JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

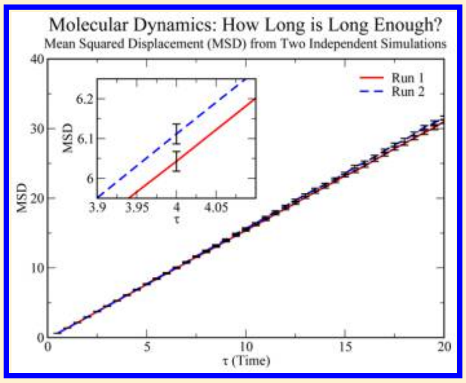
Estimating Error in Diffusion Coefficients Derived from Molecular Dynamics Simulations

Gaurav Pranami and Monica H. Lamm*

Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011, United States

S Supporting Information

ABSTRACT: The computationally expensive nature of molecular dynamics simulation limits the access to length (nanometer) and time scales (nanosecond) that are orders of magnitude smaller than the experiment it models. This limitation warrants a careful estimation of statistical uncertainty associated with the properties calculated from these simulations. The assumption that a simulation is long enough so that the ergodic hypothesis applies is often invoked in the literature for the computation of properties of interest from a single molecular dynamics simulation. Here, we demonstrate that making this assumption without validation results in poor estimates of the self-diffusion coefficient from a single molecular dynamics simulation of Lennard-Jones fluid. This problem is shown to be even more severe when the diffusion coefficient of macromolecules is calculated from a single molecular dynamics simulation. We have shown that conducting multiple independent simulations is necessary to obtain reliable estimates of diffusion coefficients and their associated statistical uncertainties. We show that even a “routine” calculation of the self-diffusion coefficient for a Lennard-Jones fluid, as determined from a linear fit of the mean squared displacement of particles as a function of time, violates the key assumptions of linear regression. A rigorous approach for addressing these issues is presented.



■ INTRODUCTION

Molecular dynamics (MD) simulation is a powerful tool for studying microscopic phenomena in complex systems that are difficult to probe experimentally or are inaccessible to theoretical treatment. In addition to gaining molecular insight into the structural and thermodynamic properties of systems of interest, MD simulations are often used for studying transport properties due to its accurate treatment of the dynamics. However, the computationally intensive nature of MD simulations limits systems sizes to, at most, a few million atoms, thus limiting accessible length and time scales to a few nanometers and nanoseconds, respectively. This limitation is even more severe in first-principles MD simulations.[1] The macroscopic properties estimated from such small samples are likely to have significant statistical uncertainty, which should be estimated for proper interpretation of the results. Limited length and time scales accessible to MD simulations are especially apparent when calculating properties that depend on observables with long-range or long-lived correlations, like the local currents of conserved quantities.[2] Additionally, these calculations are further complicated by that fact that the time required to forget the initial conditions prior to data sampling in a simulation is unknown and is difficult to estimate, especially in glassy systems.[3]

In MD simulations, the desired properties are typically evaluated as time averages over the system’s trajectory. This is based on the ergodic hypothesis, which asserts that when simulations are sufficiently long, the time average is to equal the ensemble average. However, how long a simulation needs to be for the ergodic hypothesis to be valid is unknown *a priori*. In this work, we have demonstrated the importance of addressing this issue for obtaining reliable estimates of desired properties and uncertainty associated with them.

In addition, the choice of time interval between two successive samples determines if they would be correlated to each other or not. Ensuring that successive samples are correlated is necessary when the interest is in evaluating correlation functions, for example a velocity autocorrelation function. In such cases, Zwanzig and Ailawadi showed that the error in the time correlation functions is estimated to be inversely proportional to the square root of the simulation time.[4] Statistical uncertainty in the correlated time series data has also been studied by Chodera et al.[5] However, further statistical analysis would be required when the properties of interest are derived from correlation functions, such as diffusion coefficient and viscosity. On the other hand, uncorrelated (or independent) sampling is required when calculating thermodynamic properties and the diffusion coefficient from mean squared displacement (MSD). A key point to appreciate in either case is that the minimum gap between successive samples that would ensure independent sampling is not known *a priori* and is dependent on the system being examined. The statistical error in static properties that are likely to be correlated can be

Received: June 18, 2015
Published: August 25, 2015

DOI: 10.1021/acs.jctc.5b00574

estimated using the method of block averages[6] or the renormalization group method proposed by Flyvbjerg and Petersen.[7]

The objective of this work is to highlight that the estimates of the properties calculated from MD simulations even for the simplest systems can be erroneous as typically reported in the literature. This is demonstrated via the example of the calculation of diffusion coefficient (D) of Lennard-Jones (LJ) particles comprising a LJ fluid from their mean squared displacement (MSD). We show that even for the system of a LJ fluid, MD trajectories resulting in statistically distinguishable MSD can be obtained by simply varying the initial velocities assigned to the particles. Running a sufficiently long simulation could mitigate this issue, but the length of time required is unknown *a priori*. In addition, after running a long simulation, subsequent validation is difficult for proving that it was indeed long enough. We have shown that this issue can be managed by running multiple independent simulations for measuring diffusion coefficient and the statistical uncertainty associated with it. A systematic approach for efficient independent sampling of squared displacements is proposed. It is then demonstrated that the linear regression to fit a straight line through MSD as a function of time violates the key assumptions of normal distribution and homoscedasticity, which prevents the estimation of uncertainty in the calculated diffusion coefficient. The issues arising from nonconstant variance of squared displacements have been discussed in the scientific literature but have not been broadly appreciated in the molecular simulation community.[8] Again, using the data obtained from multiple independent simulations (MIS) is proposed as a way to remedy this issue. Further complications arising from the effect of finite system size on diffusion have been highlighted. Addressing all these issues is even more important when using MD simulations for studying macromolecular systems. We emphasize this by using the diffusion of a rigid fractal aggregate at infinite dilution in a LJ fluid as an example.

The findings from this work should be applied after careful consideration of the system being studied and time and length scales involved. The problem of estimating the diffusion coefficients is nontrivial for systems with slowly evolving conformational degrees of freedom, and research to address these issues is being reported in the literature.[9,10] In fact, it may not even be appropriate to represent the dynamics with a single estimate of the diffusion depending on the system under study. For example, Thompson et al. on tracking single mRNA particles evolving in yeast cells found that its diffusion coefficient was not normally distributed, which was due to three different underlying modes of diffusion—random, confined, and directed.[11] It is beyond the scope of this manuscript to address diffusion in different types of physical systems. Instead, the goal here is to alert the practitioners of MD simulations toward the potential sources of error in data collection and subsequent analysis.

■ COMPUTATIONAL DETAILS AND THEORY

Simulation Details. All the MD simulations were performed using the LAMMPS molecular dynamics program available from Sandia National Laboratories.[12] The interaction between particles was modeled as LJ potential:

$$U(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

where $U(r)$ is the energy between two particles at a separation r, σ is the diameter or size of the particle, and ε is the well depth which characterizes the attraction between the particles. A system of units as shown in Table 1 was used in these simulations. The size (σ) and mass (m) of the particles were set to unity. The number density of particles (ρ_s) and temperature (T) were respectively set to 0.7 and 2.75 to ensure that LJ fluid was a single phase.[6] A cutoff of 2.5σ was applied. Neither was the LJ potential shifted to achieve zero potential at cutoff nor was the long-range correction applied.[13] These modifications do not affect the dynamics in the simulations conducted in this work. All the simulations were carried out in the canonical ensemble (NVT) using a Nosé–Hoover thermostat (with a time constant of 200 time steps) to ensure an identical temperature in all simulations.[6] A typical system was equilibrated for 100 000 time steps (for achieving equilibrium thermodynamic properties), followed by a production run of 1 million time steps. The trajectory of the system is recorded every 200 time steps. A time step of 0.005 dimensionless time units was used to ensure numerical stability. The correctness of the simulation approach followed in this work was established by comparison to results published in the literature. The models developed by Nuevo et al. predict a diffusion coefficient in the range 0.255–0.284 for a system of 256 LJ particles at a temperature of 2.75 and number density of 0.7.[14] This compared well with our results listed in Table 2.

Table 1. Units Used in MD Simulations

dimension	unit
length	σ
energy	ε
mass	m
time	$\sigma\sqrt{(m/\varepsilon)}$
temperature	ε/k_b

Table 2. Average Diffusion Coefficient ($\bar{D}$) Estimated from 100 MIS of N_p LJ Particles Placed in a Cubic Simulation Box of Length L[a]

N_p	L	$\bar{D}$	LCL	UCL
125	5.63	0.26168	0.26156	0.26181
216	6.76	0.26984	0.26973	0.26995
512	9.01	0.27828	0.27820	0.27836
1000	11.26	0.28227	0.28221	0.28233

[a] LCL and UCL represent lower and upper bounds of 95% confidence limit.

Statistical Uncertainty. A brief introduction to statistical uncertainty is warranted in the context of this manuscript even though any standard text on statistics would include this discussion. It is usually not possible to sample the population of a random variable (X). Therefore, the parameters describing this population, such as population mean μ and variance s^2, remain unknown. In order to characterize the population, μ and s^2 are estimated by $\bar{X}$ (sample mean) and S^2 (sample variance) respectively from a sample of N independent measurements of the random variable X:[15]

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

The uncertainty associated with the estimate of population mean ($\bar{X}$) based on a one-sided Student's t-distribution is given by a $100(1-\alpha)$ confidence interval as follows:

$$\bar{X} \pm t_{N-1,1-\alpha/2}\frac{S}{\sqrt{N}}$$

where $0 < \alpha < 1$. This approach is valid only if X is normally distributed or, as a consequence of the central limit theorem, *if the sample is sufficiently large*. In this case, the distribution of the sample mean approaches a normal distribution (with mean μ and variance s^2/N) as the sample size becomes larger regardless of the distribution of X. This is a very important consideration in reference to the calculation of MSD because squared displacement (SD) over a given time interval is not normally distributed, as discussed later in the manuscript.

There is a clear distinction between estimated standard deviation (S) and confidence interval. The error associated with $\bar{X}$ is often reported as $\bar{X} \pm S$ in the literature. However, this only characterizes the spread around the estimated mean instead of characterizing the uncertainty associated with it. The standard practice, however, is to report the 95% confidence interval ($\alpha = 0.05$) for the estimated mean. A 95% confidence interval implies that 95% of such intervals are likely to contain population mean μ. Often in the literature, $\bar{X} \pm S/\sqrt{N}$ is reported as error, but this is only a 68.2% confidence interval.

In the context of MD simulations, this treatment is applicable only if data sampled are independent. In other words, it is critical to know the time interval needed (not known *a priori*) between successive samples to ensure their independence. This is discussed in detail below.

Diffusion Coefficient. The diffusion coefficient (D) of a particle undergoing random walk (self-diffusion of LJ fluid and a rigid fractal aggregate diffusion in LJ fluid, as discussed later) is given by Einstein's relation[13]

$$D = \frac{1}{2d}\lim_{\tau\to\infty}\frac{d}{d\tau}\langle[\overrightarrow{r(\tau)} - \overrightarrow{r(0)}]^2\rangle$$

where D is the diffusion coefficient, d (= 3) is the dimension of the system, $\overrightarrow{r(\tau)}$ is the particle position at time τ, $\tau = 0$ refers to a time origin, and the angle brackets $\langle\ \rangle$ denote the average over time origins. The quantity $\left\langle\left[\overrightarrow{r(\tau)} - \overrightarrow{r(0)}\right]^2\right\rangle$ is the particle mean squared displacement (MSD) and it grows linearly with time for sufficiently large values of τ. Therefore, in a three-dimensional system, D equals one-sixth of the slope of a plot of the particle MSD versus τ.

A schematic illustrating the calculation of MSD of particles from their trajectory obtained from MD simulation is shown in Figure 1. If the displacement of N_P particles over a time interval of width τ is calculated such that the distance between the time origins of two consecutive time intervals is δ, then MSD(τ,δ) is given by

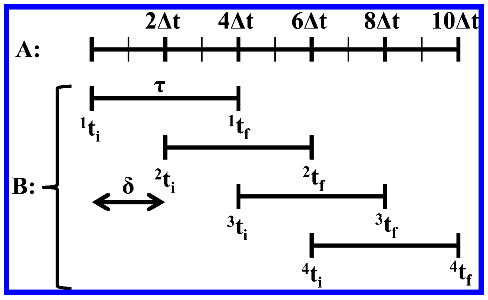


Figure 1. Schematic describing the calculation of mean squared displacement MSD(τ) over a given time interval (τ). (A) The MD trajectory where the particle coordinates are recorded at a time interval of Δt, as indicated by vertical tick marks. (B) The time intervals ${}^jt_f - {}^jt_i$ ($= \tau$) over which the displacement of particles is calculated. Mean squared displacement is obtained by taking the mean of the square of these displacements. The separation, δ, between the time origins of two consecutive intervals is given by ${}^{j+1}t_i - {}^jt_i$.

$$\mathrm{MSD}(\tau,\delta) = \frac{1}{N_\tau}\sum_{j=0}^{N_\tau}\left(\frac{1}{N_P}\sum_{i=1}^{N_P}|\overrightarrow{r_i(j\delta+\tau)} - \overrightarrow{r_i(j\delta)}|^2\right)$$

where N_τ is the number of time intervals of width τ over which the average of squared displacements (SD) is calculated, i and j indices respectively run the summation over N_P and N_τ, and $\overrightarrow{r_i(t)}$ is the position of the particle i at time t. The distance between two consecutive time origins (δ) is usually kept to a minimum to maximize the number of samples obtained from a MD simulation because of their computationally intensive nature. However, for a small δ and large τ, the consecutive values of squared displacements are likely to be correlated, thus resulting in a biased estimate of MSD(τ). This would be true for any other property of interest as well. In the context of balancing the computational cost with the requirement of uncorrelated (independent) or correlated (if required) sampling, it is critical to appreciate that the choice of an appropriate δ is unknown *a priori*, and this must be addressed for each problem being studied. The following section will address this issue for the determination of MSD.

■ RESULTS AND DISCUSSION

Sampling Squared Displacements. Independent sampling of squared displacement (SD) is required for the determination of the diffusion coefficient of randomly diffusing particles in LJ fluid from their mean squared displacement. As mentioned previously, an appropriate value of δ has to be determined for ensuring the independent sampling of SD. In order to understand the time scale at which SD is correlated, a normalized autocorrelation function of SD was defined as a function of δ and τ given by the following equation.

$$R_{\mathrm{SD}}(\tau,\delta) = \frac{\langle \mathrm{SD}(\tau,0)\,\mathrm{SD}(\tau,\delta)\rangle}{\langle \mathrm{SD}(\tau,0)\,\mathrm{SD}(\tau,0)\rangle}$$

A plot of R_{SD} as a function of δ for different values of τ is shown in Figure 2, and it shows that R_{SD} for a given value of τ decays (in other words it decorrelates) to a constant value as $\delta \to \tau$. This important result indicated that SD(τ) should be sampled such that $\delta \geq \tau$ (i.e., nonoverlapping intervals) in order to ensure independent sampling of SD in the system studied here. Using this data, MSD(τ) and associated confidence intervals were obtained as discussed earlier. Note that $R_{\mathrm{SD}}(\tau,\delta)$ does not decay to zero because $\langle \mathrm{SD}(\tau,\delta \to \infty)\rangle > 0$. This approach

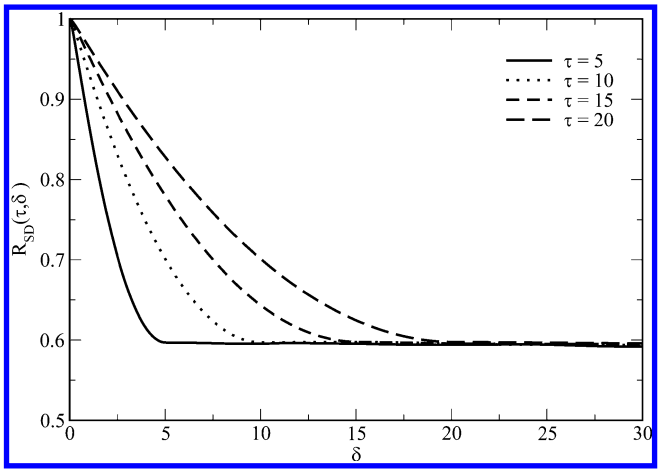


Figure 2. Normalized autocorrelation function of squared displacement samples, $R_{SD}(\tau,\delta)$ as a function of δ (time units) for different values of τ (time units). Data obtained from a MD simulation of 125 particles (N_P) comprising a LJ fluid under the conditions described earlier. The plot shows that autocorrelation function decays to a constant value as $\delta \to \tau$. Therefore, independent sampling of squared displacements would be ensured by choosing $\delta \geq \tau$.

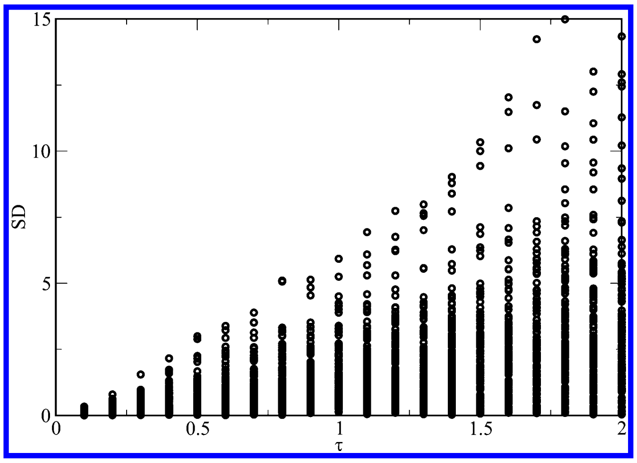


Figure 3. Plot of the squared displacements (SD) as a function of τ. Data was obtained from a MD simulation of 125 particles (N_P) comprising a LJ fluid under the conditions described earlier.

should be followed to determine the δ for independent or correlated sampling as required for calculating the properties of interest in a specific system or a problem under investigation. We would like to point out that the need for an objective approach that ensures independent sampling has been acknowledged in the literature; Calderon et al. utilized the maximum likelihood technique for determining δ needed to allow short time scale non-Markovian artifacts to average out by identifying a suitable stochastic model that described the diffusive process in their system.[10]

Linear Regression for Obtaining Diffusion Coefficient. As mentioned earlier, diffusion coefficient (D) is estimated as one-sixth of the slope of the linear fit of MSD as a function of τ. The best fitting straight line can be obtained from least-squares linear regression, which would also yield the statistical uncertainty associated with the fitting parameters under the assumptions of regression analysis. The least squared straight line fit is $\text{MSD}(\tau) = \widehat{\text{SD}(\tau)} = \widehat{\beta_0} + \widehat{\beta_1}\tau$, where $\widehat{\text{SD}(\tau)}$ is the point estimate of the mean of squared displacement, and $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the point estimates of the intercept and slope of the linear fit. For the details of the calculation of these parameters and associated uncertainty, the reader is referred to Kleinbaum et al.[15] Subsequently, the diffusion coefficient is $D = \widehat{\beta_1}/6$. An important message here is that this analysis is valid only if underlying assumptions are satisfied.

A careful evaluation of the underlying assumptions is required for the assessment of the validity of least squared linear regression for the estimation of diffusion coefficient from mean squared displacement. The assumption of the *existence* of SD(τ) with certain probability distribution having finite mean and variance is valid in this case because SD varies within a finite range for a given value of τ as shown in Figure 3. The assumption of a *linear relationship* between MSD and τ is upheld based on Einstein's relation for determining the diffusion coefficient. We have ensured the validity of the assumption of *independence* of SD samples by determining δ that ensured independent sampling. The assumption of *normal distribution* of SD is violated in this case because squared displacements are not normally distributed as verified in the SI, section 1. It is well-known that the magnitude of displacement in more than one dimension does not follow a normal distribution. This violation invalidates least-squares regression analysis in this case. For the sake of completeness, the distribution of SD also violates the assumption of *homoscedasticity* (i.e., constant variance in SD as a function of τ) as evident from the increasing spread of SD as a function of τ shown in Figure 3.

Yet, the applicability of the Einstein relation for the diffusion coefficient to the system of LJ particles studied here has a physical basis. Therefore, the straight line fit through mean squared displacements would result in a point estimate of $D = \widehat{\beta_1}/6$ even though uncertainty in D cannot be determined. An estimate of mean of D and its associated uncertainty based on a Student's t-distribution can be readily obtained from a sample of such point estimates obtained from multiple independent simulations (MIS). Evidence that D as a random variable followed normal distribution is provided in the SI, section 2. The average diffusion coefficient and associated uncertainty thus obtained from a 100 MIS for the systems of LJ particles studied here have been documented in Table 2. These simulations were identical except for the initial velocities assigned to the particles, which were sampled from a Maxwell–Boltzmann distribution at the same temperature. While we have reported the analysis for LJ fluid simulated in a canonical ensemble, we note that the difference between the average diffusion coefficients obtained in canonical and microcanonical ensembles was significant at a confidence level of 95% for the system of 125 LJ particles. Therefore, the choice of ensemble impacts the diffusion coefficient and should be considered carefully. The ideas presented in this work apply to both ensembles, as has been demonstrated through the examples of LJ fluid in canonical ensemble, and rigid fractal aggregates in microcanonical ensembles (discussed later).

The importance of conducting MIS is further emphasized by the observation that even for one of the simplest systems, LJ fluid, two independent MD simulations could give mean squared displacements statistically different from each other as shown in the following section. In other words, conducting MIS allows for more extensive exploration of the phase space, which helps in preventing erroneous conclusions that one could arrive at from a single simulation.

Need for Conducting MIS. The ergodic hypothesis states that the time average over a sufficiently long period of time and ensemble average of an observable in a system are identical. Properties of systems such as LJ fluid are usually estimated as time averages from a single MD simulation, which is assumed to be long enough for ergodic hypothesis to be applicable. However, it is not known *a priori* how long the time period must be to satisfy ergodicity.

As mentioned in the previous section, in this set of MIS that were conducted with different initial conditions (while keeping everything else the same) for the system containing N_p = 125 LJ particles, we found some MD trajectories which resulted in statistically distinguishable MSD(τ). This was surprising because these simulations were significantly longer than the simulations reported in standard texts on molecular simulations.[6,13]

The nonoverlapping 95% confidence intervals (shown as error bars) in Figure 4 indicate that MSD(τ) obtained from two such simulations were statistically different from each other. This showed that the simulation time was not long enough for the system to erase its memory of its initial state. This finding emphatically demonstrates that it is important to conduct MIS even for a simple system like LJ fluid. Estimates of desired properties should include averaging across multiple independent simulations in addition to time averaging to obtain averages that are not biased by the initial configuration.

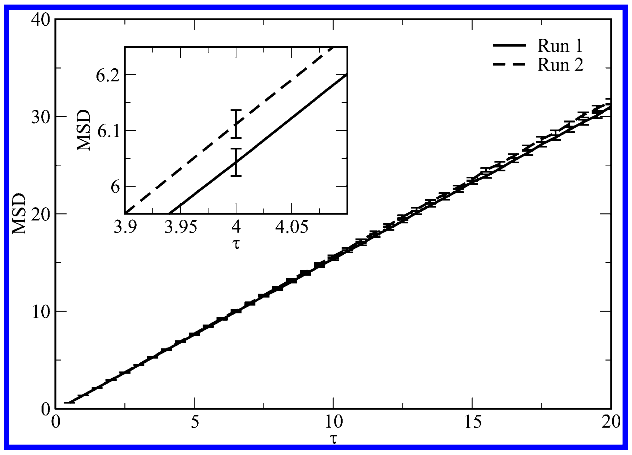


Figure 4. Plot of MSD vs τ obtained from two independent MD simulations of an identical system of 125 LJ particles. The error bars indicate 95% confidence intervals. The mean squared displacements from these two runs are statistically different as indicated by nonoverlapping 95% confidence intervals. The inset shows a zoomed-in plot to highlight a representative nonoverlapping confidence interval, which is not visible in the main figure for lower values of τ.

Here, we note that even though SD did not follow a normal distribution, the confidence interval for MSD(τ) was determined from Student's t-distribution by invoking the central limit theorem, which states that the mean of sufficiently large samples is approximately normally distributed regardless of the underlying distribution of the random variable. In this analysis, the number of SD samples was greater than 30 000 for a given value of τ, which is sufficiently large because statisticians prescribe a sample of size ~40 as a rule of thumb for the applicability of central limit theorem. To verify this, we have shown that MSD(τ) obtained from MIS indeed followed normal distribution (SI, section 2).

The effect of the initial state could potentially be mitigated by running a very long MD simulation, but such long simulations suffer from drift at long times due to finite numerical precision leading to altered dynamics in addition to high computational cost. Therefore, we recommend conducting several computationally manageable multiple independent simulations to account for initial state effects.

For more complex systems, the dependence of properties calculated from MD simulations on initial conditions has also been demonstrated in the literature. For, e.g., Calderon et al. used a mixed effects model and showed that diffusion of a charged atom diffusing in an ion channel depended upon initial conditions.[9] While the approach of generating initial states by assigning velocities to particles randomly is routinely followed in the literature, it is important to recognize that the generation of independent initial states is nontrivial for complex systems. This has been addressed in the literature to a degree of rigor through the use of techniques such as the nudged elastic band technique to generate minimum energy states.[10,16]

Diffusion as a Function of Simulation Box Length. Conducting MIS was also useful in the analysis of the effect of finite simulation box size as documented below. The motion of particles comprising a fluid induces a flow field, which is felt by other particles, resulting in an effective long-range hydrodynamic interaction. Therefore, the dynamic properties of a finite system modeled with periodic boundary conditions are affected by these long-ranged interactions between the system and its periodic images. For determining the diffusion coefficient from MD simulations, others have proposed a correction to account for the effects of a finite simulation box size L[17−19]

$$D_o = D + \frac{\xi k_b T}{6\pi\eta L}$$

where D_o is the diffusion in an infinitely large system, ξ = 2.837297 is a numerical constant, k_b is the Boltzmann constant, T is the temperature, and η is the viscosity of the solvent. Therefore, the diffusion coefficient corrected for finite size effects (D_o) is given by the intercept of a straight line fit of D vs L^{-1} data, obtained from MD simulations of systems with different box sizes.

A plot of D as a function of L^{-1} for LJ fluid particles is shown in Figure 5. One hundred multiple independent MD simulations were carried out for each L by varying the initial velocities of the LJ particles, which were sampled from Maxwell−Boltzmann distribution. Notice that if only one simulation was conducted for each L, then the linear relationship between D and L^{-1} could potentially be masked or incorrect, which would then result in an inaccurate estimate of D_o. Hypothetical curves, like l_1, l_2, and l_3 as shown in Figure 6, illustrating this point are not plotted in Figure 5 for clarity. This also underlines the importance of conducting MIS. Determining average D_o (i.e., $\overline{D_o}$) is a simple matter obtaining the intercept of a straight line fitted through D vs L^{-1} data using least squared linear regression. As mentioned earlier, all the assumptions involved in the least-squares linear regression have to be validated. The assumptions of *existence* (Figure 5), *linear relationship* (mentioned above), *independence* (MIS), and *normal distribution* of D (SI, section 3) were satisfied in this case. The assumption of *homoscedasticity* (constant variance in

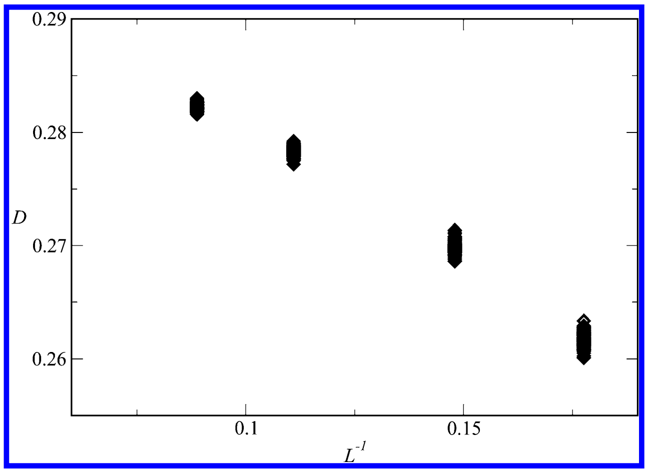


Figure 5. Variation of the diffusion coefficient (D) of LJ fluid particles as a function of the inverse of the simulation box length (L^{-1}). For each box length, 100 independent MD simulations were conducted.

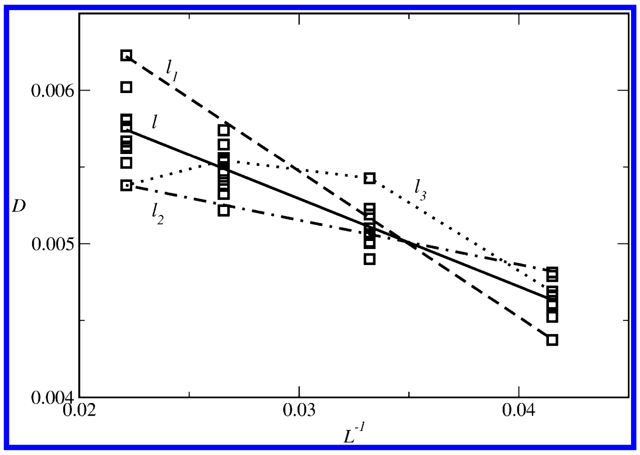


Figure 6. Variation of the diffusion coefficient (D) of a fractal aggregate containing 64 primary particles and fractal dimension (d_f) of 2.5 as a function of L^{-1}. The linear relationship between D and L^{-1} obtained from weighted least-squares regression is shown as the solid black line (l). However, if only one simulation was conducted for each box size, then D could potentially lie on one of the l_1, l_2, or l_3 curves, thus deviating significantly from l.

D for different L^{-1}) was violated here as evident from the increasing spread in D as L^{-1} increased as shown in Figure 5. The issue of heteroscedasticity was addressed by using a weighted least-squares approach. The reader is referred to SI, section 3 for least squared and weighted least-squares regression analysis of this system and other details. The average diffusion coefficient of LJ particles corrected for finite simulation box size, $\overline{D_0}$, thus obtained was 0.30267, and the 95% confidence interval was (0.30240, 0.30294). An R^2 value of 0.99 was obtained for this linear fit.

Conducting MIS for addressing the effect of finite simulation size is even more important when calculating the diffusion coefficient of macromolecules suspended in a fluid. Large size separation between the macromolecular solute particle and the solvent particle necessitates the presence of a large number of solvent particles for each solute particle, thus making MD simulation of systems containing more than a few macromolecules computationally prohibitive. This is different from the case of LJ fluid where extensive sampling of squared displacements can be obtained from a system of modest size containing 125 particles because each identical particle contributes to data sampling. This point is illustrated through MD simulations of a nanoparticle aggregate undergoing Brownian diffusion in the presence of explicit solvent particles.

An off-lattice fractal aggregate of a fractal dimension $d_f = 2.5$ containing $N = 64$ primary particles (p) was generated using the recipe proposed by Thouy and Jullien.[20,21] The aggregate was treated as a rigid body and was placed in a cubic simulation box containing explicit solvent particles (s) at a number density of $n = 0.85$. Solvent−solvent and solvent−particle interactions were modeled with LJ potential with $\sigma = 1$ and $\varepsilon = 1$. The equilibrium temperature of $T = 1.2$ was achieved by running the MD simulations for half a million time steps in the canonical (NVT) ensemble, after which the thermostat was switched off, and the production run was carried out in a microcanonical ensemble (NVE) for 10 million time steps. The MD trajectory was advanced with a time step of 0.005 in reduced units. The diffusion coefficient of the fractal aggregate was calculated from the mean squared displacement of its center of mass. The correctness of these simulations was established by a good agreement of the ratio of hydrodynamic radius of the fractal aggregate to its radius of gyration with the experimental measurements reported in the literature for aggregates with similar morphology, and the reader is referred to Pranami et al. for details.[22] We also computed the average rotational and translational kinetic energies of the fractal aggregate (N_s = 11868, and L = 24.08) to be 1.81 ± 0.03 and 1.79 ± 0.03, respectively, in a representative simulation. Here, ± represents a 95% confidence interval, and the difference between the average rotational and translational kinetic energies was statistically insignificant. This implied that the system was equilibrated sufficiently ensuring the equipartition of the energy of the rigid fractal aggregate.

The variation of the diffusion coefficient of the fractal aggregate as a function of L^{-1} obtained from ten multiple independent simulations for each L is shown in Figure 6. This data clearly illustrates the effect of finite size of MD simulation box on the diffusion of fractal aggregate. It also reinforces the importance of conducting multiple independent simulations. If only one simulation was conducted for each of the box sizes in order to determine $\overline{D_0}$ (average diffusion coefficient corrected for finite simulation size), then one could potentially obtain the hypothetical curves l_1, l_2, or l_3 describing the relationship between D and L^{-1} as shown in Figure 6. This could result in an incorrect linear relationship (l_1 and l_2) or masking the linear relationship all together (l_3) resulting in incorrect estimates of slope and intercept ($\overline{D_0}$). Therefore, we found it necessary to conduct multiple independent simulations for each L in order to extract meaningful averages.

Assumptions for weighted least-squares regression analysis for fitting a straight line through D vs L^{-1} were satisfied as shown in SI, section 4. $\overline{D_0}$ thus obtained was 0.00701, and the 95% confidence interval was (0.00677, 0.00724). An R^2 value of 0.88 was obtained for this linear fit. The average diffusion coefficient at a given simulation box length and 95% confidence interval associated with it are listed in Table 3.

■ CONCLUSIONS

Several important issues associated with MD simulations are often not addressed as rigorously as they ought to be due to the computationally demanding nature of these simulations. The answers to questions such as how long is long enough to erase

Table 3. Average Diffusion Coefficient ($\bar{D}$) Obtained from 10 MIS of an Aggregate of Fractal Dimension, d_f = 2.5, and Containing N_p = 64 Primary Particles Placed in a Cubic Simulation Box of Length L[a]

L	N_s	$\bar{D}$	LCL	UCL
24.08	11868	0.00463	0.00454	0.00472
30.10	23180	0.00512	0.00501	0.00523
37.63	45274	0.00548	0.00537	0.00559
45.15	78233	0.00575	0.00557	0.00591

[a] N_s = number of solvent particles. LCL and UCL represent lower and upper bounds of 95% confidence limit.

the effects of initial conditions and how frequently should the data be recorded to ensure correlated or independent sampling as desired are not known *a priori*. Other questions such as how are the variables of interest distributed and if the assumptions underlying the estimation of properties of interest from the MD data are satisfied are usually not even considered. However, these questions should be addressed in order to reach reliable conclusions from MD simulations.

In this work, we have probed these issues through the specific example of the estimation of diffusion coefficient of LJ fluid and a rigid macromolecule suspended in LJ fluid from Einstein's relation using MD simulations. We demonstrated how to determine the smallest interval between two successive time origins along the MD trajectory for sampling independent squared displacements, and this approach could also be used for the estimation of other properties of interest. Subsequently, we showed that the assumptions of normal distribution and homoscedasticity required for the validity of least squared linear regression for fitting a straight line through squared displacements as a function of time were violated. This has not been shown in the literature to the best of our knowledge. To our surprise, we found that MD simulations of one of the simplest systems of LJ fluid resulted in trajectories that gave mean squared displacements, which were statistically distinguishable from each other. The duration of these simulations was significantly longer than the ones reported in the literature for similar systems. This alludes to the question posed earlier—how long is long enough for the validity of an ergodic hypothesis, especially in the context of the computationally intensive nature of MD simulations? We showed that these issues could be managed by running multiple independent simulations of computationally manageable duration. Finally, we also underscored the importance of running multiple independent simulations through the exercise of correcting the diffusion coefficient for the finite simulation size.

The approach reported in this work for measuring uncertainty in MD simulations could potentially be relevant to systems studied using other simulation techniques like Brownian dynamics and dissipative particle dynamics.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00574.

> Details of statistical analysis (performed using JMP 12.0.0, www.jmp.com) and tests that form the foundation of the results and conclusions presented in this manuscript. SI section 1 documents the goodness of fit test performed to show that SD did not follow a normal distribution. SI section 2 contains the goodness of fit tests that ascertained that diffusion coefficients obtained from MIS were indeed normally distributed. SI section 3 contains the details of linear least squared regression analysis for fitting D vs $1/L$ for LJ fluid. It established the violation of the assumption of homoscedasticity, which was subsequently remedied using weighted least-squares regression in the same section. SI section 4 documents the weighted least squared regression for fitting D vs $1/L$ for a fractal aggregate. SI section 5 shows that center of mass of a system of LJ particles did not exhibit a drift during the course of MD simulation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mhlamm@iastate.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation (NIRT-0403864). We wish to thank R. Skepnek, R. M. Fernandes, D. K. Rollins, and R. D. Vigil at Iowa State University for useful discussions.

■ REFERENCES

(1) Kuo, I. F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I. *J. Chem. Theory Comput.* **2006**, *2*, 1274−1281.

(2) Belitz, D.; Kirkpatrick, T. R.; Vojta, T. *Rev. Mod. Phys.* **2005**, *77*, 579−632.

(3) Eliazar, I.; Klafter, J. *Phys. Rev.* **2009**, *79*, 021115−1−021115−15.

(4) Zwanzig, R.; Ailawadi, N. K. *Phys. Rev.* **1969**, *182*, 280−283.

(5) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. *J. Chem. Theory Comput.* **2007**, *3*, 26−41.

(6) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed.; Academic Press: Waltham, MA, 2002; Vol. *1*.

(7) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461−466.

(8) Qian, H.; Sheetz, M. P.; Elson, E. L. *Biophys. J.* **1991**, *60*, 910−921.

(9) Calderon, C. P. *J. Chem. Theory Comput.* **2011**, *7*, 280−290.

(10) Calderon, C. P.; Arora, K. *J. Chem. Theory Comput.* **2009**, *5*, 47−58.

(11) Thompson, M. A.; Casolari, J. M.; Badieirostami, M.; Brown, P. O.; Moerner, W. E. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 17864−17871.

(12) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1−19.

(13) Haile, J. M. *Molecular Dynamics Simulation: Elementary Methods*, 1st ed.; Wiley-Interscience: Hoboken, NJ, 1997.

(14) Nuevo, M. J.; Morales, J. J.; Heyes, D. M. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1997**, *55*, 4217−4224.

(15) Kleinbaum, D.; Kupper, L.; Nizam, A.; Rosenberg, E. *Applied Regression Analysis and Other Multivariable Methods*, 3rd ed.; Cengage Learning: Boston, MA, 1997.

(16) Arora, K.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18496−18501.

(17) Dünweg, B.; Kremer, K. *J. Chem. Phys.* **1993**, *99*, 6983−6997.

(18) Dünweg, B. *J. Chem. Phys.* **1993**, *99*, 6977−6982.

(19) Yeh, I. C.; Hummer, G. *J. Phys. Chem. B* **2004**, *108*, 15873−15879.

(20) Thouy, R.; Jullien, R. *J. Phys. A: Math. Gen.* **1994**, *27*, 2953−2963.

(21) Thouy, R.; Jullien, R. *J. Phys. I* **1996**, *6*, 1365−1376.

(22) Pranami, G.; Lamm, M. H.; Vigil, R. D. *Phys. Rev.* **2010**, *82*, 051402−1−051402−10.