

Non-Markov models of single-molecule dynamics from information-theoretical analysis of trajectories

Kevin Song¹, Raymond Park^{2, 3}, Atanu Das^{4, 5}, Dmitrii E. Makarov^{6, 7, *} and Etienne Vouga¹

¹Department of Computer Science, University of Texas at Austin, Austin, Texas 78712, USA.

²McKetta Department of Chemical Engineering, University of Texas at Austin, Austin, Texas 78712, USA.

³Department of Mathematics, University of Texas at Austin, Austin, Texas 78712, USA.

⁴Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pune, Maharashtra 411 008, India.

⁵Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201 002, India.

⁶Department of Chemistry, University of Texas at Austin, Austin, Texas 78712, USA.

⁷Oden Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, Texas 78712, USA.

Abstract.

Whether single-molecule trajectories, observed experimentally or in molecular simulations, can be described using simple models such as biased diffusion is a subject of considerable debate. Memory effects and anomalous diffusion have been reported in a number of studies, but directly inferring such effects from trajectories, especially given limited temporal and/or spatial resolution, has been a challenge. Recently we proposed that this can be achieved with information-theoretical analysis of trajectories, which is based on the general observation that non-Markov effects make trajectories more predictable and thus more “compressible” by lossless compression algorithms. Toy models where discrete molecular states evolve in time were shown to be amenable to such analysis, but its application to continuous trajectories presents a challenge: the trajectories need to be digitized first, and digitization itself introduces non-Markov effects that depend on the specifics of how trajectories are sampled. Here we develop a milestone-based method for information-theoretical analysis of continuous trajectories and show its utility in application to Markov and non-Markov models and to trajectories obtained from molecular simulations.

*makarov@cm.utexas.edu

1. Introduction.

Experimental single-molecule studies report on the time evolution of molecular degrees of freedom. Experimental observables are low-dimensional; the dynamics of such low-dimensional projections of high-dimensional molecular motion are known to be complex if not intractable (in practice)¹. Two key challenges in the field, then, are to (1) construct adequate models describing the dynamics of experimental observables directly from data and (2) to learn about the underlying molecular motions from the observed low-dimensional projections (aka experimental signals).

The progress toward the first of these goals has so far mostly consisted of fitting/modeling data with postulated dynamical models of increasing complexity. These range from diffusive models of barrier crossing²⁻⁶ or random walks⁷ to - more recently - models with “hidden” states⁸⁻¹¹, generalized Langevin equation with memory¹²⁻¹⁸, non-Markov master equations^{19, 20}, and other models of non-Markov dynamics²¹⁻²³. The second goal, involving solving an inverse problem, have only seen a handful of valiant efforts^{9, 24, 25}.

In a different approach to these objectives²⁶⁻²⁸, several exact results have been derived that address questions of principle. For example, instead of looking for the best fit of the observed single-molecule dynamics using the assumption that it is one-dimensional diffusive dynamics, one may ask whether the data can be described by such a model *in principle*. Indeed, some properties of the observed trajectories may be fundamentally inconsistent with diffusive dynamics²⁷. Most such studies focused on molecular transition paths - short trajectory segments crossing the transition region between metastable “reactant” and “product” states²⁹.

Another recent method³⁰ quantifies violations of the Chapman-Kolmogorov equation and compares experimental trajectories to Markov processes.

Recently, we have proposed an information-theoretical approach to detecting memory effects in discrete single-molecule trajectories³¹. In a nutshell, a trajectory that has a memory of the past states that it has visited is “more predictable” than a memoryless, Markov trajectory. As such it is more compressible when presented to a data-compression algorithm. Even before the advent of computers and compression algorithms, Shannon used this idea in his classic work on information content of printed English³²: memory of preceding letters in a text allows one to predict the subsequent letters with less uncertainty. As a result, English text can be encoded with fewer bits per character than random strings of letters. The purpose of this paper is to show how this method can be used for continuous trajectories, which are more common in single-molecule experiments. We start with summarizing our method in application to discrete data.

2. Method applied to discrete observables

Consider an ergodic single-molecule trajectory $i(t)$, where $1 \leq i \leq N$ is a discrete index enumerating the observable states of the system and t is time. In practice, the time is sampled at discrete intervals and thus is effectively discrete; without loss of generality we thus can assume that t is an integer. If $i(t)$ is a Markovian random walk, then the probability p_j to find the system in state j at time $t + 1$ obeys the master equation

$$p_j(t + 1) = \sum_i T_{ji}^{(1)} p_i(t) \quad (1)$$

Here $T_{ji}^{(1)}$ is the transition probability, i.e. the conditional probability to find the system in state j at time $t + 1$ given that it was in i at t . In particular, for a system in a steady state, we have

$$p_j^{ss} = \sum_i T_{ji}^{(1)} p_i^{ss}, \quad (2)$$

where p_j^{ss} is the steady-state probability for being in state j , which is simply the frequency of the state j appearing in the (sufficiently long) sequence $i(1), i(2), \dots$

The random walk model can be generalized by defining a k -th order Markov process in which the system has memory of its k previous steps. Let $p_{i_k, \dots, i_1}(t)$ be the joint probability of finding the system in state i_k at time t , i_{k-1} at time $t - 1$, ..., i_1 at time $t - k + 1$. This probability evolves according to

$$p_{i_{k+1}, \dots, i_2}(t + 1) = \sum_{i_1} T_{i_{k+1}, i_k, \dots, i_1}^{(k)} p_{i_k, \dots, i_1}(t), \quad (3)$$

where $T_{i_{k+1}, i_k, \dots, i_1}^{(k)}$ is the conditional probability of observing i_{k+1} given that the previous k states visited by the system were $i_1, \dots, i_k$. The steady-state solution $p_{i_k, \dots, i_1}^{ss}$ is the fixed point of the map in Eq. 3:

$$p_{i_{k+1}, \dots, i_2}^{ss} = \sum_{i_1} T_{i_{k+1}, i_k, \dots, i_1}^{(k)} p_{i_k, \dots, i_1}^{ss} \quad (4)$$

Recording a trajectory $i(1), \dots, i(M)$ on a computer without compression requires $M \log_2 N$ bits, or $\log_2 N$ bits per symbol. Shannon showed that, for a long trajectory ($M \gg 1$) obeying a k -th order Markov process, a compressed representation is possible only requiring $h^{(k)} M$ bits, or $h^{(k)}$ bits per symbol, where

$$h^{(k)} = - \sum_{i_1, i_2, \dots, i_{k+1}} T_{i_{k+1}, i_k, \dots, i_1}^{(k)} p_{i_k, \dots, i_1}^{ss} \log_2 T_{i_{k+1}, i_k, \dots, i_1}^{(k)} \quad (5)$$

is the information or entropy rate of $i(t)$.

Now consider an arbitrary ergodic trajectory $i(t)$ that is not necessarily a k -th order Markov process with a finite k . We can construct a k -th order Markov *model* of the underlying dynamics from this trajectory. Indeed, the transition probability can be estimated to any order by measuring the joint probabilities as frequencies

$$p_{i_k, \dots, i_1}^{ss} \approx \frac{\text{number of times sequence } i_1, \dots, i_k \text{ is encountered in } i(t)}{M-k+1}, \quad (6)$$

and using

$$T_{i_{k+1}, i_k, \dots, i_1}^{(k)} = \frac{p_{i_{k+1}, i_k, \dots, i_1}^{ss}}{p_{i_k, \dots, i_1}^{ss}} \quad (7)$$

Shannon has constructed such models, of increasing order k , for the case where $i(t)$ represents the sequence of characters in printed English. His key observation is that if the order k of the model is increased, the information rate decreases, $h^{(1)} \geq h^{(2)} \geq h^{(3)} \geq \dots \geq h$, where h is the true information rate which provides the ultimate limit to which the original text can be compressed. This is because accounting for more memory of the preceding states makes the trajectory $i(t)$ more predictable, thus lowering the information content of each letter. Of course if, at some value $k = k_0$, the k -th order Markov model happens to be exact, (i.e., $h^{(k_0)} = h$), then increasing k does not change the information rate, ($h^{(k)} = h$ for $k \geq k_0$).

This procedure allows us to construct increasingly more accurate higher-order Markov models of the observed experimental signal $i(t)$; moreover, for each Markov order we will have computed the coefficients $T^{(k)}$ in the master equation, Eq. 3. If, at some k_0 , $h^{(k > k_0)}$ becomes independent of k then the model can no longer be improved; the value of k_0 then quantifies the temporal extent of memory in the observed trajectory.

In practice, this program can be accomplished only for modest values of k_0 , and it is thus impractical for systems that display long memory. As an alternative to constructing high-order Markov models explicitly, we used³¹ a lossless compression algorithm, such as the one due to Lempel and Ziv³³, to estimate the true information rate h . When this estimated value is significantly lower than that of the Markov model $h^{(1)}$, we know that the process $i(t)$ is non-Markov. If, for some k_0 , $h^{(k_0)} \approx h$, we anticipate that the k_0 -th order Markov model captures most of the memory.

3. Digitizing continuous trajectories.

While information theory deals with discrete data, its practical applications often have to handle continuous signals such as the pressure of the sound wave in an audio recording. Discretization of such signals usually relies on the sampling theorem³⁴, and often involves binning of the data. In contrast, single-molecule trajectories are stochastic and not bandwidth-limited in the range of timescales of interest. Moreover, our goal here is not to accurately record or transmit a single-molecule trajectory $x(t)$, but to construct an accurate dynamical model that can generate it. As a consequence, discretization of continuous trajectories requires special care. We also note that entropy metrics directly applicable to continuous trajectories, such as the Kolmogorov-Sinai entropy, have been developed for deterministic systems; although they have also been applied to stochastic processes in some of the work³⁵, such applications were mostly focused on the general properties of the entropy for certain classes of stochastic processes rather than on approaches to its computation or on differentiating between different dynamical models.

To illustrate why discretization of continuous stochastic trajectories is a nontrivial problem, consider a trajectory $x(t)$ of a particle undergoing free diffusion. Diffusion is a Markov process, and its trajectory is a fractal, self-similar object. We wish to map $x(t)$ onto a discrete-state trajectory $m(t)$, with m numbering discrete bins of width Δx (Fig.1). On physical grounds, we hope that a discrete model of diffusion should be close to a Markov random walk. As such, the sequence of distinct states visited by this walk (i.e. one obtained by counting each continuous trajectory fragment where the state $m(t)$ does not change once) should also be a Markovian random walk).

Naïve binning of the data, $m(t) = [x(t)/\Delta x]$, however, results in a highly non-Markov sequence of discrete states. To see this, approximate $x(t)$ itself by a discrete random walk with a much finer step length $\delta x \ll \Delta x$. Thus each bin of length Δx contains $l = \Delta x/\delta x$ discrete “microsites”. One can show³¹ that if a particle enters a bin m from the left, it will exit it back to the left with probability $T_{m-1,m,m-1} = \frac{l}{l+1}$ and to the right with probability $T_{m+1,m,m-1} = \frac{1}{l+1}$. Similar expressions are obtained for particle entering the bin from the right. Moreover, the sequence of bins visited by the particle is a second-order Markov process: since $x(t)$ is Markovian only the previous bin is remembered. Using Eq. 5 one finds the information rate associated with the sequence of bins to be

$$h = h^{(2)} = -\frac{1}{l+1} \log_2 \frac{1}{l+1} - \frac{l}{l+1} \log_2 \frac{l}{l+1} \quad , \quad (8)$$

which becomes vanishingly small in the continuous limit $\delta x \rightarrow 0$ (i.e. $l \rightarrow \infty$). This rather pathological behavior of the stochastic process describing the sequence of bins visited by a diffusive particle is related to the properties of diffusive dynamics: a diffusive particle that crosses a certain point (e.g., a boundary between bins) will repeatedly cross it an infinite

number of times (on the average)³⁶. As a result, when observed with an infinite resolution, it has zero probability to traverse a bin relative to the probability of exiting the bin through the same boundary it entered it! The repeated recrossings of boundaries come in bursts, resulting in a highly non-Markov process. In contrast to the above observation, one feels that, when continuous dynamics along x is mapped onto a discrete random walk with a finite step size, the information rate should be 1 bit per step, since at each step one learns one bit of information (i.e. whether the walker has stepped left or right).

A chemical reaction rate theory expert may recognize the above issue as the “recrossing problem” encountered when attempting to identify transitions between two spatial domains³⁷ – a diffusive trajectory will, on the average, recross a boundary between the two an infinite number of times – this results in the overcounting of the number of transitions. Here we propose to solve this problem and to map continuous dynamics onto a discrete random walk using a milestone-type of approach^{36, 38-41} illustrated in Fig. 1a. In this approach, only the first crossing of each new “milestone” (i.e., the boundary between two bins) counts. Subsequent recrossings of the same milestone do not change the discrete state m of the system until a new milestone is encountered.

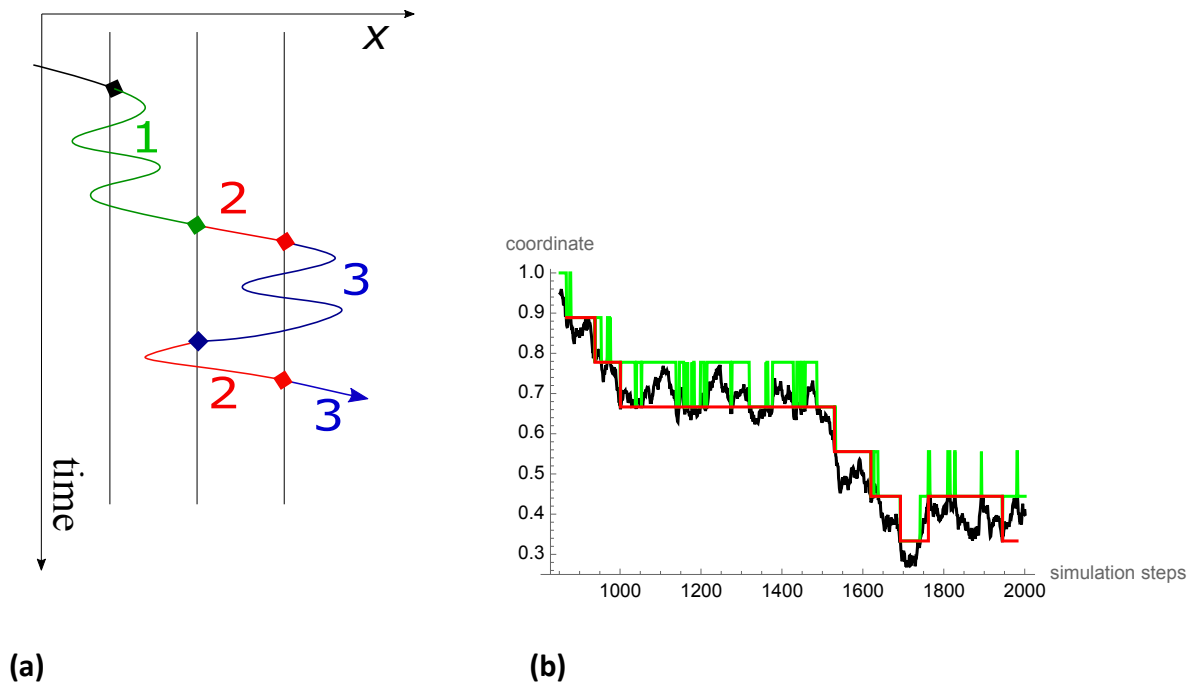


Figure 1. (a) Milestoning discretization illustrated using 3 milestones. The discrete state m of the system (indicated by a number next to the trajectory) changes only when a new (different than the previous one) milestone is crossed. **(b)** A diffusion trajectory (black) discretized using binning (green) and milestoning (red).

Figure 1b illustrates the differences between discretizing a trajectory of a diffusing particle using binning and milestoning, the latter eliminating multiple events where the trajectory exits and reenters intervals between adjacent bins.

Time discretization of trajectories is simpler and can be accomplished by sampling the milestone coordinate at a specified sampling rate. Thus a continuous trajectory, after being discretized both in time and space, can be represented as a discrete sequence $m(0), m(\Delta t), m(2\Delta t), \dots$, which is in the form amenable to the treatment of Section 2. This,

however, is not the only possible (and useful) representation: alternatively we could represent the same data, for example, as $\{m_1, \theta_1\}, \{m_2, \theta_2\}, \dots$, where $m_1, m_2, \dots$ is the sequence of milestones visited by the trajectory and $\theta_1 = \left\lceil \frac{t_1}{\Delta t} \right\rceil, \theta_2 = \left\lceil \frac{t_2}{\Delta t} \right\rceil, \dots$ are the corresponding times of arrival at those milestones measured with time resolution Δt . Although the latter representation is not a stationary process because the time θ_n increases without bounds, we can further replace it by the equivalent lag-time representation $\{m_1, \theta_2 - \theta_1\}, \{m_2, \theta_3 - \theta_2\}, \{m_3, \theta_4 - \theta_3\} \dots$, which records the sequence m_n of milestones crossed and the lagtimes $\theta_{n+1} - \theta_n$ between crossing events.

4. Information rate in a continuous-time random walk.

Continuous-time random walks (CTRWs), also known as semi-Markov processes, are an important, extensively studied class of non-Markov models^{42, 43}. If, using the milestone picture we think of $i_n = \{m_n, t_n\}$ as the state of the system (note that, for simplicity of notation, we do not differentiate between the milestone crossing time t_n and its discretized counterpart θ_n), then the transition probability matrix for a CTRW can be written as:

$$T_{\{m_{n+1}, t_{n+1}\}, \{m_n, t_n\}}^{(1)} = \phi_{m_{n+1}, m_n} \psi_{m_{n+1}, m_n}(t_{n+1} - t_n). \quad (9)$$

Here ϕ_{m_{n+1}, m_n} is the conditional probability that the next milestone crossed is m_{n+1} given that the current one is m_n , and $\psi_{m_{n+1}, m_n}(t)$ is the probability distribution of the lag time between crossing these two milestones. In this representation, CTRW is a 1st order Markov process, which implies that there is only memory of the last state:

$$T_{i_{k+1}, i_k, \dots, i_1}^{(k)} = T_{i_{k+1}, i_k}^{(1)}. \quad (10)$$

Physically, the CTRW picture is as follows: the sequence of milestones visited by the process is a (1st order) Markov random walk defined by the conditional probabilities ϕ_{m_{n+1},m_n} of visiting the next milestone m_{n+1} given the previously visited one m_n . The lag time $\tau = t_{n+1} - t_n$, i.e., the time the system spends on a milestone m_n before transitioning to m_{n+1} , is drawn from a probability density $\psi_{m_{n+1},m_n}(\tau)$ that may depend on these milestones, and the lag times are statistically independent. A CTRW trajectory $m(t)$ is “almost Markov” in that the only memory comes from non-exponentiality of the distribution $\psi_{m_{n+1},m_n}(\tau)$ [note that, when viewed as a sequence of states i_n introduced above, this process is simply Markov]. The trajectory $m(t)$ becomes Markov if this distribution is exponential and independent of m_{n+1} ⁴⁴,

$$\psi_{l,m}(\tau) = \Gamma_m e^{-\Gamma_m \tau}, \quad (11)$$

where Γ_m^{-1} is the average dwell time on a milestone m .

It is instructive to consider the case where $\psi_{l,m}(\tau) = \psi(\tau)$ is milestone-independent. In this case the milestone trajectory is completely specified by a milestone sequence $m_1, m_2, \dots$ and a lag time sequence $\tau_1, \tau_2, \dots$, with the two sequences being statistically independent of one another. The corresponding information rate is the sum of the information rates of these two processes,

$$h = h_M + h_\tau, \quad (12)$$

and thus their mutual information³⁴ $I = h - h_M - h_\tau$ is zero. Here h_M is the information rate of the milestone sequence given by (cf. Eq. 5), which can be interpreted as the “spatial component” of information rate:

$$h_M = - \sum_{m,n} \phi_{m,n} \log_2 \phi_{m,n} p_n^{ss} \text{ bits per milestone crossing, } (13)$$

with p_n^{ss} satisfying (cf. Eq. 4)

$$p_m^{ss} = \sum_n \phi_{m,n} p_n^{ss} \quad , \quad (14)$$

and h_τ , which may be interpreted as “temporal component” of the information rate, is related to the differential entropy of the distribution $\psi(\tau)$,

$$h_\tau = - \int_0^\infty d\tau \psi(\tau) \log_2 \psi(\tau) + \log_2 1/\Delta t, \quad (15)$$

where the last term results from discretization of the distribution with a time resolution of Δt ,

³⁴. Importantly, for a fixed distribution mean $\langle \tau \rangle = \int_0^\infty d\tau \tau \psi(\tau)$ the maximum value of h_τ ,

$\max h_\tau = \frac{1}{\ln 2} + \log_2 \frac{\langle \tau \rangle}{\Delta t}$, is provided when the distribution is exponential³⁴, $\psi(\tau) = \Gamma e^{-\Gamma \tau}$,

where $\Gamma = \langle \tau \rangle^{-1}$. This illustrates the fact that memory effects (i.e. non-exponentiality of the distribution) reduce the information rate of a CTRW. To further quantify this effect, it may be

beneficial to measure h_τ relative to its maximum value, that is to consider the quantity

$$\Delta h_\tau = h_\tau - \max h_\tau = - \int_0^\infty d\tau \psi(\tau) \log_2 \psi(\tau) - \log_2 \langle \tau \rangle - \frac{1}{\ln 2} \quad (15a)$$

which is no longer dependent on the time step.

Eqs. 12-15 are readily generalized when the lag time distributions depend on the milestone: Eqs. 12 -14 remain the same, but with h_τ becoming

$$h_\tau = - \int_0^\infty d\tau \sum_{m,n} \phi_{m,n} p_n^{ss} \psi_{m,n}(\tau) \log_2 \psi_{m,n}(\tau) + \log_2 1/\Delta t. \quad (16)$$

If the lag time distribution between two milestones, $\psi_{m,n}(\tau)$, depends on the initial but not the final milestone, $\psi_{m,n}(\tau) = \psi_n(\tau)$, then Eq. 16 is further simplified to give

$$h_\tau = - \int_0^\infty d\tau \sum_n p_n^{ss} \psi_n(\tau) \log_2 \psi_n(\tau) + \log_2 1/\Delta t. \quad (17)$$

5. A case study: Mapping biased continuous diffusion onto a CTRW.

We now illustrate the milestoning discretization in application to the simplest continuous Markov process, diffusion of a Brownian particle in the presence of a constant force F at temperature T . The probability $p(x, t)$ of finding the particle at coordinate x at time t is described by the Smoluchowski equation

$$\frac{\partial p(x, t)}{\partial t} + v \frac{\partial p(x, t)}{\partial x} = D \frac{\partial^2 p(x, t)}{\partial x^2} \quad (18)$$

where D is the diffusivity and $v = FD/k_B T$ is the mean drift velocity caused by the constant force F . Periodic boundary conditions are used here, which is equivalent to considering diffusion on a ring. We further use evenly spaced milestones with a distance $\Delta x = L$ between them. Since only transitions to neighboring milestones are possible, the only nonzero conditional probabilities $\phi_{m,n}$ are $\phi_+ = \phi_{m+1,m}$ and $\phi_- = \phi_{m-1,m}$. To determine these probabilities, as well as the dwell time distribution $\psi(t)$, we imagine a particle that has just crossed, at $\tau = 0$, a milestone located, say, at $x = 0$. We track this particle until it either crosses the milestone to the left (located at $x_- = -L$) or to the right (located at $x_+ = +L$) thereby exiting the interval $(-L, L)$, and we record the time τ_- or τ_+ when it happened. By repeating this experiment multiple times one can determine the distribution of the exit times τ_- or τ_+ , as well as the probabilities ϕ_- and ϕ_+ of exiting the interval $(-L, L)$ to the left or right. An analytical solution of this problem of calculating exit time distributions and probabilities has been given in the literature^{45, 46}. Although it is not immediately obvious, the distribution of the times τ_- and τ_+ are identical, and thus we drop the subscript $\pm$ to denote this distribution simply $\psi(t)$. Its Laplace transform, $\tilde{\psi}(s) = \int_0^\infty d\tau e^{-s\tau} \psi(\tau)$ is given by:

$$\tilde{\psi}(s) = \left(e^{\frac{FL}{2k_B T}} + e^{-\frac{FL}{2k_B T}} \right) \sinh L \sqrt{\frac{s}{D} + \left(\frac{F}{k_B T} \right)^2} / \cosh 2L \sqrt{\frac{s}{D} + \left(\frac{F}{k_B T} \right)^2}. \quad (19)$$

Since this is an even function of the force F , it is clear from Eq. 19 that this distribution is indeed independent of whether the next milestone crossed is in the direction of the force or opposite this direction. The probabilities $\phi_{\pm}$ are given by

$$\phi_{\pm} = \frac{1}{\left(1 + e^{\mp \frac{FL}{k_B T}}\right)}. \quad (20)$$

Because the steady-state occupation probabilities are the same for all milestones (given the periodic boundary conditions), the information rate of our process is given by Eq. 12 with

$$h_M = -\phi_- \log_2 \phi_- - \phi_+ \log_2 \phi_+ \quad (21)$$

and with h_{τ} that can be estimated using Eq. 15. Eq. 21 is simply the entropy rate of a random walk that makes a step right with a probability ϕ_+ and left with a probability $\phi_- = 1 - \phi_+$. The entropy h_{τ} of the lag-time distribution, however, is different from that expected for a Markovian process, since the distribution described by Eq. 19 is clearly not an exponential one. Indeed, since it takes a finite time to reach one milestone from another, $\psi(\tau)$ must vanish at $\tau = 0$, in contrast to the exponential distribution (Eq. 11) expected for a Markov process. As a result, the milestone description of Brownian motion is not a Markov random walk, but rather a semi-Markov process.

To get further insight into how the information rate of a discretized trajectory depends on the spacing between milestones, consider the case of free diffusion (i.e., zero force). In this case we have $\phi_- = \phi_+ = 1/2$, and $h_M = 1$ bit/milestone crossing. Notice that the Laplace-transformed distribution of Eq. 19 in this case is a function of the dimensionless frequency parameter sL^2/D . Equivalently, the lag time distribution can be written in the form

$$\psi(\tau) = \frac{D}{L^2} \rho\left(\frac{\tau D}{L^2}\right), \quad (22)$$

where $\rho(y)$ is a function that is independent of the inter-milestone spacing L or diffusivity D .

Performing the integral in Eq.15 and using Eq. 22, we now find

$$h_\tau = - \int_0^\infty \rho(y) \log_2 \rho(y) dy + \log_2 \frac{L^2}{D\Delta t} \approx 0.26 + \log_2 \frac{L^2}{D\Delta t} \quad (23)$$

It is instructive to compare this result with the case of the purely exponential distribution. To do

so, we replace the true distribution $\psi(\tau)$ with an exponential one

$$\psi_e(\tau) = e^{-\tau/\langle\tau\rangle}/\langle\tau\rangle, \quad (24)$$

where the average lag time $\langle\tau\rangle$ is the same as the one for the true distribution, Eqs. 19. To find

the latter, we write

$$\langle\tau\rangle = \left. \frac{d\tilde{\psi}(s)}{ds} \right|_{s=0} = \frac{Lk_B T}{FD} \frac{e^{\frac{FL}{k_B T}} - 1}{e^{\frac{FL}{k_B T}} + 1}, \quad (25)$$

which results, at $F \rightarrow 0$, in

$$\langle\tau\rangle = \frac{L^2}{2D}. \quad (26)$$

Also note that the asymptotic behavior of $\psi(\tau)$ for $\tau \gg L^2/D$ is precisely given by Eq.24, which

can be ascertained by considering the limit $s \ll D/L^2$ in Eq. 19. Using $\psi_e(\tau)$ instead of $\psi(\tau)$ in

Eq. 15, we find

$$h_\tau = \frac{1}{\ln 2} + \log_2 \frac{\langle\tau\rangle}{\Delta t} = \frac{1}{\ln 2} - 1 + \log_2 \frac{L^2}{D\Delta t} \approx 0.44 + \log_2 \frac{L^2}{D\Delta t} \quad (27)$$

Comparing this with Eq.23, we see that the estimates of h_τ obtained using the fully Markov

approximation (corresponding to an exponential distribution of the lag times) and the CTRW

differ by a constant.

6. Dependence on spatial resolution and infinite resolution limit.

In the above discussion of biased diffusion, we have fixed the spatial resolution L , with which the trajectory is observed. As the resolution is decreased, $L \rightarrow 0$, the observed motion looks increasingly like a random walk, with the probabilities $\phi_{\pm}$ (Eq. 20) approaching $\frac{1}{2}$, as thermal motion dominates over drift³⁶. Moreover, the random walk is self-similar, looking the same at all (sufficiently short) length scales. This is disconcerting: if the trajectory looks like a self-similar, *unbiased* random walk at all sufficiently short length scales then by analyzing the trajectory $x(t)$ with better time resolution we appear to learn less about the underlying process (particularly about the force that causes drift). We anticipate a similar problem to arise when studying trajectories $x(t)$ with memory, obeying, e.g., a generalized Langevin equation – see next Section. At sufficiently short time/length scales the dynamics, again, becomes diffusive^{36, 47}, with entropy rate thus approaching that of a Markov process despite the underlying memory.

To understand this limit better, consider our model of biased diffusion in the limit $\frac{FL}{k_B T} \ll$

1. Expanding Eq. 21 in a Taylor series in $\frac{FL}{k_B T}$, we obtain, to lowest nontrivial order

$$h_M \approx 1 - \frac{F^2 L^2}{8 \ln 2 (k_B T)^2}. \quad (28)$$

As anticipated, h_M approaches 1 bit per milestone step as L is decreased, with the biased effect of the force becoming increasingly negligible. Yet let us consider the information $H_M(t)$ gained after some finite and sufficiently long time t , during which the trajectory will perform $t/\langle\tau\rangle$ transitions between milestones:

$$H_M(t) = h_M t / \langle\tau\rangle. \quad (29)$$

Using Eq. 25, and, again, expanding in Taylor series to lowest nontrivial order, we obtain

$$H_M(t) \approx \frac{2Dt}{L^2} - \frac{F^2 D}{4 \ln 2 (k_B T)^2} = \frac{2Dt}{L^2} - \left(\frac{1}{4 \ln 2} - \frac{1}{6} \right) \frac{Fv}{k_B T} t. \quad (30)$$

The first term in Eq. 30 is the (spatial component of) the information corresponding to free diffusion. As expected, it diverges as spatial resolution is increased, i.e. as $L \rightarrow 0$. Importantly, however, Eq. 30 contains a second term, which is independent of the resolution L and is proportional to the rate $Fv/k_B T$ at which energy is dissipated by the force-driven particle. It is negative because a biased random walk is less random than an unbiased one, and thus it has lower information than an unbiased one. This result shows that resolution-independent information rate can be obtained by subtracting the “free diffusion” part $\frac{2Dt}{L^2}$. In practice, this can be achieved by considering the linear dependence of $H_m(t)$ on $1/L^2$: the second term in Eq. 30 is the intercept obtained by extrapolating this dependence to $\frac{1}{L^2} \rightarrow 0$.

Figure 2 illustrates this approach using simulations. We integrate, numerically, stochastic trajectories obeying the Langevin equation

$$\frac{k_B T}{D} \frac{dx}{dt} = -F + \zeta(t), \quad (31)$$

with $\zeta(t)$ being a Gaussian-distributed random force with zero mean satisfying the fluctuation-dissipation theorem $\langle \zeta(t) \zeta(t') \rangle = \frac{2(k_B T)^2}{D} \delta(t - t')$. Eq. 31 provides a stochastic description of trajectories whose ensemble obeys the Smoluchowski equation, Eq. 18. By applying the milestone analysis and computing the dependence of $h_M/\langle \tau \rangle$ on $1/L^2$ directly from the simulated trajectories, we, indeed, can recover the force-dependent intercept, which agrees with the prediction of Eq. 30.

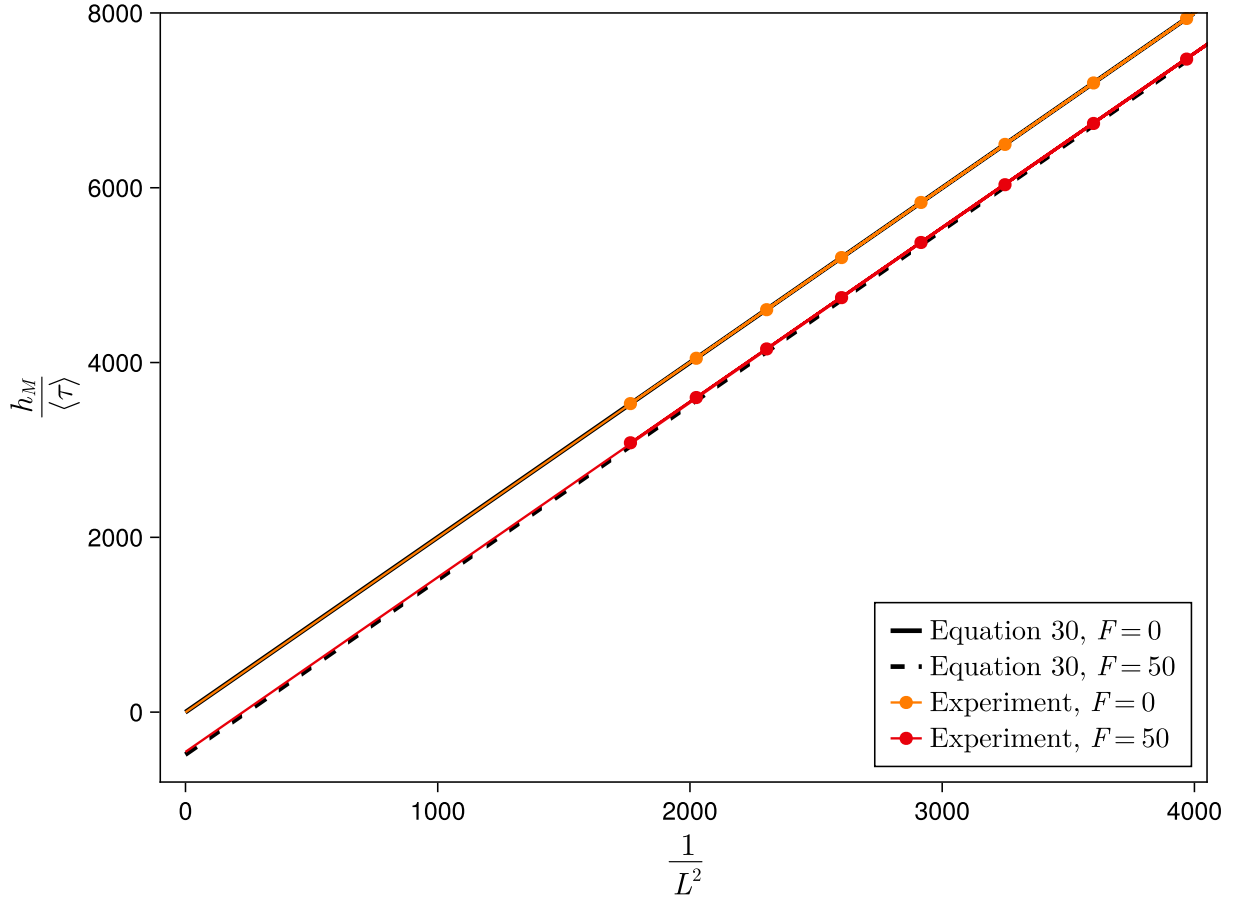


Figure 2: Information rate h_M per unit time plotted against the inverse square of the inter-milestone distance L obtained from simulations of overdamped Langevin trajectories on a ring in the presence of a constant driving force F . The distance units are R , the time units are R^2/D , and the force units are $k_B T/R$, where R is the ring length. Each system was integrated with the Euler-Maruyama integrator with a timestep of 10^{-7} units, until 5×10^6 milestone crossings were obtained. The black lines show the values predicted by Eq. 30, and the colored dots show the values estimated via milestoning. Since Equation 30 is only valid in the limit of $L \rightarrow 0$, we restrict our experiment data to milestones which are closely spaced. However, plotting data in the entire range of L (particularly for large values of $1/L^2$) makes it difficult to examine the lines of best fit and whether they agree with theory. Here, we obtain data for experiments up

to $\frac{1}{L^2} = 15876$, and generate the colored best-fit lines for all the experiments, but restrict the plot to $\frac{1}{L^2} = 4000$ in order to maintain visual clarity.

7. Inferring memory from trajectories: generalized Langevin equation

We now proceed to test the ability of the above analysis to detect memory effects in the observed non-Markovian dynamics. We first start with dynamics on a ring obeying a generalized Langevin equation of the form

$$0 = - \int_{-\infty}^t \Gamma(t - t') \dot{x}(t') dt' + \zeta(t) . \quad (32)$$

Here

$$\Gamma(t) = k_s \exp\left(-\frac{k_s t}{\gamma}\right) + 2\gamma_0 \delta(t) \quad (33)$$

is a memory kernel, and $\zeta(t)$ is a Gaussian-distributed random force with zero mean obeying the fluctuation-dissipation theorem of the form $\langle \zeta(t) \zeta(t') \rangle = k_B T \Gamma(t - t')$. Eq. 32 is obeyed by a particle that experiences an intrinsic friction force $-\gamma_0 \dot{x}$, and which is coupled to another degree of freedom y that experiences friction force $-\gamma \dot{y}$ via a harmonic potential $k_s(y - x)^2/4$. This offers a practical method of integrating Eq. 32^{36, 48}. In the results reported here we use $\gamma = 30\gamma_0$ and $k_s = 1000 \frac{k_B T}{R^2}$, where R is the length of the ring. The simulation timestep used is $\delta t = 10^{-7} \frac{R^2 \gamma_0}{k_B T}$.

Figure 3 shows the spatial component of the information rate (i.e. the information rate associated with the sequence of milestones crossed by the trajectory $x(t)$) evaluated for $k = 1, 2$, and 3. The fact that $h_M^{(3)} < h_M^{(2)} < h_M^{(1)}$ immediately informs us that we are not dealing

with a Markov process. Moreover, the dependence $h_M^{(k)}(\frac{1}{L^2})$ exhibits an intercept, when extrapolated to $\frac{1}{L^2} \rightarrow 0$, a behavior similar to that found for diffusive dynamics in the presence of force. This indicates that the observed dynamics is more predictable (i.e. has lower information) than diffusive or Markovian dynamics. We expect that, as the Markov order k is increased, $h_M^{(k)}$ will eventually converge to the “true” value h_M , but evaluation of higher order entropy rates quickly becomes expensive computationally and unreliable⁴⁹. We therefore resort to the recently reported compression-based method³¹ to estimate h_M . In brief, the information in the milestone sequence $m_1, m_2, \dots$ is estimated from the size of the data after it was compressed using a lossless compression algorithm; although compression algorithms rarely compress finite-size data to the true information limit, this error can be corrected by comparing the compression-derived information rate with the theoretically known true information rate for a suitably chosen Markov process. The result is shown in Fig. 3, where the estimated information rate h_M is found to be significantly lower than its finite-Markov-order estimates, indicating long memory.

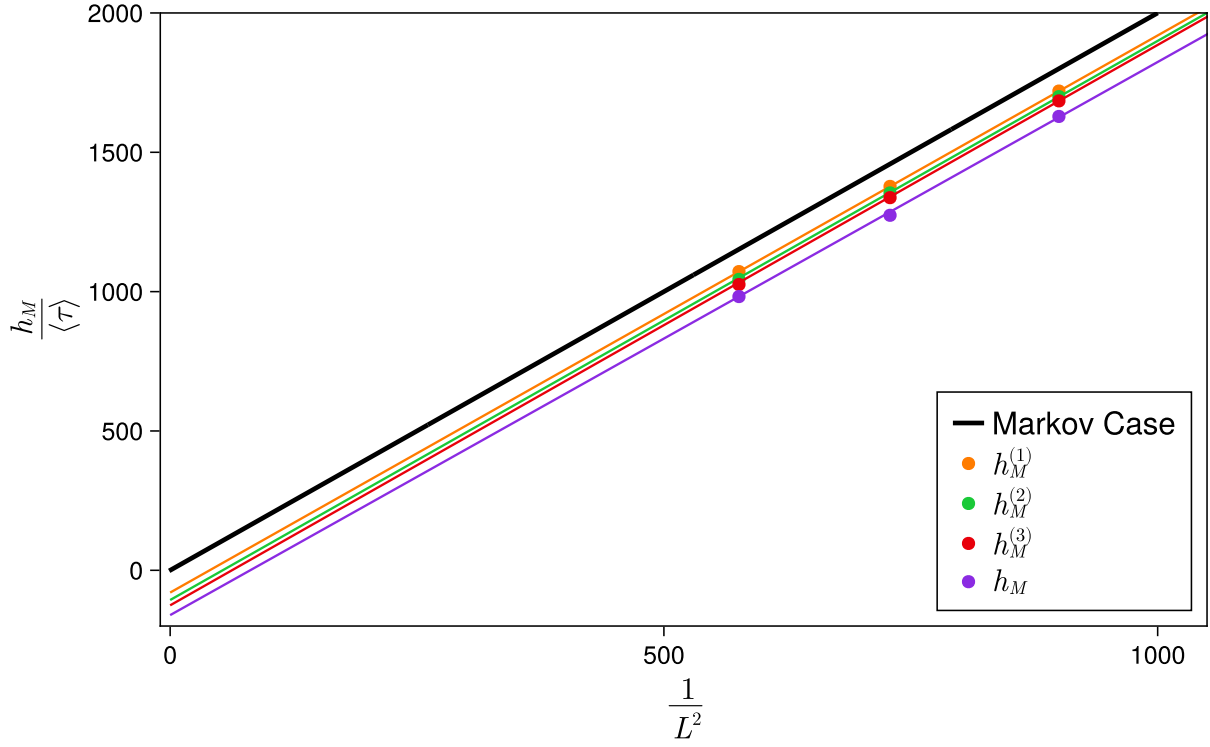


Figure 3: The spatial component of the information rate per unit time plotted against the inverse square of the inter-milestone distance, for a particle on a ring obeying the generalized Langevin equation with exponential memory, Eq. 32. The ring length R is used as the unit length, and the time is measured in units of $R^2\gamma_0/k_B T$. The integration scheme is the same as in Figure 2, with the same number of milestone crossings observed. The solid black line shows the result for ordinary Langevin equation without memory (Eq. 30 with $F = 0$; also see Fig. 2). The colored circles are estimates of the $h_M^{(k)}$, estimated with the compression method. Similarly to Figure 2, we gather data for very closely spaced milestones, up to $\frac{1}{L^2} = 3249$, but only plot results in the range $\frac{1}{L^2} \leq 1000$ in order to keep the differences between individual lines visible in the plot.

As the order k of the estimate is increased the information rate $h_M^{(k)}$ decreases toward the compression-based “true” rate h_M , indicating non-Markovianity of the dynamics.

8. Inferring memory from molecular trajectories: atomistic simulations of end-to-end dynamics of an intrinsically disordered protein

Dynamics of the distance between the ends of a biopolymer chain has been the subject of extensive experimental and theoretical work (see, e.g., refs.^{50, 51, 18, 52-55, 56-59}). While theory predicts such dynamics to be highly non-Markov^{47, 60-62}, direct demonstration of non-Markovianity in simulated or experimentally measured dynamics is often difficult and indirect: for example, analysis of transition paths⁶³, modeling of trajectories using a generalized Langevin equation¹², and analysis of the scaling behavior of the loop formation rate^{59, 61} have been invoked to probe non-Markov effects. To see if our information-theoretical analysis can directly inform us about memory effects, here we have analyzed the end-to-end dynamics of a model unstructured polypeptide, the 11-residue peptide fragment with the Gly-Ser repeat^{21, 64}, using atomistic simulations (Fig. 4).

In comparison to the above examples and to earlier studies of discrete random walks³¹, this example presents two challenges. First, the length of the trajectory (measured relative to the characteristic relaxation time of the problem) is much shorter. As a result, numerical evaluation of Eq. 5 may become unreliable⁴⁹. Likewise, the compression-based method used in the previous Section to evaluate the “true” information rate h_M may be affected by the insufficient trajectory length³¹.

Second, unlike the above examples, where we could use larger numbers of milestones for the analysis, here the time evolution of the polypeptide's end-to-end distance $x(t)$ was discretized using only seven equally spaced milestones placed as shown in Fig. 4. This particular choice of milestone spacing is dictated by practical considerations that are likely to be common to the analysis of both molecular simulations and experimental data. Specifically, the sampling rate of both types of data is often limited by data storage demands, by the speed of a camera, or by the inherent time resolution of the experiment. In the simulation described here, the molecular structures were saved every $\Delta t = 5$ ps, a time interval that is 3 orders of magnitude longer than the simulation timestep. If the spacing between the milestones is too small, multiple milestones can be crossed during this time interval. Such missed milestone crossings will lead in errors in estimating the entropy rates. While the recently proposed stochastic algorithm that reintroduces missed crossings probabilistically⁶⁵ partially remedies this problem, it does not consider the possibility that multiple milestones are crossed while the trajectory is unobserved. Note that Δt exceeds the velocity relaxation time for the trajectory considered here; thus the effect of inertia on the information rate h_M cannot be observed (similarly to most experimental studies²⁸).

To test whether our method can differentiate between Markov and non-Markov dynamics when the trajectory length and the number of milestones are limited by the above considerations, we have computed the information rates both for the original trajectory $x(t)$ and for its Markovian model, a trajectory that obeys the overdamped Langevin equation

$$\frac{k_B T}{D} \frac{dx}{dt} = -U'(x) + \zeta(t). \quad (34)$$

The potential $U(x)$ (shown in Fig. 4) was chosen such that the corresponding equilibrium distribution is the same as that for the original trajectory, and the diffusion coefficient $D = 9.04 \times 10^{-4} \text{ nm}^2/\text{ps}$ was chosen such that the mean lag time between milestone crossings is the same. The Langevin trajectory, starting from the same initial value of x as the molecular simulation, was sampled at the same time intervals and analyzed in the same way as the original trajectory using the same set of milestones, and as with the earlier figures, we assess h_M , the spatial component of the entropy. As seen from Figure 5, the information rate for the original trajectory decreases with the Markov order, while the corresponding values for the Markovian model are virtually the same. Although the errors in compression-based estimates of the information rate are quite large (see Appendix C for further error analysis, which shows that these errors originate in the relatively short length of the molecular trajectory), the method detects the non-Markov character of the molecular trajectory reliably.

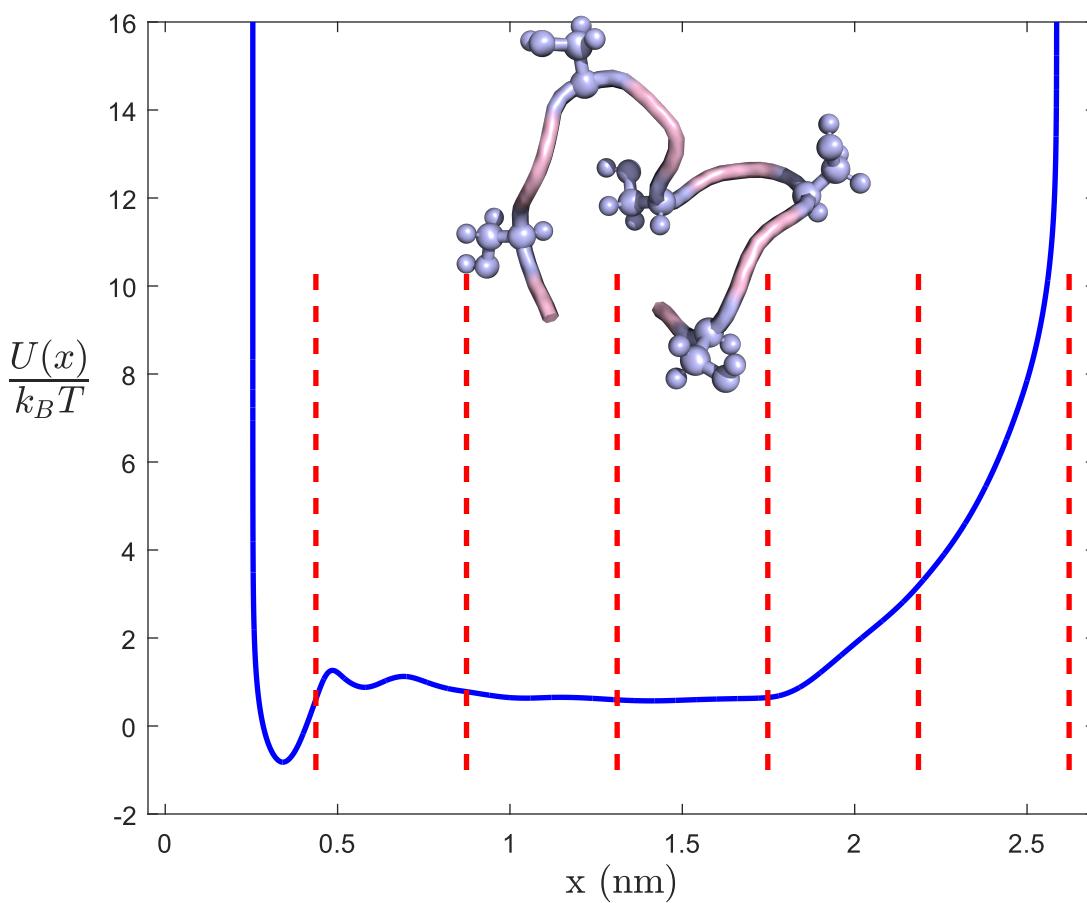


Figure 4. The potential of mean force $U(x) = -k_B T \ln p(x)$ for the end-to-end distance x of a Gly-Ser repeat peptide shown at the top. Here $p(x)$ is the equilibrium distribution of x . The locations of the milestones used to analyze the trajectory $x(t)$ are shown as vertical dashed lines.

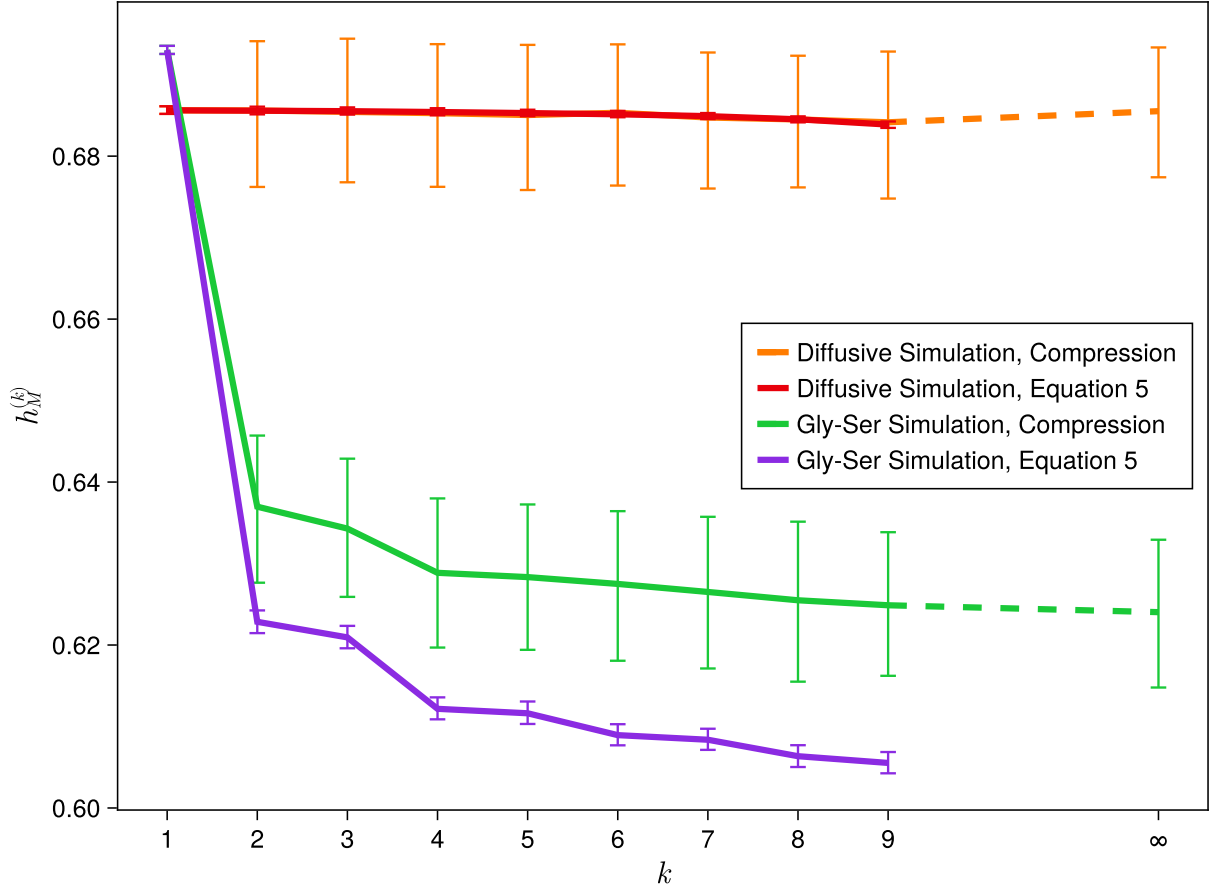


Figure 5: Compression-derived³¹ k -th order information rate $h_M^{(k)}$ (Eq.5) associated with the sequence of milestones visited by a trajectory $x(t)$ describing the time evolution of the end-to-end distance x of a Gly-Ser repeat peptide (green). The compression-derived estimate h_M for the infinite-order information rate is shown at the rightmost points, connected to the rest of the data points by a dashed line. The same information computed using Eq. 5 directly is shown in purple. Since the trajectories used are quite short, we repeatedly estimate the entropy and compute statistics over the trials to obtain the values and 95th-percentile errors shown. The exact estimation procedure is detailed in Appendix B, and the source of the difference between the two estimators is discussed in Appendix C. While the estimates do not agree with each

other numerically, $h_M^{(k)}$ decreases with increasing k regardless of the estimator used, indicating non-Markov behavior. To verify that this is a signature of non-Markovian dynamics and is not a consequence of artifacts associated with finite time resolution of the trajectory or the relatively small number of milestones used, we have simulated overdamped Langevin dynamics of a one-dimensional particle that has the same potential of mean force $U(x)$ as the peptide, and analyzed the resulting trajectory, sampled at the time intervals of the same duration, using the same methods. The estimated information rate for such a Langevin trajectory is virtually independent of the Markov order k , as expected for a Markov process.

9. Summary

This work extends the recent information-theory-based method³¹ for detecting and quantifying memory effects in discrete random walks to continuous trajectories such as those observed in molecular simulations and single-molecule experiments. Doing so requires solving two related problems. First, this requires discretization of trajectories, which, in general, introduces spurious memory effects that are absent in true dynamics. Second, for a stochastic process the information in a trajectory diverges as the resolution with which it is sampled is increased⁶⁶.

Our approach to solve these problems is to use milestoning, a method introduced originally as a computational tool, as a data analysis tool. Milestoning naturally maps continuous trajectories onto continuous-time discrete random walks, whose information rates can then be computed either directly or using compression algorithms³¹. Although those information rates depend on the resolution with which the observed trajectory is discretized and diverge when the distance between the milestones goes to zero, the divergent part, when it results from diffusive

dynamics at short timescales, is well understood and can be subtracted to get a resolution-independent component of the entropy rate resulting from non-Markov effects. We have shown the utility of this method in detecting memory effects in generalized Langevin equation dynamics and in end-to-end dynamics of an intrinsically disordered peptide.

In principle, this method automatically produces a hierarchy of k -th order continuous-time random walk models of the observed trajectory^{67, 68}, whose quality can be judged by the convergence of the information rate $h^{(k)}$ toward the $k = \infty$ value²⁴. While not necessarily “physical”, such models may provide clues to the physical nature of the observed process and/or discriminate between its alternative physical explanations – an illustration of that is given in ref.³¹, where information theory was used to discriminate between static and dynamic disorder.

Finally, let us note that using the information rate $h^{(k)}$ to assess the quality of a k -th order Markov model is not the only possible approach. One can, for example, ascertain the validity of the k -th order Markov model by verifying that the conditional probabilities

$T_{i_{k+1}, i_k, \dots, i_1, i_0}^{(k)}$ are independent of i_0 (i.e. there is only memory of k previously visited states).

Such an alternative approach is briefly explored in Appendix D. Although we have no proof that the information-theory based approach is always superior, we find this to be anecdotally the case for all the systems studied here (Appendix D and Fig. 7).

Acknowledgments

Discussions with Ron Elber and Aljaž Godec are gratefully acknowledged. This work was supported by the Robert A. Welch Foundation (Grant No. F- 1514 to DEM), the National Science Foundation (Grant Nos. CHE 1955552 to DEM and IIS 2212048 to EV), and by the Adobe Inc. AD acknowledges the computing facilities provided by the CSIR-NCL.

Appendix A: Details of molecular simulations of Gly-Ser repeat peptide.

All-atom molecular dynamics simulations of this peptide were performed using the GROMACS software package⁶⁹ with the CHARMM27 parameter set⁷⁰. The peptide was solvated with TIP3P water molecules in a cubic box so that the minimal distance of the peptide from any edge of the cubic box was at least 10 Å, which gave us a box dimension of $36.6 \times 36.6 \times 36.6$ Å³. Electroneutrality of the system was achieved by randomly replacing water molecules with Na⁺ and Cl⁻ ions keeping the salt concentration fixed at 100 mM using the genion module of GROMACS. The structure of the peptide was energy-minimized by using the steepest descent algorithm to generate a configuration with no steric hindrance. This conformation was then subjected to two consecutive equilibration phases – 500 ps of NVT equilibration at 300 K using the modified Berendsen weak coupling method⁷¹ and 500 ps of NPT equilibration at 300 K and 1.0 bar using the Parrinello–Rahman barostat⁷². For both equilibration phases, the peptide was kept at the center of the cubic box using a force constant of $1000 \text{ kJ mol}^{-1} \text{ nm}^{-2}$, and position restraints applied to all the heavy atoms throughout. The system was then allowed to evolve freely for 12 microseconds. Other simulation parameters are as follows: (1) 2 fs integration time step; (2) 5 ps resolution for storing snapshots, (3) 10 step resolution for the nonbonded interactions list update, (4) LINCS algorithm⁷³ to constrain hydrogen atom containing bonds,

with a warning angle of 30° , (5) temperature baths with a relaxation constant of 0.1 ps, (6) barostats with a relaxation constant of 1.0 ps and isothermal compressibility of $4.5 \times 10^{-5} \text{ bar}^{-1}$, (7) periodic boundary conditions with the standard minimum image convention in all three directions, (8) particle mesh Ewald method with a real-space cutoff at 9 Å to treat the long-range electrostatics, and (9) a 9 Å cutoff for Lennard-Jones interactions with a 1.2 Å Fourier spacing.

Appendix B. Details of entropy estimation and error analysis from a molecular trajectory.

Although we are working with a single molecular trajectory for a Gly-Ser repeat peptide, the method of ref. ⁶⁵ introduces additional stochastic component by adding missed milestone crossings probabilistically. Starting with the same molecular trajectory, we thus apply this method to generate 500 different milestone sequences. For each such sequence, we estimate the k -th order entropy by (i) measuring the probabilities $T^{(k)}$ (c.f. Eq. 3), (ii) sampling a k -th order Markov process according to the measured transition probabilities, and (iii) applying the compression method as described in ref. ³¹ to obtain an estimate for $h_M^{(k)}$ (green data in Fig. 5). The infinite-order estimates for h_M are obtained by applying the compression method to the original milestone sequences. The “histogram estimates” (purple data in Fig. 5) are obtained by applying Eq. 5 to the measured transition probabilities $T^{(k)}$. The error bars reported in Fig. 5 represent the 5th and 95th percentile values. We emphasize that those only reflect the statistical errors arising due the stochastic procedure of ref. ⁶⁵. Other errors such as

compression algorithm errors or the statistical errors arising from the finite length of the molecular trajectory itself cannot be estimated this way. Finally, the same procedure is applied to estimate the information rates corresponding to a Langevin trajectory (Eq. 34) of the same length as the molecular trajectory (red and orange curves in Fig. 5).

Appendix C: Accuracy of entropy estimates from short trajectories

As observed in Fig. 5, the compression-based estimates of the entropy rates for a molecular trajectory are significantly higher than the estimates based on the direct use of Eq. 5. We propose that this discrepancy results from an insufficient simulation time of the Gly-Ser peptide. As much longer molecular simulation trajectories are unavailable to us here, below we explore how the finite trajectory length affects the performance of both methods using the previously studied model of a single-file random walk³¹, for which very long trajectories can be obtained. This system has three discrete sites on a ring (Fig. 6). At every discrete timestep, one of the two walkers (with equal probability) moves into the vacant site. The trajectory of one selected walker is a highly non-Markov process, with the exact information rate known to be $h_M = 1$ bit per step³¹.

We simulate such a trajectory for $N=45000$ timesteps, to create data of a similar length to the milestone trajectory, and then compare the compression-based and Equation 5-estimated entropy values in Figure 6. A comparison between the information rates estimated, as in Fig. 5, for a short trajectory ($N = 45000$) with that for a much longer trajectory ($N = 2 \times 10^7$) shows

that, indeed, the compression algorithm overestimates the information rates, but its error is significantly reduced for longer trajectories, a behavior similar to that observed in Fig. 5.

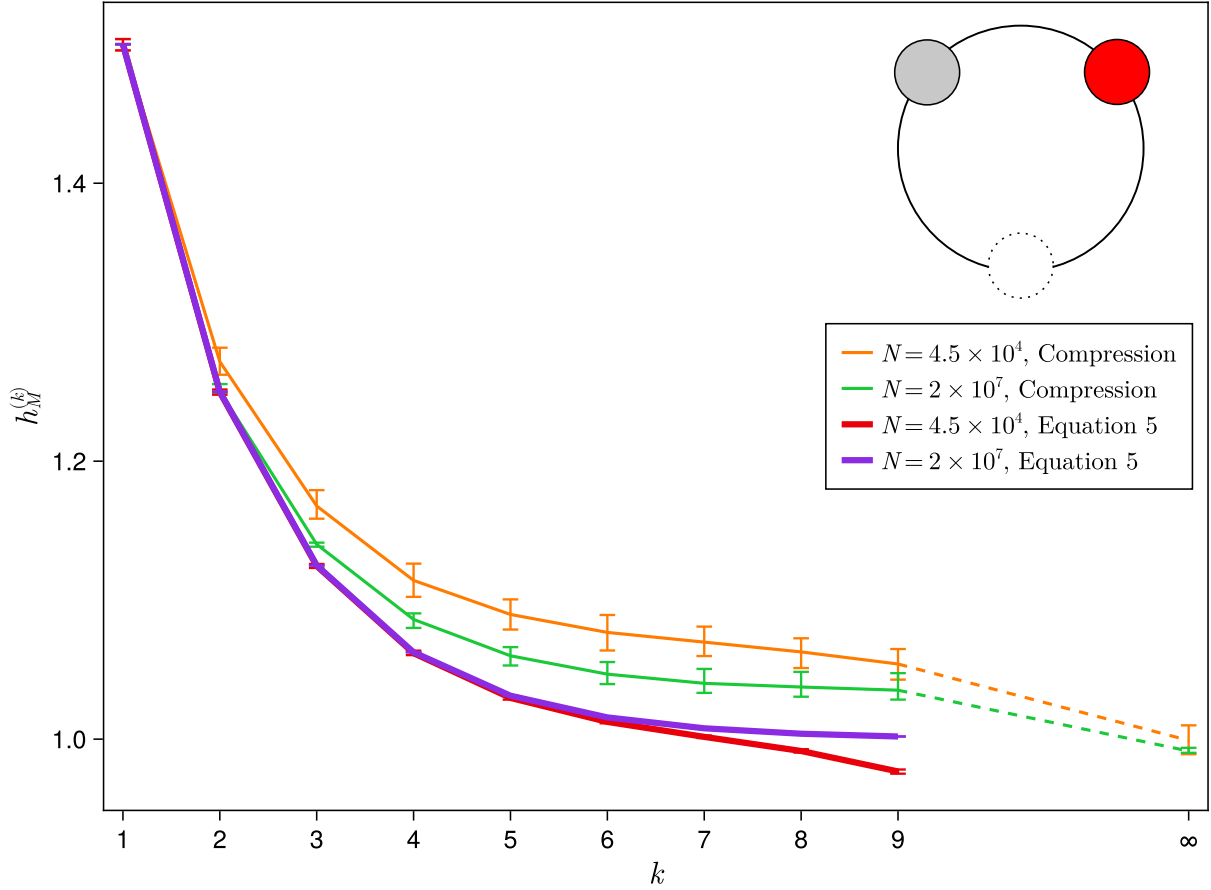


Figure 6: Entropy rates for a discrete single-file random walk of a tagged particle on a ring – see inset. Comparison of Eq.5-based and compressor-based estimates for the entropy rates computed for a short trajectory ($N=45000$) and long trajectory ($N = 2 \times 10^7$) shows that the discrepancy between the two estimates is due to insufficient length of the trajectory. Note that, for the longer trajectory the $h_M^{(k)}$ values approach the theoretical value ($h_M = 1$) as the Markov order k increases. The error bars reported here have been obtained from 100 simulated single-file random walk trajectories of given length N .

Appendix D: Alternative Markov order estimates

Once we have obtained estimates of $T_{i_{k+1}, i_k, \dots, i_1}^{(k)}$ and $p_{i_k, \dots, i_1}^{ss}$, there are several possible estimators of Markov order that can be computed. For example, a k^{th} order Markov process obeys the k^{th} order Markov property, namely that:

$$T_{i_{k+1}, i_k, \dots, i_1}^{(k)} = T_{i_{k+1}, i_k, \dots, i_1, i_0}^{(k+1)} \quad (\text{D1})$$

It also must obey the generalized form of the Chapman-Kolmogorov equations, the simplest form of which can be written as

$$T_{i_{k+1}, i_{k-1}, \dots, i_0}^{(k)} = \sum_{i_k} T_{i_{k+1}, i_k, \dots, i_1}^{(k)} T_{i_k, i_{k-1}, \dots, i_0}^{(k)}, \quad (\text{D2})$$

where $T_{i_{k+1}, i_{k-1}, \dots, i_0}^{(k)}$ is defined as the probability of i_{k+1} conditional on observing the sequence $i_0, \dots, i_{k-1}$.

By summing these quantities over all possible values and taking the difference between the right hand side and left hand side, we can quantify how non-Markov the process is at order k . For Equation D1, we can write this as

$$\Delta M^{(k)} = \sum_{i_{k+1}, i_k, \dots, i_0} |T_{i_{k+1}, i_k, \dots, i_1}^{(k)} - T_{i_{k+1}, i_k, \dots, i_1, i_0}^{(k+1)}| \quad (\text{D3}).$$

If a process is k^{th} order Markov, then $\Delta M^{(k)}$ should be zero (note, however, that this is a necessary but not sufficient condition). A similar expression can be written for Equation D2 – in the special case of $k = 2$, this would be equivalent to taking a matrix norm.

We examined these estimators for the three systems studied in this paper, as well as for a coarse-grained random walk, which is known to be second-order Markov³¹. As shown in Figure 7, the estimator correctly indicates that the overdamped Langevin system is at least 1st order, and the coarse-grained random walk is at least second order. In fact, the former is exactly 1st order and the latter is exactly 2nd order³¹, and thus we should expect that $\Delta M^{(k)}$ should be zero for all higher orders. However, in practice, the error rapidly accumulates past $k=3$, and the estimator incorrectly reports the Langevin system as non-Markov at third or higher orders. In contrast, comparing these observations to Figure 5, we can see that both Equation 5 and the compression method correctly report that the Langevin system is Markov to order 9 without any significant loss of statistics.

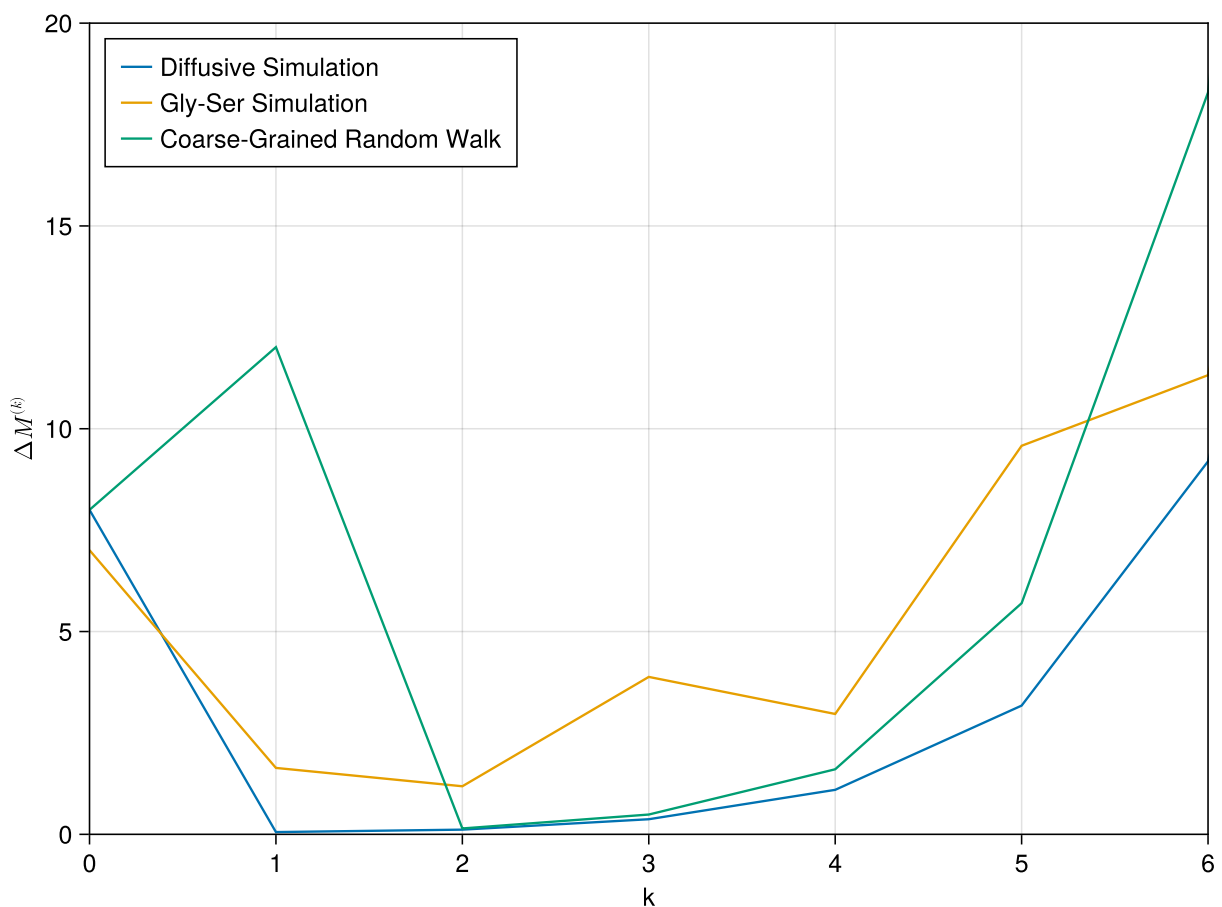


Figure 7: An attempt to quantify the Markov order of various systems using Equation

D3. A nonzero value at k indicates that the process is not Markov at that order. The estimator is accurate at low orders, correctly indicating that the overdamped Langevin system is 1st order Markov and the coarse-grained random walk is second-order³¹. However, the error rapidly grows for higher orders, and incorrectly reports that these systems are not Markov to any order higher than $k = 3$. Similar behavior was observed when using the Chapman-Kolmogorov equations (Eq. D2). In contrast, Equation 5 and the compression technique both correctly show that overdamped Langevin dynamics is Markov even at 9th order (Fig. 5).

References

1. R. Zwanzig, *Nonequilibrium Statistical Mechanics*. (Oxford University Press, 2001).
2. H. A. Kramers, *Physica* **7**, 284-304 (1940).
3. A. Szabo, K. Schulten and Z. Schulten, *J Chem Phys* **72**, 4350 (1980).
4. O. K. Dudko, G. Hummer and A. Szabo, *Phys Rev Lett* **96** (10), 108101 (2006).
5. O. K. Dudko, G. Hummer and A. Szabo, *Proceedings of the National Academy of Sciences of the United States of America* **105** (41), 15755-15760 (2008).
6. R. B. Best and G. Hummer, *Phys Chem Chem Phys* **13** (38), 16902-16911 (2011).
7. A. B. Kolomeisky, *Motor Proteins and Molecular Motors*. (CRC Press, 2015).
8. H. Y. Aviram, M. Pirchi, Y. Barak, I. Riven and G. Haran, *J Chem Phys* **148** (12), 123303 (2018).
9. J. N. Taylor, M. Pirchi, G. Haran and T. Komatsuzaki, *J Chem Phys* **148** (12), 123325 (2018).
10. Z. Kilic, I. Sgouralis, W. Heo, K. Ishii, T. Tahara and S. Presse, *Cell Rep Phys Sci* **2** (5), 100409 (2021).
11. Z. Kilic, I. Sgouralis and S. Presse, *Biophys J* **120** (9), 1665-1679 (2021).
12. R. Satija and D. E. Makarov, *J Phys Chem B* **123** (4), 802-810 (2019).
13. J. O. Daldrop, J. Kappler, F. N. Brunig and R. R. Netz, *Proc Natl Acad Sci U S A* **115** (20), 5169-5174 (2018).
14. A. Lapolla and A. Godec, *J Chem Phys* **153** (19), 194104 (2020).
15. F. Glatzel and T. Schilling, *Europhysics Letters* **136** (3), 36001 (2021).
16. C. Widder, F. Koch and T. Schilling, *J Chem Phys* **157** (19), 194107 (2022).
17. C. Ayaz, L. Tepper, F. N. Brunig, J. Kappler, J. O. Daldrop and R. R. Netz, *Proc Natl Acad Sci U S A* **118** (31) (2021).

18. R. Satija, A. Das, S. Muhle, J. Enderlein and D. E. Makarov, *J Phys Chem B* **124** (17), 3482-3493 (2020).
19. S. Presse, J. Lee and K. A. Dill, *J Phys Chem B* **117** (2), 495-502 (2013).
20. S. Presse, J. Peterson, J. Lee, P. Elms, J. L. MacCallum, S. Marqusee, C. Bustamante and K. Dill, *J Phys Chem B* **118** (24), 6597-6603 (2014).
21. S. M. Avdoshenko, A. Das, R. Satija, G. A. Papoian and D. E. Makarov, *Sci Rep* **7** (1), 269 (2017).
22. R. Satija, A. Das and D. E. Makarov, *The Journal of Chemical Physics* **147** (15), 152707 (2017).
23. I. Grossman-Haham, G. Rosenblum, T. Namani and H. Hofmann, *Proc Natl Acad Sci U S A* **115** (3), 513-518 (2018).
24. C. B. Li, H. Yang and T. Komatsuzaki, *Proc Natl Acad Sci U S A* **105** (2), 536-541 (2008).
25. C. B. Li, H. Yang and T. Komatsuzaki, *J Phys Chem B* **113** (44), 14732-14741 (2009).
26. A. M. Berezhkovskii and D. E. Makarov, *J Phys Chem Lett* **9** (9), 2190-2195 (2018).
27. R. Satija, A. M. Berezhkovskii and D. E. Makarov, *Proc Natl Acad Sci U S A* **117** (44), 27116-27123 (2020).
28. N. Zijlstra, D. Nettels, R. Satija, D. E. Makarov and B. Schuler, *Phys. Rev. Letters* **125**, 146001 (2020).
29. D. E. Makarov, *J Phys Chem B* **125** (10), 2467-2476 (2021).
30. A. Lapolla and A. Godec, *Physical Review Research* **3** (2), L022018 (2021).
31. K. Song, D. E. Makarov and E. Vouga, *Physical Review Research* **5** (1), L012026 (2023).
32. C. E. Shannon, *Bell System Technical Journal* **January**, 57-64 (1951).
33. J. Ziv and A. Lempel, *IEEE transactions on information theory* **23** (3), 337 (1977).
34. J. V. Stone, *Information theory: A Tutorial Introduction*. (Sebtel Press, 2015).
35. P. Gaspard and X.-J. Wang, *Physics Reports* **235** (6), 291-343 (1993).
36. R. Elber, D. E. Makarov and H. Orland, *Molecular Kinetics in Condense Phases: Theory, Simulation, and Analysis*. (Wiley and Sons, 2020).
37. D. E. Makarov, *Single Molecule Science: Physical Principles and Models*. (CRC Press, Taylor & Francis Group, Boca Raton, 2015).
38. R. Elber, *J Chem Phys* **144** (6), 060901 (2016).
39. C. Schutte, F. Noe, J. Lu, M. Sarich and E. Vanden-Eijnden, *J Chem Phys* **134** (20), 204105 (2011).
40. D. Hartich and A. Godec, *Phys. Rev. X* **11**, 041047 (2021).
41. D. Bhatt and D. M. Zuckerman, *J Chem Theory Comput* **7** (8), 2520-2527 (2011).
42. J. Klafter and I. M. Sokolov, *First steps in random walks : from tools to applications*. (Oxford University Press, Oxford ; New York, 2011).
43. R. Metzler and J. Klafter, *Physics Reports* **339**, 1-77 (2000).
44. A. M. Berezhkovskii and A. Szabo, *J Chem Phys* **150** (5), 054106 (2019).
45. S. Redner, *A Guide to First Passage Times*. (Cambridge University Press, 2001).
46. A. M. Berezhkovskii and D. E. Makarov, *Biophysical Reports* **1**, 100029 (2021).
47. D. E. Makarov, *J. Chem. Phys.* **138**, 014102 (2013).
48. E. Medina, R. Satija and D. E. Makarov, *J Phys Chem B* **122** (49), 11400-11413 (2018).
49. T. Schurmann and P. Grassberger, *Chaos* **6** (3), 414-427 (1996).

50. L. J. Lapidus, W. A. Eaton and J. Hofrichter, *Proc. Natl. Acad. Sci. USA* **97** (13), 7220 (2000).
51. O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello and T. Kiefhaber, *Proceedings of the National Academy of Sciences of the United States of America* **96** (17), 9597-9601 (1999).
52. A. Soranno, B. Buchli, D. Nettels, R. R. Cheng, S. Muller-Spath, S. H. Pfeil, A. Hoffmann, E. A. Lipman, D. E. Makarov and B. Schuler, *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17800-17806 (2012).
53. A. Soranno, A. Holla, F. Dingfelder, D. Nettels, D. E. Makarov and B. Schuler, *Proc Natl Acad Sci U S A* **114** (10), E1833-E1839 (2017).
54. Z. Wang and D. E. Makarov, *J. Phys. Chem. B* **107**, 5617 (2003).
55. D. Thirumalai, *J. Phys. Chem. B* **103**, 608 (1999).
56. R. R. Cheng, T. Uzawa, K. W. Plaxco and D. E. Makarov, *The journal of physical chemistry. B* **113** (42), 14026-14034 (2009).
57. R. R. Cheng, T. Uzawa, K. W. Plaxco and D. E. Makarov, *Biophysical Journal* **99** (12), 3959-3968 (2010).
58. R. W. Pastor, R. Zwanzig and A. Szabo, *J. Chem. Phys.* **105** (9), 3878-3882 (1996).
59. N. M. Toan, G. Morrison, C. Hyeon and D. Thirumalai, *J Phys Chem B* **112** (19), 6094-6106 (2008).
60. Y. Kantor and M. Kardar, *Physical review. E, Statistical, nonlinear, and soft matter physics* **76** (6 Pt 1), 061121 (2007).
61. B. Friedman and B. O'Shaughnessy, *Macromolecules* **26**, 4888-4898 (1993).
62. R. R. Cheng, A. T. Hawk and D. E. Makarov, *J. Chem. Phys.* **138**, 074112 (2013).
63. R. R. Cheng and D. E. Makarov, *The Journal of chemical physics* **134** (8), 085104 (2011).
64. A. Das and D. E. Makarov, *J Phys Chem B* **122**, 9049-9060 (2018).
65. K. Song, D. E. Makarov and E. Vouga, *The Journal of Chemical Physics* **158** (11), 111101 (2023).
66. G. Boffetta, M. Cencini, M. Falcioni and A. Vulpiani, *Physics Reports* **356** (6), 367-474 (2002).
67. A. T. Hawk, *J Chem Phys* **138** (15), 154105 (2013).
68. A. T. Hawk and D. E. Makarov, *The Journal of chemical physics* **135** (22), 224109 (2011).
69. S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess and E. Lindahl, *Bioinformatics* **29** (7), 845-854 (2013).
70. P. Bjelkmar, P. Larsson, M. A. Cuendet, B. Hess and E. Lindahl, *J Chem Theory Comput* **6** (2), 459-466 (2010).
71. G. Bussi, D. Donadio and M. Parrinello, *J Chem Phys* **126** (1), 014101 (2007).
72. M. Parrinello and A. Rahman, *Journal of Applied Physics* **52** (12), 7182-7190 (1981).
73. B. Hess, *J Chem Theory Comput* **4** (1), 116-122 (2008).