Structural comparison of homomolecular systems on surfaces using a fingerprint-based method

William Margerit, Nathalie Tarrat, Juan Cortés, Cathy Maugis-Rabusseau

The Journal of Chemical Physics

AIP Publishing

Structural Comparison of Homomolecular Systems on Surfaces Using a Fingerprint-based Method

William Margerit,[1,2] Nathalie Tarrat,[2, a)] Juan Cortés,[1, b)] and Cathy Maugis-Rabusseau[3, c)]

[1)] *LAAS-CNRS, Université de Toulouse, CNRS, F-31400 Toulouse, France*

[2)] *CEMES, Université de Toulouse, CNRS, 31055 Toulouse, France*

[3)] *Institut de Mathématiques de Toulouse; UMR5219; Université de Toulouse; CNRS; INSA, F-31077 Toulouse, France*

(Dated: 18 April 2025)

This work presents an adaptation of the Smooth Overlap of Atomic Positions (SOAP) method to improve the efficiency of (dis)similarity quantification in homogeneous molecular (homomolecular) systems. SOAP, a fingerprint-based approach, is widely used to measure molecular similarity. We propose variants of SOAP kernels that leverage the structural architecture of homomolecular systems to minimize irrelevant comparisons of atomic environments. To evaluate its performance, we apply this adapted SOAP-based method to a synthetic dataset consisting of two identical tripeptides deposited on a copper surface, simulating different molecular states. The results demonstrate that the adapted method not only improves computational efficiency but also yields more meaningful clustering outcomes by better capturing the key structural differences between states. These findings suggest that the proposed method is well-suited for the study of homomolecular systems, particularly those involving surface interactions, and has the potential to enhance the use of diverse types of molecular modeling and analysis methods that rely on (dis)similarity measures.

[a)] Electronic mail: nathalie.tarrat@cemes.fr

[b)] Electronic mail: juan.cortes@laas.fr

[c)] Electronic mail: cathy.maugis@insa-toulouse.fr

I. INTRODUCTION

Homogeneous molecular systems, also called homomolecular (HM) systems, composed of molecules of a single type, are ubiquitous. This is the case of diverse types of natural and synthetic polymeric materials. For example, the self-assembly capabilities of polypeptides and polysaccharides are exploited for the creation of novel biomaterials with remarkable properties in terms of biocompatibility and biodegradability that can be exploited for applications in tissue engineering, drug delivery, or for the development of biosourced materials[1–3]. Homomolecular systems are also important in the life sciences. A clear example are phospholipids that aggregate to form cell membranes[4]. Another relevant instance are aggregates of amyloid-β peptides, which have been extensively studied due to their association with neurodegenerative diseases such as Alzheimer's disease[5]. We can also mention systems composed of molecules bonded to inorganic materials. For instance, the self-assembly of peptides on metal surfaces, such as gold or silver, enables the creation of biofunctionalized surfaces with potential applications in biosensors[6], catalysis[7], and electronic devices[8]. Additionally, the functionalization of nanoparticles with peptides can enhance their stability[9], biocompatibility[10], and targeting capabilities[11], making them highly effective for diagnostics[12], therapy[13] or theranostics[14].

The investigation of all these systems is usually based in multi-scale approaches. At the atomic or sub-atomic scale, theoretical studies use to focus on subsets of the whole system involving a few molecules[15–19]. One of the challenges in this context is the quantification of the (dis)similarity between states of the system. This is required for the analysis of results obtained from molecular dynamics simulation techniques[20,21] or other conformational exploration/prediction methods[22,23]. Moreover, some of these methods require distances or (dis)similarity measures to guide/bias the exploration[24–27]. Efficient approaches to assessing (dis)similarity between states are also essential for the implementation of predictive and generative machine/deep learning methods[23,28].

When studying a single molecule, the most widely used approach for quantifying the (dis)similarity between two states is to perform a structural alignment minimizing the root mean square deviation (RMSD) of the atomic coordinates[29,30]. The main drawback of this technique is that finding the optimal structural alignment can be computationally expensive. Several approaches have been proposed to address this issue, using alternative descriptors

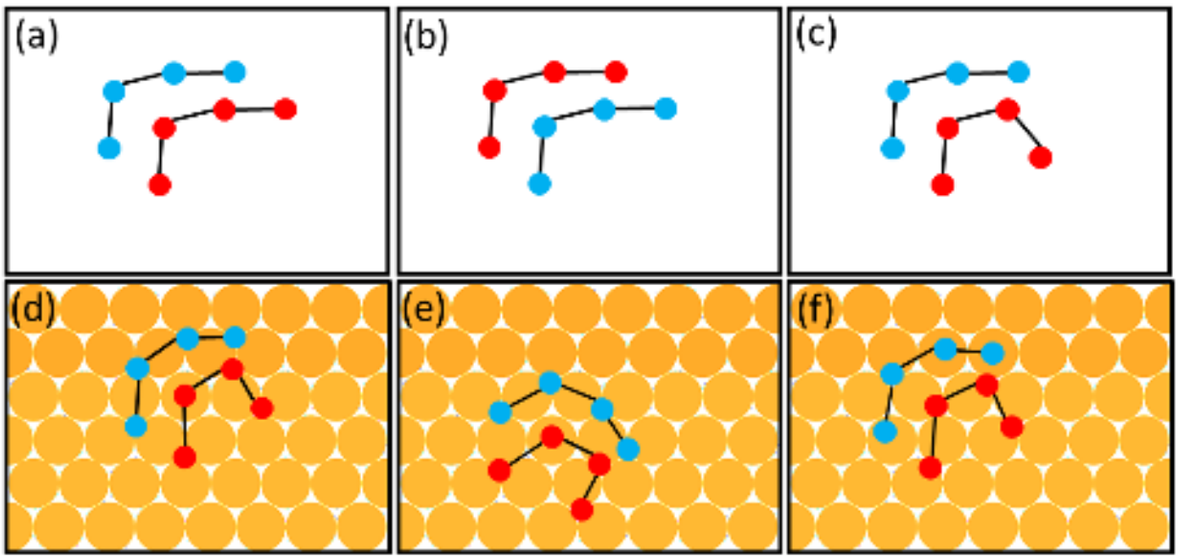

FIG. 1: Toy examples illustrating different types of difficulties encountered when comparing states of homogeneous molecular systems. The images represent states of a dimer composed of two identical molecules, one in blue and the other in red. Top: toy dimer systems in vacuum. Dimer states in (a) and (b) are equivalent by molecular permutation, while the dimer state in (c) is different from the first two due to a conformational change of one of the molecules. Bottom: toy dimer systems on a metal surface. Dimer states in (e) and (f) are two different states obtained by rigid body motion (rotation and translation) of the state in (d). The state in (e) is equivalent to (d) since it respects the periodicity of the surface, unlike the state in (f).

instead of atomic coordinates and relying on dimensionality reduction methods[21,29]. However, additional problems arise when homomolecular systems are involved. One of them is that the result of the (dis)similarity quantification should be invariant to the permutation of molecules. In addition, if we consider systems where molecules interact with a surface, (dis)similarity assessment should take into account symmetries induced by the periodicity of the surface. These two problems are illustrated with toy examples in Figure 1.

Extending RMSD-based metrics to take account of invariances due to permutations and periodicity is a challenge, and probably not the most efficient solution. A different type of methods is based on the concept of *fingerprints*. This idea was initially developed for the analysis of the local environment of atoms, with different types of applications such as structural classification or the development of energy/scoring functions, and is now a key

element of machine learning methods in this context[31].

Various types of local descriptors and associated (dis)similarity measures have been proposed on the basis of graph theory, theoretical physics concepts or statistical tools such as kernel methods[29,32,33]. One of the main advantages of these methods is that they are intrinsically invariant to permutations and periodicity.

Here, we focus on a fingerprint-based method called Smooth Overlap of Atomic Positions (SOAP), originally proposed for fitting potential energy surfaces[33]. SOAP represents atomic environments through smooth continuous functions of the local atomic density. These functions are then compared using kernel methods, where a kernel quantifies the similarity between atomic environments by integrating the overlap of their densities. By summing these kernel values across pairs of atoms in two conformational states, SOAP provides a robust measure of structural similarity that accounts for both geometric and chemical details (e.g. atom types). Several variants of the method, using different types of kernels, have been proposed to compare the structure of molecules and bulk materials[34]. The principles of the method will be briefly explained in Section II A.

We propose an adaptation of dissimilarities based on SOAP to the case of homomolecular systems. To do this, we exploit knowledge of the architecture of a homomolecular system to adapt operations and reduce the number of environment comparisons to those relevant to this type of system. The main aim is to reduce "noise" by neglecting the irrelevant environment comparisons. While the term noise can have various interpretations in different contexts, in this article, we use it to refer specifically to the irrelevant contributions in the kernel-dissimilarity that do not provide meaningful information for comparing molecular states. In addition, this also reduces the computational cost of the method. Details are provided in Section II B.

Although the proposed approach can be applied to homomolecular systems in general, we are particularly interested in molecules deposited on surfaces. For the evaluation of the methods, we considered a system consisting of two identical tripeptides (His-Pro-Phe) deposited on a copper surface. We generated a synthetic dataset based on four classes of states of this system (details are provided in Section III A). We compared the two adapted dissimilarities with the two original SOAP-based dissimilarities using our synthetic dataset, and we studied the impact of such dissimilarities in a hierarchical clustering process. Results are presented in Section III.

II. METHODS

A. Summary of SOAP fingerprints and the derived dissimilarity measures

We begin by remembering the main outlines of the SOAP[33,34] kernel construction. This kernel corresponds to the so-called *fingerprint* descriptor. It is based on the power spectrum of the environment of each atom in a system. For this summary, let us consider a system X composed of one molecule with $n(X)$ atoms. This system X can admit several states S, and X_i^S denotes the position of the i-th atom of the system X in the state S. The SOAP kernel being based on the atomic neighborhoods, we denote by $\mathcal{X}_i^S$ the atom environment around the i-th atom in the state S of X.

The first step consists of defining the atomic neighbor density function of the environment $\mathcal{X}_i^S$ by the sum of Gaussian functions with variance σ^2 centered on each of the neighboring atoms j of the central atom X_i^S by

$$\rho_{\mathcal{X}_i^S}(\mathbf{r}) = \sum_{j \in \mathcal{X}_i^S} \exp\left(-\frac{\|X_j^S - \mathbf{r}\|^2}{2\sigma^2}\right), \ \forall \mathbf{r} \in \mathbb{R}^3. \tag{1}$$

Then the similarity of two atomic environments can be defined as the inner product

$$\langle \rho_{\mathcal{X}_i^S}(\mathbf{r}), \rho_{\mathcal{X}_j^{S'}}(\mathbf{r}) \rangle = \int_{\mathbb{R}^3} \rho_{\mathcal{X}_i^S}(\mathbf{r}) \rho_{\mathcal{X}_j^{S'}}(\mathbf{r}) d\mathbf{r}. \tag{2}$$

In order to have a rotational invariant similarity kernel, Equation (2) is integrated over all the three-dimensional rotations $\hat{R}$:

$$\tilde{K}(\mathcal{X}_i^S, \mathcal{X}_j^{S'}) = \int \left| \int \rho_{\mathcal{X}_i^S}(\mathbf{r}) \rho_{\mathcal{X}_j^{S'}}(\hat{R}\mathbf{r}) d\mathbf{r} \right|^2 d\hat{R}. \tag{3}$$

This kernel is then normalized to obtain the SOAP kernel

$$K(\mathcal{X}_i^S, \mathcal{X}_j^{S'}) = \frac{\tilde{K}(\mathcal{X}_i^S, \mathcal{X}_j^{S'})}{\sqrt{\tilde{K}(\mathcal{X}_i^S, \mathcal{X}_i^S) \tilde{K}(\mathcal{X}_j^{S'}, \mathcal{X}_j^{S'})}}.$$

This SOAP kernel K can be expressed as an inner-product. To do this, each point $\mathbf{r} \in \mathbb{R}^3$ is represented by its spherical coordinate $(|\mathbf{r}|, \hat{\mathbf{r}})$ and the atomic neighbor density function can be expanded as

$$\rho_{\mathcal{X}_i^S}(\mathbf{r}) = \sum_{b \in \mathbb{N}} \sum_{\ell \in \mathbb{N}} \sum_{m=-\ell}^{\ell} C_{b\ell m}^{\mathcal{X}_i^S} \, g_b(|\mathbf{r}|) \, Y_{\ell m}(\hat{\mathbf{r}}), \tag{4}$$

where $(g_b(.))_b$ is a set of orthogonal radial basis functions and $(Y_{\ell m}(.))_{\ell m}$ is a basis of spherical harmonics[35]. Then the rotationally invariant power spectrum is defined by

$$p(\mathcal{X}_i^S)_{b_1 b_2 \ell} = \pi \sqrt{\frac{8}{2\ell+1}} \sum_{m=-\ell}^{\ell} (C_{b_1 \ell m}^{\mathcal{X}_i^S})^\dagger C_{b_2 \ell m}^{\mathcal{X}_i^S}, \ \forall (b_1, b_2, \ell) \in \mathbb{N}^3,$$

and the elements of the power spectrum are collected into a unit-length vector $\mathbf{p}(\mathcal{X}_i^S)$. Using (4) in (3) allows to reformulate the SOAP kernel as a inner-product

$$K(\mathcal{X}_i^S, \mathcal{X}_j^{S'}) = \mathbf{p}(\mathcal{X}_i^S) \cdot \mathbf{p}(\mathcal{X}_j^{S'}) \tag{5}$$

and a distance, based on the SOAP kernel K, can be deduced as

$$d(\mathcal{X}_i^S, \mathcal{X}_j^{S'}) = \sqrt{2 - 2K(\mathcal{X}_i^S, \mathcal{X}_j^{S'})}. \tag{6}$$

When the system contains multiple atomic species, a separate density of atomic neighbors is computed as in (1) for each atomic species, and the SOAP kernel is defined by matching separately the different species in (5) as explained by De et al.[34] (Equation 15). To avoid cumbersome notation, we do not mention the pairs of atomic-species indexes in (5) and in the definition of the four kernels in the following. In practice, the fingerprint is approximated by truncating $\mathbf{p}(\mathcal{X}_i^S)$ according to two thresholds ℓ_{max} and b_{max}[34,36]. In the following, we retain the same notation for readability reasons.

Based on the SOAP kernel, a similarity kernel can be constructed to compare two states S and S' of a system X by the average of all the similarities of each possible environment pairing $K(\mathcal{X}_i^S, \mathcal{X}_j^{S'})$:

$$K_{\text{AS}}(X^S, X^{S'}) \ = \ \frac{1}{n(X)^2} \sum_{i=1}^{n(X)} \sum_{j=1}^{n(X)} K(\mathcal{X}_i^S, \mathcal{X}_j^{S'})$$

$$= \left[\tfrac{1}{n(X)} \textstyle\sum_{i=1}^{n(X)} \mathbf{p}(\mathcal{X}_i^S)\right] \cdot \left[\tfrac{1}{n(X)} \textstyle\sum_{j=1}^{n(X)} \mathbf{p}(\mathcal{X}_j^{S'})\right]. \tag{7}$$

This kernel therefore induces a dissimilarity as in Equation (6) and it is robust at any atom position permutation.

Another possibility is to consider the so-called *best-match kernel* K_{BM}[34], which identifies the best match between the environments of two states

$$K_{\text{BM}}(X^S, X^{S'}) = \max_{\pi} \left(\frac{1}{n(X)} \sum_{i=1}^{n(X)} K\left(\mathcal{X}_i^S, \mathcal{X}_{\pi(i)}^{S'}\right) \right) \tag{8}$$

where the maximum is evaluated over all permutations π of $\{1, \ldots, n(X)\}$. This kernel can be computed using the Munkres algorithm[37], with a computational complexity of $\mathcal{O}(n(X)^3)$. This process ensures that only the most relevant fingerprint comparisons contribute to the kernel calculation. However, K_{BM} is not guaranteed to be a positive definite kernel.

B. Extensions to homomolecular systems

In our framework, we assume that a homomolecular system X is composed of $N(X)$ identical molecules, all composed of the same number of atoms $\eta(X)$. Then, $X^S_{a,i}$ corresponds to the position of the i-th atom of the molecule a in the system X in its state S, and the environment of this atom is denoted $\mathcal{X}^S_{a,i}$. With this notation, the total number of atoms in system X is $n(X) = N(X)\eta(X)$. In the following, we also denote X_i the position of the i-th atom among the $n(X)$ atoms without worrying about its belonging to one of the $N(X)$ molecules.

In this context, the kernel K_{AS} defined in Equation (7) is extended for two states S and S' of a homomolecular system X by

$$\begin{aligned} K_{\text{AS}}(X^S, X^{S'}) &= \frac{1}{N(X)^2\eta(X)^2} \sum_{a,b=1}^{N(X)} \sum_{i,j=1}^{\eta(X)} K(\mathcal{X}^S_{a,i}, \mathcal{X}^{S'}_{b,j}) \\ &= \frac{1}{N(X)^2\eta(X)^2} \sum_{a,b=1}^{N(X)} \sum_{i,j=1}^{\eta(X)} \mathbf{p}(\mathcal{X}^S_{a,i}) \cdot \mathbf{p}(\mathcal{X}^{S'}_{b,j}). \end{aligned} \tag{9}$$

However, since in our framework the molecules composing the system are identical, we propose a homomolecular kernel ($K_{\text{HM-AS}}$) that only averages the similarities between the environments of equivalent atoms in molecules a and b:

$$\begin{aligned} K_{\text{HM-AS}}(X^S, X^{S'}) &= \frac{1}{N(X)^2\eta(X)} \sum_{a,b=1}^{N(X)} \sum_{i=1}^{\eta(X)} K(\mathcal{X}^S_{a,i}, \mathcal{X}^{S'}_{b,i}) \\ &= \frac{1}{N(X)^2\eta(X)} \sum_{a,b=1}^{N(X)} \sum_{i=1}^{\eta(X)} \mathbf{p}(\mathcal{X}^S_{a,i}) \cdot \mathbf{p}(\mathcal{X}^{S'}_{b,i}) \\ &= \frac{1}{N(X)^2\eta(X)} \sum_{a=1}^{N(X)} \sum_{b=1}^{N(X)} \mathbf{P}(X^S_a) \cdot \mathbf{P}(X^{S'}_b) \\ &= \frac{1}{N(X)^2\eta(X)} \left[\sum_{a=1}^{N(X)} \mathbf{P}(X^S_a)\right] \cdot \left[\sum_{b=1}^{N(X)} \mathbf{P}(X^{S'}_b)\right]. \end{aligned} \tag{10}$$

where $\mathbf{P}(X_a^S) = (\mathbf{p}(\mathcal{X}_{a,i}^S); i \in \{1, ..., \eta(X)\})$.

Compared with K_{AS}, $K_{\text{HM-AS}}$ avoids irrelevant environment comparison information. By deleting these comparisons, which add noise to the evaluation of dissimilarities between two states of a homomolecular system, the $K_{\text{HM-AS}}$ kernel should give better results.

In the same spirit, the best-match kernel K_{BM} can be used to compare two states of a homomolecular system X as follows:

$$\begin{aligned} K_{\text{BM}}(X^S, X^{S'}) &= \max_{\pi} \left(\frac{1}{n(X)} \sum_{i=1}^{n(X)} K(\mathcal{X}_i^S, \mathcal{X}_{\pi(i)}^{S'}) \right) \\ &= \max_{\pi} \left(\frac{1}{n(X)} \sum_{i=1}^{n(X)} \mathbf{p}(\mathcal{X}_i^S) \cdot \mathbf{p}(\mathcal{X}_{\pi(i)}^{S'}) \right) \end{aligned} \tag{11}$$

where the maximum is evaluated on all the permutations of $\{1, \ldots, n(X)\}$. Note that in Equation (11), the fact that an atom belongs to one of the molecules is not taken into account. To overcome the potential computational cost associated with the K_{BM} kernel and to take into account the homomolecular context, we propose to adapt this kernel, relying instead on the best match between molecule fingerprints. Thus, we define the following kernel

$$K_{\text{HM-BM}}(X^S, X^{S'}) = \max_{\Pi} \left(\frac{1}{n(X)} \sum_{a=1}^{N(X)} \mathbf{P}(X_a^S) \cdot \mathbf{P}(X_{\Pi(a)}^{S'}) \right) \tag{12}$$

where the maximum is evaluated on all the permutations Π of $\{1, \ldots, N(X)\}$. Here the kernel searches for the best index match between molecules.

Note that, when studying a homomolecular system on a surface, we can use the dissimilarities associated with the four kernels defined above. The atoms on the surface are used to calculate the density of the environment for each atom in the system.

III. RESULTS AND DISCUSSION

A. Case study presentation

A dataset was generated using a homomolecular system composed of two identical polypeptides, namely His-Pro-Phe. This system was positioned on a copper surface Cu(111) consisting of three copper layers, each containing 256 atoms. This polypeptide was chosen

for its propensity to self-assemble on metallic surfaces. Firstly, four initial states were created for this system, denoted $X^{t.r:0}$ with $t \in \{1, \ldots, 4\}$. These states were derived from two different minimized structures of the monomer, called *min1* and *min2* (see Figure 2), which were obtained by an exploration of the conformational energy landscape using the IGLOO algorithm[27]. States $X^{1.r:0}$ and $X^{2.r:0}$ were both constructed using two *min1* structures with different relative positions, State $X^{3.r:0}$ was composed of one *min1* structure and one *min2* structure, and State $X^{4.r:0}$ was built with two *min2* structures. Secondly, for each initial state $X^{t.r:0}$, two additional states were generated, denoted as $X^{t.p:0}$ and $X^{t.d:0}$: State $X^{t.p:0}$ was derived from $X^{t.r:0}$ by a permutation of chain labels, while State $X^{t.d:0}$ was induced by a displacement of molecules on the surface that respects the periodicity of the surface.

Then, 20 collision-free conformations were randomly generated $\{X^{t.u:1}, \ldots, X^{t.u:20}\}$ from each of the 12 root states $X^{t.u:0}$ using a local sampling. More precisely, sampled states were obtained by perturbing dihedral angles of the backbone in a range of ± 0.5 rad, adjusting the position of the molecule (centered on the carbon α of the proline residue) in a square of 1 Å for each molecule, and altering their orientations along the three axes of rotation within ± 0.5 rad. For collision tests, the radius of all atoms was set to 0.75% of their van der Waals radius.

This simulation workflow is illustrated in Figure 2. In summary, the simulation workflow generated a sample $\{X^{t.u:v}; t \in \{1, \ldots, 4\}, u \in \{r, p, d\}, v \in \{0, \ldots, 20\}\}$ of size $n = 252$. These states were divided into $L = 4$ classes, according to the initial state. This partition is denoted $\mathcal{Z} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_4\}$ in the following.

B. Kernel computation and performance

By definition of the implemented kernels, the first step consists of computing the fingerprint of each state. For each His-Pro-Phe molecule, only the heavy atoms of the polypeptide backbone and the side chain of the proline were considered.

The fingerprint of each state was computed using the `DScribe`[36] Python library version 2.0.1, taking into account the species-dependent information (see Section II A). The cutoff for local region r_{cut} was set to 10 Å, to take into account some atoms of the second molecule and of the surface in the fingerprint of the atoms of the first molecule. To limit the impact of the surface on the fingerprint of each atom, only a subset of the atoms in the first copper

K_{AS}	$K_{\text{HM-AS}}$	K_{BM}	$K_{\text{HM-BM}}$
$1.60\ 10^{-2} \pm 0.001$	$1.17\ 10^{-3} \pm 0.00009$	$1.74\ 10^{-2} \pm 0.0008$	$2.53\ 10^{-3} \pm 0.0002$

TABLE I: Average CPU time ($\pm$ standard deviation) in seconds for the computation of the 31626 elements of the dissimilarity matrix for each kernel.

(Cu) layer of the surface were considered. More precisely, this subset is the union of the three closest Cu atoms to each atom in the molecule. The number of radial basis functions was set to $b_{max} = 10$ and the parameters σ^2 and l_{max} were fixed to the default values (1 and 6 respectively) in the `DScribe` library. Once the vectors $\mathbf{p}(\mathcal{X}_i^S)$ were computed, the dissimilarities for each pair of states associated to the four kernels K_{AS}, $K_{\text{HM-AS}}$, K_{BM} and $K_{\text{HM-BM}}$ were calculated according to Equations (9), (10), (11) and (12), respectively.

The computing time (CPU time, in seconds) using each of the four kernels is presented in Table I. This table presents the average and the standard deviation over the $\frac{n(n-1)}{2} = 31626$ elements of each dissimilarity matrix, using a single core of a 11$^{\text{tm}}$ Gen Intel® Core™ i9-11950H @ 2.60GHz processor. These results clearly show the important computational gain of the proposed extensions with respect to the general kernels. $K_{\text{HM-AS}}$ is around 14 times faster than K_{AS}, and $K_{\text{HM-BM}}$ 7 times faster than K_{BM}. Moreover, kernels adapted for homomolecular systems are more stable (much lower standard deviation). This is due to the very small number of fingerprint comparisons required by the homomolecular kernels, which also enables more optimal processing and avoids noise fluctuations in dissimilarity calculations. As expected, best-match-based kernels are more expensive, since they require to solve an optimization problem.

Nevertheless, as it will be shown below, they provide more accurate results.

C. Comparison of representative conformations

We performed a comparative analysis of the dissimilarities associated to the four kernels on the 12 root states $(X^{t.u:0})_{t\in\{1,\dots,4\},u\in\{r,d,p\}}$ of our simulated dataset. The four dissimilarity matrices are presented in Figure 3. The zero diagonal blocks for the four dissimilarities allow us to check that the kernels are indeed invariant by permutation or displacement. Recall that the states of Classes 1 and 2 are close because they are built with two *min1* structures.

By virtue of its composition and the positioning of the molecules, the states of Class 3 are between those of Classes 1-2 and those of Class 4. This structural proximity, denoted by $\mathcal{Z}_1 \prec \mathcal{Z}_2 \prec \mathcal{Z}_3 \prec \mathcal{Z}_4$, is found in the four dissimilarity matrices. This can also be seen in Figure 4, where the values available in each dissimilarity matrix in Figure 3 are normalized by the maximal value for each kernel. The kernel K_{AS} has more difficulty distinguishing between the states of $\mathcal{Z}_1$ and $\mathcal{Z}_2$. The kernel $K_{\text{HM-BM}}$, adapted for homomolecular system, takes on slightly more distinct values when comparing the different classes. In addition, we find with varying degrees of distinction that the states of $\mathcal{Z}_3$ are "halfway" between those of the other classes by construction of synthetic roots.

Based on the complete dissimilarity matrices (comparison of the 252 states), the distribution of the dissimilarities between two classes ($\mathcal{Z}_t$ versus $\mathcal{Z}_{t'}$) is represented in Figure 5. For each kernel, we obtain the same variation of the dissimilarities between states of the same root ($t' = t$) and find values in line with class order $\mathcal{Z}_1 \prec \mathcal{Z}_2 \prec \mathcal{Z}_3 \prec \mathcal{Z}_4$. The kernels K_{BM} and $K_{\text{HM-BM}}$ enable us to better distinguish between the states of Classes 1, 2 and 3, compared with the kernels K_{AS} and $K_{\text{HM-AS}}$.

D. Clustering performance

In order to study the impact of the four dissimilarities on a clustering process of the states of a homomolecular system, a hierarchical ascending clustering (HAC) with Ward's linkage measure was applied on the simulated dataset. The associated dendrograms for the four kernels are represented in Figure 6. HAC with the complete linkage and the average linkage were also tested, the results are reported in Figures S2 and S4 of the Supplementary Material, respectively.

We used three indicators to compare the clusterings obtained by cutting the dendrograms from 2 to 15 clusters with the true partition of the data $\mathcal{Z}$: Adjusted Rand Index (ARI)[38], homogeneity and completeness[39]. The mathematical definitions of these three indicators are reminded in Section A of the Supplementary Material. The ARI is based on the contingency table between a clustering and the true partition $\mathcal{Z}$, and takes values in $[-1, 1]$. The closer the value is to 1, the better the match between the clustering and $\mathcal{Z}$. The homogeneity and completeness indicators, taking values in $[0, 1]$, are based on the notion of conditional entropy. The first one measures if clusters are made up of a majority of individuals from

the same class, the second one assesses if the majority of individuals of the same class are assigned to the same cluster. These three indicators provide an overview of the link between a clustering and the true partition. The results for the four kernels are given in Figure 7.

In the sense of the ARI criterion, the clustering closest to the true partition $\mathcal{Z}$ is a clustering into three clusters for the kernels K_{AS}, $K_{\text{HM-AS}}$ and K_{BM}, whereas the kernel $K_{\text{HM-BM}}$ retrieves the true partition into four classes. This closest clustering is plotted below the dendrogram for each kernel, along with the true partition ("True labels") and the clustering into eight clusters.

The homogeneity and completeness indicators show that the HAC with the kernels $K_{\text{HM-AS}}$ and K_{BM} produce clusterings with fairly similar behavior to the true partition. They achieve an homogeneity of over 0.8 from a clustering with at least six clusters. Using the dendrogram, we can see that these two kernels recover $\mathcal{Z}_3$ and $\mathcal{Z}_4$, but form clusters where the states of $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are mixed. For the clustering with three clusters, states of $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are grouped together in the same cluster.

With the kernel K_{AS}, all three indicators show that the clusterings obtained by cutting the dendrogram are poorly matched to the true partition. The states of $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are very difficult to distinguish in the hierarchy. In addition, elements of $\mathcal{Z}_3$ and $\mathcal{Z}_4$ are related to states of $\mathcal{Z}_1$ and $\mathcal{Z}_2$. The clustering closest to the true partition in the ARI sense is composed of three clusters: the first contains all states of $\mathcal{Z}_1$ and $\mathcal{Z}_2$ (except 1) and 17% of states of $\mathcal{Z}_3$, the second contains 79% of $\mathcal{Z}_3$ states, 17% of $\mathcal{Z}_4$ states and one state of $\mathcal{Z}_2$; the third contains 82% of $\mathcal{Z}_4$ elements and 2 states of $\mathcal{Z}_3$.

The HAC obtained with the average and complete linkages (see Sections B and C of the Supplementary Material) lead to fairly similar conclusions. The clustering most in line with the true partition is obtained with the kernel $K_{\text{HM-BM}}$. The kernels $K_{\text{HM-AS}}$ and K_{BM} find $\mathcal{Z}_3$ and $\mathcal{Z}_4$ fairly well, but have some difficulty distinguishing $\mathcal{Z}_1$ and $\mathcal{Z}_2$. Results with K_{AS} are further from the true partition.

IV. CONCLUSION

This study presents a significant advancement in the application of the SOAP method for homomolecular systems. By introducing an adaptation that avoids irrelevant comparisons of atomic environments, we demonstrated improved computational efficiency and enhanced

accuracy of molecular state differentiation. The results from a dataset involving two identical tripeptides deposited on a copper surface show that the adapted kernels give better (dis)similarity measurements than the original kernels. It should be noted that although the method was presented on a specific system, it can be applied to any type of molecule and surface without any additional information on their properties other than the position and type of the atoms composing them. Its ability to better quantify structural differences, being robust to molecular permutations and symmetries introduced by the periodicity of the surface, can be exploited within numerous molecular modeling and analysis methods. The promising results pave the way towards the evaluation of the dissimilarity between more complex, and thus more realistic, systems.

Finally, we should mention that our approach builds on the original linear SOAP kernel, which effectively captures three-body structural correlations. While recent extensions of the SOAP framework[40,41] aim to incorporate higher-order invariant interactions, integrating these more expressive descriptors into our methodology is not straightforward. Future work will be required to explore such adaptations, with particular attention to maintaining computational efficiency, an essential requirement for the large-scale evaluation of dissimilarity matrices.

V. SUPPLEMENTARY MATERIAL

Definition of three indicators for clustering performances (homogeneity, completeness and ARI indicators) in Section A. Hierarchical clustering results obtained with the complete and average linkages (Sections B and C respectively).

VI. DATA AND SOFTWARE AVAILABILITY

The dataset and the software used in this work to compute the dissimilarity matrices are available at: https://gitlab.laas.fr/moma/methods/analysis/soap_hm

VII. CONFLICT OF INTEREST STATEMENT

The authors have no conflicts to disclose.

VIII. ACKNOWLEDGMENTS

This work was supported by the "Institut National des Sciences Appliquées de Toulouse" through the DEFIANT project.

REFERENCES

[1]K. Rajagopal and J. P. Schneider, "Self-assembling peptides and proteins for nanotechnological applications," Curr. Opin. Struct. Biol. **14**, 480–486 (2004).

[2]K. Sato, M. P. Hendricks, L. C. Palmer, and S. I. Stupp, "Peptide supramolecular materials for therapeutics," Chem. Soc. Rev. **47**, 7539–7551 (2018).

[3]S. Djalali, N. Yadav, and M. Delbianco, "Towards glycan foldamers and programmable assemblies," Nat. Rev. Mater. **9**, 190–201 (2024).

[4]T. Harayama and H. Riezman, "Understanding the diversity of membrane lipid composition," Nat. Rev. Mol. Cell. Biol. **19**, 281–296 (2018).

[5]M. P. Murphy and H. LeVine III, "Alzheimer's disease and the amyloid-beta peptide," J. Alzheimers Dis. **19**, 311–323 (2010).

[6]B. Yang, D. J. Adams, M. Marlow, and M. Zelzer, "Surface-mediated supramolecular self-assembly of protein, peptide, and nucleoside derivatives: From surface design to the underlying mechanism and tailored functions," Langmuir **34**, 15109–15125 (2018), pMID: 30032622, https://doi.org/10.1021/acs.langmuir.8b01165.

[7]T. Sun, Y. Feng, J. Peng, Y. Hao, L. Zhang, and L. Liu, "Cofactors-like peptide self-assembly exhibiting the enhanced catalytic activity in the peptide-metal nanocatalysts," J. Colloid. Interface Sci. **617**, 511–524 (2022).

[8]S. Sek, A. Tolak, A. Misicka, B. Palys, and R. Bilewicz, "Asymmetry of electron transmission through monolayers of helical polyalanine adsorbed on gold surfaces," J. Phys. Chem. B **109**, 18433–18438 (2005), pMID: 16853373, https://doi.org/10.1021/jp052157p.

[9]S. Locarno, R. Bucci, E. Impresari, M. L. Gelmi, S. Pellegrino, and F. Clerici, "Ultrashort peptides and gold nanoparticles: Influence of constrained amino acids on colloidal stability," Front. Chem. **9** (2021), 10.3389/fchem.2021.736519.

[10]K. K and P. MH., "Role of functionalized peptides in nanomedicine for effective cancer therapy," Biomedicines **16**, 202 (2024).

[11]J. Dai, M. Ashrafizadeh, A. R. Aref, G. Sethi, and Y. N. Ertas, "Peptide-functionalized, -assembled and -loaded nanoparticles in cancer therapy," Drug Discov. Today **29**, 103981 (2024).

[12]J. Zong, S. L. Cobb, and N. R. Cameron, "Peptide-functionalized gold nanoparticles: versatile biomaterials for diagnostic and therapeutic applications," Biomater. Sci. **5**, 872–886 (2017).

[13]R. Sharma, S. J. Borah, Bhawna, S. Kumar, A. Gupta, P. Singh, V. K. Goel, R. Kumar, and V. Kumar, "Functionalized peptide-based nanoparticles for targeted cancer nanotherapeutics: A state-of-the-art review," ACS Omega **7**, 36092–36107 (2022), https://doi.org/10.1021/acsomega.2c03974.

[14]S. Kim, Y. H. No, R. Sluyter, K. Konstantinov, Y. H. Kim, and J. H. Kim, "Peptide-nanoparticle conjugates as a theranostic platform," Coord. Chem. Rev. **500**, 215530 (2024).

[15]A. M. Brown and D. R. Bevan, "Molecular dynamics simulations of amyloid -peptide: Tetramer formation and membrane interactions," Biophys. J. **111**, 937–949 (2016).

[16]D. Costa, L. Savio, and C.-M. Pradier, "Adsorption of amino acids and peptides on metal and oxide surfaces in water environment: A synthetic and prospective review," J. Phys. Chem. B **120**, 7039–7052 (2016), pMID: 27366959, https://doi.org/10.1021/acs.jpcb.6b05954.

[17]X. Fenouillet, M. Benoit, and N. Tarrat, "On the role of intermolecular interactions in stabilizing aunp@ampicillin nano-antibiotics," Materialia **4**, 297–309 (2018).

[18]S. Abb, N. Tarrat, J. Cortés, B. Andriyevsky, L. Harnau, J. C. Schön, S. Rauschenbach, and K. Kern, "Carbohydrate self-assembly at surfaces: STM imaging of sucrose conformation and ordering on Cu(100)," Angew. Chem. Int. Ed. **58**, 8336–8340 (2019).

[19]M. Khavani, A. Mehranfar, and M. R. K. Mofrad, "On the interactions of peptides with gold nanoparticles: effects of sequence and size," J. Biomol. Struct. Dyn. **42**, 4429–4441 (2024).

[20]R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L. P. Wang, T. J. Lane, and V. S. Pande, "MDTraj: A modern open library for the analysis of molecular dynamics trajectories," Biophys. J. **109**, 1528–1532 (2015).

[21]A. Conev, M. M. Rigo, D. Devaurs, A. F. Fonseca, H. Kalavadwala, M. V. de Freitas, C. Clementi, G. Zanatta, D. A. Antunes, and L. E. Kavraki, "EnGens: a computational

framework for generation and analysis of representative protein conformational ensembles," Brief. Bioinform. **24**, bbad242 (2023).

[22]Nicy, J. W. R. Morgan, and D. J. Wales, "Energy landscapes for clusters of hexapeptides," J. Chem. Phys. **161**, 054112 (2024).

[23]H. Jung, L. Sauerland1, S. Stocker, K. Reuter, and J. T. Margraf, "Machine-learning driven global optimization of surface adsorbate geometries," npj Comput. Mater. **9**, 114 (2023).

[24]L. Jaillet, F. J. Corcho, J.-J. Pérez, and J. Cortés, "Randomized tree construction algorithm to explore energy landscapes," J. Comput. Chem. **32**, 3464–3474 (2011).

[25]C.-A. Roth, T. Dreyfus, C. H. Robert, and F. Cazals, "Hybridizing rapidly exploring random trees and basin hopping yields an improved exploration of energy landscapes," J. Comput. Chem. **37**, 739–752 (2016).

[26]S. Nandi, S. R. McAnanama-Brereton, M. P. Waller, and A. Anoop, "A tabu-search based strategy for modeling molecular aggregates and binary reactions," Comput. Theor. Chem. **1111**, 69–81 (2017).

[27]W. Margerit, A. Charpentier, C. Maugis-Rabusseau, J. C. Schön, N. Tarrat, and J. Cortés, "IGLOO: An iterative global exploration and local optimization algorithm to find diverse low-energy conformations of flexible molecules," Algorithms **16**, 476 (2023).

[28]C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, and K. F. Jensen, "Generative models for molecular discovery: Recent advances and challenges," Wiley Interdiscip. Rev. Comput. Mol. Sci. **12**, e1608 (2022).

[29]A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, and S. Goedecker, "Metrics for measuring distances in configuration spaces," J. Chem. Phys. **139**, 184118 (2013).

[30]I. Kufareva and R. Abagyan, "Methods of protein structure comparison," in *Homology Modeling: Methods and Protocols*, edited by A. J. W. Orry and R. Abagyan (Humana Press, Totowa, NJ, 2012) pp. 231–257.

[31]F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Physics-inspired structural representations for molecules and materials," Chem. Rev. **121**, 9759–9815 (2021).

[32]F. Pietrucci and W. Andreoni, "Graph theory meets ab initio molecular dynamics: Atomic structures and transformations at the nanoscale," Phys. Rev. Lett. **107**, 085504 (2011).

[33]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).

[34]S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).

[35]K. Kaufmann and W. Baumeister, "Single-centre expansion of gaussian basis functions and the angular decomposition of their overlap integrals," J. Phys. B: At. Mol. Opt. Phys. **22**, 1 (1989).

[36]L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," Comput. Phys. Commun. **247**, 106949 (2020).

[37]H. W. Kuhn, "The hungarian method for the assignment problem," Nav. Res. Logist. Q. **2**, 83–97 (1955).

[38]L. Hubert and P. Arabie, "Comparing partitions," J. Classif. **2**, 193–218 (1985).

[39]A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (2007) pp. 410–420.

[40]V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," Chem. Rev. **121**, 10073–10141 (2021).

[41]F. Bigi, S. N. Pozdnyakov, and M. Ceriotti, "Wigner kernels: body-ordered equivariant machine learning without a basis," J. Chem. Phys. **161** (2024).

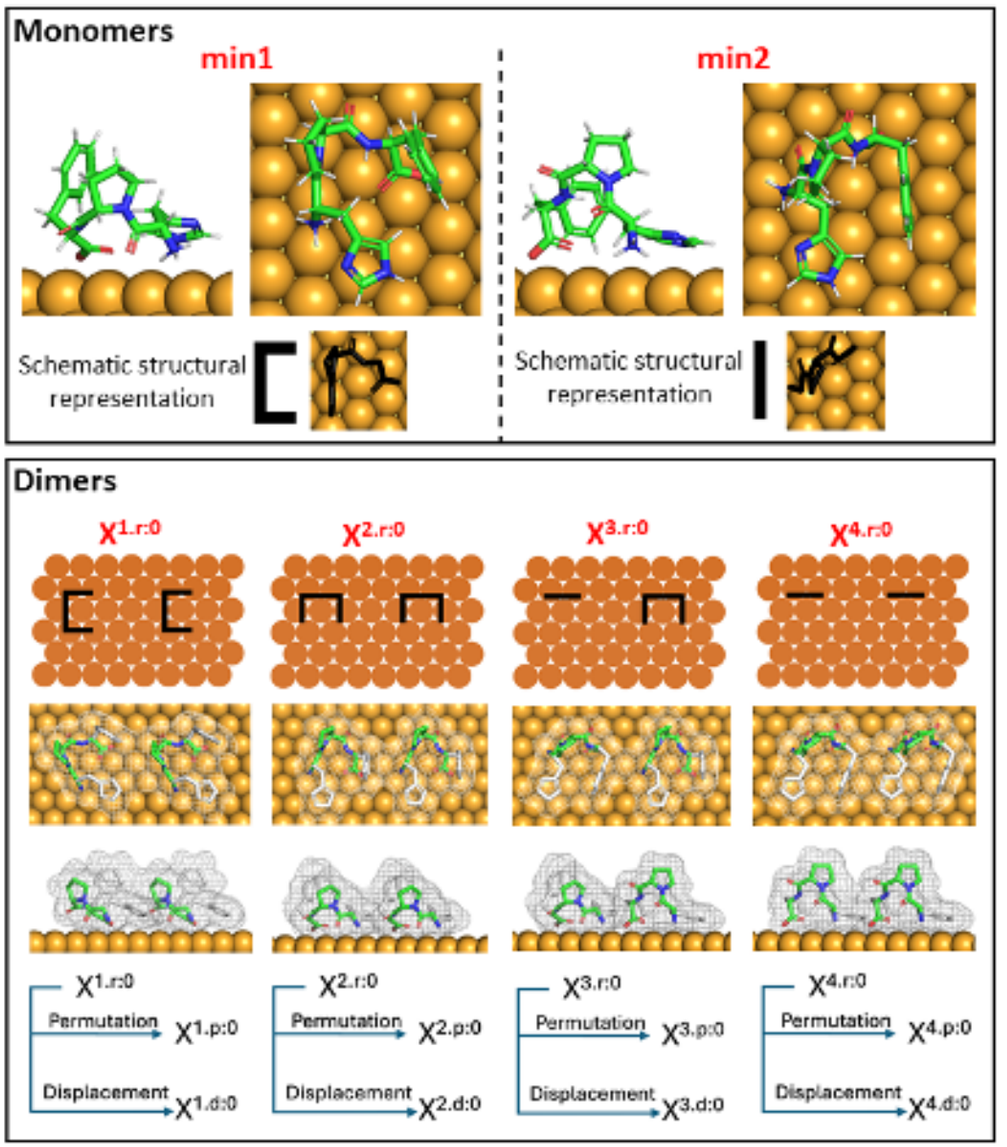

FIG. 2: Illustration of the protocol to generate the 12 root states $X^{t.r:0}$ of our dataset. Top: Two structures of His-Pro-Phe monomer on a copper surface Cu(111), called *min1* and *min2*, obtained by a global optimization algorithm, together with their associated schematic structural representation. Bottom: Schematic representation of the four initial homomolecular dimer states ($X^{t.r:0}$) based on *min1* and *min2* structures positioned on a copper surface Cu(111), top and side views of each dimer, and notation of their permuted ($X^{t.p:0}$) and displaced states ($X^{t.d:0}$).

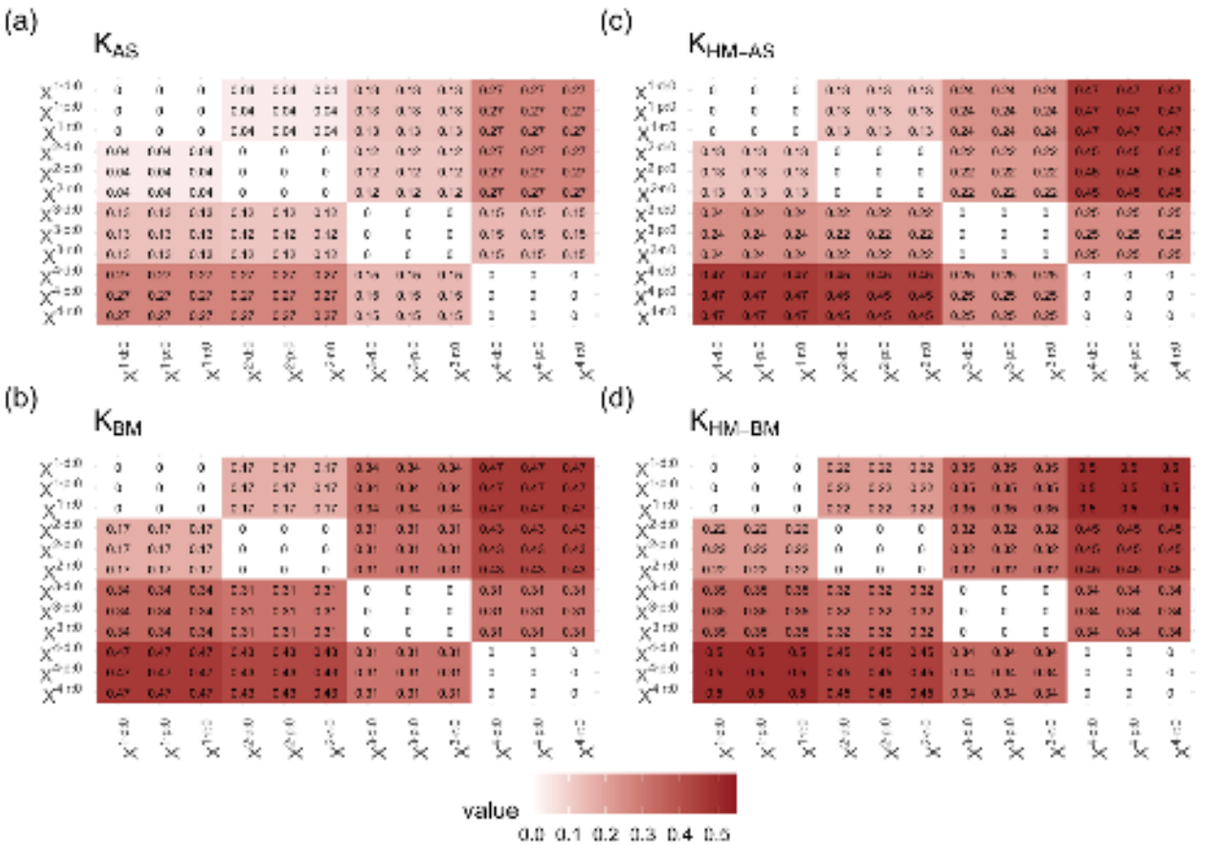

FIG. 3: Dissimilarity matrix between the 12 root states based on the kernel K_{AS} (a), K_{BM} (b), $K_{\text{HM-AS}}$ (c) and $K_{\text{HM-BM}}$ (d), respectively.

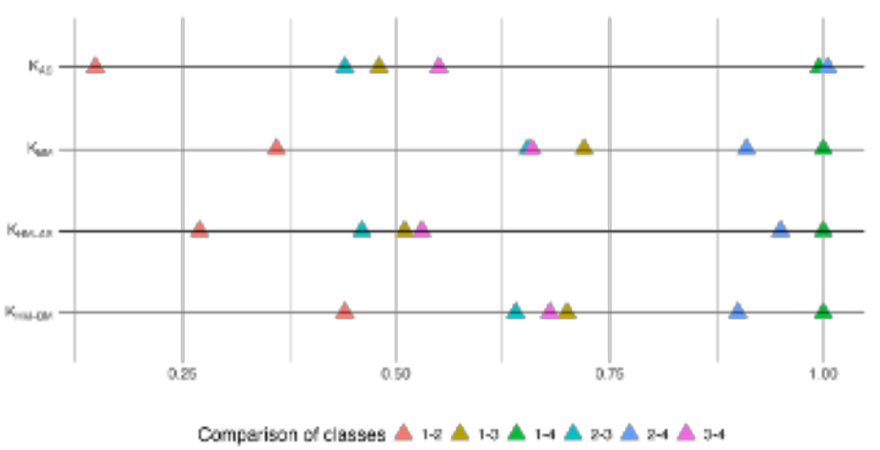

FIG. 4: Values available in each dissimilarity matrix between the 12 roots states (see Figure 3), normalized by the maximal value for each kernel, in order to compare the four classes.

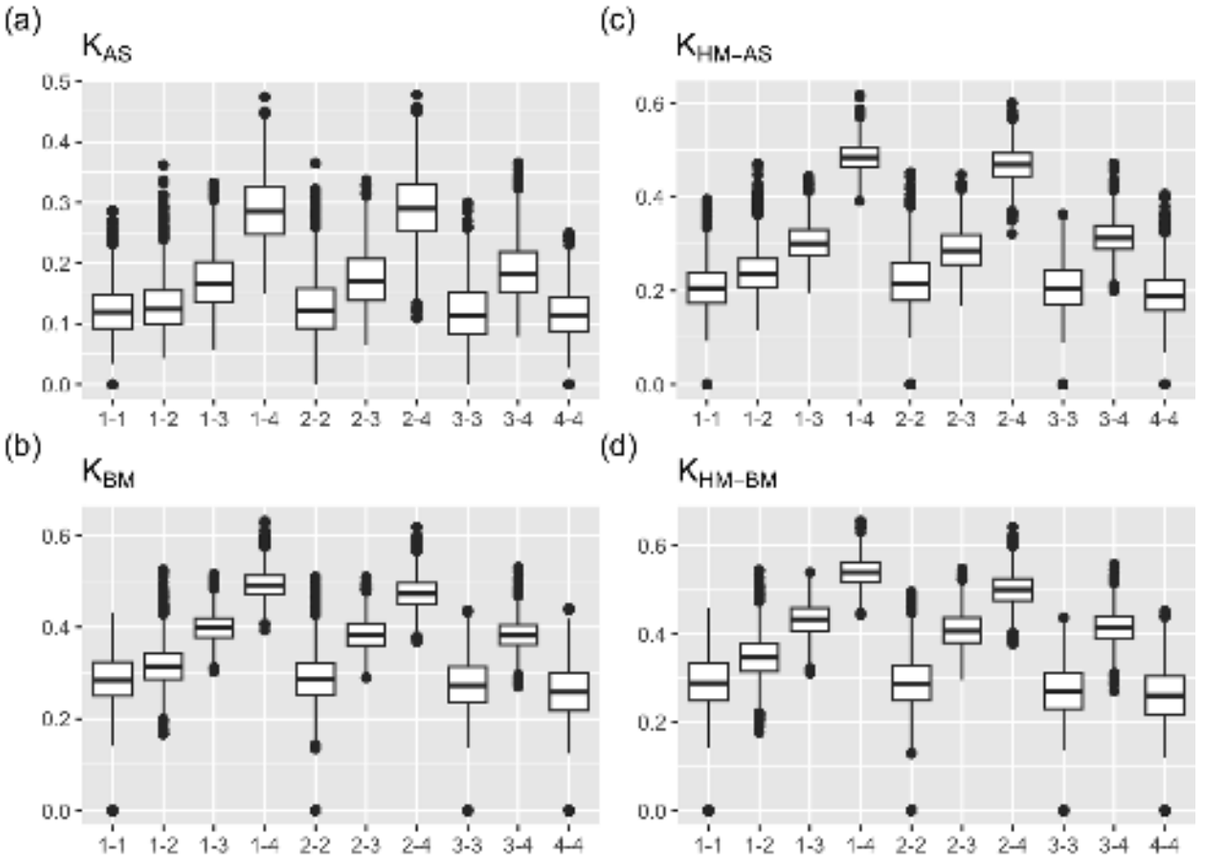

FIG. 5: Boxplot of the dissimilarities between states $\left(X^{t.u:v}, X^{t'.u':v'}\right)_{u,u' \in \{r,p,d\}, v,v' \in \{0,\ldots,20\}}$ for each couple $(t, t') \in \{1, \ldots, 4\}^2$ based on the four kernels.

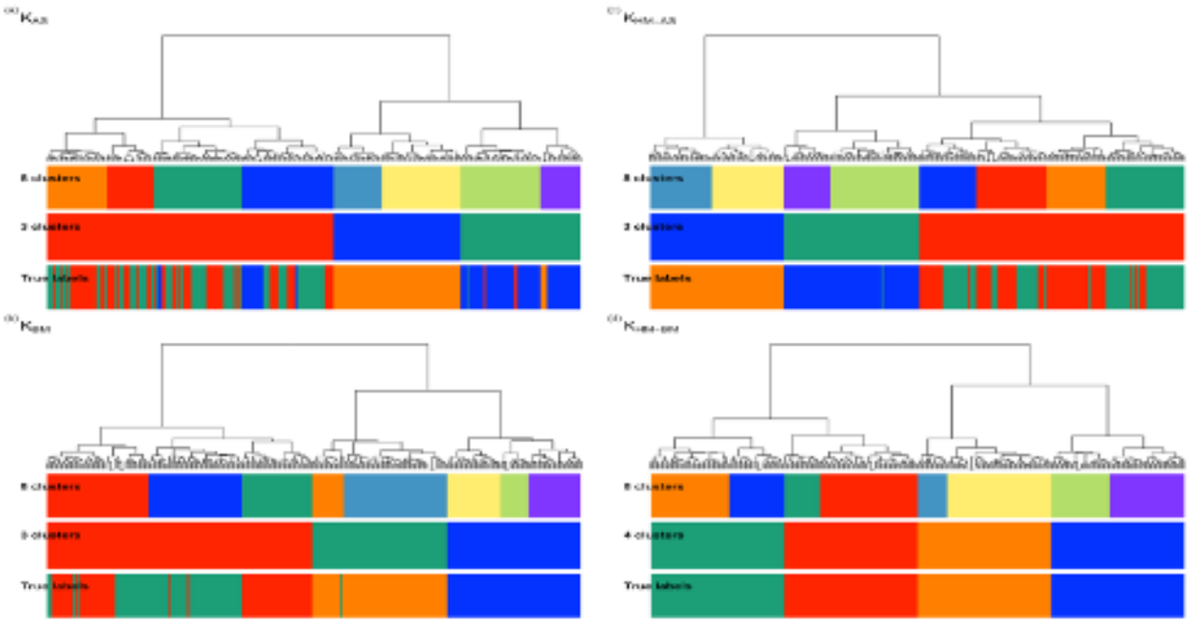

FIG. 6: Dendrograms of hierarchical clustering (HAC) with the Ward's linkage for the four kernels. Below, the distribution of states according to the clustering with 8 clusters, the clustering closest to the true partition according to the ARI indicator, and the true partition $\mathcal{Z}$ of individuals. On the "True labels" line, the states of Class 1 are in red, those of Class 2 in green, Class 3 in blue and Class 4 in orange.

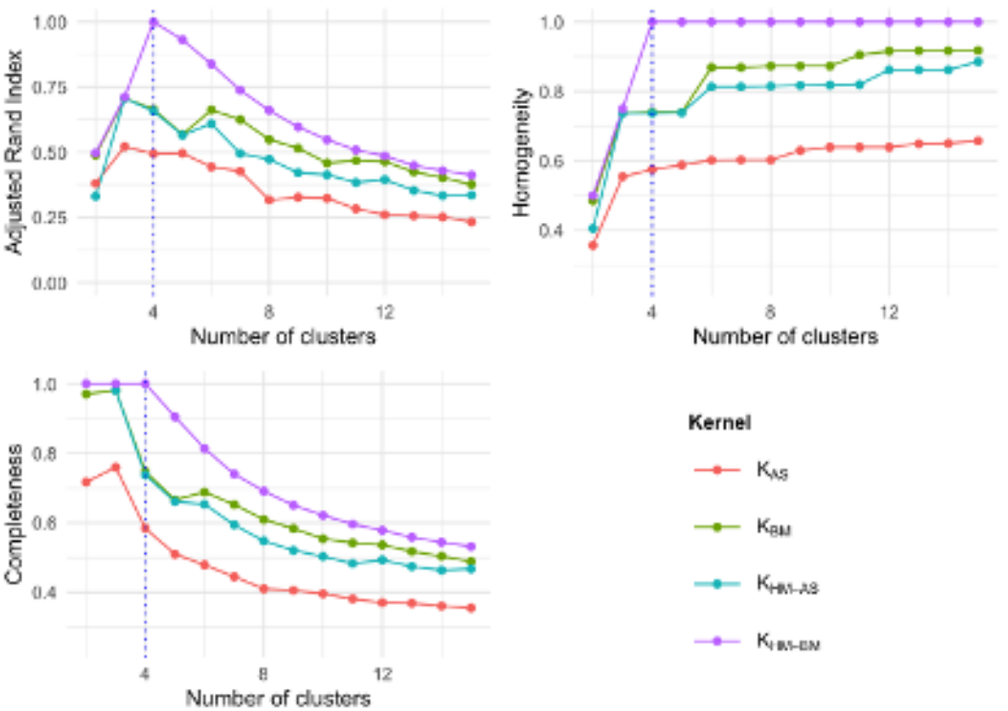

FIG. 7: Values of the Adjusted Rand Index (top left), the homogeneity (top right) and the completeness (bottom left) between the true partition and the clustering obtained for each kernel by cutting the dendrogram of the HAC with the Ward's linkage, for a number of clusters varying between 2 and 15.