Accurate Estimation of Diffusion Coefficients and their Uncertainties from Computer Simulation

Andrew R. McCluskey,[1,2,*] Samuel W. Coles,[3,4] and Benjamin J. Morgan[3,4,†]

[1] *School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, United Kingdom*
[2] *European Spallation Source ERIC, Ole Maaløes vej 3, 2200 København N, DK*
[3] *Department of Chemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK*
[4] *The Faraday Institution, Quad One, Harwell Science and Innovation Campus, Didcot, OX11 0RA, UK*

Self-diffusion coefficients, D^*, are routinely estimated from molecular dynamics simulations by fitting a linear model to the observed mean-squared displacements (MSDs) of mobile species. MSDs derived from simulation suffer from statistical noise, which introduces uncertainty in the resulting estimate of D^*. An optimal scheme for estimating D^* will minimise this uncertainty, i.e., will have high statistical efficiency, and will give an accurate estimate of the uncertainty itself. We present a scheme for estimating D^* from a single simulation trajectory with high statistical efficiency and accurately estimating the uncertainty in the predicted value. The statistical distribution of MSDs observable from a given simulation is modelled as a multivariate normal distribution using an analytical covariance matrix for an equivalent system of freely diffusing particles, which we parameterise from the available simulation data. We then perform Bayesian regression to sample the distribution of linear models that are compatible with this model multivariate normal distribution, to obtain a statistically efficient estimate of D^* and an accurate estimate of the associated statistical uncertainty.

I. INTRODUCTION

Mass transport is a fundamental physical process that is central to our understanding of fluids [1–3] and plays a critical role in biochemical systems [4, 5], and in solid-state devices such as batteries, fuel cells, and chemical sensors [6–8]. Molecular dynamics simulations are widely used to study microscopic transport processes, as they give direct insight into atomic-scale transport mechanisms and can be used to calculate macroscopic transport coefficients [9–14]. These transport coefficients are formally defined in terms of ensemble averages. Dynamical simulations, however, sample the full ensemble space stochastically, and parameters derived from simulation data are therefore only estimates of the true parameter of interest. The statistical uncertainty associated with such estimates depends on the details of the simulation—e.g., size and timescale—and on the choice of estimation method. An optimal estimation method will minimise the uncertainty in the computed quantity, i.e., it will have high statistical efficiency, and will also allow this uncertainty to be accurately estimated.

One commonly used parameter for quantifying atomic-scale mass transport is the self-diffusion coefficient, D^*, which describes diffusion in the absence of a chemical potential gradient. D^* is related to the ensemble-average mean squared displacement (MSD), $\left\langle \Delta\mathbf{r}(t)^2 \right\rangle$, via the Einstein relation [15, 16],

$$D^* = \lim_{t\to\infty} \frac{\left\langle \Delta\mathbf{r}(t)^2 \right\rangle}{6t}, \tag{1}$$

where t is elapsed time. Because numerical simulations are finite in both time and space, MSDs obtained from simulation data always deviate from the true ensemble average MSD. One can, however, compute an estimate of the self-diffusion coefficient, $\widehat{D}^*$, by fitting a linear model to the observed MSD and using the gradient of this fitted model in place of $\left\langle \Delta\mathbf{r}(t)^2 \right\rangle/t$ in Eqn. 1.

The simplest approach to fitting a linear model to MSD data from simulation is ordinary least squares regression (OLS). OLS gives analytical expressions for the "best fit" regression coefficients (the slope and intercept) and their respective uncertainties, making it easy to implement and quick to perform. This procedure, however, is appropriate only for data that are both statistically independent and identically distributed. Neither of these conditions hold for MSD data obtained from simulation, which instead are serially correlated and usually have unequal variances. As a consequence, OLS is statistically inefficient, giving a relatively large statistical uncertainty in $\widehat{D}^*$. Furthermore, using the textbook OLS expression for the uncertainty in $\widehat{D}^*$ significantly underestimates the true uncertainty in this estimate. This underestimated uncertainty may cause overconfidence in the accuracy of values of D^* estimated using OLS, and using these data in downstream analyses may result in faulty inferences. While the uncertainty associated with OLS estimates of D^* can, in principle, be accurately estimated by directly sampling over multiple repeated simulations, this approach greatly increases the total computational cost and therefore is often not practical.

Here, we describe an approximate Bayesian regression method for estimating D^* with near-maximal statistical efficiency while also accurately estimating the corresponding statistical uncertainty, using data from a single

[*] andrew.mccluskey@bristol.ac.uk
[†] b.j.morgan@bath.ac.uk

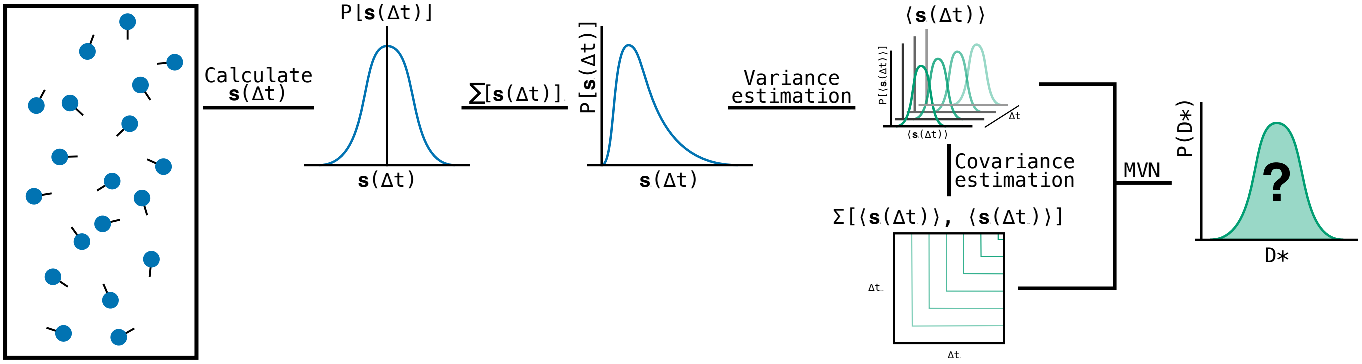

FIG. 1. A schematic diagram of the Bayesian regression method described in this work, running from our simulations through the sampling of displacement, variance and covariance estimation and final the sampling process to give the marginal posterior distribution for D^*.

simulation (Fig. 1). We model the statistical population of simulation MSDs as a multivariate normal distribution, using an analytical covariance matrix derived for an equivalent system of freely diffusing particles, with this covariance matrix parameterised from the observed simulation data. We then use Markov-chain Monte Carlo to sample the posterior distribution of linear models compatible with this multivariate normal model. The resulting posterior distribution provides an efficient estimate for $\widehat{D}^*$ and allows the associated statistical uncertainty in $\widehat{D}^*$ to be accurately quantified. This method is implemented in the open-source Python package KINISI [17].

II. RESULTS

A. Background

For a simulation of equivalent particles, the observed mean squared displacement as a function of time, $x(t)$, can be computed as an average over equivalent particles and time origins:

$$x(t) = \frac{1}{N(t)} \sum_{j=1}^{N(t)} \left[\Delta \mathbf{r}_j(t)\right]^2, \tag{2}$$

where $N(t)$ is the total number of observed squared-displacements at time t. The resulting observed MSD is a vector, $\boldsymbol{x}$, with individual elements x_i. Each element of this vector differs from the true ensemble-average MSD for that time by some unknown amount. Fitting a linear model to $\boldsymbol{x}$ gives an estimated self-diffusion coefficient, $\widehat{D}^*$, which again differs from the true self-diffusion coefficient, D^*, by some unknown amount.

Performing repeated simulations starting from different random seeds or with different histories will produce a set of replica trajectories, where each trajectory gives a different, statistically equivalent, observed MSD. The set of all possible replica trajectories defines a population of hypothetical observed MSDs, and the MSD obtained from any one trajectory can be considered a random sample, $\boldsymbol{X}$, drawn from the multivariate probability distribution that describes this population, i.e, $\boldsymbol{X} \sim p(\boldsymbol{x})$. Each potential MSD sample could, in principle, be fitted to a linear model to obtain a corresponding estimate for the self-diffusion coefficient; $\boldsymbol{X} \mapsto \widehat{D}^*$. The population of all such estimates therefore defines a probability distribution $p(\widehat{D}^*)$. The estimated diffusion coefficient obtained from a single simulation corresponds to a random sample drawn from this distribution, while the uncertainty in $\widehat{D}^*$ is described by the shape of the full distribution $p(\widehat{D}^*)$.

The statistical properties of $p(\widehat{D}^*)$ depend on both the input MSD data and the choice of regression scheme used to obtain a "best fit" linear model. An optimal estimation scheme for D^* should be unbiased, i.e., the expected value, $\mathbb{E}(\widehat{D}^*)$, should equal the true self-diffusion coefficient D^*, and should be maximally statistically efficient, i.e., the spread of $p(\widehat{D}^*)$ around D^* should be minimised. An estimation scheme should also provide an accurate estimate of the uncertainty in $\widehat{D}^*$, to allow this estimated parameter to be used in subsequent inferential analysis.

For data that are both statistically independent and identically normally distributed, ordinary least squares regression (OLS) is unbiased and statistically efficient, and gives accurate estimates of the uncertainties in the resulting regression coefficients. MSD data obtained from simulation, however, are neither statistically independent nor identically distributed. The variances, $\sigma^2[x_i]$, are correlated, since the displacement of each particle at time $t + \Delta t$ is necessarily similar to its displacement at time t, and hence, $x(t)$ is similar to $x(t + \Delta t)$. These variances are also typically unequal—the data are heteroscedastic [18**?** , 19]. Because the key assumptions of the OLS method are not valid for MSD data, OLS gives statistically inefficient estimates of D^*, while the estimated regression uncertainties obtained from the standard OLS statistical formulae significantly underestimate the true uncertainty in $p(\widehat{D}^*_{\mathrm{OLS}})$ (Fig. 2a).

Some improvement can be made by using weighted least squares (WLS) (Fig. 2b), where the residual for

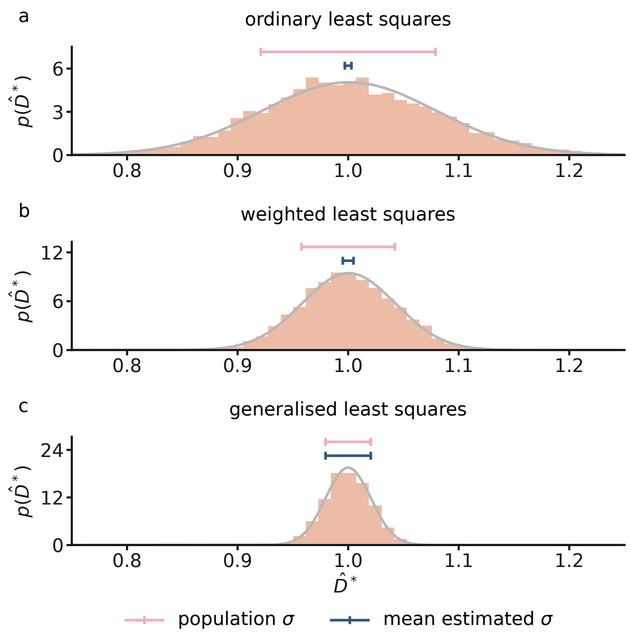

FIG. 2. Example distributions of estimated self-diffusion coefficients, $\widehat{D}^*$, calculated using (a) ordinary least squares (OLS), (b) weighted least squares (WLS), and (c) generalised least squares (GLS), from MSD data from 4096 individual simulations of 128 particles undergoing a 128 step 3D lattice random walk, with a step size chosen so that the true diffusion coefficient $D^* = 1$. In each panel, the grey curve shows the best-fit normal distribution for the simulation data, the upper horizontal bar shows the standard deviation of this distribution, and the lower horizontal bar shows the average estimated standard distribution given by the analytical expression for $\sigma[p(\widehat{D}^*)]$ for each regression method.

each observed MSD value is weighted by the reciprocal of its variance, $1/(\sigma^2[x_i])$. Like OLS, WLS is an unbiased estimator, and for heteroscedastic data it has higher statistical efficiency than OLS. WLS still disregards correlations in $\boldsymbol{x}$, however, and is therefore statistically inefficient, while the WLS estimated uncertainties for the regression coefficients still underestimate the true uncertainty in $p(\widehat{D}^*_{\mathrm{WLS}})$.

To optimally estimate the true ensemble-average MSD, and hence D^*, from simulation data, it is necessary to account for both the changing variance and correlation structure of $\boldsymbol{x}$. Within the framework of linear regression, this can be achieved using generalised least squares (GLS). GLS gives estimated regression coefficients, $\widehat{\beta}$, via

$$\widehat{\beta} = \left(\mathbf{A}^\top \boldsymbol{\Sigma}^{-1} \mathbf{A}\right)^{-1} \mathbf{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}, \tag{3}$$

where $\mathbf{A}$ is the model matrix $\begin{bmatrix}\mathbf{1} & \boldsymbol{t}\end{bmatrix}$, with $\boldsymbol{t}$ the vector of observed times, and $\boldsymbol{\Sigma}$ is the covariance matrix for the observed MSD values. For correlated heteroscedastic data, such as MSD data, GLS offers the theoretical maximum statistical efficiency—it achieves the Cramér–Rao bound [20–24]—and provides accurate analytical estimates of the uncertainty in the predicted regression coefficients (Fig. 2c).

An alternative method for estimating the ensemble-average MSD, and thus $\widehat{D}^*$, from simulation data is Bayesian regression. Like GLS, Bayesian regression can take into account both the changing variance and the correlation structure inherent in the data. Rather than providing a singular "best-fit" estimate like GLS, Bayesian regression produces a posterior probability distribution for the regression coefficients. The mean of this distribution serves as the point estimate of the coefficients and, in the absence of additional prior information, is equivalent to the GLS estimate, while the spread of the distribution quantifies the uncertainty in these estimates. For data that is both heteroscedastic and correlated, such as MSD data from simulations, Bayesian regression, like GLS, is formally fully statistically efficient.

To estimate D^* from some observed MSD data, $\boldsymbol{x}$, using Bayesian regression, we compute the posterior probability distribution $p(\boldsymbol{m}|\boldsymbol{x})$ for a linear model $\boldsymbol{m} = 6D^*\boldsymbol{t} + c$, where D^* and c are parameters to be estimated. This posterior distribution is described by Bayes' theorem,

$$p(\boldsymbol{m}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{m})p(\boldsymbol{m})}{p(\boldsymbol{x})}, \tag{4}$$

where $p(\boldsymbol{x}|\boldsymbol{m})$ is the probability of observing data $\boldsymbol{x}$ given model $\boldsymbol{m}$, often described as the "likelihood", and $p(\boldsymbol{x})$ is the marginal probability of the observed data $\boldsymbol{x}$. Integrating over $p(\boldsymbol{m}|\boldsymbol{x})$ with respect to c yields the marginal posterior distribution $p(D^*|\boldsymbol{x})$, from which the best point-estimate $\widehat{D}^*$ and distribution variance $\widehat{\sigma}^2[\widehat{D}^*]$ can be computed.

Given a sufficiently large number of observed squared displacements at each time t, the central limit theorem applies, and $\boldsymbol{x}$ can be considered a sample from a multivariate normal distribution with log-likelihood

$$\begin{aligned} \ln p(\boldsymbol{x}|\boldsymbol{m}) = -\frac{1}{2}\big[&\ln(|\boldsymbol{\Sigma}|) + (\boldsymbol{x}-\boldsymbol{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{m}) \\ &+ k\ln(2\pi)\big], \end{aligned} \tag{5}$$

where $\boldsymbol{\Sigma}$ is the observed MSD covariance matrix and k is the length of the vector $\boldsymbol{x}$, i.e., the number of time intervals for which we have observed MSD data. Providing this likelihood function can be calculated, we can compute the posterior distribution $p(\boldsymbol{m}|\boldsymbol{x})$, which gives an optimally efficient point-estimate for D^* and a complete description of the associated uncertainty in $\widehat{D}^*$.

B. Approximating $\boldsymbol{\Sigma}$ from simulation data

For Bayesian regression and GLS, we require the covariance matrix for the observed MSD, $\boldsymbol{\Sigma}$, which is generally unknown. To proceed, we replace $\boldsymbol{\Sigma}$ with a model covariance matrix, $\boldsymbol{\Sigma}'$, with a known analytical form, that we parameterise from the available simulation data. Providing the correlation structure of $\boldsymbol{\Sigma}'$ is similar to that of $\boldsymbol{\Sigma}$, this model correlation matrix can be used in approximate Bayesian or GLS schemes to estimate the ensemble-average MSD, and hence D^*, with high efficiency and accurate estimated uncertainties.

We model the covariance matrix for the observed MSD from a given simulation using the covariance matrix for the MSD of an equivalent system of freely diffusing particles, $\boldsymbol{\Sigma}'$. For observed MSDs computed by averaging over numerically-independent sub-trajectories, the covariance matrix $\boldsymbol{\Sigma}'$, in the long time limit, has elements (see SI)

$$\Sigma'\left[x_i, x_j\right] = \Sigma'\left[x_j, x_i\right] = \sigma^2[x_i]\frac{N_i'}{N_j'}, \quad \forall\, i \leq j, \tag{6}$$

where $\sigma^2[x_i]$ are the time-dependent variances of the observed MSD, and N_i' is the total number of numerically-independent observed squared-displacements for time-interval i. We estimate the variances $\sigma^2[x_i]$ using the standard result that the variance of the mean of a sample scales inversely with the number of independent constituent observations. Specifically, we approximate the variance $\widehat{\sigma}^2[x_i]$ (see SI) by rescaling the observed variance of the squared displacement for time interval i by the number of numerically-independent contributing sub-trajectories, N_i';

$$\widehat{\sigma}^2[x_i] \approx \frac{1}{N_i'}\sigma^2[\Delta r_i^2]. \tag{7}$$

Rescaling by the number of numerically-independent contributing sub-trajectories has the effect of renormalising the variance of the observed squared displacements to account for correlations between particle squared displacements computed from overlapping time windows. An alternative approach to renormalising $\sigma^2[x_i]$ is to use a non-parametric block-averaging procedure [25–27], which gives undesirable results for a random walk (see SI). The block-averaging approach, additionally, requires numerical convergence with respect to block size, which is not guaranteed for time windows with few independent observations. We require $\sigma^2[x_i]$ at all observed time intervals, i, making block averaging on large data sets computationally prohibitive.

The estimated variance $\widehat{\sigma}^2[\boldsymbol{x}]$ can be calculated from a single simulation trajectory, and provides an accurate estimate of the true variance $\sigma^2[\boldsymbol{x}]$. To demonstrate this, we performed 4096 independent simulations of 128 particles undergoing a three-dimensional cubic-lattice random walk of 128 steps per particle. Using data from all 4096 simulations, we first compute the true simulation MSD and its variance (Fig. 3a). We also compute the MSD and estimated variance using data from a single simulation trajectory (Fig. 3b), using the scheme described above. A quantitative comparison between the true MSD variance and the single-trajectory estimated MSD variance is made in Fig. 3c: the close numerical agreement confirms that Eqn 7 can be used to estimate $\sigma^2[\boldsymbol{x}]$, which can then be used to parameterise the model covariance matrix $\boldsymbol{\Sigma}'$ via Eqn. 6.

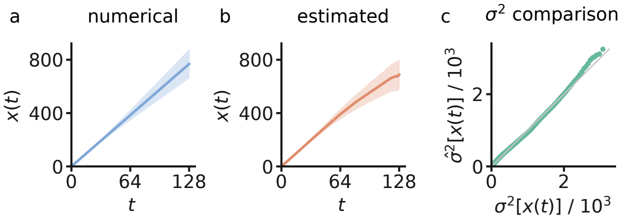

FIG. 3. Comparison of the numerical variance in observed MSD from multiple replica simulations and the estimated variance in observed MSD given by rescaling the variance in observed squared displacements (Eqn. 7). Panel (a) shows the mean observed MSD from 4096 simulations of 128 particles undergoing a 3D lattice random walk of 128 steps per particle, with error bars of $\pm 2\sigma[x_i]$. Panel (b) shows the MSD from just one simulation, with error bars of $\pm 2\widehat{\sigma}[x_i]$, obtained via Eqn. 7. Panel (c) plots the numerical variance against the estimated variance from a single simulation as a function of timestep i.

The practical implementation of both GLS and Bayesian regression requires that the covariance matrix $\boldsymbol{\Sigma}'$ is invertible (positive definite); see Eqns. 3 and 5. The estimated MSD variances derived from simulation data via Eqn. 7 are statistically noisy and using these to directly parameterise $\boldsymbol{\Sigma}'$ can yield non-invertible singular matrices. To make our scheme numerically tractable, we therefore fit our estimated MSD variances to the analytical variance for an analogous system of particles undergoing random walks [18];

$$\sigma^2[x_i] = a\frac{t_i^2}{N_i'}, \tag{8}$$

where a is a scaling parameter determined by fitting Eqn. 8 to the directly estimated MSD variances. This smoothing of $\sigma^2[x_i]$ guarantees that the resulting model covariance matrix $\boldsymbol{\Sigma}'$ is invertible and thus suitable for GLS or Bayesian regression.

To illustrate the complete numerical procedure for deriving the model covariance matrix, $\boldsymbol{\Sigma}'$, we present in Fig. 4 the MSD covariance matrix for 4096 random-walk simulations, as described above, at three differing levels of approximation: the numerically converged covariance matrix, $\boldsymbol{\Sigma}$, computed using the data from all 4096 simulations (Fig. 4a); the corresponding analytical model covariance matrix, $\boldsymbol{\Sigma}'$, defined by Eqn. 6 and parametrised using analytical variances $\sigma^2[x_i]$ (Fig. 4b); and the average model covariance matrix obtained by parametrising Eqn. 6 using smoothed variances estimated from individual simulation trajectories, and averaging over the resulting set of all 4096 matrices (Fig. 4c).

While the analytical and average estimated covariance matrices show some systematic deviation from the numerically converged covariance matrix, the general cor-

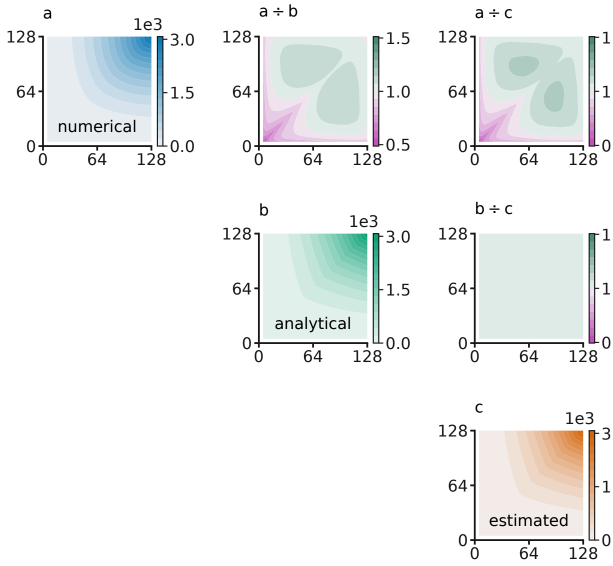

FIG. 4. (a) The numerical MSD covariance matrix $\mathbf{\Sigma}$ calculated using MSD data from 4096 simulations of 128 particles undergoing a 3D lattice random walk of 128 steps per particle. (b) The analytical MSD covariance matrix $\mathbf{\Sigma}'$ (Eqn. 6), parametrised using analytical random-walk variances $\sigma^2[x_i]$. (c) The MSD covariance matrix obtained applying the numerical scheme described in the main text to each individual random walk simulation, averaged over all 4096 such simulations. Colour bars in (a–c) show the covariance, $\Sigma\,[x_i, x_j]$. The off-diagonal panels show difference plots, computed as per-element ratios between pairs of covariance matrices (a–c).

relation structure is preserved. The discrepancy between the model and numerical covariance matrices largely stems from the approximation made in deriving the analytical form that t is large, which leads to an overestimation of the variance at low t. Despite this, the average estimated covariance matrix reproduces well the correlation structure of the true numerical covariance matrix, indicating that the covariance matrices estimated from individual simulation trajectories may be used within approximate GLS or Bayesian regression schemes to estimate D^* and $\sigma^2[\widehat{D}^*]$.

C. Validation

To demonstrate the complete approximate Bayesian regression scheme, as described above, we present two distinct examples. First, we consider a simple 3D-lattice random walk, where the true self-diffusion coefficient D^* is specified by the simulation parameters, and a well-converged numerical covariance matrix can be obtained with relatively low computational cost, which allows us to directly compare the estimates produced by our method to “best case” estimates from a hypothetical method with access to the true covariance matrix. Second, we consider an example real-world system—the lithium-ion solid electrolyte $Li_7La_3Zr_2O_{12}$ (LLZO)—which represents an application of our method to a well-studied material of practical interest for solid-state lithium-ion batteries [28–31].

Fig. 5a shows the observed MSD from a single 3D-lattice random-walk simulation, along with the estimated posterior distribution of linear models compatible with the observed MSD data, $p(\boldsymbol{m}|\boldsymbol{x})$, calculated via Eqns. 4 and 5. The corresponding marginal posterior distribution of estimated diffusion coefficients $p(D^*|\boldsymbol{x})$ is shown in Fig. 5b; this distribution is approximately Gaussian and is centred on the true self-diffusion coefficient $D^* = 1$, demonstrating that for this example trajectory we obtain a good point-estimate of D^*.

To evaluate the overall performance of our method, we repeat our analysis on the full set of 4096 random-walk simulations. Fig. 5c presents a histogram of the resulting point estimates of D^*, with each estimate derived as the mean of the posterior distribution $p(D^*|\boldsymbol{x})$ using input data from an individual simulation. We also show the probability distribution of estimated diffusion coefficients obtained using Bayesian regression with a mean vector and covariance matrix derived numerically from all 4096 simulations (solid line). This latter distribution represents the distribution of “best possible” estimates of D^* and exhibits the minimum possible theoretical variance. The close agreement between these two distributions demonstrates that our approximate Bayesian regression scheme yields nearly optimal estimates of D^* using data from individual simulations. The distribution of estimated diffusion coefficients from single simulations is slightly broader than the exact numerical results. This minor deviation is a consequence of the overestimation of $\widehat{\sigma}^2[x_i]$ at short times, noted above, which results from our use of the long-time limit in the derivation of the analytical model covariance matrix.

We next consider the degree to which our method can quantify the uncertainty in $\widehat{D}^*$ when using input data from a single simulation. Fig. 5d shows the distribution of estimated variances $\widehat{\sigma}^2[\widehat{D}^*]$, with each sample calculated from an individual simulation trajectory. We also show the true variance of individual point estimates, $\sigma^2[\widehat{D}^*]$, which characterises the spread of the histogram in Fig. 5c. The distribution of estimated variances is biased relative to the true variance and skewed, due to numerical differences between the true covariance matrix $\mathbf{\Sigma}$ and the model covariance matrix $\mathbf{\Sigma}'$ (further details are provided in the SI). In general, however, the distribution of the estimated variance shows good agreement with the true sample variance. Notably, the precision of this estimate is significantly greater than obtained using OLS or WLS and their corresponding textbook statistical formulae.

We next benchmark our method using data from simulations of the lithium-ion solid electrolyte cubic $Li_7La_3Zr_2O_{12}$ (c-LLZO). We performed a single simulation of 1536 atoms (448 Li ions) at 1000 K for 1.6 ns (full

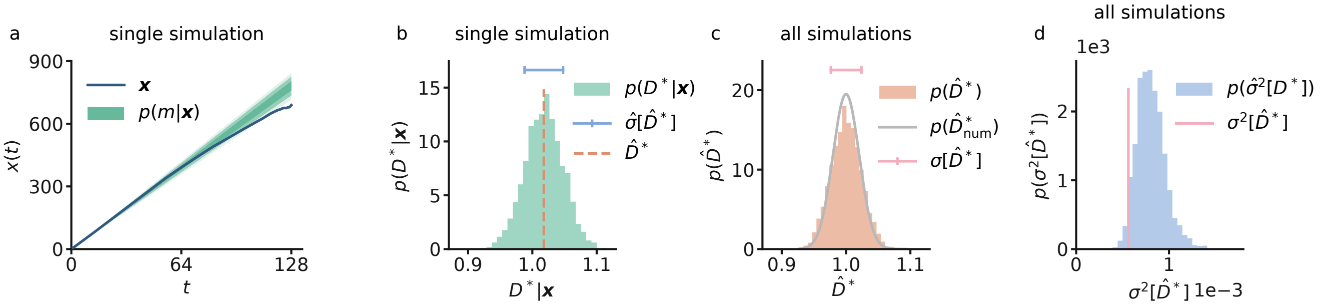

FIG. 5. (a) Observed MSD from a single simulation of 128 particles undergoing a 3D-lattice random walk of 128 steps per particle (dark line). The green shading shows the corresponding posterior distribution $p(\boldsymbol{m}|\boldsymbol{x})$ of linear models compatible with the observed MSD data $\boldsymbol{x}$, calculated using the scheme described in the main text. The variegated shading indicates compatibility intervals of (1, 2, and 3) $\sigma[p(\boldsymbol{m}|\boldsymbol{x})]$. (b) The marginal posterior distribution $p(\widehat{D}^*|\boldsymbol{x})$ obtained from the posterior distribution of linear models in (a). The mean of this distribution gives the point estimate $\widehat{D}^*$ for this simulation input data. The blue horizontal bar shows an interval of one standard deviation in $p(\widehat{D}^*|\boldsymbol{x})$. (c) Probability distribution of point-estimates $p(\widehat{D}^*)$ obtained from 4096 individual random-walk simulations. Each simulation has been analysed as in (a) and (b) to yield a single corresponding point estimate $\widehat{D}^*$. The grey line shows the distribution of point estimates, $p(\widehat{D}^*_{\mathrm{num}})$, obtained using Bayesian regression with a mean vector and numerical covariance matrix derived from the complete dataset of all 4096 simulations. The pink horizontal bar shows an interval of one standard deviation in $p(\widehat{D}^*)$. (d) Probability distribution of estimated variances, $\widehat{\sigma}^2[\widehat{D}^*]$, for individual random-walk simulations, compared to the true sample variance (pink vertical line) $\sigma^2[\widehat{D}^*]$.

simulation details are provided in the Methods section). To generate multiple statistically equivalent trajectories, the resulting simulation data was partitioned into 512 effective trajectories, each approximately ~ 25 ps in length, and containing data for 56 lithium ions. We then perform the same approximate Bayesian regression analysis as above on each effective trajectory, excluding the first 10 ps of MSD data in each case to remove short-time data corresponding to ballistic and sub-diffusive regimes [19].

The resulting distribution of the point estimates, $\widehat{D}^*$, from analysis of all 512 effective trajectories is shown in Fig. 6a. Again, the corresponding distribution of $\widehat{D}^*$ estimates derived using Bayesian regression and a well-converged numerical covariance matrix calculated from the full LLZO dataset is also shown for comparison. The distribution $p(\widehat{D}^*)$ obtained using the model covariance matrix and parametrised separately for each individual effective simulation is highly similar to that obtained using the aggregate numerical covariance matrix calculated from the complete simulation dataset. This close agreement mirrors the results for our random walk simulations (see the SI for a similar comparison of the OLS, WLS, and GLS as shown in Fig. 2), and confirms that our method yields accurate and statistically efficient estimates for D^*, even for real-world simulation data.

We also consider the probability distribution of estimates of the variance in $\widehat{D}^*$ calculated for each effective trajectory (Fig. 6b), which we compare to the true variance in $\widehat{D}^*$ for our method; i.e., the variance of the histogram in Fig. 6a. While the estimated variances deviate somewhat from the true distribution $p(\sigma^2[\widehat{D}^*])$, the agreement is reasonable and mirrors our results for the random walk simulations. Hence, our method provides reasonably accurate estimates of the uncertainty in $\widehat{D}^*$ for our c-LLZO dataset, even when applied to single effective trajectories with limited displacement data (only 56 mobile ions, and 25 ps simulation length).

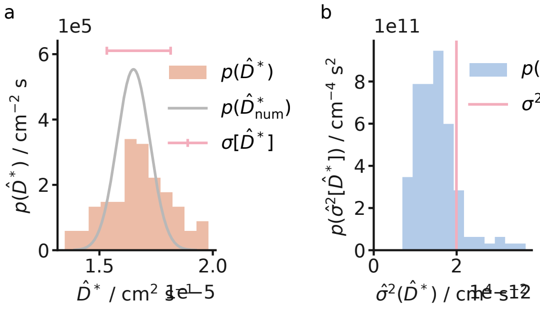

FIG. 6. (a) Probability distribution of point estimates $p(\widehat{D}^*)$ for 512 effective simulations of LLZO (orange histogram). The grey line shows the distribution $p(\widehat{D}^*_{\mathrm{num}})$ obtained using Bayesian regression with the complete LLZO dataset as input. The pink bar shows an interval of one standard deviation $\sigma[p(\widehat{D}^*)]$. (b) Probability distribution of estimated variances, $\widehat{\sigma}^2[\widehat{D}^*]$, for individual LLZO effective simulations, compared to the true sample variance (pink vertical line) $\sigma^2[\widehat{D}^*]$.

D. $\sigma^2[\widehat{D}^*]$ scaling and comparison to OLS, WLS, and GLS

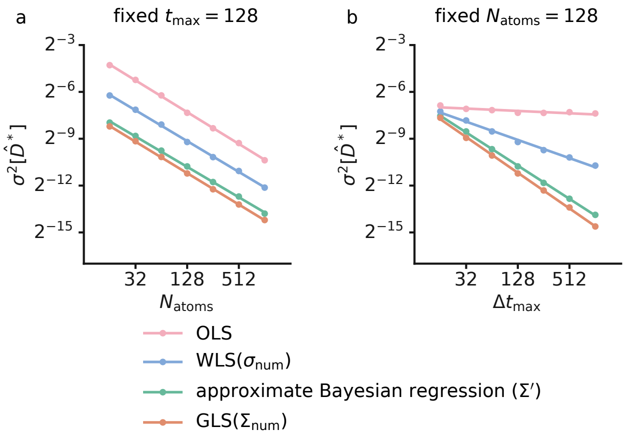

FIG. 7. Scaling of $\sigma^2[\widehat{D}^*]$ with simulation size for OLS (pink), WLS (blue), our approximate Bayesian regression method (green), and GLS (orange). (a) Scaling versus number of mobile particles, N_{atoms}. (b) Scaling versus total simulation time, t_{max}. Solid lines show fitted power law relationships for each dataset. The WLS and GLS data are obtained using numerically determined variances and covariance, respectively, from a set of 512 repeat simulations for each combination of N_{atoms} and t_{max}.

Fig. 7 presents an analysis of the variation in $\sigma^2[\widehat{D}^*]$ as the number of mobile particles (Fig. 7a) and the total simulation time (number of steps) (Fig. 7b) are changed. We compare four methods for estimating D^* from the observed MSD data: OLS, WLS, the approximate Bayesian regression method described here, and GLS. When estimating D^* using WLS and GLS, we calculate the variances and the covariance matrix, respectively, numerically, using the complete set of 512 simulations. Each data point in Fig. 7 represents the variance across point-estimates of D^* derived from 512 individual 3D-lattice random walk simulations, for each combination of N_{atoms} and t_{max}. The GLS dataset corresponds to an optimally efficient estimator for linear regression of observed MSD data, equivalent to performing Bayesian regression with the numerical covariance matrix and an uninformative prior.

Our approximate Bayesian regression method performs similarly to GLS, with a numerically converged covariance matrix, and gives significantly reduced uncertainty in $\widehat{D}^*$ compared to OLS or WLS, for all simulation sizes and lengths considered. Moreover, our method scales better than OLS or WLS as the total simulation time is increased. The approximate Bayesian regression method in this manuscript therefore presents a significant improvement over more conventional methods such as OLS and WLS, by enabling more precise estimates of D^* across varied simulation sizes at equivalent computational cost.

III. SUMMARY AND DISCUSSION

We have introduced and evaluated an approximate Bayesian regression method for estimating the self-diffusion coefficient, D^*, from molecular dynamics simulation data. We consider the observed mean-squared displacement data from a single simulation as a random sample, $\boldsymbol{X}$, from a population of potential MSDs generated by equivalent replica simulations, $\boldsymbol{X} \sim p(\boldsymbol{x})$. We model this population using a multivariate normal distribution, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$, with mean vector $\boldsymbol{m} = 6D^*\boldsymbol{t} + c$, where D^* and c are model parameters to be determined.

To model the covariance matrix, we use an analytical solution derived for an equivalent system of freely diffusing particles. To parameterise this model covariance matrix, we rescale the variance of the observed squared displacements from the input simulation trajectory, followed by a smoothing step to ensure a positive-definite matrix. The resulting model covariance matrix preserves the correlation structure of the true simulation MSD covariance matrix, and gives a multivariate normal model for the population of observable simulation MSDs that depends solely on the model parameters, D^* and c.

We use Markov-Chain Monte Carlo to sample the posterior distribution of linear models compatible with the observed MSD data. This approach yields a marginal posterior distribution, $p(D^*|\boldsymbol{x})$, that gives a statistically efficient point estimate for D^* and allows the associated statistical uncertainty, $\sigma^2[\widehat{D}^*]$, to be quantified.

We have benchmarked our approach using simulation data for an ideal 3D lattice random walk and for the lithium-ion solid electrolyte $\mathrm{Li_7La_3Zr_2O_{12}}$ (LLZO). In both cases, we obtain a distribution of estimates for D^* that closely matches the theoretically optimal distribution obtained using a numerical covariance matrix derived from a large number of replica simulation trajectories.

We obtain estimates for D^* that are unbiased, with near-optimal statistical efficiency, using input data from single simulation trajectories. The approximate Bayesian regression scheme therefore provides more accurate single-point estimates of the self-diffusion coefficient than the commonly used OLS or WLS methods, when applied to the same input simulation data. The improved statistical efficiency of this method, when compared to OLS or WLS, enables the estimation of D^* with equivalent accuracy from considerably smaller simulations—either in terms of timescale or system size. This reduces the overall computational cost when compared to studies that use OLS or WLS for estimating a linear fit to simulation MSD data. Alternatively, this approach provides the possibility to estimate D^* with greater precision, given simulation trajectories of equal size.

Our method also provides reasonable estimates of the statistical uncertainty in the estimated value $\widehat{D}^*$, in contrast to OLS and WLS which systematically significantly underestimate the uncertainty in regression coefficients when applied to simulated MSD data. While these estimated statistical uncertainties can still differ from the true (but unknown) uncertainty in $\widehat{D}^*$, particularly when using short-timescale simulation data, they allow for scientifically meaningful comparisons to be made between estimated diffusion coefficients across different materials or under varying conditions, such as changes in temperature, or between computational findings and experimental results. Furthermore, these uncertainties allow for quantitative downstream analysis, such as the application of Arrhenius (on non-Arrhenius) type models to describe the temperature dependence of self-diffusion.

The approximate Bayesian regression scheme presented here provides a statistically efficient means of estimating the self-diffusion coefficient, D^*, from molecular dynamics simulation data. It improves upon textbook approaches by providing accurate point estimates of D^* with near-optimal statistical efficiency, while also providing a reasonable description of the uncertainty in these estimates. The high statistical efficiency of our method allows for the use of smaller simulations, which can significantly reduce computational costs. Overall, our method offers significant advantages over more conventional methods of estimating self-diffusion coefficients from atomistic simulations. We have implemented this procedure in the open-source package KINISI [17], which we hope will support its use within the broader simulation community across a range of materials science contexts.

IV. METHODS

A. Numerical implementation in KINISI

We have implemented the approximate Bayesian estimation method described in the main text in the open-source Python package KINISI [17], under the MIT licence.

KINISI uses overlapping sliding window sampling when calculating the observed mean squared displacement at each time interval t (see Eqn. 2). For a given time interval, t, the maximum number of observations is $N_{\text{atoms}} \times (N_t - i)$ displacements, where N_{atoms} is the number of atoms, N_t is the total number of timesteps, and i is the index of the timestep (where 1 is the index for the shortest timestep). To estimate the variance of the observed MSD, we rescale the variance of observed squared displacements by the number of numerically-independent sub-trajectories in the simulation, $N_i' = N_{\text{atoms}} \times N_t / i$, as presented in Eqn. 7.

The parametrisation of the covariance matrix from the variances $\sigma^2[x_i]$ and the number of independent observations N_i' is defined by Eqn. 6. The covariance matrix is only constructed for values of t where the particle motion is considered to be in the long-time diffusive limit, with this threshold set by the user to a value appropriate for their system and simulation data. For the examples presented in the main manuscript, we consider particles to be in the diffusive regime from $t = 4$ for the random walk trajectories and from $t = 10\,\text{ps}$ for the LLZO simulations.

KINISI uses ordinary least squares to obtain an initial guess for the gradient and intercept of the linear model describing the observed MSD. This initial guess is then used as the starting point for minimising the negative maximum a posteriori (the peak of the posterior distribution as per Eqn. 4), with the improper prior that $D^* \geq 0$ [32–35]. The log-likelihood calculation (Eqn. 5) uses the Moore-Penrose generalisation of the inverse of a Hermitian matrix [36–38].

To sample the joint posterior probability distribution of the linear model, KINISI uses the EMCEE package [39], which implements Goodman and Weare's affine invariant Markov chain Monte Carlo ensemble sampler [40]. When sampling $p(D^*|\boldsymbol{m})$ we again apply the improper prior $D^* \geq 0$. The sampling process uses 32 walkers for 1500 steps, with the first 500 steps discarded as a burn-in period. The sampled chains are thinned such that only every 10th value is retained, yielding 3200 points sampled from the posterior distribution $p(D^*|\boldsymbol{m})$. These points can then be plotted as a histogram (as in Fig 5b), and summary statistics $\widehat{D}^*$ and $\widehat{\sigma}^2[\widehat{D}^*]$ can be derived.

B. LLZO simulations

Classical molecular dynamics were run using the METALWALLS code [41]. We used the DIPPIM polarisable ion force field, as parameterised by Burbano *et al.* [29], due to its proven accuracy in accounting for the effect of ion polarisability on diffusion [29, 42]. We simulated the cubic phase of LLZO in NVT ensemble at a temperature of 1000 K. Simulations were run for 2 ns with a 2 fs timestep. To control temperature, we used a Nosé-Hoover thermostat, with a relaxation time of 121 fs (5000 $\hbar/E_h$) [43–45]. Simulations were performed using $2\times2\times2$ supercells with 1536 atoms following the same protocol as in Ref. 29.

DATA & CODE AVAILABILITY

Electronic Supplementary Information (ESI) available: A complete set of analysis/plotting scripts allowing for a fully reproducible and automated analysis workflow, using SHOWYOURWORK [46], for this work is available at Ref. [47] under an MIT license. All raw simulation files are available on Zenodo shared under CC BY-SA 4.0 licences, the random walk simulations and other analysis-linked data [48] and the LLZO raw simulation trajectories [49]. The method outlined in this work is implemented in the open-source Python package KINISI [17], which is available under an MIT license, and can be accessed via `https://github.com/bjmorgan/kinisi`.

AUTHOR CONTRIBUTION STATEMENT

A.R.M.: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualisation, Writing—original draft. S.W.C.: Methodology, Resources, Writing—review and editing. B.J.M.: Conceptualization, Methodology, Software, Writing—review and editing.

ACKNOWLEDGEMENTS

The authors thank Jacob M. Dean and Tim Rogers for their valuable input in checking the mathematical derivations that make up the appendices, the beta-testers for the kinisi package, and Rodrigo Luger and Daniel Foreman-Mackey for their help using showyourwork [46]. This work used the Isambard 2 UK National Tier-2 HPC Service (http://gw4.ac.uk/isambard/) operated by GW4 and the UK Met Office, and funded by EPSRC (EP/T022078/1). The authors acknowledge the University of Bath's Research Computing Group for their support in running the LLZO molecular dynamics simulations. Other simulations and analyses were carried out using the Data Management and Software Centre computing cluster at the European Spallation Source ERIC. S.W.C. and B.J.M. acknowledge the support of the Faraday Institution through the CATMAT project (grant number FIRG016). B.J.M. acknowledges support from the Royal Society (UF130329 and URF\R\191006). The funders listed above played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

COMPETING INTERESTS

All authors declare no financial or non-financial competing interests.

[1] Sendner, C., Horinek, D., Bocquet, L. & Netz, R. R. Interfacial water at hydrophobic and hydrophilic surfaces: Slip, viscosity, and diffusion. *Langmuir* **25**, 10768–10781 (2009).

[2] Shimizu, K. *et al.* Structural and aggregate analyses of (Li salt + glyme) mixtures: the complex nature of solvate ionic liquids. *Phys. Chem. Chem. Phys.* **17**, 22321–22335 (2015).

[3] Ghoufi, A., Szymczyk, A. & Malfreyt, P. Ultrafast diffusion of ionic liquids confined in carbon nanotubes. *Sci. Rep.* **6** (2016).

[4] McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).

[5] Robertson, R. M., Laib, S. & Smith, D. E. Diffusion of isolated DNA molecules: Dependence on length and topology. *Proceedings of the National Academy of Sciences* **103**, 7310–7314 (2006).

[6] Eames, C. *et al.* Ionic transport in hybrid lead iodide perovskite solar cells. *Nature Commun.* **6** (2015).

[7] Morgan, B. J. Understanding fast-ion conduction in solid electrolytes. *Phil. Trans. Roy. Soc. A* **379** (2021).

[8] Walsh, A. & Stranks, S. D. Taking Control of Ion Transport in Halide Perovskite Solar Cells. *ACS Energy Lett.* **3**, 1983–1990 (2018).

[9] Morgan, B. J. & Madden, P. A. Relationships between atomic diffusion mechanisms and ensemble transport coefficients in crystalline polymorphs. *Phys. Rev. Lett.* **112** (2014).

[10] Morgan, B. J. Mechanistic origin of superionic lithium diffusion in anion-disordered Li$_6$PS$_5X$ argyrodites. *Chem. Mater.* **33**, 2004–2018 (2021).

[11] Poletayev, A. D., Dawson, J. A., Islam, M. S. & Lindenberg, A. M. Defect-driven anomalous transport in fast-ion conducting solid electrolytes. *Nat. Mater.* **21**, 1066–1073 (2022).

[12] Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Op. Struct. Biol.* **19**, 120–127 (2009).

[13] Wang, J. & Hou, T. Application of molecular dynamics simulations in molecular property prediction II: Diffusion coefficient. *J. Comput. Chem.* **32**, 3505–3519 (2011).

[14] Zelovich, T. *et al.* Hydroxide Ion Diffusion in Anion-Exchange Membranes at Low Hydration: Insights from Ab Initio Molecular Dynamics. *Chem. Mater.* **31**, 5778–5787 (2019).

[15] Einstein, A. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Ann. Phys. (Berl.)* **322**, 549–560 (1905).

[16] Helfand, E. Transport coefficients from dissipation in a canonical ensemble. *Physical Review* **119**, 1–9 (1960).

[17] McCluskey, A. R. & Morgan, B. J. kinisi-0.6.3. https://github.com/bjmorgan/kinisi (2023).

[18] Smith, W. & Gillan, M. J. The Random Walk and the Mean Squared Displacement. *Inf. Q. Comput. Simul. Condens. Phases* 54–64 (1996).

[19] He, X., Zhu, Y., Epstein, A. & Mo, Y. Statistical variances of diffusional properties from ab initio molecular dynamics simulations. *npj Comput. Mater.* **4**, 18 (2018).

[20] Cramér, H. *Mathematical methods of statistics (PMS-9), volume 9* (Princeton University Press, 1946).

[21] Rao, C. R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–89 (1945).

[22] Rao, C. R. *Selected papers of C. R. Rao* (John Wiley & Sons, 1994).

[23] Darmois, G. Sur les limites de la dispersion de certaines estimations. *Rev. Int. Inst. Statist.* **13**, 9 (1945).

[24] Aitken, A. C. & Silverstone, H. XV. – on the estimation of statistical parameters. *Proc. R. Soc. Edinb. A.* **61**, 186–194 (1942).

[25] Flyvbjerg, H. & Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **91**, 461 (1989).

[26] Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications* (Academic Press, 2023), 3 edn.

[27] Materzanini, G., Kahle, L., Marcolongo, A. & Marzari, N. High li-ion conductivity in tetragonal LGPO: A comparative first-principles study against known LISICON and LGPS phases. *Phys. Rev. Mater.* **5** (2021).

[28] Murugan, R., Thangadurai, V. & Weppner, W. Fast lithium ion conduction in garnet-type $Li_7La_3Zr_2O_{12}$. *Angew. Chem. Int. Ed.* **46**, 7778–7781 (2007).

[29] Burbano, M., Carlier, D., Boucher, F., Morgan, B. J. & Salanne, M. Sparse cyclic excitations explain the low ionic conductivity of stoichiometric $Li_7La_3Zr_2O_{12}$. *Phys. Rev. Lett.* **116** (2016).

[30] Morgan, B. J. Lattice-geometry effects in garnet solid electrolytes: a lattice-gas Monte Carlo simulation study. *R. Soc. Open Sci.* **4**, 170824 (2017).

[31] Squires, A. G. *et al.* Low electronic conductivity of $Li_7La_3Zr_2O_{12}$ solid electrolytes from first principles. *Phys. Rev. Mater.* **6** (2022).

[32] Broyden, C. G. The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA J. Appl. Math.* **6**, 76–90 (1970).

[33] Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **13**, 317–322 (1970).

[34] Goldfarb, D. A family of variable-metric methods derived by variational means. *Math. Comput.* **24**, 23–26 (1970).

[35] Shanno, D. F. Conditioning of quasi-newton methods for function minimization. *Math. Comput.* **24**, 647–656 (1970).

[36] Moore, E. H. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.* **26**, 394–395 (1920).

[37] Bjerhammar, A. Application of calculus of matrices to method of least squares; with special references to geodetic calculations. *Kungl. Tekn. Högsk. Hand. Stockholm* **49** (1951).

[38] Penrose, R. A generalized inverse for matrices. *Math. Proc. Camb. Philos. Soc.* **51**, 406–413 (1955).

[39] Foreman-Mackey, D. *et al.* emcee v3: A Python ensemble sampling toolkit for affine-invariant MCMC. *J. Open Source Softw.* **4**, 1864 (2019).

[40] Goodman, J. & Weare, J. Ensemble samplers with affine invariance. *Comm. App. Math. Comp. Sci.* **5**, 65–80 (2010).

[41] Marin-Laflèche, A. *et al.* MetalWalls: A classical molecular dynamics software dedicated to the simulation of electrochemical systems. *J. Open Source Softw.* **5**, 2373 (2020).

[42] Wilson, M. & Madden, P. A. Polarization effects in ionic systems from first principles. *J. Phys.: Condens. Matter* **5**, 2687–2706 (1993).

[43] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **81**, 511–519 (1984).

[44] Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695–1697 (1985).

[45] Martyna, G. J., Klein, M. L. & Tuckerman, M. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635–2643 (1992).

[46] Luger, R. showyourwork. https://github.com/rodluger/showyourwork (2021).

[47] McCluskey, A. R., Coles, S. W. & Morgan, B. J. msd-errors-0.0.1. https://github.com/arm61/msd-errors (2023).

[48] McCluskey, A. R. Data for github.com:arm61/msd-errors. DOI: 10.5072/zenodo.1206872 (2022).

[49] Coles, S. W. LLZO simulation trajectories (2022).

[50] Howard, R. E. & Lidiard, A. B. Matter transport in solids. *Rep. Prog. Phys.* **27**, 161–240 (1964).

[51] Spencer, J., Eikås, R. D. R., Neufeld, V. & Poole, T. pyblock-0.6. https://github.com/jsspencer/pyblock (2020).

Supplemental Material for "Accurate Estimation of Diffusion Coefficients and their Uncertainties from Computer Simulation"

Andrew R. McCluskey,[1, 2, *] Samuel W. Coles,[3, 4] and Benjamin J. Morgan[3, 4, †]

[1] *School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, United Kingdom*
[2] *European Spallation Source ERIC, Ole Maaløes vej 3, 2200 København N, DK*
[3] *Department of Chemistry, University of Bath, Claverton Down, Bath, BA2 7AY, UK*
[4] *The Faraday Institution, Quad One, Harwell Science and Innovation Campus, Didcot, OX11 0RA, UK*

This document presents supplementary material for the manuscript "Accurate Estimation of Diffusion Coefficients and their Uncertainties from Computer Simulation". It contains the following sections:

1. The derivation of the covariance matrix in the long-time limit for freely diffusion particles.

2. Discussion of the origin of bias and skew in the distribution of the estimated variance of the estimated diffusion coefficient, $p(\widehat{\sigma}^2[\widehat{D}^*])$.

3. Details of the LLZO molecular dynamics simulations.

4. Details of the implementation of the model covariance method in the KINISI package.

A repository containing the analysis and plotting code used to generate all results and figures in the main manuscript and this supplemental material document is available at `www.github.com/arm61/msd-errors` [47], under MIT (code) and CC BY-SA 4.0 (figures and text) licenses. This repository includes a fully reproducible showyourwork workflow, which allows complete reproduction of the analysis, plotting of figures and compilation of the manuscripts. The corresponding input datasets are openly available under the CC BY-SA 4.0 licence [48, 49].

SI.I: Derivation of the long-time limit covariance matrix for a system of freely diffusing particles.

In the main manuscript we present the result that the covariance matrix for a system of freely diffusing particles, in the long-time limit, has the form

$$\Sigma'\left[x_i, x_j\right] = \Sigma'\left[x_j, x_i\right] = \sigma^2[x_i]\frac{N'_i}{N'_j}, \quad \forall\, i \leq j, \tag{SI.1}$$

where x_i is the observed mean-squared displacement (MSD) for time interval i and N'_i is the number of statistically independent observed squared displacements averaged over to compute the mean value.

To derive this result, we first present a derivation of the expected variance for the MSD at timestep i, $\sigma^2[\boldsymbol{x}]$, following the approach of Smith and Gillan [18]. We then derive an expression for the covariance $\Sigma'\left[x_i, x_j\right]$ to obtain the result above.

For a single particle undergoing a one-dimensional random walk with step size κ, each step gives a displacement $h = \pm\kappa$. After n steps, the MSD, x_n, is given by

$$\begin{aligned} x_n &= \left[\sum_i^n h_i\right]^2 \\ &= \sum_i^n \sum_j^n h_i h_j \\ &= \sum_i^n h_i^2 + \sum_i^n \sum_{j\neq i}^n h_i h_j. \end{aligned} \tag{SI.2}$$

The expected MSD in the long-time limit, $\mathbb{E}(x_n) = \langle x_n \rangle$, is obtained by averaging over all permutations of h_i and h_j:

$$\langle x_n \rangle = \sum_i^n \left\langle h_i^2 \right\rangle + \sum_i^n \sum_{j\neq i}^n \left\langle h_i h_j \right\rangle. \tag{SI.3}$$

For a random walk, the second term averages to zero for all h_i and h_j, and

$$\begin{aligned} \langle x_n \rangle &= \sum_i^n \left\langle h_i^2 \right\rangle \\ &= n\kappa^2. \end{aligned} \tag{SI.4}$$

Hence the expected value for the mean-squared displacement increases linearly with the number of steps taken.

The variance in the observed MSD, $\sigma^2[x_n]$, is given by the standard statistical formula

$$\sigma^2[x_n] = \left\langle \left[x_n - \langle x_n \rangle\right]^2 \right\rangle, \tag{SI.5}$$

which can be expanded as

$$\begin{aligned} \sigma^2[x_n] &= \left\langle x_n^2 \right\rangle - 2\left\langle x_n \right\rangle \left\langle x_n \right\rangle + \left\langle x_n \right\rangle^2, \\ &= \left\langle x_n^2 \right\rangle - \left\langle x_n \right\rangle^2. \end{aligned} \tag{SI.6}$$

[*] andrew.mccluskey@bristol.ac.uk
[†] b.j.morgan@bath.ac.uk

The first term can be expanded in terms of displacements h as

$$\left\langle x_n^2 \right\rangle = \left\langle \sum_i^n \sum_j^n \sum_k^n \sum_l^n h_i h_j h_k h_l \right\rangle, \tag{SI.7}$$

which can be simplified by noting that h_i, h_j, h_k, and h_l are uncorrelated when $i \neq j \neq k \neq l$, and the only terms that contribute to the average are those where $h_i h_j h_k h_l$ is guaranteed to be non-zero:

(a) $i = j = k = l$;

(b) $(i = j) \neq (k = l)$;

(c) $(i = k) \neq (j = l)$;

(d) $(i = l) \neq (j = k)$.

From (a) we obtain

$$\left\langle \sum_i^n h_i^4 \right\rangle = n\kappa^4, \tag{SI.8}$$

and from (b), (c), and (d), which are equivalent, we obtain

$$\left\langle \sum_i^n \sum_j^n h_i^2 h_j^2 \right\rangle = (n\kappa^2)^2 = n^2\kappa^4. \tag{SI.9}$$

This gives

$$\left\langle x_n^2 \right\rangle = (3n^2 + n)\kappa^4, \tag{SI.10}$$

which, in the limit $n \to \infty$, approaches

$$\left\langle x_n^2 \right\rangle = 3n^2\kappa^4. \tag{SI.11}$$

Combining this result with Eqn. SI.4, we can express the variance in the mean-squared displacement as

$$\sigma^2[x_n] = 3n^2\kappa^4 - n^2\kappa^4 = 2n^2\kappa^4, \tag{SI.12}$$

i.e., $\sigma^2[x_n]$ increases quadratically with the number of steps taken, or, equivalently, with time.

Eqn. SI.12 gives the variance of the mean squared displacement for a single particle considering a single time-origin. We can obtain improved statistics by averaging over statistically equivalent observed squared displacements (see Eqn. 2 in the main text), which can be achieved by averaging over mobile particles or by averaging over time origins. This averaging over equivalent observations reduces the variance in the observed MSD to

$$\sigma^2[x_n] = \frac{2n^2\kappa^4}{N'_n}, \tag{SI.13}$$

where N'_n is the total number of statistically independent (non-overlapping) squared displacements that contribute to x_i. In the long-time limit, N'_n is given by the product of the number of mobile particles and the number of numerically-independent sub-trajectories of length i in our simulation trajectory. Note that N'_n considers numerically-independent sub-trajectories, since mutually overlapping time-windows give correlated squared displacements. Where overlapping time-windows are used, Eqn. SI.13 approximates the observed variance (the variance for overlapping and non-overlapping samples are not the same) with accuracy that increases as a function of time-interval length. This leads to the approximation in Eqn. 7.

The results for a one-dimensional lattice above (Eqns. SI.4 & SI.13) can be extended to a d-dimensional lattice, to give

$$\langle x_n \rangle_d = \sum^d \frac{n\kappa^2}{d} = n\kappa^2, \tag{SI.14}$$

with variance

$$\sigma^2[x_n]_d = \sum^d \frac{2n^2\kappa^4}{d^2 N'_n} = \frac{2n^2\kappa^4}{d N'_n}, \tag{SI.15}$$

Because each step is equally likely to move a particle along each of the d dimensions, the term n in Eqns. SI.4 & SI.13 is replaced here with n/d.

The analysis above can be extended to consider the covariance between two different numbers of steps, n and $n+m$, in the random walk where the expected MSDs will be

$$\begin{aligned} \langle x_n \rangle &= n\kappa^2; \\ \langle x_{n+m} \rangle &= (n+m)\kappa^2. \end{aligned} \tag{SI.16}$$

The covariance between these is defined as

$$\Sigma\left[x_n, x_{n+m}\right] = \left\langle \left[x_n - \langle x_n \rangle\right] \left[x_{n+m} - \langle x_{n+m} \rangle\right] \right\rangle, \tag{SI.17}$$

which can be expanded as

$$\begin{aligned} \Sigma\left[x_n, x_{n+m}\right] = \big\langle & x_n x_{n+m} - x_n \left\langle x_{n+m} \right\rangle \\ & - \left\langle x_n \right\rangle x_{n+m} + \left\langle x_n \right\rangle \left\langle x_{n+m} \right\rangle \big\rangle, \end{aligned} \tag{SI.18}$$

and then reformulated to give

$$\Sigma\left[x_n, x_{n+m}\right] = \left\langle x_n x_{n+m} \right\rangle - \left\langle x_n \right\rangle \left\langle x_{n+m} \right\rangle, \tag{SI.19}$$

where

$$\begin{aligned} \left\langle x_n \right\rangle \left\langle x_{n+m} \right\rangle &= x_n x_{n+m} \\ &= n\kappa^2 (n+m)\kappa^2 \\ &= n(n+m)\kappa^4 \end{aligned} \tag{SI.20}$$

and, by analogy to Eqn. SI.7,

$$\left\langle x_n x_{n+m} \right\rangle = \left\langle \sum_i^n \sum_j^n \sum_k^{n+m} \sum_l^{n+m} h_i h_j h_k h_l \right\rangle, \tag{SI.21}$$

which we can rewrite as

$$\begin{aligned}\langle x_n x_{n+m}\rangle = \Bigg\langle & \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} h_i h_j h_k h_l \\ & + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=n+1}^{n+m} h_i h_j h_k h_l \\ & + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=n+1}^{n+m}\sum_{l=1}^{n} h_i h_j h_k h_l \\ & + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=n+1}^{n+m}\sum_{l=n+1}^{n+m} h_i h_j h_k h_l \Bigg\rangle .\end{aligned} \tag{SI.22}$$

The second and third terms in Eqn. SI.22 tend to zero as there is an equal probability of positive and negative displacements. This reduces Eqn. SI.22 to

$$\begin{aligned}\langle x_n x_{n+m}\rangle = & \Bigg\langle \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} h_i h_j h_k h_l \Bigg\rangle \\ & + \Bigg\langle \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=n+1}^{n+m}\sum_{l=n+1}^{n+m} h_i h_j h_k h_l \Bigg\rangle ,\end{aligned} \tag{SI.23}$$

and using Eqn. SI.11 gives

$$\langle x_n x_{n+m}\rangle = 3n^2\kappa^4 + \Bigg\langle \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=n+1}^{n+m}\sum_{l=n+1}^{n+m} h_i h_j h_k h_l \Bigg\rangle . \tag{SI.24}$$

We can rewrite this as

$$\langle x_n x_{n+m}\rangle = 3n^2\kappa^4 + \Bigg\langle \sum_{i=1}^{n}\sum_{j=1}^{n} h_i h_j \Bigg\rangle \Bigg\langle \sum_{k=n+1}^{n+m}\sum_{l=n+1}^{n+m} h_k h_l \Bigg\rangle , \tag{SI.25}$$

where the following holds,

$$\begin{aligned}\langle x_n x_{n+m}\rangle &= 3n^2\kappa^4 + n\kappa^2 m\kappa^2 \\ &= 3n\kappa^4 + nm\kappa^4 .\end{aligned} \tag{SI.26}$$

Putting this result into Eqn. SI.19 allows the covariance to be written as

$$\begin{aligned}\Sigma'\left[x_n, x_{n+m}\right] &= 3n^2\kappa^4 + nm\kappa^4 - n(n+m)\kappa^4 \\ &= 3n^2\kappa^4 - n^2\kappa^4 = 2n^2\kappa^4 ,\end{aligned} \tag{SI.27}$$

where we use the Σ' notation to identify that this is in the long-time limit.

In this case, the covariance depends only on the number of overlapping points, n, between the two time intervals. We can rationalise this by noting that for a random walk any numerically-independent points will be completely uncorrelated and therefore have a covariance of 0. Similar to the case for the variance, the covariance derived in Eqn. SI.27 is that for a single particle at a single time origin. The number of independent observed squared displacements for a given covariance should be the minimum number of shared independent observed squared displacements between the two time intervals, which is N'_{n+m}. Therefore, the covariance, scaled by the number of contributing independent observations, in the long-time limit, is

$$\Sigma'\left[x_n, x_{n+m}\right] = \frac{2n^2\kappa^4}{N'_{n+m}} . \tag{SI.28}$$

Similar to the MSD and the variance, the covariance can be written for d-dimensions as

$$\Sigma'\left[x_n, x_{n+m}\right] = \frac{2n^2\kappa^4}{dN'_{n+m}} . \tag{SI.29}$$

The covariance can be calculated directly from the variance by recognising that both depend on the number of overlapping points, n, as follows

$$\Sigma'\left[x_n, x_{n+m}\right] = \sigma^2[x_n]\frac{N'_n}{N'_{n+m}} . \tag{SI.30}$$

This is then rewritten in terms of i and j to give, Eqn. SI.1.

Using the equivalence of $2dD^*t \equiv n\kappa^2$ [50], Eqns. SI.4 & SI.1 can be rewritten in terms of t (or t_1 and t_2) and the diffusion coefficient, for any dimensionality of lattice random walk,

$$x(t) = 2dD^*t, \tag{SI.31}$$

and

$$\Sigma'\left[x(t_1), x(t_2)\right] = 8d(D^*)^2 t_1^2 \frac{N'(t_2)}{N'(t_1)}, \quad \forall\, t_1 \leq t_2 . \tag{SI.32}$$

SI.II: Comparison With Block Averaging Approach

Eqn. 7 enables the approximate estimation of the variance in x_i, as mentioned in the main text, it is also possible to obtain this from a block averaging approach [25, 26]. Fig. SI.1 compares the D^* estimate distributions using variances from approximated by Eqn. 7 with those from the block averaging approach (using the PYBLOCK Python package [51]) both with the fitting of the variances to Eqn. 8 and without. When the fitting is not performed, the resulting covariance matrix is numerically unstable, leading to estimated values of D^* at extreme values. While when the fitting is performed and the blocking-estimated variances as used, there is no improvement in the estimation of D^* and the resulting distribution of estimates in the variance of $p(D^*)$ is broader than when Eqn. 7 is used.

SI.III: Bias and skew in $p(\widehat{\sigma}^2[\widehat{D}^*])$

In the main manuscript, we present results for a set of 4096 3D-lattice random walk simulations, each consisting of 128 particles undergoing 128 steps (Fig. 5).

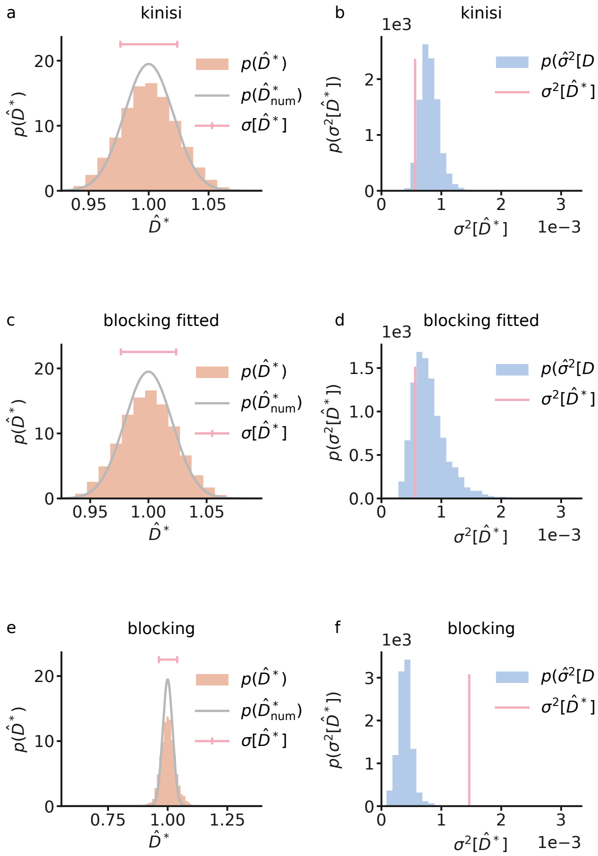

FIG. SI.1. Comparison of the distribution of $\widehat{D}^*$ (a, c, & e) and the estimated variances (b, d, & f) from 4096 individual random walk simulations using variances approximated by Eqn. 7 (a & b), from block averaging with fitting to Eqn. 8 (c & d), and without fitting (e & f). Note that the bounds in a, c, & e are defined by the extrema of the estimated values, showing the fact that without the fitting to Eqn. 8, the resulting covariance matrix is numerically unstable.

Our approximate Bayesian regression scheme allows us to estimate the variance in $\widehat{D}^*$, denoted as $\widehat{\sigma}^2[\widehat{D}^*]$, that would be obtained over a large number of repeat simulations. This estimate is calculated from the variance of the marginal posterior distribution $p(D^*|\boldsymbol{m})$, which we derive from analysis of a single simulation trajectory. As shown in Fig. 5d, our estimate for the population variance $\sigma^2[\widehat{D}^*]$, obtained from a single simulation, aligns reasonably with the true value. When considering the distribution of estimated variance, $p(\widehat{\sigma}^2[\widehat{D}^*])$, however, we observe a systematic overestimation (bias) of the true value, along with visible skewness.

This bias and skew arise from our use of estimated variances $\widehat{\sigma}^2[x_i]$ when parametrising the model covariance matrix $\boldsymbol{\Sigma'}$. Fig. SI.2 presents equivalent results for $p(\widehat{D}^*)$ and $p(\widehat{\sigma}^2[\widehat{D}^*])$ for the same 4096 individual simulations, but calculated using a numerical covariance matrix, $\boldsymbol{\Sigma}_{\mathrm{num}}$ derived from all 4096 observed MSDs. The resulting distribution $p(\widehat{\sigma}^2[\widehat{D}^*])$ (Fig. SI.2b) is no longer biased or skewed. Furthermore, the distribution $p(\widehat{D}^*)$ agrees even more closely with the numerically converged distribution obtained when combining data from all 4096 simulations (Fig SI.2a), contrasting with the results presented in Fig. 5b, where our approximate Bayesian regression scheme yields a slightly broadened distribution due to the use of the long-time limit in the derivation of the analytical form for $\boldsymbol{\Sigma'}$.

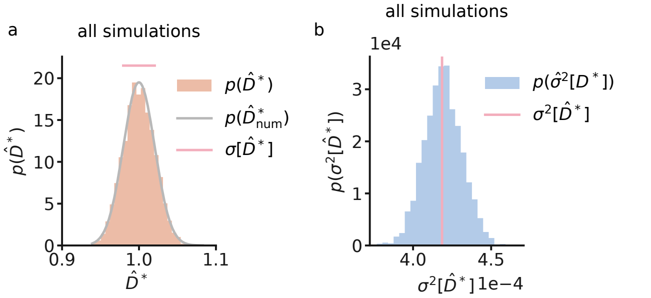

FIG. SI.2. (a) Probability distribution of point-estimates $p(\widehat{D}^*)$ obtained from 4096 individual random-walk simulations, using the numerical covariance matrix $\boldsymbol{\Sigma}_{\mathrm{num}}$. Each simulation has been analysed as in Fig. 2(a) and (b) to yield a single corresponding point estimate $\widehat{D}^*$. The grey line shows the distribution of point estimates, $p(\widehat{D}^*_{\mathrm{num}})$, obtained using Bayesian regression with a mean vector and numerical covariance matrix derived from the complete dataset of all 4096 simulations. The pink horizontal bar shows an interval of one standard deviation in $p(\widehat{D}^*)$. (b) Probability distribution of estimated variances, $\widehat{\sigma}^2[\widehat{D}^*]$, for individual random-walk simulations, using the numerical covariance matrix $\boldsymbol{\Sigma}_{\mathrm{num}}$, compared to the true sample variance (pink vertical line) $\sigma^2[\widehat{D}^*]$.

SI.IV: Evaluation of OLS, WLS, and GLS for LLZO System

In Fig. 2, it is shown that for a 3D lattice random walk the heteroscedastic and correlated nature of the data requires generalised least squares to optimally estimate the true ensemble-average MSD, and therefore estimate D^*. This is also true for the real materials, such as the $\mathrm{Li_7La_3Zr_2O_{12}}$ (LLZO) investigated by classical molecular dynamics simulation in this work (Fig. SI.3). Once again, GLS (or, indeed, the Bayesian regression equivalent, shown in Fig. 6) offers both the most statistically efficient estimation of D^* and is required to obtain an accurate estimate of the statistical uncertainty in D^* from a single simulation.

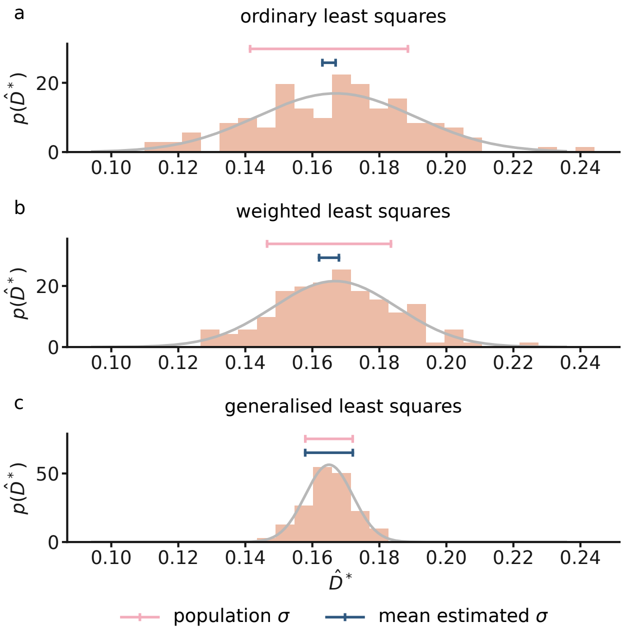

FIG. SI.3. Example distributions of estimated self-diffusion coefficients, $\widehat{D}^*$, calculated using (a) ordinary least squares (OLS), (b) weighted least squares (WLS), and (c) generalised least squares (GLS), from MSD data from 512 effective simulations of LLZO of ~25 ps with 56 lithium ions. In each panel, the grey curve shows the best-fit normal distribution for the simulation data, the upper horizontal bar shows the standard deviation of this distribution, and the lower horizontal bar shows the average estimated standard distribution given by the analytical expression for $\sigma[p(\widehat{D}^*)]$ for each regression method.