Computational Mechanics of molecular systems

Dmitry Nerukh
Department of Chemistry, Cambridge University, Cambridge CB2 1EW, UK

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

A framework that connects Computational Mechanics and molecular dynamics has been developed and described. As the key part 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, 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 these areas of the phase space can be divided into two classes according to their Poincare return times. The first class is characterised by predominantly short time returns, typical to quasi-periodic trajectories of the dynamical system. This class includes a limited number of areas that are robust with respect to different total length of the molecular trajectory. The second class has a chaotic behaviour of the return times distributed exponentially in accordance with the Poincare theorem. In contrast to the first class, the number of such areas grows logarithmically with the length of the trajectory. We put forward and numerically illustrate a hypothesis that explains this behaviour by the presence of temporal non-stationarity in molecular trajectory.

I. INTRODUCTION

The dynamics 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 the coordinates and velocities of all the particles in the analysed volume. The problem of describing and predicting the appearance of such patterns is crucially important since they ultimately define the functionality of the systems and fundamental properties of such processes as, for example, protein folding. There is, however, a fundamental 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 if the motions occur in the phase space of dimension higher than $\approx$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 cases 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 systems with complex non-trivial behaviour.

Computational Mechanics (CM), a promising new concept aimed at building a statistical and at the same time dynamical description, has been recently proposed. 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.

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 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 in contrast to other, non-resonant areas where the trajectories move randomly filling the phase space almost uniformly.

It should be noted that there is another reason to observe areas in the phase space that are non-uniformly filled with trajectories. It comes from a necessity to analyse low-dimensional projections of the full-dimensional phase space. Since performing the analysis in the space of dimension of several thousands is infeasible in any re-

alistic computer experiment, the focus of the research shifts naturally to low-dimensional projections. Dense areas can appear in projections if the motion occupies a compact volume in the subspace embedded into the whole-dimensional phase space. The process of projecting a high-dimensional object to lower dimensions produces dense areas in the middle of the analysed volume and relatively sparse areas adjacent to its borders. Note, however, that this happens only if non-trivial geometrical structures do exist in the full-dimensional space when the motion occupies an area with well-defined borders. Any projection (i.e. linear function) of, say, a volume filled with Gaussian noise would bring another Gaussian noise, in other words projecting a randomly filled high-dimensional spherical object does not induce any areas of excessive concentration in the subspace. Therefore, we don't make any distinction between the non-uniformities caused by the projection and those that appear due to stickiness, regarding any deviation from uniformity in the projection space as manifestation of intrinsic structures present in the phase space.

In order to quantify the distinction between uniform and non-uniform filling of the phase space we utilize the Takens embedding procedure [1] combined with the approach of surrogate time series, a methodology widely used for detecting geometric structures in the reconstructed phase space of nonlinear dynamical systems. The surrogate time series technique (also known as bootstrapping statistic) compares a characteristic of the analysed time series (discriminating statistic) to the same value, but calculated for a set of computer simulated data (surrogates). The surrogate data are similar to the original time series in certain properties (autocorrelation function, power spectrum, probability distribution function) but differ in other characteristics of special interest. In our analysis we focus on the properties of correlation and uniform phase space filling, keeping other properties of the data intact. For the discriminating statistic we take the value of Statistical Complexity (SC), a measure introduced in CM [2] that quantifies clustering in the statistical sense considering the short time predictability of the phase space trajectories. We would like to stress that other statistics, like correlation dimension, Lyapunov exponents, or Kolmogorov-Sinai entropy traditionally used in the field of time series analysis are unable to discriminate between molecular dynamics and surrogate time series. For all the above mentioned characteristics the necessary amount of data grows exponentially with the dimensionality of the considered projection of the phase space making the reliable calculation infeasible in the subspaces of dimension higher then ≈10. On the other hand, using the complexity based measures [3, 4] and analysing the probabilistic properties of symbolic sequences corresponding to the phase space trajectories is a promising alternative for detecting structures in the phase space.

It is also interesting to note that the non-uniform covering of the phase space by the trajectories leads to anomalous transport properties of the trajectories in the phase space. This issue attracted a lot of attention recently [5] and it has been demonstrated that important insights into the details of the transport can be achieved in terms of the Poincare theorem of returns. Using this approach we show how the analysis of molecular trajectories by SC provides a link from a purely statistical description with Markov chain-type modelling to the dynamical systems theory based on Poincare recurrence analysis.

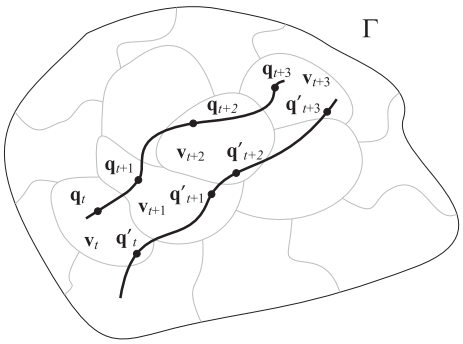

FIG. 1: Illustration of the degeneracy of a macro-observable projection of the full-dimensional phase space trajectory. The same sequence of the observable (the velocity) $\{\mathbf{v}_t\mathbf{v}_{t+1}\mathbf{v}_{t+2}\mathbf{v}_{t+3}\}$ is generated by two different pieces of the phase space trajectory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

II. MOLECULAR PHASE SPACE TRAJECTORY AS A COMPLEX DYNAMICAL SYSTEM

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

In order to analyse huge volumes of data corresponding to the high-dimensional trajectory, low-dimensional observables (macro-observables) have to be considered. For example, the velocity of an atom $\mathbf{v}$ can be taken for such an observable time series. So defined $\mathbf{v}$ is a projection of the full-dimensional trajectory onto a low-dimensional subspace. Because of the discrete nature of the time sampled trajectory and the finite tolerance of the measurements of $\mathbf{v}$ the analysed data cover a large, but finite set of possible values.

This leads to the situation that different realisations of the trajectory in the full-dimensional phase space may

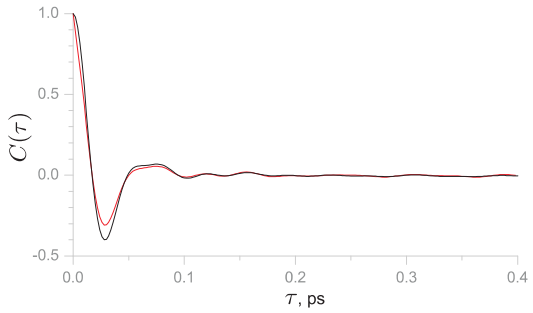

FIG. 2: Autocorrelation functions $C(\tau) \equiv \frac{1}{T}\sum_t^T \mathbf{v}_t \cdot \mathbf{v}_{t+\tau}$ for the original velocity of the hydrogen of bulk water (black) and the signal made of 27 symbols (red, see text for details)

produce the same experimental time series due to the degeneracy of the observable illustrated in Fig. 1. In other words, many different phase space points correspond to the same value of $\mathbf{v}$ obtained in the simulation. The whole phase space Γ is thus partitioned into the areas such that on each of them the macroscopic observable $\mathbf{v}$ takes a unique value while the full-dimensional points $\mathbf{q}_i$ can be different (Fig. 1).

III. THE PROBLEM OF SYMBOLISATION

The values of the observables that we analyse are discrete and take a finite number of values. This means that the trajectory can be interpreted as a sequence of values taken from a countable set of numbers or *symbols*. In the case of, for example, the typical computer representation of data as floating point numbers, the amount of equivalent symbolic value is large but limited and defined by the precision used in the simulation (single, double, etc). Note, however, that sometimes the high precision and hence large number of symbols in data representation is not an essential requirement for further analysis. In many cases even a very coarse representation of $\mathbf{v}$ produces statistical indicators characterising the dynamics of the molecular system very close to their true values. Fig. 2 shows an example of such a characteristic, the common velocity autocorrelation function calculated for a signal where the true (double precision) velocity coordinates are replaced (rounded) with only one ternary digit of v_x,v_y, and v_z, such that $\{x \equiv -1, \text{if } x < -1; x \equiv 0, \text{if } -1 \leq x < 1; x \equiv 1, \text{if } x \geq 1\}$, where x represents v_x,v_y, and v_z. The total number of possible values of the resulting coarse grained velocity vector is $3^3 = 27$, that is the signal is represented by only 27 symbols. Nevertheless, the autocorrelation function calculated from such a rough approximation of the original signal is very similar to that calculated from the double precision values of $\mathbf{v}$.

The 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]. We here show one more time that this appears to be a very useful framework that allows to make unexpected conclusions about the phase space trajectory of the molecular system. The common sense perception could be that such a few symbol representation is an oversimplified description of the trajectory that can provide only very approximate conclusions about the system. However, somewhat counter intuitive, this is not the case, especially when the *dynamics* is analysed, that is when the sequences of the symbols (words) are considered, rather than the symbols separately. Moreover, it can be proven that for a specially chosen partitioning of the phase space any symbolisation, even the most coarse grained one, the binary, contains the same information about the dynamics as the original signal, if infinitely long sequences of the data points are analysed (see bellow).

Thus, the "default" partitioning of the phase space provided by the finite precision of the computer representation of floating point numbers can be further coarse grained in order to reduce the number of symbols to just a few. An obvious question is: how this symbolisation should be performed and in what respect will the resulting sequence of symbols be different from the original continuous signal? There is a rigorous answer to this question. A natural choice for the symbolising partitioning is the so called generating partition (GP) [7] that has the property of a one-to-one correspondence between the continuous trajectory and the generated symbolic sequence. That is, for an infinitely long trajectory all information is retained after the symbolisation. It has been proven that such a partition exists for any dynamical system [8]. It is defined as follows.

Consider a dynamical system $\mathbf{x}_{i+1} = \mathbf{f}(\mathbf{x}_i), \mathbf{f} : M \to M$ and a finite collection of disjoint open sets $\{B_k\}_{k=1}^K$, partition elements, such that for their closures $M = \cup_{k=1}^K \bar{B}_k$. Given an initial condition $\mathbf{x}_0$, the trajectory $\{\mathbf{x}_i\}_{i=-n}^n$ defines a sequence of visited partition elements $\{B_{\mathbf{x}_i}\}_{i=-n}^n$ or $\{s_i\}_{i=-n}^n$, where s_i are symbols from the alphabet that mark the elements where $\mathbf{x}_i \in B_i$. For a generating partition the intersection of all images and pre-images of these elements is, in the limit $n \to \infty$, a single point: $\cap_{i=-n}^n \mathbf{f}^{(-i)}(B_{\mathbf{x}_i})$.

This elegant mathematical construct has a very important disadvantage when applied to realistic signals. An algorithm for calculating a GP in a general case is unknown. Recently methods for finding approximations for GP are reported. The method from [9] is shown to reproduce GP for several known low-dimensional systems and could be designed to treat multi-dimensional data. Its applications for a high-dimensional systems remains, however, unknown.

IV. HOW COMPUTATIONAL MECHANICS CAN BE USED TO FIND A SUITABLE SYMBOLISATION

Thus, the best possible partition, the GP, is very difficult, if possible at all, to find for a general dynamical signal. Other criteria can be used to find an approximation of GP or other partition suitable for symbolisation. For example, a partition can be constructed using a statistical argument that the properly designed symbolic sequence maximises the Shannon entropy, which is indeed an approach often used in symbolic dynamics. In this section, we study a different approach [10] that allows to utilise CM for finding a suitable partition in both the dynamical and statistical senses. Specifically, CM allows to find a partition (ϵ-machine) that is optimal from two statistical viewpoints: (i) it transmits as much information as possible from the continuous signal to the symbolic sequence and (ii) it is optimal for the purpose of statistical prediction of the signal given all the the information contained in the past. In the following we analyse three stages of coarse graining used for symbolising the trajectory and constructing the ϵ-machine that corresponds to a specific partition of the phase space of the system.

A. The dynamics makes the partition finer

Suppose an arbitrary initial partition of the phase space is chosen and consider one of its elements. Dynamical trajectories starting from every point of this area move to new locations at the next time step. The set of these new points represents the dynamical image of the partition element generated by the system's equations of motion. Similar transformation happens to all the partition elements thus changing the initial partition into a new one.

More specifically, the evolution of the phase space points $\mathbf{q}$, sampled at times t, is governed by an operator $\mathbf{T}$: $\mathbf{q}_{t+1} = \mathbf{T}\mathbf{q}_t$. $\mathbf{T}$ is a dynamical operator that advances the phase space points in time according to Hamiltonian equations of motion. The same transformation can be considered in the statistical sense, as a sequence of random points. Because of the determinism implied by the Hamilton equations of motion the dynamical points $\{\mathbf{q}_t\}$ form a Markov chain, that is the value of $\mathbf{q}$ at the next time step is completely defined by the current value of $\mathbf{q}$. Considering an ensemble of such dynamical systems (that is a collection of all possible realisations of the trajectories of the system), 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 $\mathbf{T}$. The probability distribution of $\mathbf{Q}$ is also known as the invariant measure [11].

Consider now not an arbitrary partitioning of the phase space, but the one defined by the choice of the low-dimensional observable taken for the analysis (projection) and the finite precision of it's measurement, as discussed in the previous section. The macroscopically observed variable A is a function f of the microstate $\mathbf{Q}$. For example, this could be the instantaneous temperature $\frac{1}{Nk}\sum_i m_i \mathbf{v}_i^2$, where N is the number of atoms, 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 (a particular phase space point corresponds to only one value of the macro-observable) and jointly exhaustive (the union of all the partition elements gives the whole phase space) sets, on each of which f takes a unique value. Denote the partition of Γ induced by f as $\mathcal{F}$. The observed process is equivalent to a phase space trajectory, i.e. $A_t = f(\mathbf{q}_t)$. Because of the degeneracy of the macro-observable described in the previous section, it is not necessarily Markovian. Fig. 3 gives an illustration of how the degeneracy of A can lead to a situation when both current and previous values of A are needed to predict the value of A at the next time step.

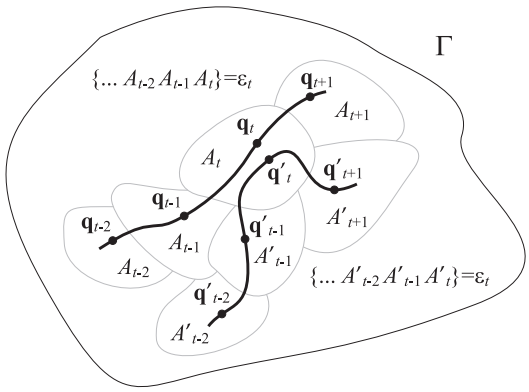

FIG. 3: Schematic illustration of the sequences used to define formula (2). Phase space points $\{\mathbf{q}_i\}$ and $\{\mathbf{q}'_i\}$ of two pieces of the trajectory form Markov sequences. The corresponding observation sequences $\{A_i\}$ and $\{A'_i\}$ are not Markovian since the same value A_t leads to different A_{t+1} and A'_{t+1} depending on the previous values A_{t-1} and A'_{t-1}. However, if both histories $\{\dots A_{t-2}A_{t-1}A_t\}$ and $\{\dots A'_{t-2}A'_{t-1}A_t\}$ 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

Now, what happens to the partition $\mathcal{F}$ when we consider the sequences of A_t instead of the individual values of A? Take an observation at time t, A_t, and its partition element $\mathcal{F}_t$ of Γ. For a sequence of two consecutive observations at the current and previous time moments the corresponding area of the phase space is

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

which is a refinement of the partition $\mathcal{F}$. This new area corresponds to the situation when we consider the value of the macro- observable A_t at the current moment *and* the value A_{t-1} at the previous moment. In other words, the area now corresponds to a *sequence* of two observations, rather than each observation independently. Note

also that this new area is not larger than each of the two areas (partition elements) representing single symbols because of the intersection operator. When considering the partition as a whole this means that the partition typically becomes finer when one considers longer symbolic sequences. This procedure can be repeated any countable number of times thus providing the refined partitions for the symbolic words (histories) of the macro-observable A. Thus, the sequences generated by the system's dynamics make the initial partition induced by the macro-observable finer, the longer the sequence $\{A_t\}$ (the "history") the finer the partition induced by the sequence.

If the initial partition $\mathcal{F}$ is a generating partition, the finest possible partition corresponds to the infinitely long trajectory, and it is equivalent to a set of original data points covering the accessible phase space.

It is appropriate to put here an example that illustrates the partition refinement process by considering longer and longer histories. For this purpose we use a well-documented two-dimensional map called the Standard map (or sometimes the Taylor-Greene-Chirikov map) [12]. It is defined as a transformation of the plane to itself

$$\begin{aligned} P_{n+1} &= P_n + K \sin\theta_n \\ \theta_{n+1} &= \theta_n + P_n + K \sin\theta_n \end{aligned}$$

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. An example of a chaotic trajectory in this system at the value of $K = 6.908745$ is given in Fig. 4, where one can see a large chaotic area with only two large stability islands symmetrically located with respect to the origin. Infinite number of smaller islands exist in the vicinity of the large ones, but they are not visible at the given picture resolution. Finding the GP even for such a simply looking system as the Standard map is not a trivial task. Therefore, we applied a simple uniform partitioning in the variable θ. In other words, considering one time step histories results in the partition shown in Fig. 5a. Fig. 5$b-d$ demonstrate the effect of partition refinement by considering the histories of increasing length. Here different colours correspond to different symbolic histories and borders separating the areas with different colours are the partition elements delimiters. It is evident that as the histories become longer the partition becomes finer and the number of partition elements grows exponentially with the length of the symbolic words.

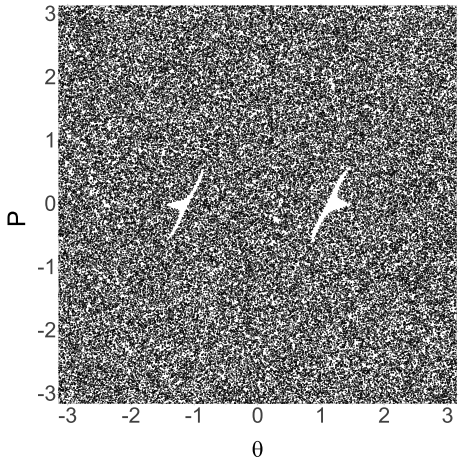

FIG. 4: The Standard map trajectory in chaotic regime

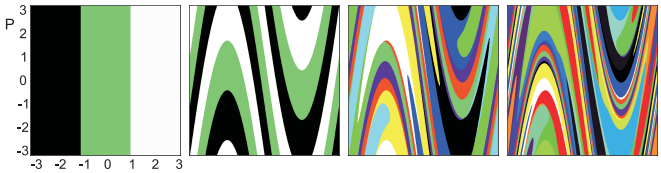

FIG. 5: The refinement of a partition caused by considering longer symbolic sequences for the Standard map. a) the original three symbol partitioning corresponding to the one time step history; b0 - d) two to four symbol histories induced partitioning

B. Computational Mechanics coarsens the partition

So far we have considered two different stages of defining the phase space partition used in our numerical experiments. The first one is produced by the choice of macro-observable and the finite precision of the simulation procedure. This partition can be further coarse grained for simplifying the analysis, down to a few symbols representing the dynamical trajectory. The second stage of building the partition is the refinement of the one obtained at the first stage by considering the symbolic sequences (words) instead of the individual symbols.

The next step of the analysis is to apply a statistical description called Computational Mechanics (CM) [2] to the observed sequences of A. The rigorous definition of CM can be found in the original works by Crutchfield and co-authors. Here we briefly reproduce the principal steps of the approach relevant for the present study.

All past A_i^- and future A_i^+ halves of bi-infinite sequences of the macro-observable centred 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, $H[P(\nu)] \equiv -\sum_\nu P(\nu)\log_2 P(\nu)$.

Thus, the essence of CM consists of grouping the histories $\{A_t^-\}$ into causal states. In terms of the partitions of the phase space this process corresponds to joining the partition elements (areas) of Γ induced by the histories with the same (in statistical sense) future (see the previous section). Importantly, the new partition elements allow to construct a Markovian process 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 $\{\epsilon_t\}$ constitutes a Markov chain (Fig. 3).

C. The partition generated by Computational Mechanics is the most informative one

The causal states $\{\epsilon_i\}$ are now used for the analysis of dynamics instead of the symbolic sequences (histories) defining the phase space partition elements. Shalizi and Moore [10] show that in this setting the Statistical Complexity (the characteristic of the causal states $\{\epsilon_i\}$) has a clear physical meaning: it quantifies the amount of information contained in the new constructed macro-observable process $\{\epsilon_i\}$ about the microstate $\mathbf{Q}$:

$$C_\mu = I[\mathbf{Q};\epsilon], \tag{2}$$

where I is the mutual information between 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: $H[X|Y] = -\sum P(X)\sum P(X|Y)\log_2 P(X|Y)$.

Eq. 2 is obtained using the fact that the knowledge of the microstate specifies the macro-observable precisely: $H[\epsilon|\mathbf{Q}] = 0$. All histories contained in ϵ_t and the corresponding partition of $\mathbf{Q}$ uniquely define the next state ϵ_{t+1} (the ϵ-machine definition). Using this property and the equality $H[X] + H[Y|X] = H[Y] + H[X|Y]$ equation (2) follows:

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

Due to the properties of the ϵ-machine this partition contains the maximal information that is possible to extract from the chosen macro-observable and the specified initial partition of the phase space Γ.

D. Three stages of symbolisation

Summarising, the phase space partition we use in numerical experiments is obtained in three stages. They are schematically illustrated in Fig. 6.

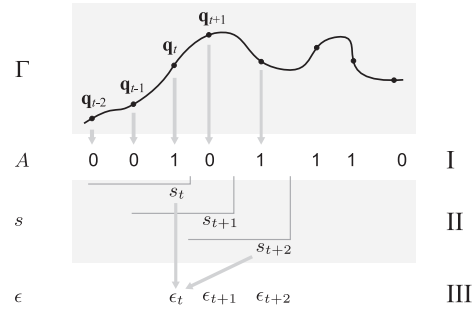

FIG. 6: Three stages of symbolisation: I - converting the continuous phase space points $\mathbf{q}_i$ to symbols using the partition induced by the macro-observable; II - forming the symbolic histories from the symbols $s_t \equiv \{\ldots A_{t-2}A_{t-1}A_t\}$; III - joining the histories into the causal states $\epsilon_t \equiv \{s_i \ldots s_j\}$

I. 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.

II. The partition elements of this partition are refined by the dynamics, when we consider words (histories) instead of single symbols (1). 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 [1]. 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.

III. 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).

Molecular signals are chaotic. In statistical terms this means that all two-point correlations quickly (exponentially) go to zero with time. Let us fix a time moment t_i. Analysing the observable forward in time we can reach a moment t_{m^+} after which the correlations vanish and we can regard the signal as essentially random. The same behaviour of correlations can be expected when considering the signal backwards in time with the boundary of randomness t_{m^-}. Thus, the times t_{m^-} and t_{m^+} define the interval beyond which the sequence $\{\ldots A_{i-m^-}\ldots A_{i-1}A_iA_{i+1}\ldots A_{i+m^+}\ldots\}$ is undistinguishable from a random noise by the correlation analysis. This implies that beyond these times all the sequences are statistically similar. In other words, there is a subsequence of a minimal length $t_m = t_{m^+} - t_{m^-}$ that differs from random noise and, thus, reflects non-random correlation properties in the system.

An important advantage of using the CM formalism for analysing symbolic sequences the possibility of finding the interval of substantial multi-point correlations by

considering the histories of different length. A history of length l induces a partition $\mathcal{F}^l$, and longer histories induce finer partitions. However, if the "tails" of the symbolic histories corresponding to finer partitions are statistically indistinguishable from the shorter histories because of the absence of correlations, the causal states group the corresponding partitions $\mathcal{F}^l$ back into *the same* partition $\mathcal{F}^{min}$ that contains all statistical information and, therefore, is optimal from the future prediction point of view. This reflects the fact that the ϵ-machine contains the complete information that the observable possesses about the microstate (2).

V. 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 [13] molecules was simulated using the `GROMACS` molecular dynamics [14] package. The temperature of the systems was kept constant at 300K using Berendsen [15] or Nose-Hoover [16] 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.

B. Symbolisation

As described above, 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. 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 [9]. The example of calculations with this method for the two-dimensional cross-section of our (tree-dimensional) velocity data using 2, 3, 4, and 5 partition elements are shown in Fig. 7. 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 $(x, y) \to (\rho, \varphi)$ and estimating the probability density $w(\varphi)$ for the random variable φ. The histogram corresponding to such $w(\varphi)$ distribution is given in Fig. 8. Almost perfect uniformity of the distribution function is obvious, thus justifying the choice of centrally symmetric partitions that we used in all subsequent calculations.

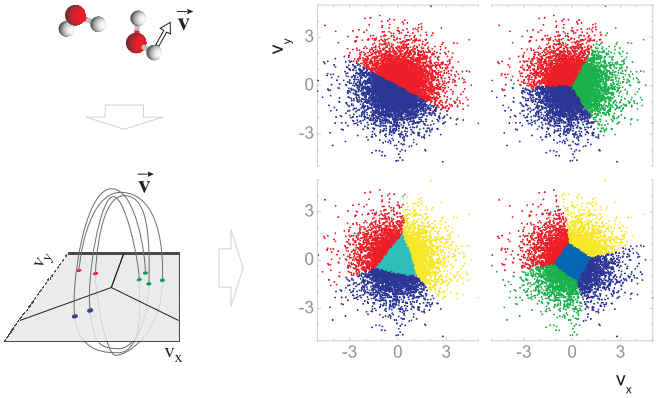

FIG. 7: The process of converting the continuous atomic velocity signal $\mathbf{v}$ into symbolic sequence. On the right the symbolisation with 2, 3, 4, and 5 symbols are shown

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 [17]. 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

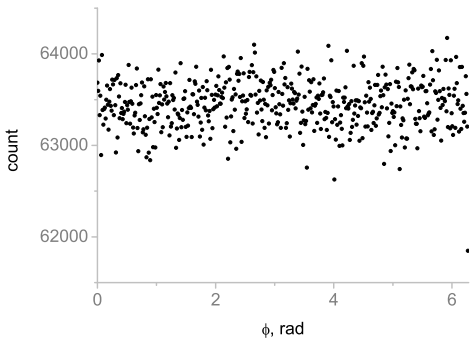

FIG. 8: The histogram of the random variable φ illustrating the uniform symmetric distribution of points in the $v_z = 0$ cross-section plane of the hydrogen velocity trajectory

$P(\overrightarrow{s}|\,\overline{s}_i)$ used for grouping the histories into causal states by their predictive properies (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 $P(\overrightarrow{s}|\,\overline{s}_i)$ 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 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, Fig. 9.

D. Surrogate time series

As we already mentioned in the previous chapter the computation procedure implementing the idea of building an ϵ-machine and estimating the SC-value contains several control parameters that require fine tuning. The careful selection of the parameters is necessary for the purpose of good algorithm convergence, as well as independence of the found SC-value on the details of the computation process. In addition to the parameters σ and l that control, respectively, the robustness of partitioning the symbolic phase space to the causal states and the symbolic space dimension (in the sense of Takens [1]), there are also other parameters that should be taken into account. Such parameters include, for example, the size of the alphabet used for the symbolisation purpose, or the parameters specifying the position and orientation of

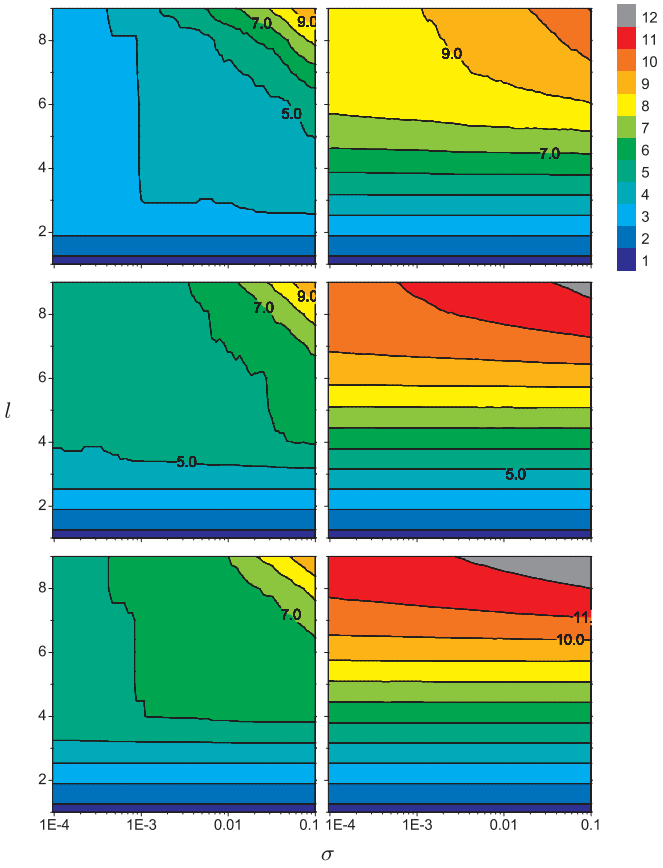

FIG. 9: The number of causal states in the ϵ-machines as a function of the history length l, the tolerance σ, and the duration of the time series; left: the surrogate time series, right: the molecular signal; top: time series duration of 60ns, middle: 450ns, bottom: 1μs

the cross-section plane utilized for discretising (decimating) the initially continuous time series (calculated at the time increments equal to the numerical integration step).

As a consequence, there is a number of potential sources for random deviations and biases that appear in the calculated value of SC and, hence, may produce a spurious indication of the presence of clustering. A straightforward way of avoiding wrong conclusions from biased estimates is a careful error analysis based on the calculation of corresponding distribution functions for the estimated values. This approach, however, encounters serious technical difficulties due to the complexity of the calculation procedure and multiple possible choices for the control parameters. We, therefore, accepted a different, much simpler way of obtaining error estimates, widely used in the literature devoted to the analysis of time series. In the works devoted to statistical data analysis the method is known as "bootstrap" technique [18], whereas in the papers discussing nonlinear dynamics based analysis [19] it is called "surrogate data" method. Throughout this paper we employ the latter term as the name for the artificial time series used for obtaining the error estimates.

The idea of the surrogate data approach is briefly described in the Introduction and it consists of testing the molecular time series against a hypothesis that it is produced by a liner stochastic process like, for example, a white noise passed through a band-pass filter (coloured noise) or an autoregressive moving average process (ARMA), etc. [20].

In practice the algorithm implementing the idea of surrogate data includes several steps. First, we choose the SC as the discriminating statistics, i.e. the measure used for detecting the clustering in the phase space. Second, the surrogate data series are produced by using a random number generator that converts the initially random sequence of numbers to a time series with required properties. The surrogates, thus, preserve some well controlled statistical characteristics of the analysed data but lack the property of interest. In the context of the analysis of deterministic dynamics the surrogates are often chosen to have the power spectrum identical to the original data series, but, due to the way the artificial data set is constructed, do not possess any property imposed by deterministic dynamics, such as, for example, the finite value of correlation dimension [21] or others [22]. Since the surrogate time series is characterized by an infinite value of the correlation dimension it fills the phase space almost uniformly. At the final stage the discriminating statistics for the original data is compared to a set of corresponding values calculated from the surrogate time series. The detection of significant discrepancy can be interpreted as an indication of essential differences between the surrogates and the original time series in the analysed property of the phase space filling.

In this work we use two types of time series for surrogates:

- Phase shuffled surrogate. This is a standard method of building the surrogate where time series is obtained via the phase shuffling algorithm [19]. The surrogate data generated 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 calculated by estimating the Fourier spectrum of the original data and assigning random values to all the phases of the Fourier components. After calculating the inverse Fourier transform the artificial data series (surrogate) has the unchanged amplitude spectrum but it becomes random, i.e. belonging to the class of Gaussian linear stochastic processes.

- White noise passed through a low pass filter. Contrary to the phase shuffled surrogate, which preserves exactly the shape of the power spectrum and autocorrelation function of the original data series, this surrogate is simply a coloured noise at the output of a filter with a flat frequency response function.

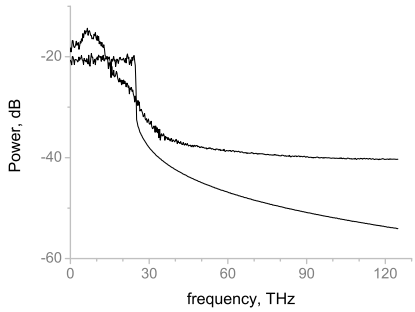

FIG. 10: Power spectra of the two surrogates: phase-shuffling algorithm (same shape of the spectrum as that of the original time series for x-component of the Hydrogen velocity) and coloured noise of approximately same effective bandwidth.

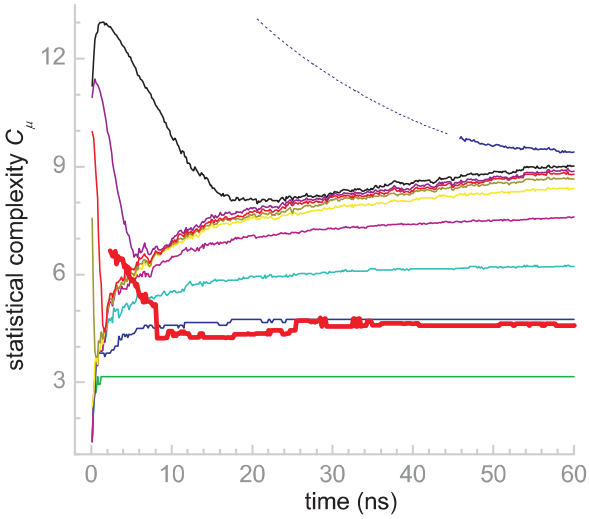

FIG. 11: Statistical complexity against time for the hydrogen velocity signal and the surrogate. The curves, from bottom to top, correspond to the values of the history length l from 2 to 11. The $l = 11$ curve does not settle on the logarithmic part within the shown area but seems to follow the same 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

The power spectra for the two surrogates generated with the algorithms described above are shown in Fig. 10.

VI. RESULTS

A. ϵ-machine grows with the length of time series

The calculated values of C_μ against the trajectory length are shown in Fig. 11 for different history lengths l. The heavy red curve corresponds to the phase shuffled surrogate time series.

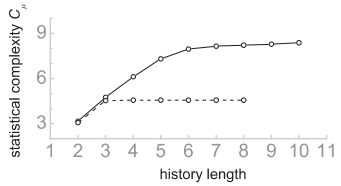

FIG. 12: Statistical complexity vs. the length of histories (dimension of the phase space) for the molecular data (solid line) and the surrogate (dashed line), data length: 30 ns

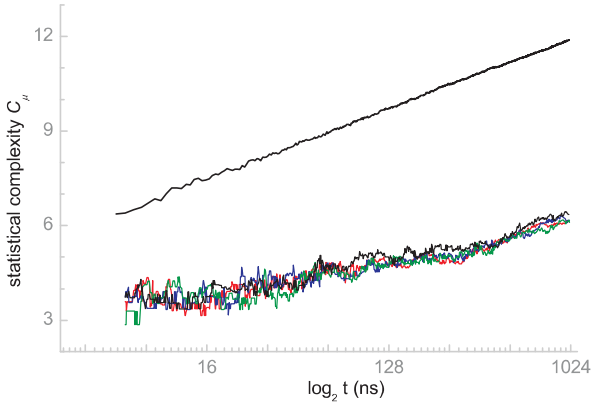

FIG. 13: Statistical complexity vs. the (log of) length of the analysis interval for the hydrogen velocity time series (top curve) and four surrogate time series (bottom curves): three independent realisations of the phase-shuffling algorithm (red, green, and blue), and single time series of a white noise passed through a low-pass linear filter (black). The frequency response functions of the equivalent filters used in generating the surrogates are shown in Fig.10

The dependence of C_μ on time does not converge, but first goes through a maximum and then settles on the $\log_2 t$-like curve. The maximum at the small times 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 at these times 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 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. 12, (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 noted that the time intervals discussed here are very long compared to the correlation time (Fig. 2) or any other time period where a non-trivial (i.e. non-Brownian) statistics can be expected to exist.

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:

$$C_\mu = a + h_Q \log_2 t \tag{3}$$

We put forward a conjecture that the coefficient h_Q can be used as a robust and universal characteristic of of molecular trajectories in terms of the Statistical Complexity, 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). Moreover, we attribute the large values of Statistical Complexity at long times to the deterministically chaotic nature of the molecular trajectories. In order to justify this hypothesis we compare the plots of C_μ vs time for the molecular trajectory and the phase-shuffled surrogates that lack the features of determinism characteristic to chaotic motion.

In Fig. 13 we plot the curve of C_μ for a much longer time interval of 1000 ns. It is evident that the equation (3) holds even for very long periods of time and the ϵ-machine continues to grow in the whole interval of the analysis. It is also clear from Fig. 13 that the C_μ curves for the surrogate time series display some growth at large time values but the differences between the curves corresponding to the molecular dynamics and the surrogate time series are evident. At any given time moment both the values of Statistical Complexity and its growth rate are much higher for the molecular trajectory compared to those of the surrogate time series. The origin of the growth phenomenon in the values of C_μ for the surrogate time series, although not yet fully understood, most probably lies in the deterministic properties of the algorithms used for the generating surrogates. It has been recently reported [23] that noise driven linear systems may produce trajectory patterns in the phase space that resemble those of chaotic systems. Phase shuffled surrogate time series is, in fact, white noise passed through a linear filter with a frequency reponse, identical to the Fourier spectrum of the molecular trajectory. Since any filter is a linear system, the surrogate obtained as the output of such a filter may possess certain features characteristic to deterministic dynamics.

The statistical complexity as a function of σ and l are shown in Fig. 9. We have found that the CSSR algorithm converges producing stable ϵ-machine architectures for all cases studied (six examples are represented in Fig. 9). The optimal values of the causal states and the corresponding history lengths are those that produce large areas of the same colour on the plots since the larger ones correspond to more stable size of ϵ-machine.

B. Analysis of the causal states

The presence of structures in the phase space of dynamical systems can be interpreted as the existence of nonuniformities in the invariant measure [11]. 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. In Hamiltonian systems the abundant resonances between natural oscillations of various frequencies create so-called islands of stability in the phase space that are known, on the one hand, as strong sources of non-uniformity in the invariant measure and, on the other hand, lead to breaking the ergodicity due to the formation of impermeable and "sticky" barriers in their vicinity [5]. The islands typically have a fractal structure and the finer is the scale of sub-islands the more "sticky" are their borders for trajectories, i.e. 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 Poincare recurrence theory [24]. 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

$$P(\tau) = lim_{\Delta\Gamma\to 0} P(\Delta\Gamma, \tau)/\Delta\Gamma \tag{4}$$

For a typical chaotic trajectory the following asymptotic relation holds

$$P(\tau) = \frac{1}{\langle\tau\rangle} exp(-\tau/\langle\tau\rangle), \tag{5}$$

where $\langle\tau\rangle$ is the average recurrence time over the distribution $P(\tau)$. Eq. (5) 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 under the assumption of ergodicity the sizes and shapes of the partition elements do not matter and, therefore, they can be made arbitrary depending on the details and convenience of the analysis of a particular problem.

As it has been stated in Section IV A 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 Statistical Complexity through the distribution function of their occurrence rates $P(\epsilon_i)$. To get a further insight into the link between the Statistical Complexity 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 plotted the histograms of the recurrence times (periods) for each of them. This analysis reveals that the causal states demonstrate a clear separation into two classes that we will refer to as "periodic" states (those defined by short time recurrences) and "chaotic" states (those without a well defined characteristic time scale of recurrence). "Periodic" states are characterised by a clearly developed peak at the value of about 0.1 ps (see Fig. 14b), while the rest of the causal states are characterized by an exponential distribution of the recurrence times (Fig. 14e,f).

In order to quantify the difference between the two classes we introduce a dimensionless parameter G that quantifies the presence of the peak in the interval of the recurrences ≤ 1 ps (compared to the interval 1 ps $\leq t \leq$ 2 ps)

$$G = \frac{max(h_1) - m_{12}}{\sigma_{12}}, \tag{6}$$

where h_1 is the values of the recurrence time histogram in the time interval $t \leq 1$ ps, m_{12}, σ_{12} are the median and the standard deviation values for the histogram in the interval 1 ps $\leq t \leq 2$ 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 quasi-periodic nature of the corresponding causal state. The causal states characterised by a low value of G have exponential distribution of the return times and do not have pronounced low-order periodicity. In Fig. 15 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 illustration of splitting 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 in the vicinity of the characteristic period of ≈ 1 ps, as well as around ≈ 0.03 ps where the corresponding autocorrelation function reaches its firs zero value. "Chaotic" states have "white noise" type of the power spectrum with approximately uniform spectral density function, Fig. 14a-c. This finding suggests that the processes with characteristic time scales of $\approx$ 0.03 ps corresponding to the first zero of the correlation function as well as ≈ 1 ps corresponding to the peak of

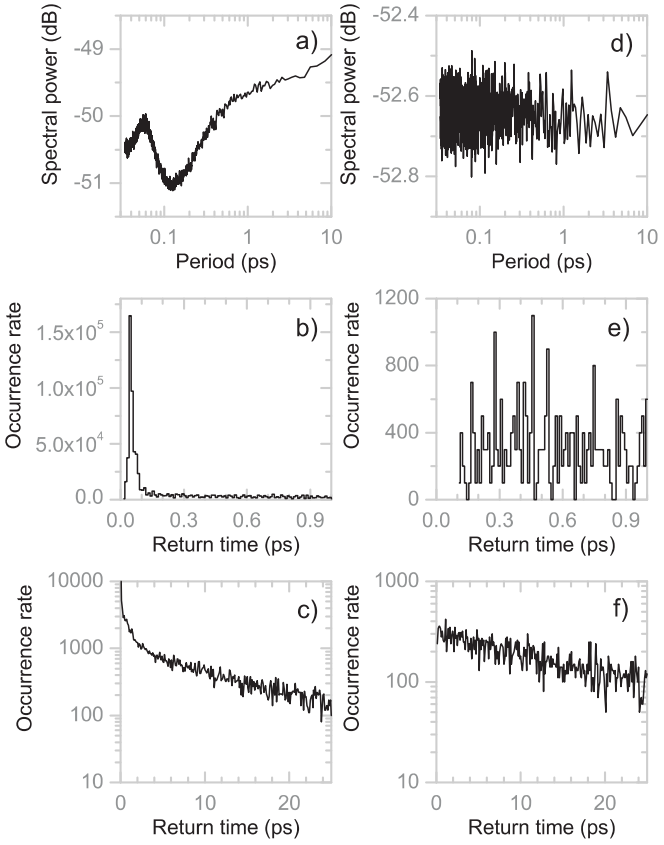

FIG. 14: Power spectra (a,d) and histograms of recurrence times (b,c,e,f) for typical causal states belonging to different types: a "periodic" state (a-c) and a "chaotic" state (d-f). The histograms on (c,f) are zoomed and smoothed fragments of those shown in (b,e). Spectra in (a,d) are the functions of inverse frequency

the power spectrum are mainly defined by the "periodic" causal states.

Summarizing, the "periodic" states are always present in the analysed velocity time series of the hydrogen atom, whatever is the length of the time series or the location of the analysis time window on the time axis. The presence of both the "periodic" and "chaotic" states is important for the formation of the conditional probability distribution functions defining the causal states. The "chaotic" states of the ϵ-machine represent non-trivial, non-linear, long-term processes that describe the way the system explores the phase space. For a molecular trajectory 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 in low-dimensional nonlinear dynamical systems [5].

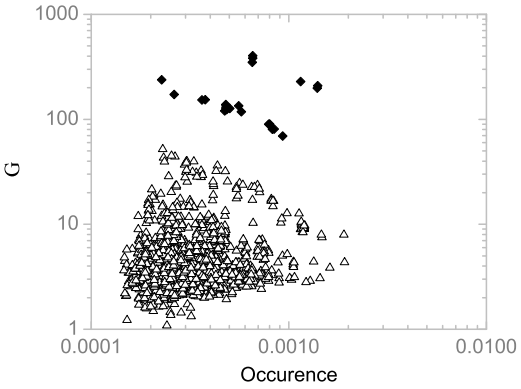

FIG. 15: Clustering of the causal states for the hydrogen atom velocity time series into "periodic" (diamonds) and "chaotic" (triangles) classes. Parameter G is plotted vs. occurrence rates of causal states

C. Non-stationary model of growing ϵ-machine

In order to get further insight into the mechanism providing the growth of the ϵ-machine with the length of the time series we consider in some detail the process of grouping the histories into the causal states performed by the CSSR algorithm. The SC is defined as the Shannon entropy of the distribution of the causal states probabilities, hence, the observed increase in its value should be related to the changes in the number and probabilities of the corresponding groups of histories (causal states). Our numerical experiments indicate that the main factor responsible for the growth is the process of splitting the causal states, i.e. regrouping the histories within causal states into pairs of sub-groups. The splitting occurs from time to time with different causal states causing the overall growth in their number and rearranging the distribution of associated probabilities. As a result, the ϵ-machine grows, as well as its complexity measure, the Statistical Complexity.

The key step in grouping the histories into the causal states consists of estimating the distributions of conditional probabilities for each history. The estimates are made by the analysis of the occurrence rates for the symbol following a given history: $P(\mathbf{v}_{i+1}|s_i)$, where $s_i \equiv \{\mathbf{v}_{i-l+1}\dots\mathbf{v}_{i-1}\mathbf{v}_i\}$. Since we have chosen the 3-symbol alphabet the distribution of the conditional probabilities can be illustrated with two-dimensional scatter plots of $P(0|s_i)$ vs. $P(1|s_i)$ (the probability of the third symbol is defined by the first two). Such two-dimensional diagrams for the water and surrogate time series are shown in Fig. 16. It is clear from Fig. 16 that: (i) the distributions are significantly different for the two cases and (ii) the former converges extremely slow compared to the latter, even at the scale as long as hundreds of nanoseconds. This slow convergence causes perpetual regrouping of the histories belonging to different causal states, most often resulting in numerous splitting of the

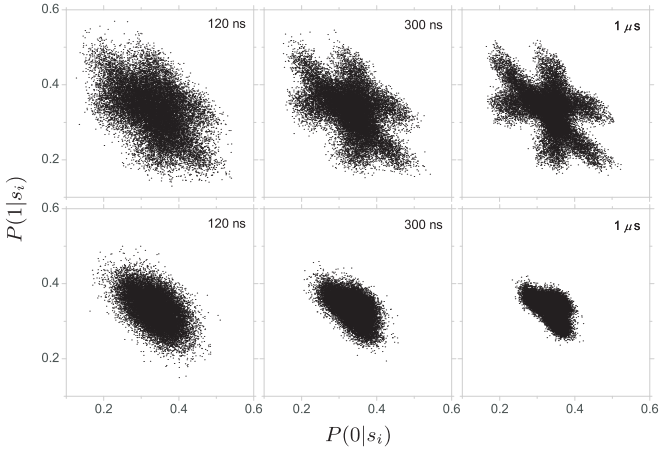

FIG. 16: Conditional probabilities $P(0|s_i)$ versus $P(1|s_i)$ for the analysed signal, where s_i are all sequences of 9 symbols from the three symbol alphabet {012}; upper row: molecular signal, lower row: the surrogate; the time shown on the panels is the length of the trajectory used to calculate the plot

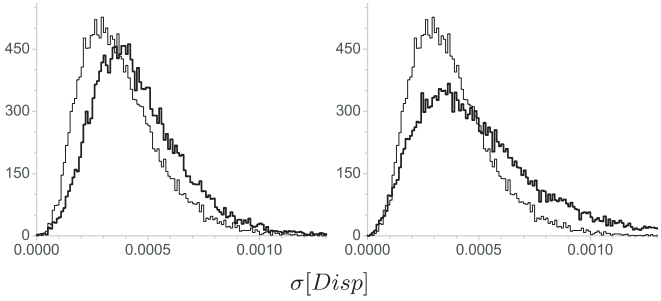

FIG. 17: The histograms of the standard deviations σ of the dependence $Disp(N)$ in the interval $N \in [30 \times 10^6; 31 \times 10^6]$, see text; histograms were calculated for all 9-symbol words occurring in the symbolic sequences of **left**: a molecular trajectory (thick line) and corresponding surrogate (thin line); **right**: the surrogate obtained from the molecular trajectory (thin line, same as that at the left) and an artificial symbolic sequence with non-stationary conditional probabilities (thick line), see text; periodic non-stationarity with the period of 5000 symbols was used

causal states to two or more sub-groups.

In order to analyse the convergence process in conditional probabilities $P(\mathbf{v}_{i+1}|s_i)$ we studied their dependence on the length N of the symbolic string. For this we introduced a parameter $Disp$ as the deviation of points around the straight line approximating the dependence of $P(\mathbf{v}_{i+1}|s_i)$ on N at large values of N (the last 3% of the total simulation interval from 0 to N). The standard deviations of $Disp$ were plotted as histograms for the molecular signal and the surrogate, Fig. 17, left. The curve corresponding to the molecular signal demonstrates pronounced fluctuations shifted to larger values of the variance that implies poorer convergence of the probabilities $P(\mathbf{v}_{i+1}|s_i)$ for the molecular signal.

To provide an explanation for the appearance of slow

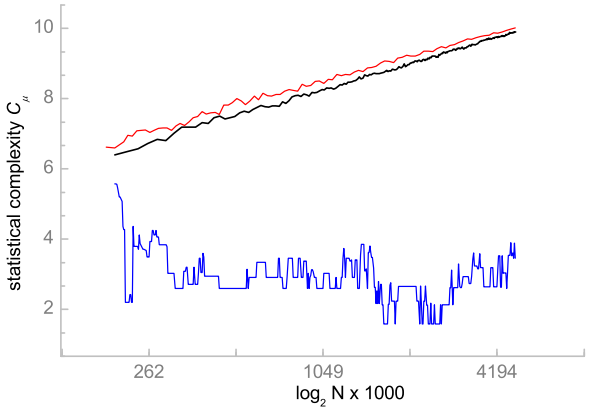

FIG. 18: The dependence of C_μ vs. N for original time series (black), and two simulated time series of stationary (blue) and non-stationary (red) Markov chain models

convergence in the conditional probabilities of symbolic sub-sequences we have designed a simple model that has trivial statistical Markov chain type properties, while demonstrating significant deviations of C_μ from zero and the growth of the C_μ value with N. The model is a ternary random sequence with the following properties. The probabilities of any of the three symbols in the alphabet {0,1,2} are equal to each other $P(0) = P(1) = P(2) = 1/3$, as well as the probabilities for any sequences of two symbols $P(00) = P(10) = \ldots$. The conditional probabilities of the third symbol given a two previous symbol word are assigned different values and, moreover, they are made time dependent, i.e. the simulated symbol string becomes non-stationary. The introduction of non-stationarity appears to be a necessary element in the model that is capable of producing the symbolic strings with the desired property (i.e. demonstrating the growth of the Statistical Complexity with the data volume). We have found that the introduction of a periodic modulation as non-stationarity in defining the conditional probabilities in three symbol words causes the slow convergence in the conditional probabilities, produces a shift in the histograms of the parameter $Disp$ (Fig. 17, right), and also results in the plot of the the C_μ vs. N very similar to that of the molecular signal, Fig. 18.

Moreover, the complexity growth rate depends on the period of the introduced non-statiaonarity, being negligible at short-periodic modulation, and becoming substantial at the time scales of the order of 100 ps. This is in sharp contrast to the case of the similar Markov chain with stationary conditional probabilities that always produced fast convergence and stationary value of C_μ. Thus, we believe that it is the non-stationarity in the transition probabilities that produces the growth of C_μ and exhibits non-trivial behaviour in their distributions $P(\mathbf{v}_{i+1}|s_i)$.

VII. 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 of the system 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 defined by the causal states possess unexpected properties in the dynamical sense. The Poincare returns statistic for these areas allows to classify them into quasi-periodic and chaotic types. The "periodic" ones are robust with respect to increasing the molecular trajectory total length, while the number of "chaotic" ones increases with the size of the time series. We further suggest a non-stationary Markov type model that is capable of reproducing this behaviour. The non-stationarity of transition probabilities in the Markov chain is a necessary attribute of the model that appears to be responsible for the increase in the number of "chaotic" causal states and, hence, the growth of the ϵ-machine.

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

[1] F. Takens, *Detecting strange attractors in turbulence* (Springer Berlin / Heidelberg, 1981), vol. 898 of *Lecture Notes in Mathematics*, chap. Detecting strange attractors in turbulence, pp. 366 – 381, URL `http://www.springerlink.com/content/b254x77553874745`.

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

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

[4] R. Wackerbauer, A. Witt, H. Atmanspacher, J. Kurths, and H. Scheingraber, Chaos, Solitons & Fractals **4**, 133 (1994).

[5] G. Zaslavsky, Phys. Repts **371**, 461 (2001).

[6] D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding* (Cambridge University Press, New York, NY, USA, 1995), ISBN 0-521-55900-6.

[7] S. Wiggins, *Introduction to applied nonlinear dynamical systems and chaos* (Springer, New York, 1990).

[8] V. M. Alekseev and M. V. Yakobson, Physics Reports **75**, 290 (1981), ISSN 0370-1573, URL `http://www.sciencedirect.com/science/article/B6TVP-46T4X1J-8M/2/c0fcca49070a7864a9101d557145155e`.

[9] M. Buhl and M. B. Kennel, Physical Review E **71**, 046213 (2005).

[10] C. R. Shalizi and C. Moore, Studies in History and Philosophy of Modern Physics **submitted** (2003), URL `http://arxiv.org/abs/cond-mat/0303625`.

[11] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).

[12] B. V. Chirikov, Phys. Rep. **52**, 264 (1979).

[13] 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.

[14] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comp. Chem. **26**, 17011718 (2005).

[15] H. J. C. Berendsen, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis (Kluwer, 1991), p. 139155.

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

[17] 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, URL `http://arxiv.org/abs/cs.LG/0406011`.

[18] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, Cambridge series on statistical and probabilistic mathematics (Cambridge Univ. Press, Cambridge, 1997).

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

[20] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, 2005).

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

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

[23] N. J. Corron, S. T. Hayes, S. D. Pethel, and J. N. Blakely, Phys. Rev. Lett. **97**, 024101 (2006).

[24] V. Afraimovich and G. Zaslavsky, Phys. Rev. E **55**, 5418 (1997).