Motional displacements in proteins: The origin of wave-vector-dependent values

Derya Vural, Liang Hong, Jeremy C. Smith, and Henry R. Glyde

Phys. Rev. E **91**, 052705 — Published 14 May 2015

DOI: 10.1103/PhysRevE.91.052705

Motional displacements in proteins, the origin of Q dependent values

Derya Vural,[1] Liang Hong,[2] Jeremy C. Smith,[2] and Henry R. Glyde[1]

[1]*Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716-2570, USA*

[2]*UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, P.O. Box 2008, Tennessee 37831 USA*

(Dated: April 20, 2015)

The average mean square displacement, $\langle r^2\rangle$, of H atoms in a protein is frequently determined using incoherent neutron scattering experiments. $\langle r^2\rangle$ is obtained from the observed elastic incoherent dynamic structure factor, $S_i(Q,\omega=0)$, assuming the form $S_i(Q,\omega=0) = \exp(-\,Q^2\langle\, r^2\rangle/3)$. This is often referred to as the Gaussian approximation (GA) to $S_i(Q,\omega=0)$. $\langle r^2\rangle$ obtained in this way depends on the value of the wave vector, Q considered. Equivalently, the observed $S_i(Q,\omega=0)$ deviates from the GA. We investigate the origin of the Q dependence of $\langle r^2\rangle$ by evaluating the scattering functions in different approximations using molecular dynamics (MD) simulation of the protein lysozyme. We find that keeping only the Gaussian term in a cumulant expansion of $S(Q,\omega)$ is an accurate approximation and is not the origin of the Q dependence of $\langle r^2\rangle$. This is demonstrated by showing that the term beyond the Gaussian is negligible and that the GA is valid for an individual atom in the protein. Rather, the Q dependence (deviation from the GA) arises from the dynamical heterogeneity of the H in the protein. Specifically it arises from representing, in the analysis of data, this diverse dynamics by a single average scattering center that has a single, average $\langle r^2\rangle$. The observed Q dependence of $\langle r^2\rangle$ can be used to provide information on the dynamical heterogeneity in proteins.

I. INTRODUCTION

The averaged mean-square motional displacement (MSD) of nuclei in proteins is extensively investigated using neutron scattering methods [1–12]. An MSD, $\langle r^2\rangle$, can be expressed in terms of the elastic ($\omega = 0$) component of the incoherent dynamic structure factor (DSF), $S_i(Q,\omega=0)$, where Q is the wave vector transfer in the scattering. For simplicity and generality, the data are typically analyzed using the approximate relation

$$S_i(Q,\omega=0) = AI_i(Q,t=\infty) \simeq A\exp(-\frac{1}{3}Q^2\langle r^2\rangle), \quad (1)$$

where $\langle r^2\rangle$ is an average MSD of the nuclei in the protein. $A = S_i(Q\ =0,\omega\ =0)$ is a convenient normalizing constant, the elastic DSF at $Q=0$. $I_i(\mathbf{Q},t=\infty)$ is the infinite time limit of the incoherent intermediate scattering function,

$$\begin{aligned} I_i(\mathbf{Q},t) &= \frac{1}{N}\sum_{j=1}^{N}\langle \mathrm{e}^{-i\mathbf{Q}\cdot\mathbf{r}_j(t)}\mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}_j(0)}\rangle && (2)\\ &= \frac{1}{N}\sum_{j=1}^{N}\langle \mathrm{e}^{-i\mathbf{Q}\cdot(\mathbf{r}_j(t)-\mathbf{r}_j(0))}\rangle. && (3)\end{aligned}$$

The second expression for $I_i(\mathbf{Q},t=\infty)$ holds in the classical limit. Since hydrogen (the proton) has a large incoherent scattering cross - section, 10 - 20 times that of other nuclei in the protein, the sum over j in Eq. (2) is well represented by a sum over the H in the protein. Hence, the average MSD, $\langle r^2\rangle$ is dominated by that of H in the protein (and in its associated hydration water, if any). $S_i(Q,\omega=0) = AI_i(Q,t=\infty)$ is often denoted the elastic incoherent structure factor (EISF).

Based on Eq. (1), the MSD $\langle r^2\rangle$ can be obtained from the observed $S_i(Q,\omega=0)$ as,

$$\langle r^2\rangle = -3\frac{\mathrm{d}\ln S_i(Q,\omega=0)}{\mathrm{d}Q^2}. \quad (4)$$

Similarly, based on Eq. (1), $\ln S_i(Q,\omega=0)$ vs. Q^2 will be a straight line. Using this approach, much progress has been made in determining the increase of $\langle r^2\rangle$ with temperature. In a wide range of hydrated proteins, there is a dynamical transition at a temperature $T_D \simeq 220$ K to a rapid increase of $\langle r^2\rangle$ with temperature. The large $\langle r^2\rangle$ values at higher temperature have been associated with protein function.

A number of assumptions are made in Eq. (1). For example, the use of Eq. (1) assumes that $S_i(Q,\omega=0)$ can be measured. In practice, since neutron instruments have a finite energy resolution width, W, $S_i(Q,\omega)$ around $\omega=0$ is always incorporated. As a result, a reduced $\langle r^2(\tau_R)\rangle$ that has had only a limited time $\tau_R \simeq \hbar/W$ to develop is observed, rather than the fully-developed, long time, intrinsic $\langle r^2\rangle$. The time scales, τ_R over which $\langle r^2(\tau_R)\rangle$ is observed have been extensively discussed [6, 9, 10, 13]. Methods have been developed to extract the intrinsic $\langle r^2\rangle$ from data taken at finite W and from $I_i(\mathbf{Q},t)$ calculated out to limited times only [14, 15]. These show, for example, that in lysozyme the intrinsic $\langle r^2\rangle$ is roughly twice the $\langle r^2(\tau_R)\rangle$ obtained from Eq. (1) at low Q for a resolution width $W = 1$ μeV ($\tau_R \simeq$ 1 ns). IN16 at Institut Laue-Langevin (ILL) and HFBS at National Institute of Standards and Technology (NIST) with $W = 1$ μeV are the instruments having the highest energy resolution.

The MSD obtained using Eq. (1) often depends on the value of Q at which the derivative in Eq. (4) is taken. Equivalently, the observed $\ln S_i(Q,\omega=0)$ vs. Q^2 is not a straight line. In short, the observed $\langle r^2\rangle$ obtained from Eq. (1) is Q dependent [1, 10, 16, 17]; $S_i(Q,\omega =$

0) deviates from the Gaussian approximation assumed in Eq. (1). Fig. 1 (Top) shows the ln$S_i(Q,\omega=0)$ vs. Q^2 observed by Daniel *et al.* in glutamate dehydrogenase [18]. The lines are a guide to the eye through the observed $S_i(Q,\omega=0)$. Clearly, a straight line ln$S_i(Q,\omega=0)$ vs. Q^2, as would be required to obtain a Q-independent $\langle r^2\rangle$ from Eq. (1), is not observed. The observed $\langle r^2\rangle$ generally decreases with increasing Q. Fig. 1 (Bottom) shows a Q-dependent $\langle r^2\rangle$ obtained by Calandrini *et al.* [19] from a molecular dynamics (MD) simulation-derived $S_i(Q,\omega=0)$ in lysozyme at ambient temperature. In Fig. 2 we show values of $\langle r^2\rangle$ obtained by us [15] from fits of a model $I(Q,t)$ that contains $I(Q,t\to\infty) = \exp[-\frac{1}{3}Q^2\langle r^2\rangle]$ to an MD simulation-derived $I_i(Q,t)$ for lysozyme. The fitted values of $\langle r^2\rangle$ clearly decrease with increasing Q as do observed values, especially at high temperature.

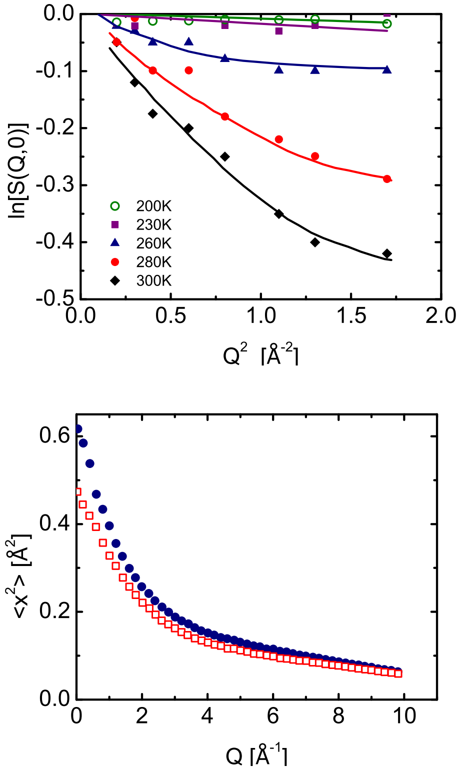

FIG. 1: (Color online) (Top): Elastic component ($\omega=0$) of the dynamical structure factor, $S(Q,\omega)$, as a function of wave vector Q of glutamate dehydrogenase observed by Daniel *et al.* [18] (also reproduced in Becker *et al.* [20]). The solid lines are a guide to the eye. (Bottom): An MSD obtained from fitting to $S(Q,\omega)$ derived from MD simulation of lysozyme in Calandrini *et al.*, which shows a strong Q dependence [19]). The solid circles and open squares represent the MSD for $p=0.1$ MPa and $p=300$ MPa.

The dependence of $\langle r^2\rangle$ on Q is unlikely to be physical. Rather, it arises because Eq. (1) is an approximate expression. The goal of the present paper is to examine the approximations in Eq. (1) and to determine the origin of this Q dependence. Specifically, our goal is to assess the validity of the approximations for the long-time, intrinsic $\langle r^2\rangle$ rather than on the resolution dependent/time limited MSD $\langle r^2(\tau_R)\rangle$. In the following subsection we identify the approximations and highlight some assessments of them.

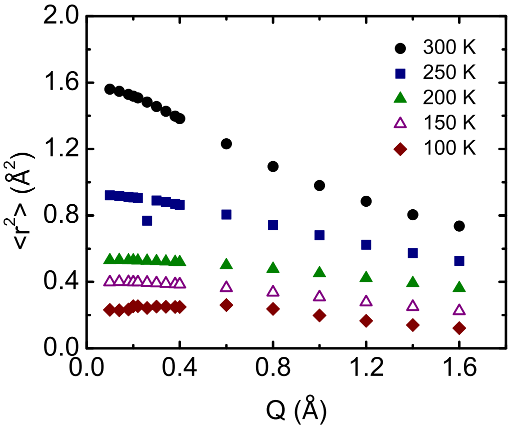

FIG. 2: (Color online) An intrinsic, long-time MSD in lysozyme which shows a strong Q dependence at 300 K. The MSD is obtained [15] by fitting a model intermediate scattering function $I(Q,t)$ to a calculated, $I_i(Q,t)$, obtained from a 1 μs MD simulation

1. *Approximations and their validity*

Beginning from Eq. (2), the first approximation (I) made to obtain Eq. (1) is a cumulant expansion of the individual terms (j) in $I_i(\mathbf{Q},t)$ and $I_i(\mathbf{Q},t=\infty)$ and retention of only the lowest-order, Gaussian term. With this approximation, the classical limit of $I_i(\mathbf{Q},t)$ in Eq. (2) reduces to $I_{iG}(\mathbf{Q},t)$ given by,

$$I_{iG}(\mathbf{Q},t) = \frac{1}{N}\sum_{j=1}^{N}\exp(-\frac{1}{2}\langle[\mathbf{Q}\cdot(\mathbf{r}_j(t)-\mathbf{r}_j(0))]^2\rangle)(5)$$

$$\simeq \frac{1}{N}\sum_{j=1}^{N}\exp(-\frac{1}{6}Q^2\Delta_j^2(t)), \qquad (6)$$

where $\Delta_j^2(t) = \langle(\mathbf{r}_j(t)-\mathbf{r}_j(0))^2\rangle$. In the second approximation (II), we assume that the motional distribution of the individual H has cubic or spherical symmetry so that $\langle[\mathbf{Q}\cdot(\mathbf{r}_j(t)-\mathbf{r}_j(0))]^2\rangle = Q^2\Delta_j^2(t)/3$. With approximation II, Eq. (5) reduces to Eq. (6). In the limit $t\to\infty$, which is assumed in Eq. (1), $\Delta_j^2(t)$ is,

$$\Delta_j^2(t=\infty) = \langle r_j^2(\infty)\rangle + \langle r_j^2(0)\rangle = 2\langle r_j^2\rangle. \qquad (7)$$

At $t\to\infty$, $I_{iG}(\mathbf{Q},t)$ in Eq. (6) is,

$$I_{iG}(\mathbf{Q},t=\infty) = \frac{1}{N}\sum_{j=1}^{N}\exp(-\frac{1}{3}Q^2\langle r_j^2\rangle). \qquad (8)$$

The third approximation (III) is to neglect the motional heterogeneity of H in the protein and assume that the $\langle r_j^2\rangle$ of all H is the same, $\langle r_j^2\rangle = \langle r^2\rangle$. With this approximation $I_{iG}(\mathbf{Q}, t = \infty)$ in Eq. (8) reduces to $I_i(\mathbf{Q}, t = \infty)$ in Eq. (1). However, even though each individual term in Eq. (8) is Gaussian the sum is not Gaussian if the $\langle r_j^2\rangle$ are not the same. With these three approximations and in the limit $t \to \infty$ we arrive at Eq. (1).

It is also useful to introduce an MSD that can be calculated directly from a MD simulation,

$$\Delta^2(t) = \frac{1}{N}\sum_{j=1}^{N}\Delta_j^2(t) = \frac{1}{N}\sum_{j=1}^{N}\langle[\mathbf{r}_j(t) - \mathbf{r}_j(0)]^2\rangle. \quad (9)$$

In the limit $t \to \infty$,

$$\Delta^2(t = \infty) = \frac{2}{N}\sum_{j=1}^{N}\langle r_j^2\rangle = 2\langle r^2\rangle_{MD}. \quad (10)$$

We note that the MSD $\langle r^2\rangle_{MD}$ in Eq. (10) is not the same as an average $\langle r^2\rangle$ calculated from Eq. (1) since the $\langle r^2\rangle$ in Eq. (1) arises from an average over exponentials as in Eq. (8).

To investigate the accuracy of retaining only the Gaussian term in the cumulant expansion, Hayward and Smith [21] calculated the full $I_i(\mathbf{Q}, t)$ for individual H (individual terms j in Eq. (2)) in an MD simulation of dry bovine pancreatic trypsin inhibitor. The $I_i(\mathbf{Q}, t)$ at long time was evaluated. If only the Gaussian cumulant in $I_i(\mathbf{Q}, t)$ is significant (i.e. Eq. (5) is valid), then $\ln I_i(\mathbf{Q}, t)$ vs. Q^2 should be a straight line. They found that for many individual H a straight line fitted well, for a significant number of H there was a modest deviation and for a handful of H having a large $\langle r_j^2\rangle$ there was a significant deviation. They concluded that keeping only the Gaussian cumulant was accurate at low Q. They investigated the dynamical heterogeneity (DH) by calculating an average $\Delta^2(t)$ from the slope of $I_{iG}(Q, t)$ in Eq. (6) vs. Q^2 and an average $\Delta^2(t)$ from Eq. (9). If there is no DH then the two $\Delta^2(t)$ will be the same. They found that the average $\Delta^2(t)$ calculated from Eq. (9) was 70 % larger than that obtained from Eq. (6) if the slope was averaged over a range $0 < Q < 4$ Å^{-1}. DH was clearly significant.

DH can also be described within equilibrium statistical mechanics using an H in a rugged potential landscape $V(r)$ representing the different environments seen by H. In this way DH was illustrated by Bicout [22] and Bicout and Zaccai [23] using a model of H in two different sites (cages). An equation similar to Eq. (8) above was used with j = 1,2. The temperature dependence was particularly developed.

Similarly, to model DH, Daniel *et al.* [3] and Becker and Smith [20] expressed the $I_i(\mathbf{Q}, t)$ in Eq. (2) at long times in terms of an average $\langle r^2\rangle$ as in Eq. (1) and a standard deviation, σ, of $\langle r_j^2\rangle$ from the average. They found that they could fit the curved $\ln S_i(Q, \omega = 0)$ vs. Q^2 well with this expression as shown in Fig. 1 (Top) where the curvature arises from the finite σ. Similarly, Yi *et al.* [24] determined σ for cytochrome P450cam from fits to experimental data.

Tokuhisa *et al.* [25] evaluated (1) two higher-order terms (fourth and sixth order) beyond the Gaussian in a cumulant expansion of $I_i(\mathbf{Q}, t)$ for individual atoms, (2) the degree of anisotropy in the motion of individual atoms and (3) the degree of dynamical heterogeneity among the atoms in the sum over j in Eq. (6), all from an MD simulation of staphylococal nuclase (SNase). They found that the two higher-order cumulants are negligible, and that the motional distribution of 90 % of the atoms is isotropic but again that DH is significant. They concluded that the deviation of $S_i(Q, \omega = 0)$ from a Gaussian can be used to assess the nature and degree of DH in SNase. In a more recent study Kneller and Chevrot [26] found that the motional anisotropy of individual H in lysozyme to be non-negligible.

In treating DH Meinhold *et al.* [27] and Kneller and Hinsen [28] have replaced the sum over j in Eq. (8) by an integral over a continuous distribution, $\rho(\langle r^2\rangle)$, of $\langle r^2\rangle$ values. Meinhold *et al.* employed a Weibull distribution that has two parameters. They fitted $\rho(\langle r^2\rangle)$ to MD-derived values of $\langle r_j^2\rangle$ at long time, calculated $\ln I_{iG}(Q, t)$ vs. Q^2, and found a good fit to the observed $S_i(Q, \omega = 0)$ vs. Q^2 except at $Q > 3$ Å^{-1}. Kneller and Hinsen used a Gamma distribution for $\rho(\langle r^2\rangle)$, to calculate $I_{iG}(Q, t)$ at $t \to \infty$ ($S_i(Q, \omega = 0)$) from the equivalent of Eq. (8). They found a good fit to an $I_i(\mathbf{Q}, t)$ at long time calculated from a MD simulation of lysozyme. In *Appendix C*, we note that $I_i(Q, t)$ a along time t (*e.g.* $t = 1 - 10$ ns) generally does not represent $I_i(Q, t = \infty)$ in the EISF $S_1(Q, \omega = 0) = AI_i(Q, t = \infty)$ well.

In further work, Peters and Kneller [29] fitted the Gamma distribution model of $S_i(Q, \omega = 0)$, to $S_i(Q, \omega = 0)$ of human acetylcholinesterase (hSChE) observed on 3 neutron spectrometers at the Institut Laue-Langevin, IN6 ($W = 50$ μeV), IN13 ($W = 8$ μeV) and IN16 ($W = 0.9$ μeV). They were able to obtain good fits to the data for which a Gaussian with a single $\langle r^2\rangle$, as in Eq. (1), failed. Also the average MSD, $\langle r^2\rangle$, when DH is incorporated showed a sharper and more clearly defined increase with T at the dynamical transition (DT). This means the DH widened above the DT. The average MSD was approximately 30-40 % larger at high temperature than that obtained using a fit with a single "average" $\langle r^2\rangle$. All the $\langle r^2\rangle$ quoted are "resolution broadened" values.

Finally, we remark that there could be a genuine dependence of $\langle r^2\rangle$ on Q. That is, a specific Q value implies that length scales up to $l = 2\pi/Q$ only can be sampled. If the motional displacement $\langle r^2\rangle^{1/2}$ is longer than l, the observed $\langle r^2\rangle^{1/2}$ could limited to l and therefore limited by Q. For example, at $Q = 2$ Å^{-1}, $\langle r^2\rangle$ observed would be limited to lengths of order $\langle r^2\rangle < \pi^2$ Å^2. However, from this argument we expect only very large MSD would be limited by using a Q of, say 2, Å^{-1}.

In the following sections we investigate the impact of making the Gaussian approximation, of neglecting dynamical heterogeneity and of possible limits to $\langle r^2\rangle$ at specific Q values. We investigate these approximations to $I_i(Q,t)$ itself and in fits to $I_i(Q,t)$.

II. SIMULATION AND SCATTERING FUNCTIONS

A. Molecular dynamics simulation of lysozyme

Two lysozyme molecules (1AKI [30]) were arbitrarily oriented and placed in a simulation box of dimensions 6.5 nm × 3.4 nm × 3.6 nm. The lysozyme molecules were surrounded by 636 water molecules, corresponding to a hydration level $h = 0.4$ g water/g protein. The box was replicated using periodic boundary conditions to mimic the environment of an experimental powder sample. The system was simulated using GROMACS 4.5.1 [31]. The OPLS-AA force field [32] was used for the protein and TIP4P [33] for the water. The van der Waals interactions were truncated at 1.4 nm, and the electrostatic interactions represented using the Particle Mesh Ewald method [34] with a real-space cutoff of 0.9 nm. All bonds including hydrogen bonds were constrained with a linear constraints solver algorithm (LINCS) [35]. The energy of the system was first minimized using 50000 steepest descent steps. The system was then equilibrated in the NVT (mole-volume-temperature) ensemble at each temperature investigated for 10 ns and in the NPT (mole-pressure-temperature) ensemble at 1 bar for 10 ns. The Nose-Hoover algorithm [36] with a coupling time $\tau = 1$ ps and the Parrinello-Rahman algorithm [37] with a coupling time $\tau = 3$ ps were used for the temperature coupling and pressure coupling, respectively.

Simulations of 100 ns length were performed at 300 K. The data were collected every 10 ps.

B. Scattering functions and MSD

The full incoherent intermediate scattering function, $I_i(\mathbf{Q},t)$ is given by Eq. (2). We evaluated $I_i(\mathbf{Q},t)$ arising from all the H in lysozyme including those that can exchange with H in the hydration water. The trajectories $\mathbf{r}_j(t)$ of the H were taken from the simulation described above. The ensemble average in Eq. (2) is an average over typically 100 time slices of the 100 ns simulation.

In programs such as GROMACS and SASSENA, the $I_i(\mathbf{Q},t)$ is averaged over many directions in the protein to obtain an $I_i(Q,t)$ that is a function of the absolute value of Q only. We found that the $I_i(Q,t)$ averaged over many directions in these programs could be accurately represented by an average of Q over just three perpendicular directions, as discussed in *Appendix B*. Hence, in what follows, we used $I_i(Q,t)$ averaged over these three directions only.

$I_i(Q,t)$ in the Gaussian approximation, $I_{iG}(\mathbf{Q},t)$, defined in Eq. (5), was also calculated from the same simulation and averaged over three perpendicular directions (exactly as was $I_i(\mathbf{Q},t)$) to obtain a Gaussian approximation $I_{iG}(Q,t)$(see *Appendix B*). In this way, $I_i(Q,t)$ and $I_{iG}(Q,t)$ can be compared directly.

To analyse the simulation-derived data we fit a model of a single scatterer intermediate scattering function,

$$I(\mathbf{Q},t) = \langle \exp(-i\mathbf{Q}.\mathbf{r}(t)) \exp(i\mathbf{Q}.\mathbf{r}(0))\rangle. \qquad (11)$$

to the full $I_i(Q,t)$ and to individual terms j in $I_i(Q,t)$. The model is [14, 15],

$$I(Q,t) = I_\infty(Q) + (1 - I_\infty(Q))C(Q,t), \qquad (12)$$

where

$$I_\infty = I(Q,t=\infty) = \exp(-\frac{1}{3}Q^2\langle\ r^2\rangle). \qquad (13)$$

is the infinite time limit of $I(Q,t)$ that defines an infinite time, intrinsic MSD, $\langle r^2\rangle$. In the previous work we used this model to obtain intrinsic long-time MSDs from experiment and simulations [14, 15].

The model is essentially a separation of $I(\mathbf{Q},t)$ into a time independent part, I_∞ and a time dependent part, $I'(\mathbf{Q},t) = (1 - I_\infty)C(Q,t)$, where $C(Q,t)$ has the limits $C(Q,t=0) = 1$, $C(Q,t\to\infty) = 0$. $C(Q,t)$ is represented by a stretched exponential function,

$$C(Q,t) = \exp(-(\lambda t)^\beta), \qquad (14)$$

where λ and β are constants. $C(Q,t)$ represents the decay of correlations in the protein and has the desired limits.

The time independent part is the $t \to \infty$ limit of $I(\mathbf{Q},t)$,

$$\begin{aligned} & I(\mathbf{Q},t=\infty) & (15)\\ =\ & \langle\exp(-i\mathbf{Q}\cdot\mathbf{r}(\infty))\rangle\langle\exp(i\mathbf{Q}\cdot\mathbf{r}(0))\rangle \\ =\ & \langle\exp(-i\mathbf{Q}\cdot\mathbf{r})\rangle\langle\exp(i\mathbf{Q}\cdot\mathbf{r})\rangle \\ =\ & \langle\exp[-\langle[\mathbf{Q}\cdot\mathbf{r}]^2\rangle + \frac{1}{12}(\langle[\mathbf{Q}\cdot\mathbf{r}]^4\rangle - 3\langle[\mathbf{Q}\cdot\mathbf{r}]^2\rangle^2) + \cdots]. \end{aligned}$$

To obtain the last line of Eq. (15) we have assumed (1) that $r(\infty)$ and $r(0)$ are statistically independent, and (2) that $\langle r(\infty)\rangle = \langle r(0)\rangle = \langle r\rangle$ is independent of time and (3) made a cumulant expansion [14] of $\langle\exp(\pm i\mathbf{Q}\cdot\mathbf{r})\rangle$. The expression $I_\infty = \exp[-\frac{1}{3}Q^2\langle r^2\rangle]$ is obtained by neglecting all the higher-order cumulants in Eq. (15) beyond the second order, Gaussian cumulant and assuming cubic or spherical symmetry so that $\langle[\mathbf{Q}\cdot\mathbf{r}]^2\rangle = \frac{1}{3}Q^2\langle r^2\rangle$.

The model was fitted to calculated $I_i(Q,t)$ and $I_{iG}(Q,t)$ with $\langle r^2\rangle$, λ and β treated as free fitting parameters to be determined. In this way we obtain an intrinsic MSD, $\langle r^2\rangle$, from fits to simulations of $I_i(Q,t)$ in much the same way that $\langle r^2\rangle$ is determined from experimental data.

III. RESULTS

A. Higher cumulants and the Gaussian approximation

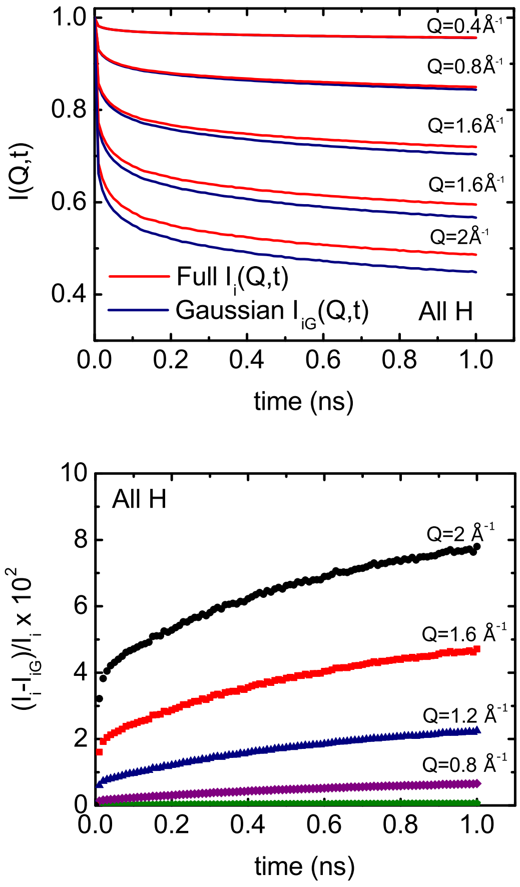

FIG. 3: (Color online) (Top): Comparison of the full intermediate scattering function (ISF), $I_i(Q,t)$, containing all cumulants (red solid lines) and the Gaussian limit, $I_{iG}(Q,t)$, containing only the second order cumulant (blue solid lines) calculated for all H in lysozyme at 300 K for $0 < t < 1$ ns obtained from a 100 ns MD simulation. From top to bottom the Q values are 0.4, 0.8, 1.2, 1.6 and 2 Å^{-1}. (Bottom) The percentage difference between $I_i(Q,t)$ and $I_{iG}(Q,t)$ for Q values 2 to 0.4 Å^{-1}.

In this section, we assess the importance of higher cumulants in the DSF. To do this, we firstly compare the full ISF, $I_i(Q,t)$, given by Eq. (3), which contains all the cumulants, with that obtained making the Gaussian approximation for each individual atom and summing the Gaussians (Eq. (6)). We may view this as comparing the full and Gaussian ISF for each individual H in the protein and summing over all H, $j = 1$ to N. Fig. 3(top) shows the full $I_i(Q,t)$ and the Gaussian $I_{iG}(Q,t)$ arising from all H in lysozyme at 300 K. At low Q, as expected, the difference between the two is negligible. However, at $Q = 2$ Å^{-1}, an observable difference develops with time and amounts to 8% after 1 ns (Fig. 3(bottom)).

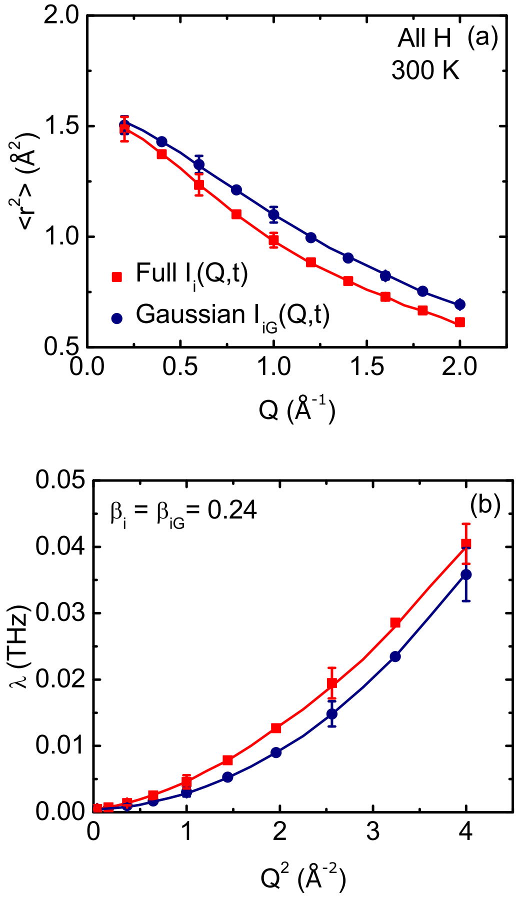

FIG. 4: (Color online) The MSD, $\langle r^2\rangle$, (Top (a)) and the relaxation parameter, λ, (Bottom (b)) obtained from fits of the model $I(Q,t)$ given by Eq. (12) to the calculated full $I_i(Q,t)$ (red solid squares) and to the Gaussian approximation $I_{iG}(Q,t)$ given by Eq. (6) (blue solid circles). The MSD obtained from a fit to the Gaussian approximation $I_{iG}(Q,t)$ remains Q dependent and similar to that obtained from a fit to the full $I_i(Q,t)$.

To test the impact of this difference, we fit the model function $I(Q,t)$ given by Eq. (3) to both $I_i(Q,t)$ and $I_{iG}(Q,t)$. The aim is to determine the intrinsic MSD, $\langle r^2\rangle$, and decay parameters λ and β in the stretched exponential and see how much they are affected by omitting the higher-order cumulants. The $\langle r^2\rangle$ and λ obtained from the fits are shown in Fig. 4. A detailed discussion of the fits can be found in Ref. [15]. Error bars for $\langle r^2\rangle$ and λ are shown in Fig. 4 for some Q values, e.g. $Q = 0.2$ and 2.0 Å^{-1} for $\langle r^2\rangle$ and $Q = 1.6$ and 2.0 Å^{-1} for λ. These error bars were obtained directly from a fit program and as such are not always precise. The smooth variation of both $\langle r^2\rangle$ and λ with Q in Fig. 4 suggests an error bar comparable to the size of the points. The

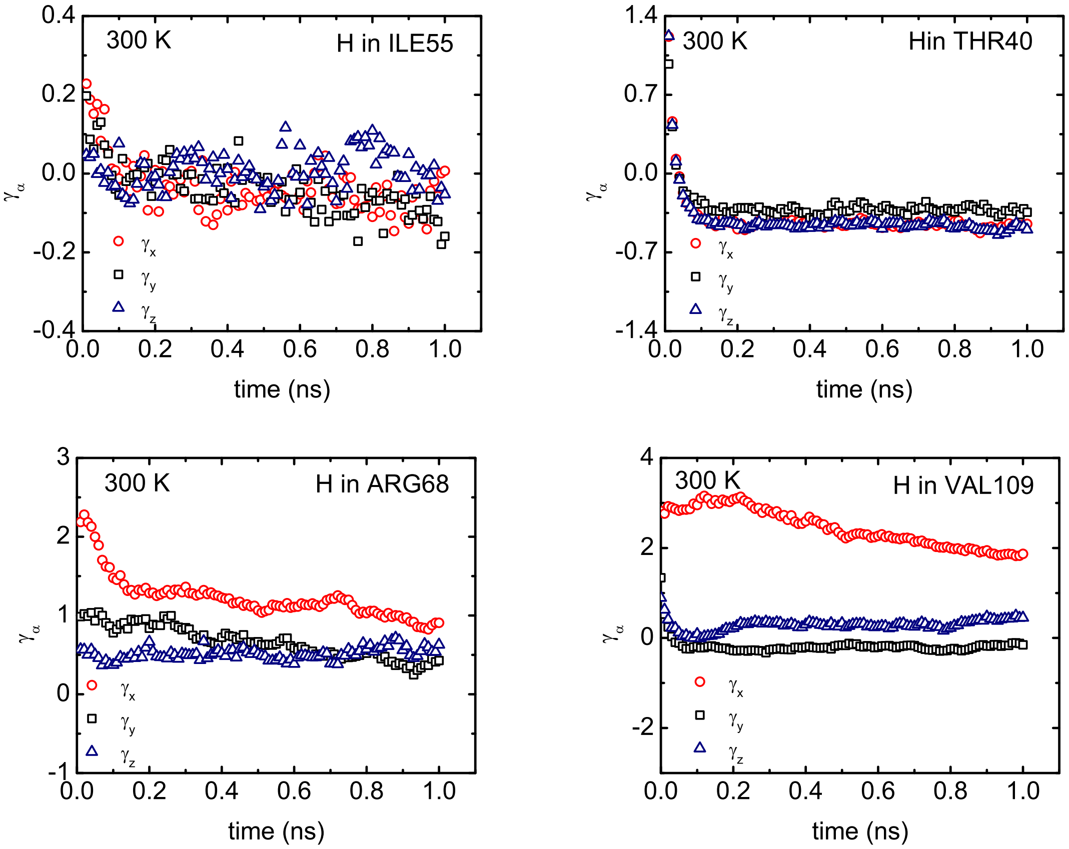

FIG. 5: (Color online) The kurtosis γ_α of the 4th cumulant defined in Eq. (19) along x (red open circles), y (black open squares) and z (blue open triangle) for a single H in lysozyme at 300 K using $r(t)$ calculated in a 100 ns MD simulation. In ILE55 and in ARG68, the single H is bonded to C_β and in THR40 and in VAL109 the single H is in $C_\gamma H_3$ bonded to C_β.

best fit β was found to be largely independent of Q, at $\beta \simeq 0.24$. Hence, for consistency, β was held fixed at $\beta = 0.24$ in both cases. The decay parameter, λ, obtained is not exactly proportional to Q^2. A Q^2 dependence is expected for purely diffusive motions, e.g, the ISF for translational diffusion is given by $I(Q,t) = \exp(-\lambda t)$ with $\lambda = \tau^{-1} \propto DQ^2$ where D is diffusion constant [38]. The best fit values of the intrinsic $\langle r^2\rangle$ are clearly similar for both $I_i(Q,t)$ and $I_{iG}(Q,t)$ which shows that making the Gaussian approximation has only a small impact on the MSD extracted (Fig. 4). The $\langle r^2\rangle$ remains Q dependent in both cases (Fig. 4(top)). In particular, if we fit a model $I(Q,t)$ in which the Gaussian approximation is made to an $I_{iG}(Q,t)$ in which the Gaussian approximation is also made, we still obtain a Q dependent MSD.

A second test can be made: we can explicitly calculate the magnitude of the higher cumulants for the individual H in the full $I_i(Q,t)$ in Eq. (2) and determine whether they are significant compared to the second order, Gaussian term. The next cumulant beyond the Gaussian is that of fourth-order. Retaining the fourth cumulant (see *Appendix A*), the ISF at ($t \to \infty$) is,

$$I_i(\mathbf{Q}, t=\infty) = \frac{1}{N}\sum_{j=1}^{N} \exp(-\frac{1}{3}Q^2\langle r^2\rangle_j[1 - a_jQ^2]) \quad (16)$$

where a_j represents the magnitude of the 4th cumulant of atom j. From *Appendix A* we see that for a specific H $a = \frac{1}{36}\langle r^2\rangle\gamma_\alpha$ where $\langle r^2\rangle$ is the MSD and γ_α is the kurtosis of the distribution projected along axis α of the specific H. The fourth cumulant is negligible if $aQ^2 <<$ 1. We evaluated the MSD and kurtosis of many H atoms in lysozyme at 300 K.

Fig. 5 shows γ_α for four of these H. H in ILE55 and ARG68 refer to H bonded to C_β, and H in THR40 and VAL109 refer to H in $C_\gamma H_3$ in these residues. In For H in ILE55, γ_α is small ($\gamma_\alpha < 0.1$). Indeed, most H in lysozyme have a small γ_α, comparable to that in ILE55. In contrast, the γ_α for H in VAL 109 is exceptionally large. Generally, the larger the MSD the larger is γ_α (see Fig. 9 for a histogram of the MSDs). For the H in ILE55, taking $\gamma_\alpha = \simeq 0.1$ and $\langle r^2\rangle\ = 0.4$ Å^2 (see Fig. 7), we have at $Q = 2$ Å^{-1}, $aQ^2 \simeq 0.005$. In comparison, for the H atoms in THR40, ARG68 and VAL109 aQ^2 is 0.01, 0.07 and 0.50, respectively. We found that, of the 1918 H in lysozyme, only for LYS116 and VAL109 is aQ^2 significant compared to the Gaussian term ($aQ^2 \gtrsim 0.1$). As a result, when summed over all the H in lysozyme, the fourth-order cumulant is negligible. The kurtosis and all higher cumulants are exactly zero if the motional distribution of the H is Gaussian. In summary, then, from a direct calculation, we find the fourth cumulant can practically be neglected.

B. Dynamical heterogeneity

To test the impact of dynamical heterogeneity we calculate $I_i(Q,t)$ for individual H in the lysozyme. Fig. 6 shows $I_i(Q,t)$ and $I_{iG}(Q,t)$ for a single specific H in ILE55 in lysozyme at 300 K. The $I_i(Q,t)$ for this single H has a similar shape to that for averaged over all H in lysozyme (Fig. 3(top)) but reaches a plateau in a shorter time. We investigated many single H. For each of these, $I_i(Q,t)$ and $I_{iG}(Q,t)$ are very similar if $\langle r^2\rangle$ of the H is small (as for ILE55) but there is some difference between them when the $\langle r^2\rangle$ is larger (e.g. $\langle r^2\rangle \gtrsim 2.0$ Å^2).

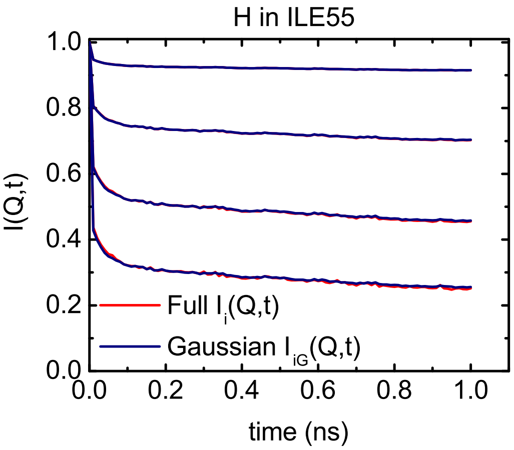

FIG. 6: (Color online) The $I_i(Q,t)$ (red solid lines) and $I_{iG}(Q,t)$ (blue solid lines) as in Fig. 3 calculated for a single H in lysozyme at 300 K, denoted H in ILE55, versus t obtained from a 100 ns MD simulation . The Q values are 1 to 4 Å^{-1}, top to bottom.

Fig. 7 shows the $\langle r^2\rangle$ and λ obtained by fitting the model $I(Q,t)$ to the calculated $I_i(Q,t)$ for the single H in ILE55. The $\langle r^2\rangle$ is quite independent of Q. The value of λ obtained from fits to $I_i(Q,t)$ and $I_{iG}(Q,t)$ are somewhat different but β is small in each case. A small β means that the motional decay ($C(Q,t)$) has a long tail extending out to long times.

Fig. 8 shows $\langle r^2\rangle$ of several individual H obtained from fits to $I_i(Q,t)$. Again, as shown in the top frame of Fig. 8, the fitted $\langle r^2\rangle$ are quite independent of Q if the $\langle r^2\rangle$ is small, $\langle r^2\rangle \lesssim 2$ Å^2. This is the case for the vast majority of the H in lysozyme. In contrast, for the few individual H that have very large $\langle r^2\rangle$, $\langle r^2\rangle \gtrsim 3$ Å^2, the $\langle r^2\rangle$ is found to be Q dependent, as shown in the bottom frame of Fig. 8. However, there are so few H that have large MSD, $\langle r^2\rangle \gtrsim 3$ Å^2, that these exceptional H atoms contribute little to the $I_i(Q,t)$ summed over all H (some 1918 H).

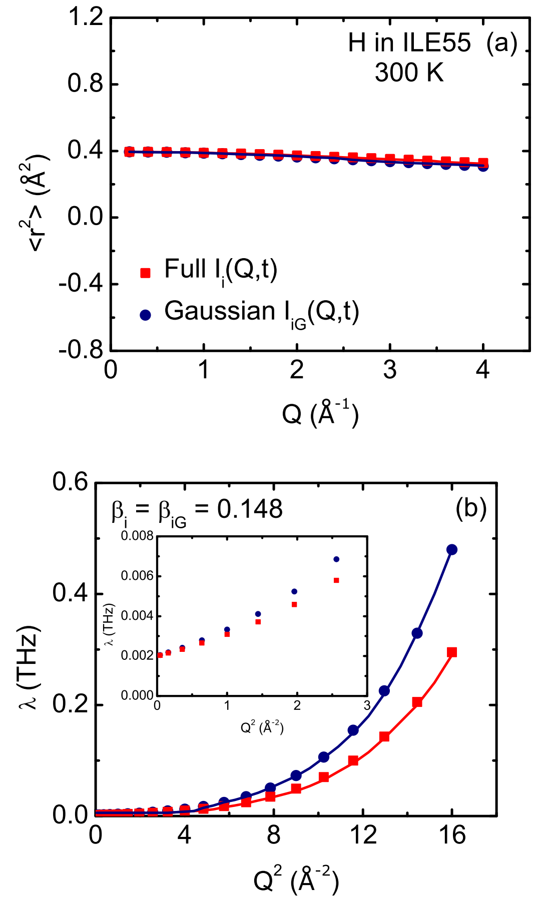

FIG. 7: (Color online) The MSD, $\langle r^2\rangle$, (Top (a)) and the relaxation parameter, λ, (Bottom (b)) obtained from fits of the model $I(Q,t)$ given by Eq. (12) to the calculated $I_i(Q,t)$ (red solid squares) and Gaussian approximation $I_{iG}(Q,t)$ (blue solid circles) for a single H in lysozyme at 300 K, (the same single H as shown in Fig. 6, i.e. a H in ILE55). The MSD, $\langle r^2\rangle$, obtained is independent of Q in both cases.

In Fig. 9 (Top) we show $\langle r^2\rangle$ for two individual H (in $C_\beta H_3$ in ALA122 and in $C_\gamma H_3$ in THR40) for which both $\langle r^2\rangle$ and the kurtosis are large. For these two H there is a small but significant difference between $\langle r^2\rangle$ obtained from $I_i(Q,t)$ and $I_{iG}(Q,t)$. The results of the section above suggest that this difference arises from the higher cumulants in $I_i(Q,t)$ that are neglected in $I_{iG}(Q,t)$. However, the difference is small and the $\langle r^2\rangle$ are approximately independent of Q, confirming that the large observed Q dependence does not arise from neglecting higher cumulants.

Fig. 9 (Bottom) shows a histogram of the MSD

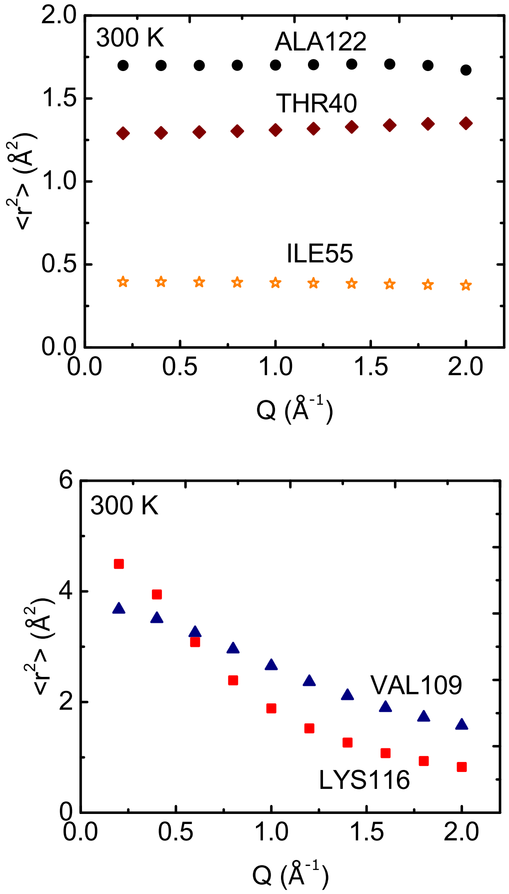

FIG. 8: (Color online) The MSD, $\langle r^2\rangle$, obtained from fits of the model $I(Q,t)$ to simulations of $I_i(Q,t)$ for a single H in lysozyme (no dynamical heterogeneity). (Top): The MSD of H in ALA122 (solid circles), of H in THR40 (solid diamonds) and H in ILE55 (open stars) which are of moderate size and independent of Q. (Bottom): The MSD of two single H that have large and Q dependent MSDs, H in LYS116 (solid squares) and H in VAL109 (solid triangles).

$\Delta_j^2(t)$ defined in Eq. (7) calculated from the 100 ns simulation at time $t = 1$ ns. The $\Delta_j^2(t)/2$ at $t = 1$ ns are lower but approximately equal to $\langle r_j^2\rangle$. For example, for H in THR40 and in ALA122 the $\Delta_j^2(t)/2$ lie 1% and 5% below $\langle r_j^2\rangle$, respectively. From the histogram we see that the H in ILE55 lies in the main peak of the distribution, i.e., H that have MSDs $\langle r_j^2\rangle \simeq 0.5$ Å^2. The H in ALA122 and THR40 belong to a smaller group of H that have MSDs in the range $\langle r_j^2\rangle \approx 1.5$ Å^2. In contrast the H in VAL109 belongs to a very small group that has exceptionally large MSDs, $\langle r_j^2\rangle \simeq 2.5 - 3.5$ Å^2. In Fig. 9 (Bottom), the H in ILE55 which lies in the main peak is embedded in lysozyme, whereas the H in ALA122 which lies in the second peak at 1.3 Å^2 is located on the surface of lysozyme. We also observed a small peak in the distribution function at 2.7 Å^2, and H in VAL109 lying in this peak has a large MSD and is located on the surface

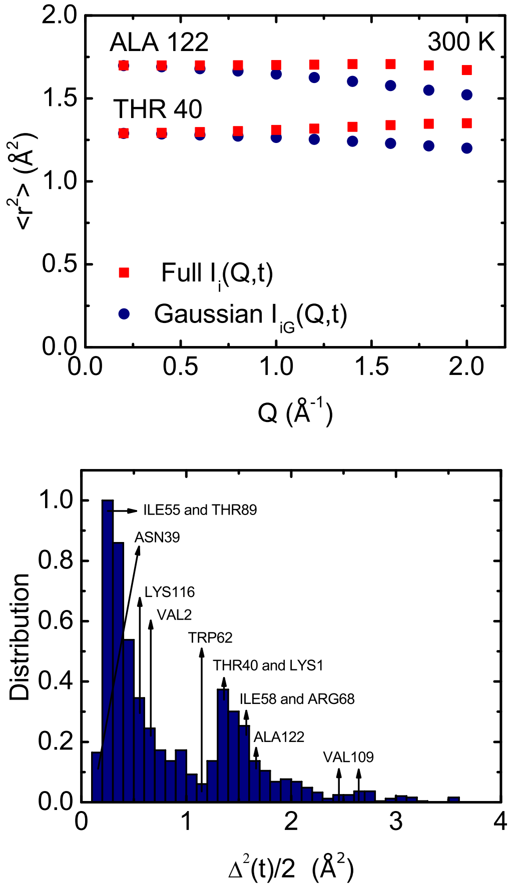

FIG. 9: (Color online) (Top): The $\langle r^2\rangle$ obtained from the fit of the model $I(Q,t)$ to the Full $I_i(Q,t)$ (solid squares) and the Gaussian approximation $I_{iG}(Q,t)$ (solid circles) for individual H in ALA122 and THR40 that have large MSD and kurtosis. There is some difference in the MSD obtained from $I_i(Q,t)$ and $I_{iG}(Q,t)$ but the difference is small. (Bottom): The distribution of the MSD of individual H in lysozyme at 300 K calculated using $\Delta_j^2(t)/2 = \langle(r_j(t) - r_j(0))^2\rangle/2$ at time $t = 1$ ns from a 100 ns MD simulation.

.

of the mouth part of lysozyme.

From these results, we may conclude that the Q dependence of $\langle r^2\rangle$ obtained by fitting the model $I(Q,t)$ to $I_i(Q,t)$ for all H in lysozyme (e.g. Fig. 4) arises from the dynamical heterogeneity of the H in $I_i(Q,t)$, a diversity that is not contained in the model $I(Q,t)$. When the same model is fitted to $I_i(Q,t)$ of a single H, a Q independent $\langle r^2\rangle$ is obtained (e.g. Figs. 7 and 9).

IV. DISCUSSION

A. Cumulant Expansion

In the previous section we assessed the accuracy of keeping only the lowest order (second order), Gaussian term in a cumulant expansion of the incoherent DSF of a globular protein. We found it to be an accurate approximation up to $Q = 2 - 3$ Å^{-1}. We tested the accuracy in two ways. Firstly, we compared the full incoherent ISF, $I_i(Q,t)$, given by Eq. (2), which contains all the cumulants, with the ISF, $I_{iG}(Q,t)$, given by Eq. (6) which contains only the lowest order, Gaussian cumulant. There is a small difference between the two which increases with increasing time (see Fig. 3). At 300 K and $Q = 2$ Å^{-1} the difference reached 8 % after $t = 1$ ns, the longest time considered. The difference between the MSD, $\langle r^2 \rangle$, obtained by fitting to $I_i(Q,t)$ and to $I_{iG}(Q,t)$ was not significant (see Fig. 4). Both the associated $\langle r^2 \rangle$ depend markedly on Q. This shows that omitting all higher cumulants does not change the Q dependence of $\langle r^2 \rangle$ and that the higher cumulants are not responsible for the deviation of $\langle r^2 \rangle$ in Eq. (4) from a constant.

Secondly, for the individual H(j), we evaluated the leading cumulant beyond the Gaussian (fourth-order) in the cumulant expansion of the EISF = $I_i(Q, t = \infty)$. The fourth-order cumulant for individual H was found to be quite negligible except for a few of the 1918 H atoms in lysozyme. Consequently, when summed over all H at Q = 2 Å^{-1}, the fourth cumulant is negligible compared to the Gaussian term.

As noted in the Introduction, Tokuhisa and coworkers [25] evaluated the cumulant expansion of the EISF for individual atoms in Staphylococcal Nuclase (SNase) keeping terms up to 4th and 6th order. They found that the 4th and 6th order terms were negligible compared to the Gaussian when the average over all the H in the protein was taken. Since the higher cumulants are negligible in both SNase and lysozyme, it is tempting to extrapolate this finding and conclude that retaining only the Gaussian term in a cumulant expansion of the EISF or $I_i(\mathbf{Q},t)$ is an excellent approximation for H in most folded proteins. It would be of interest to examine whether this observation also holds for denatured or intrinsically disordered proteins.

The higher cumulants beyond the Gaussian are exactly zero for a Gaussian motional distribution. The above results suggests that the motional distributions of H in proteins are indeed nearly Gaussian. For example, Hong *et al.* [39] find that the motion of H at 300 K in lysozyme is well represented by a Gaussian motional distribution in different sites with rapid jumps between these sites. Diffusion in a harmonic potential has a Gaussian spatial distribution.

Tokuhisa and coworkers [25] also find that the isotropic approximation, in which $\langle [\mathbf{Q} \cdot \mathbf{r}_j(t)]^2 \rangle$ is replaced by $Q^2 \langle {r_j}^2 \rangle / 3$, is valid in SNase and not responsible for any deviation of the EISF from a Gaussian. Although some 10 % of H in SNase show an anisotropic $\langle {r_j}^2 \rangle$ the 10 % have small $\langle r^2 \rangle$ values and contribute little to the EISF. We did not evaluate this approximation in lysozyme. We did, however, find that $I_i(\mathbf{Q},t)$ itself is quite isotropic. Specifically $I_i(\mathbf{Q},t)$ evaluated for three perpendicular $\mathbf{Q}$ values are indistinguishable from one another and independent of direction (see *Appendix B*). We attribute this isotropy of $I_i(\mathbf{Q},t)$ to there being a large number of H in similar structures that have different orientations in lysozyme.

B. Dynamical heterogeneity

We find that the Q dependence of the average $\langle r^2 \rangle$ obtained from Eq. (4), or equivalently the deviation of $S_i(Q, \omega = 0)$ from a Gaussian in Q, arises from the dynamical heterogeneity of H in lysozyme. We showed this by fitting the same "representative atom" model that led to a Q dependent $\langle r^2 \rangle$ to the calculated $I_i(Q,t)$ of single H atoms where there can be no dynamical heterogeneity, leading to a Q independent $\langle (r_j)^2 \rangle$ for the single H atoms. Tokuhisa and coworkers [25] reached an equivalent conclusion for SNase. This opens the way to using the Q dependence of $\langle r^2 \rangle$ to test models of distribution of the $\langle (r_j)^2 \rangle$ without concern that the Q dependence could arise from other factors. Several studies aimed at extracting the distribution of $\langle r^2 \rangle$ values, or moments of the distribution, have already been proposed [3, 24, 28, 29, 40]. For example, Nakagawa *et al.* [40] find a bimodal distribution of MSDs is consistent with data. A somewhat bimodal distribution of MSDs is found in simulations of SNase [25] and a clearly bimodal distribution here in Fig. 9 for lysozyme. Daniel *et al.* [3] and Yi *et al.* [24] determined the second moment of the distribution of $\langle (r_j)^2 \rangle$ from fits to data.

Peters and Kneller [29] have made fits to observed EISF using a model in which the $\langle r^2 \rangle$ have a Gamma distribution [28]. In this distribution a large fraction of the H have small $\langle r^2 \rangle$ values with a tail in the distribution reaching large $\langle r^2 \rangle$ values. For example, the histogram of $\langle r^2 \rangle \simeq \Delta^2/2$ in lysozyme shown in Fig. 9 has a peak at $\langle r^2 \rangle \simeq 0.25$ Å^2 and a tail reaching up to $\langle r^2 \rangle = 3 - 4$ Å^2. A much better fit was found to the observed EISF out to $Q = 4.5$ Å^{-1} using a Gamma distribution than using a single average $\langle r^2 \rangle$ [29]. Furthermore, when a distribution of $\langle r^2 \rangle$ was used, the average $\langle r^2 \rangle$ obtained was also much larger and the dynamical transition temperature, T_D, more consistently determined. In Fig. 2, we showed an average $\langle r^2 \rangle$ in lysozyme that has a large dependence on Q at high temperature (300 K) indicating high dynamical heterogeneity. In Ref. [24, 29], both the average $\langle r^2 \rangle$ and the mean square deviation of the $\langle r^2 \rangle$ from the average increase significantly with temperature at temperatures above the dynamical transition temperature, T_D.

V. CONCLUSION

We have investigated the origin of Q dependent values of the average MSD, $\langle r^2 \rangle$, observed in neutron scattering experiments. A Q dependent $\langle r^2 \rangle$ is found assuming that the elastic part of the normalized, incoherent DSF,

$S_i(Q,\omega\ = 0) = AI_i(Q,t = \infty)$, is given by Eq. (1). A Q dependent MSD is equivalent to finding a deviation of $I_i(Q,t)$ from a Gaussian. The origin of the deviation was investigated by calculating $I_i(\mathbf{Q},t)$ exactly and with approximations from a simulation of lysozyme.

We find that the deviation from a Gaussian does not arise from neglecting the higher order cumulants in the elastic incoherent DSF. Rather, we find that retaining only the second order, Gaussian term in the cumulant expansion is an accurate approximation out to $Q = 2$ - 3 Å^{-1}. The approximation was tested firstly by calculating the full incoherent function $I_i(Q,t)$ and the Gaussian approximation to it, $I_{iG}(Q,t)$, for all H and showing that the difference between the two is small and that the MSD $\langle r^2\rangle$ obtained from the two differs insignificantly. It was secondly tested by calculating the leading term beyond the Gaussian, the fourth cumulant, and showing that this is negligible. The Q dependence of the MSD obtained from experiments does not arise from keeping only the Gaussian cumulant in $S(\mathbf{Q},\omega=0)$.

We find that the apparent Q dependence of the observed MSD arises from neglecting the dynamical heterogeneity of H in a protein in the analysis of data. It is the use of a single scatterer or "representative atom" model to fit data arising from thousands H that have a wide spectrum of MSD values that leads to a Q dependent average MSD. If the same model is fitted to $I_i(Q,t)$ arising from a single H, then a Q independent $\langle r^2\rangle$ is obtained. This holds except for a few individual H that have very large MSD and may have a non-Gaussian distribution. There are so few such H that this does not affect the observed $\langle r^2\rangle$.

The finding that the deviation of $S_i(Q,\omega = 0)$ from a Gaussian and the Q dependence of $\langle r^2\rangle$ arises solely from the dynamical heterogeneity opens the way for confident empirical determination of the distribution of $\langle r^2\rangle$ values in proteins from neutron scattering data.

VI. APPENDIX A : THE IMPORTANCE OF FOURTH CUMULANT TERM

To assess the importance of higher cumulants, we calculate the fourth cumulant explicitly for individual H in Eq. (2) in lysozyme in the long time limit. The intermediate scattering function $I(\mathbf{Q},t)$ at infinite time ($t \to \infty$) for a specific H in Eq. (2) is, as in Eq. (15),

$$I(\mathbf{Q},\infty) = \exp[-\langle[\mathbf{Q}\cdot\mathbf{r}]^2\rangle+\frac{1}{12}[\langle[\mathbf{Q}\cdot\mathbf{r}]^4\rangle-3\langle[\mathbf{Q}\cdot\mathbf{r}]^2\rangle^2]+\cdots] \tag{17}$$

For $\mathbf{Q}$ parallel to the direction α, $\langle[\mathbf{Q}\cdot\mathbf{r}]^2\rangle = Q^2\langle r_\alpha^2\rangle$, $\langle[\mathbf{Q}\cdot\mathbf{r}]^4\rangle = Q^4\langle r_\alpha^4\rangle$, and

$$\begin{aligned} I(\mathbf{Q},\infty) &= \exp[-Q^2\langle r_\alpha^2\rangle + \frac{1}{12}Q^4[\langle r_\alpha^4\rangle - 3\langle r_\alpha^2\rangle^2]] \\ &= \exp[-Q^2\langle r_\alpha^2\rangle[1 - \frac{1}{12}Q^2\gamma_\alpha\langle r_\alpha^2\rangle]] \end{aligned} \tag{18}$$

where

$$\gamma_\alpha = \frac{\langle r_\alpha^4\rangle - 3\langle r_\alpha^2\rangle^2}{\langle r_\alpha^2\rangle^2} \tag{19}$$

is the kurtosis of the motional distribution.

The 4th cumulant is negligible if the kurtosis γ_α is small enough that the term $\frac{1}{12}Q^2\gamma_\alpha\langle r_\alpha^2\rangle$ in Eq. (18) is small compared to 1. Note $\langle r_\alpha^2\rangle \approx \frac{1}{3}\langle r^2\rangle$. For example, for $\gamma_\alpha \sim 1$, $\langle r^2\rangle \sim 1$ Å^2, $Q^2 \sim 4$ Å^{-2}, we obtain $\frac{1}{36}Q^2\gamma_\alpha\langle r^2\rangle \sim \frac{2}{18} \sim \frac{1}{10}$. For almost all H in lysozyme, as shown in the top half of Fig. 5 for ILE55 and THR40, γ_α is much less than unity, $\langle r^2\rangle < 1$, and the 4th cumulant is negligible at $Q \sim 2$ Å^{-1}. Only in those very few exceptional H, such as VAL 109 for which $\langle r^2\rangle \sim 3$ Å^2 and $\gamma_\alpha \sim 1 - 2$, is the 4th cumulant of any significance at $Q \sim 2$ Å^{-1}.

VII. APPENDIX B : CALCULATING THE FULL $I_i(\mathbf{Q},t)$ AND GAUSSIAN $I_{iG}(\mathbf{Q},t)$

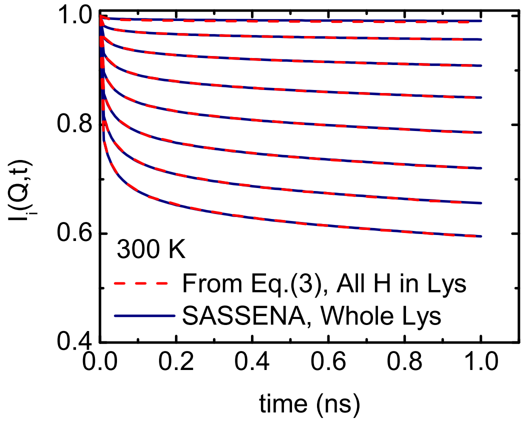

FIG. 10: Comparison of the ISF, $I_i(Q,t)$, arising from all H in lysozyme (red solid lines) calculated from Eq. (20) and from all nuclei in lysozyme (black dashed lines) calculated using SASSENA.

The incoherent ISF $I_i(\mathbf{Q},t)$ in Eq. (2) depends on the vector $\mathbf{Q}$. The $I_i(\mathbf{Q},t)$ observed in neutron scattering experiments is an average of $I_i(\mathbf{Q},t)$ over all directions so that $I_i(\mathbf{Q},t)$ depends only on the magnitude of Q. In the calculation of $I_i(\mathbf{Q},t)$, we make this average over directions by choosing Q as $\mathbf{Q} = Q\mathbf{x}$, $\mathbf{Q} = Q\mathbf{y}$ and $\mathbf{Q} = Q\mathbf{z}$, and taking an average over these three directions:

$$I_i(Q,t) = \frac{1}{3}[I_i(Q\mathbf{x},t) + I_i(Q\mathbf{y},t) + I_i(Q\mathbf{z},t)]. \tag{20}$$

where $I_i(Q\mathbf{x},t))$, $I_i(Q\mathbf{y},t))$ and $I_i(Q\mathbf{z},t))$ are calculated from Eq. (3). Fig. 10 shows that the $I_i(\mathbf{Q},t)$ calculated from Eq. (20) for all H averaged over three directions agrees well with the $I_i(\mathbf{Q},t)$ averaged over many directions as calculated by SASSENA. We believe these averages over directions agree well because the sum over all H includes H in many orientations.

The average of $I_{iG}(\mathbf{Q},t)$ over directions was calculated in the same way as $I_i(\mathbf{Q},t)$ so that a direct comparison of the two can be made. i.e. $I_{iG}(Q,t)$ was obtained as

$$I_{iG}(Q,t) = \frac{1}{3}[I_{iG}(Q\mathbf{x},t) + I_{iG}(Q\mathbf{y},t) + I_{iG}(Q\mathbf{z},t)]. \quad (21)$$

where $I_{iG}(Q\mathbf{x},t))$, $I_{iG}(Q\mathbf{y},t))$ and $I_{iG}(Q\mathbf{z},t))$ are calculated from Eq. (6).

Explicitly, Fig. 11 shows the $I_i(\mathbf{Q},t)$ calculated from Eq. (2) for all H for $\mathbf{Q} = Q\mathbf{x}$, $\mathbf{Q} = Q\mathbf{y}$ and $\mathbf{Q} = Q\mathbf{z}$, and the $I_i(\mathbf{Q},t)$ averaged over these three directions. The $I_i(\mathbf{Q},t)$ for each direction are the same within statisical error. This indicates that $I_i(\mathbf{Q},t)$ is independent of the direction selected for the vector $\mathbf{Q}$ within present statistical error.

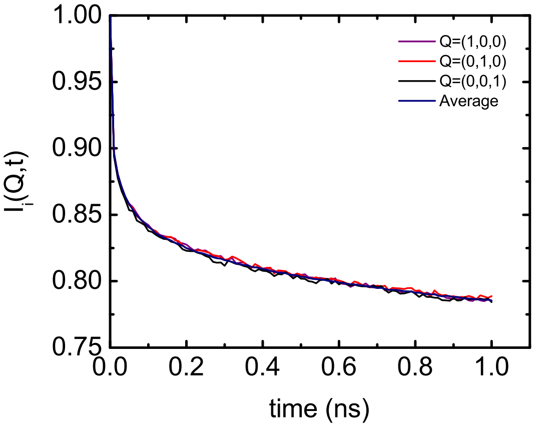

FIG. 11: Comparison of the ISF, $I_i(\mathbf{Q},t)$, arising from all H in lysozyme for $\mathbf{Q} = Q\mathbf{x}$ (purple solid line), $\mathbf{Q} = Q\mathbf{y}$ (red solid line), $\mathbf{Q} = Q\mathbf{z}$ (black solid line) and average over these there $\mathbf{Q}$ (blue solid lines), which are calculated from Eq. (2).

VIII. APPENDIX C : ON CALCULATING THE EISF

In this appendix we discuss calculation of the elastic component of the DSF, $S(Q,\omega=0) = \mathrm{A}I_i(\mathbf{Q},t=\infty)$, denoted the EISF. We propose a method of calculating the EISF which is more accurate than is often found in the literature.

The DSF is defined as

$$S(Q,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt \exp(i\omega t) I(Q,t) \quad (22)$$

with $I(Q,t)$ given by Eq. (2). We have dropped the subscript i. The elastic DSF is

$$S(Q,\omega=0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt I(Q,t) \quad (23)$$

which cannot be calculated directly from a MD simulation because we cannot extend the simulation out to long times. We can introduce, as in experiment, an instrument resolution function $R(\omega)$ and convolute $S(Q,\omega)$ with the resolution function. The convolution broadened $S_R(Q,\omega=0)$ written in time space is

$$S_R(Q,\omega=0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt I(Q,t) R(t) \quad (24)$$

where

$$R(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt \exp(i\omega t) R(\omega) \quad (25)$$

is the Fourier transform of $R(\omega)$. Essentially $R(t)$ serves to cut off the integration after a finite time. If $R(\omega) = \pi^{-1}\Gamma/(\omega^2+\Gamma^2)$ is a Lorentzian function, $R(t) = \exp(-\Gamma t)$ and the integral is cut off after a time $\tau_R \sim \Gamma^{-1}$.

To identify the EISF it is convenient to formally separate $I(Q,t)$ into a time independent and a time dependent part,

$$I(Q,t) = I(Q,\infty) + [I(Q,t) - I(Q,\infty)]. \quad (26)$$

Substituting Eq. (26) into Eq. (24), we obtain

$$\begin{aligned} S_R(Q,\omega=0) &= I(Q,\infty)\int_{-\infty}^{\infty} dt R(t) \\ &+ \int_{-\infty}^{\infty} dt[I(Q,t) - I(Q,\infty)]R(t) \\ &= R(\omega=0) I(Q,\infty) \\ &+ \int_{-\infty}^{\infty} dt[I(Q,t) - I(Q,\infty)]R(t) \quad (27) \end{aligned}$$

As done in experiment, and to obtain regular functions when we go to high resolution ($W = 2\Gamma \to 0$) where $R(\omega) = \delta(\omega)$, we normalize $S_R(Q,\omega=0)$ by $S_R(Q=0,\omega=0) \equiv S_R(0,\omega=0)$. Using $I(Q=0,t)=1$ at all times, we obtain, substituting $I(0,t)=1$ into Eq. (27),

$$S_R(0,\omega=0) = R(\omega=0) \quad (28)$$

and

$$\begin{aligned} \frac{S(Q,0)}{S(0,0)} &= I(Q,\infty) \\ &+ \int_{-\infty}^{\infty} dt[I(Q,t) - I(Q,\infty)]\frac{R(t)}{R(\omega=0)} \quad (29) \end{aligned}$$

In the limit of infinitely high resolution where $R(\omega) = \delta(\omega)$ the last term is zero and we obtain

$$\frac{S(Q,\omega=0)}{S(0,\omega=0)} = I(Q,\infty) = lim_{t\to\infty} I(Q,t) \quad (30)$$

which is our desired result relating $S(Q,\omega = 0)$ and $I(Q,\infty)$.

In the present paper we used $I(Q,\infty) = \exp(-\frac{1}{3}Q^2\langle r^2\rangle)$ which is valid (1) keeping Gaussian terms only in the cumulant expansion, (2) neglecting dynamical heterogeneity and (3) assuming cubic motional symmetry. We can obtain $I(Q,\infty)$ by fitting a model of the form Eq. (26) such as the model used in section IIB,

$$I(Q,t) = I_\infty + [1 - I_\infty]C(Q,t), \tag{31}$$

where $I_\infty \equiv I(Q,\infty)$, to calculated values of $I(Q,t)$ computed from a simulation. $I(Q,\infty)$ ($\langle r^2\rangle$) is obtained as a fitting parameter. We believe this represents $I(Q,\infty)$ within the approximations above and the limits of the simple model.

The EISF is often calculated as

$$S(Q,\omega = 0)/S(0,\omega = 0) \simeq I(Q,\tau) \tag{32}$$

where τ is a reasonably long time of order $1 - 10$ ns. We can estimate the error in this latter estimate of the EISF using the simple model (Eq. (31)) which gives

$$\frac{I(Q,\tau)}{I_\infty} = 1 + \frac{1 - I_\infty}{I_\infty}C(\tau) \tag{33}$$

and

$$I(Q,\tau) - I_\infty = (1 - I_\infty)C(\tau). \tag{34}$$

From the $C(Q,t)$ obtained from fits to the simulated $I(Q,t)$ in lysozyme (see Figs. 2 and 6 of Ref. [15]), we see that $C(Q,t) \simeq 0.25$ even for long times up to 10 ns, typically longer than τ used in Eq. (32). $C(Q,t)$ has not reached to zero as would be required for Eq. (32) to be accurate. A similar value of $C(Q,t)$ at $t = 10$ ns was obtained earlier by Calandrini and Kneller [41] using the Mittag-Leffler function. Employing $I_\infty = \exp(-\frac{1}{3}Q^2\langle r^2\rangle)$ to estimate I_∞ and selecting $\langle r^2\rangle = 1$ Å^2, we see that at $Q \to 0$, $I_\infty \to 1$ and there is no error in using $I(Q,\tau)$. However, at $Q = 3$ Å^{-1} where $I_\infty \simeq 0.05$, we see from Eqs. (33) and (34) that $I(Q,\tau)/I_\infty \simeq 5$ and $I(Q,\tau) - I(Q,\infty) = 0.25$. At $Q = 3$ Å^{-1}, $I(Q,\tau)$ is dominated by the error and does not represent $I(Q,\infty)$ at all. Thus, we believe it is important to use a fitting procedure or something similar to obtain $I(Q,\infty)$, rather than Eq. (32). This is particularly the case if the goal is to assess properties of the protein from the Q dependence of the EISF calculated from a simulation.

IX. ACKNOWLEDGEMENTS

It is a pleasure to acknowledge valuable discussions with Mark Johnson, Giuseppi Zaccai and Dominique Bicout. HRG acknowledges the hospitality of the Institut Laue Langevin where much of this paper was written. This work was supported by the DOE, Office of Basic Energy Sciences, under contract No ER46680 (DV and HRG) and by NSF grant number MCB-0842871 (LH and JCS).

[1] W. Doster, S. Cusack, and W. Petry, Nature (London) **337**, 754 (1989).

[2] W. Doster, S. Cusack, and W. Petry, Phys. Rev. Lett. **65**, 1080 (1990).

[3] R. M. Daniel, J. C. Smith, M. Ferrand, S. Héry, R. Dunn, and J. L. Finney, Biophys. Journ. **75**, 2504 (1998).

[4] M. Tarek and D. J. Tobias, Biophys. Journ. **79**, 3244 (2000).

[5] G. Zaccai, Science **288**, 1604 (2000).

[6] J. H. Roh, V. N. Novikov, R. B. Gregory, J. E. Curtis, Z. Chowdhuri, and A. P. Sokolov, Phys. Rev. Lett. **95**, 038101 (2005).

[7] J. H. Roh, J. E. Curtis, S. Azzam, V. N. Novikov, I. Peral, Z. Chowdhuri, R. B. Gregory, and A. P. Sokolov, Biophys. Journ. **91**, 2573 (2006).

[8] J. H. Roh, R. M. Briber, A. Damjanovic, D. Thirumalai, S. A. Woodson, and A. P. Sokolov, Biophys. Journ. **96**, 2755 (2009).

[9] M. Jasnin, L. van Eijck, M. M. Koza, J. Peters, C. Laguri, H. Lortat-Jacob, and G. Zaccai, Phys. Chem. Chem. Phys. **12**, 3360 (2010).

[10] H. Nakagawa, H. Kamikubo, and M. Kataoka, Biochim. Biophys. Acta **1804**, 27 (2010).

[11] V. G. Sakai and A. Arbe, Current Opinion in Colloid & Interface Science **14**, 381 (2009).

[12] S.-H. Chen, M. Lagi, X.-Q. Chu, Y. Zhang, C. Kim, A. Faraone, E. Fratini, and P. Baglioni, Spectroscopy **24**, 1 (2010).

[13] K. Wood, C. Caronna, P. Fouquet, W. Haussler, F. Natali, J. Ollivier, A. Orecchini, M. Plazanet, and G. Zaccai, Chem. Phys. **345**, 305 (2008).

[14] D. Vural and H. R. Glyde, Phys. Rev. E **86**, 011926 (2012).

[15] D. Vural, L. Hong, J. C. Smith, and H. R. Glyde, Phys. Rev. E **88**, 052706 (2013).

[16] U. Lehnert, V. Reat, M. Weik, G. Zaccai, and C. Pfister, Biophys. Journ. **75**, 1945 (1998).

[17] A. Paciaroni, S. Cinelli, and G. Onori, Biophys. Journ. **83**, 1157 (2002).

[18] R. M. Daniel, J. L. Finney, V. Reat, R. Dunn, M. Ferrand, and J. C. Smith, Biophys. Journ. **77**, 2184 (1999).

[19] V. Calandrini, V. Hamon, K. Hinsen, P. Calligari, M.-C. Bellissent-Funel, and G. R. Kneller, Chem. Phys. **345**, 289 (2008).

[20] T. Becker and J. C. Smith, Phys. Rev. E **67**, 021904 (2003).

[21] J. A. Hayward and J. C. Smith, Biophys. Journ. **82**, 1216 (2002).

[22] D. J. Bicout, Phys. Rev. E **62**, 261 (2000).

[23] D. J. Bicout and G. Zaccai, Biophys. Journ. **80**, 1115 (2001).

[24] Z. Yi, Y. Miao, J. Baudry, N. Jain, and J. C. Smith, J. Phys. Chem. B **116**, 5028 (2012).

[25] A. Tokuhisa, Y. Joti, H. Nakagawa, A. Kitao, and M. Kataoka, Phys. Rev. E **75**, 041912 (2007).

[26] G. R. Kneller and G. Chevrot, J. Chem. Phys. **137**, 225101 (2012).

[27] L. Meinhold, D. Clement, M. Tehei, R. Daniel, J. L. Finney, and J. C. Smith, Biophys. Journ. **94**, 4812 (2008).

[28] G. R. Kneller and K. Hinsen, J. Chem. Phys. **131**, 045104 (2009).

[29] J. Peters and G. R. Kneller, J. Chem. Phys. **139**, 165102 (2013).

[30] P. J. Artymiuk, C. C. F. Blake, D. W. Rice, and K. S. Wilson, Acta. Cryst. **B38**, 778 (1982).

[31] B. Hess, C. Kutzner, D. V. Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).

[32] W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).

[33] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon, J. Chem. Phys. **120**, 9665 (2004).

[34] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).

[35] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comp. Chem. **18**, 1463 (1997).

[36] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[37] M. Parrinello and A. Rahman, J. App. Phys. **52**, 7182 (1981).

[38] M. Lagi, P. Baglioni, and S. H. Chen, Phys. Rev. Lett. **103**, 108102 (2009).

[39] L. Hong, N. Smolin, B. Lindner, A. P. Sokolov, and J. C. Smith, Phys. Rev. Lett. **107**, 148102 (2011).

[40] H. Nakagawa, A. Tokuhisa, H. Kamikubo, Y. Joti, A. Kitao, and M. Kataoka, **442**, 356 (2006).

[41] V. Calandrini and G. R. Kneller, J. Chem. Phys. **128**, 065102 (2008).