PNAS RESEARCH ARTICLE | APPLIED PHYSICAL SCIENCES

Likelihood-based non-Markovian models from molecular dynamics

Hadrien Vroylandt[a], Ludovic Goudenège[b], Pierre Monmarché[c,d], Fabio Pietrucci[e], and Benjamin Rotenberg[f,1]

Edited by Peter Rossky, Rice University, Houston, TX; received October 14, 2021; accepted February 16, 2022

We introduce a method to accurately and efficiently estimate the effective dynamics of collective variables in molecular simulations. Such reduced dynamics play an essential role in the study of a broad class of processes, ranging from chemical reactions in solution to conformational changes in biomolecules or phase transitions in condensed matter systems. The standard Markovian approximation often breaks down due to the lack of a proper separation of time scales, and memory effects must be taken into account. Using a parametrization based on hidden auxiliary variables, we obtain a generalized Langevin equation by maximizing the statistical likelihood of the observed trajectories. Both the memory kernel and random noise are correctly recovered by this procedure. This data-driven approach provides a reduced dynamical model for multidimensional collective variables, enabling the accurate sampling of their long-time dynamical properties at a computational cost drastically reduced with respect to all-atom numerical simulations. The present strategy, based on the reproduction of the dynamics of trajectories rather than the memory kernel or the velocity-autocorrelation function, conveniently provides other observables beyond these two, including, e.g., stationary currents in nonequilibrium situations or the distribution of first passage times between metastable states.

generalized Langevin equation | coarse-grained models | maximum likelihood | data-driven parametrization

Significance

The analysis of complex systems with many degrees of freedom generally involves the definition of low-dimensional collective variables more amenable to physical understanding. Their dynamics can be modeled by generalized Langevin equations, whose coefficients have to be estimated from simulations of the initial high-dimensional system. These equations feature a memory kernel describing the mutual influence of the low-dimensional variables and their environment. We introduce and implement an approach where the generalized Langevin equation is designed to maximize the statistical likelihood of the observed data. This provides an efficient way to generate reduced models to study dynamical properties of complex processes such as chemical reactions in solution, conformational changes in biomolecules, or phase transitions in condensed matter systems.

In different branches of science, the interpretation and mathematical modeling of both experimental and computational data requires the analysis of the system dynamics in terms of a reduced set of collective variables (CVs), or order parameters. Prominent examples include chemical reactions in solution, conformational changes in biomolecules, or phase transitions in condensed matter systems. A standard approach is to approximate the evolution of the CVs by an effective dynamics, namely, a closed equation in which the degrees of freedom beyond the CVs (forming the so-called environment or "bath") do not appear explicitly. Such coarse-grained models not only provide a physical interpretation more accessible to understanding than the full system but also, from a numerical perspective, enable one to recover the desired dynamical properties with long but cheap simulations of the reduced system (while only shorter simulations of the large system are used to determine the effective dynamics).

The most widespread model for this task is the Langevin equation, which can be derived—in some particular cases—from the Hamiltonian dynamics of a small system interacting with a large environment. It describes the evolution of a Markov process, which requires that the decorrelation time of the environment be short compared to the characteristic times of the reduced system. However, many cases do not enter the validity range of this approximation, displaying memory effects (1–8). To go beyond the Markovian approximation, a popular class of processes is given by the generalized Langevin equation (GLE) (9–15)

$$\begin{cases} \dot{x}(t) = v(t) \\ M\dot{v}(t) = F_{\text{eff}}(x(t)) - \displaystyle\int_0^t K(t-\tau)v(\tau)\mathrm{d}\tau + R(t), \end{cases} \tag{1}$$

where $x(t)$ is the value of the d-dimensional CV at time t; $v(t)$ is its time derivative; M is an effective mass, F_{eff} is a mean force, usually deriving from a potential V identified with the free energy; K is a memory kernel; and $R(t)$ is a (colored) noise.

This form of the GLE can be motivated from the dynamics of the original full system following the Mori–Zwanzig formalism (9, 16–18), even though it cannot be formally obtained as a controlled approximation of the exact coarse-grained dynamics, since a rigorous derivation generally results in a memory kernel that depends on the CVs (19, 20). Nevertheless, in practice, this simple form is the most widely used effective dynamics. While an analytical derivation of the memory kernel is possible only in a few

Author contributions: H.V., L.G., P.M., F.P., and B.R. designed research; H.V. performed research; H.V., L.G., P.M., F.P., and B.R. analyzed data; and H.V., L.G., P.M., F.P., and B.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: benjamin.rotenberg@sorbonne-universite.fr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2117586119/-/DCSupplemental.

Published March 23, 2022.

cases (21), for more general systems, K can be estimated from a data-driven approach. In most cases, the goal is to extract the memory kernel from trajectories of the CV computed with all-atom simulations (4, 22–37).

As already mentioned, the solutions of Eq. **1** are not Markov processes, except when K is the Dirac δ function and R is a white noise. Both for fitting the model and then for generating new trajectories of the effective dynamics, it is convenient to consider the subclass of models where an extended process (x, v, h) is Markovian, with h some hidden auxiliary variables (12, 38–45). Restricting further to the case where the evolution of the hidden variables and the coupling with the observed variables are linear, this leads to an equation of the form

$$\begin{cases} \dot{x} = v \\ \dot{v} = M^{-1}F_{\text{eff}}(x) - A_{vh}h - A_{vv}v + \sigma_{vv}\xi(t) + \sigma_{vh}W(t) \\ \dot{h} = \qquad\qquad\quad -A_{hh}h - A_{hv}v + \sigma_{vh}^{\mathrm{T}}\xi(t) + \sigma_{hh}W(t), \end{cases} \qquad [2]$$

where $A_{vh}, A_{vv}, A_{hh}, A_{hv}, \sigma_{vv}, \sigma_{vh}, \sigma_{hh}$ are constant matrices and ξ and W are independent standard white noises. This gives a convenient class of models parametrized by the dimension d_h of h, the corresponding matrices, and the (rescaled) effective force $M^{-1}F_{\text{eff}}$. For equilibrium processes, the coefficients of Eq. **2** are related by the so-called fluctuation–dissipation relation (40). Although we could enforce this condition, thereby reducing the number of parameters, we do not, since we also consider nonequilibrium systems in the following.

Integrating over the hidden variables, we recover Eq. **1** with a memory kernel of the form of a finite Prony series (40, 41)

$$K(\tau) = w_0\delta(\tau) + \sum_{k=1}^{d_h} w_k e^{-\lambda_k \tau}, \qquad [3]$$

where w_k and λ_k are (possibly complex) coefficients of the series derived from the matrices $A_{vh}, A_{vv}, A_{hh}, A_{hv}$. In principle, on all finite time intervals, any kernel given as the sum of a Dirac function at zero and of a continuous function can be approximated arbitrarily accurately by a sum of the form Eq. **3**. However, in practice d_h is relatively small, and a memory kernel with, e.g., an algebraic tail can only be approximated on a small time interval (38, 45).

The use of auxiliary variables in the form of Eq. **2** has been abundantly used and studied as it allows efficient integration of GLE Eq. **1** (40, 42, 46), even though other methods exist (28, 29, 33, 47, 48). The estimation of GLE parameters from simulations is an active field of research. The main method consists in a nonparametric estimation of the memory kernel via the Volterra integral equation (12, 25, 34, 37, 39, 46, 49, 50), but other methods have also been proposed (24, 32, 44, 45, 51). In the present work, we 1) introduce a parametric estimator of GLE coefficients, based on a maximum likelihood approach, and 2) show that it allows building faithful coarse-grained models of MD simulations in a cost-effective way (i.e., starting from a relatively small training data set), such that the dynamics is well reproduced.

Data-Driven Approach on Extended Dynamics

In statistics, a standard method to deal with hidden variables is the expectation-maximization (EM) algorithm, which belongs to the category of likelihood maximization algorithms (52, 53). It is frequently used to estimate parameters of time series models in the case of partial or noisy observations of the system, for either hidden Markov models (54) or state-space models (55). A first application in the context of GLE was proposed in ref. 38 to reconstruct the memory kernel in the absence of effective force $F_{\text{eff}}(x)$ and under more restrictive conditions than the method presented below.

The algorithm proceeds by alternating steps: in the E step, one determines the conditional probability law of the hidden variables given the observed ones at fixed parameters; in the M step, one optimizes the parameters to maximize the log-likelihood averaged with respect to these conditional laws. In the following we denote as Θ_j the whole set of parameters estimated after j iterations of the algorithm, which includes the mean force projected on some functional basis (which can be very large in general or reduced if prior knowledge on the system is available), the coefficients of the matrices A, σ of Eq. **2**, and, for technical reasons discussed below, the mean value at time 0 of the hidden variables, $\langle h_0 \rangle$.

EM Algorithm. The available data, obtained from all-atom simulations, consists of a set of independent trajectories. For simplicity of the notation, we introduce the algorithm with only one trajectory $\{x\}_{0:N} = \{x(k\Delta t), k \in [\![0, N]\!]\}$ for some time step Δt and simulation time $T = N\Delta t$, the extension to the general case being straightforward. The statistical models we consider are Euler–Maruyama discretizations of Eq. **2** with the same time step Δt, for a fixed dimension d_h of auxiliary variables h. The state of the system at time $t = k\Delta t$ will be denoted $(x_k, v_k, h_k) = (\widetilde{X}, h)_k = X_k$, and we write $\{X\}_{0:N}$ as a complete trajectory of the system. Hence, $\widetilde{X}$ is the value of the known variables since from the choice of the Euler–Maruyama scheme, the velocity can be computed as $v_k = (x_{k+1} - x_k)/\Delta t$. In the following we write $\pi(z)$ as the probability density of a variable z; $\pi(z|u)$ as the conditional probability density of z with respect to u; and, in both cases, π_Θ to make explicit the value of the parameters if needed.

As the extended system is Markovian, we have for the probability density of a trajectory

$$\pi\left(\{X\}_{0:N}\right) = \pi\left(X_0\right) \times \prod_{k=0}^{N-1} \pi(X_{k+1}|X_k), \qquad [4]$$

and the form of Eq. **2** and of the Euler–Maruyama scheme lead to a Gaussian transition kernel, characterized by its mean μ and variance Σ (*SI Appendix*).

E Step. The first step is to compute the conditional law of the hidden variables given the observed variables at the current guess of the parameters, i.e., $\pi_{\Theta_j}(\{h\}_{0:N}|\{\widetilde{X}\}_{0:N})$. Due to the Markovianity of the extended system, it is sufficient to compute the mean and variance of the Gaussian marginal laws $\pi(h_k, h_{k+1}|\{\widetilde{X}\}_{0:N})$ for all $k \in [\![0, N-1]\!]$. Taking advantage of the explicit form of the transition probability (*SI Appendix*, Eq. **3**), we apply an iterative predictor–corrector–smoother approach (also known as Kalman filter and Rauch–Tung–Striebel smoother) (56). Starting from the trajectory up to step $k-1$, we determine the law of the hidden variable h_k conditioned on the past information $\{\widetilde{X}\}_{0:k-1}$ only. We then use the expression of the transition probability $\pi(X_k|X_{k-1})$ to determine the current value of $\pi\left(h_k|\{\widetilde{X}\}_{0:k}\right)$. These are the prediction and correction parts that are run forward on the trajectory, i.e., from $k = 0$ to $k = N$ (arrow 1 in Fig. 1). The initial guess at $k = 0$ of $\pi\left(h_0\right)$ uses the measured $\langle h_0 \rangle$ vector as the mean and an arbitrary variance (identity matrix). Such initial guess could be optimized, but we did not observe any influence on the final results. The second part of the E step, called the smoother part, computes $\pi(h_{k-1}|h_k, \{\widetilde{X}\}_{0:N})$

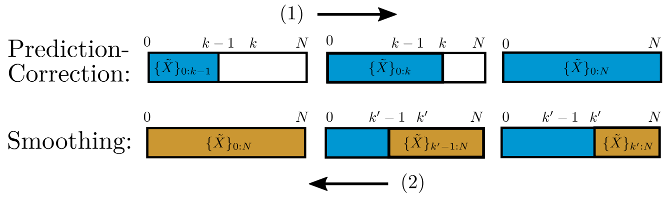

Fig. 1. E step. (1) We first predict iteratively the history of h_k for the whole trajectory, using a predictor-corrector. (2) The values of $h_{k'}$ are then smoothed iteratively backward from the end of the trajectory (see *Data-Driven Approach on Extended Dynamics*).

and is run backward, i.e., from $k = N$ to $k = 0$ (arrow 2 in Fig. 1), which finally gives the required probability law of h_{k-1}, h_k conditioned to the full observed trajectory. Detailed formulas are presented in *SI Appendix*.

M Step. For any set of parameters Θ, introduce the evidence lower bound $\mathcal{L}^j_{LB}$ after the jth iteration of the algorithm as the expectation with respect to $\pi_{\Theta_j}(\{h\}_{0:N}|\{\widetilde{X}\}_{0:N})$ of the log-likelihood of the full trajectory $\{X\}_{0:N}$ with parameter Θ, namely (see derivation in *SI Appendix*),

$$\begin{aligned}\mathcal{L}^j_{LB}(\Theta) &= \int \pi_{\Theta_j}(\{h\}_{0:N}|\{\widetilde{X}\}_{0:N}) \ln \pi_\Theta(\{X\}_{0:N})\, \mathrm{d}\{h\}_{0:N}\\ &= \int \pi_{\Theta_j}(h_0|\{\widetilde{X}\}_{0:N}) \ln \pi_\Theta(X_0) \mathrm{d}h_0\\ &\quad + \sum_{k=0}^{N-1} \int \pi_{\Theta_j}(h_k, h_{k+1}|\{\widetilde{X}\}_{0:N})\\ &\quad \times \ln \pi_\Theta(X_{k+1}|X_k) \mathrm{d}h_k \mathrm{d}h_{k+1}. \qquad [5]\end{aligned}$$

The M step consists in setting Θ_{j+1} to be the maximizer of this quantity. Notice that due to the particular form of Eq. **2**, $\mathcal{L}^j_{LB}(\Theta)$ is an explicit function of Θ that can be easily optimized as described in *SI Appendix*.

Full Algorithm. The algorithm is then run as follows. An initial random or informed guess Θ_0 is taken for the parameters. Such an informed guess could come from a previous execution of the algorithm with a different number of hidden dimensions. From parameters Θ_j, a new set of parameters Θ_{j+1} is computed through an iteration of E and M steps. Since maximizing the evidence lower bound $\mathcal{L}^j_{LB}$ increases the observed likelihood, the method is iterated until either a prescribed maximum number of EM steps or a convergence criterion is reached.

Assessing the Quality of a Given Model. The number of hidden dimensions d_h is an important parameter of the algorithm. It can be chosen using a model validation approach, classically by dividing the set of trajectories between a training and a validation set. However, here we simply compute the optimal parameters for several values of d_h and compare the predictions of the corresponding models for a number of observable properties, such as the memory kernel, velocity-autocorrelation functions (VACF), or first passage times (FPTs). Similarly, the quality of the model depends on the time step used for the coarse-grained dynamics. This choice depends among other things on the numerical scheme for the propagator. For a given underlying dynamics of the full system, the most accurate choice for the coarse-grained one is to use the same time step Δt_{full}, but as a compromise for the amount of data, one can also use $\Delta t = m\Delta t_{full}$ (i.e., using only every m step), with m a small integer.

Efficient Sampling of New Trajectories. Once the model has been optimized by the EM algorithm, it can be used to generate new trajectories in the CV space. Due to their limited computational cost compared to MD trajectories, such synthetic data grant easier access to well-converged average properties, in the form of static and dynamic observables. As an example, in *Results* the mean FPTs (as well as their probability densities) of a Lennard–Jones (LJ) dimer in a bath are estimated based on the GLE model and compared with the corresponding ones extracted from expensive MD simulations.

Results

We first present the result of the algorithm on a simple yet nontrivial test case with a one-dimensional (1D) system following the extended dynamics of Eq. **2**, with five hidden dimensions and a quadratic potential well $V(x) = x^2/2$, using 20 trajectories of 2.5×10^4 steps and a time step of 5×10^{-3}. The effective force is fitted as a linear function of *x*. Fig. 2*A* compares the result of our algorithm to the true memory function that can be computed from Eq. **3** and the one obtained by the Volterra method (*Materials and Methods*). It demonstrates that the present EM method is able to reproduce the true memory kernel. Furthermore, the parametric structure of the fitted model enforces the decay to zero of the memory kernel, whereas the Volterra method is unstable for times longer than a few tens. Fig. 2*B* finally shows that the method accurately reproduces the VACF.

The algorithm also applies to multidimensional and nonequilibrium systems. This is illustrated in Fig. 3 for a 2D system with two different thermal noises along each axis, with temper-

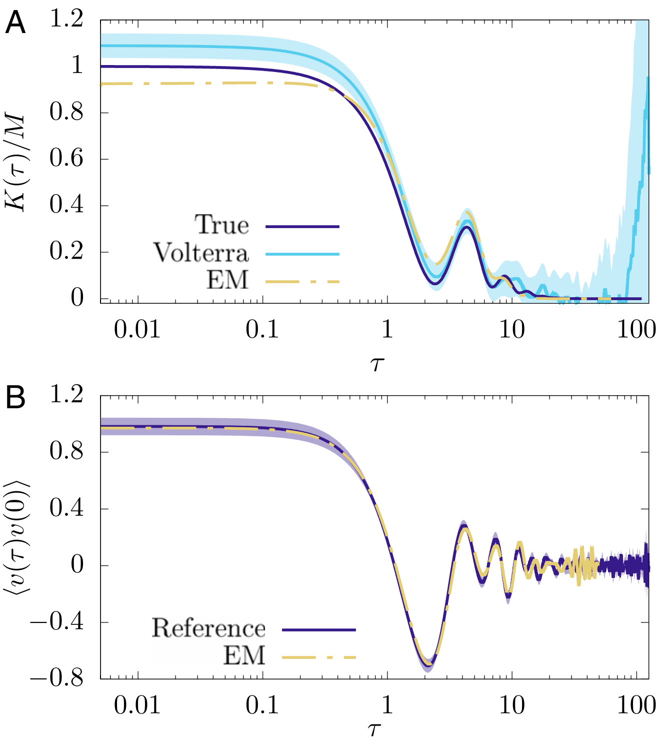

Fig. 2. Equilibrium 1D case. (*A*) Memory kernel $K(\tau)$ divided by the mass *M*: the true kernel used to generate the reference trajectories (dark blue line) is compared with the predictions of the Volterra method (cyan solid line, with shaded area indicating uncertainties computed from a bootstrap analysis) and of the present EM method (dash-dotted yellow line). (*B*) Velocity autocorrelation function, from the reference trajectories (dark blue line) and from new trajectories sampled using the fitted EM model (dash-dotted yellow line).

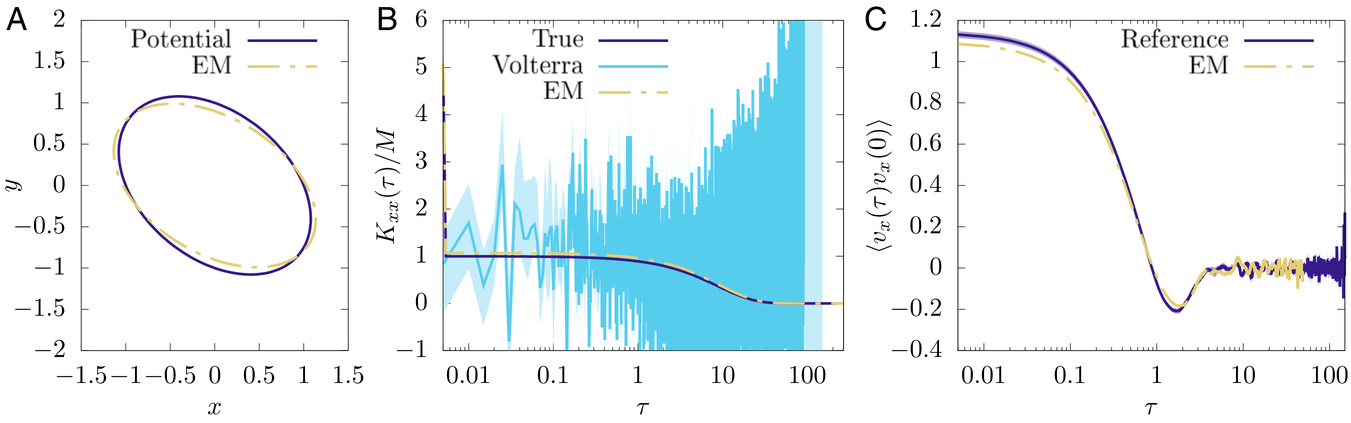

Fig. 3. Nonequilibrium 2D case. (*A*) Locus of $V(x, y) = 1$ for the original potential and the one estimated by the EM algorithm. (*B*) *xx* component of the reconstructed memory kernel $K_{xx}(\tau)$ divided by the mass *M*: the true kernel used to generate the reference trajectories (dark blue line) is compared with the predictions of the Volterra method (cyan solid line, with shaded area indicating uncertainties computed from a bootstrap analysis) and of the present EM method (dash-dotted yellow line); the left peak represents the Dirac function of Eq. **3**. (*C*) Velocity autocorrelation function (for the *x* component of the velocity), from the reference trajectories (dark blue line) and from new trajectories sampled using the fitted EM model (dash-dotted yellow line).

atures $T_x = 1$ and $T_y = 5$ and a quadratic potential $V(x, y) = \frac{1}{2}\left(x^2 + \frac{3}{4}xy + y^2\right)$ whose principal axes are not aligned with the *x* and *y* axes, leading to nonequilibrium conditions. This setup is inspired by a similar Markovian model used to describe nonequilibrium experiments on cold atoms (57). We run 20 trajectories of 3×10^4 steps with a time step of 5×10^{-3}. The effective 2D force is fitted as a linear combination of *x* and *y*. The corresponding quadratic potential, illustrated in Fig. 3*A*, is in good agreement with the one used to generate the trajectories. Fig. 3*B* then shows that the algorithm correctly estimates the memory kernel (in the present case, a simple one with a single hidden dimension for each visible dimension). In particular, the presence of a strong Markovian component is captured by the algorithm but missed by the Volterra method. Finally, the dynamics of the system is well reproduced, as demonstrated for the VACF in Fig. 3*C*.

The present approach, based on the reproduction of the dynamics of trajectories rather than the memory kernel or the VACF, conveniently provides other observables beyond these two. Indeed, by generating new trajectories corresponding to the fitted GLE model, one has in principle access to all properties that can be computed from the time evolution of the CVs. As an illustration, Fig. 4 shows for the same nonequilibrium 2D case the stationary probability distribution and the average velocity as a function of the position, estimated using either the initial trajectories used to fit the GLE model (Fig. 4*A*) or the same number of trajectories generated with the latter (Fig. 4*B*). Despite the relatively small number (only 20) of original trajectories used to fit the model and to compute the properties, those computed from the fitted GLE model are in very good agreement with the original ones.

As a final illustration, we apply our algorithm to a more realistic 3D system composed of 512 LJ particles at reduced temperature $\widehat{T} = k_B T/\epsilon = 1$ and reduced density $\widehat{\rho} = \rho\sigma^3 = 1$. Two of the LJ particles are singled out to form a dimer (24), the others constituting the solvent. The CV of interest is the distance *r* between the two particles forming the dimer. LJ parameters for all interactions are taken as $\epsilon = 1$ and $\sigma = 1$ (in LJ units), except between the two particles forming the dimer, with $\epsilon_d = 2$ and $\sigma = 1$. The size of the cubic simulation box is 8σ, with periodic boundary conditions in all directions. The dynamics is integrated with a time step of $\Delta t_{MD} = 10^{-3}$ (in LJ units) in the microcanonical ensemble (NVE; constant number of particles *N*, volume *V* and energy *E*) using the LAMMPS simulation package (58). We run 20 trajectories with length of 10^5 time steps, and CV values are extracted every 2 steps.

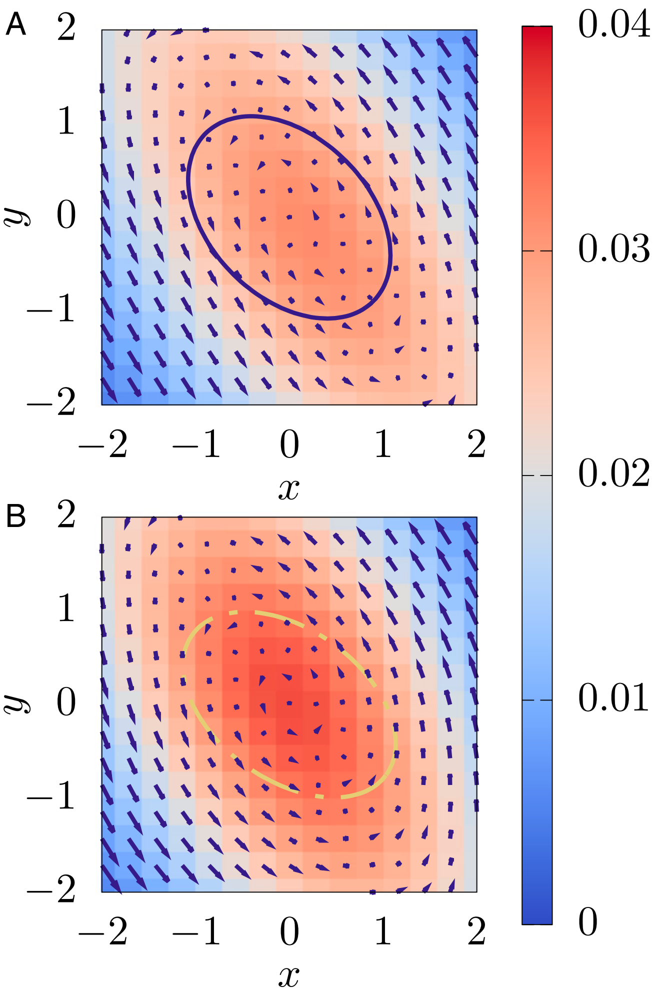

Fig. 4. Nonequilibrium 2D case: beyond the kernel and the VACF. Stationary probability distribution (colors) and average velocity (arrows) as function of the position for (*A*) the original dynamics and (*B*) the GLE model estimated by the EM algorithm. The two ellipses are the same as in Fig. 3*A* and represent the locus of $V(x, y) = 1$ for the original potential (blue line) and the one estimated by the EM algorithm (dashed yellow line).

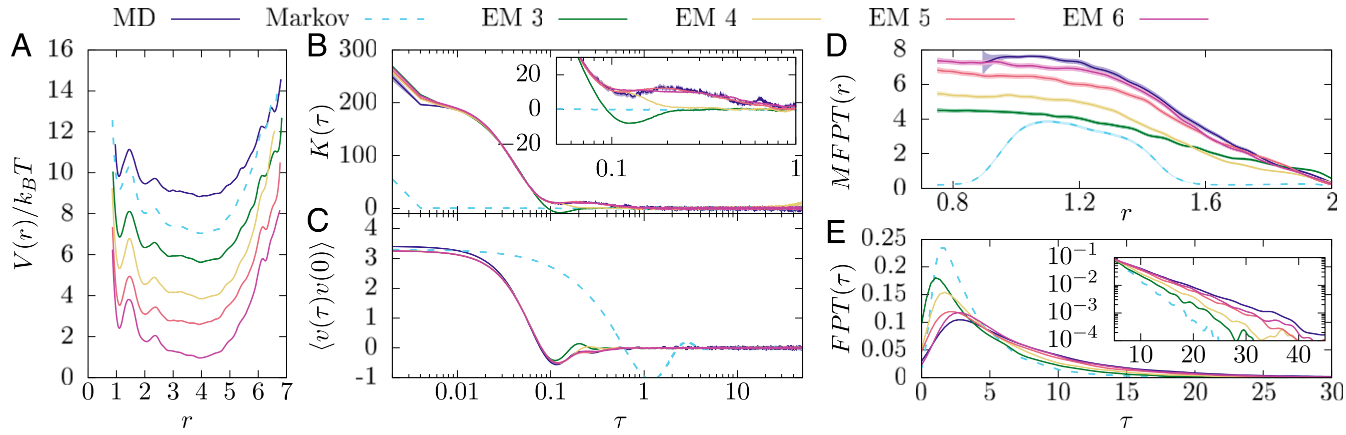

Fig. 5. LJ fluid: two solutes in an explicit solvent. In all panels, results are shown for the reference MD trajectory (dark blue line) and for trajectories generated by the estimated Markov model (dashed cyan lines) and EM models with three to six hidden dimensions (from dark green to purple). Unless specified, all quantities are in LJ units. (*A*) Free energy (in units of the thermal energy, k_BT) estimated from histograms of the distance r between the two solutes. The various curves are shifted by physically irrelevant constants for clarity. (*B*) Memory kernel and (*C*) velocity autocorrelation function. *Inset* in *B* shows a zoom on intermediate times. (*D*) Mean FPT to reach $r = 2.0$ starting from the distance r. (*E*) Distribution of the FPT for trajectories starting at $r = 2^{1/6}$ and ending at $r = 2.0$ (*Inset* shows a zoom on the tail of the distributions, on semilogarithmic scale).

We fit GLE models defined by Eq. **2** with the EM algorithm for a number of hidden dimensions ranging from three to six. In all cases, the effective force $F_{\text{eff}}(r)$ determined from the MD trajectory is used as the single function of the above-mentioned functional basis, so that fitting this part reduces to determining a single prefactor. Our aim is to test the ability of these models to reproduce, in the statistical sense, the properties of the original simulations. In order to check the importance of the hidden variables, we also provide an analysis for a Markovian model, fitted using a maximum likelihood algorithm with 0 hidden dimensions (corresponding to the M step of the above EM algorithm). For each fitted GLE model, we generate 75 new trajectories of length 10^5 time steps using Eq. **2**, to compute the observable properties and compare them with those obtained from the original set of MD trajectories.

We first compare the stationary distribution for the various GLE models in Fig. 5*A*, which shows the free energy as a function of the r coordinate computed from the histogram of each set of new trajectories (i.e., not the one corresponding to the fitted effective force). The good agreement with the MD free energy profile demonstrates 1) that the coefficient multiplying the model free energy profile of each model is fitted precisely and 2) that the numerical integration of the GLE models is performed accurately. Notice that free energy beyond $r = 4\sigma$ is affected by the size of the periodic box. The free energy displays two potential wells at $r = 1.12\sigma$ and $r = 2.00\sigma$, corresponding to the contact pair (CP, i.e., the dimer) and the solvent shared pair (SSP, with solvent atoms belonging to the solvation shells of both solutes), whose dynamics is investigated below.

We then consider dynamical observables in Fig. 5*B*, which shows the memory kernel estimated from the Volterra method (4) for MD as well as GLE trajectories, and Fig. 5*C*, which illustrates the VACF (the velocities being computed numerically from positions both in the MD and GLE trajectories). In both cases, increasing the number of hidden dimensions increases the fidelity of the model with respect to the original data. The latter are correctly reproduced for five and six hidden dimensions. The plot also shows the poor quality of the Markovian model, which confirms the necessity of introducing some hidden variables (37).

Finally, we study the transition kinetics between the CP and SSP states, as a stringent test requiring accurate reproduction of both thermodynamic and dynamic properties of the system. Fig. 5*D* represents the mean FPT to reach the SSP state starting from smaller r distances, whereas Fig. 5*E* represents the FPT distribution for trajectories starting from the CP state and reaching the SSP state. Clearly, a sufficient number of hidden dimensions (in this case five to six) allows us to quantitatively reproduce the detailed transition statistics. This demonstrates again both the importance of memory effects and the ability of the present algorithm to reconstruct an accurate GLE model.

Conclusions

In this work we addressed the construction of reduced mathematical models of the dynamics of complex molecular systems. Projecting the phase-space trajectories on a reduced set of CVs leads to a powerful framework for the prediction of thermodynamic and kinetic properties of experimental interest. However, the key problems in this context consist of the identification of a suitable dynamical equation and its parametrization. We developed an approach combining GLEs, their numerically efficient representation via Markovian equations including hidden variables, and a powerful machine-learning algorithm borrowed from the field of statistical modeling and data science. Starting from non-Markovian trajectories (e.g., projected all-atom molecular dynamics trajectories in condensed-matter applications), we maximize the likelihood of an extended Markovian model employing the EM algorithm. The advantage of obtaining an explicit parametrization allows for inexpensive sampling of synthetic trajectories, which can be used for the direct computation of quantitative observables (beyond the standard memory kernel and VACF) such as stationary currents in nonequilibrium situations or the distribution of FPTs between metastable states, generally hard to access through atomistic simulations.

Several features distinguish our approach from others existing in the literature. First, the model we optimize includes an explicit parametrization of both the friction and the noise, ensuring consistency between the analysis of the MD trajectories and the generation of new projected trajectories. Second, our method is based on a maximum likelihood procedure, which is well justified from a mathematical perspective. In particular, instead of estimating a nonparametric kernel which is then parametrized (as e.g., in Volterra-based approaches), the parametric model is directly fitted on the data; this should limit the accumulation of errors. Third, we do not enforce equilibrium conditions (such as the fluctuation dissipation theorem) on the model, so that the present approach offers the possibility to investigate nonequilibrium systems. Finally, the present approach readily applies to multidimensional CVs and corresponding matrix memory kernels.

The maximum likelihood approach offers a versatile strategy to implement various extended Markovian models, which could be extended in particular to position-dependent GLEs and higher-order discretization schemes. Overall, the present work provides an efficient way to generate reduced dynamical models for multidimensional CVs, with the same memory kernels as the underlying complex system, enabling the accurate sampling of the long-time dynamics of the latter at a dramatically reduced computational cost.

Materials and Methods

Estimate of the (Potential of) Mean Force. In the first two examples, the coefficients of the quadratic potentials in the EM method follow from those of the corresponding forces, which are the ones determined numerically along with the parameters related to the memory (*SI Appendix*). Potentials of mean force in the Volterra method result from quadratic fits of the logarithm of histograms of the position. We obtain the results for the memory kernel with the Volterra approach using the memtools package [https://github.com/jandaldrop/memtools (4)] in the 1D case and the multidimensional version of ref. 39 (see our implementation at https://github.com/HadrienNU/VolterraBasis) in the 2D case.

MD Simulation Details for the LJ Dimer. The dynamics is integrated with a time step of $\Delta t_{MD} = 0.001$ (in LJ units) in the NVE ensemble with the velocity Verlet algorithm using the LAMMPS simulation package (58). We run 20 trajectories of 10^5 time steps, and CV values are extracted every 2 steps.

EM Convergence. Initial values of the parameters are taken randomly. For all examples, we stop the EM iterations if the difference of log-likelihood between two EM steps is less than 10^{-8} or if the number of EM steps exceeds 2,000.

Density and Average Velocity for the Nonequilibrium 2D Case. The density is estimated by kernel density estimation using the positions along the trajectories. The average velocities are estimated conditionally on the positions using kernel regression. The same Gaussian kernel is used in both cases, with a bandwidth of 1. For Fig. 4*B*, 20 new trajectories of 3×10^4 steps with a time step of 5×10^{-3} were sampled from the fitted GLE model and compared to the original 20 trajectories of Fig. 4*A*.

Mean FPT Estimation. The FPT is estimated for molecular dynamics starting by restraining the initial position with a parabolic potential as a function of r using PLUMED (59). Two thousand trajectories are generated from different restrained positions. Gaussian kernel estimates (with bandwidth of $1/\Delta t_{MD}$) of the mean FPT as well as the FPT density are then obtained conditioned on the realized starting position. The FPT and mean FPT from the fitted models are computed using 1,500 trajectories per initial value of the distance, again employing kernel estimates.

Data and Code Availability. A python package to perform the analysis introduced in the present work is available at GitHub, https://github.com/HadrienNU/GLE_AnalysisEM. Our implementation of the 2D Volterra method is available at GitHub, https://github.com/HadrienNU/VolterraBasis. These data and code are also available at Zenodo (DOI: 10.5281/zenodo.5536561).

ACKNOWLEDGMENTS. We thank Sara Bonella, Michele Casula, Arthur France-Lanord, Marco Saitta, Mathieu Salanne, and Rodolphe Vuilleumier for fruitful discussions within the materials for energy through stochastic sampling and high-performance computing (MAESTRO) collaboration. This project received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement 863473).

Author affiliations: [a]Institut des Sciences du Calcul et des Données, Sorbonne Université, F-75005 Paris, France; [b]CNRS, FR 3487, Fédération de Mathématiques de CentraleSupélec, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France; [c]Laboratoire Jacques-Louis Lions, Sorbonne Université, F-75005 Paris, France; [d]Laboratoire de Chimie Théorique, Sorbonne Université, F-75005 Paris, France; [e]Muséum National d'Histoire Naturelle, UMR CNRS 7590, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, Sorbonne Université, F-75005 Paris, France; and [f]Physicochimie des Électrolytes et Nanosystèmes Interfaciaux, Sorbonne Université, CNRS, F-75005 Paris, France

1. J. T. Hynes, Chemical reaction dynamics in solution. *Annu. Rev. Phys. Chem.* **36**, 573–597 (1985).
2. J. P. Bergsma, B. J. Gertner, K. R. Wilson, J. T. Hynes, Molecular dynamics of a model sn2 reaction in water. *J. Chem. Phys.* **86**, 1356–1376 (1987).
3. L. Bocquet, J. Piasecki, Microscopic derivation of non-Markovian thermalization of a Brownian particle. *J. Stat. Phys.* **87**, 1005–1035 (1997).
4. J. O. Daldrop, J. Kappler, F. N. Brünig, R. R. Netz, Butane dihedral angle dynamics in water is dominated by internal friction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 5169–5174 (2018).
5. W. Min, G. Luo, B. J. Cherayil, S. C. Kou, X. S. Xie, Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Phys. Rev. Lett.* **94**, 198302 (2005).
6. S. Kheifets, A. Simha, K. Melin, T. Li, M. G. Raizen, Observation of Brownian motion in liquids at short times: Instantaneous velocity and memory loss. *Science* **343**, 1493–1496 (2014).
7. M. Lysy *et al.*, Model comparison and assessment for single particle tracking in biological fluids. *J. Am. Stat. Assoc.* **111**, 1413–1426 (2016).
8. B. G. Mitterwallner, C. Schreiber, J. O. Daldrop, J. O. Rädler, R. R. Netz, Non-Markovian data-driven modeling of single-cell motility. *Phys. Rev. E* **101**, 032408 (2020).
9. R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).
10. A. J. Chorin, O. H. Hald, R. Kupferman, Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 2968–2973 (2000).
11. A. J. Chorin, O. H. Hald, R. Kupferman, Optimal prediction with memory. *Physica D* **166**, 239–257 (2002).
12. L. Ma, X. Li, C. Liu, The derivation and approximation of coarse-grained dynamics from Langevin dynamics. *J. Chem. Phys.* **145**, 204117 (2016).
13. S. H. Chung, M. Roper, Generalized Langevin equation: An introductory review for biophysicists. *Biophys. Rev. Lett.* **14**, 171–196 (2019).
14. E. Darve, J. Solomon, A. Kia, Computing generalized Langevin equations and generalized Fokker-Planck equations. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10884–10889 (2009).
15. S. Izvekov, Microscopic derivation of particle-based coarse-grained dynamics. *J. Chem. Phys.* **138**, 134106 (2013).
16. R. Zwanzig, Nonlinear generalized Langevin equations. *J. Stat. Phys.* **9**, 215–220 (1973).
17. H. Mori, A continued-fraction representation of the time-correlation functions. *Prog. Theor. Phys.* **34**, 399–416 (1965).
18. H. Mori, Transport, collective motion, and Brownian motion. *Prog. Theor. Phys.* **33**, 423–455 (1965).
19. F. Glatzel, T. Schilling, The interplay between memory and potentials of mean force: A discussion on the structure of equations of motion for coarse grained observables. *Europhys. Lett.* **136**, 36001 (2021).
20. H. Vroylandt, P. Monmarché, *Position-dependent memory kernel in generalized Langevin equations: Theory and numerical estimation.* arXiv [Preprint] (2022). https://arxiv.org/abs/2201.02457 (Accessed 10 January 2022).
21. T. J. Doerries, S. A. M. Loos, S. H. L. Klapp, Correlation functions of non-Markovian systems out of equilibrium: Analytical expressions beyond single-exponential memory. *J. Stat. Mech.* **2021**, 033202 (2021).
22. M. Berkowitz, J. D. Morgan, D. J. Kouri, J. A. McCammon, Memory kernels from molecular dynamics. *J. Chem. Phys.* **75**, 2462–2463 (1981).
23. B. J. Berne, G. D. Harp, "On the calculation of time correlation functions" in *Advances in Chemical Physics*, I. Prigogine, S. A. Rice, Eds. (John Wiley & Sons, Ltd, 1970), pp. 63–227.
24. B. J. Berne, M. E. Tuckerman, J. E. Straub, A. L. R. Bug, Dynamic friction on rigid and flexible bonds. *J. Chem. Phys.* **93**, 5084–5095 (1990).
25. H. Lei, N. A. Baker, X. Li, Data-driven parameterization of the generalized Langevin equation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14183–14188 (2016).
26. A. Carof, R. Vuilleumier, B. Rotenberg, Two algorithms to compute projected correlation functions in molecular dynamics simulations. *J. Chem. Phys.* **140**, 124103 (2014).
27. D. Lesnicki, R. Vuilleumier, A. Carof, B. Rotenberg, Molecular hydrodynamics from memory kernels. *Phys. Rev. Lett.* **116**, 147804 (2016).
28. G. Jung, M. Hanke, F. Schmid, Iterative reconstruction of memory kernels. *J. Chem. Theory Comput.* **13**, 2481–2488 (2017).
29. G. Jung, M. Hanke, F. Schmid, Generalized Langevin dynamics: Construction and numerical integration of non-Markovian particle-based models. *Soft Matter* **14**, 9368–9382 (2018).
30. V. Klippenstein, M. Tripathy, G. Jung, F. Schmid, N. F. A. van der Vegt, Introducing memory in coarse-grained molecular simulations. *J. Phys. Chem. B* **125**, 4931–4954 (2021).
31. H. Lei, B. Caswell, G. E. Karniadakis, Direct construction of mesoscopic models from microscopic simulations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **81**, 026704 (2010).
32. A. Davtyan, J. F. Dama, G. A. Voth, H. C. Andersen, Dynamic force matching: A method for constructing dynamical coarse-grained models with realistic time dependence. *J. Chem. Phys.* **142**, 154104 (2015).
33. Z. Li, X. Bian, X. Li, G. E. Karniadakis, Incorporation of memory effects in coarse-grained modeling via the Mori-Zwanzig formalism. *J. Chem. Phys.* **143**, 243128 (2015).
34. Z. Li, H. S. Lee, E. Darve, G. E. Karniadakis, Computing the non-Markovian coarse-grained interactions derived from the Mori-Zwanzig formalism in molecular systems: Application to polymer melts. *J. Chem. Phys.* **146**, 014104 (2017).
35. Y. Yoshimoto, Z. Li, I. Kinefuchi, G. E. Karniadakis, Construction of non-Markovian coarse-grained models employing the Mori-Zwanzig formalism and iterative Boltzmann inversion. *J. Chem. Phys.*

147, 244110 (2017).

36. A. V. Straube, B. G. Kowalik, R. R. Netz, F. Höfling, Rapid onset of molecular friction in liquids bridging between the atomistic and hydrodynamic pictures. *Commun. Phys.* **3**, 1–11 (2020).
37. C. Ayaz *et al.*, Non-Markovian modeling of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023856118 (2021).
38. J. Fricks, L. Yao, T. C. Elston, A. M. Gregory Forest, Time-domain methods for diffusive transport in soft matter. *SIAM J. Appl. Math.* **69**, 1277–1308 (2009).
39. H. S. Lee, S. H. Ahn, E. F. Darve, The multi-dimensional generalized Langevin equation for conformational motion of proteins. *J. Chem. Phys.* **150**, 174113 (2019).
40. M. Ceriotti, G. Bussi, M. Parrinello, Colored-noise thermostats à la carte. *J. Chem. Theory Comput.* **6**, 1170–1180 (2010).
41. A. D. Baczewski, S. D. Bond, Numerical integration of the extended variable generalized Langevin equation with a positive Prony representable memory kernel. *J. Chem. Phys.* **139**, 044107 (2013).
42. G. Ciccotti, J. P. Ryckaert, Computer simulation of the generalized Brownian motion. *Mol. Phys.* **40**, 141–159 (1980).
43. L. Stella, C. D. Lorenz, L. Kantorovich, Generalized Langevin equation: An efficient approach to nonequilibrium molecular dynamics of open systems. *Phys. Rev. B Condens. Matter Mater. Phys.* **89**, 134303 (2014).
44. S. Wang, Z. Ma, W. Pan, Data-driven coarse-grained modeling of polymers in solution with structural and dynamic properties conserved. *Soft Matter* **16**, 8330–8344 (2020).
45. N. Bockius, J. Shea, G. Jung, F. Schmid, M. Hanke, Model reduction techniques for the computation of extended Markov parameterizations for generalized Langevin equations. *J. Phys. Condens. Matter* **33**, 214003 (2021).
46. L. Ma, X. Li, C. Liu, Coarse-graining Langevin dynamics using reduced-order techniques. *J. Comput. Phys.* **380**, 170–190 (2019).
47. M. Berkowitz, J. D. Morgan, J. A. McCammon, Generalized Langevin dynamics simulations with arbitrary time-dependent memory kernels. *J. Chem. Phys.* **78**, 3256–3261 (1983).
48. J. L. Barrat, D. Rodney, Portable implementation of a quantum thermal bath for molecular dynamics simulations. *J. Stat. Phys.* **144**, 679–689 (2011).
49. S. Wang, Z. Li, W. Pan, Implicit-solvent coarse-grained modeling for polymer solutions via Mori-Zwanzig formalism. *Soft Matter* **15**, 7567–7582 (2019).
50. F. Gottwald, S. Karsten, S. D. Ivanov, O. Kühn, Parametrizing linear generalized Langevin dynamics from explicit molecular dynamics simulations. *J. Chem. Phys.* **142**, 244110 (2015).
51. A. Russo, M. A. Durán-Olivencia, I. G. Kevrekidis, S. Kalliadasis, Machine learning memory kernels as closure for non-Markovian stochastic processes. arXiv [Preprint] (2019). https://arxiv.org/abs/1903.09562.
52. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977).
53. R. Little, D. Rubin, *Statistical Analysis with Missing Data* (Wiley, ed. 3, 2019).
54. L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
55. A. Dembo, O. Zeitouni, Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stochastic Process. Appl.* **23**, 91–113 (1986).
56. R. Fildes, Forecasting, structural time series models and the Kalman filter: Bayesian forecasting and dynamic models. *J. Oper. Res. Soc.* **42**, 1031–1033 (1991).
57. V. Mancois, B. Marcos, P. Viot, D. Wilkowski, Two-temperature Brownian dynamics of a particle in a confining potential. *Phys. Rev. E* **97**, 052121 (2018).
58. S. Plimpton, Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
59. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, Plumed 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).