Complex temporal patterns in molecular dynamics: A direct measure of the phase-space exploration by the trajectory at macroscopic time scales

Dmitry Nerukh[*]
Unilever Centre for Molecular Informatics, Department of Chemistry, Cambridge University, Cambridge CB2 1EW, United Kingdom

Vladimir Ryabov
*School of Systems Information Science, Department of Complex Systems, 116-2 Kamedanakano-cho, Hakodate-shi,
041-8655 Hakodate, Hokkaido, Japan*

Robert C. Glen
Unilever Centre for Molecular Informatics, Department of Chemistry, Cambridge University, Cambridge CB2 1EW, United Kingdom

(Received 18 July 2007; revised manuscript received 13 November 2007; published 28 March 2008)

Computer simulated trajectories of bulk water molecules form complex spatiotemporal structures at the picosecond time scale. This intrinsic complexity, which underlies the formation of molecular structures at longer time scales, has been quantified using a measure of statistical complexity. The method estimates the information contained in the molecular trajectory by detecting and quantifying temporal patterns present in the simulated data (velocity time series). Two types of temporal patterns are found. The first, defined by the short-time correlations corresponding to the velocity autocorrelation decay times (≤ 0.1 ps), remains asymptotically stable for time intervals longer than several tens of nanoseconds. The second is caused by previously unknown longer-time correlations (found at longer than the nanoseconds time scales) leading to a value of statistical complexity that slowly increases with time. A direct measure based on the notion of statistical complexity that describes how the trajectory explores the phase space and independent from the particular molecular signal used as the observed time series is introduced.

DOI: 10.1103/PhysRevE.77.036225 PACS number(s): 05.45.Jn, 83.10.Rs, 61.20.Ja, 02.70.Ns

I. INTRODUCTION

Molecular dynamics (MD) trajectories representing a diffusion process in liquids form complex patterns in the phase space. Because of the system's high dimensionality (defined by the number of interacting molecules in the analyzed volume), single molecule trajectories in the long-time limit are usually assumed to be indistinguishable from correlated noise by standard statistical methods [1]. However, the equations of motion of each particle are deterministic, therefore, the local nonlinear dynamics at the picosecond time scale may lead to nontrivial behavior and the emergence of molecular structures over much longer (nanoseconds) periods. Discovering and quantifying this nontrivial long-term behavior and intrinsic complexity of molecular trajectories (over various time scales) that can shed light on the origin of emergent molecular behavior is the subject of the present work.

The computational mechanics approach suggested by Crutchfield *et al.* [2–4] seems to be most suitable for the task. The approach is useful for detecting dynamical structures in observed time series, since it is based on information-theoretic concepts, without detailed assumptions on the geometric properties of the structures originating from the dynamics in the phase space. The authors introduced the idea of an ϵ machine working in the space of so-called causal states that catch the principal dynamics of the system. One of the central concepts of the formalism is a statistical complexity measure, a characteristic of the ϵ machine, that describes how complex the underlying process is by quantifying the amount of information stored and processed by the system. The advantage of computational mechanics is that, besides intuitive (and at the same time mathematically rigorous) statistical properties that quantify the complexity, it captures complete statistical information contained in the signal (see the Appendix).

It should be noted that there exist several alternative definitions of complexity used in different contexts. The analysis of the metric properties of sets of trajectories in phase space [5] or the study of time separation between initially close trajectories [6] should be distinguished from the statistical complexity used in the present work. The latter is aimed at purely probabilistic modeling of the time series, irrespective of the geometry of trajectories in the phase space or local instability of the dynamics. Approaches for estimating the physical complexity of classical trajectories using their information-theoretic contents include, for example, work by Brudno [7] that relates Kolmogorov's algorithmic complexity and the metric entropy of an ergodic dynamical system. Recently, Segre [8] showed that chaotic dynamical systems are simple from Bennett's "logical depth" point of view. Benci *et al.* [9] have investigated the amount of information necessary to describe the chaotic orbits and find more than logarithmic increase with time for weakly chaotic cases.

An alternative approach to the analysis of molecular trajectories could be the application of the concepts and methods of time series analysis that appeared recently in the field of nonlinear dynamics for characterizing intrinsically deterministic processes [10,11]. However, this group of time series analysis techniques seems to be feasible only for systems

[*] dn232@cam.ac.uk

with few degrees of freedom. The limitation of all nonlinear dynamics methods is the assumption of the existence of a low-dimensional manifold (attractor) in the phase space where the essential dynamics occur. It is, however, unclear whether this concept can be effectively utilized in a numerical analysis of signals obtained from highly dimensional systems, such as fully developed turbulence, the human brain, or molecular ensembles. It is, therefore, necessary to search for new techniques that would discover inherent signatures of dynamics rather than assume the existence of structures in the phase space.

From the viewpoint of classical physics, the motion of a single particle (hydrogen atom in our case) in bulk water without impurities can be well approximated by a stochastic model of Brownian motion. Characterizing a molecular trajectory in terms of diffusion theory reveals linear time growth of the mean square displacement [12], that approximately corresponds to theoretical predictions from the classical theory of Brownian motion [13]. Although deviations from Gaussianity of corresponding distribution functions can be detected by the analysis of higher moments [1,14] the experimental (or numerically simulated) time series becomes indistinguishable from a stochastic Gaussian process at time periods longer than ≈100 ps (at room temperature) [1,15]. This implies a sufficient description of molecular trajectories in terms of correlation functions and/or power spectra and almost trivial behavior on the time scales larger than "correlation time" defined by, e.g., the first zero of the autocorrelation function.

Although such a stochastic approach provides a satisfactory description of liquids at the macroscopic scale, there is no clear understanding of how the observed macroscopic randomness is produced by purely deterministic equations of motion of every atom, i.e., how the microscopically ordered motion is transformed to the macroscopic disorder. There were several attempts recently to demonstrate that at a microscopic level the motion of molecules is chaotic [16], and the randomness due to local instability of trajectories in the phase space is transformed to a random walk motion of Brownian particles observed in experiments. It has been argued, however, in later works [17] that similar random behavior at macro level can be caused by nonchaotic systems that do not possess the property of local instability at microscales. Therefore, the question on the microscopic origin of macroscopic randomness remains open.

On the other hand, it became clear recently [18] that, even if the dynamics of a single microparticle is chaotic its temporal behavior may be nontrivial due to the presence of resonances in the phase space and particles can demonstrate anomalous diffusion. When a chaotically moving particle comes close to any of the resonance zones, it can spend an abnormally long time there due to the so-called "stickiness" of the border of the resonances. As a result of such intermittent behavior, the phase space becomes strongly nonuniform and processes with significantly different characteristic temporal scales appear in the time series representing the trajectories. Consequently, particle trajectories may possess much longer memories than can be expected from a simple analysis of the autocorrelation function.

All the above considerations provide motivation for performing the analysis of molecular trajectories over very long-time periods compared to time scales defined by the autocorrelation functions (≈1 ps), aimed at detecting nontrivial (i.e., different from a pure Gaussian noise) temporal structures. In order to quantify the deterministic origin of the dynamics of particles in a high-dimensional phase space corresponding to MD trajectories of bulk water we apply a combination of the computational mechanics approach and a surrogate data method from nonlinear dynamics [19]. Our goal is to detect and quantify complex temporal structures present in the water trajectory defined by the deterministic dynamics. We show that, because of existing long-time correlations, the structure of the groups of histories, the ϵ machine, in the MD signal is qualitatively different from that obtained for an artificial random time series (surrogate) with identical correlation and/or spectral properties.

Our work can thus serve to provide insights into two important open problems: (i) is it possible to apply the computational mechanics approach to realistic molecular systems; (ii) does computational mechanics give a possibility to quantify nontrivial temporal structures in liquid water.

The next section describes the specifics about the molecular model for water used in our calculations, as well as other details of numerical procedures used to obtain the molecular trajectory. The main idea and the methodology of the further analysis aimed at quantifying the informatic-theoretical content of the calculated time series are given in Sec. III. The obtained results are presented in Secs. IV and V, while their interpretation is provided in Sec. VI.

II. MOLECULAR MODEL AND MOLECULAR DYNAMICS SIMULATION DETAILS

Water, being a complex liquid, has arguably one of the most developed simulation models [20–27]. Numerous MD models of water differ in sophistication depending on the specific task of the simulation. For us, the combination of the simulation speed and the potential ability of modeling protein folding was decisive in choosing the molecular model. We, therefore, focused on simple point charge (SPC) [28] water while checking other flavors for the consistency of the results. We expect that the main conclusions of this work will hold for other liquids, the extensive study of which is, however, a subject of a separate publication.

Bulk water consisting of 392 or 878 SPC, simple point charge extended (SPC-E) [28], or transferable intermolecular potential 3 point (TIP3P) molecules was simulated using the GROMACS molecular dynamics [29] package. The temperature of the system was kept constant at 275, 300, or 380 K using Berendsen [30] or Nose-Hoover [31] thermostats, with a coupling time of 0.1 ps, whose combination with various coupling constants was investigated. Pressure coupling was also applied to a pressure bath with a reference pressure of 1 bar and a coupling time of 0.1 ps. A 1 nm cutoff distance for both van der Waals and Coulomb potentials was used. An equilibration until the potential and kinetic energies reach constant levels of fluctuations was performed before collecting data for analysis. The velocity of the oxygen and hydrogen atoms of one of the water molecules was used as a three-dimensional signal for the complexity analysis. Instant temperature, $T_{inst}=\frac{1}{N_{df}k}\Sigma_i m_i \mathbf{v}_i^2$, where the summation is over all atoms, N_{df} is the number of degrees of freedom and k is the Boltzman constant, was also used for calculating the complexity values.

Classical molecular systems are Hamiltonian. However, numerical errors associated with the model potential and thermostatting algorithm make the simulated MD system non-Hamiltonian. Therefore, we use a selected velocity subspace for building a symbolic sequence and reconstructing the phase space without recourse to the Hamiltonian properties of the underlying molecular dynamics.

III. IDEA AND METHODOLOGY

A. Signal analysis

1. Diffusion process

Traditionally, motion of a single particle in a liquid that appears random can be characterized by the time dependence of its mean squared displacement $\langle x^2(t)\rangle$, which demonstrates a power law behavior

$$\langle x^2(t)\rangle \propto t^{\alpha} \tag{1}$$

at sufficiently long times. Here $x(t)$ corresponds to the deviation of the coordinate x from the arbitrary initial condition $x(0)$ and the averaging is performed either over an ensemble of trajectories, or, under the assumption of ergodicity of the diffusion process, over an ensemble of initial conditions along a single trajectory. Normal diffusion (Brownian motion) then corresponds to the value of the diffusion coefficient $\alpha=1$, whereas values different from unity indicate the presence of anomalous diffusion. The case of $0<\alpha<1$ is called subdiffusion, whereas the range of values $1<\alpha<2$ is attributed to a superdiffusion process. The limiting value of $\alpha=2$ characterizes the ballistic regime typical to free motion of the particles over short distances. Note that the distribution of $x(t)$ is Gaussian only in the asymptotic limit of $t\to\infty$ and at small t significant deviations from Gaussianity can be detected by the analysis of higher moments [1]. The indicator $\sigma(t)$ is often used for quantifying the non-Gaussian behavior [12,14]

$$\sigma(t)=\frac{\langle x^4(t)\rangle}{3\langle x^2(t)\rangle}-1. \tag{2}$$

Due to the fact that for a Gaussian distribution $\langle x^4(t)\rangle =3\langle x^2(t)\rangle$, the deviation of the indicator $\sigma(t)$ from zero is a manifestation of intrinsic non-Gaussianity of the process $x(t)$.

2. Phase-space partitioning

In this study the velocity of one of the atoms is mainly used as a three-dimensional signal for the complexity analysis. The domain of the signal values appearing in the simulation has the shape of a ball centered at the origin with the radius of ≈4 nm/ps. The approximately centrally symmetric distribution of the data points makes the velocities a convenient signal for symbolization and data accumulation in contrast to, for example, the coordinates that diffuse very slowly in the phase space.

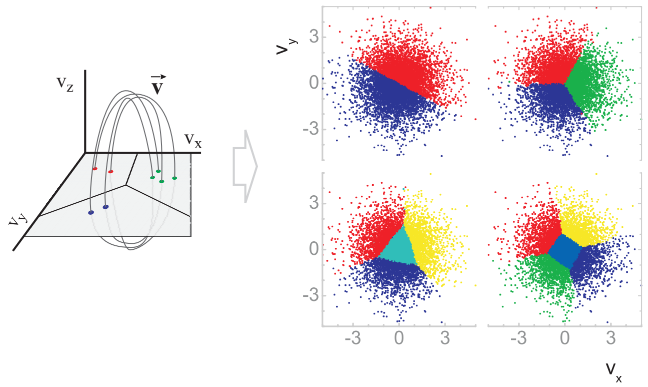

FIG. 1. (Color) Approximations for generating partitions obtained using the method by Buhl and Kennel [33] for the cross section of the hydrogen velocity space for the partitions corresponding to two-, three-, four-, and five-symbols alphabet.

At the initial stage of data analysis, the original three-dimensional vector time series is transformed to a scalar symbolic sequence. The symbols are produced by applying a special phase-space coarse graining procedure on a suitably chosen cross section plane in the velocity space. It turns out that if the $v_z=0$ plane is used as a section surface, the average time interval between the resulting data points (≈0.032 ps) roughly corresponds to the first minimum on the autocorrelation function of the original three-dimensional signal. A natural choice for the phase-space partitioning used for symbolization would be the generating partition (GP) [32] that has the property of a one-to-one correspondence between the continuous trajectory and the generated symbolic sequence. That is, all information is retained after the symbolization. However, because of the very high dimensionality of the system it is infeasible to find a GP in our case. There are methods for calculating approximations to the GP. Using one of them [33] we obtained indications of what an optimal partition would look like. For our velocities data, the resulting approximated GPs were centrally symmetric for all tested number of partitions; the cases for two, three, four, and five partitions are shown in Fig. 1. Thus, for all the data analyzed we used centrally symmetrical partitions for converting the continuous data into symbolic sequences.

Summarizing, the following procedure was used for symbolization (Fig. 1): (i) the velocities of the hydrogen (or oxygen) atoms were used as a continuous three-dimensional signal; (ii) at the locations where the velocity pierces the xy plane the points of the map were generated; (iii) using the centrally symmetric partitions the map was converted into the symbolic sequence. The size K of the alphabet to be used in the conversion process is empirically found to be sufficient at the value $K=3$ for good convergence and reproducibility of numerical results.

3. Reconstruction of the ϵ machine

Computational mechanics detects hidden order within random looking symbolic sequences by building a linked structure of probabilistically related states in the phase space, the ϵ machine. The phase space is obtained by a procedure similar to the attractor reconstruction technique in dynamical systems theory based on the Takens embedding theorem [34], but with a symbolic sequence taken for an original time series. Every point in an l-dimensional reconstructed phase space then corresponds to a set of l successive symbols from the (scalar) symbolic sequence.

Such sets of symbols (heretofore, histories of length l) are grouped according to their ability to predict one step forward. If the time step between successive symbols approximately equals the correlation decay time, then, e.g., for a completely random process, all the histories are grouped into a single causal state. Due to the absence of predictive power on the time scales longer than the decay time of correlations, this simply reflects the fact that all histories predict the same random futures.

Contrary to completely random signals, the groups of histories with similar futures are numerous for the molecular signal, and corresponding causal states can be characterized by their occurrence rates $P(\epsilon_i)$ (see the Appendix). The statistical complexity C_μ is then a statistical measure quantifying the difference of the distribution of $P(\epsilon_i)$ from the uniform one expected for a purely stochastic process. Therefore, as will be shown in subsequent sections, the difference of the molecular signal from a random signal consists of both the large number of unique causal states found in the phase space, and the corresponding high value of the statistical complexity.

4. Surrogate time series analysis

It should be noted that the algorithms used for numerical identification of the set of causal states and estimating the value of statistical complexity include several computational steps as well as many internal parameters that control the computation precision and statistical confidence of the results. As a consequence, there are a number of potential sources for statistical errors and biases that require special care. A straightforward approach would be direct analysis of corresponding distribution functions for the estimated values, with the subsequent calculation of moments (including dispersion) and the associated error bars. However, this type of analysis encounters serious technical difficulties due to the complexity of the calculation procedure that is, in fact, a combination of several algorithms. We therefore accepted a different, much simpler way of getting error estimates, widely used in the literature on the analysis of time series. In works devoted to statistical data analysis, it is also known as the "bootstrap" method [35], whereas in papers discussing nonlinear dynamics based techniques [19] it is called the "surrogate data" method. Throughout this paper, we use the latter term as the name for artificial time series.

The idea of the surrogate data methodology consists of the comparison of the experimental data to a set of artificially created time series (surrogates), which lack some intrinsic property of interest but have very similar (or even identical) probability density function and/or power spectrum to those of the original data. This method thus provides a kind of "control experiment" that allows testing of the experimental time series against a hypothesis that it is produced by a linear stochastic process like, e.g., an autoregressive moving average process (ARMA) [36].

The practical algorithm implementing the idea of surrogate data consists of several steps. First, a statistical indicator called discriminating statistics has to be defined. From a general viewpoint of time series analysis, this could be any real number calculated from the data, like, for example, a high-order moment, autocorrelation, fractal dimension, or statistical complexity, depending on the particular property of interest that the analysis is focused on. At the second stage, the surrogate data series are produced by using a random number generator combined with some original algorithm of data transformation. In other words, the algorithm converts the random sequence of numbers to a time series with required properties. The surrogates preserve some well-controlled statistical characteristics of the analyzed data, but lack the property of interest. For example, in the context of the analysis of deterministic dynamics, the surrogates are usually chosen to have the power spectrum identical to the original data series, but, by its definition, do not possess any property imposed by deterministic dynamics such as, e.g., finite value of correlation dimension [37] or others [38]. At the final stage, the discriminating statistics have to be calculated for original data and compared to a set of corresponding values calculated from a set of surrogate time series. Significant discrepancy in the calculated values can be considered as an indication of an essential difference between the surrogates and the original time series in the analyzed property.

In this work the discriminating statistics are the number of causal states n_{st} and the value of statistical complexity C_μ. For surrogates, we use two types of time series:

(1) Phase-shuffled surrogate. This is a standard surrogate time series obtained via the phase-shuffling algorithm [19]. The data obtained with this method possess identical power spectrum (and, hence, autocorrelation function) to the original time series, but lack the property of dynamic correlation between the data points. It is generated by calculating the Fourier spectrum of the original data and assigning random values to all the phases of Fourier components. After calculating the inverse Fourier transform, the artificial data series (surrogate) has the unchanged power spectrum but is completely random, i.e., it belongs to the class of Gaussian linear stochastic processes.

(2) Temporal pattern-shuffled surrogate. This surrogate data set is created by introducing random perturbations to the original time series that do not change dynamic temporal patterns in the data up to a certain time period T_r. It is obtained by random rotation (by a random angle) of the data segments containing n_r consecutive points in the phase-space cross-section plane (see Fig. 1). The surrogates obtained by this method allow an analysis of the relationship between the characteristic periods in the analyzed data and the structure of causal states in the ϵ machine.

IV. RESULTS

A. Overview of the dynamics and statistical analysis

1. Spectral characteristics

The velocity autocorrelation functions (vacf) of water atoms, obtained as a time average over 2 ns (Fig. 2), show that

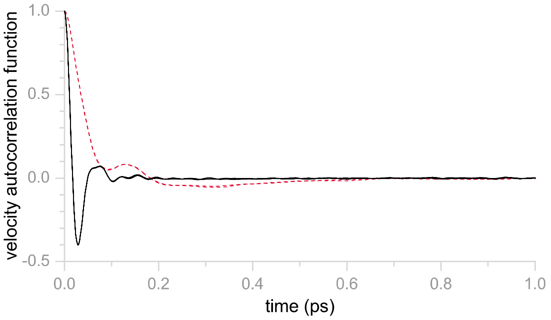

FIG. 2. (Color online) Velocity autocorrelation function for oxygen (dashed line) and hydrogen atoms of two water molecules calculated as time average over 2 ns. The curves for different atoms of the same type are practically indistinguishable.

linear time correlations last for ≈0.2 ps for the hydrogens and ≈0.5 ps for the oxygens. From a viewpoint of linear correlation theory, there are no long-range temporal correlations in the analyzed data.

On the other hand, the power spectrum of the velocity fluctuations (Fig. 3) shows a nontrivial feature at somewhat longer time scales: a broad low frequency peak is present at ≈1 ps.

This fact can also be visualized by plotting the trajectory corresponding to the time evolution of the orientation vector of a water molecule (Fig. 4). The trajectory shown in Fig. 4 clearly displays complicated intermittent dynamical structures over different time scales. A typical temporal pattern observed for such data can be roughly described as follows. It tends to fluctuate around some fixed value for a time period of ≈1 ps demonstrating quick jumps to other areas of similar fluctuations within a much shorter time interval (compared to 1 ps). Note, that such behavior can be a manifestation of a fractional kinetics process typical to Hamiltonian systems typically containing resonances [39].

2. Diffusion

The calculation of the squared displacement for a hydrogen atom shows a power law behavior, implying a diffusion process with the diffusion constant slightly less than unity, Fig. 5. The largest deviation from the normal diffusion with

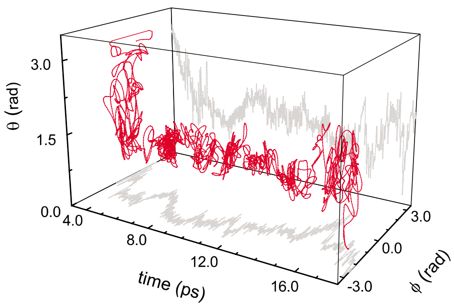

FIG. 4. (Color online) Dipole moment orientation (ϕ and θ angles in the spherical coordinates) as a function of time for the selected water molecule.

$\alpha=1$ is observed at the time scales approximately corresponding to the maximum in the power spectrum, i.e., at about 1 ps. Therefore, the Brownian motion of the hydrogen atom can be roughly classified as normal diffusion at time scales longer than ≈100 ps. At long time scales (≳100 ps) the motion of the atom becomes similar to a classical Brownian particle, therefore, it can be expected that at large time intervals the trajectory is well approximated by a stochastic Gaussian process. This fact is further illustrated in Fig. 6 where we plot the time dependence of the non-Gaussianity parameter $\sigma(t)$ [Eq. (2)]. The MD trajectory shows significant deviations from the Gaussian behavior only at small time intervals, whereas at $t\gtrsim 10$ ps it becomes indistinguishable from a surrogate data (Gaussian process). However, as it will be shown below, the molecular trajectories of water particles present nontrivial dynamics and a significant difference from the Gaussian surrogates even at much longer time scales (tens of nanoseconds) that can be clearly demonstrated by the analysis of statistical complexity.

B. Time dependence of C_μ: A universal exponent of logarithmic growth

The calculated values of C_μ against time t are shown in Fig. 7. The red heavy curve at the bottom corresponds to the phase-shuffled surrogate time series. The other curves, calculated at different values of the parameter l (phase-space

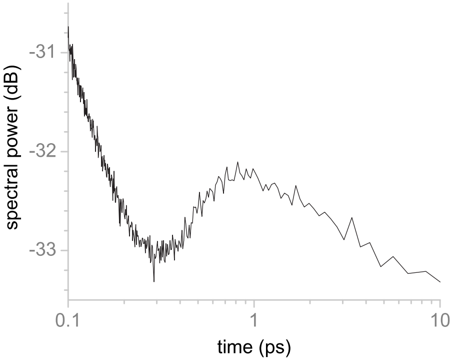

FIG. 3. Spectrum of the velocity x component of a hydrogen atom in bulk SPC water at 300 K.

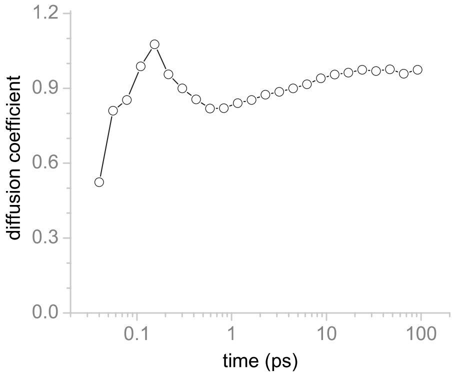

FIG. 5. Diffusion coefficient of the x component of a hydrogen atom in bulk SPC water at 300 K.

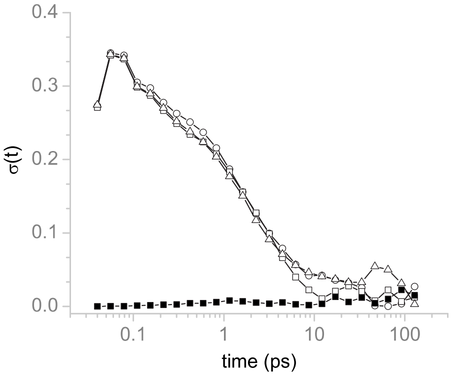

FIG. 6. Non-Gaussianity parameter for three Cartesian components of the displacement for a hydrogen atom (x—circles, y—squares, z—triangles) and phase-shuffled surrogate for the x component (filled squares).

dimension, history length) demonstrate the convergence of C_μ with increasing l. Starting from the history length of about seven symbols (l=7), the calculated value of the statistical complexity (at any fixed moment of time) saturates and does not change with a further increase in l.

Note, however, that the dependence of C_μ on t does not converge, but first goes through a maximum and then settles on the $\log_2 t$-like curve (this behavior is clearly seen, especially for high l values, Fig. 7). The maximum at the small t is due to the lack of statistics, when the algorithm finds too many causal states considering almost every history s^- as a unique causal state. The number of causal states n_{st} at these values of t is abnormally high, and each causal state consists of only a few histories s^-. This part of the curve is, therefore, of little interest for the present analysis and in the following we focus the analysis only on the logarithmic part of the curves.

While C_μ practically converges at any sufficiently large time moment with l (for $l>7$), Fig. 8 (and has values significantly higher than those for a corresponding phase-shuffled surrogate time series), its logarithmic dependence on time requires special consideration. It should be emphasized that the time intervals discussed here are very long compared to the correlation time (Fig. 2) or any other time period where nontrivial (i.e., non-Brownian) statistics can be expected to exist (Fig. 6).

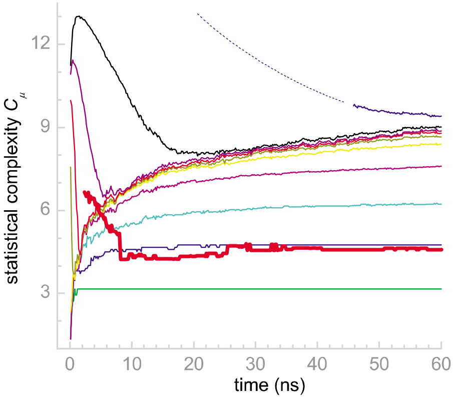

FIG. 7. (Color online) Statistical complexity against time for the hydrogen velocity signal and the surrogate. The curves, from bottom to top, correspond to the values of the history length l from 2 to 11. The l=11 curve does not settle on the logarithmic part within the shown area but seems to follow the same trend. The thick line is the C_μ values for the phase-shuffled surrogate signal (l=9). For all curves the alphabet size K is equal to 3.

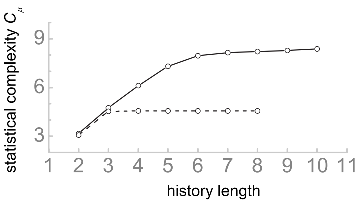

FIG. 8. Statistical complexity vs the length of histories (dimension of the phase space) for the original data (solid line) and the surrogate (dashed line), t=30 ns.

Since the growth of C_μ has a clear logarithmic character, we propose to introduce a coefficient (h_Q) that can measure the growth rate as follows:

$$C_\mu = a + h_Q \log_2 t. \quad (3)$$

We would like to emphasize that the coefficient h_Q can be used as a robust and universal characteristic of the statistical complexity of molecular trajectories since it seems not to depend on the particular numerical model, details of computational procedure, size of the molecular ensemble, and type of the test atom (hydrogen or oxygen).

To ensure the reproducibility of the phenomenon a number of tests have been performed. The tests provide evidence that the $\log_2$ dependence is not an artifact of the numerical methods used but rather an inherent property of the water molecular system. Different MD models, parameters of integration, signal processing, and symbolization procedure produce statistically the same results.

The effects of various methods of phase-space partitioning in the algorithm of discretizing the continuous time series and producing the symbolic time series have been checked by applying nonsymmetric (shifted along the x and/or y axes) partitioning and varying the position of the cross-section plane along the z axis. Except for the trivial cases characterized by poor statistics of the data points, the logarithmic growth of C_μ was present with the same (within numerical errors) values of h_Q.

The influence of a particular MD numerical simulation model on the detected phenomenon were insignificant. (i) Both Nose-Hoover and Berendsen thermostats produced almost identical results in C_μ with the same $\log_2$-like behavior. Varying the coupling constant of the Berendsen thermostat by two orders of magnitude did not change the results. (ii) SPC-E and TIP3P water models produced slightly different absolute values of C_μ than SPC while keeping the same overall logarithmic behavior of the curves unchanged. (iii) Systems containing 392 and 878 water molecules resulted in the same values of the complexity parameters.

Finally, different values of the second adjustable parameter of the ϵ-machine reconstruction algorithm, the signifi-

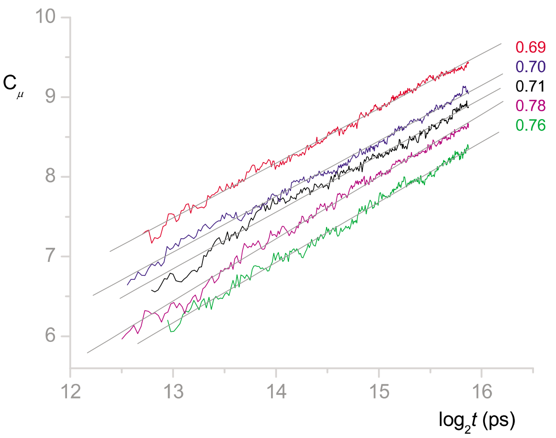

FIG. 9. (Color online) h_Q values (indicated on the right) for SPCE (h_Q=0.69), TIP3P (h_Q=0.78), and SPC (h_Q=0.71) models of water. The values for the SPC model at 275, 300, and 380 K are labeled 0.70, 0.71, and 0.76, respectively.

cance level for the χ-squared significance test, 0.001, 0.01, and 0.1, reproduced the same behavior of C_μ vs t curve.

The results of various numerical tests performed on velocity time series data of hydrogen atom are presented in Fig. 9. Within statistical fluctuations, the value of h_Q remained in the interval 0.74 ± 0.07 under any combination of physical parameters and details of the calculation procedure.

C. Comparison to surrogate data: Characteristic time scale of dynamical patterns

1. Phase-shuffled surrogate

A qualitatively different result was obtained for the phase-shuffled surrogate data. The complexity quickly settles at a constant value that roughly corresponds to the $l=3$ curve for the original signal (Fig. 7). The number of causal states as well as their occurrence rates also do not change for time intervals longer than ≈ 10 ns. The significant decrease in the statistical complexity of the surrogate signal can be interpreted as the absence of most of the dynamical patterns that constitute the causal states and contribute to a high value of C_μ in the molecular trajectories.

2. Temporal pattern-shuffled surrogate

An important test supporting the validity of our calculations is the numerical experiments with the surrogate of the second type. The idea behind introducing this kind of artificial time series consists of an attempt to destroy temporal correlations (dynamic patterns) present in the data, while preserving the overall distribution of points in the cross-section plane. The key parameter used in the generation procedure was the length of the temporal patterns to be preserved, thus the temporal scales responsible for the high value of the statistical complexity (compared, e.g., to the surrogate of the first type) could be distinguished. The time series have been obtained as follows: before performing the symbolization in the cross-section plane, the groups of n_r consecutive points $\mathbf{v}_i$ were rotated by a random angle α around the origin. The results for $n_r=3, 4, 10, 20, 50$ are presented in Fig. 10. For $n_r \leq 4$ the $\log_2$ behavior in the temporal profile of the statistical complexity was completely destroyed and C_μ and the number of causal states saturated at a constant value. For higher n_r the logarithmic behavior was retained, although the absolute values of C_μ were somewhat lower than for the case of original data. Starting from the value of $n_r \geq 40$, i.e., for the time scales longer than ≥ 1 ps, the value of statistical complexity remains unchanged. This result thus indicates that the statistical complexity at long time intervals (≥ 1 ns) is completely defined by three orders of magnitude shorter dynamical patterns (≤ 1 ps).

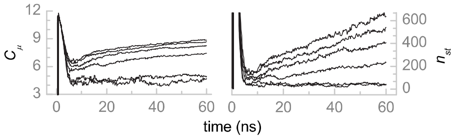

FIG. 10. Statistical complexity C_μ and number of states n_{st} for the temporal pattern-shuffled surrogate data with $l=9$. From bottom to top, $n_r=3, 4, 10, 20, 50$, and the original data.

On the other hand, if the interval of the perturbations is less than the linear correlation time of 0.1 ps ($n_r \leq 5$) this destroys all longer-time correlations, making the signal similar to the phase-shuffled surrogate characterized by very low values of C_μ.

For disturbance intervals longer than the correlation time ($n_r=10, 20, 50$) the number n_{st} of the states is much less than that in the original data leading to reduced values of statistical complexity.

V. CAUSAL STATES CLUSTERING: "CORE" STATES

To get a further insight into the link between the statistical complexity and the characteristic periods in the time series responsible for its high value, we analyzed the sets of causal states that constitute the ϵ machine and, hence, define the statistical complexity through the distribution function of their occurrence rates $P(\epsilon_i)$. In order to distinguish between various time scales, we studied the time intervals between successive appearances of a causal state in the symbolic time series. For all the analyzed time series, we first identified the set of causal states and then plotted the histograms of recurrence times (periods) for each of them. This analysis reveals that the causal states demonstrate a clear separation into two classes, which we will refer to as "core" states (those defined by short-time recurrences) and "noncore" states (those without the well-defined characteristic time scale of recurrence). Core states are characterized by a clearly developed peak at the value of about 0.1 ps [see Fig. 12(b)], while the rest of the causal states are characterized by an exponential distribution of the recurrence times [Figs. 12(e) and 12(f)]. In order to quantify the difference between the two classes, we introduce a dimensionless parameter G that characterizes the presence of the peak in the interval of the recurrences ≤ 1 ps (compared to the interval 1 ps$\leq t \leq 2$ ps),

$$G = \frac{\max(h_1) - m_{12}}{\sigma_{12}}, \tag{4}$$

where h_1 is the value of the recurrence time histogram in the time interval $t \leq 1$ ps, and m_{12}, σ_{12} are the median and the

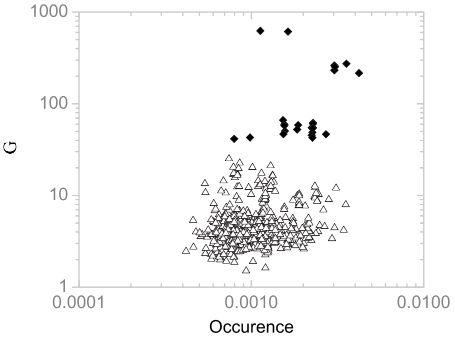

FIG. 11. Clustering of the causal states for the hydrogen atom velocity time series into core (diamonds) and noncore (triangles) classes. Parameter G is plotted vs occurrence rates of corresponding causal states.

standard deviation values for the histogram in the interval $1\ \mathrm{ps} \leq t \leq 2\ \mathrm{ps}$. G can be used as a characteristic of each of the causal states. Its large value indicates high probability of the short-time recurrences, or, in other words, the quasiperiodic nature of the corresponding causal state. The causal states characterized by a low value of G have "exponential" distribution of the return times and do not have pronounced low-order periodicity. In Fig. 11 we plot the scatter diagram representing the apparent clustering of the causal states into two classes with respect to the parameter G. The horizontal axis approximates the occurrence rate [or probability $P(\epsilon_i)$] of the causal states, i.e., for each of them we counted the number of its appearances in the symbolic time series and estimated the probability $P(\epsilon_i)$ by dividing it to the total length of the symbolic series.

Additional support to the observation of two qualitatively different classes of causal states is provided by Fourier analysis. For each of the causal states we generated a binary time series that contained "1" at those time moments where the given causal state was observed, and "0" elsewhere. By calculating the power spectra for the time series corresponding to the causal states we obtain an alternative indication of the difference between the core states and the rest of the set. Core states have a comparatively high level of spectral density in the vicinity of the characteristic period of ≈ 1 ps, whereas "noncore" states have a pronounced gap at this value, Figs. 12(a) and 12(d). This finding implies that the processes with characteristic time scales of ≈ 0.1 ps calculated from the first zero of the correlation function as well as ≈ 1 ps corresponding to the peak of the power spectrum are defined by the core causal states.

As for the phase-shuffled surrogate time series, the set of causal states is found to consist completely from the core states, i.e., the total number, probabilities, and even the symbolic sequences constituting the states approximately coincide with those of the core states of the ϵ machine of the original signal. We can, therefore, conclude that the noncore causal states are the main reason for the high value of the statistical complexity in the MD signal.

The number of noncore states in the original signal grows approximately linearly with time. They are, therefore, responsible for the phenomenon of the logarithmic growth of C_μ with time. The noncore states are defined by the long-time nonlinear correlations that are not captured by the (linear) autocorrelation analysis and completely absent in the surrogate signal.

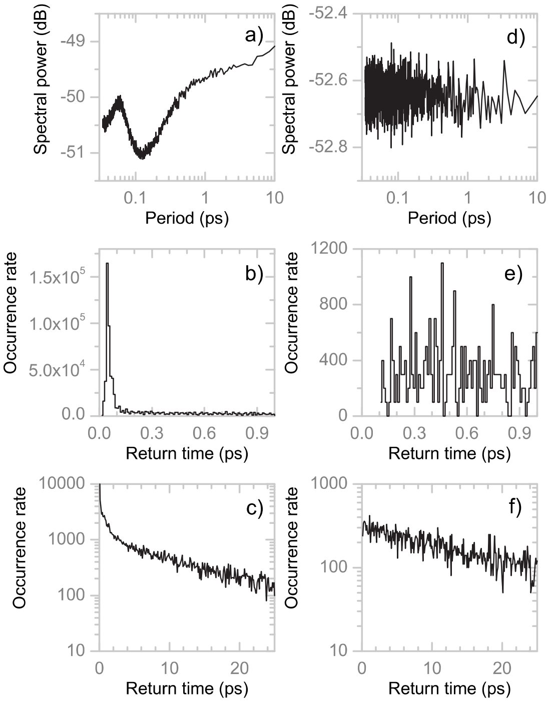

FIG. 12. Power spectra [(a) and (d)] and histograms of recurrence times [(b), (c), (e), and (f)] for typical causal states belonging to different types: a core state [(a)–(c)] and a noncore state [(d)–(f)]. The histograms on (c) and (f) are zoomed and smoothed fragments of those shown in (b) and (e). Spectra in (a) and (d) are the functions of inverse frequency.

Summarizing, the core states are always present, whatever the length of the time series or the location of the time window on the time axis. The fact of the invariant presence of the core states indicates their key role in the formation of the power spectrum and correlation function. The rest of the ϵ machine represents nontrivial, nonlinear, long-term processes that describe the way the system explores the phase space. For a molecular trajectory, the number of noncore states is high, indicating a perpetual process of exploring new areas in the phase space, whereas the absence of such states in the case of surrogate time series shows statistical stationarity of the latter and uniform phase-space coverage property of the surrogate trajectory.

VI. DISCUSSION AND CONCLUSIONS

The computational mechanics approach utilizing information theoretic concepts of the ϵ machine and statistical complexity are used for describing the high-dimensional molecular dynamics of an ensemble of 392 water molecules.

The problem of finding hidden regular patterns in a time series that appears to be a random process is addressed by the analysis of the phase-space filling property by an individual trajectory. The presence of patterns is judged by sig-

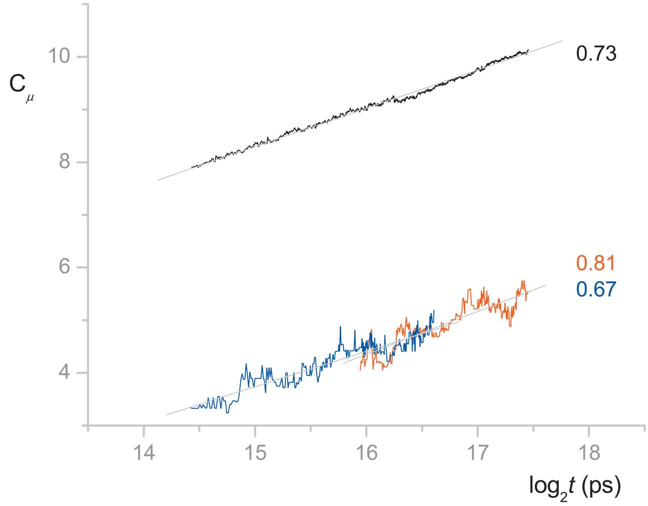

FIG. 13. (Color) h_Q values (indicated on the right) for various observables: black—the hydrogen velocity, red—the oxygen velocity, and blue—the instantaneous temperature.

nificant deviations from the uniform coverage of the phase space expected for the case of a random process.

Long-range memories present in the molecular dynamics simulations are detected and investigated by the means of statistical complexity analysis. It is shown that arbitrary long memories (much longer than one can expect from a spectral or correlation analysis) are present in the recorded time series, manifesting themselves as groups of causal states in the velocity-defined phase space.

The noncore causal states change with the length of the analyzed molecular signal reflecting subtle differences in the statistics of the sequences of data points over time scales that are orders of magnitude longer than the common microdynamics correlations. It should be stressed that these long-range correlations cannot be detected using the usual linear two-point statistics: the correlation function is essentially zero at all times for the data points spaced with intervals longer than a few picoseconds.

The time dependence of the statistical complexity value presumably comes from the fact that the microstate sampling is a slow process due to the extremely high value of the dimension of phase space. Since this time dependence originates from the dimensionality of the phase space, the rate of the complexity change, h_Q [Eq. (3)], should be an invariant for any microscopic observable (provided that this observable is exhaustively sampled at these times). We have tested this hypothesis by comparing the complexity values obtained for the velocities of oxygen and hydrogen atoms to the time series of the instantaneous temperature T_{inst}. The result is presented in Fig. 13 where we plot the dependencies of C_μ on time for one of the MD simulations. It is clear that the slopes of all the curves shown in Fig. 13 are indeed the same within numerical tolerance.

Finally, statistical complexity turns out to be a universal measure of dynamical structures present in the observed data. A comparison to surrogate data sets with broken dynamical correlations supports the hypothesis that the patterns are not caused by the details of the computational procedure, intrinsic statistical errors, or insufficient data, but by the complex dynamics of the system.

The rate of the temporal change in the statistical complexity value reflects the way the phase space of the system is explored (filled) by the trajectory. It can be conjectured that the exponent h_Q represents a universal physical constant characterizing water, since it does not depend on the specific macroscopic observable analyzed, parameters of the system or simulation model, such as temperature, number of molecules, or numerical model employed. It is also independent of the details of the data processing such as the choice of the phase-space partition used in the symbolization of the time series, the number of symbols, or the length of the histories used for reconstruction of the ϵ machines. However, it does depend on the particular substance used in the simulations. For example, we have also investigated the case of a very different molecular system, liquid argon (Lennard-Jones liquid) and found significantly different values of h_Q compared to the case of water. This will be the subject of future publications.

The main results obtained in this paper by the analysis of the statistical complexity of a single MD trajectory (not an ensemble of independent trajectories) can be roughly summarized as follows: (i) over a long (nanoseconds) time scale nontrivial structures in the probabilistic space (far beyond the decay time of a conventional linear autocorrelation function) are detected and analyzed; (ii) the probabilistic (causal) states demonstrate apparent clustering by the parameter G characterizing their periodicity (recurrence times); (iii) a measure is introduced that quantifies the growth of statistical complexity, i.e., the way the molecular system explores the phase space; (iv) the analysis of surrogate data reveals the absence of any significant causal states structures in the surrogate time series, thus indicating the dynamic nature of the temporal patterns that form causal states in the MD time series.

ACKNOWLEDGMENTS

This work was supported by Unilever and the European Commission (EC Contract No. 012835—EMBIO). The authors are grateful to an anonymous referee for careful reading of the manuscript and constructive comments that resulted in significant improvement of the paper.

APPENDIX: COMPUTATIONAL MECHANICS

In nonlinear dynamics, common dynamic invariants such as dimensions, entropies, and Lyapunov exponents, are, in essence, simply sets of numbers used for characterizing the dynamics of the system. Obviously, it is hopeless to expect that one number or some small set of numbers can adequately describe a multidimensional complicated dynamical system such as water. In contrast, computational mechanics extracts all statistically significant information from the signal at the same time achieving the maximal information compression (see below), thus, providing a very desirable description of the dynamics.

Computational mechanics analyzes symbolic dynamics, that is, a sequence of symbols, $\ldots s_{-2}s_{-1}s_0s_1s_2\ldots$, from a finite alphabet of size K. All past s_i^- and future s_i^+ halves of bi-infinite symbolic sequences centered at times i are considered. Two pasts s_1^- and s_2^- are defined equivalent if the conditional distributions over their futures $P(s^+|s_1^-)$ and $P(s^+|s_2^-)$ are equal. A *causal state* $\epsilon(s_i^-)$ is a set of all pasts equivalent to s_i^-: $\epsilon_i \equiv \epsilon(s_i^-)=\{\lambda: \ P(s^+|\lambda)=P(s^+|s_i^-)\}$. At a given moment the system is at one of the causal states, and moves to the next one with the probability given by the transition matrix $T_{ij}\equiv P(\epsilon_j|\epsilon_i)$. The transition matrix determines the asymptotic causal state probabilities as its left eigenvector $P(\epsilon_i)T=P(\epsilon_i)$, where $\Sigma_i P(\epsilon_i)=1$. The collection of the causal states together with the transition probabilities define an ϵ machine.

It is proven [40] that the ϵ machine is

(1) a *sufficient* statistic, that is, it contains the complete statistical information about the data;

(2) a *minimal sufficient* statistic, therefore the causal states cannot be subdivided into smaller states; and

(3) a *unique minimal sufficient* statistic, any other one simply relabels the same states.

The *statistical complexity* $C_\mu = 0$ is the information measure of the size of the ϵ machine that quantifies the amount of information about the past of the system that is needed to predict its future dynamics: $C_\mu = H[P(\epsilon_i)]$, where H is the Shannon entropy of the distribution of a random variable X, $H[P(X)]\equiv -\Sigma_X P(X)\log_2 P(X)$. ϵ machines can be reconstructed from observed data using the CSSR algorithm described and implemented in Ref. [41].

Statistical complexity measures the informational content of the dynamics by searching and quantifying dynamical patterns in the signal. $C_\mu = 0$ both for completely random (all values are independent) and completely ordered (constant value) signals. For the intermediate cases the level of the order is estimated. However, in contrast to, for example, Fourier analysis, the shape of the patterns is not prescribed. Any patterns present are *discovered*.

For subsequences s^-, s^+ of a finite length l the upper limit of the statistical complexity realized is when all of them are unique. In this case the number of causal states equals K^l with probabilities $1/K^l$, where K is the alphabet size. Thus, the maximum possible value of the complexity is $l \log_2 K$.

[1] A. M. Berezhkovskii and G. Sutmann, Phys. Rev. E **65**, 060201(R) (2002).

[2] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).

[3] J. P. Crutchfield and K. Young, *Entropy, Complexity, and Physics of Information, SFI Studies in the Sciences of Complexity, VIII*, edited by by W. Zurek (Addison-Wesley, Reading, Massachusetts, 1990).

[4] J. P. Crutchfield, Physica D **75**, 11 (1994).

[5] A. Cohen and I. Procaccia, Phys. Rev. A **31**, 1872 (1985).

[6] V. Afraimovich and G. Zaslavsky, Chaos **13**, 519 (2003).

[7] A. A. Brudno, Trans. Mosc. Math. Soc. **2**, 127 (1983).

[8] S. Segre, Int. J. Theor. Phys. **43**, 1371 (2004).

[9] V. Benci, C. Bonanno, S. Galatolo, G. Menconi, and M. Virgilio, Discrete Contin. Dyn. Syst., Ser. B **4**, 935 (2004).

[10] *Coping with Chaos. Analysis of Chaotic Data and the Exploitation of Chaotic Systems*, edited by E. Ott, T. Sauer, and J. A. Yorke (John Wiley, New York, 1994).

[11] *Dimensions and Entropies in Chaotic Systems*, edited by G. M. Kress (Springer, New York, 1986).

[12] J. Cao, Phys. Rev. E **63**, 041101 (2001).

[13] R. Cukier and J. Deutch, Phys. Rev. **177**, 240 (1969).

[14] A. Rahman, Phys. Rev. **136**, A405 (1964).

[15] M. G. Mazza, N. Giovambattista, F. W. Starr, and H. E. Stanley, Phys. Rev. Lett. **96**, 057803 (2006).

[16] P. Gaspard, M. E. Briggs, M. K. Francis, J. V. Sengers, R. W. Gammon, J. R. Dorfman, and R. V. Calabrese, Nature (London) **394**, 865 (1998).

[17] C. P. Dettmann and E. G. D. Cohen, J. Stat. Phys. **101**, 775 (2000).

[18] G. M. Zaslavsky, Phys. Rep. **371**, 461 (2002).

[19] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer, Physica D **58**, 77 (1992).

[20] J. Boon and S. Yip, *Molecular Hydrodynamics* (McGraw-Hill, New York, 1980).

[21] U. Balucani and M. Zoppi, *Dynamics of the Liquid State* (Oxford University Press, Oxford, 1994).

[22] B. M. Ladanyi and M. S. Skaf, Annu. Rev. Phys. Chem. **44**, 335 (1993).

[23] P. Grassberger and H. Kantz, Phys. Lett. A **113**, 235 (1985).

[24] J. Plumecoq and M. Lefranc, Physica D **144**, 231 (2000).

[25] M. A. Ricci, D. Rocca, G. Ruocco, and R. Vallauri, Phys. Rev. A **40**, 7226 (1989).

[26] M. Wojcik and E. Clementi, J. Chem. Phys. **85**, 6085 (1986).

[27] J. Teixeira, M. C. Bellissent-Funel, S. H. Chen, and B. Dorner, Phys. Rev. Lett. **54**, 2681 (1985).

[28] H. J. C. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans, in *Intermolecular Forces*, edited by B. Pullman (D. Reidel Publishing Company, Dordrecht, 1981), pp. 331–342.

[29] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).

[30] H. J. C. Berendsen, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis (Kluwer, Dordrecht, 1991), pp. 139–155.

[31] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[32] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (Springer, New York, 1990).

[33] M. Buhl and M. B. Kennel, Phys. Rev. E **71**, 046213 (2005).

[34] F. Takens, in *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics 898, edited by D. A. Rand and L.-S. Young (Springer-Verlag, New York, 1981), pp. 366–381.

[35] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*, Cambridge Series on Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, England, 1997).

[36] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, New York, 2003).

[37] P. Grassberger and I. Procaccia, Physica D **9**, 189 (1983).

[38] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).

[39] G. Zaslavsky, Phys. Rep. **371**, 461 (2002).

[40] C. R. Shalizi, K. L. Shalizi, and R. Haslinger, Phys. Rev. Lett. **93**, 118701 (2004).

[41] C. R. Shalizi, K. L. Shalizi, and R. Haslinger, in *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference*, edited by M. Chickering and J. Halpern (AUAI Press, Arlington, Virginia, 2004), pp. 504–511; URL http://arxiv.org/abs/cs.LG/0406011.