proteins. *Proteins* 29(2):153–160.
19. Furnham N, Blundell TL, DePristo MA, Terwilliger TC (2006) Is one solution good enough? *Nat Struct Mol Biol* 13(3):184–185, discussion 185.
20. Kuzmanic A, Pannu NS, Zagrovic B (2014) X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat Commun* 5:3220.
21. Zachariasen W (1945) *Theory of X-Ray Diffraction in Crystals* (Wiley, New York).
22. James R (1948) *The Optical Principles of the Diffraction of X-Rays* (Bell, London).
23. Wooster WA (1962) *Diffuse X-Ray Reflections from Crystals* (Oxford Univ Press, Oxford).
24. Guinier A (1963) *X-ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies* (W. H. Freeman and Company, San Francisco).
25. Amorós JL, Amorós M (1968) *Molecular Crystals; Their Transforms and Diffuse Scattering* (Wiley, New York).
26. Warren BE (1969) *X-Ray Diffraction* (Addison-Wesley, Reading, MA).
27. Willis BTM, Pryor AW (1975) *Thermal Vibrations in Crystallography* (Cambridge Univ Press, Cambridge, UK).
28. Welberry TR (2004) *Diffuse X-Ray Scattering and Models of Disorder* (Oxford Univ Press, Oxford).
29. Wall ME, Adams PD, Fraser JS, Sauter NK (2014) Diffuse X-ray scattering to model protein motions. *Structure* 22(2):182–184.
30. Moore PB (2009) On the relationship between diffraction patterns and motions in macromolecular crystals. *Structure* 17(10):1307–1315.
31. Wilson MA (2013) Visualizing networks of mobility in proteins. *Nat Methods* 10(9): 835–837.
32. Caspar DL, Clarage J, Salunke DM, Clarage M (1988) Liquid-like movements in crystalline insulin. *Nature* 332(6165):659–662.
33. Chacko S, Phillips GN, Jr (1992) Diffuse x-ray scattering from tropomyosin crystals. *Biophys J* 61(5):1256–1266.
34. Clarage JB, Clarage MS, Phillips WC, Sweet RM, Caspar DL (1992) Correlations of atomic movements in lysozyme crystals. *Proteins* 12(2):145–157.
35. Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN, Jr (1995) A sampling problem in molecular dynamics simulations of macromolecules. *Proc Natl Acad Sci USA* 92(8):3288–3292.
36. Doucet J, Benoit JP (1987) Molecular dynamics studied by analysis of the X-ray diffuse scattering from lysozyme crystals. *Nature* 325(6105):643–646.
37. Faure P, et al. (1994) Correlated intramolecular motions and diffuse X-ray scattering in lysozyme. *Nat Struct Biol* 1(2):124–128.
38. Glover ID, Harris GW, Helliwell JR, Moss DS (1991) The variety of X-ray diffuse-scattering from macromolecular crystals and its respective components. *Acta Crystallogr B Struct Sci* 47(Pt 6):960–968.
39. Helliwell JR, Glover ID, Jones A, Pantos E, Moss DS (1986) Protein dynamics—use of computer-graphics and protein crystal diffuse-scattering recorded with synchrotron X-radiation. *Biochem Soc Trans* 14(3):653–655.
40. Héry S, Genest D, Smith JC (1998) X-ray diffuse scattering and rigid-body motion in crystalline lysozyme probed by molecular dynamics simulation. *J Mol Biol* 279(1): 303–319.
41. Kolatkar AR, Clarage JB, Phillips GN, Jr (1994) Analysis of diffuse scattering from yeast initiator tRNA crystals. *Acta Crystallogr D Biol Crystallogr* 50(Pt 2):210–218.
42. Meinhold L, Smith JC (2005) Fluctuations and correlations in crystalline protein dynamics: A simulation analysis of staphylococcal nuclease. *Biophys J* 88(4):2554–2563.
43. Meinhold L, Smith JC (2005) Correlated dynamics determining x-ray diffuse scattering from a crystalline protein revealed by molecular dynamics simulation. *Phys Rev Lett* 95(21):218103.
44. Meinhold L, Smith JC (2007) Protein dynamics from X-ray crystallography: Anisotropic, global motion in diffuse scattering patterns. *Proteins* 66(4):941–953.
45. Meinhold L, Merzel F, Smith JC (2007) Lattice dynamics of a protein crystal. *Phys Rev Lett* 99(13):138101.
46. Mizuguchi K, Kidera A, Gō N (1994) Collective motions in proteins investigated by X-ray diffuse scattering. *Proteins* 18(1):34–48.
47. Phillips GN, Jr, Fillers JP, Cohen C (1980) Motions of tropomyosin. Crystal as metaphor. *Biophys J* 32(1):485–502.
48. Riccardi D, Cui Q, Phillips GN, Jr (2010) Evaluating elastic network models of crystalline biological molecules with temperature factors, correlated motions, and diffuse x-ray scattering. *Biophys J* 99(8):2616–2625.
49. Wall ME, Ealick SE, Gruner SM (1997) Three-dimensional diffuse x-ray scattering from crystals of *Staphylococcal* nuclease. *Proc Natl Acad Sci USA* 94(12):6180–6184.
50. Janowski PA, Cerutti DS, Holton J, Case DA (2013) Peptide crystal simulations reveal hidden dynamics. *J Am Chem Soc* 135(21):7938–7948.
51. Grissom CB, Markley JL (1989) Staphylococcal nuclease active-site amino acids: pH dependence of tyrosines and arginines by ^{13}C NMR and correlation with kinetic studies. *Biochemistry* 28(5):2116–2124.
52. Cotton FA, Hazen EE, Jr, Legg MJ (1979) Staphylococcal nuclease: Proposed mechanism of action based on structure of enzyme-thymidine 3′,5′-bisphosphate-calcium ion complex at 1.5-Å resolution. *Proc Natl Acad Sci USA* 76(6):2551–2555.
53. Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE (2012) Biomolecular simulation: A computational microscope for molecular biology. *Annu Rev Biophys* 41:429–452.
54. Earl DJ, Deem MW (2005) Parallel tempering: Theory, applications, and new perspectives. *Phys Chem Chem Phys* 7(23):3910–3916.
55. Pande VS, Beauchamp K, Bowman GR (2010) Everything you wanted to know about Markov state models but were afraid to ask. *Methods* 52(1):99–105.
56. Martínez E, Uberuaga BP, Voter AF (2014) Sublattice parallel replica dynamics. *Phys Rev E Stat Nonlin Soft Matter Phys* 89(6):063308.
57. Voter AF (1998) Parallel replica method for dynamics of infrequent events. *Phys Rev B Condens Matter* 57(22):13985–13988.
58. Chatfield DC, Szabo A, Brooks BR (1998) Molecular dynamics of staphylococcal nuclease: Comparison of simulation with ^{15}N and ^{13}C NMR relaxation data. *J Am Chem Soc* 120(21):5301–5311.
59. Showalter SA, Brüschweiler R (2007) Validation of molecular dynamics simulations of biomolecules using NMR spin relaxation as benchmarks: Application to the AMBER99SB force field. *J Chem Theory Comput* 3(3):961–975.
60. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78(8):1950–1958.
61. Lindorff-Larsen K, et al. (2012) Systematic validation of protein force fields against experimental data. *PLoS One* 7(2):e32131.
62. Beauchamp KA, Lin YS, Das R, Pande VS (2012) Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* 8(4):1409–1414.
63. Wall ME (2009) Methods and software for diffuse X-ray scattering from protein crystals. *Methods Mol Biol* 544:269–279.
64. Berman HM, et al. (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242.
65. Hynes TR, Fox RO (1991) The crystal structure of staphylococcal nuclease refined at 1.7 Å resolution. *Proteins* 10(2):92–105.
66. DeLano WL, The PyMOL Molecular Graphics System (Schrödinger, LLC, New York), Version 1.7.1.1.
67. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91(1-3):43–56.
68. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 118(45):11225–11236.
69. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 105(28):6474–6487.
70. Grosse-Kunstleve RW, Sauter NK, Moriarty NW, Adams PD (2002) The *Computational Crystallography Toolbox*: Crystallographic algorithms in a reusable software framework. *J Appl Cryst* 35(Pt 1):126–136.
71. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66(Pt 2):213–221.
72. Wall ME (1996) Diffuse features in X-ray diffraction from protein crystals. PhD thesis (Princeton University, Princeton).