Molecular phase space transport in water: non-stationary random walk model

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

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

Makoto Taiji
Computational Systems Biology Group, Advanced Science Institute, RIKEN, 61-1 Onocho, Tsurumi, Yokohama, Kanagawa 230-0046, Japan

Molecular transport in phase space is crucial for chemical reactions because it defines how pre-reactive molecular configurations are found during the time evolution of the system. Using Molecular Dynamics (MD) simulated atomistic trajectories we test the assumption of the normal diffusion in the phase space for bulk water at ambient conditions by checking the equivalence of the transport to the random walk model. Contrary to common expectations we have found that some statistical features of the transport in the phase space differ from those of the normal diffusion models. This implies a non-random character of the path search process by the reacting complexes in water solutions. Our further numerical experiments show that significant long period non-stationarity in the transition probabilities of the segments of molecular trajectories can account for the observed non-uniform filling of the phase space. Surprisingly, the characteristic periods in the model non-stationarity constitute hundreds of nanoseconds, that is much longer time scales compared to typical lifetime of known liquid water molecular structures (several picoseconds).

I. INTRODUCTION

Molecular transport in liquids can be considered from two seemingly different viewpoints. On the one hand, the familiar diffusion, the mean squared displacement of atoms in Euclidean three dimensional space, describes how, on average, the atoms move in the liquid. On the other, the high-dimensional phase space transport of the dynamical system trajectories comprising all the coordinates and velocities of all the particles in the volume of interest can be analysed. It should be noted that both approaches are closely related, since the Cartesian coordinates defining the three-dimensional volume in the first case are the phase space coordinates as well. Therefore, the phenomenon of the first kind is just a projection of that of the second. It is, however, unclear whether the properties of a low-dimensional projection carry over to higher dimensional subspaces. Therefore, it appears interesting to look at the diffusion process in dimensions higher then three.

High-dimensional transport in molecular systems plays a crucial role in, for example, defining the rates of chemical reactions. The usual three-dimensional motion describes how the reacting molecules are brought to physical contact (Fig. 1, left), while the high-dimensional transport defines the details of the mutual rearrangements of the reacting species and their relative motion during the chemical reaction (Fig. 1, right). In other words, it characterises how the molecules explore different conformations and reciprocal movements (through the inclusion of the atomistic velocities). These details are fundamentally important. For example, the analysis of the molecular arrangements and the angle of attack necessary for the reaction leads to a special, commonly accepted class of molecular structures called "near attack conformers". These are "special substrate conformations in which the bond forming atoms are at the van der Waals distance and at the angle near the one in the transition state" [1, 2]. Near attack conformers can be investigated at the classical level of description, before the beginning of the quantum mechanical process of the actual reaction. Nevertheless, they are shown to be the necessary prerequisites for many reaction to happen (see, for example, [3]). Another example is the role of water in the process of heme degradation by heme oxygenase, where a very special water cluster is necessary for the O-O cleavage and the O-C_{meso} bond formation to take place [4].

It is believed that three-dimensional diffusion in simple liquids such as common solvents or small molecules in solution can be well approximated by a random Brownian motion, at least at the longer than few picosecond time scale relevant for chemical reactivity. This is a postulate of one of the cornerstones of the reaction rates theories, the Kramers theory [5], where the reaction is considered as a diffusion problem and the solvent influence is described as a white noise. This view implies that the reactants find each other as well as the necessary arrangements and angles of attack *by pure chance*. If this is experimentally and theoretically supported in the case of classical three-dimensional diffusion, this point of view is

*Electronic address: dn232@cam.ac.uk

apparently unable to describe the emergence of complicated structures and specific reciprocal movements of reacting atoms in chemical reactions. In the latter case the investigation of the high-dimensional phase space transport becomes promising, since it allows studying in detail the whole sequence of molecular transformations leading to the pre-reactive complexes. It should be noted in this respect that the total volume of the phase space is extremely large and in many cases it seems very unlikely that the required (actually formed) unique configuration of molecules corresponding to a small area in the phase space can be found through a simple random search, Fig. 1.

We are not aware of any detailed investigation of the high-dimensional phase space transport in molecular systems. This is partly because of the substantial technical difficulties in calculating reliable statistics on variables with very large range of values (it grows exponentially with the investigated subspace dimensionality of the phase space) and partly because of the unjustified assumption that the phase space coordinates are Gaussian random variables beyond the correlation time and, thus, do not possess any additional information compared to the usual two-point correlation function.

In this paper we analyse the statistical properties of the phase space transport for bulk water at room temperature. Contrary to the three-dimensional space we find significant deviations from the purely random character of motion. Unlike the standard diffusion model where the displacements of any atom are adequately represented by a Gaussian random process, the studied molecular system exhibits preferable routes in the phase space implying the existence of more probable molecular configurations and reciprocal motions defined by the mutual interactions between the particles. Most surprisingly, we found that the transition probabilities defining these phase space routes as the probabilities of "futures" given specific "pasts" slowly change with time that can be attributed to substantial non-stationarity of the atomistic trajectories. Moreover, we show that this behaviour of the trajectories in the phase space can be reproduced with a non-stationary Markov chain-type model by augmenting it with periodic time modulation of the transition probabilities of the period of 100 ps and longer.

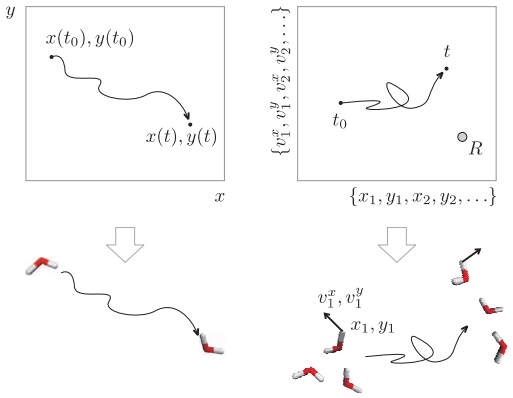

FIG. 1: **Left**: the spatial diffusion in three dimensional space of atom's coordinates x, y, z defines how on average the atom moves from time t_0 to time t; **right**: the phase space transport in the multidimensional space of the coordinates and momenta (velocities) of all the atoms $\{x_1, y_1, x_2, y_2, \ldots\}, \{v_1^x, v_1^y, v_2^x, v_2^y, \ldots\}$ takes into account the mutual arrangements of the molecules, their velocities and the process of the rearrangements of the molecular complexes; a chemical reaction takes place when the phase space trajectory finds a small area R in the phase space that corresponds to a particular pre-reacting arrangement of the atoms

II. INVESTIGATED SYSTEM AND MOLECULAR SIGNAL

We analyse the trajectories of MD simulated bulk water at room temperature (see appendix A for the MD simulation details) sampled at discrete time moments. To be specific, we analyse the motion of one of the hydrogen atoms (signals from other atoms as well as their combinations result in qualitatively the same conclusions, see later) of a randomly selected molecule in the ensemble of 392 water molecules. In the case of normal diffusion described by a standard random walk model the displacement of the atom along any of the three coordinates is well approximated by a randomly distributed Gaussian variable. To avoid the influence of short time correlations that have the pronounced influence on the statistical properties of the analysed trajectories, we downsample the initially three-dimensional trajectory by introducing a two-dimensional cross-section (see appendix B) that approximately corresponds to sampling the data at a rate defined by the first minimum of the autocorrelation function. It can be further shown that for studying the properties of *trajectories* in the phase space, it appears convenient (and sufficient) to replace the continuous signal with symbols from an alphabet of few values, for example, the binary alphabet $\{0, 1\}$ or the three symbol alphabet that we used in our studies $\{0, 1, 2\}$, in other words, introduce a partition.

Thus, we take the velocity of one of the hydrogens as the experimental trajectory in the three-dimensional space and construct its symbolic representation [6, 7] (coordinates and other molecular signals produce the same conclusions, see later). In this way the time series of the velocities of an atom $\{\ldots \mathbf{v}_{i-3}\mathbf{v}_{i-2}\mathbf{v}_{i-1}\mathbf{v}_i\}$ is converted into a symbolic sequence $\{\ldots 0010\}$. Considering symbols instead of the coordinate values for the signal gives us an advantage of a simpler and more robust analysis without essential loss in the statistical information content [8]. Moreover, by considering symbolic subsequences of finite length, we can reconstruct the high-dimensional dynamics of the water molecules ensemble similar to the procedure based on the famous Takens embedding theorem that allows reconstructing the vector dynamics from a scalar observable [9].

The algorithm of symbolisation is not a trivial procedure since it requires an approximation to the generating partition of the phase space [10]. For this purpose it is important to perform a proper selection of an observable, i.e. a variable or a function of variables that allows the efficient approximation procedure to be developed. With this in mind we followed the work [8] and aimed at obtaining a symbolic sequence of maximum Shannon entropy. For this we select the (uniformly distributed) angle φ_{xy} between the components v^x, v^y as a variable of interest and sample it at a rate defined by the cross-section condition $v^z = 0$. Dividing the area of the angular variable φ_{xy} values into three equal segments provides an efficient partitioning for the purpose of symbolisation with 3-symbol alphabet due to the uniformity of the probability distribution of the variable φ_{xy}.

The obtained symbolic sequence can be used to study the transport in the phase space by analysing the symbolic sub-sequences (or "histories") that start at times $t - l$ and end at times t, with t covering the whole simulation period (up to 1μs) and l being the length of the sub-sequence. In this representation each sub-sequence (word) is an l dimensional projection of the whole dimensional molecular phase space. Varying the word length l we investigate various projections of the whole dimensional phase space trajectory. We have found that the results reported here are robust with respect to various projections (velocities, coordinates, etc., and $l \approx$ 5-7 to 13-15) as well as symbolisation schemes (see appendix D).

We would like to stress again that even though we used low dimensional signals for initial symbolisation of the trajectory (the most straightforward were three dimensional velocities or coordinates of individual atoms, but we also analysed many-atom signals, like, for example, instantaneous temperature) the analysed l-symbol words correctly represent the l-dimensional subspaces of the whole phase space corresponding to the molecular system. Two considerations support this point. First, according to Takens embedding theorem, a trajectory of a high-dimensional dynamical system can be reconstructed by applying the time delay procedure to a properly selected scalar observable [11]. Second, representing the selected observable with only few symbols is not an unjustified oversimplification. This is because *sequences* of the data points are considered, when the dynamics "cut out" a "tube" in the phase space that contains only a small fraction of admissible trajectories, Fig. 2. The "cross section" area of the tube depends on the length of the sequences and in the limit of infinitely long sequences converges to a single point (in the special case of a "generating" partition). We have found that for our numerically simulated molecular signal, such as an atomistic velocity, the symbolic sub-sequences of length 7 or higher produce a "tube" that provides essentially the same statistics on the sequences of 3 symbols as the original, continuous valued sequences.

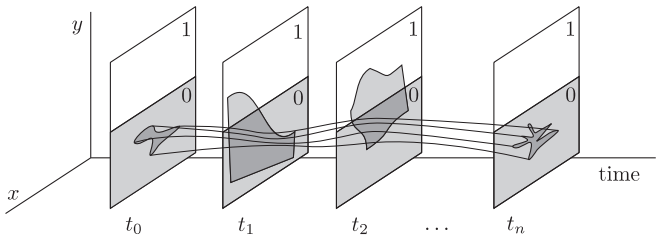

FIG. 2: The process of converting the atomistic trajectory (two dimensional in this example) into a sequence of symbols from the alphabet $\{0, 1\}$; the area marked "0" at time t_0 is transformed by the dynamics at the next time step; all trajectories that pass through the dark shaded area at time t_1 have the symbolic representation 00; the corresponding area is transformed at the next time step t_2 and the dark shaded area here represents all trajectories encoded as 000; the area is non-increasing and for an infinitely long sequence shrinks into a single point (if the partition is a special "generating partition")

III. THE TEST ON RANDOMNESS OF THE DIFFUSION PROCESS

Any l-symbol word represents a point in the l-dimensional symbolic projection of the phase space. To investigate the way the words diffuse in the phase space we calculate various statistical characteristics over the long symbol trajectories generated by the system during its evolution in time. The same characteristics were calculated for an ensemble of artificial random signals obtained by the methods known in the field of signal processing as "surrogate signal analysis" or "bootstrap technique" [12]. The procedure of building a "surrogate" time series generates a signal indistinguishable from the original signal in terms of common statistical characteristics, such as the correlation function, the mean, and the variance, but random by definition (that is it does not contain the dynamic correlations originating from the deterministic motion of atoms). The comparison of the results for the original molecular signal with the random surrogate provides a definitive test on the randomness of the dynamics.

IV. RESULTS OF NUMERICAL ANALYSIS

The simplest statistic characterising the words (phase space points) is their occurrence rate in the long symbolic trajectories corresponding to the whole analysed time series. An example of the occurrences for a typical water trajectory is shown in Fig. 3. For a random symbolic sequence, the occurrence probabilities should be uniform, i.e. constitute a uniform distribution. This is indeed observed for the case of the surrogate signal, Fig. 3. However, the same statistic calculated from the molecular signal shows significant non-uniformity in the corresponding distributions. There are several frequent

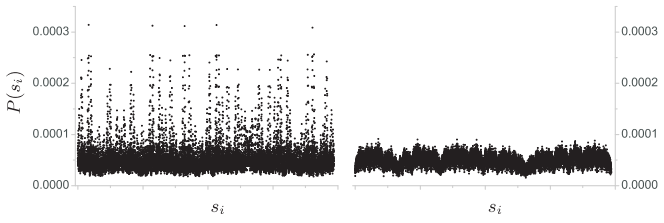

FIG. 3: The probabilities $P(s_i)$ of 9-symbol words in the symbolic sequence obtained from the hydrogen velocities of a 1μs long molecular trajectory (left) and the surrogate of the same length (right); the alphabet of 3 symbols {012} was used; words s_i are numbered arbitrarily

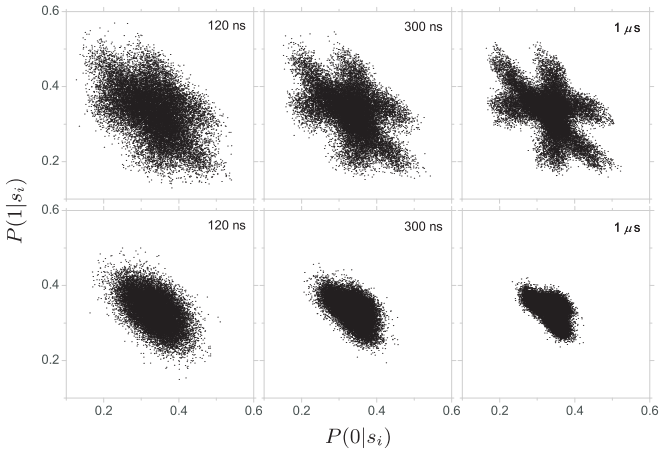

FIG. 4: Conditional probabilities $P(0|s_i)$ versus $P(1|s_i)$ for the signal of Fig. 3, where s_i are all sequences of 9 symbols from the three symbol alphabet {012}; upper row: molecular signal, lower row: the surrogate; the time shown on the panels is the length of the trajectory used to calculate the plot

words that signify preferable routes in the phase space of the water dynamical system.

A more elaborate statistic describes the conditional probabilities of the symbol following each sequence: $P(\mathbf{v}_{i+1}|s_i)$, 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, this characteristic can be easily visualised with two-dimensional scatter plots by plotting the probabilities of $P(0|s_i)$ versus $P(1|s_i)$ (the probability of the third symbol is defined by the first two). The results for the molecular and the corresponding surrogate time series are shown in Fig. 4. Two features become clear after the analysis: (i) the distributions are significantly different for the two cases and (ii) the statistic converges extremely slow even at the time scale as long as hundreds of nanoseconds. The difference in the shapes of the observed patterns thus quantifies the deviation of probabilities $P(\mathbf{v}_{i+1}|s_i)$ in our molecular time series from those expected for a purely random signal.

In order to analyse the slow convergence in conditional probabilities $P(\mathbf{v}_{i+1}|s_i)$ we studied their dependence on the length N of the symbolic series. 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. 5, 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.

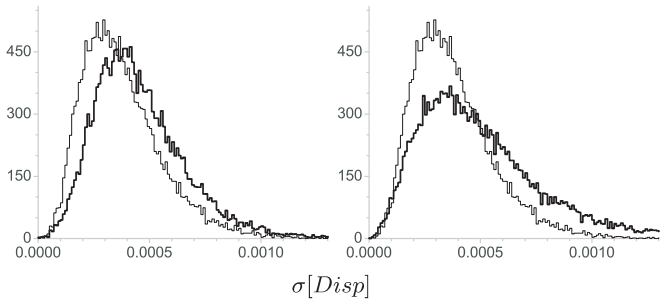

FIG. 5: The histograms of the standard deviations σ of the dependence $Disp(N)$ in the interval $N \in [30\times10^6; 31\times10^6]$, see text; histograms were calculated for all 9-symbol words occurring in the symbolic sequences of **left**: a molecular trajectory (thick line) and corresponding surrogate (thin line); **right**: the surrogate obtained from the molecular trajectory (thin line) 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; the molecular trajectory and the corresponding surrogate are the same as used in Fig. 3

Finally, in order do provide a quantitative description of the detected poor convergence in the conditional probabilities of the water time series we utilised the technique known as Computational Mechanics (CM) approach [13]. The CM analysis introduces a characteristic, Statistical Complexity C_μ, that is calculated over the distribution of conditional probabilities $P(\mathbf{v}_{i+1}|s_i)$ (see appendix C for details) and estimates its Shannon entropy. By definition, the Statistical Complexity equals to zero for either purely random symbolic sequences (all conditional probabilities are equal, i.e. uniform distribution is observed) or trivial periodic ones (strictly predictable sequence, that is one conditional probability equals to unity, others vanish), taking non-zero values for "structured" symbolic sequences (non-trivial distribution of conditional probabilities).

The analysis in the CM framework has confirmed that (i) the molecular signal is indeed different from the random surrogates (the Statistical Complexity is significantly higher for water) and, very unexpectedly, (ii) the value of C_μ *never converges* with the length of the simulation N, up to the longest simulations we tried, 1μs. This is in contrast to the surrogate symbolic sequence which demonstrates a quick convergence to a constant value, significantly lower than that for the molecular signal. We

thus confirm the conjecture that the stable growth of C_μ with N is the result of the slow convergence of the conditional probabilities in the water time series, Fig. 4. It is defined not by short time correlations that could be detected by a standard correlation analysis, but by long time statistical properties of the trajectories of atoms in the phase space.

Here we would like to make a short remark concerning a possibility that the detected effect is a numerical artefact of the algorithm used in our MD simulations. For example, a natural question to ask would be: **is the phenomenon a simulation artefact due to the numerical errors and/or the thermostat**? We addressed this and other similar questions by a thorough check of the numerical inaccuracies of the simulation protocol and the use of the thermostat. For this purpose we varied the thermostat parameters (in the range of several orders of magnitude) as well as we made experiments with several different types of the thermostat (deterministic rescaling of velocities or stochastic), as described in Appendix A. In addition, the simulations with single as well as double floating point precision were compared. In all cases the calculations brought statistically identical result, thus, confirming the genuine dynamical origin of the found non-randomeness in the molecular time series.

V. NON-STATIONARY DIFFUSION MODEL

To provide an explanation for the slow convergence of the conditional probabilities of symbolic sub-sequences we have constructed a simple Markov chain-type model that, on the one hand, has trivial statistical measures, and on the other, demonstrate significant deviation of C_μ from zero. The application of CM approach to such symbolic sequences reveals similar results to what was observed for the molecular signal. 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 $P(0) = P(1) = P(2) = 1/3$, as well as the probabilities for any combinations of two symbols $P(00) = P(10) = \ldots = 1/3$. The conditional probabilities of the third symbol given a two symbol word are made different and, moreover, they are time dependent. In other words, the resulting symbolic string is *non-stationary*.

The introduction of non-stationarity appears to be necessary for producing the symbolic strings with the desired property (demonstrating the growth of the Statistical Complexity with the data volume). We have found that the introduction of a periodic modulation as a non-stationarity in defining the conditional probabilities in three symbol words causes the shift in the histograms of the parameter $Disp$ (Fig. 5, right) and also makes C_μ grow with N very similar to the molecular signal. Moreover, the effect depends on the period of the introduced non-stationarity, being negligible for short period modulation, and becoming pronounced at the time scales of the order of 100ps modulation. This was in sharp contrast to the case of the Markov chain with stationary conditional probabilities, that always produced fast convergence and constant 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 $P(\mathbf{v}_{i+1}|s_i)$.

VI. CONCLUSIONS

We have found that molecular trajectories of bulk water fill the phase space in a very non-uniform manner and hence not randomly. This contradicts the simple assumption of the applicability of the random walk model traditionally made in the case of normal diffusion. The assumption on random motion of atoms is fundamental for many theories of chemical reaction rates in molecular systems and, thus, our funding may have significant consequences for them. More specifically, this implies very different waiting times until the required pre-reactive complexes and angles of attack occur compared to the commonly assumed random search mechanism. Our results show that the mechanism leading to the non-randomness of the phase space search is the existence of preferable routes in the phase space that leads to the appearance of more probable pieces of the phase space trajectories. We have also found that one of the possible mechanisms for such an unexpected behaviour can be the modified random walk where the transition probabilities conditioned on the pieces of the trajectory change in time at the scale of 100 ps and longer.

It is worth noting that several characteristic motions of molecules are known to exist in water (however, specific molecular structures and mechanisms are still the subject of active discussion, see, for example, [14, 15]). Classified by the time scale, they include librations ($<$ 1ps), rearrangements within the "cage" of the nearest molecules ($\approx$1ps), and long-term motions causing the hydrodynamic "tails" in the correlation function (tens of picoseconds). However, *all these phenomena can be unambiguously identified by the standard analysis with the autocorrelation function* and have the time scales much shorter that the discussed convergence rates of conditional probabilities in the symbolic time series. Finally, we have also found that the described phenomenon is probably not linked to any unique properties if water, since it is also found in other molecular liquids (we tested liquid argon, octanol, and octanol-water mixtures). Therefore, it seems to be a rather general property of liquid molecular systems.

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

APPENDIX A: SIMULATION DETAILS

Bulk water consisting of 392 or 878 SPC, SPC-E (Simple Point Charge Extended) [16], or TIP3P (Transferable Intermolecular Potential 3 Point) molecules was simulated using the `GROMACS` molecular dynamics [17] package. The temperature of the system was kept constant at 300K using Berendsen [18] or Nose-Hoover [19] thermostats, with a coupling time of 0.1 ps, whose combination with various coupling constants was investigated. Pressure coupling was also applied to a pressure bath with reference pressure of 1 bar and a coupling time of 0.1 ps. A 1 nm cutoff distance for both van der Waals and Coulomb potentials was used. An equilibration until the potential and kinetic energies reach constant levels of fluctuations was performed before collecting data for analysis. The velocity or the coordinate of the oxygen and hydrogen atoms of one of the water molecules was used as a 3-dimensional signal for the analysis. Instant temperature, $T_{inst} = \frac{1}{N_{df}k}\sum_i m_i \mathbf{v}_i^2$, where the summation is over all atoms, N_{df} is the number of degrees of freedom and k is the Boltzman constant, was also used for the analysis. The simulation time step was 2 fs and all time points were used in the analysis.

APPENDIX B: CONVERTING THE MOLECULAR SIGNAL INTO A SYMBOLIC SEQUENCE

To discretise the three-dimensional velocity trajectories of individual atoms of the molecular system we used its intersections with the xy cross-section plane (similar to Poincare section in dynamical systems theory). For hydrogen water atoms, for example, the average time interval between the intersections was equal to 0.032 ps. Very conveniently it roughly corresponds to the first minimum on the autocorrelation function, obeying the general rule for time sampling of signals. The resulting two-dimensional points approximately uniformly cover the area and form a centrally-symmetric distribution of points, Fig. 6.

In order to convert the points at the cross-section into a sequence of symbols from a finite alphabet, an appropriate partitioning of the continuous space is required. A natural choice for such partitioning is the generating partition (GP) [20] that has the property of a one-to-one correspondence between the continuous trajectory and the generated symbolic sequence. That is, in the ideal case of GP, all information is retained after the symbolisation.

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 two disadvantages when applied to realistic molecular signals. First, an algorithm for calculating a GP in a general case is unknown. Second, it is shown for simple tent maps [21] that the values of statistical complexity for different GPs of the same system are different (a system can have many GPs, not to confuse with the uniqueness of a symbolic representation of a trajectory for a given GP).

Recently methods for finding approximations for GP are reported. The method from [10] is shown to reproduce GP for known systems and can treat multi-dimensional observed time-series data. The results of the application of this method to our velocity data using 2, 3, 4, and 5 partitions are shown in Fig. 6. For all cases the resulting approximations to GP are centrally symmetric (reflecting the central symmetry of the data points distribution). Thus, for our signals we used centrally symmetric 3-partitions in all subsequent calculations.

Summarising, in converting the three-dimensional molecular trajectories into symbolic sequences we, first, built a two-dimensional map by finding the intersections of the trajectory with the xy-plane and, second, assigned a symbol to each point of the map depending to what segment of the partition the point belongs (Fig. 6).

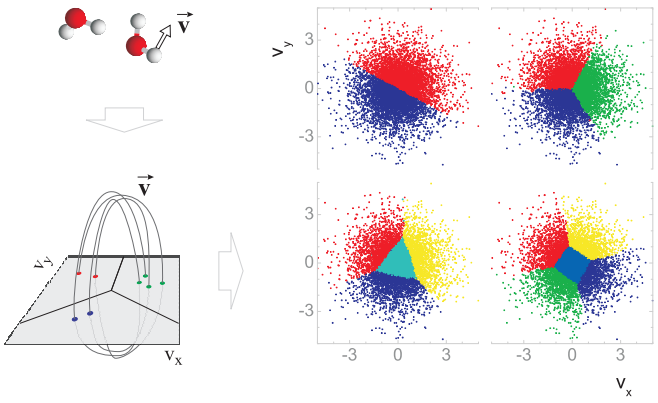

FIG. 6: The process of converting the continuous atomic velocity signal $\mathbf{v}$ into symbolic sequence. On the right the symbolisation with 2, 3, 4, and 5 symbols are shown

APPENDIX C: COMPUTATIONAL MECHANICS

Computational Mechanics analyses symbolic sequences that represent a temporal dynamics of some system. All past A_i^- and future A_i^+ halves of bi-infinite symbolic sequences centred at times i are considered. Two pasts A_1^- and A_2^- are defined as 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 time moment the system is in 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 *sufficient* statistic, that is it contains the complete statistical information about the data;
- a *minimal sufficient* statistic, therefore the causal states can not be subdivided into smaller states;
- a *unique minimal sufficient* statistic, any other one simply re-labels the same states.

The *Statistical Complexity* is the information-theoretic measure of the size of the ϵ-machine and quantifies the amount of information about the past of the system that is needed to predict its future dynamics: $C_\mu = H[P(\epsilon_i)]$, where H is the Shannon entropy.

APPENDIX D: ROBUSTNESS OF THE RESULTS WITH RESPECT TO VARIOUS PROJECTION OF THE PHASE SPACE

Two parameters of the algorithm should be set in calculating C_μ of a signal of given length, the alphabet size K and the length l of the histories s^- used by the ϵ-machine reconstruction algorithm `CSSR`.

The dependence of C_μ on both parameters is shown in Table I. The convergence with l is excellent, so that for $l \geq 6$ the algorithm produces almost identical results. Reliable results for large alphabet sizes K are more difficult to obtain because for higher K much longer signals are required. This explains the somewhat increased values of C_μ for $K = 5$ in Table I.

Varying the position of the Poincare section plane along the z axes did not lead to any significant change in the results. The effect of various partitionings of the continuous space has been checked by applying non-symmetric (same as symmetric but shifted along the x and/or y axes) partitions. In all shifted partitioning cases this resulted in somewhat lower values of C_μ. Any variants of centrally symmetric partitioning produced identical results.

Finally, different values of the adjustable parameter of the `CSSR` algorithm, the significance level for the χ-squared test that quantifies the statistical equivalence of the histories, has been checked. For the values of 0.001, 0.01, and 0.1 the same qualitative behaviour of C_μ has been reproduced.

TABLE I: Statistical Complexity C_μ vs. the length of histories l ($K = 3$) and the alphabet size K ($l = 6$) for bulk water hydrogen velocity 60 ns long signal

l	C_μ	K	C_μ
2	3.17	2	5.22
3	4.75	3	7.95
4	6.11	4	8.23
5	7.31	5	8.68
6	7.95		
7	8.15		
8	8.21		
9	8.29		
10	8.37		

[1] T. C. Bruice and S. J. Benkovic, Biochemistry **39**, 6267 (2000), http://pubs.acs.org/doi/pdf/10.1021/bi0003689, URL `http://pubs.acs.org/doi/abs/10.1021/bi0003689`.

[2] T. C. Bruice, Accounts of Chemical Research **35**, 139 (2002), http://pubs.acs.org/doi/pdf/10.1021/ar0001665, URL `http://pubs.acs.org/doi/abs/10.1021/ar0001665`.

[3] P. Hirunsit and P. B. Balbuena, The Journal of Physical Chemistry A **112**, 4483 (2008), http://pubs.acs.org/doi/pdf/10.1021/jp711101b, URL `http://pubs.acs.org/doi/abs/10.1021/jp711101b`.

[4] H. Chen, Y. Moreau, E. Derat, and S. Shaik, Journal of the American Chemical Society **130**, 1953 (2008), http://pubs.acs.org/doi/pdf/10.1021/ja076679p, URL `http://pubs.acs.org/doi/abs/10.1021/ja076679p`.

[5] H. A. Kramers, Physica **7**, 284 (1940), ISSN 0031-8914, URL `http://www.sciencedirect.com/science/article/B6X42-4CB752H-3G/2/18dd09fed8a9142aed637597660731c5`.

[6] C. S. Daw, C. E. A. Finney, and E. R. Tracy, Review of Scientific Instruments **74**, 915 (2003), URL `http://link.aip.org/link/?RSI/74/915/1`.

[7] D. Nerukh, V. Ryabov, and R. C. Glen, Physical Review E **77**, 036225 (2008).

[8] M. Lehrman, A. B. Rechester, and R. B. White, Phys. Rev. Lett. **78**, 54 (1997).

[9] 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`.

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

[11] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).

[12] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer, Physica D **58**, 77 (1992).
[13] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).
[14] D. Laage and J. T. Hynes, Science **311**, 832 (2006), http://www.sciencemag.org/cgi/reprint/311/5762/832.pdf, URL `http://www.sciencemag.org/cgi/content/abstract/311/5762/832`.
[15] D. Laage and J. T. Hynes, The Journal of Physical Chemistry B **112**, 14230 (2008), http://pubs.acs.org/doi/pdf/10.1021/jp805217u, URL `http://pubs.acs.org/doi/abs/10.1021/jp805217u`.
[16] 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.
[17] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comp. Chem. **26**, 17011718 (2005).
[18] H. J. C. Berendsen, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis (Kluwer, 1991), p. 139155.
[19] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[20] S. Wiggins, *Introduction to applied nonlinear dynamical systems and chaos* (Springer, New York, 1990).
[21] O. Gornerup and K. Lindgren, personal communication (2006).
[22] C. Shalizi, K. Shalizi, and R. Haslinger, Physical Review Letters **93**, 118701 (2004).