

Communication: A multiscale Bayesian inference approach to analyzing subdiffusion in particle trajectories

Konrad Hinsén and Gerald R. Kneller

Citation: *The Journal of Chemical Physics* **145**, 151101 (2016); doi: 10.1063/1.4965881

View online: <http://dx.doi.org/10.1063/1.4965881>

View Table of Contents: <http://scitation.aip.org/content/aip/journal/jcp/145/15?ver=pdfcov>

Published by the [AIP Publishing](#)

Articles you may be interested in

[Subdiffusion of proteins and oligomers on membranes](#)

J. Chem. Phys. **137**, 175102 (2012); 10.1063/1.4764305

[Communication: Consistent picture of lateral subdiffusion in lipid bilayers: Molecular dynamics simulation and exact results](#)

J. Chem. Phys. **135**, 141105 (2011); 10.1063/1.3651800

[A Bayesian statistics approach to multiscale coarse graining](#)

J. Chem. Phys. **129**, 214114 (2008); 10.1063/1.3033218

[Bayesian inference for inverse problems](#)

AIP Conf. Proc. **617**, 477 (2002); 10.1063/1.1477067

[Information theoretic approach to Bayesian inference](#)

AIP Conf. Proc. **617**, 433 (2002); 10.1063/1.1477064



NEW Special Topic Sections

NOW ONLINE
Lithium Niobate Properties and Applications:
Reviews of Emerging Trends

AIP | Applied Physics
Reviews

Communication: A multiscale Bayesian inference approach to analyzing subdiffusion in particle trajectories

Konrad Hinsen^{1,2,a)} and Gerald R. Kneller^{1,2,3,b)}

¹Centre de Biophysique Moléculaire, CNRS, Rue Charles Sadron, 45071 Orléans, France

²Synchrotron SOLEIL, L'Orme des Merisiers, 91192 Gif-sur-Yvette, France

³Université d'Orléans, Chateau de la Source, Ave. du Parc Floral, 45067 Orléans, France

(Received 12 September 2016; accepted 7 October 2016; published online 20 October 2016)

Anomalous diffusion is characterized by its asymptotic behavior for $t \rightarrow \infty$. This makes it difficult to detect and describe in particle trajectories from experiments or computer simulations, which are necessarily of finite length. We propose a new approach using Bayesian inference applied directly to the observed trajectories sampled at different time scales. We illustrate the performance of this approach using random trajectories with known statistical properties and then use it for analyzing the motion of lipid molecules in the plane of a lipid bilayer. *Published by AIP Publishing*. [<http://dx.doi.org/10.1063/1.4965881>]

Anomalous diffusion has been observed in a wide range of physical systems and has also been the subject of theoretical investigations.^{1–4} The term “anomalous” refers to a non-linear (asymptotic) growth of the mean square displacement (MSD) of the diffusing particles with time, and thus to a deviation from Einstein’s classical diffusion model.⁵ Most anomalous diffusion processes display subdiffusion, where the MSD grows like $2D_\alpha t^\alpha$, with $0 < \alpha < 1$ for large t , and D_α is the fractional diffusion constant.

A central problem in the study of anomalous diffusion is the estimation of the parameters D_α and α from experimental or simulated particle trajectories. Since any observed trajectory consists of a finite set of points, it is not even obvious that it makes sense to discuss its asymptotic behavior for $t \rightarrow \infty$. The most common approach is to compute a time-averaged MSD over all position pairs with identical lag time t , complemented by a statistical average over multiple trajectories that are assumed to be equivalent. Since there are fewer position pairs for large values of t than for small ones, the statistical quality of the MSD decreases with t although it is exactly the large- t data that are most relevant for the asymptotic behavior. In the case of normal diffusion, this approach works reasonably well because the characteristic linear growth of the MSD with time is easy to spot visually on a plot, but for anomalous diffusion, the identification of a potential asymptotic regime is not reliable.

Another approach is to assume a concrete model for the joined probability distribution functions of the sampled trajectories and use inference techniques to estimate the model parameters. There are several fundamentally different physical and mathematical models that give rise to the same $2D_\alpha t^\alpha$ form of the MSD for long times, which raises the question of how to choose the most appropriate one for a given dataset. Much recent work on particle trajectory analysis for subdiffusion has been dedicated to this question.⁶ Moreover, none of these

models can be expected to be valid on all time scales. Diffusive dynamics gives way to ballistic motion at time scales smaller than the atomic or molecular collision times, meaning that for sufficiently small t , the MSD always grows as t^2 .

In this work, we address these issues with a new data analysis technique that is based on two core ideas: (1) a multiscale approach in which the data are analyzed at different sampling time steps and (2) the use of Bayesian inference to obtain model parameters and their uncertainties directly from observed particle positions, without the computation of intermediate quantities such as an MSD. We will limit ourselves to descriptions of subdiffusion in molecular systems, but our method is not limited to this type of diffusion. In the first step, we explore the performance of our method for ideal trajectory ensembles generated randomly from a precisely known model. This allows us to see how the inference process converges as more input data are added, and to understand how it can be used to identify the asymptotic behavior in the presence of distinct short-time dynamics. In the second step, we analyze the lateral diffusion of lipid molecules in a computer simulation of a hydrated lipid bilayer.

The starting point of our approach is an observed particle trajectory $X(i\delta t)$, $i = 0, \dots, K$. From this trajectory we extract subsamples of length L at sampling time step $\Delta t = s\delta t$, $s = 1, 2, \dots$, i.e., the points $X_j = X(j\Delta t)$, $j = 0, \dots, L$. We interpret these trajectories as realizations of an L -step stochastic process that is defined by its probability distribution

$$P(X_0, \dots, X_L | \phi_1, \dots, \phi_M), \quad (1)$$

where ϕ_1, \dots, ϕ_M are the numerical parameters of the model whose values we wish to estimate. For our intended application to molecular diffusion, we need a model for a continuous diffusion process that is sampled at arbitrarily chosen discrete time values. This excludes models such as the Continuous Time Random Walk (CTRW),⁷ which describes a hopping process with random waiting times, but also mathematical models such as the autoregressive fractionally integrated moving average (ARFIMA) process,^{8,9} which assume from

^{a)}konrad.hinsen@cnrs.fr

^{b)}gerald.kneller@cnrs.fr

the start a discrete time series and thus cannot accommodate different choices of sampling times. The simplest continuous stochastic process that gives rise to anomalous diffusion is *fractional Brownian Motion* (fBM). Its probability distribution is given by^{8,9}

$$P(\mathbf{X}|\alpha, D_\alpha) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2} \mathbf{X} \cdot \Sigma^{-1} \cdot \mathbf{X}\right), \quad (2)$$

where $\mathbf{X} = (X_0, \dots, X_L)$ and

$$\Sigma_{i,j} = D_\alpha \Delta t^\alpha (i^\alpha + j^\alpha + |i-j|^\alpha) \quad (3)$$

are the components of the covariance matrix Σ of the process. For $\alpha = 1$, fBM reduces to standard Brownian motion, whereas $0 < \alpha < 1$ describes subdiffusion and $1 < \alpha < 2$ superdiffusion. Although the choice of the fBM model is motivated by its capacity to describe asymptotic subdiffusion, we use it here as an analysis tool for *all* time scales present in the observed trajectories.

Bayesian inference is based on the idea of introducing a probability distribution not only for the data but also for the parameters of the model describing them.¹⁰ This distribution is taken to be a description of the knowledge one has about these parameters, in contrast to the more common use of probabilities in statistical physics for describing natural phenomena by random processes. Our data are equidistantly sampled trajectories \mathbf{X} , which are treated as “physical” random variables, and the model parameters α and D_α , which are treated as “informational” random variables, leading to an interpretation of Eq. (2) as a conditional probability distribution.

The starting point of Bayesian inference is a *prior probability distribution* $P_0(\alpha, D_\alpha)$, which describes the prior information one has about the parameters, before exploiting any observations. We use a uniform distribution for α in the interval $(0, 2)$ and a uniform distribution for D_α in the interval $(0, D_{\max})$ with a large but finite D_{\max} to ensure normalizability. Each observed L -step trajectory $\mathbf{X}^{(j)}$ (for $j = 1, \dots, N$) adds information about α and D_α . The posterior probability distribution that integrates the information from N trajectories is derived via Bayes’ theorem and is given by

$$P_N(\alpha, D_\alpha) = \prod_{j=1}^N \frac{P(\mathbf{X}^{(j)}|\alpha, D_\alpha)}{\int d\alpha \int dD_\alpha P(\mathbf{X}^{(j)}|\alpha, D_\alpha)} P_0(\alpha, D_\alpha), \quad (4)$$

assuming that the trajectories are independent realizations of the fBM process. The factor $P(\mathbf{X}^{(j)}|\alpha, D_\alpha)$ is called the likelihood of the observation $\mathbf{X}^{(j)}$ and is given by Eq. (2). It describes the information about α and D_α contributed by a single trajectory. The denominator in Eq. (4) is not important for our application, it can be regarded as a normalization factor for $P_N(\alpha, D_\alpha)$.

Bayesian inference has several advantages over the traditional approaches for studying diffusion processes. First of all, it yields probability distributions for the parameters and thus information about their uncertainties, in addition to estimates for their values. Second, it allows a comparative evaluation of the quality of different models through the computation of their Bayes factor,¹¹ although we will not

discuss this aspect here. Third, it makes all the assumptions that have an impact on the results explicit.

Although we have performed simultaneous estimation of the parameters α and D_α , yielding the two-dimensional posterior distribution from Eq. (4), we report only single-parameter Bayesian inference on α in this communication. The two parameters turn out to be only weakly correlated, with $2D_\alpha \Delta t^\alpha$ being narrowly distributed around the straightforward estimate,

$$(2D_\alpha \Delta t^\alpha)_{est} = \left\langle (X_{j+1}^{(k)} - X_j^{(k)})^2 \right\rangle_{j,k}. \quad (5)$$

In the following, we use this estimate for D_α , allowing us to concentrate on the more difficult and more important inference for α .

Before applying Bayesian inference to the estimation of fBM parameters for observed particle trajectories, we must develop an understanding of how it is affected by the quality of the input data. To this end, we analyze synthetic trajectories whose statistical properties are known by construction. In the first step, we generate these trajectories as numerical samples of the distribution given by Eq. (2), choosing $2D_\alpha \Delta t^\alpha = 1$ for simplicity and $\alpha = 0.6$ because that is the order of magnitude we will find for subdiffusion in our lipid bilayer simulations. We thus study an ideal case in which the fBM model is known to be exact.

Fig. 1 shows how the posterior distribution narrows as more trajectories are supplied to the inference algorithm. For short trajectories (10 steps, upper panel), a single trajectory provides little information, the distributions being wide and their maxima often far away from the known value for α . Several hundred trajectories are required before the maximum of the distribution stabilizes near the correct value. As is to be expected, longer trajectories (100 steps, lower panel) yield individual parameter distributions that are narrower and closer to the true value. As a consequence, a reliable estimate for α can be obtained from about 200 input trajectories. We use trajectory lengths of $L = 100$ everywhere in the rest of this work.

As we explained earlier, trajectories for any real physical system cannot be expected to be exactly described by the fBM model, in particular for short times, when the dynamics of the underlying microscopic processes become visible. In order to study the impact of a different short-time behavior on our inference approach, we analyze another set of synthetic trajectories, generated from a modification of the fBM model that has different short-time properties.

To construct such a model, we consider the increments of the fBM process, defined by $\Delta X_i = X_{i+1} - X_i$. The increments form a Gaussian stochastic process as well, which is often called fractional Brownian noise. Its covariance matrix is given by

$$\begin{aligned} \Sigma_{i,i}^{(inc)} &= 1, \\ \Sigma_{i,i+k}^{(inc)} &= D_\alpha \Delta t^\alpha [(k+1)^\alpha - 2k^\alpha + (k-1)^\alpha] \text{ for } k \neq 0. \end{aligned}$$

For any Gaussian process, $\Sigma_{i,j}$ and $\Sigma_{i,j}^{(inc)}$ are related by

$$\Sigma_{i,j} = \sum_{k=1}^i \sum_{l=1}^j \Sigma_{k,l}^{(inc)}. \quad (6)$$

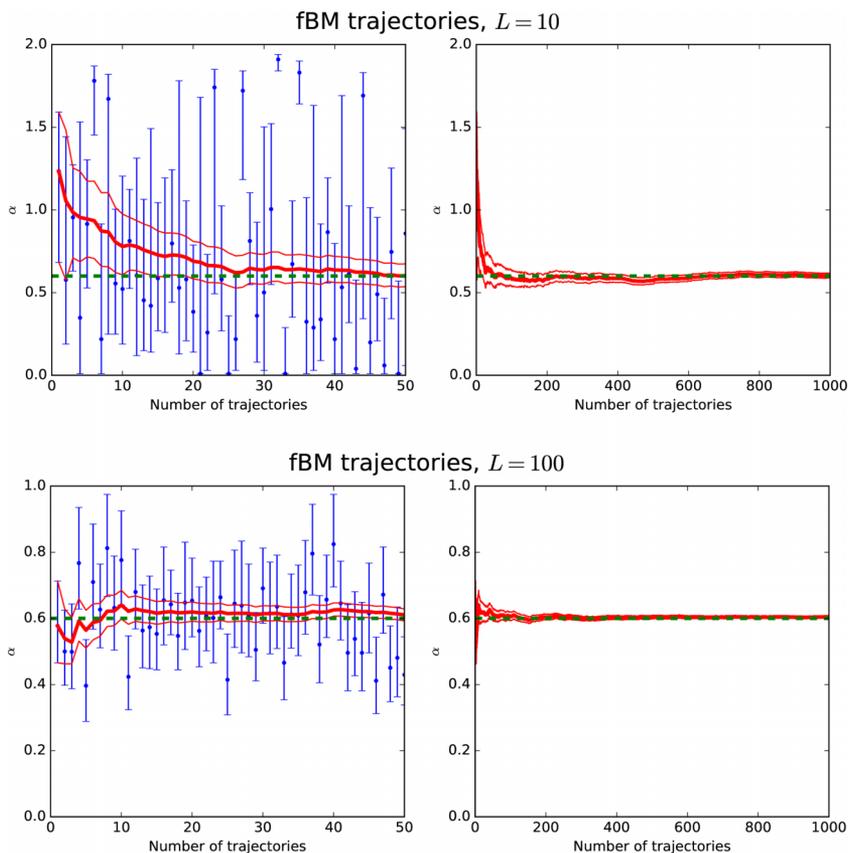


FIG. 1. Convergence of the inference process as a function of the number of trajectories. Upper panel: short trajectories ($L = 10$). Lower panel: long trajectories ($L = 100$). On the left, each blue bar describes the likelihood $P(\alpha)$ for a single input trajectory. The dot is the location of the maximum, which is the maximum-likelihood estimate for α . The ends of the bar are the half-maximum points. The red lines provide the same information for the cumulative posterior distribution. The green dashed line shows the value of α used for generating the trajectories.

Contrary to the fBM process itself, the increment process is stationary, and therefore its autocorrelation function $\langle \Delta X_i \Delta X_j \rangle = \Sigma_{i,j}^{(inc)}$ depends only on the time lag $(i - j)\Delta t$. Its physical meaning is very similar to that of a velocity autocorrelation function in atomic or molecular liquids. In fact, if we assume the existence of a microscopic velocity $v(t)$ such that $\Delta X(t) = \int_0^{\Delta t} d\tau v(t + \tau)$, then $\langle \Delta X_i \Delta X_j \rangle$ is given by the microscopic velocity autocorrelation function convoluted with a triangular weight function of width $2\Delta t$. The increment autocorrelation function has been used by Jeon *et al.*¹² to characterize subdiffusion in lipid bilayers, and their analysis shows clear deviations from fBM for short lag times.

A straightforward way to modify the short-time behavior of a process is to add a term $\Sigma_{i,j}^{(st)}$ to the increment covariance matrix that is non-zero only for small $|i - j|$. According to Eq. (6), it is sufficient to respect the condition $\sum_k \Sigma_{i,i+k}^{(st)} = 0$

to ensure that the asymptotic behavior of the process remains unaltered. We choose a form with only two non-zero terms, $\Sigma_{i,i+1}^{(st)} = -\Sigma_{i,i+2}^{(st)} = 1/2(\Sigma_{i,i+1}^{(inc)} + \Sigma_{i,i+2}^{(inc)})$, i.e., we replace the points at $k = 1$ and $k = 2$ by their average. This leads to a wider and shallower minimum, as shown in the leftmost panel of Fig. 2, which shows the increment autocorrelation functions for the fBM process and for our modified process. The middle panel of Fig. 2 shows the MSD for the standard and modified fBM processes. There is a pronounced difference in the short-time behavior that persists over a much longer time range than for the increment correlations. However, the asymptotic subdiffusive behavior of the two processes is the same.

We now use the modified process for generating synthetic trajectories of $L = 100$ steps and perform parameter inference on these trajectories using the standard fBM process from

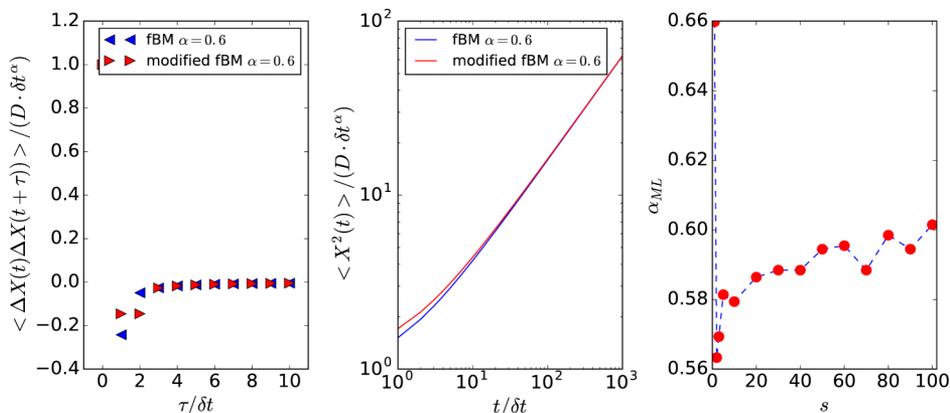


FIG. 2. Left: the increment correlation of the standard fBM process (blue) and our short-time modification (red). Middle: the mean-square displacements for the standard and modified fBM processes. Although the modification changes the correlations of the increments only up to $2\Delta t$, the processes differ significantly over a much longer time range. Right: maximum-likelihood estimates for α obtained from 500 trajectories of 100 steps generated from the fBM model with a modified short-time behavior, sampled with a time step of $s\delta t$.

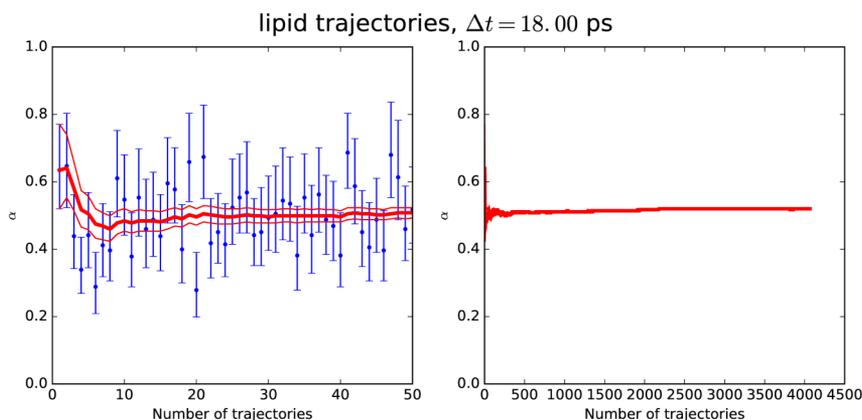


FIG. 3. Convergence of the inference process for lipid center-of-mass diffusion, as a function of the number of trajectories.

Eq. (2) as a model. The inference process converges with no sign of a problem (see figures in the [supplementary material](#)), but yields an estimate for α of 0.66, which is rather far from the input value of 0.6. This shows that short-time effects cannot simply be neglected. It also illustrates a well-known limitation of uncertainty estimation by Bayesian inference, which captures the uncertainty due to noise in the input data, but not the uncertainty due to a possible mismatch between the data and the chosen model (fBM in our case). The convergence of the inference process to $\alpha = 0.66$ thus does not permit the conclusion that fBM with $\alpha = 0.66$ is a good description of the data. In particular, it does not mean that the asymptotic diffusive behavior of our trajectories corresponds to $\alpha = 0.66$. In fact, we know by construction that its long-time diffusion is described by $\alpha = 0.6$.

However, we can still say that “fBM with $\alpha = 0.66$ ” is a well-defined characterization of the data, since Bayesian inference yields a posterior probability distribution that is sharply peaked around this value. We then address the question of the asymptotic behavior by applying fBM-based inference to our data at different time scales, sampling the trajectories from the modified fBM process at time steps $\Delta t = s\delta t$, for s ranging from 1 to 100. For each s , we generate trajectories with $L = 100$ points, such that the total amount of data received by the inference procedure is always the same. The length of each trajectory in physical time is $sL\delta t$. With increasing s , we thus include less short-time information and more long-time information. As Fig. 2 shows, this change of sampling time yields estimates for α that converge towards the input value of 0.6 around $s = 100$, i.e., for sampling times that are 50 times longer than the short-time perturbation we added to our model.

We now apply our Bayesian inference method to computer simulation trajectories for a hydrated lipid bilayer. The simulation data we use have been described and analyzed earlier.^{13,14} The simulated system consists of 2033 POPC (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine) molecules and 57 952 water beads (equivalent to 231 808 water molecules), using the coarse-grained MARTINI force field. The simulation was performed in an NVT ensemble at $T = 320$ K. Two trajectory files were generated during the simulation run: (1) a short-time trajectory of length 300 ps, sampled every 0.03 ps and (2) a long-time trajectory of length 600 ns, sampled every 18 ps. Extracts of these trajectories containing the center-of-mass trajectories of the

lipid molecules have been published^{15,16} and are the basis for the present analysis.

We use only the x and y components of the center-of-mass trajectories, discarding the z coordinate which is perpendicular to the membrane plane. Assuming that the x and y components for all molecules are statistically equivalent and independent, we thus have 4066 single-coordinate input trajectories for inference. We use subsets consisting of $L = 100$ steps, with different sampling time steps $\Delta t = s\delta t$.

Fig. 3 shows that the inference procedure converges much like for the synthetic trajectories. The dependence of the maximum-likelihood estimate for α on the sampling step size is shown in Fig. 4. On a very short time scale, we find $\alpha \approx 2$, i.e., we see the nearly ballistic motion typical for times shorter than the mean time between collisions. Increasing the sampling step size, we see a rapid decrease of α , ending in a stable plateau with $\alpha \approx 0.55$ that extends over three decades. In Fig. 4 we also show the values of $2D_\alpha\Delta t^\alpha$ estimated using Eq. (5). Its behavior for large sampling times also

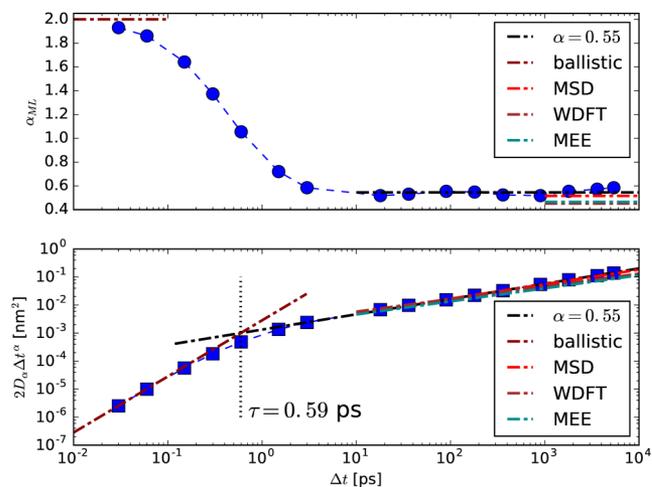


FIG. 4. Maximum-likelihood estimates for α (upper panel) and $2D_\alpha\Delta t^\alpha$ (lower panel) describing lipid center-of-mass diffusion in a membrane. Each estimate was obtained using 4066 single-coordinate trajectories of 100 steps, with varying sampling step sizes Δt . The fBM model with $\alpha = 0.55$ and $D_\alpha = 0.66 \cdot 10^{-4} \text{ nm}^2/\text{ps}^\alpha$ is shown in black, whereas the short-time ballistic motion ($\alpha = 2$, $D_\alpha = \frac{k_B T}{2m}$) is shown in red. The intersection of these pure power-law behaviors defines a characteristic time scale τ for the diffusion process. For comparison, we also show the fBM parameterizations from Ref. 14.

indicates a stable asymptotic fBM regime for $\Delta t > \approx 10$ ps. For short times, we see a ballistic regime that remains a good approximation up to $\Delta t < \approx 0.2$ ps. The transition between the two regimes can be used to define a time scale τ for the diffusion process through the relation $D_2\tau^2 = D_\alpha\tau^\alpha$, yielding a value of $\tau = 0.59$ ps for our trajectories. Up to a factor of $2^{1/(2-\alpha)}$, this time scale is the same as the time scale $\tau_{\text{VACF}} = (D_\alpha/\langle|\mathbf{v}|^2\rangle)^{1/(2-\alpha)}$ introduced earlier by Kneller *et al.*¹⁷

For comparison, we have added to Fig. 4 asymptotic behavior of the three fBM model fits from Ref. 14. Their “MSD” fit, which fits the long-time part of a computed MSD to its fBM form, yields parameters that are very similar to those found in this work. The other two approaches (WDFT = Windowed Discrete Fourier Transform, MEE = Maximum Entropy Estimation) fit the frequency spectrum of the velocity autocorrelation function to its known functional form for the fBM process and lead to noticeably smaller values for α .

This analysis of computer simulations of lipid dynamics shows that Bayesian inference on multiple time scales is a powerful method to gain insight into subdiffusion processes. In comparison to the most common approach of fitting long-time MSD data to the asymptotic $2D_\alpha t^\alpha$ form, it has several advantages. The application of Bayesian inference directly to the observed particle positions makes it possible to focus on a small and well-defined time scale range. In our application, we chose a time scale range from $[\Delta t \dots 100\Delta t]$, for many different values of Δt . The only arbitrary parameter choice is $L = 100$, but different choices lead to similar conclusions (see the [supplementary material](#)). Furthermore, monitoring the narrowing of the posterior distribution as more trajectory data were injected into the procedure allowed us to evaluate the statistical quality of our dataset, which is always a problematic

aspect in MSD-based fits. We can therefore have confidence in our observation of a stable $\alpha \approx 0.55$ for time scales ranging from 10 to 10.000 ps. For an analysis of observed trajectories, which are necessarily of finite length, this is as close as one can get to identifying asymptotic long-time behavior. We note, however, that this conclusion could only be reached because our simulation trajectories spanned an exceptionally long time range.

The complete source code of our analysis software and additional figures are available as the [supplementary material](#) to this article.

¹*Anomalous Transport: Foundations and Applications*, edited by R. Klages, G. Radons, and I. Sokolov (Wiley-VCH Verlag, Weinheim, Germany, 2009).

²I. M. Sokolov, *Soft Matter* **8**, 9043 (2012).

³R. Metzler, J.-H. Jeon, A. G. Cherstvy, and E. Barkai, *Phys. Chem. Chem. Phys.* **16**, 24128 (2014).

⁴F. Höfling and T. Franosch, *Rep. Prog. Phys.* **76**, 046602 (2013).

⁵A. Einstein, *Ann. Phys.* **322**, 549 (1905).

⁶Y. Meroz and I. M. Sokolov, “A toolbox for determining subdiffusive mechanisms,” *Phys. Rep.* **573**, 1 (2015).

⁷H. Scher and E. W. Montroll, *Phys. Rev. B* **12**, 2455 (1975).

⁸J. Beran, Y. Feng, S. Ghosh, and R. Kulik, *Long-Memory Processes* (Springer Verlag, Berlin, Heidelberg, 2013).

⁹J. Beran, *Statistics for Long-Memory Processes* (Chapman and Hall, 1994).

¹⁰E. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003).

¹¹R. E. Kass and A. E. Raftery, *J. Am. Stat. Assoc.* **90**, 773 (1995).

¹²J.-H. Jeon, H. M.-S. Monne, M. Javanainen, and R. Metzler, *Phys. Rev. Lett.* **109**, 188103 (2012).

¹³S. Stachura and G. R. Kneller, *Mol. Simul.* **40**, 245 (2014).

¹⁴S. Stachura and G. R. Kneller, *J. Chem. Phys.* **143**, 191103 (2015).

¹⁵S. Stachura and G. R. Kneller, *Zenodo* (2016).

¹⁶S. Stachura and G. R. Kneller, *Zenodo* (2016).

¹⁷G. R. Kneller, K. Baczynski, and M. Pasenkiewicz-Gierula, *J. Chem. Phys.* **135**, 141105 (2011).