Maximum information extraction via clustering and minimization of Shannon entropy

MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

ACCEPTED MANUSCRIPT • OPEN ACCESS

Maximum information extraction via clustering and minimization of Shannon entropy

To cite this article before publication: Matteo Becchi *et al* 2025 *Mach. Learn.: Sci. Technol.* in press https://doi.org/10.1088/2632-2153/ae2dbb

Maximum Information Extraction Via Clustering and Minimization of Shannon Entropy

Matteo Becchi
Department of Applied Science and Technology
Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129 Torino

Giovanni M. Pavan*
Department of Applied Science and Technology
Politecnico di Torino
Corso Duca degli Abruzzi, 24, 10129 Torino

November 18, 2025

Abstract

In the analysis of any type of system, granting maximum information extraction from its data is non-trivial. Confidence in successful information extraction typically builds on prior knowledge of the studied system or on the user's experience. However, a robust and objective criterion for ensuring maximum information extraction from data is difficult to define. Here, we introduce a data-driven approach that employs Shannon entropy as a transferable metric to assess and quantify Maximum Information Extraction (MInE) from data via their clustering into statistically-relevant micro-domains. The method is general and can be applied virtually to any type of data or system. We demonstrate its efficiency by analyzing, as a first example, time-series data extracted from molecular dynamics simulations of water and ice coexisting at the solid/liquid transition temperature. The method allows quantifying the information contained in the data distributions (time-independent component) and the additional information gain attainable by analyzing data as time-series (i.e., accounting for the information contained in data time-correlations). The different micro-domains that can be effectively resolved and classified in the system are characterized by own entropy, which are found consistent with experimentally known thermodynamic parameters. A second test case demonstrates how the MInE approach is also effective for high-dimensional datasets and clearly shows how including little informative, but noisy, extra components/features in high-dimensional analyses may be not only useless, but even detrimental to maximum information extraction. This provides a robust parameter-free approach and quantitative metrics for data-analysis, and for the study of any type of system from its data.

Keywords Information gain · Shannon entropy · Data clustering · Time-series · Data science

Introduction

Scientific data analysis often requires critical methodological choices – such as the selection of observables, resolution, and analysis parameters – that significantly influence both the extraction and interpretation of physical information [1]–[3]. These choices are frequently guided by heuristic conventions, prior experience, or implicit assumptions, which can introduce bias and compromise the robustness and reliability of the results [4], [5]. An automatic method capable of learning directly from the data the maximum information effectively extractable from them would be a breakthrough in many fields, from data science to the study of the physical behavior of complex systems. As notable examples, various algorithms have been developed to optimize specific analysis parameters [6], [7]. While effective, these methods typically focus on isolated steps in data acquisition or processing – such as, e.g., feature selection [8]–[11], clustering [12], [13], or coarse-graining [14], [15] – making it difficult to assess and compare the impact of different choices using a unified transferable metric. However, understanding the physics of complex systems or extracting the maximum amount of information from their data would require a unified framework/method that allows systematically tackling such problems in a more general, holistic, and comprehensive way.

*Corresponding author: giovanni.pavan@polito.it

Here we present a robust, data-driven approach based on information theory to optimize such methodological choices, thereby maximizing the extraction of relevant information through the clustering of data into statistically relevant micro-domains. This method, herein referred to as "Maximum Information Extraction" (MInE), is versatile and applicable to both uni- and multivariate datasets and to study the physics of virtually any type of system from its data. The MInE approach employs a Shannon entropy-based metric to quantify the information gain (i.e., entropy decrease) achieved via clustering across various cases and setups, providing a principled criterion for identifying the optimal methodological choices and analysis setup for maximizing information extraction. We demonstrate the effectiveness of MInE in maximizing the extraction of information from data using different case studies. In particular, we show how the method is particularly efficient and useful in cases that are typically non-trivial, such as in noisy datasets and time-series.

By using Shannon entropy as a transferable metric, MInE compares the information extractable as a function of the micro-clusters that can be effectively resolved. The optimal resolution [16] and best analysis setup for maximum information extraction are identified as those minimizing Shannon entropy. We demonstrate the effectiveness of MInE through two case studies. First, we compare the information extractable from univariate time-series data – obtained using different descriptors [17]–[19] from Molecular Dynamics (MD) trajectories of water and ice phases coexisting at the melting temperature – as a function of resolution and/or descriptor choice. We demonstrate how the different micro-clusters that are discovered from the data correspond to domains that are characterized by their own internal entropy – proving their physical relevance –, whose differences are found consistent with known thermodynamic quantities for aqueous systems (e.g., entropy of fusion). Second, in a model Langevin dynamics on a bi-dimensional energy landscape, we show how MInE can assess how information extraction is influenced by the interplay between the number of dimensions considered and the impact of noisy data and frustrated information phenomena [20] in multivariate analyses. These case studies highlight MInE as a physically robust and general method for optimizing data analysis and for maximizing the knowledge attainable from virtually any type of system and/or data.

Methods

Theoretical Framework

The Shannon entropy [21] of an observable x, which takes values in a set $\mathcal{X}$ with probability distribution $p(x)$, is defined as:

$$H(x) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \tag{1}$$

This quantity measures the uncertainty in x, with higher entropy indicating a broader distribution and lower entropy corresponding to more sharply concentrated distributions. By normalizing the Shannon entropy between 0 and 1 (see Methods section), we can define the information available from the data as:

$$I = 1 - H \tag{2}$$

In this way, the information contained in the data takes values in the range $0 \leq I \leq 1$. In the limit case where the data contain no uncertainty (e.g., all data points exactly equal to one specific single value), I equals to 1, while $I = 0$ corresponds to the case where data are completely random (pure entropy), containing no structure, and no statistically releyant information can be extracted from them.

In practice, the information $I_0 = 1 - H_0$ contained in the data, where H_0 is the Shannon entropy of the raw data, can be increased by clustering them into statistically distinct micro-domains. For a dataset clustered into K clusters with probability distributions $p_k(x)$, the Shannon entropy of the clustered system is:

$$H_{\text{clust}}(x) = \sum_{k=1}^{K} f_k H_k \tag{3}$$

where f_k is the fraction of data points in cluster k, and H_k is the Shannon entropy of the data within that cluster. By concavity, clustering never increases entropy, and the corresponding information gain is thus defined as:

$$\Delta I = -\Delta H = H_0 - H_{\text{clust}} \geq 0 \tag{4}$$

where H_0 is the Shannon entropy of the original data distribution.

This quantity effectively measures the reduction in uncertainty achieved via data clustering [22], and correspond to the Mutual Information [23] between the descriptor and the cluster labels. Possible alternative definitions are discussed in Fig. S5 in the SI, also showing the robustness of this definition. Based on Eq. 4, $I_{\text{clust}} = I_0 + \Delta I$. Optimizing the clustering of the data thus allows to maximize the information gain following clustering, and consequently the information effectively extractable from the data, I_{clust}. A trivial clustering – where all points belong to a single cluster, or are assigned randomly – yields $\Delta I \sim 0$, indicating little to no information gain compared to that attainable simply form the distribution of the data. In contrast, an effective clustering maximizes ΔI, maximizing the translation of the data structure into information.

Detailed information on the entropy calculation are provided in the Materials and Methods section, as well as in the SI. Our tests demonstrate that the MInE method is robust with respect to, e.g., changing the binning in the calculation of entropy (see Fig. S6), or using a different entropy estimator (e.g., the Kozachenko-Leonenko density-based

estimator – see Fig. S6). Note that, while here Shannon entropy is our primary measure for building our method, alternative information-theory based metrics (such as, when dealing with time-series, Approximate Entropy [24] or Sample Entropy [25], [26]) can also be employed within the same framework.

In short, MInE shows the maximum amount of information attainable from data, and unveils how to effectively extract it. Note that the MInE workflow thus substantially differs from entropy-minimization based clustering algorithms [27]. Shannon entropy minimization is not used in MInE to guide the clustering itself, but it is rather applied a posteriori to evaluate the effectiveness of the methodological choices for the clustering and to quantify the information attainable from it. While the MInE approach is general and can be used to reveal the maximum extractable information from virtually any type of data (e.g., static, dynamic), here we demonstrate its effectiveness by applying it to prototypical challenging cases, such as highly noisy time-series data generated by complex dynamical systems of various types.

MD simulations

Complete details on the MD simulation performed for this study are reported in the SI.

Results and Discussion

From entropy to information and back

As a first example, we tested the MInE framework to analyze MD simulation trajectories of 2048 TIP4P/ICE [28] molecules coexisting in dynamic equilibrium between solid and liquid phases at the melting temperature (Fig. 1a). After equilibration, we performed a $\tau = 50$ ns production run under NPT conditions, sampling trajectories every 0.04 ns (complete simulation details are provided in the SI). We used various single-particle descriptors to extract univariate time-series data from the MD trajectories, each capturing distinct structural and/or dynamical features of the system, with its own signal-to-noise ratio [29], [30]. For example, Figs. 1 and 2 show results obtained using the Smooth Overlap of Atomic Positions (SOAP) [17], which provides a rotationally invariant, high-dimensional representation of the local molecular density around each molecule (Fig. 1b). The SOAP power spectrum offers a fingerprint of the spatial arrangement of neighboring molecules within a sphere of radius r_c around each molecule. Fig. 2 also contains analyses on the Local Environment and Neighbors Shuffling (LENS) descriptor [19], which captures the entity and speed of the reshuffling of the neighboring molecules of each molecule.

In the case of noisy time-series data, such as these ones, the ability to detect/discriminate statistically relevant micro-domains through single point clustering depends on the resolution Δt used to segment and analyze the data [16], [20], [31]. Among the available algorithms for single point time-series analysis [32], here we use Onion clustering [31], an unsupervised method that identifies all statistically relevant micro-domains (including hidden ones) that can be resolved in a time-series of length τ as a function of Δt, as well as the fraction of unclassifiable data due to insufficient resolution. More details about Onion clustering are available in the SI. While Onion clustering is particularly well suited to for handling noisy time-series data, and it is thus chosen as a basis for the demonstrations presented herein, note that the MInE method is general and can be applied, in principle, in combination with any other clustering technique (see Fig. S1, which shows how MInE can reproduce standard clustering evaluation approaches, like the Bayesian Information Criterion [33], when available).

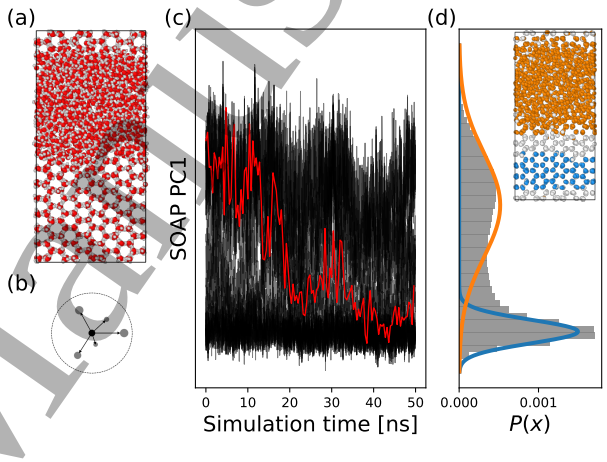

Figure 1: (a) Snapshot of the water/ice atomistic model system used as a case study. (b) Schematic of the SOAP descriptor, which provides a high-dimensional representation of the density, order/disorder, and symmetry of the arrangement of neighboring molecules around each molecule. (c) Denoised [34] PC1 SOAP time-series [20], [30] for the 2048 molecules as a function of simulation time. In red: the signal of a representative molecule undergoing a water-to-ice transition. (d) Probability distribution $P(x)$ of the SOAP PC1 signals (gray). Two main clusters, corresponding to the liquid and ice phases (inset), are identified by the two maxima in $P(x)$ (orange and blue Gaussian fits). These clusters are readily detected using pattern recognition methods (see SI for details).

In the following, we show as a representative example the results obtained by analyzing denoised [34] time-series data of the first principal component (PC1) of the SOAP vectors, which is an interesting informative descriptor for such systems [20], [30] and for the demonstrations that we are providing herein (complete description for the SOAP analysis are provided in the Supplementary Material). Data obtained with other descriptors [19] and/or by reducing the dimensionality of the SOAP spectra with other methods (e.g., via Time-lagged Independent Component Analysis [35], [36]) are reported later in the paper, and in the Supplementary Material. Fig. 1c shows the SOAP PC1 time-series data for all molecules along the trajectory (in

black), with one molecule undergoing an ice-to-water transition highlighted in red. The probability distribution $P(x)$ of the entire SOAP PC1 dataset is shown in Fig. 1d (gray). The two prominent density peaks in $P(x)$ allow typical clustering methods to readily identify two main clusters in the data. These clusters, represented by the orange and blue Gaussian curves centered on the $P(x)$ peaks in Fig. 1d, correspond to the liquid and solid phases, as shown in the snapshot inset with matching color coding.

However, additional information is nested inside these data, particularly in their temporal correlations [31]. This is evident from the results obtained by Onion clustering at smaller Δt (i.e., increasing the temporal resolution). Complete Onion clustering results are shown in Fig. S1 in the SI, showing that the number of resolvable clusters is maximized in the resolution interval 2 ns $\lesssim \Delta t \lesssim$ 20 ns. Within this range, the method identifies three statistically distinct clusters corresponding to the ice and water phases, and the ice/water interface (see Fig. S2 in the SI).

Note that results such as, e.g., the number and quality of detected clusters, as well as the fraction of unclassified data, are descriptor-dependent. Each descriptor is characterized by its own signal-to-noise ratio, and its feature space is defined by metrics that differ from those of other descriptors. As a result, it is not trivial to unambiguously infer, for example, whether the information captured by the SOAP PC1 descriptor represents the maximum extractable information, or how one descriptor compares to another. MInE overcomes this limitation by using entropy as a transferable metric and exploiting Shannon entropy minimization to assess the maximum information that can be extracted from the data following clustering, as defined by Eq. 4.

Fig. 2a shows the information I attainable from the data before and after clustering. The information contained in the data probability distribution $P(x)$ (prior to clustering) can be quantified by its Shannon entropy, H_0 (calculated via Eq. 1), as $I_0 = 1 - H_0 \sim 0.151$ (black horizontal dashed line in Fig. 2a). Since this value depends solely on the data distribution, it is independent of the time resolution Δt used in the subsequent clustering analysis. In contrast, the information content after clustering is given by $I_{\text{clust}}(\Delta t) = 1 - H_{\text{clust}}(\Delta t)$, represented by the solid black curve in Fig. 2a. The gray-shaded area between these two curves represents the total information gain extractable, in this case, by Onion clustering. Shannon entropy minimization is achieved at $\Delta t^* = 1.8$ ns (vertical red dashed line), where the maximum information $I_{\text{max}} = I_{\text{clust}}(\Delta t^*) \sim 0.41$ is attained (confidence intervals for the entropy values are shown in Fig. S3). At all resolutions, the sum of the resolved information I and the residual entropy H remains constant (by definition). Above the solid black line in Fig. 2a, the weighted Shannon entropy of the different resolved environments (micro-clusters), $f_k H_k$, is shown for each value of Δt. As seen in Fig. 2a, within the range of Δt where the ice/water interface (green area)is resolved as a distinct micro-cluster, the information I_{clust} is maximized when the fraction of unclassified data points (in violet) is minimized. To ensure that the entropy values obtained are not overfitting on the single trajectory, we performed all the analyses also on the two halves of the trajectory, obtaining quantitatively identical results, see Fig. S4.

From an information-theoretic perspective, the disappearance of the interface cluster (green) leads to an increase in the residual entropy $f_k H_k$ of the remaining clusters (Fig. 2a). Conversely, when the interface is detected, the total entropy in Eq. 3 decreases, despite the introduction of an additional term. This is because the reduction in $f_k H_k$ for the ice and liquid phases more than compensates for the added entropy contribution of the interface. Physically, the optimal range 2 ns $\lesssim \Delta t \lesssim$ 20 ns provides the best balance between resolution and noise, enabling the identification of the ice/water interface, which has a characteristic residence time of ~ 10 ns [37]. At lower time resolutions ($\Delta t \geq 20$ ns), an increasing fraction of data points remain unclassified, resulting in a raise in the residual entropy H_{clust} and a corresponding decrease in the information gain (gray area in Fig. 2a). On the other hand, at higher resolutions ($\Delta t \leq 1$ ns), the interface can no longer be resolved as a distinct environment. This occurs because excessively fine temporal segmentation emphasizes short-time molecular vibrations –interpreted as noise– at the expense of capturing the physically meaningful transitions between phases [20].

To estimate the contribution of time correlations to the total information I_{clust}, we repeated the Onion clustering analysis after randomly reshuffling the time-series frames. This procedure preserves the data probability distribution $P(x)$, while effectively removing temporal correlations, thereby eliminating any information that could be extracted from them. We performed this reshuffling ten times and computed the average values of I_{clust} as a function of Δt; the variance in the estimation was consistently below 0.1% for all Δt. The results are shown in Fig. 2a as a dotted curve. Notably, this curve is nearly horizontal, indicating that – once time correlations are removed – the extractable information becomes largely independent of the resolution Δt. Furthermore, the value of I_{clust} obtained from the reshuffled data closely matches that obtained from the original time-series when analyzed at the lowest possible resolution ($\Delta t = \tau = 50$ ns). In Fig. 2a, the difference between the I_0 baseline (dashed line) and the reshuffled I_{clust} (dotted line) represents the information gain achievable via clustering under a purely ergodic approximation – that is, neglecting time correlations. Finally, the difference between this dotted I_{clust} curve and the solid black $I_{\text{clust}}(\Delta t)$ curve (corresponding to the original time-ordered series) quantifies the additional information that is uniquely encoded in the temporal correlations – highlighted by the diagonally hatched gray area in Fig. 2a. These results are particularly interesting because, while time-correlations are kept into account by Onion Clustering, they are then ignored in the Shannon entropy calculations. Despite that, the results of Fig. 2a show that the MInE framework is capable to evaluate and quantify the extra information gained

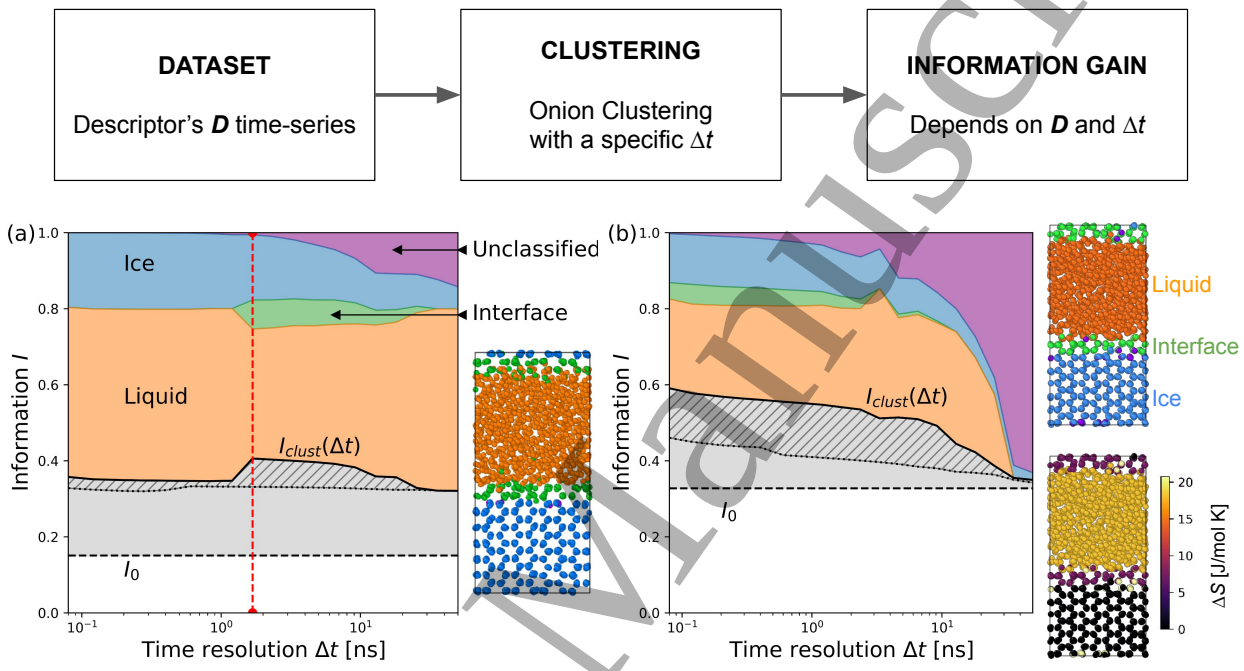

Figure 2: Information gain from Onion clustering as a function of the time resolution Δt used in the analysis. Top: schematic representation of a typical MInE workflow. Starting from a dataset (in this case, time-series of single-particle descriptors such as SOAP or LENS), clustering is performed (with Onion Clustering, setting the time resolution Δt), and the information gain for each descriptor and each Δt is evaluated. Bottom: Panel (a) shows results for the SOAP PC1 dataset with spatial averaging; panel (b) shows results for the LENS dataset. In both panels, the horizontal dashed line (I_0) represents the information content of the raw data distribution. The solid black curve (I_{clust}) represents the information retained after clustering, which varies with Δt. The gray-shaded area corresponds to the increase in information content achieved through clustering. The dotted line shows the information gain obtained when the dataset frames are randomly reshuffled, while the diagonally hatched area highlights the additional gain attributable to time correlations. The width of the colored regions reflects the weighted Shannon entropy $f_k H_k$ of each environment, illustrating how entropy varies with Δt. In panel (a), I_{clust} reaches a maximum at $\Delta t \sim 2$ ns (vertical red dashed line), indicating the optimal time resolution for extracting information from the SOAP PC1 time-series. The simulation snapshot is colored according to Onion clustering at this resolution: bulk ice and liquid water appear in blue and orange, respectively, with the solid–liquid interface in green. A small fraction of unclassifiable points appear as sparse molecules in purple (color coding matches that in the plot). On the right: The top snapshot shows the corresponding clustering, using the same color scheme as in panel (a). The bottom snapshot is the same one, colored by entropy difference calculated relative to that of the bulk of ice and converted in units of [J mol^{-1} K^{-1}].

when also considering the time correlations in the data during clustering.

Fig. 2b shows the same analysis on time-series data obtained from the same MD trajectories, but using a different single-particle descriptor, LENS [19]. LENS – which generally exhibits a higher intrinsic signal-to-noise ratio compared to SOAP [16], [30], [31], [38] – enables correct detection of the interface from the highest time resolution, $\Delta t = 0.08$ ns, up to approximately $\Delta t \sim 10$ ns.

To further assess the physical relevance of the Shannon entropy values computed in this analysis, we attempted to relate them to the thermodynamic entropy of the system under study. This comparison can be made following, for instance, the approach presented in Ref. [39], under the assumption that the chosen descriptor captures all of the system's degrees of freedom and their correlations. While this assumption may not strictly hold – since any descriptor is, to some extent, incomplete – we show below that, for the LENS data analyzed here, the results obtained are reasonable and physically meaningful. Complete details on the assumptions needed for this calculations and comparison can be found in the SI.

Absolute entropy values computed on continuous variables, such as LENS in our case, are generally not meaningful, as they depend strongly on the binning used in the Shannon entropy estimation. However, entropy differences can still carry significant physical meaning. Using the LENS signals we compute the entropy difference between the solid ice and liquid water micro-domains, obtaining a value of $\Delta S \sim 18$ J mol^{-1} K^{-1}. This is very close to the experimental entropy of fusion of water, $\Delta S_{\text{melt}} \sim 22$ J mol^{-1} K^{-1} [40] (especially considering the accuracy that can be expected from such simulations, and the description allowed by the LENS descriptor). This demonstrates the robustness of the approach and the physical significance of the results obtained, in a purely data-driven way, using the MInE method.

Noteworthy, this approach also allows us to estimate the entropy difference between any other micro-state in the system – in this case, the bulk solid ice vs. the solid–liquid interface environments, a quantity that is generally difficult to determine experimentally or otherwise. For the ice/water interface, we obtain a value of $\Delta S_{\text{interf}} \sim 7.2$ J mol^{-1} K^{-1} compared to the bulk of ice. The lower snapshot in Fig. 2b is colored according to the entropy difference between each environment vs. solid ice. This is a significant result, as it clearly highlights the physical relevance of the interface environment in such a system. The fact that this micro-cluster corresponds to a dynamically distinct molecular environment, with its own entropy – different from those of both ice and liquid water –, implies that it should not be confused with either of these two other phases. In particular, any method or analysis that fails to identify the interface environment as a separate and distinct cluster (i) loses important physical information and (ii) compromises the statistical characterization of the ice and liquid domains (especially when interface water molecules are incorrectly assigned to one of the two bulk phases [41]).

Note that other tested descriptors, which account for different degrees of freedom, yield different entropy values (although they all remain within the same order of magnitude). For example, using SOAP PC1, we obtain a ice-water $\Delta S \sim 10.7$ J mol^{-1} K^{-1}. This difference between LENS and SOAP can be rationalized by considering that these descriptors capture only part of the system's internal degrees of freedom. In particular, these results demonstrate how the difference between the solid ice and liquid water phases is better captured by differences in the local diffusivity of the molecules populating these two environments (measured by LENS) rather than by differences in their local structural vibrations (measured by SOAP).

It is worth underlining that the MInE approach is not limited to a specific clustering method (e.g., Onion clustering: Fig. 2) but it can be used in combination with virtually any other clustering technique. In typical Onion clustering analyses, depending on the chosen time resolution Δt used to segment the time-series, some data points cannot be reliably assigned to dynamically persistent micro-clusters. Such faster-scattering points are grouped into an additional cluster that collects all "unclassifiable" data. The entropy of this cluster – typically higher than that of the dynamically persistent micro-clusters due to its greater variability – must be included in the total entropy calculation to preserve probability mass. Note that this requirement is automatically fulfilled when using clustering methods that assign all data points to clusters (e.g., Gaussian Mixture Models – see Fig. S1).

The information captured by different descriptors

By assessing the maximum amount of information that can be extracted from a given dataset, MInE also enables a direct comparison, and ranking, of different descriptors based on their effectiveness in capturing relevant information from the same trajectories. This comparison is made possible by the transferable nature of the Shannon entropy–based metric used. As shown in Fig. 3a, we applied MInE to time-series data derived from different descriptors computed on the same MD trajectories of the ice/water model system. In addition, we employed alternative dimensionality-reduction techniques to analyze the high-dimensional SOAP spectra.

For this specific system, the results of these analyses, shown in Fig. 3a, reveal that some descriptors enable the conversion of MD trajectories into time-series from which significantly more information can be extracted than others. In particular, these descriptors yield higher relative information gain $\Delta I/H_0$. For example, the LENS descriptor [19] and the first time-lagged Independent Component (tIC1) [35], [36] of the SOAP spectra, when analyzed using Onion clustering, provide the highest information gain (up to time-resolutions of $\Delta t \leq 2$ ns) among the descriptors tested: $\Delta I/H_0 \sim 0.35 - 0.4$. This relative information

gain highlights the effectiveness of these descriptors in retaining and extracting physically relevant information encoded in the molecular trajectories. Note that all four descriptors used in this comparative analysis were computed with a cutoff radius of $r_c = 10$ Å, corresponding to the third solvation shell. This parameter sets the interparticle distance over which correlations, reconfigurations, and structural features are probed. As such, r_c is a fundamental parameter that determines the spatial resolution of the analysis and directly influences the amount and type of information that can be effectively captured.

Optimal resolution for maximum information extraction

Many widely used single-particle descriptors are designed to capture the behavior of neighboring units surrounding each particle in the system. The neighborhood of each molecule is typically defined as a sphere with a cutoff radius r_c, which sets the spatial scale over which correlations, rearrangements, symmetries, and other structural features are analyzed. In this way, r_c effectively determines the spatial resolution of the analysis and, consequently, the physical information that can be extracted from the system. Recent work has shown that the optimal spatial resolution for studying a given system depends on whether the dominant physical phenomena are local or collective in nature [16]. The MInE approach offers a quantitative framework to identify this optimal resolution by evaluating how much relevant information can be extracted at different values of r_c.

As a proof of concept, we tested this approach on the same MD trajectory of the TIP4P/ice solid–liquid coexistence system. We computed both LENS and SOAP descriptors using different cutoff radii, r_c, chosen to correspond to characteristic minima in the water–water radial distribution function, $g(r)$ (see Fig. 3b, top; red dots). Fig. 3b (bottom) shows how the maximum relative information gain, $\Delta I/H_0$, achievable through Onion clustering varies as a function of the cutoff radius used to compute the descriptor time-series. For both descriptors, $\Delta I/H_0$ exhibits a clear maximum at $r_c \sim$ 10–13 Å, thereby identifying this as the optimal resolution for maximizing information extraction in this system.

As discussed in detail in Ref. [16], this result reflects the characteristic length scale of the collective structural dynamics dominating aqueous systems of this kind: accurately capturing such phenomena requires including at least up to the third solvation shell. Smaller cutoff values fail to capture essential mid-range correlations, resulting in datasets dominated by local noise and fast vibrations; conversely, excessively large cutoffs cause information loss due to over-averaging. More broadly, the results shown in Fig. 3 illustrate how MInE offers a robust, quantitative framework for optimizing key analysis parameters – such as descriptor type, cutoff radius, and spatial and temporal resolutions – to ensure maximal information extraction from molecular trajectory data.

Application to multivariate datasets

MInE can also be applied to high-dimensional datasets and multivariate time-series. When using multivariate descriptors, quantifying information becomes particularly relevant in tasks such as feature selection and dimensionality reduction. These approaches aim at reduce data complexity but often involve trade-offs, as decreasing the number of variables can lead to information loss. To explore this, we tested the MInE framework on a simple multivariate model dataset. Specifically, we simulated the Langevin dynamics of 100 particles in two distinct bi-dimensional potential energy landscapes: one with four minima (A to D), requiring both (x, y) coordinates for proper identification, and another with only two minima (A and B), distinguishable using the y coordinate alone. In both cases, all minima are subject to identical Gaussian noise (Fig. 4a). We then applied the same MInE procedure as described above to four different datasets, obtained by analyzing both systems using either the full (x, y) coordinates or only the y coordinate through Onion clustering (see SI for full details on these test cases).

The information gain obtained after clustering for these model systems is shown in Fig. 4b. For the system with four energy minima (Fig. 4b, left), using both (x, y) coordinates all four minima to be identified. In contrast, using only the y coordinate does not distinguishes between A and C or between B and D, which collapse into two degenerate (and doubly-populated) A+C and B+D states. Comparing the blue and orange curves highlights that the maximum attainable information gain (reached in this system for $\Delta t < 5$ frames) is twice as high when performing a bi-dimensional analysis (blue) compared to a mono-dimensional one (orange). For the system with just two minima (Fig. 4b, right), the maximum attainable information gain is the same wether using both (x, y) or just the y coordinate (green vs. red curves): in this case, both minima are correctly identified regardless of dimensionality, making the second variable unnecessary. Importantly, the maximum information gain is identical in the orange curve (mono-dimensional analysis of the four-minima system) and the red and green curves (two-minima system), despite representing different systems and analyses. This demonstrates how Shannon entropy within the MInE framework serves as an objective, transferable, and quantitative metric for assessing and comparing the extractable information across datasets and analysis strategies.

Interestingly, in both systems, the information gain (and thus clustering performance) decreases more rapidly with increasing Δt in the bi-dimensional analyses (blue and green curves) compared to the mono-dimensional ones (orange and red). In the four-minima system, one initial hypothesis attributed this behavior to the "apparent" longer residence times of particles in the degenerate A+C and B+D states identified by the mono-dimensional analysis, compared to the shorter residence times in the individual A, B, C, and D states resolved by the bi-dimensional one. However, this explanation does not hold for the two-

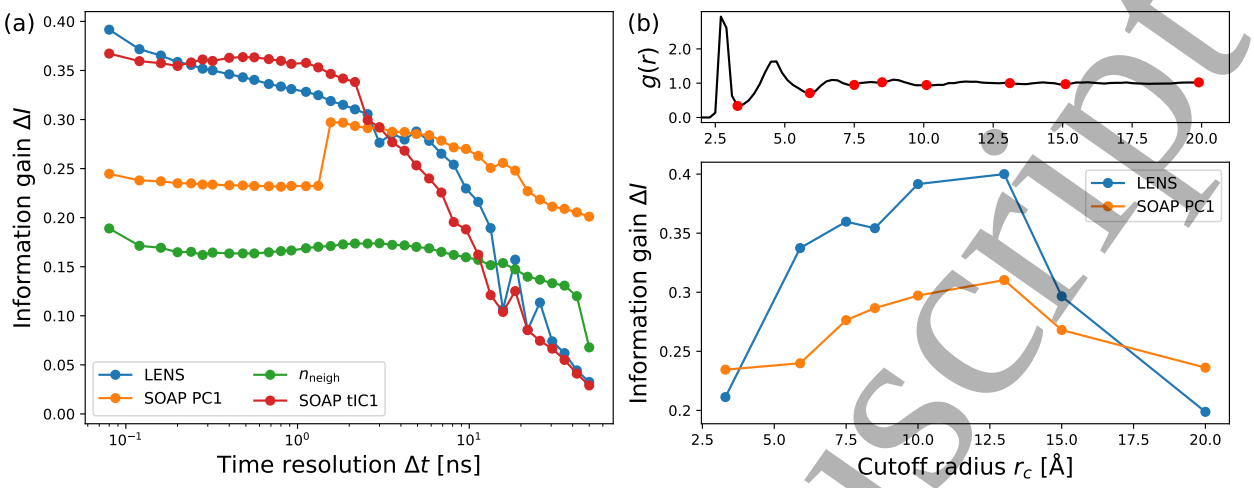

Figure 3: Optimizing descriptor choice and spatiotemporal resolutions for maximum information extraction. (a) Relative information gain $\Delta I/H_0$, obtained by applying Onion clustering to time-series data derived from different descriptors: LENS (blue), the first principal component (PC1) of SOAP (orange), the first time-lagged independent component (tIC1) of SOAP (red), and the number of neighbors n_{neigh} (green). (b) Top: Radial distribution function $g(r)$ of the water molecules (oxygen atoms only), computed over the entire simulation trajectory. Red dots mark the minima in $g(r)$, corresponding to solvation shells, used here as characteristic cutoff values r_c for computing descriptor time-series. Bottom: Maximum relative information gain $\Delta I/H_0$ attainable via Onion clustering on LENS and SOAP PC1 time-series calculated using different cutoff radii r_c.

minima system (Fig. 4b, right), where both A and B states are correctly identified in both types of analyses – yet the same faster decrease in information gain with increasing Δt is observed in the bi-dimensional case. These findings for the two-minima case point instead to the role of noise-addition phenomena that can affect high-dimensional analyses: specifically, even though the x component in this case does not provide relevant information, it still introduces noise [20]. In higher-dimensional spaces, clustering algorithms tend to leave a larger fraction of data points unclassified [20], [42], as also shown in Fig. S2 of the SI. As discussed elsewhere [20], these results underscore how including additional (noisy) dimensions may be not only unhelpful, but actually detrimental to the effectiveness of high-dimensional analyses.

To further validate these effects, we performed an additional test in both systems by introducing a third dimension z, which – like x in the previous test – does not carry any meaningful information but only contributes Gaussian noise (i.e., all minima are characterized by identical, uncorrelated noise in all directions). The gray curves in Fig. 4b,c show that while the maximum attainable information remains unchanged compared to the lower-dimensional cases, the range of Δt values over which this information is effectively extractable becomes further reduced. This confirms that adding increasingly noisy components hinders the extraction of meaningful information, particularly that embedded in the time-correlations of the data. These findings reflect a manifestation of the so-called "curse of dimensionality" [43], [44], highlighting that in clustering analyses there exists a threshold beyond which the inclusion of additional dimensions shifts from enriching the information content to merely introducing noise. MInE proves particularly valuable in this context, as it enables probing and quantifying these effects in a rigorous and interpretable way.

We underline that, while the results above focus on time-series data – where temporal correlations play a key role – the MInE approach is general and applicable to other types of datasets as well. For example, in the case of static data distributions (or in situations where time correlations are irrelevant or difficult to resolve), the I_{clust} becomes independent of the Δt. In such contexts, MInE can be thus used to identify, e.g., the most effective clustering method, or parameter set (descriptor(s)) that allows minimizing Shannon entropy and maximizing information extraction from the data distributions.

Conclusions

We introduce a general data-driven framework to quantify the information extractable from virtually any type of data and to assess Maximum Information Extraction (MInE). The method relies on Shannon entropy minimization to quantify how much structure in the data can be effectively resolved via data clustering. In a purely unbiased and data-driven way, MInE allows to optimize data analysis (e.g., the choice of the best descriptor, or of the optimal spatial and temporal resolutions) to maximize the extraction of physically meaningful information. The framework is applicable to both continuous and categorical data and pro-

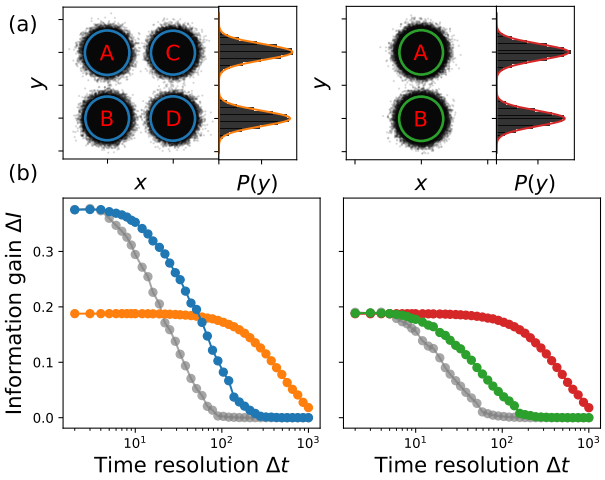

Figure 4: (a) From left to right, the trajectories of 100 particles in a bi-dimensional energy landscape with 4 and 2 minima respectively, and the probability distribution of the y coordinate $P(y)$. The clusters identified by Onion clustering in the two systems, using the full (x, y) trajectories or only the y coordinate, are shown in blue, orange, green and red respectively. (b) Information gain ΔI from Onion clustering as a function of time resolution Δt (in simulation time-steps) used for the analysis. For each of the two example model systems (left and right panels), three different datasets are analyzed: different color curves matching those in panel (a). For the system with four minima (left), clustering with (x, y) (blue curve) allows resolving all four A-D minima, yielding at maximum twice the information gain compared to that attainable when clustering using variable y only (orange: which allows resolving only two minima), but the performance degrades an order of magnitude earlier ($\Delta t \sim 10$ vs. 100 frames). For the two-minima system (right), clustering with (x, y) (green) and with y (red) achieves the same maximum gain (allowing to resolve in both cases both A-B minima), though degradation is again faster for the bi-variate case. In both systems, we added a third case (gray curve) demonstrating how adding a third z coordinate, which does not bring additional relevant information but just noise (the minima become spherical in three dimension), does not increase the attainable information gain but also accelerates analysis degradation confining it to lower Δt.

vides a transferable metric to guide analysis across diverse systems.

Here, we first test the efficiency of the MInE approach to analyze molecular dynamics simulations of solid–liquid water coexistence. Single point time-series clustering conducted by Onion clustering allows maximizing the Shannon entropy reduction following to clustering. MInE reveals the best suited descriptors (among a set of tested ones) and the optimal resolutions to maximum information extraction. We demonstrate how this maximum information extraction aligns with the robust detection of the solid–liquid interface as a distinct environment (Fig. 2).

In one shot, MInE estimates the Shannon entropy of the various detected environments, and it allows to reconduct it to thermodynamic entropy differences. As a notable example, the comparison between the entropy of the bulk ice and liquid water provides an entropy difference between the solid and liquid phases in the system that is in good agreement with the experimentally known thermodynamic entropy of fusion of water (Fig. 2b). This is a remarkable result, considered the typical accuracy expected from such simulations and the level of description allowed by one single descriptor, i.e., LENS. Any individual descriptor is, per se, incomplete to some extent, and different descriptors may provide different entropy difference estimations. This demonstrates the extent to which different representations encode physically relevant degrees of freedom and proves how MInE can be a fundamental tool to compare, assess, and rank descriptors based on their completeness and ability to maximize information extraction (Fig. 3).

Noteworthy, MInE also allows estimating the entropy of the solid–liquid interface domain relative to bulk ice (Fig. 2b), a quantity that is difficult to access experimentally. MInE proves the relevance of this micro-state in the system, demonstrating how this has its own entropy, which is different from those of ice and liquid phases. This demonstrates how MInE can not only assist the discovery of new relevant micro-states in the system, but it also reveals their relative entropy values and quantifies the information gain associated to their discovery. This is useful for validating the obtained results in the case of known systems (as in the aqueous system studied herein, for which experimental entropy variations are available). Also, this is especially useful for systems about which little is known a priori, for which MInE can reveal precious and robust quantitative information on their internal microscopic-level physics.

The results of Fig. 4 demonstrate the efficiency of MInE in analyzing also high-dimensional dataset, revealing the information loss encountered when, e.g., missing one fundamental information. This provides a robust quantitative handle to support dimensionality reduction approaches as well as feature selection to maximum information extraction.

The MInE method is simple to implement, physically interpretable, and it does not rely on any system-specific

assumptions. This provides a general framework to quantify information gain in data analysis that can be applied to a wide range of problems, from atomistic to biological systems, as well as to other complex datasets. We expect that MInE will become a fundamental framework for studying and resolving the physics of systems from their data, as well as to guide the discovery of new knowledge via data analysis in general.

Data availability statement

All the code and data necessary to reproduce the analysis of this work are available on a Zenodo repository at https://doi.org/10.5281/zenodo.15236523.

Acknowledgments

G.M.P. acknowledges the support received by the European Research Council (ERC) under the Horizon 2020 research and innovation program (grant agreement no. 818776 - DYNAPOL).

References

[1] J. Liepe, P. Kirk, S. Filippi, T. Toni, C. P. Barnes, and M. P. Stumpf, "A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation," *Nature protocols*, vol. 9, no. 2, pp. 439–456, 2014. DOI: 10.1038/nprot.2014.025.

[2] G. Bonaccorso, *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd, 2018.

[3] J. Ding, V. Tarokh, and Y. Yang, "Model selection techniques: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 6, pp. 16–34, 2018. DOI: 10.1109/MSP.2018.2867638.

[4] T. E. Sweeney, A. C. Chen, and O. Gevaert, "Combined mapping of multiple clustering algorithms (communal): A robust method for selection of cluster number, k," *Scientific reports*, vol. 5, no. 1, p. 16 971, 2015. DOI: 10.1038/srep16971.

[5] M. C. Thrun, "Distance-based clustering challenges for unbiased benchmarking studies," *Scientific reports*, vol. 11, no. 1, p. 18 988, 2021. DOI: 0.1038/s41598-021-98126-1.

[6] H. Kohjitani, S. Koda, Y. Himeno, *et al.*, "Gradient-based parameter optimization method to determine membrane ionic current composition in human induced pluripotent stem cell-derived cardiomyocytes," *Scientific Reports*, vol. 12, no. 1, p. 19 110, 2022. DOI: 10.1038/s41598-022-23398-0.

[7] B. Bischl, M. Binder, M. Lang, *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 2, e1484, 2023. DOI: 10.1002/widm.1484.

[8] W. Siedlecki and J. Sklansky, "On automatic feature selection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 02, pp. 197–220, 1988. DOI: 10.1142/S0218001488000145.

[9] A. Arauzo-Azofra, J. M. Benitez, and J. L. Castro, "Consistency measures for feature selection," *Journal of Intelligent Information Systems*, vol. 30, pp. 273–292, 2008. DOI: 10.1007/s10844-007-0037-0.

[10] A. Glielmo, C. Zeni, B. Cheng, G. Csányi, and A. Laio, "Ranking the information content of distance measures," *PNAS nexus*, vol. 1, no. 2, pgac039, 2022. DOI: 10.1093/pnasnexus/pgac039.

[11] R. Wild, E. Sozio, R. G. Margiotta, *et al.*, "Maximally informative feature selection using information imbalance: Application to covid-19 severity prediction," *Scientific Reports*, vol. 14, no. 1, p. 10 744, 2024. DOI: 10.1038/s41598-024-61334-6.

[12] M. Sugiyama, M. Yamada, M. Kimura, and H. Hachiya, "On information-maximization clustering: Tuning parameter selection and analytic solution," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 65–72.

[13] E. Aldana-Bobadilla and A. Kuri-Morales, "A clustering method based on the maximum entropy principle," *Entropy*, vol. 17, no. 1, pp. 151–180, 2015. DOI: 10.3390/e17010151.

[14] P. Gkeka, G. Stoltz, A. Barati Farimani, *et al.*, "Machine learning force fields and coarse-grained variables in molecular dynamics: Application to materials and biological systems," *Journal of chemical theory and computation*, vol. 16, no. 8, pp. 4757–4775, 2020. DOI: 10.1021/acs.jctc.0c00355.

[15] S. Y. Joshi and S. A. Deshmukh, "A review of advancements in coarse-grained molecular dynamics simulations," *Molecular Simulation*, vol. 47, no. 10-11, pp. 786–803, 2021. DOI: 10.1080/08927022.2020.1828583.

[16] D. Doria, S. Martino, M. Becchi, and G. M. Pavan, "Data-driven assessment of optimal spatiotemporal resolutions for information extraction in noisy time series data," *The Journal of Chemical Physics*, vol. 162, p. 234 110, 2025. DOI: 10.1063/5.0261449.

[17] A. P. Bartok, R. Kondor, and G. Csanyi, "On representing chemical environments," *Physical Review B - Condensed Matter and Materials Physics*, vol. 87, no. 18, p. 184 115, 2013. DOI: 10.1103/PhysRevB.87.184115.

[18] C. Caruso, A. Cardellini, M. Crippa, D. Rapetti, and G. M. Pavan, "Timesoap: Tracking high-dimensional fluctuations in complex molecular systems via time variations of soap spectra," *Journal of Chemical Physics*, vol. 158, p. 214 302, 2023. DOI: 10.1063/5.0147025.

[19] M. Crippa, A. Cardellini, C. Caruso, and G. M. Pavan, "Detecting dynamic domains and local fluctuations in complex molecular systems via time-lapse neighbors shuffling," *Proceedings of the National Academy of Sciences*, vol. 120, no. 30, e2300565120, 2023. DOI: 10.1073/pnas.2300565120.

[20] C. Lionello, M. Becchi, S. Martino, and G. M. Pavan, "Relevant, hidden, and frustrated information in high-dimensional analyses of complex dynamical systems with internal noise," *Journal of Chemical Theory and Computation*, vol. 21, no. 14, pp. 6683–6697, 2025. DOI: 10.1021/acs.jctc.5c00374.

[21] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[22] E. S. Soofi, H. Zhao, and D. L. Nazareth, "Information measures," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 75–86, 2010. DOI: 10.1002/wics.62.

[23] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[24] S. M. Pincus, "Approximate entropy as a measure of system complexity.," *Proceedings of the national academy of sciences*, vol. 88, no. 6, pp. 2297–2301, 1991. DOI: 10.1073/pnas.88.6.2297.

[25] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American journal of physiology-heart and circulatory physiology*, vol. 278, no. 6, H2039–H2049, 2000. DOI: 10.1152/ajpheart.2000.278.6.H2039.

[26] J. S. Richman, D. E. Lake, and J. R. Moorman, "Sample entropy," in *Methods in enzymology*, vol. 384, Elsevier, 2004, pp. 172–184. DOI: 10.1016/S0076-6879(04)84011-4.

[27] G. Palubinskas, X. Descombes, and F. Kruggel, "An unsupervised clustering method using the entropy minimization," in *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*, vol. 2, IEEE, 1998, pp. 1816–1818. DOI: 10.1109/ICPR.1998.712082.

[28] J. Abascal, E. Sanz, R. García Fernández, and C. Vega, "A potential model for the study of ices and amorphous water: Tip4p/ice," *Journal of Chemical Physics*, vol. 122, p. 234 511, 2005. DOI: 10.1063/1.1931662.

[29] E. D. Donkor, A. Laio, and A. Hassanali, "Do machine-learning atomic descriptors and order parameters tell the same story? the case of liquid water," *Journal of Chemical Theory and Computation*, vol. 19, no. 14, pp. 4596–4605, 2023. DOI: 10.1021/acs.jctc.2c01205.

[30] S. Martino, D. Doria, C. Lionello, M. Becchi, and G. M. Pavan, "A data driven approach to classify descriptors based on their efficiency in translating noisy trajectories into physically-relevant information," *Machine Learning: Science and Technology*, vol. 6, no. 3, p. 035 039, 2025. DOI: 10.1088/2632-2153/adfa66.

[31] M. Becchi, F. Fantolino, and G. M. Pavan, "Layer-by-layer unsupervised clustering of statistically relevant fluctuations in noisy time-series data of complex dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 121, no. 33, e2403771121, 2024. DOI: 10.1073/pnas.2403771121.

[32] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information systems*, vol. 53, pp. 16–38, 2015. DOI: 10.1016/j.is.2015.04.007.

[33] A. A. Neath and J. E. Cavanaugh, "The bayesian information criterion: Background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199–203, 2012. DOI: 10.1002/wics.199.

[34] E. D. Donkor, A. Offei-Danso, A. Rodriguez, F. Sciortino, and A. Hassanali, "Beyond local structures in critical supercooled water through unsupervised learning," *Journal of Physical Chemistry Letters*, vol. 15, no. 15, pp. 3996–4005, 2024. DOI: 10.1021/acs.jpclett.4c00383.

[35] C. R. Schwantes and V. S. Pande, "Modeling molecular kinetics with tica and the kernel trick," *Journal of chemical theory and computation*, vol. 11, no. 2, pp. 600–608, 2015. DOI: 10.1021/ct5007357.

[36] M. Hoffmann, M. Scherer, T. Hempel, *et al.*, "Deeptime: A python library for machine learning dynamical models from time series data," *Machine Learning: Science and Technology*, vol. 3, no. 1, p. 015 009, 2021. DOI: 10.1088/2632-2153/ac3de0.

[37] O. A. Karim and A. Haymet, "The ice/water interface: A molecular dynamics simulation study," *Journal of Chemical Physics*, vol. 89, no. 11, pp. 6889–6896, 1988. DOI: 10.1063/1.455363.

[38] C. Caruso, M. Crippa, A. Cardellini, *et al.*, "Classification and spatiotemporal correlation of dominant fluctuations in complex dynamical systems," *PNAS nexus*, vol. 4, no. 2, pgaf038, 2025. DOI: 10.1093/pnasnexus/pgaf038.

[39] Q.-J. Hong and Z.-K. Liu, "Generalized approach for rapid entropy calculation of liquids and solids," *Physical Review Research*, vol. 7, no. 1, p. L012030, 2025. DOI: 10.1103/PhysRevResearch.7.L012030.

[40] D. R. Lide, *CRC handbook of chemistry and physics: a ready-reference book of chemical and physical data*. CRC press, 1995.

[41] M. Crippa, A. Cardellini, M. Cioni, G. Csányi, and G. M. Pavan, "Machine learning of microscopic structure-dynamics relationships in complex molecular systems," *Machine Learning: Science and Technology*, vol. 4, no. 4, p. 045 044, 2023. DOI: 10.1088/2632-2153/ad0fa5.

[42] I. Assent, "Clustering high dimensional data," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 4, pp. 340–350, 2012. DOI: 10.1002/widm.1062.

[43] N. Kouiroukidis and G. Evangelidis, "The effects of dimensionality curse in high dimensional knn search," in *2011 15th panhellenic conference on informatics*, IEEE, 2011, pp. 41–45. DOI: 10.1109/PCI.2011.45.

[44] N. Altman and M. Krzywinski, "The curse (s) of dimensionality," *Nat Methods*, vol. 15, no. 6, pp. 399–400, 2018. DOI: 10.1038/s41592-018-0019-x.