Computational mechanics of molecular systems: Quantifying high-dimensional dynamics by distribution of Poincaré recurrence times

Vladimir Ryabov, and Dmitry Nerukh

Citation: Chaos **21**, 037113 (2011); doi: 10.1063/1.3608125
View online: https://doi.org/10.1063/1.3608125
View Table of Contents: http://aip.scitation.org/toc/cha/21/3
Published by the American Institute of Physics

Computational mechanics of molecular systems: Quantifying high-dimensional dynamics by distribution of Poincaré recurrence times

Vladimir Ryabov[1] and Dmitry Nerukh[2]
[1]*Future University-Hakodate, School of Systems Information Science, Department of Complex System, 116-2 Kamedanakano-cho, Hakodate-shi, 041-8655 Hakodate, Hokkaido, Japan*
[2]*Department of Chemistry, Cambridge University, Cambridge CB2 1EW, United Kingdom*

(Received 24 August 2010; accepted 17 May 2011; published online 30 September 2011)

A framework that connects computational mechanics and molecular dynamics has been developed and described. As the key parts of the framework, the problem of symbolising molecular trajectory and the associated interrelation between microscopic phase space variables and macroscopic observables of the molecular system are considered. Following Shalizi and Moore, it is shown that causal states, the constituent parts of the main construct of computational mechanics, the ϵ-machine, define areas of the phase space that are optimal in the sense of transferring information from the micro-variables to the macro-observables. We have demonstrated that, based on the decay of their Poincaré return times, these areas can be divided into two classes that characterise the separation of the phase space into resonant and chaotic areas. The first class is characterised by predominantly short time returns, typical to quasi-periodic or periodic trajectories. This class includes a countable number of areas corresponding to resonances. The second class includes trajectories with chaotic behaviour characterised by the exponential decay of return times in accordance with the Poincaré theorem. [doi:10.1063/1.3608125]

Complex dynamics in systems with multiple degrees of freedom, like, for example, an ensemble of water molecules, seems to be indistinguishable from noise by any standard statistical method. At the same time, the motion of every atom is described by deterministic differential equations; hence, the signatures of deterministic Hamiltonian dynamics are contained in the time dependent coordinates and momenta. We consider computational mechanics as a bridge between deterministic chaos in nonlinear dynamical systems with few degrees of freedom and apparently random trajectories in the high-dimensional phase space. The construction of an ϵ-machine allows decomposing the phase space into non-overlapping elementary areas of two qualitatively different classes, depending on the decay law for Poincaré recurrence times. By an analogy with standard map, they can be attributed to "chaotic sea" and quasiperiodic motions in the vicinity of ("sticky") periodic islands. The proposed method of identifying the areas with sticky dynamics in the high-dimensional phase space has far reaching implications in understanding the molecular transport, including the anomalous diffusion process. It is important, for example, for elucidating general regularities underlying the complex motions of protein atoms in the process of folding or other self-organising biomolecular dynamics.

I. INTRODUCTION

The trajectories of atoms and molecules in liquids can be described by Newtonian ordinary differential equations of motion. Therefore, any complex patterns formed by the molecules due to their mutual interactions have geometric counterparts in the phase space defined by their coordinates and velocities. The problem of identifying and classifying the patterns as well as predicting their appearance is crucially important since they ultimately define the functionality of the systems and can provide keys to understanding the fundamental properties of, for example, protein folding. There is, however, a profound difficulty in the dynamical picture of molecular systems related to high dimensionality of their phase space. Commonly used approaches from non-linear dynamics, such as Lyapunov exponents, dimensions, and entropies fail in most cases when the motion occurs in the phase space of dimension higher than ≈ 10. Therefore, new conceptually different methodologies have to be developed for high-dimensional systems.

An alternative description in terms of probability and statistics can be and has been successfully applied in many situations to systems with too complicated behaviour. However, due to the way the probability theory is built, that is its axiomatic assumption of a priori given distribution functions, it has limited potential of understanding the dynamic patterns in open systems demonstrating highly complex non-stationary behaviour.

Computational mechanics (CM), a promising new concept aimed at building a statistical and at the same time dynamical description, has been recently proposed.[1] It combines the well-developed theoretical framework of generalised Markov chains, called ϵ-machines, with the concept of short time predictability characteristic to dynamical systems.

Since typical motions of molecules ultimately define their conformational rearrangements, complete quantitative

analysis of the patterns in the trajectory provided by CM gives new insight into molecular mechanisms. Our goal, thus, is to find persistent structures in the phase space formed by the trajectories and interpret typical behaviour of such structures in terms of both the statistical theory and the dynamical systems approach. We analyse trajectories of molecular dynamics (MD) simulated systems where the coordinates and momenta of the atoms can be obtained with any reasonable precision. Using the complexity based measures[1,2,7] and analysing the probabilistic properties of symbolic sequences corresponding to the phase space trajectories are promising alternatives for detecting structures in the phase space.

One of the most difficult problems in the analysis of the high-dimensional molecular trajectories is the definition of the notion of "structure" or "cluster" in the phase space. We address this issue in a broad statistical sense considering deviations from the uniform phase space filling by a typical trajectory as clusters. The presence of structures in the phase space of dynamical systems can be interpreted as the existence of nonuniformities in the invariant measure.[3] The latter defines the probabilities of visiting various parts of the phase space by trajectories or, under the assumption of ergodicity, by a typical trajectory observed for a long enough period of time. The clusters appear in the phase space due to the presence of abundant resonances that arise as a result of nonlinear interactions between atoms. The borders of resonant areas are known to be "sticky" in a sense that any trajectory spends a long time in their vicinity. This is in contrast to other, non-resonant areas, where the trajectories evolve randomly filling the phase space almost uniformly.

In simple Hamiltonian systems like, for example, low-dimensional area preserving maps, the resonant areas appear as the islands of stability in the phase space. They are known, on the one hand, as sources of nonuniformity in the invariant measure and, on the other hand, they lead to breaking the ergodicity due to the formation of impermeable and "sticky" barriers in their vicinity.[4] The islands typically have a fractal structure, and the finer is the scale of subislands the more "sticky" are their borders for trajectories, that is a typical trajectory, once trapped by such a structure, remains there for a very long time.

A quantitative description of the nonuniformity of the phase space covering by the trajectories can be achieved via the Poincaré recurrence theory.[5] Consider a small element $\Delta\Gamma$ of the phase space Γ of a Hamiltonian system located around the point $\mathbf{x}$. A trajectory wanders in the chaotic area visiting the element $\Delta\Gamma$ from time to time (recurring to it). Denoting the time between successive recurrences as τ, the probability distribution function of recurrence times $P(\Delta\Gamma, \mathbf{x}, \tau)$ can be introduced that depends on the phase volume and the position of the element $\Delta\Gamma$, as well as the value of τ itself. If the motion is ergodic, the dependence of τ on the coordinates $\mathbf{x}$ becomes inessential and one can introduce the distribution function

$$P(\tau) = lim_{\Delta\Gamma\to 0} P(\Delta\Gamma, \tau)/\Delta\Gamma. \quad (1)$$

For a typical chaotic trajectory, the following asymptotic relation holds

$$P(\tau) = \frac{1}{\langle\tau\rangle} exp(-\tau/\langle\tau\rangle), \quad (2)$$

where $\langle\tau\rangle$ is the average recurrence time over the distribution $P(\tau)$. Equation (2) can be used, in principle, for distinguishing areas with chaotic motion from those close to sticky areas by introducing a partition of the phase space into non-overlapping volumes and analyzing the distributions $P(\tau)$ for each of them. Note also, that the problem of choice of the sizes and shapes of the partition elements is not a trivial one and, in the general case, the distribution of the Poincaré recurrence times can depend on the location and shape of the area $\Delta\Gamma$.

It is also important that stickiness of resonant areas leads to anomalous transport properties of the trajectories in the phase space. This issue attracted a lot of attention recently[4] and it has been demonstrated that key insights into the details of the transport can be found in terms of the Poincaré theorem of returns.

In this paper, we show how the analysis of molecular trajectories in terms of ϵ-machine associated with the notion of statistical complexity (SC) provides a link from a purely statistical description with Markov chain type modeling to the dynamical systems theory based on Poincaré recurrence analysis. One of the pressing questions in the analysis of very high-dimensional trajectories by Poincaré recurrences is the choice of areas of interest. In other words, it is not clear how to create a partition in the phase space that would give a meaningful description of the dynamics in terms of the recurrence times of trajectories. We demonstrate that causal states, that constitute a core of an ϵ-machine, provide a "natural" partitioning in the phase space in the sense that for each causal state the Poincaré recurrence times are distributed in accordance with Eq. (2). Therefore, every causal state can be considered as an element of the phase space that contains a set of Poincaré cycles. Moreover, considering the deviations from Eq. (2) allows identifying different types of motion in the phase space, that is finding the causal states with special properties that correspond to the resonance areas within chaotic sea in the phase space. The goal of making a quantitative distinction between the areas with qualitatively different dynamics can thus be achieved by comparing the decay times of Poincaré recurrences to causal states.

II. COMPUTATIONAL MECHANICS DEALS WITH CERTAIN PARTITIONING OF THE MOLECULAR PHASE SPACE

A. Initial symbolisation produces a very coarse grained partition

A molecular trajectory obtained in the simulation experiment is a series of $2N$-dimensional phase space points $\mathbf{q}_i$, where N is the number of degrees of freedom of the system, i.e., the number of atoms multiplied by 3 minus various constraints such as fixed bond lengths, angles, etc. N is of the order of several thousands for realistic MD simulations. Thus, the molecular trajectory is a very high-dimensional object. The points are generated by the system along the trajectory at fixed time moments (Fig. 1).

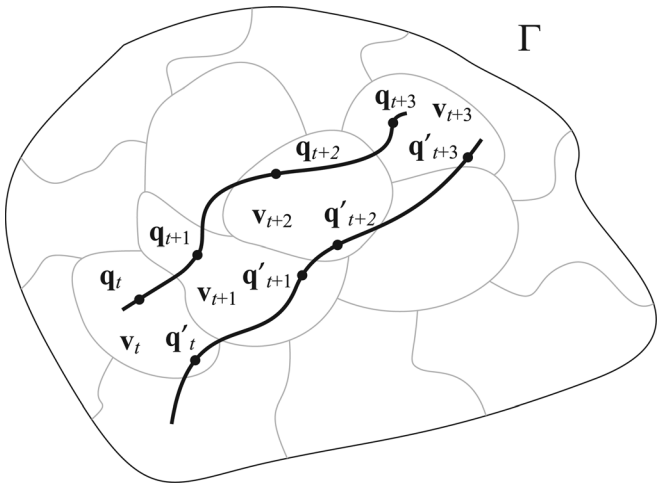

FIG. 1. Illustration of the degeneracy of the macro-observable projection of the full-dimensional phase space trajectory. The same sequence of the observable (the velocity) $\{\mathbf{v}_t\mathbf{v}_{t+1}\mathbf{v}_{t+2}\mathbf{v}_{t+3}\}$ is generated by two different pieces of the phase space trajectory $\{\mathbf{q}_t\mathbf{q}_{t+1}\mathbf{q}_{t+2}\mathbf{q}_{t+3}\}$ and $\{\mathbf{q}'_t\mathbf{q}'_{t+1}\mathbf{q}'_{t+2}\mathbf{q}'_{t+3}\}$.

To analyse the data inevitably low-dimensional observables (macro-observables) are considered, for example, a velocity of an atom $\mathbf{v}$. This is a projection of the full-dimensional trajectory onto a low-dimensional observable, in this case $\mathbf{v}$. This low-dimensional projection of the phase space trajectory is degenerate that is very many different realisations of the trajectory produce the same series of values of the low-dimensional projection $\mathbf{v}$ (Fig. 1). This is caused by (i) the discrete time sampling of the trajectory, (ii) the finite tolerance of the measurements of $\mathbf{v}$, and (iii) the independence of $\mathbf{v}$ at each individual time moment from some other degrees of freedom, for example the positions and velocities of atoms at a large distance. Therefore, the whole phase space Γ is partitioned into the areas such that on each of them, the macroscopic observable $\mathbf{v}$ takes a single value while the full-dimensional points $\mathbf{q}_i$ can have different values (Fig. 1).

The values of the observable variables that we analyse are discrete and finite. In other words, we deal with a set of countable number of *symbols*. In the case of the computer floating point representation, for example, the number of symbols is large but limited and defined by the precision used in the simulation (single, double, etc). The finite precision of $\mathbf{v}$ results in a finite (but large) set of its possible values. Moreover, it is easy to check that even a very coarse representation of $\mathbf{v}$ produces almost the same characteristics of the analysed molecular signals. Figure 2 shows one of such characteristics, the common velocity autocorrelation function for a signal where the velocity coordinates are replaced by only three values in $\mathbf{v}_x$, $\mathbf{v}_y$, and $\mathbf{v}_z$, such that $\{x \equiv -1, \text{if} x < -1; x \equiv 0, \text{if} -1 \le x < 1; x \equiv 1, \text{if} x \ge 1\}$, where x represent $\mathbf{v}_x$, $\mathbf{v}_y$, and $\mathbf{v}_z$. The total number of possible values of the resulting coarse grained vector is $3^3 = 27$ that is the signal can be represented by only 27 symbols. Nevertheless, the autocorrelation function of this signal is very similar to the original one, calculated from the double precision values of $\mathbf{v}$.

This representation of the dynamics in terms of symbols from a finite size alphabet is called "symbolic dynamics" and is the subject of the mathematical field with the same name.[6]

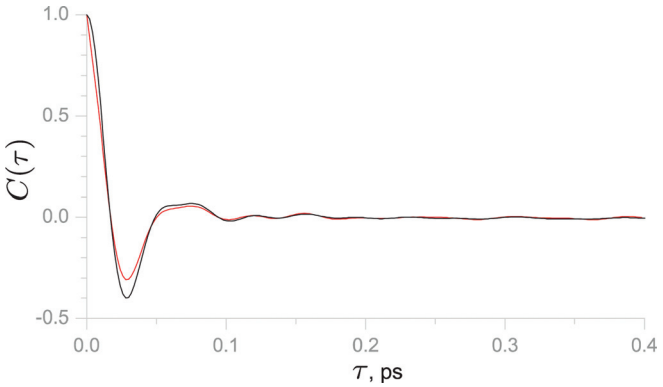

FIG. 2. (Color) Autocorrelation functions $C(\tau) \equiv \frac{1}{T}\sum_t^T \mathbf{v}_t \cdot \mathbf{v}_{t+\tau}$ for the original velocity of the hydrogen of bulk water (black) and the signal made of 27 symbols (red, see text for details).

Summarising, by converting the continuous trajectory into the symbolic sequence, a coarse grained partition of the phase space is produced.

B. The dynamics makes the partition finer

The evolution of the phase space points $\mathbf{q}$, sampled at times t, is governed by an operator $\mathbf{T}$: $\mathbf{q}_{t+1} = \mathbf{T}\mathbf{q}_t$. Considering an ensemble of such dynamical systems, denote a random variable representing the current microstate as $\mathbf{Q}$, that is a set of all possible values of the phase space points having probabilities generated by the dynamics $\mathbf{T}$.

A macroscopic observed variable A is a function f of the microstate $\mathbf{Q}$ (for example, the instantaneous temperature $\frac{1}{Nk}\sum_i m_i\mathbf{v}_i^2$, where N is the number of degrees of freedom, k is the Boltzmann constant, m_i are the atoms' masses, and $\mathbf{v}_i$ are their velocities). As discussed before, the function f partitions the phase-space Γ into mutually exclusive and jointly exhaustive sets, on each of which f takes a unique value.

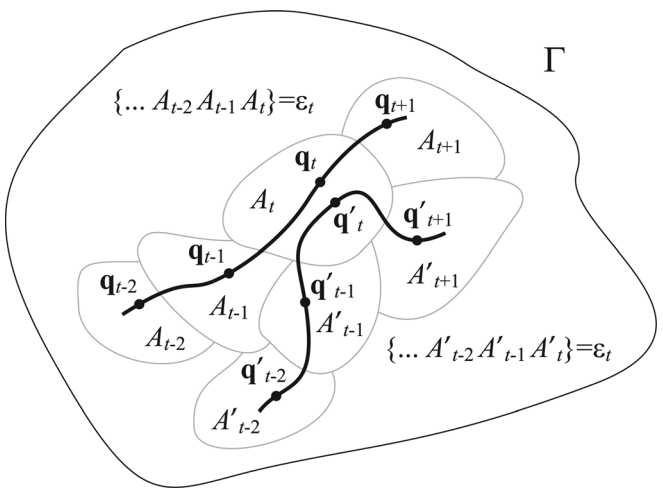

FIG. 3. Schematic illustration of the sequences used to define formula (4). Phase space points $\{\mathbf{q}\}$ and $\{\mathbf{q}'\}$ of two pieces of the trajectory form Markov sequences. The corresponding observation sequences A and A' are not Markovian since the same value A_t leads to different A_{t+1} and $A_{t+1}\prime$ depending on the previous values A_{t-1} and $A_{t-1}\prime$. However, if both histories $\{\ldots A_{t-2}A_{t-1}A_t\}$ and $\{\ldots A_{t-2}\prime A_{t-1}\prime A_t\}$ belong to the same causal state ϵ_t than the next causal state ϵ_{t+1} is defined without knowing ϵ_{t-1}, thus making $\{\epsilon\}$ a Markov sequence.

Denote the partition of Γ induced by f as $\mathcal{F}$. The observed process is $A_t = f(\mathbf{Q})$ and it is not necessarily Markovian (Fig. 3).

Now, what happens to this partition when the sequences of A_t are considered instead of the individual values of A? Take an observation at time t, A_t. The corresponding set of points in Γ is $\mathcal{F}_t$. For a sequence of two observations at the current and previous time moments, the set of points is

$$\mathcal{F}_t \cap \mathbf{T}\mathcal{F}_{t-1}, \tag{3}$$

which is a refinement of the partition $\mathcal{F}$. This procedure can be repeated any countable number of times thus providing the refined partitions for the histories of the macro-observable A. Thus, the dynamics makes the initial partition induced by the macro-observable finer, the longer the sequence $\{A_t\}$ (the "history") the finer the partition generated by the sequence is.

C. Computational mechanics coarsens the partition

The next step is to apply a special statistic, called CM,[7] to the observable A. The rigorous definition of CM is given in Appendix A. Here, we provide the part of the approach necessary for answering the main question formulated in the Introduction.

All past A_i^- and future A_i^+ halves of bi-infinite sequences of the macro observable centered at times i are considered. Two pasts A_1^- and A_2^- are defined equivalent if the conditional distributions over their futures $P(A^+|A_1^-)$ and $P(A^+|A_2^-)$ are equal. A *causal state* $\epsilon(A_i^-)$ is a set of all pasts equivalent to A_i^-: $\epsilon_i \equiv \epsilon(A_i^-) = \{\lambda : P(A^+|\lambda) = P(A^+|A_i^-)\}$. 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 $\sum_i P(\epsilon_i) = 1$. The collection of the causal states together with the transition probabilities define an *ϵ-machine*. The *statistical complexity* is the informational measure of the size of the ϵ-machine: $C_\mu = H[P(\epsilon_i)]$, where P are the probabilities of the causal states and H is the Shannon entropy of the distribution of a random variable ν, $H[P(\nu)] \equiv -\sum P(\nu)\log_2 P(\nu)$.

Thus, the essence of CM is in grouping the histories $\{A_t\}$ into causal states. In terms of the partitions of the phase space, this corresponds to joining together the cells of Γ induced by the dynamics. Importantly, the new cells represent a Markovian process constructed from the observed process A_t by building the ϵ-machine on A. Now, by the ϵ-machine definition, the sequence of the causal states $\{\epsilon_t\}$ makes a Markov chain (Fig. 3).

D. The partition generated by computational mechanics is the most informative one

Shalizi and Moore[8] show that, in this setting, the statistical complexity has a clear physical meaning: it quantifies the amount of information contained in the new constructed macro-observable process $\{\epsilon_i\}$ about the microstate

$$C_\mu = I[\mathbf{Q}; \epsilon], \tag{4}$$

where I is the mutual information between the random variables X and Y: $I[X;Y] = H[X] - H[X|Y]$; and $H[X|Y]$ is a conditional entropy of X given Y: $H[X|Y] = -\sum P(X)\sum P(X|Y)\log_2 P(X|Y)$.

This is because the knowledge of the microstate would specify the macro observable precisely: $H[\epsilon|\mathbf{Q}] = 0$, because all histories contained in ϵ_t and the corresponding partition of $\mathbf{Q}$ would uniquely define the next state ϵ_{t+1} (the ϵ-machine definition). Using this and the equality $H[X] + H[Y|X] = H[Y] + H[X|Y]$, the Eq. (4) follows:

$$\begin{aligned} H[\mathbf{Q}|\epsilon] + H[\epsilon] &= H[\epsilon|\mathbf{Q}] + H[\mathbf{Q}], \\ H[\mathbf{Q}|\epsilon] + C_\mu &= H[\mathbf{Q}], \\ C_\mu &= H[\mathbf{Q}] - H[\mathbf{Q}|\epsilon], \\ C_\mu &= I[\mathbf{Q}; \epsilon]. \end{aligned}$$

Because of the properties of the ϵ-machine, this is the maximal information that is possible to extract from the chosen macro-observable and the specified initial partition of it.

E. Three stages of symbolisation

Summarising, the phase space partition we use in numerical experiments is obtained in three stages.

1. The observed macro variable induces the initial (usually very coarse grained) partition of the phase space defined by the procedures of projection, measurement uncertainty, and symbolisation.
2. The partition elements of this partition are refined by the dynamics, when we consider words (histories) instead of single symbols (3). Note also that considering words instead of symbols is similar to reconstructing the high-dimensional phase space from the scalar time series by the Takens embedding procedure.[9] In terms of the embedding, the histories correspond to different points in the phase space, while the history length l is equal to the embedding dimension.
3. The refined partition elements (histories) are further grouped by the process of ϵ-machine reconstruction, thus providing the final partition that is the minimal, unique, and most informative one (given the initial partition of Γ).

III. IMPLEMENTATION

A. Molecular dynamics simulation

In subsequent sections, we apply the developed theoretical framework to the analysis of dynamics in the ensemble of interacting water molecules. Molecular dynamics is a technique for numerically solving the Newton equations describing the time changes of the atomic coordinates $\mathbf{x}$ and velocities $\mathbf{v} = \dot{\mathbf{x}} : \dot{\mathbf{v}} = \frac{1}{m}\mathbf{F}$. The force $\mathbf{F}$ is derived from the prescribed interatomic interaction potential V (also called the "forcefield"): $\mathbf{F} = -\nabla V(\mathbf{x}_i)|_{i=1..N}$, which is a function of all the coordinates of the atoms. Commonly used forcefields are empirical functions that are the results of careful balance between the sophistication of reproducing realistic interatomic interactions and computational effectiveness. The parameters of forcefields are calibrated to reproduce either

rigorous quantum mechanical calculations or experimental thermodynamical data.

In this work, bulk water (periodic boundary conditions) consisting of 392 or 878 SPC or SPC-E (Ref. 10) molecules was simulated using the GROMACS molecular dynamics[11] package. The temperature of the systems was kept constant at 300 K using Berendsen[12] or Nose-Hoover[13] thermostats whose combination with various coupling constants was investigated. A sufficient equilibration was performed before collecting data for analysis. The velocity of the hydrogen atom of one of the water molecules was used. At the locations where the velocity pierces the xy plane, the points of a two-dimensional map were generated and used as the original continuous signal for analysis.

We have found that the results do not depend on the parameters of molecular simulations such as the forcefield, the temperature, the type of the thermostat, the number of molecules, etc.[15]

B. Symbolisation

The best possible initial partition for converting the floating point double precision time series data to a symbolic string can be achieved using the generating partition (GP). Although it is not possible to find it exactly, there are methods for computing approximations to it. GP provides a partition that preserves all information in the signal. Therefore, the closer approximation to GP is used the more information is transferred from the continuous signal to the symbolic sequence.

For an initial approximation to GP, we have chosen the partition provided by the application of the method described in Ref. 14. The example of calculations with this method for the two-dimensional cross-section of our (three-dimensional) velocity data using 2, 3, 4, and 5 partition elements are shown in Fig. 4. For all cases, the resulting approximations to GP are centrally symmetric (probably, because of the central symmetry of the data points distribution). The symmetry of the two-dimensional set of points can be further illustrated by transforming the data to the polar coordinates $(x, y) \rightarrow (\rho, \varphi)$ and estimating the probability density $w(\varphi)$ for the random variable φ. The histogram corresponding to such $w(\varphi)$ distribution is given in Fig. 5. Almost perfect uniformity of the distribution function is obvious, thus justifying the choice of centrally symmetric partitions that we used in all subsequent calculations. The results for three and more partitions (the number of symbols) are qualitatively the same, see Ref. 15 for the discussion on the number of symbols. We used three symbol alphabet in all reported results. The simulated trajectory of 1 μs long resulted in approximately 3×10^7 data points (symbols).

C. ϵ-machine reconstruction: CSSR

At the next step of the analysis, we change the description from considering the separate symbols in the symbolic string to the study of histories (symbolic words of finite length) and building the ϵ-machine. For this purpose, we use the method developed by Shalizi with co-authors who also proposed an algorithm of reconstructing the ϵ-machine from the given data series.[16] In a general case, CM is formulated using the assumption of infinitely long pasts and futures. In practice, a finite history length l has to be chosen and this is one of the adjustable parameters of the CSSR algorithm. The number of possible histories grows exponentially with the history length. Therefore, for long histories, an exponential increase in the number of data points is also needed.

The second parameter of the CSSR algorithm is the significance level σ used in comparing the distributions $P(\vec{s} \mid \overleftarrow{s_i})$ for grouping the histories into causal states by their predictive properties (the Kolmogorov-Smirnov test is used). Too large σ values (too strict threshold for two distributions to be considered equivalent) lead to artificially too many causal states. This is equivalent to under-sampling the histories. The same situation takes place for too long history length since the number of possible histories is too large and, for moderately long experimental time series, the distributions $P(\vec{s} \mid \overleftarrow{s_i})$ become not statistically significant.

Therefore, for obtaining the robust results, it appears necessary to perform the analysis of the ϵ-machine as a function of these two parameters. Too long a history or too large a σ

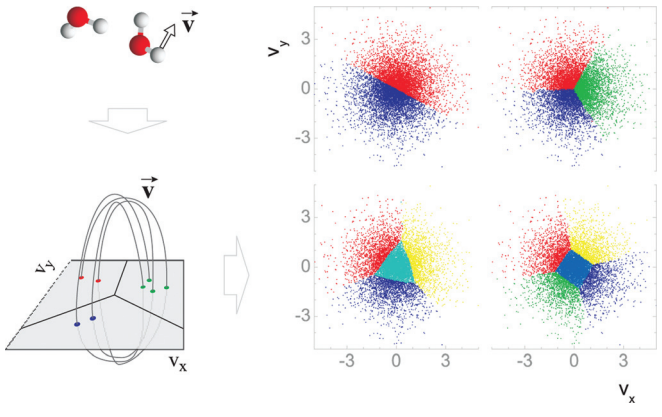

FIG. 4. (Color) The process of converting the continuous atomic velocity signal **v** into symbolic sequence. On the right, the symbolisation with 2, 3, 4, and 5 symbols are shown. 3 symbol alphabet was used in all subsequent calculations.

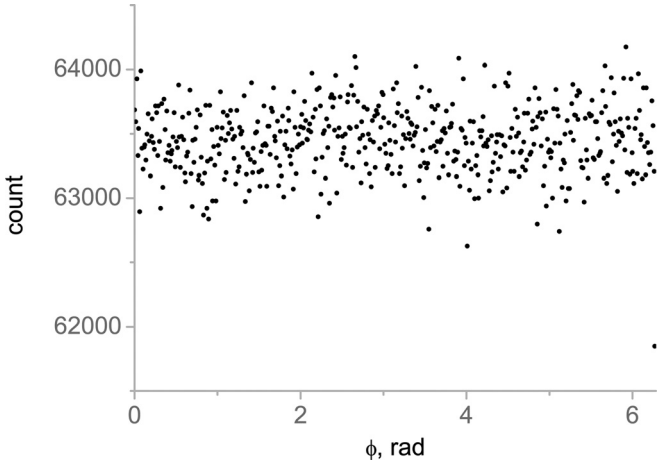

FIG. 5. The histogram of the random variable φ illustrating the uniform symmetric distribution of points in the $v_z = 0$ cross-section plane of the hydrogen velocity trajectory.

value leads to statistically incorrect results. As the authors of CSSR recommend, the value of σ should be chosen such that there is a "plateau" in the number of causal states as a function of l. If there are several such values of l, then the lowest one has to be chosen (according to the minimality principle of CM). This constant value of l is the "true" value of the history length for a stable ϵ-machine architecture.

Our analysis of the convergence of CSSR algorithm with history length is presented in Fig. 6. Here, we compare the time series from MD simulation of water to a so-called surrogate data that has identical power spectrum and hence autocorrelation function, but is stochastic in the sense of absent dynamic correlations between adjacent points.[19] The results suggest that the choice of a plateau is not a trivial task. Surrogate data (left panel) exhibits a clear plateau at any given value of the length of time series, but the minimum value of l (at the smallest σ-value) grows from ≈ 3 to ≈ 6 when the data length increases from 60 ns ($\approx$ 2 000 000 symbols) to 1 μ s ($\approx$ 30 000 000 symbols). The plateau for the water signal (right panel in Fig. 6) is less pronounced, but still visible at the low values of σ that corresponds to, e.g., $l = 10$ for the data length above $\approx 0.45\,\mu s$. The very slow convergence of the ϵ-machine with data size for molecular signals has been discussed in Refs. 15, 17, 18. The results for $l = 10$ are at the limit of statistical reliability for the used data lengths. We, therefore, used $l = 9$ in the analysis discussed here. It is worth noting however, that the main result of this work, i.e., the splitting of the causal states into two groups, does not depend on the value of l starting from the value of $l \approx 6$.

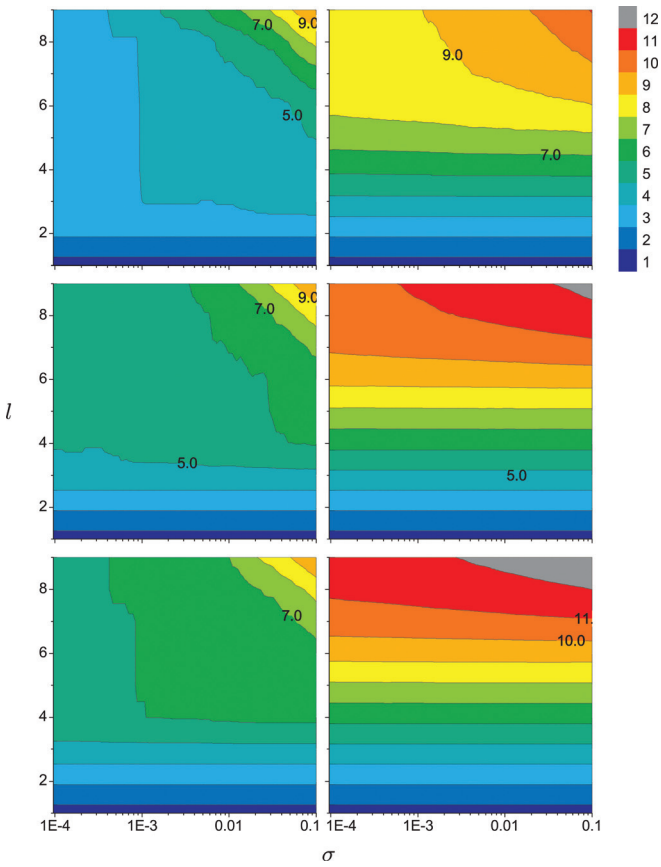

FIG. 6. (Color) The logarithm of the number of causal states $\log_2 n_{st}$ in the ϵ-machines as a function of the history length l, the tolerance σ, and the duration of the time series; left: the Fourier surrogate time series,[19] right: the molecular signal; top: time series duration of 60 ns, middle: 450 ns, bottom: 1 μ s.

IV. RESULTS

As it has been stated in Sec. II B, the symbolic words (histories) that we analyse correspond to the elements of partitioning the phase space into non-overlapping areas. Further, joining the histories into the causal states produces a more coarse grained partition that possesses certain Markovian properties and defines the ϵ-machine through the distribution function of their occurrence rates $P(\epsilon_i)$. To get a further insight into the link between the ϵ-machine and the dynamics, we analyse the distribution of recurrence times for the set of causal states considering them as elements of the phase space partitioning. In order to introduce the recurrence times, we looked at the time intervals between the successive appearances of a causal state in the symbolic time series. For all the analysed data, we first identified the set of causal states and then analysed the histograms of the recurrence times (periods) for each of them.

As it follows from the Poincaré theory, the recurrence times for chaos in systems where only chaotic motions exist are distributed in accordance with Poissonian distribution (Eq. (2)). This turns out to be different in the case of chaotic dynamics in area preserving maps with divided phase space[20] where areas of chaos coexist with periodic behavior (periodic islands). In such systems, the exponential decay in the distribution function turns into a power law

$$P(\tau) \propto \tau^{-\gamma}, \tag{5}$$

at very large times, as it is shown in Fig. 7.

In numerical analysis, the function $P(\tau)$ depicted in Fig. 7 is usually calculated by averaging the corresponding distribution functions over many trajectories randomly chosen by specifying arbitrary initial conditions. This corresponds to the analysis of the recurrence times in many randomly located small areas centered around the randomly selected centers. In our calculation, we assume the ergodicity of the system and calculate similar distributions considering a single time series only. We also select the areas for the analysis of Poincaré recurrences coinciding with the phase space partitioning imposed by the set of the causal states. Surprisingly, we found that for each causal state, the distribution of

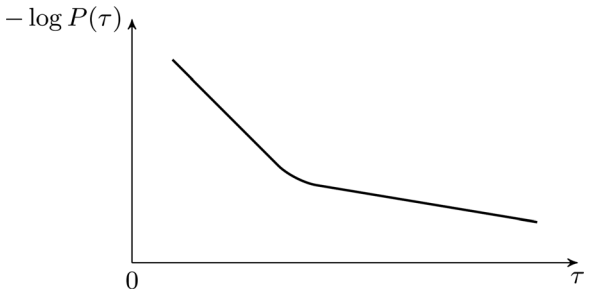

FIG. 7. A schematic of the probability distribution function for the Poincaré recurrence time τ. The exponential decay at small τ changes to apparently slower decay law for long times.

recurrence times at large enough times is well approximated by exponential function, i.e., every causal state contains a set of statistically independent Poincaré cycles.

In order to compare our results pertaining to the analysis of the MD trajectories in water to those available in the literature, we have chosen a well documented two-dimensional area preserving system known as Standard map (or sometimes called the Taylor-Greene-Chirikov map).[21] It is defined as a transformation of the plane to itself

$$P_{n+1} = P_n + K \sin\theta_n,$$
$$\theta_{n+1} = \theta_n + P_n + K \sin\theta_n,$$

where P and θ are computed mod 2π and K is a positive parameter that controls different kinds of behaviour that the system can demonstrate. For an example of chaotic trajectory in this system, we calculated the time series at the value of parameter $K = 6.908745$, where the phase portrait has a large chaotic area containing two stability islands symmetrically located with respect to the origin. An infinite number of smaller islands also exists in the vicinity of the large ones, making the dynamics in their vicinity complicated. Furthermore, we applied a simple uniform partitioning in the variable θ into three equal intervals, i.e., we have chosen a three-symbol alphabet in a similar manner to the case of water time series analysis.

Our numerical experiments for both the water time series and that of the Standard map reveal that the causal states demonstrate a clear separation into two classes that we will refer to as "periodic" states (those defined by Poincaré recurrence times decaying much slower compared to Eq. (2)) and "chaotic" ones (that demonstrate an exponential decay in accordance with Poincaré law). In order to quantify the difference between the two classes, we introduce a dimensionless parameter D, equal to the discrepancy between the decay exponent λ calculated from the histogram of recurrence times and its "normal" value $\frac{1}{\langle\tau\rangle}$ defined by the Eq. (2)

$$D = \frac{1}{\lambda\langle\tau\rangle} - 1, \tag{6}$$

where λ is the exponent defining the shape of the distribution function

$$P(\tau) \propto exp(-\lambda\tau), \tag{7}$$

found numerically. Large D values indicate strong discrepancy between the calculated value of the exponent in Eq. (7)) and the expected value of $1/\langle\tau\rangle$.

The power law tail in the distribution of the recurrence times (Eq. (5)) can be clearly visualised for the case of the Standard map, if we plot the corresponding distribution averaged over all causal states that constitute the ϵ-machine (Fig. 8(a)). Two segments in the distribution function corresponding to normal chaotic behaviour defined by Eq. (2) and abnormally long recurrences in the tail are evident. However, in the case of water time series, there is no apparent distinction between the two types of behaviour as it is shown in Fig. 8(b). In the simplest approximation, this could be interpreted as the absence of periodic islands in the phase space of water because of the breaking of all invariant tori that occurs due to interaction between resonances in the multiple degrees of freedom system. It should be noted, however, that the calculation of the average histogram over all causal states is equivalent to averaging the recurrence rates over all accessible areas of the phase space. Therefore, the motion in the vicinity of several periodic islands visited moderately often by the trajectory may be masked by more frequent chaotic motions. Our analysis of the ensemble of causal states in terms of the D parameter thus provides an alternative approach that allows detecting the periodic islands in the chaotic sea by making a more subtle distinction between the periodic and chaotic phases of motion.

In Fig. 10, we plot the scatter diagrams representing the apparent clustering of the causal states into two classes with respect to the parameter D. The horizontal axis approximates the occurrence rate (or probability $P(\epsilon_i)$) of the causal states, that is 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 by the total length of the symbolic series.

It is also interesting to note that the histogram for "periodic" states possesses a clearly developed peak at the value of about 0.1 ps (see Fig. 9(b)), while those for the rest of the causal states are characterized just by a mere exponential function (Figs. 9(e) and 9(f)). The combination of the peak at short times with the slowly decaying tail at long times in the histogram evidences that the trajectory visits the vicinity of periodic islands quite rarely, but, once it is trapped by a

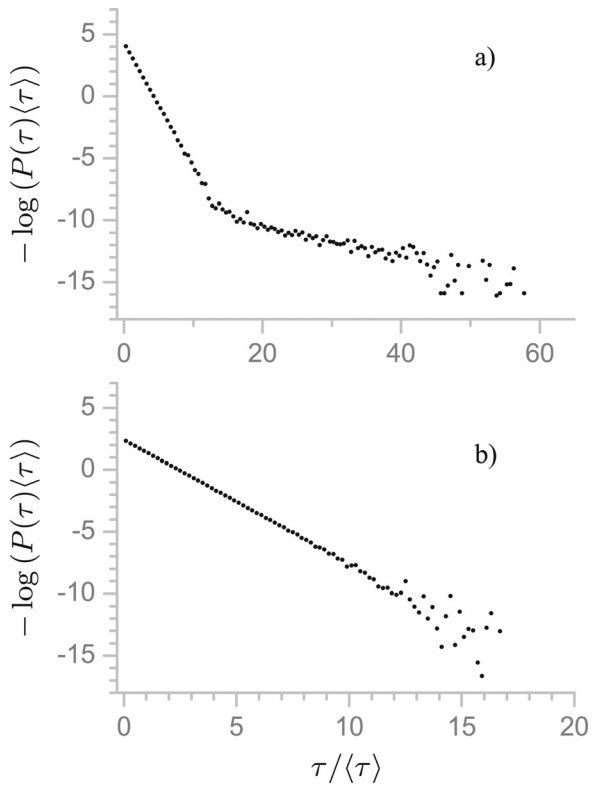

FIG. 8. Distribution of Poincaré return times averaged over all causal states. (a) Standard map and (b) water time series corresponding to the velocity of a hydrogen atom.

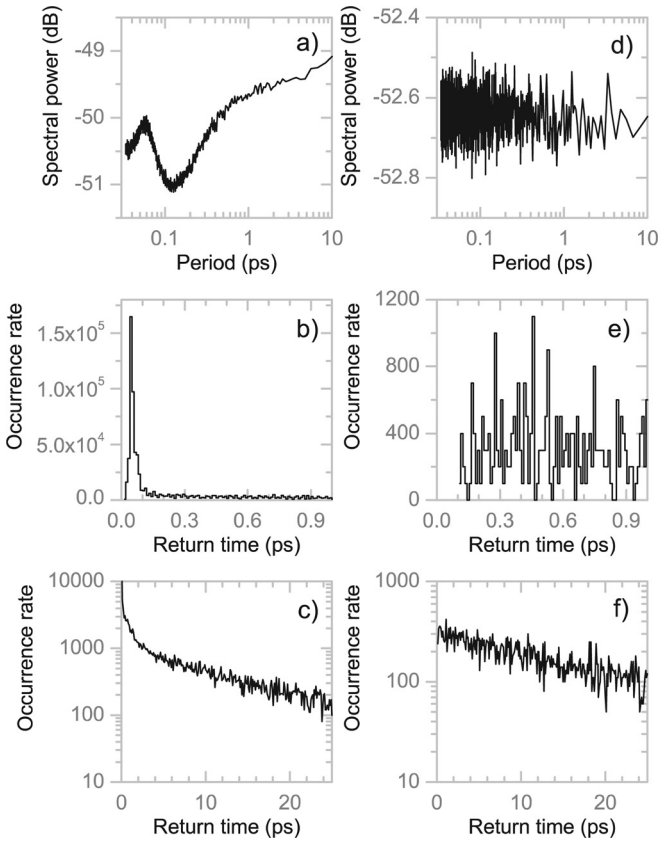

FIG. 9. Fourier analysis of water time series. Power spectra (a,d) and histograms of recurrence times (b,c,e,f) for typical causal states belonging to different types: a "periodic" state (a-c) and a "chaotic" state (d-f). The histograms on (c,f) are zoomed and smoothed fragments of those shown in (b,e). Spectra in (a,d) are the functions of inverse frequency.

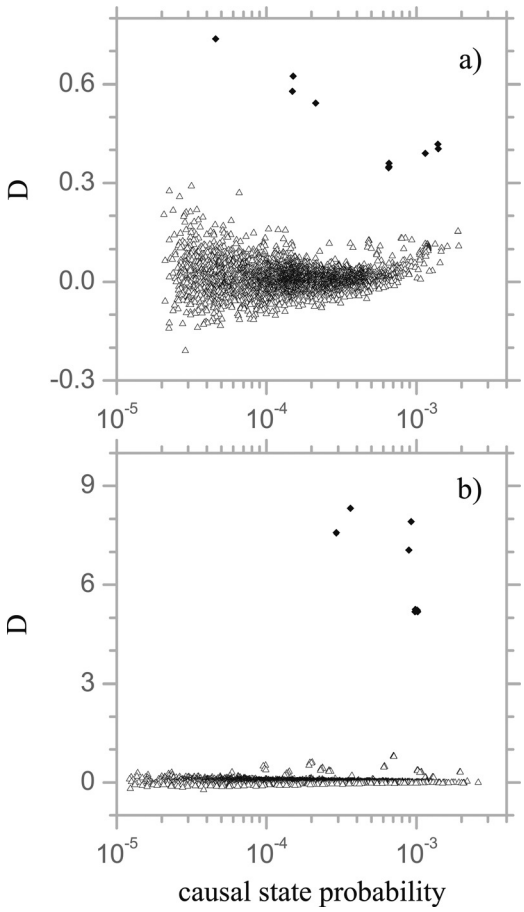

FIG. 10. Clustering of the causal states for the hydrogen atom velocity time series (a) and the Standard map (b) into "periodic" (diamonds) and "chaotic" (triangles) classes. Parameter D is plotted vs. occurrence rates of the causal states.

sticky boundary, it experiences a sequence of several short time returns. The causal states characterised by a low value of D have strictly exponential distribution of the return times and do not have pronounced low order periodicity.

The short-time returns of the high D-value causal states are caused by their quasi-periodic character, apparent in the symbolic sequences constituting these states. They are: '000000000', '222222222', '001001001', '010010010', '100100100', '122122122', '212212212', '221221221'. The repetition with the maximum period of 3 is evident in all of them.

Additional illustration of the splitting of the set of the causal states into two qualitatively different classes can be 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 binary time series corresponding to each of the causal states, we obtain an alternative indication of the difference between the "periodic" states and the rest of the set. "Periodic" states have a comparatively high level of spectral density above the characteristic period of ≈ 1 ps, as well as around ≈ 0.03 ps where the corresponding autocorrelation function reaches its first zero value. "Chaotic" states have "white noise" type of the power spectrum with approximately uniform spectral density function, Figs. 9(d)–9(f). This finding suggests that the processes with characteristic time scales of ≈ 0.03 ps corresponding to the first zero of the correlation function as well as ≈ 1 ps corresponding to the peak of the power spectrum are mainly defined by the "periodic" causal states.

Summarizing, the two classes of "periodic" and "chaotic" states are present in the analysed time series of both the velocity of the hydrogen atom and the Standard map. The "chaotic" states of the ϵ-machine represent long term mixing processes that describe the way the system explores the phase space. The number of "chaotic" states is high indicating a prevalence of the areas of chaotic motions (chaotic sea) over the periodic components (resonance islands), a rather typical picture previously reported only in low-dimensional nonlinear dynamical systems.[4]

V. CONCLUSIONS

We have analysed the application of computational mechanics to Hamiltonian dynamics of molecular systems. A conceptually important connection of the causal states of the ϵ-machine built on an initially symbolised trajectory to the areas of phase space that are optimal in the sense of predicting the trajectory's behaviour has been analysed. It has been shown that the areas in the phase space defined by the causal states possess special properties in the dynamical sense, that is their recurrence time distributions follow Poincaré law with two distinct exponents. This allows classifying

the causal states into quasi-periodic and chaotic types. Comparing our findings to a well studied case of the Standard map, one can conclude that our approach provides a new quantitative characteristic that allows to separate the motion in the phase space into two distinct classes.

Our result on the distribution of recurrence rates over the ensemble of causal states suggests that the phase space of the dynamical system corresponding to water has more complex structure than can be concluded from average statistical analysis of return times. The parameter D introduced as an indicator of deviation from Poincaré law thus provides a more subtle distinction between periodic and chaotic phases of motion, compared to the single histogram analysis presented in Fig. 8. Several causal states demonstrate much slower decay rate than can be expected from Poincaré law. This fact evidences the presence of the areas in the phase space where the trajectory spends longer time compared to the rest of the accessible volume. Such areas can not be detected easily by other methods, most probably due to abundance of resonant areas in the high-dimensional phase space that makes difficult a clear distinction between chaotic and quasi-periodic motions.

From a different perspective, our method also has a special importance for the problem of quantifying transport properties in high-dimensional molecular systems, since it reveals a (small) number of areas in the phase space playing crucial importance for particle motion through the phase space. Finding such areas from the analysis of a single scalar time series can be very useful in numerical experiments with large number of interacting particles that typically generate huge volumes of data. Extracting the essential information from the trajectory of a single test particle thus looks a promising approach, for example, in modeling the process of protein folding or dynamics of complex biomolecules.

ACKNOWLEDGMENTS

The work is supported by Unilever and the European Commission (EC Contract Number 012835-EMBIO).

APPENDIX: COMPUTATIONAL MECHANICS

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 $\sum_i P(\epsilon_i) = 1$. The collection of the causal states together with the transition probabilities define an ϵ-*machine*.

It is proven[22] that the ϵ-machine is

- a *minimal sufficient* statistic, therefore, the causal states can not be subdivided into smaller states; and
- a *unique minimal sufficient* statistic, any other one simply re-labels the same states.

[1]J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).
[2]R. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, and H. Scheingraber, Chaos, Solitons Fractals **4**, 133 (1994).
[3]J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[4]G. Zaslavsky, Phys. Rep. **371**, 461 (2001).
[5]V. Afraimovich and G. Zaslavsky, Phys. Rev. E **55**, 5418 (1997).
[6]D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding* (Cambridge University Press, New York, 1995), ISBN 0-521-55900-6.
[7]V. Afraimovich and G. Zaslavsky, Chaos **13**, 519 (2003).
[8]C. R. Shalizi and C. Moore, e-print arXiv:cond-mat/0303625.
[9]F. Takens, *Detecting Strange Attractors in Turbulence* (Springer, Berlin, Heidelberg, 1981), Vol. 898, pp. 366–381.
[10]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.
[11]D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 17011718 (2005).
[12]H. J. C. Berendsen, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis (Kluwer, Dordrecht, 1991), pp. 139–155.
[13]W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[14]M. Buhl and M. B. Kennel, Phys. Rev. E **71**, 046213 (2005).
[15]D. Nerukh, V. Ryabov, and R. C. Glen, Phys. Rev. E **77**, 036225 (2008).
[16]C. R. Shalizi and K. L. Shalizi, 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.
[17]D. Nerukh, V. Ryabov, and M. Taiji, Physica A **388**, 4719 (2009).
[18]V. Ryabov and D. Nerukh, J. Mol. Liq. **159**(1), 99 (2011).
[19]J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[20]B. V. Chirikov and D. L. Shepelyansky, Phys. Rev. Lett. **82**, 528 (1999).
[21]B. V. Chirikov, Phys. Rep. **52**, 264 (1979).
[22]C. Shalizi, K. Shalizi, and R. Haslinger, Phys. Rev. Lett. **93**, 118701 (2004).