A data driven approach to classify descriptors based on their efficiency in translating noisy trajectories into physically-relevant information

MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

PAPER • OPEN ACCESS

A data driven approach to classify descriptors based on their efficiency in translating noisy trajectories into physically-relevant information

To cite this article: Simone Martino *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 035039

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 035039 https://doi.org/10.1088/2632-2153/adfa66

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
10 January 2025

REVISED
9 June 2025

ACCEPTED FOR PUBLICATION
11 August 2025

PUBLISHED
26 August 2025

PAPER

A data driven approach to classify descriptors based on their efficiency in translating noisy trajectories into physically-relevant information

Simone Martino, Domiziano Doria, Chiara Lionello, Matteo Becchi and Giovanni M Pavan*

Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy
* Author to whom any correspondence should be addressed.

E-mail: giovanni.pavan@polito.it

Keywords: high-dimensional analysis, dimensionality reduction, information extraction, time-series analysis, unsupervised clustering, descriptors, local noise reduction

Supplementary material for this article is available online

Abstract

Reconstructing the physical complexity of many-body dynamical systems can be a hard task. Starting from the trajectories of their constitutive units (raw data), typical approaches require choosing adequate parameters/descriptors to convert them into time-series that are then analyzed to extract human-interpretable information. However, identifying the best descriptor is often far from being trivial. Here we report a data-driven approach that allows to compare the efficiency of different types of descriptors in extracting information from noisy trajectories and translating them into physically-relevant information. As a prototypical example of a system with non-trivial internal complexity, we analyze molecular dynamics trajectories of an atomistic model system where ice and water coexist dynamically in correspondence of the solid/liquid transition temperature. We compare different types of general or specific descriptors often used to study aqueous systems, e.g. number of neighbors, molecular velocities, smooth overlap of atomic positions (SOAP), local environments and neighbors shuffling (LENS), orientational tetrahedral order, and distance from the fifth neighbor (d_5). We use Onion clustering (an efficient unsupervised clustering method for timeseries analysis) to assess the maximum amount of information that can be extracted from the noisy trajectories by the various descriptors, which we then rank via a high-dimensional metric. Our results demonstrate how advanced descriptors, such as SOAP and LENS, outperform classical ones thanks to higher signal-to-noise ratios. Nonetheless, even the simplest descriptor can become as efficient (and even more) as advanced ones upon local-denoising of their signal. This is the case of, e.g. d_5, among the worst performing descriptors, which becomes following to denoising by far the best one in resolving the non-strictly-local dynamical complexity of such an ice/water system. This work highlights the critical role of noise in the process of information extraction and it offers a data-driven approach to identify optimal descriptors for systems with characteristic internal complexity.

1. Introduction

The study of complex molecular systems composed of numerous constituent units, especially under conditions where they exhibit non-trivial internal dynamics, can be quite challenging. In such ensembles, interacting molecules can occupy a wide range of microstates, constantly exchanging among them. Untangling the complexity of such dynamical network is crucial for understanding the underlying physics of these systems, but it is often a difficult task [1–6]. Experimentally, it is typically challenging to obtain microscopic-level information with the spatial and temporal resolution necessary to discriminate and comprehend the processes occurring within these systems. In contrast, computational approaches such as molecular dynamics (MDs) simulations can generate detailed trajectories for a given system, providing

information on the individual trajectories of the units composing it, their mutual arrangements, and more. However, the extraction of meaningful information from these raw and often noisy datasets is often non-trivial.

Using collective variables, or descriptors, can help to extract and retain key information from the raw MD trajectories, making them interpretable and useful for describing the phenomena and processes that characterize the system [7–9]. Typical approaches often study the behavior of a system by relying on intuitive, physically meaningful descriptors, such as distances, number of neighbors, and geometrical shape orders [10]. However, a potential drawback of these approaches is that relying on descriptors based on human intuition can lead to biased or incomplete analysis, revealing only what was initially expected and potentially overlooking other important information. Another limitation of human-based descriptors is that they are often not transferable between different systems, which makes it more challenging to draw meaningful comparisons.

To address these limitations, recent efforts have focused on developing more abstract and 'agnostic' descriptors, which do not rely on preconceived knowledge or physical intuition about the system [11, 12]. Some of these descriptors, based on particle density expansion, efficiently capture information about the local order and disorder in the positions of neighboring atoms (or molecules/units) in a given system. Notable examples include the smooth overlap of atomic positions (SOAP) [13], the atomic cluster expansion [14] or the N-body iterative contraction of equivariants (NICE) [15] frameworks.

Alternatively, there are descriptors of a different nature that do not quantify static features at each time step, but rather track how local features evolve over time—the so-called dynamic descriptors. A relevant example of such abstract dynamic descriptor is the local environments and neighbors shuffling (LENS) [16], which tracks how the identity of the particles in the local environment surrounding each particle changes over time. LENS quantifies the dynamics of local environments, capturing significant fluctuations and providing insights into the system's microscopic dynamical homogeneity or inhomogeneity [16–19]. Another relevant example is *Time*SOAP (*t*SOAP) [20], a one-dimensional quantity that tracks changes in the SOAP spectra of the particles along the trajectory. While LENS depends on the identities of neighboring particles without considering spatial coordinates, *t*SOAP is sensitive to their positions, so that the two descriptors capture different physical aspects of the particles' environment [18, 21].

Aside from the ability of a specific descriptor to extract relevant information for a given system, another important point is finding the best approach to analyze the dataset once all the data are collected. Recent studies have shown, for example, that performing a global pattern recognition analysis on such datasets in the attempt of identify the microscopic environments within a system may only uncover statistically dominant patterns, often overlooking crucial information (such as their early emergence, or the presence of other less statistically relevant domains) [21]. Information loss can occur when time correlations within the time-series are neglected or when sparsely populated domains are masked by noise from dominant ones. Thus, incorporating time correlations into the analysis of each particle's signal can reveal events that might otherwise be missed by traditional pattern recognition approaches [17, 22].

Given the large variety and diversity of descriptors that can be used to translate raw data (MD trajectories) into a dataset to analyze, a key question is which one is the best suited for a given system. In general, identifying the most appropriate descriptor is a crucial step, as preconceived choices—often guided by prior experiences or previous studies—may compromise or bias the final interpretation of the results. To address these important point, we present a purely data-driven and agnostic approach to effectively compare the efficiency of different descriptors in extracting and resolving the information contained in a trajectory. As a prototypical test case, we use a molecular trajectory with known non-trivial internal dynamical complexity—specifically, a water system at the solid/liquid coexistence point [16, 20].

This system comprises various environments (solid ice, liquid water, liquid-solid interface, etc) that are highly heterogeneous both structurally and dynamically, making it an ideal case study for assessing the effectiveness of local descriptors in resolving the system's complexity. We compare different types of static and dynamic descriptors, ranging from simpler human-based and physics-inspired ones to more abstract and data-driven ones that are general and agnostic. By employing a purely data-driven metric, we can quantify and compare the information extracted from the MD trajectory by each descriptor. This allows us to classify them based on their similarities and differences, as well as their efficiency in extracting and classifying information from the analyzed trajectories. Since a descriptor efficiency is determined by its signal-to-noise ratio, we also assess the impact of noise reduction across all explored descriptors using a recently reported approach [23]. Our results show that even the simplest descriptors, once denoised, can be as efficient as the most advanced ones. This work highlights that it is more appropriate to discuss the best analysis framework rather than the best descriptor for extracting information from a specific system. Additionally, it provides a general, agnostic, physically interpretable, and data-driven approach to identify this framework based on a maximum resolved information criterion.

2. Results

2.1. Extracting information from trajectories

Extracting information from a trajectory primarily involves two main steps. First, starting with the raw trajectory (the collection of $\vec{r}_i(t)$ coordinates for every unit i at every time-step t), the data are converted in a set of time-series $D_i(t)$ by selecting an appropriate descriptor D. Second, these time-series $D_i(t)$ are analyzed in various ways to extract meaningful information (figure 1(A)). This general approach applies to the analysis of any trajectory, not limited to those obtained from simulations; here, as a prototypical example we consider a 50 ns-long MD trajectory of an atomistic model system composed of 2048 TIP4P/ICE molecules [24] in a rectangular simulation box, as shown in figure 1(B). The system is simulated with periodic boundary conditions, starting from a configuration in which 50% of the molecules are in the liquid state and the remaining 50% is in the solid state, arranged in a hexagonal ice crystalline structure (*Ih*) [24]. After equilibrating the system at the melting temperature, we conducted a 50 ns production run, saving the molecules' coordinates every $\delta t = 0.1$ ns (additional simulation details can be found in the methods section 4.1).

Various descriptors can be used to extract information from these raw MD trajectories, broadly distinct between static (figure 1(C)) and dynamic (figure 1(D)). Static descriptors provide characteristic fingerprints of each molecule's local environment at each time-step; they are functions of the molecular coordinates at a specific time frame: $D_i(t) = D(\vec{r}(t))$. Common descriptors used to study aqueous systems include the number of neighbors (N_{neigh}), the distance from the fifth neighbor (d_5), the orientational tetrahedral order parameter (q_{tet}), and the previously mentioned SOAP. In contrast, dynamic descriptors depend on the change in molecular environment between consecutive time-steps: $D_i(t) = D(\vec{r}(t), \vec{r}(t+\delta t))$. An example is LENS, which captures the local dynamics and diffusivity of each molecule's environment by assessing the reshuffling and exchange among neighboring molecules within a certain cutoff distance [16]. LENS primarily detects two types of changes in local environments over time: fluctuations in the number of neighbors (addition or departure of neighbors within δt) and changes in neighbor identity (swapping of molecules inside the cutoff with others outside within δt). This enables LENS to identify dynamically diverse microscopic (local) environments in a system, along with the fluctuations between them.

Figure 1(E) depicts the LENS time-series computed for all water molecules in the system. The cumulative kernel density estimation (KDE) of the LENS data points reveals two density peaks: one at LENS ~ 0.1, indicating a more static environment (corresponding to solid ice), and another at LENS ~ 0.4, identifying molecules in a more dynamic environment (corresponding to liquid water). However, there is more information in these data beyond the high-density peaks of the cumulative KDE. Analyzing these time series in detail requires distinguishing meaningful fluctuations from noise. Typical approaches involve single-point clustering of the time-series, which are examined over time rather than as a set of uncorrelated frames. Single point time-series clustering is a class of methods for grouping individual time-series data points based on their similarity in features, patterns, or dynamics. This approach allows for identifying and categorizing distinct temporal behaviors within a complex dataset, revealing underlying structures or states in the system [22, 25–27].

Many different (supervised or unsupervised) learning approaches can be used, in principle, to analyze datasets and extract information from them. Here, we base our analysis on the recently published Onion clustering method [17] and on the unsupervised detection and classification of signal fluctuations, which has been recently demonstrated to be very effective in the analysis of noisy data (for comparisons with widely used methods, we refer the interested readers to dedicated papers on this topic [6, 16, 18, 20, 28]). A relevant feature of Onion clustering is that it requires the choice of a time resolution Δt, representing the minimum lifetime for an environment to be considered stable. The key factor we highlight here is that, instead of relying on a specific *a priori* choice of Δt (which can lead to a potentially risky black-box approach), Onion clustering automatically performs single-point clustering across all possible Δt values—from the highest possible resolution ($\Delta t = 2$ frames) to the lowest (the entire trajectory length, $\Delta t = 500$ frames for the current simulation). For each Δt, it outputs both the number of clusters that can be statistically robustly distinguished at that resolution and the fraction that cannot be classified due to insufficient resolution (indicating dynamical events occurring faster than Δt). For a detailed description of Onion clustering, we refer readers to [17].

Figure 1(F) shows the results of Onion clustering for different Δt values in the analysis of the LENS time-series extracted for this system under study. The number of classifiable clusters at each Δt is shown in blue, while the fraction of unclassified data points (lost information) is colored in orange. The Onion results demonstrate that, down to a resolution of $\Delta t < 20$ ns, three statistically-relevant environments can be resolved in the LENS time-series. Figure 1(G) colors the water molecules based on their assigned clusters at the example resolution of $\Delta t = 0.3$ ns, clearly identifying three physically relevant and dynamically different

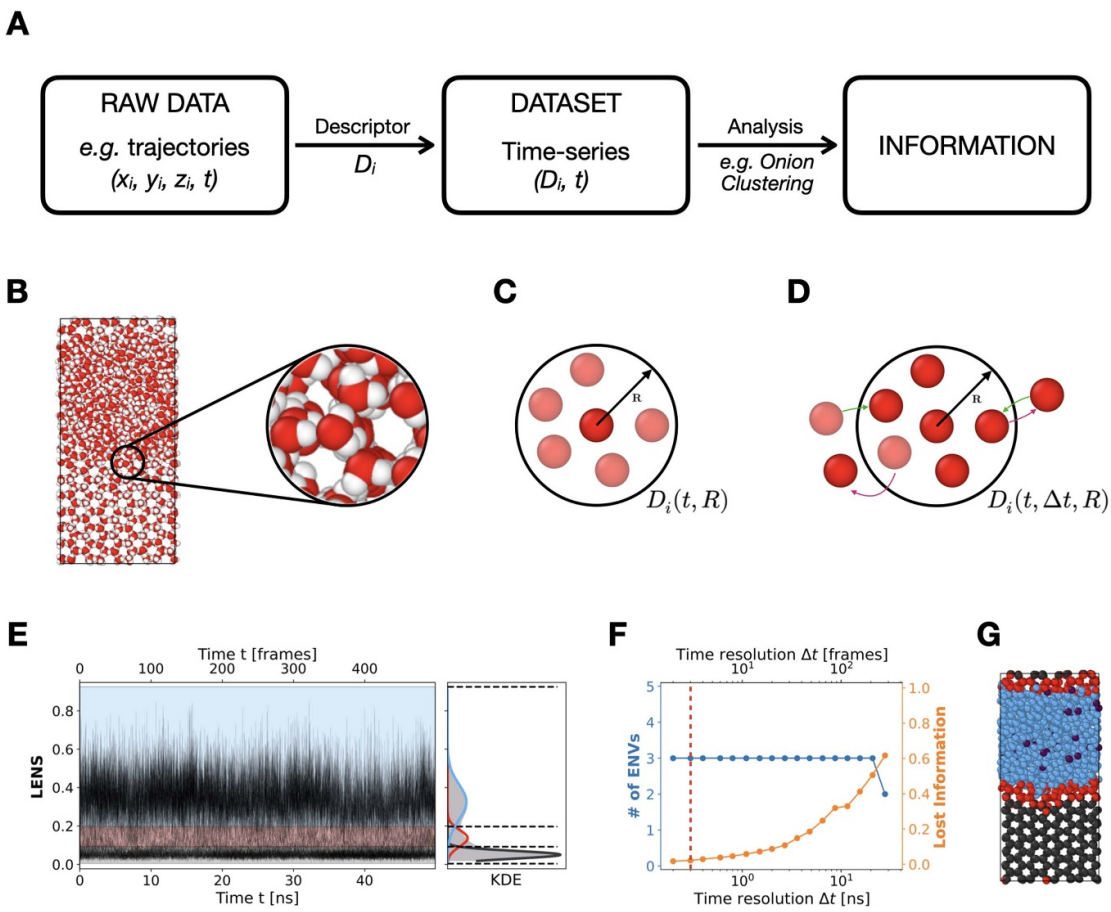

Figure 1. Ice/liquid water coexistence MD simulation, descriptors and LENS clustering. (A) Flowchart of the information extracting procedure through the use of a generic descriptor D_i. (B) Snapshot of the simulation. Oxygen atoms are colored red, and hydrogen atoms are colored white. (C) Schematic representation of a static descriptor, $D_i(t, R)$, which depends on the coordinates and/or identities of molecules within a cutoff radius R at time t. (D) Schematic representation of a dynamic descriptor (LENS), which depends on the variation in the molecules' identities within a cutoff radius R at times t and $t + \delta t$. (E) LENS signal time-series, for each particle as a function of simulation time. The background is colored according to the thresholds between the three identified clusters: solid ice (gray), liquid water (blue) and interface (red). The KDE of the signals (gray shaded area) is overlaid with the Gaussian distributions fitted by the clustering algorithm. Dashed lines represent the thresholds between the clusters. (F) Typical Onion clustering output plot, showing the number of environments detected (blue line) and the fraction of unclassified data points (orange line), as a function of the time resolution Δt. The red dashed lines indicates the time resolution ($\Delta t = 0.3$ ns) used for the clustering shown in this figure. (G) Simulation snapshot where molecules are colored according to their cluster assignment (for $\Delta t = 0.3$ ns). Unclassified molecules are colored purple.

clusters: solid ice (dark gray), the ice-water interface (red), and liquid water (blue). The small fraction of unclassified molecules is represented in purple, corresponding to molecules undergoing transitions faster than the selected time resolution of the analysis. Notably, beyond $\Delta t = 20$ ns, the efficiency decreases, and the number of distinguishable clusters drops to 2, with the ability to distinguish the interface being lost. To confirm the physical meaning of the clusters obtained by the Onion method, we computed the average self-diffusion coefficient of the water molecules belonging to every different environment, obtaining a good agreement with already published results [29, 30]. In particular, the water and ice environments show diffusion coefficients that are consistent with those of water in the liquid and solid states (see figure S1). Noteworthy, the ice/water interface environment is composed of molecules with a roughly intermediate diffusivity. This underlines the importance of detecting such an environment as a separate/distinct one from the other two, as it emerges from the Onion clustering analysis of the LENS time-series data (but neglected in many classical pattern recognition approaches based on data density).

2.2. Comparison between different descriptors

In this study, we compared various descriptors commonly used in water simulations to assess their effectiveness in capturing the system's features. As examples of more general descriptors we selected the number of neighbors within a defined cutoff distance N_{neigh} (figure 2(A)), and the modulus of each molecule's velocity v (figure 2(C)). For descriptors specifically designed for aqueous systems, we considered

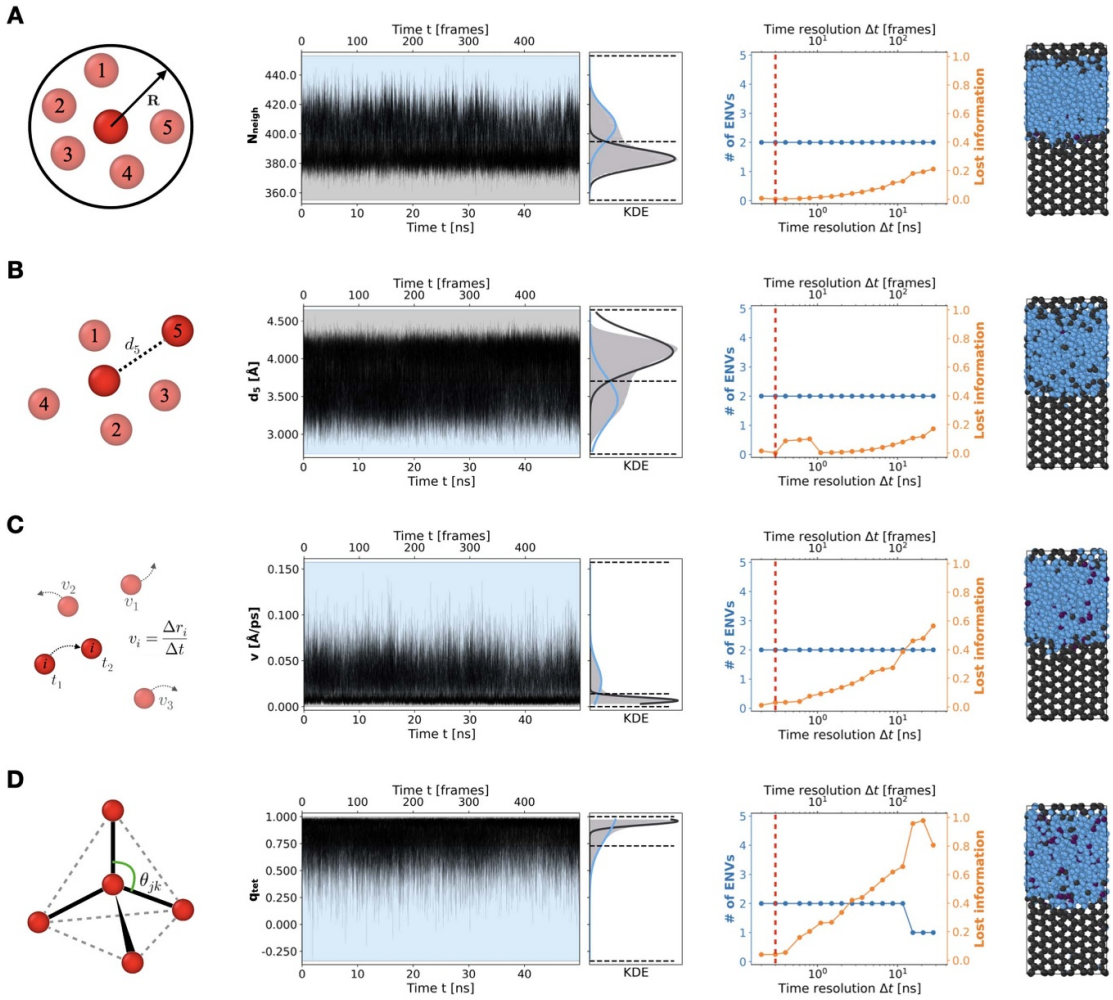

Figure 2. Comparison between descriptors. For the different descriptors — (A) number of neighbors N_{neigh}, (B) distance from the fifth atom d_5, (C) molecule velocity modulus v and (D) orientational tetrahedral order parameter q_{tet}—we show, from left to right: a schematic representation, the signal distribution over the entire simulation, clustered KDE, the number of resolved environments (blue line) and fraction of unclassified data (orange line), and a clustering-colored snapshot in correspondence to the time resolution highlighted with the red line.

the distance from the fifth neighbor d_5 (figure 2(B)), and the orientational tetrahedral order parameter q_{tet} (figure 2(D)). More details on the computation of these descriptors can be found in section 4.2.

Figures 2(A)–(D) displays the time-series generated from the MD trajectory for each of these physics-inspired descriptors, together with the clusters identified by Onion clustering, shown with solid colored lines on the KDE of the signals. As can be seen from the plots and snapshots in figure 2, for all the descriptors at most two different environments (solid ice and liquid water) are identified. With none of these descriptors the clustering algorithm is able to resolve the solid/liquid interface, differently from what we found using LENS. These results indicate that these descriptors are less effective than LENS in characterizing the system's internal physics. Specifically, while they reliably distinguish between ice and liquid phases, they fail to capture the distinct structural and dynamical features of the interface region. This outcome suggests that descriptors specifically tailored for a particular system—in this case, aqueous environments—do not necessarily outperform more abstract, general descriptors. Furthermore, both static and dynamic descriptors yielded similar results, with no clear advantage in using one type over the other for identifying the key features in this system.

2.3. A more advanced high-dimensional descriptor: SOAP

One powerful and widely used descriptor in the study of complex molecular systems, including aqueous ones, is the SOAP descriptor. In particular, it provides a rotational-invariant decomposition of the local particle density of the neighborhood around each particle within a defined cutoff radius [13]. For each particle i at each simulation frame t, SOAP generates a spectrum of coefficients that encode a specific fingerprint of the relative positions of neighboring particles.

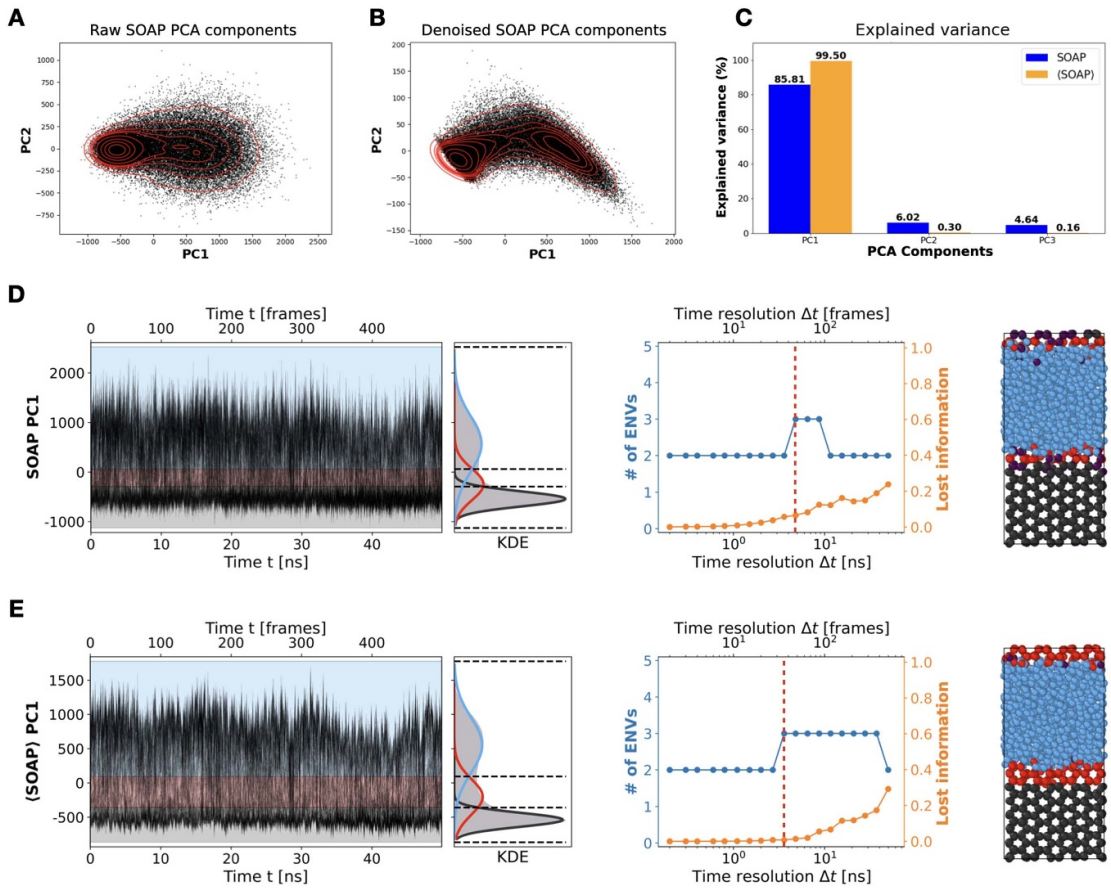

Figure 3. Clustering results on SOAP. (A) Projection of SOAP on the first two Principal Components (PCs). (B) Projection of spatially averaged SOAP over the first two PCs. (C) First three PCs' explained variance of raw SOAP (blue bars) and spatially averaged (orange bars). (D) Onion clustering output on SOAP; from left to right, signal distribution, clustered kernel density estimation (KDE), number of resolved environments (blue line), fraction of unclassified data (orange line) and clustering-colored snapshot in correspondence to the time resolution highlighted with the red line. (E) Onion clustering output on spatially averaged SOAP; same as in panel (D). Note that angle brackets notation $\langle \ldots \rangle$ is used to denote the version of the descriptor after the application of local denoising.

Recently, it has been shown [23] that in systems with high structural heterogeneity (for example, supercooled water below the liquid–liquid phase transition [31]), the ability of SOAP to identify different molecular environments can be greatly improved by local noise reduction. SOAP spectra are often 'overloaded' with local structural information, as well as noise from surrounding environments. This issue can be mitigated by averaging the SOAP vectors of molecule i with those of its neighbors within the cutoff sphere. By doing so, the local heterogeneity within the microscopic environment surrounding each particle is smoothed, reducing noise and enhancing SOAP's ability to distinguish between different environments, particularly those with non-local differences. This noise reduction approach has recently been shown to be highly efficient, enabling, for example, the distinction of the presence and coexistence of two liquid phases in aqueous systems—an otherwise challenging task.

Figure 3(B) shows the principal component analysis (PCA) projection of the SOAP spectra of water molecules after smoothing. Compared to the raw SOAP dataset in figure 3(A), it is clear that smoothing the local noise facilitates the detection of the liquid domain in our water-ice system. Figure 3(C) shows the variance explained by the first three principal components (PCs). In the raw SOAP dataset, the sum of the first three PCs explains approximately 96, 5% of the variance, which makes it a good approximation of the entire dataset. We spatially averaged (smoothed) the SOAP spectra of each molecule with those of the neighboring molecules within a cutoff radius of 10 Å (the same used for the SOAP calculation). This cutoff distance, which is the typical one often used in the analysis of aqueous systems, has been recently demonstrated to be a well-suited one, as this is consistent with the typical characteristic length scale of the collective microscopic dynamical events characterizing the dynamics of such aqueous systems [32] and in this case is safe to exclude finite size effects. As a result, the first three components explains approximately 99.96% of the total variance (orange bars in 3(C)), with PC1 alone accounting for about 99.5%.

Recently, it has been shown [28] that performing a single-point time-series clustering (using, for example, Onion clustering) on PC1 alone typically provides more insightful information than a pattern recognition analysis applied to the entire dataset. For instance, figures 3(D)–(E) shows the results of Onion clustering on the PC1 time-series computed for both the raw (figure 3(D)) and spatially-smoothed (figure 3(E)) SOAP datasets. Onion clustering successfully identifies three distinct environments in these time-series: solid ice (dark gray), liquid water (blue), and the solid-liquid interface (red). Notably, the application of local noise reduction via spatial averaging increases the separation between the KDE peaks of these different environments, thus improving detection and classification. This effect is reflected in the PCA components, where local noise reduction boosts the explained variance in PC1 from approximately 85.8% to 99.5% in PC1 (figure 3(C)).

Furthermore, regarding the dependency of the clustering results on the time resolution (Δt), figures 3(D)–(E) shows that the interface is only detected within a specific range of Δt: $4.8 < \Delta t < 8.7$ ns in the raw trajectories, compared to $3.6 < \Delta t < 37.2$ ns in the denoised ones. The lower resolution limit is determined by the minimum observation time required to gather sufficient information to distinguish one environment from another. The upper resolution limit, which is determined by the average lifetime of the particles inside the interface, is also affected by the noise in the dataset. As the level of noise in the data increases, the overlap between the signals from different environments grows, making it more challenging to reliably classify longer segments of the trajectories into distinct environments. As shown in the plots, denoising the SOAP data shifts the upper limit to much higher values, improving the robustness of interface detection. This is due to the reduction in noise, which increases the signal-to-noise ratio in the smoothed PC1 SOAP time-series compared to the raw data. In the raw SOAP data, the noise from the two most populated environments (solid ice and liquid water) is so large that it becomes impossible to discern the interface signal for most choices of Δt.

Note that in all the analyzes conducted herein, we are dealing with MD trajectories sampled every $\delta t = 100$ ps. Previous tests demonstrated that such a temporal sampling constitutes a good compromise between avoiding oversampling issues (important especially for descriptors suffering from local noise, such as SOAP) and maintaining a sufficiently high sampling resolution to capture the information contained in the time correlations between the data (see also figure S2) [28].

Since each different descriptor D_i has its own signal-to-noise ratio, methods such as this one, based on spatial averaging, provide a viable strategy to create denoised versions of virtually any descriptor. This opens up the possibility to expand studies assessing the effect of noise on the efficiency of different descriptors in capturing the complex physics of systems like the one examined here.

2.4. Cleaning descriptors from local noise

The improvement observed in SOAP results suggests that a similar denoising approach could be applied to all previously discussed descriptors. The outcomes, displayed in figure 4, reveal that spatial averaging significantly enhances the performance of all descriptors. After local noise reduction, descriptors such as N_{neigh}, v, and q_{tet} become capable of detecting the interface as a distinct environment with efficiency comparable to (or exceeding) that of more sophisticated descriptors like LENS and SOAP. An especially intriguing result is obtained with d_5: it not only identifies the interface but also distinguishes between two subregions—one exposed to liquid and the other to ice.

Interestingly, spatial averaging yields a less pronounced benefit for LENS. In the raw data, LENS already stands out as a top-performing descriptor, with a distribution that is easily classified by Onion clustering without further processing. Consequently, spatial averaging provides only a slight improvement, similar to what is observed for SOAP. Both LENS and SOAP possess a high intrinsic signal-to-noise ratio, and while spatial averaging does increase clarity, the impact is far more limited than with simpler descriptors. Denoised LENS results can be found in figure S3.

2.5. An 'evaluation space' for comparing descriptors

Comparing the results of figure 4 with those of figures 2 and 3 it is possible to obtain precious information on how noisy the various descriptors are. In principle, in typical analyzes, the effect of noise can have various origins (e.g. thermal noise, sampling, descriptor-based). The thermal noise corresponds to the intrinsic, temperature-driven fluctuations of atoms/molecules within the ensemble (its amplitude being essentially related to the temperature). This is a constant in all the datasets compared herein, as the trajectory analyzed is always the same. Sampling issues may be due to the fact that each individual descriptor possesses its own signal-to-noise ratio, which requires its own optimal sampling/resolution to maximize the extraction of information over noise. This is, nonetheless, already accounted for by the Onion clustering analysis in our case, which reveals the optimal sampling/resolution for the information extraction [32, 33]. In this way, the results of figures 2–4 essentially contain a number of features useful to assess and compare the noise proper

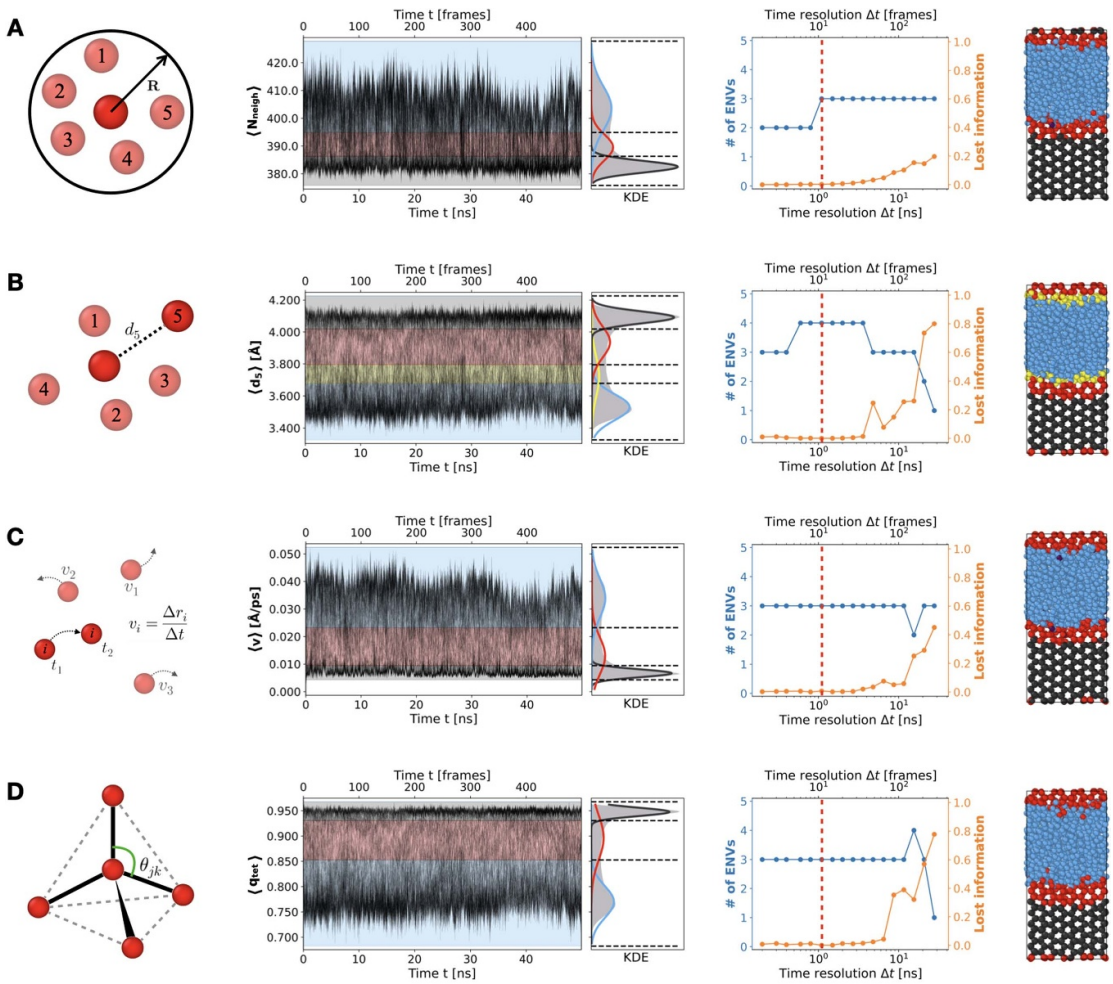

Figure 4. Comparison between descriptors after local noise reduction. For each spatially averaged descriptor — (A) number of neighbors N_{neigh}, (B) distance from the fifth atom d_5, (C) molecule velocity modulus v and (D) orientational tetrahedral order parameter q_{tet}—we show, from left to right: a schematic representation, the signal distribution over the entire simulation, clustered KDE, the number of resolved environments (blue line) and fraction of unclassified data (orange line), and a clustering-colored snapshot in correspondence to the time resolution highlighted with the red line. Note that angle brackets notation $\langle \ldots \rangle$ is used to denote the version of the descriptors after the application of local denoising.

of each different descriptor and, thus, how efficient they are in extracting information from the noisy trajectories analyzed herein as a test case.

Building on this, we leverage the extensive information provided by Onion clustering to construct an 'evaluation space' — a framework where each tested descriptor can be positioned and quantitatively compared using a data-driven metric. Similar approaches have recently been applied to, for example, compare lipid bilayers modeled with different force fields [34] or classify different types of self-assembled soft materials by assessing the similarity of their local molecular (SOAP) environments [35].

The Onion clustering method provides in output two main pieces of information. First, it reveals the fraction of data effectively classifiable as a function of time-resolution of the analysis, as $1 - f_0(\Delta t)$ (being f_0 the fraction of unclassified data points for insufficient resolution): the higher the term, the more statistically robust is the analysis, as it can rely on more data points. Second, the method unveils in how many statistically different micro-clusters the retained/analyzable data can be effectively sub-divided into: typically, the higher is the n_{env}, the more effective is the analysis. In this sense, both terms (n_{env} and $1 - f_0(\Delta t)$) are important to assess the efficiency of the analysis. While different types of approaches (e.g. normalizations, weighting) can be used to build different metrics to compare different analysis setups, here we show how even a very simple linear combination between these two parameters is informative on how good a certain descriptor is, compared to the others, in extracting and classifying information from trajectory data. In particular, we use the following metric:

$$\chi(\Delta t) = n_{\text{env}}(\Delta t) \cdot [1 - f_0(\Delta t)].$$

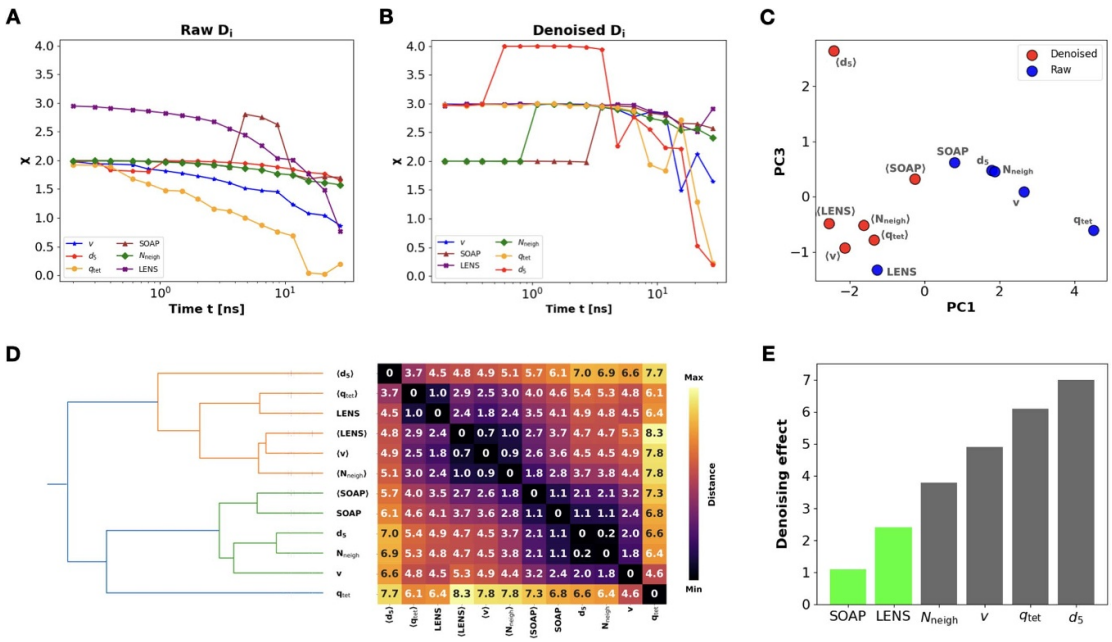

Figure 5. Evaluation space. (A) χ parameter for each raw descriptor, as a function of the time resolution Δt. (B) χ parameter for each denoised descriptor, as a function of the time resolution Δt. (C) Two dimensional projection of the 'evaluation space', using the first and the third PCs. (D) Hierarchical clustering results and distance matrix between all the descriptors studied. (E) Amplitude of the denoising effect, obtained computing the distances between raw version of the descriptors and their respective denoised one. Note that angle brackets notation $\langle \ldots \rangle$ is used to denote the version of the descriptor after the application of local denoising.

While relatively simple, this metric χ thus reflects both the fraction of data that the analysis can effectively classify and how effective the classification of the retained data into micro-clusters is—note that both parameters are directly related to the information that is effectively extracted from the data [33]. Figures 5(A) and (B) shows the χ values across different time resolutions and descriptors used.

To build an 'evaluation dataset' capturing each descriptor's performance in resolving system complexity, we selected specific features output by Onion clustering. We included seven key features: (i) the mode of the number of states identified, (ii) the number of occurrences of each state count, (iii) the time resolution at which more than 50% of the information is lost, (iv) the mean fraction of information lost before it exceeds 50%, (v) the standard deviation of the information lost fraction before it exceeds 50%, (vi) the maximum χ value, and (vii) the average χ value before the information lost fraction surpasses 50% (for a more detailed description of this dataset see SI). This process creates a multi-dimensional dataset which enables an in-depth comparison of descriptors, with flexibility to add further features as needed.

To create the 'evaluation space', we performed a PCA dimensionality reduction of this dataset. Figure 5(C) shows the first and the third PCs, color-coded to distinguish between the raw (blue) and denoised (red) versions of each descriptor. PC1 and PC3 were chosen for better visual clarity, the plots of the other PCs are shown in figure S4. This representation allows for a hierarchical clustering analysis [36, 37] based on distances between the PCA scores, enabling a comparative view of the descriptors (figure 5(D)).

It is important to emphasize that this method does not directly rank descriptors from best to worst; rather, it defines a hierarchy of similarity and difference, showing which descriptors are closely related in terms of extracted information and which ones stand apart quantitatively. In this analysis, we observe a clear distinction between denoised and raw descriptors, except for the raw LENS descriptor, which already stands out in its ability to identify the water/ice interface. At the extremes of this evaluation space, denoised d_5 and raw q_{tet} are the most distinctive descriptors, for contrasting reasons. Denoised d_5 uniquely captures the second interface across a significant range of time resolution, while q_{tet} is a descriptor with low signal-to-noise ratio that struggles to capture relevant information across the whole time resolution range.

3. Conclusions

In the study of complex many-body systems, selecting an effective analysis framework and identifying the most informative descriptors are critical steps for extracting meaningful information from inherently noisy trajectories of individual components. While choosing the optimal descriptor for a given system might be

challenging, in this work we introduced a data-driven, parameter-free approach to address this crucial aspect, which is fundamental to virtually all types of analyzes in complex system research.

To evaluate the effectiveness of different descriptors in capturing and resolving information, we exploited Onion clustering, a single-point time-series clustering algorithm with two key advantages. Through an iterative find-classify-archive approach, Onion clustering reveals all classifiable information based on the time resolution (Δt), thereby automatically identifying the optimal resolution for analysis. This also allows it to estimate the amount of data that cannot be classified in a statistically robust manner due to resolution limitations. By leveraging this information, we compared various descriptors in characterizing molecular environments in a water/ice coexistence system at the atomic level.

Our results demonstrate that general-purpose descriptors, such as LENS and SOAP, are able to identify and classify physically relevant molecular microstates more effectively than descriptors specifically tailored for this system, thanks to their inherently higher signal-to-noise ratio. Additionally, we show that performance across many descriptors can be sensibly enhanced through local denoising *via* spatial averaging. This approach boosts the signal-to-noise ratio in time-series data, enabling even simple descriptors like the number of neighbors N_{neigh} to perform comparably to more advanced ones such as SOAP and LENS, after denoising.

A quantitative comparisons of the descriptors was achieved by constructing an 'evaluation space' that enables the use of an Onion clustering-based metric to quantify the similarities between descriptors in terms of their effectiveness in resolving noisy trajectories and extracting physically meaningful information. The method is completely parameter-free and data-driven, and it does not require arbitrary choices or prior knowledge when assessing performance. Additionally, the number of features included in the evaluation dataset can be expanded by incorporating additional parameters as needed.

The results reported here also demonstrate the extent to which different descriptors can be improved through local denoising. This suggests that, for molecular systems with intricate internal structures, it is more productive to pursue a customized analysis framework than to rely on any single 'best' descriptor.

While here we demonstrate the effectiveness of the approach for one specific case (ice and water in dynamical coexistence), the implications of this work are broader. Similar dimensionality reduction methods have been recently used, for example, to create metrics to attain a data-driven comparisons and classifications in different types of problems: e.g. to compare (lipids or water) force fields [34, 38] but also to compare and classify different types of hard matter [6, 39, 40] and soft assemblies systems [4, 35] based on their similarity. Given the abstractness of the approach, we can thus speculate that this approach may have a general character, and it could be applied, in principle, to compare also how different descriptors perform with different systems.

We believe this approach offers a valuable framework for optimizing the selection of descriptors and analysis methods in the study of complex dynamical systems. Its versatility extends beyond molecular systems, making it suitable for applications involving time-series data from mesoscopic and macroscopic systems, as well as noisy trajectory datasets obtained experimentally [41–44]. The method's generality and data-driven nature thus provide a flexible tool for extracting meaningful insights across a wide range of scales and disciplines.

4. Methods

4.1. Details on the water/ice simulation

The molecular ice/water system is simulated using the direct co-existence technique, with the TIP4P/Ice water model [24]. The trajectory is obtained starting from a configuration of 50% ice 50% liquid, at $T = 268$ K. After equilibration, a production run is performed for $t = 50$ ns, sampled and analyzed every 0.1 ns. GROMACS software is used to run the simulation [45]. More detailed information can be found in [16].

4.2. Details on the descriptors

4.2.1. SOAP

The SOAP descriptor [13] provides a representation of the atomic neighbor density around a particle

$$\rho(\mathbf{r}) = \sum_{i=1}^{n_{\text{cut}}} \delta(\mathbf{r} - \mathbf{r}_i)$$

where i runs over the n_{cut} particles closer than a cutoff radius r_{cut} from the central particle (in this work $r_{\text{cut}} = 10$ Å), and $\mathbf{r}_i$ is the position of the i-th particle. $\rho(\mathbf{r})$ can be approximated in terms of radial basis

functions $g_n(r)$ and spherical harmonics $Y_{lm}(\theta,\phi)$ as

$$\rho(\mathbf{r}) \approx \sum_{n=0}^{n_{\max}-1} \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} c_{nlm} \cdot g_n(r) \cdot Y_{lm}(\theta,\phi)$$

where c_{nlm} are the expansion coefficients. In this work $n_{\max}$ and $l_{\max}$ have been set to 8. From this expansion, it is possible to compute a rotational-invariant power spectrum $\mathbf{p}$, whose components are

$$p_{nn'l} = \pi \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^\dagger c_{n'lm}$$

which contains the structural information on the particle's environment. SOAP power spectra were computed using the DScribe [46, 47] and Dynsight [48] packages.

4.2.2. LENS

The value of LENS between time t and $t+\Delta t$ is defined as

$$\delta_i(t) = \frac{\#\left(C_i^t \cup C_i^{t+\Delta t}\right) - \#\left(C_i^t \cap C_i^{t+\Delta t}\right)}{\#\left(C_i^t\right) + \#\left(C_i^{t+\Delta t}\right)}$$

and represents the set difference between the mathematical union and intersection of neighbor IDs lists present within r_{cut} (10 Å in this work) from particle i at the two consecutive time steps t and $t+\Delta t$, normalized by the total length of the neighbor ID lists. Here, C_i^t is the list of the IDs of the neighboring particles of particle i, and $\#(C_i^t)$ is its cardinality. The result is a value between 0 and 1, where 0 means that no changes happened in the neighbor list, while 1 means that all the neighbors' identities changed. Full details are available in [16]. LENS was computed using the Dynsight [48] package.

4.2.3. Orientational tetrahedral order parameter q_{tet}

This quantity has been often used [49, 50] to measures how much a local environment resembles a perfect tetrahedron; it is computed as

$$q_{\text{tet}} = 1 - \frac{3}{8} \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left(\cos\phi_{jk} + \frac{1}{3}\right)^2$$

where ϕ_{jk} is the angle between the central molecule and its neighbors j and k. q_{tet} equals 1 if the molecules form a perfect tetrahedron, while it is closer to 0 for non-tetrahedral environments.

4.2.4. Other descriptors

The use of distance from the fifth neighbor (d_5) descriptor to characterize different environments in water is shown, for instance, in [12, 51]. Number of neighbors (N_{neigh}) and the modulus of the instantaneous particles velocity (v) are widely used general descriptors in various field of application [12]. With regards to the number of neighbors, in this work $R = 10$ Å is used. A final summary of the descriptors used is provided in table 1.

4.3. The Onion clustering algorithm

Onion clustering is an algorithm for single-point clustering of time-series data. It performs a series of clustering analyzes, each one with a different time-resolution Δt, which is the minimum lifetime required for a cluster to be characterized as a stable environment. The clustering proceeds in an iterative way. At each iteration, the maximum of the cumulative distribution of data points is identified as a Gaussian state (meaning, a state characterized by the mean value and the variance of the signal inside it). The time-series signals are sliced in consecutive windows of duration Δt, and the windows close enough to the state's mean are classified as belonging to that state. These signals are then removed from the analysis, in order to enhance the resolution on the still unclassified signals at the next iteration. At the end of the process each signal windows is thus either classified in one of the identified states, or labeled as 'unclassified' at that specific time resolution. To discard statistically irrelevant clusters, populations below 1% in this study are removed from the classified data fractions.

Performing this analysis at different values of the time resolution Δt allows then to automatically identify the optimal choice of Δt that maximizes the number of environments correctly separated, and minimizes the fraction of unclassified points. Note that while Onion clustering performs well in identifying dominant and

Table 1. Brief description and dimensionality of the descriptors used in this work.

Used descriptors list		
Descriptor name	Dimensionality	Description
Distance from the 5th neighbor	Monodimensional	The distance from the target particle to the 5th distant neighbor in Å.
Number of neighbors	Monodimensional	The number of particle within a certain spherical volume fixed by a specific cutoff radius.
Particle velocities	Monodimensional	Modulus of the instantaneous particle velocities.
Orientational tetrahedral order	Monodimensional	Parameter that measures the geometrical dispositions of the first four neighbors in order to quantify the resembling with a perfect tetrahedron.
LENS	Monodimensional	Detects local diffusive fluctuations by measuring the frequencies of shuffling introduction and removal of neighbors around a defined cutoff.
SOAP	Multidimensional	Detects local structural environments encoding regions of atomic geometries by using a local expansion of a Gaussian smeared atomic density with orthonormal functions.
TimeSOAP	Monodimensional	Detects local structural fluctuations by measuring differences along the time of the SOAP descriptor power spectra.

rare states, it also presents some limitations. In particular, the division of the time-series into a fixed sequence of length Δt introduces arbitrary segmentation in the single-point clustering Δt that can affect the stability of clustering across timescales. Additionally, the fitting procedure may become less accurate for highly multivariate or non-Gaussian distributions. Complete details can be found in [17]. Onion clustering was performed using the Dynsight [48] package.

4.4. The descriptors' 'evaluation space'

PCs analysis was computed using the Scikit-learn python package [52]. Distances between PCA scores are finally classified using SciPy euclidean hierarchical clustering tool [36, 37].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.16892938 [53].

Acknowledgments

G M P acknowledges the funding received by the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 818776-DYNAPOL).

ORCID iDs

Simone Martino 0009-0009-7369-3809
Domiziano Doria 0000-0002-6176-3576
Chiara Lionello 0000-0002-7491-8952
Matteo Becchi 0000-0002-6306-5229
Giovanni M Pavan 0000-0002-3473-8471

References

[1] ten Wolde P R and Frenkel D 1997 Enhancement of protein crystal nucleation by critical density fluctuations *Science* **277** 1975–8
[2] Baletto F 2019 Structural properties of sub-nanometer metallic clusters *J. Phys.: Condens. Matter* **31** 113001
[3] Bochicchio D, Kwangmettatam S, Kudernac T and Pavan G M 2019 How defects control the out-of-equilibrium dissipative evolution of a supramolecular tubule *ACS Nano* **13** 4322–34

[4] Gasparotto P, Bochicchio D, Ceriotti M and Pavan G M 2020 Identifying and tracking defects in dynamic supramolecular polymers *J. Phys. Chem. B* **124** 589–99
[5] de Marco A L, Bochicchio D, Gardin A, Doni G and Pavan G M 2021 Controlling exchange pathways in dynamic supramolecular polymers by controlling defects *ACS Nano* **15** 14229–41
[6] Cioni M, Polino D, Rapetti D, Pesce L, Piane M D and Pavan G M 2023 Innate dynamics and identity crisis of a metal surface unveiled by machine learning of atomic environments *J. Chem. Phys.* **158** 124701
[7] Kathirgamanathan B and Cunningham P 2020 A Feature Selection Method for Multi-dimension Time-Series Data *Advanced Analytics and Learning on Temporal Data* (Springer) pp 220–31 (available at: http://dx.doi.org/10.1007/978-3-030-65742-0_15)
[8] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 Physics-inspired structural representations for molecules and materials *Chem. Rev.* **121** 9759–815
[9] Schmidt J, Piringer H, Mühlbacher T and Bernard J 2023 *Human-Based and Automatic Feature Ideation for Time Series Data: a Comparative Study* (Eurographics Association) (https://doi.org/10.2312/eurova.20231089)
[10] Nayar D, Agarwal M and Chakravarty C 2011 Comparison of tetrahedral order, liquid state anomalies and hydration behavior of mTIP3p and TIP4p water models *J. Chem. Theory Comput.* **7** 3354–67
[11] Uhrin M 2021 Through the eyes of a descriptor: constructing complete, invertible descriptions of atomic environments *Phys. Rev. B* **104** 144110
[12] Donkor E D, Laio A and Hassanali A 2023 Do machine-learning atomic descriptors and order parameters tell the same story? the case of liquid water *J. Chem. Theory Comput.* **19** 4596–605
[13] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev. B* **87** 184115
[14] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev. B* **99** 014104
[15] Nigam J, Pozdnyakov S and Ceriotti M 2020 Recursive evaluation and iterative contraction of n-body equivariant features *J. Chem. Phys.* **153** 121101
[16] Crippa M, Cardellini A, Caruso C and Pavan G M 2023 Detecting dynamic domains and local fluctuations in complex molecular systems via timelapse neighbors shuffling *Proc. Natl Acad. Sci.* **120** e2300565120
[17] Becchi M, Fantolino F and Pavan G M 2024 Layer-by-layer unsupervised clustering of statistically relevant fluctuations in noisy time-series data of complex dynamical systems *Proc. Natl Acad. Sci.* **121** e2403771121
[18] Caruso C, Crippa M, Cardellini A, Cioni M, Perrone M, Piane M D and Pavan G M 2025 Classification and spatiotemporal correlation of dominant fluctuations in complex dynamical systems *PNAS Nexus* **4** pgaf03
[19] Perrone M, Cioni M, Piane M D and Pavan G M 2025 Unsupervised tracking of local and collective defects dynamics in metals under deformation *J. Chem. Phys.* **162** 214507
[20] Caruso C, Cardellini A, Crippa M, Rapetti D and Pavan G M 2023 TimeSOAP: tracking high-dimensional fluctuations in complex molecular systems via time variations of SOAP spectra *J. Chem. Phys.* **158** 214302
[21] Crippa M, Cardellini A, Cioni M, Csányi G and Pavan G M 2023 Machine learning of microscopic structure-dynamics relationships in complex molecular systems *Mach. Learn.: Sci. Technol.* **4** 045044
[22] Butler B L, Fijan D and Glotzer S C 2024 Change point detection of events in molecular simulations using dupin *Comput. Phys. Commun.* **304** 109297
[23] Donkor E D, Offei-Danso A, Rodriguez A, Sciortino F and Hassanali A 2024 Beyond local structures in critical supercooled water through unsupervised learning *J. Phys. Chem. Lett.* **15** 3996–4005
[24] Abascal J L F, Sanz E, García Fernández R and Vega C 2005 A potential model for the study of ices and amorphous water: TIP4p/ice *J. Chem. Phys.* **122** 234511
[25] Aghabozorgi S, Shirkhorshidi A S and Wah T Y 2015 Time-series clustering – a decade review *Inf. Syst.* **53** 16–38
[26] Aminikhanghahi S and Cook D J 2017 A survey of methods for time series change point detection *Knowl. Inf. Syst.* **51** 339–67
[27] Shiraj M M B, Rahman M M, Al-Imran M, Liza M Z A, Murshed M M and Akhter N 2024 Anomaly detection in financial time series data via mapper algorithm and DBSCAN clustering *World J. Adv. Eng. Technol. Sci.* **13** 070–84
[28] Lionello C, Becchi M, Martino S and Pavan G M 2025 Relevant, hidden, and frustrated information in high-dimensional analyses of complex dynamical systems with internal noise *J. Chem. Theory Comput.* **21** 6683–97
[29] ŁBaran W R and MacDowell L G 2023 Self-diffusion and shear viscosity for the tip4p/ice water model *J. Chem. Phys.* **158** 064503
[30] Lupi L and Gallo P 2023 Glassy dynamics of water in tip4p/ice aqueous solutions of trehalose in comparison with the bulk phase *J. Chem. Phys.* **159** 154504
[31] Harrington S, Zhang R, Poole P H, Sciortino F and Stanley H E 1997 Liquid-liquid phase transition: evidence from simulations *Phys. Rev. Lett.* **78** 2409–12
[32] Doria D, Martino S, Becchi M and Pavan G M 2025 Data-driven assessment of optimal spatiotemporal resolutions for information extraction in noisy time series data *J. Chem. Phys.* **162** 234110
[33] Becchi M and Pavan G M 2025 Maximum information extraction from noisy data via shannon entropy minimization (arXiv:2504.12990)
[34] Capelli R, Gardin A, Empereur-Mot C, Doni G and Pavan G M 2021 A data-driven dimensionality reduction approach to compare and classify lipid force fields *J. Phys. Chem. B* **125** 7785–96
[35] Gardin A, Perego C, Doni G and Pavan G M 2022 Classifying soft self-assembled materials via unsupervised machine learning of defects *Commun. Chem.* **5** 1–15
[36] Bar-Joseph Z, Gifford D K and Jaakkola T S 2001 Fast optimal leaf ordering for hierarchical clustering *Bioinformatics* **17** S22–S29
[37] Müllner D 2011 Modern hierarchical, agglomerative clustering algorithms (arXiv:1109.2378)
[38] Perrone M, Capelli R, Empereur-Mot C, Hassanali A and Pavan G M 2023 Lessons learned from multiobjective automatic optimizations of classical three-site rigid water models using microscopic and macroscopic target experimental observables *J. Chem. Eng. Data* **68** 3228–41
[39] Cioni M, Piane M D, Polino D, Rapetti D, Crippa M, Irmak E A, Van Aert S, Bals S and Pavan G M 2024 Sampling real-time atomic dynamics in metal nanoparticles by combining experiments, simulations and machine learning *Adv. Sci.* **11** 2307261
[40] Rapetti D, Piane M D, Cioni M, Polino D, Ferrando R and Pavan G M 2023 Machine learning of atomic dynamics and statistical surface identities in gold nanoparticles *Commun. Chem.* **6** 143
[41] Keogh E, Lonardi S and Chiu B-c 2002 Finding surprising patterns in a time series database in linear time and space *Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (*Edmonton, Alberta, Canada*) pp 550–6 (*KDD '02*) (Association for Computing Machinery)
[42] Gupta M, Gao J, Aggarwal C C and Han J 2013 Outlier detection for temporal data: a survey *IEEE Trans. Knowl. Data Eng.* **26** 2250–67

[43] Mantegna R N and Stanley H E 1999 *Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge university press)
[44] Nagy M, Ákos Z, Biro D and Vicsek T 2010 Hierarchical group dynamics in pigeon flocks *Nature* **464** 890–3
[45] Abraham M J, Murtola T, Schulz R, Páll S, Smith J C, Hess B and Lindahl E 2015 GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers *SoftwareX* **1–2** 19–25
[46] Himanen L, Jäger M O J, Morooka E V, Canova F F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 DScribe: library of descriptors for machine learning in materials science *Comput. Phys. Commun.* **247** 106949
[47] Laakso J, Himanen L, Homm H, Morooka E V, Jäger M O J, Todorović M and Rinke P 2023 Updates to the DScribe library: new descriptors and derivatives *J. Chem. Phys.* **158** 234802
[48] dynsight: simplifies analysis of molecular dynamics simulations (available at: https://github.com/GMPavanLab/dynsight)
[49] Chau P-L and Hardwick A J 1998 A new order parameter for tetrahedral configurations *Mol. Phys.* **93** 511–8
[50] Giovambattista N, Debenedetti P G, Sciortino F and Stanley H E 2005 Structural order in glassy water *Phys. Rev. E* **71** 061505
[51] Cuthbertson M J and Poole P H 2011 Mixturelike behavior near a liquid-liquid phase transition in simulations of supercooled water *Phys. Rev. Lett.* **106** 115706
[52] Pedregosa F *et al* 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30 (available at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page)
[53] Martino S, Doria D, Lionello C, Becchi M and Pavan G M 2025 Research data supporting: A data driven approach to classify descriptors based on their efficiency in translating noisy trajectories into physically- relevant information *Zenodo* (https://doi.org/10.5281/zenodo.16892938)