Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information: Unlocking the Predictive Power of Quantum-Inspired Representations for Intermolecular Properties in Machine Learning

Raul Santiago,[1,*] Sergi Vela,[1] Mercè Deumal,[1] Jordi Ribas-Arino[1]

[1] *Departament de Ciència de Materials i Química Física and Institut de Química Teòrica i Computacional (IQTCUB), Universitat de Barcelona, Martí i Franquès, 1, 08028 Barcelona, Spain.*

[*] email: raul.sant.1972@gmail.com

1 Assessment of the effect of the basis set in MODA

The quality of the MODA representation significantly depends on both the chosen theoretical level used to compute the density matrix and the selected basis set for expanding the Hilbert space. Consistent with prior research, MODA employs the Superposition of Atomic Densities (SAD) guess, offering a balanced compromise between computational efficiency and the reliability of electronic structure representation. To address the basis set effect, we benchmarked MODA's performance using five different sets, selecting them based on a progressive increase in their complexity and accuracy. We start with STO-6G (*i.e.*, a minimal basis set) that, while less accurate, keeps computational demands in check. The 6-31G set provides a more precise representation, thanks to a split-valence approach that offers a nuanced description of valence electrons. The basis sets

6-31G* and 6-31+G* further refine accuracy by integrating polarization and diffuse functions, capturing shifts in electron cloud shape when atoms bond, and thereby enhancing computational precision. The final set, aug-cc-pVDZ, is part of the correlation-consistent basis set family, designed for systematic convergence to the complete basis set limit. Collectively, these basis sets provide various accuracy-to-computational cost trade-off options.

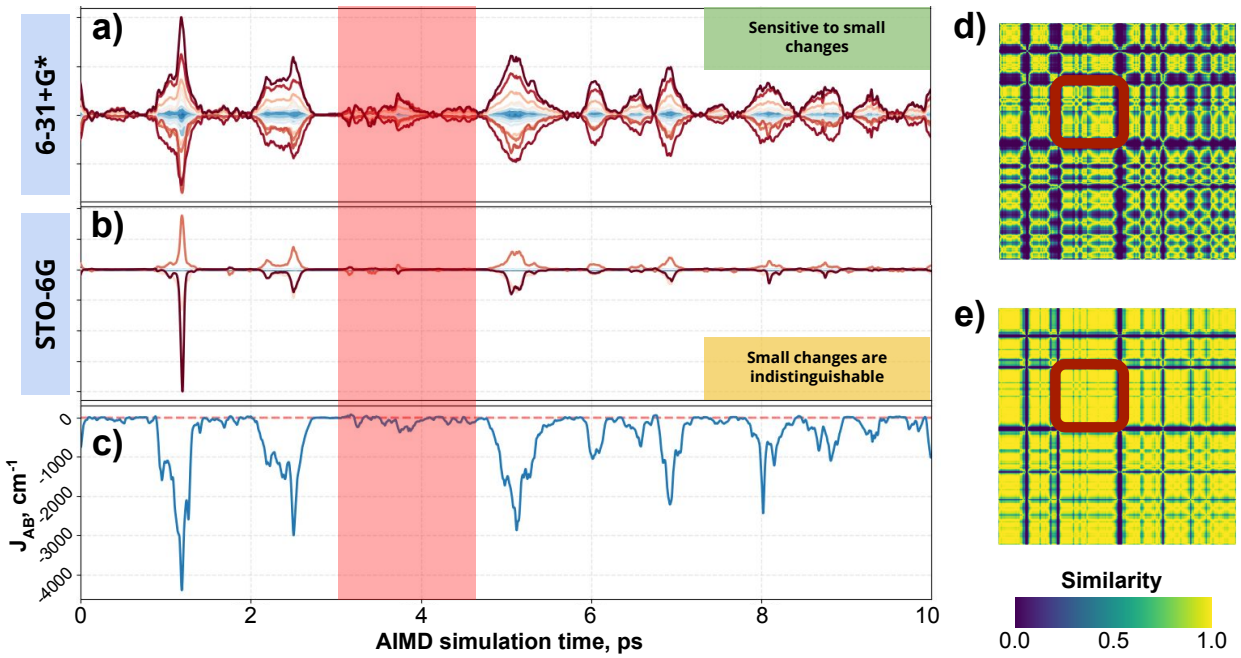

Figure 1: Evolution of the intermolecular components of MODA with (a) 6-31+G* and (b) STO-6G basis sets. (c) Time-resolved evolution of J_{AB} along 10 ps of AIMD simulation. Sample's similarity matrices for (d) 6-31+G* and (e) STO-6G basis sets. The highlighted region in pink of a-c corresponds to the framed region in d) and e).

Following the same approach employed throughout this work, we evaluated the performance of the basis set choice through both qualitative and quantitative analyses. Regarding the qualitative aspect, Fig. 1 illustrates the intermolecular components of MODA, computed via 6-31+G* and STO-6G basis sets, along with the time-resolved evolution of J_{AB} for a specific TTTA dimer. Upon inspection of the time-dependent MODA evolution (Fig. 1a-b), the highest sensitivity of the 6-31+G* basis set becomes apparent. This sensitivity is exhibited as a proportional shift in MODA components in response to minor variations in J_{AB}, and *vice versa*. In stark contrast, the STO-6G basis set displays significantly lower responsiveness. For instance, the evolution of the region highlighted in Fig. 1 covers a relatively narrow span of J_{AB} values. The 6-31+G* results

Basis	Train	Test	Validation
STO-6G	30.26	47.78	53.64
6-31G	8.12	20.14	25.43
6-31G*	8.81	20.08	25.46
6-31+G*	1.96	13.35	17.57
aug-cc-pvdz	1.77	13.58	17.42

Table 1: MAE (in cm^{-1}) associated to the prediction of J_{AB} values for the TTTA dataset. The MAE value of each column is provided at the optimal cross-validated set of hyperparameters using 25% of samples for training.

reveal these small variations, whereas the MODA components computed with STO-6G remain constant throughout the entire range, which makes a ML model unable to distinguish between these structures. The similarity maps of MODA using both basis sets (Fig. 1d-e) support these conclusions.

As for the quantitative analysis, we trained a KRR model using MODA data informed by different basis sets. The Mean Absolute Error (MAE) for the train, test, and validation sets is calculated. Our results indicate a progressive reduction of the MAE when transitioning from the STO-6G to the aug-cc-pvdz basis sets (see Table 1), which reinforces the importance of the basis set in MODA as a quantum-informed representation. It is to be noted that MAEs for 6-31+G* and aug-cc-pvdz compare well. However, since the computational cost of MODA using 6-31+G* is lower, this basis set is selected for further discussions in the main text.

2 Decoupling formalism of BoB and SOAP descriptors

In this section we present the formalism to decouple intra- and intermolecular components of BoB and SOAP representations. Additionally, we describe a scheme to automatically detect molecules in multi-moiety systems.

2.1 Decoupling of BoB

The Bag of Bonds (BoB) is a variant of the Coulomb Matrix descriptor, where instead of a matrix representation, atomic pairs are sorted into "bags" based on atomic types, with each bag containing the sorted Coulombic interactions of those atomic pairs. Let us consider a molecule

with N atoms, where each atom i is described by a nuclear charge Z_i and a position vector $\mathbf{R_i}$ in the three-dimensional space. The elements M_{ij} of the Coulomb matrix are defined[1] by

$$M_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{if } i = j, \\ \frac{Z_i Z_j}{|\mathbf{R_i} - \mathbf{R_j}|} & \text{if } i \neq j \end{cases} \tag{1}$$

where Z_i and Z_j are the nuclear charges, and $\mathbf{R_i}$ and $\mathbf{R_j}$ are the positions of atoms i and j. Next, in the BoB representation, we partition the off-diagonal elements of the Coulomb matrix into different "bags" corresponding to distinct pairs of atomic species. For example, all elements M_{ij} with i corresponding to a carbon atom and j to a hydrogen atom are placed into the "C-H" bag. Then, for each bag, we sort the interaction values in descending order. The final BoB representation is the concatenation of these sorted bags. It should be noted that in order to handle molecules of different sizes, zero padding is often used to ensure a consistent size of the feature vector across all molecules. The BoB method maintains a direct link to chemical intuition, as its elements can be associated with pairwise interactions between atoms. Importantly, this representation captures the invariance to atom indexing (like the Coulomb matrix) but also the total number of distinct atom pairs.

In the Bag of Bonds (BoB) representation, a decoupling of intra- and intermolecular components can be accomplished by partitioning each "bag" into two distinct sections. For a given bag corresponding to a specific atomic pair type, say X-Y, the associated intra- and intermolecular components can be differentiated by examining the individual atomic pairs contributing to that bag. Specifically, if both atoms X and Y belong to the same molecule, their interaction is an intramolecular component. On the other hand, if X and Y are found in different molecules, their interaction represents an intermolecular component. By applying this differentiation to all atomic pairs in a given bag, one can effectively divide the bag into two parts: one part containing all intramolecular components and the other part containing all intermolecular components. This process is repeated for all bags, creating a doubled set of bags, each bag pair representing the intra- and intermolecular components of a particular atomic pair type. The final representation, thus, consists of the sorted arrays of these intra- and intermolecular components for all types of

atomic pairs present in the system, providing a more detailed account of the atomic interactions in the system.

2.2 Decoupling of SOAP

SOAP, as implemented in many packages, is a 3-body descriptor, and thus, the intra/inter decoupling strategy presented for MODA and BoB (based on allocation of 2-body terms with respect of the interaction type) cannot be directly applied. For this reason, we have adopted for SOAP a decoupling strategy highly inspired in the work of Cersonsky *et al.*.[2] Here, we first illustrate the approach for the structure-average version of SOAP and later provide some specifications for the application of this approach in a local version of SOAP.

Consider x_a and x_b as distinct intramolecular sub-systems of the dimer, X. The respective structure-average SOAP representations for these sub-systems, denoted as ρ_{x_a} and ρ_{x_b}, encode intramolecular information. The SOAP representation of the dimer, ρ_X, on the other hand, encompasses both intra- and intermolecular information, which cannot be directly segregated in intra- and intermolecular components. We can directly compute the intramolecular components of the dimer, ρ_X^I, as the sum of the sub-systems x_a and x_b components:

$$\rho_X^I = \rho_{x_a} + \rho_{x_b} \tag{2}$$

an then, the intermolecular components (denoted by ρ_X^i) can be obtained as:

$$\rho_X^i = \rho_X - \rho_X^I \tag{3}$$

where intramolecular components are subtracted from the total dimer representation, still retaining the 3-body encoding of SOAP.

When considering a local SOAP representation, one major issue related to the dimensionality of X and x_i representations emerges. The dimer's SOAP representation, ρ_X, can be represented by a matrix of shape $(N_{a+b} \times M)$, where N_{a+b} represents the total atom count in the sub-systems x_a

and x_b, and M is a function of selected hyperparameters, among other factors. Conversely, the respective sub-systems' shapes are $N_a \times M$ and $N_b \times M$ for x_a and x_b, thus making inadequate to calculate the intermolecular components as stated formerly. To mitigate this discrepancy, it is proposed to concatenate the subsystem components of Eq. 2, rather than summing them ($\rho_X^I = \rho_{x_a}|\rho_{x_b}$). From this point, operations can be carried out as usual. Care should be taken, however, to ensure that the components of ρ_X and ρ_X^I are identically ordered prior to perform the subtraction.

2.3 Automatic detection of moieties for arbitrary chemical systems

The decoupling strategy described for BoB, SOAP, and MODA generally relies on partitioning the contributions arising from either the same or different moiety. Consequently, the algorithm designed to generate any of these descriptors must understand the connectivity of every atom in the system to establish intra/intermolecular criteria. Datasets typically consist of hundreds or thousands of data samples, thus it is required the automation of the process to detect different moieties from a set of coordinates. In this work, we have tested two approaches to tackle this problem, both based on graph theory algorithms. Before delving into the details, the graph structure must be derived from the atomic coordinates.

A graph $\mathcal{G} = (V, E)$ is defined by the vertices V and edges E. Analogously, a molecule possesses a set of atoms and bonds that can be directly mapped to V and E, respectively, by establishing a distance criterion to determine whether two atoms are close enough to be mapped into $\mathcal{G}$ as connected or disconnected. By iteratively performing this process, the graph adjacency matrix, A, can be built, where the element a_{ij} is either 1 or 0 depending on the existence of a connection. Graph theory provides useful techniques to search for sub-graphs ($\mathcal{G}_i \subseteq \mathcal{G}$ of moieties) from a given graph (the dimer or multi-moiety system), with A being its cornerstone. Among others, depth-first search and spectral clustering-based (SC) techniques are typical choices. In this work, we have implemented both approaches, but after experimentation, we found that the SC-based algorithm yielded higher performance with lower computation times and more stable results.

Let L be the Laplacian matrix, obtainable as $L = D - A$, where D represents the degree matrix and A is the adjacency matrix. The matrix D is diagonal and can be easily derived from A, as every

diagonal element (d_{ii}) consists of the number of edges attached to the vertex v_i (*i.e.*, $d_{ii} = deg(v_i)$). Now, let the eigen-decomposition of L be

$$L = U\mathcal{L}U^{\dagger} \tag{4}$$

where U contains the eigenvectors in columns and $\mathcal{L}$ is a diagonal matrix containing the eigenvalues, λ_i. Moreover, L is a positive semidefinite matrix, which must satisfy $\lambda_i \geq 0$. Spectral clustering theory states that the number of eigenvalues satisfying $\lambda_k = 0$ equals to the number of sub-graphs (or moieties) of the total graph. Moreover, the nodes (or atoms) of each moiety can be easily identified by inspecting the values of the eigenvectors, u_i, with an associated $\lambda_i = 0$. The components of these eigenvectors must have the following values:

$$u_{ij} = \begin{cases} \frac{1}{\sqrt{N_i}} & \text{if } v_j \in \mathcal{G}_i, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

In other words, the j-th component of u_i is $1/\sqrt{N_i}$ if the vertex v_j belongs to the i-th sub-graph ($\mathcal{G}_i$), and it is 0 otherwise. In the last equation, N_i represents the total number of nodes of the i-th subgraph. Building upon that, the algorithm to detect moieties from a multi-moiety system consists of the following steps:

1. Build the adjacency matrix A based on the distances of all atoms with the rest and a threshold value serving as a distance cutoff (2.0 Å in our case based on Van der Waals radii).
2. Calculate the Laplacian matrix L as $L = D - A$, where D is easily obtained from A.
3. Diagonalize L to obtain the eigenvectors u_i and their associated eigenvalues λ_i.
4. Identify the number of eigenvalues that satisfy $\lambda_i = 0$ and extract their associated eigenvectors.
5. Group together all the nodes with non-zero components belonging to each sub-graph, $\mathcal{G}_i$, based on the values of their associated eigenvectors.

3 Description of TTTA dataset

3.1 TTTA unit cell and the origin of PED

TTTA crystals present two stable phases depending on the temperature. On heating above 300 K TTTA arranges in a paramagnetic, monoclinic phase (high-temperature or HT phase) containing four columns of TTTA units (see "CX" labels in Fig. 2a) that stack equidistant on top of each other (see "DX" labels in Fig. 2b) due to π-π interactions between the TTTA radicals. The observed uniform stacking is a consequence of a rapid intra-stack Pair Exchange Dynamics (PED), stemming from thermal fluctuations. Here, TTTA radicals continually swap the adjacent TTTA neighbor (either upper or lower) with which they form a dimer (see Fig. 2c).

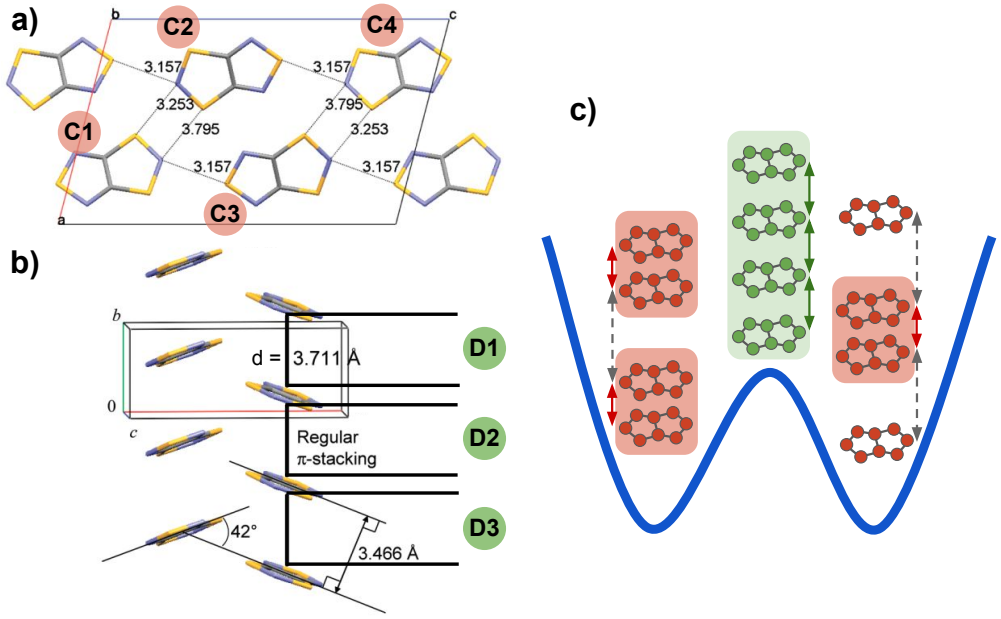

Figure 2: (a) Top and (b) side views of the monoclinic polymorph unit cell of TTTA. Insets in (a) and (b) provide labels to identify the TTTA dimers throughout the unit cell, where CX and DX stand for column and dimer number X. (c) Schematic depiction of the Pair Exchange Dynamics (PED) phenomenon. The path along the Potential Energy Surface (blue) features a double well, illustrating that the dimerized TTTA π-stacks occupy a lower energy state than the equidistant columns. It is noteworthy that the latter emerge as a time-resolved average when the thermal energy is sufficient to facilitate transitioning between both dimerized ground-state configurations.

Following the same strategy of our previous works,[3,4] the PED phenomenon can be studied through solid-state AIMD simulations. In this case, the primitive HT polymorph cell of TTTA is extended along the b (π-stacking direction) and c lattice parameters (see Fig. 2a-b), facilitating free move-

ment of the TTTA units. This results in a total of 32 TTTA units. Subsequently, the positions of all atoms within the TTTA dimer conformations (D1-D3) from each column (C1-C4) are extracted from the AIMD trajectory, and the magnetic exchange coupling J_{AB} is computed for each structure. Due to computational cost limitations, the time-evolution of J_{AB} interactions was evaluated only for a specific subset of radical pairs among the 32 TTTA units in the supercells. Specifically, we focused on adjacent TTTA dimers along the π-stacking direction (D1 and D3 from Fig. 2b) from the first and second columns, *i.e.*, C1 and C2 columns of Fig. 2a. These are identifiable via the dimer and column tags D1C1, D2C2, D3C1. Fig. 3a shows the time resolved J_{AB} for the selected dimers extracted from AIMD carried out at two different temperatures (250 K and 300 K), and Fig. 3b indicates the J_{AB} distribution of each dimer.

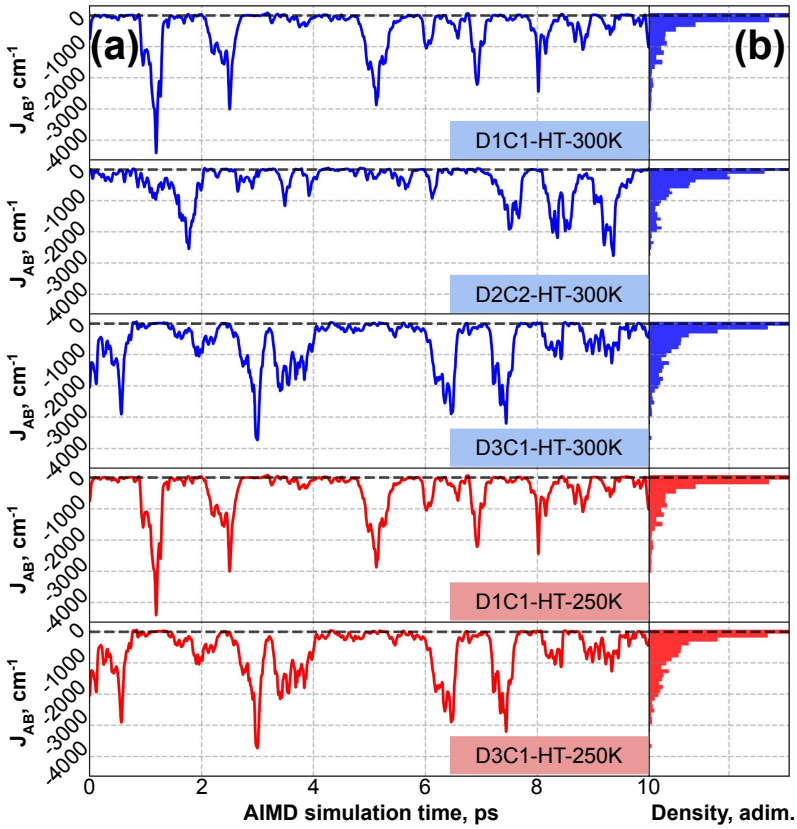

Figure 3: (a) Time-resolved J_{AB} values across a 10 ps AIMD simulation for each TTTA pair examined in this study. (b) Density distribution of J_{AB}. Blue curves indicate an AIMD temperature of 300 K, whereas the red curves correspond to 250 K.

3.2 Two-samples Kolmogorov-Smirnov test

In order to verify that sampled TTTA dimers from HT-300K and HT-250K explore a different region of the thermally-available configurational space and, thus, some degree of extrapolation exists from a model trained with HT-300K when predicting J_{AB} for dimers extracted from HT-250K, we have applied the 2-samples Kolmogorov-Smirnov (KS) test,[5] which is a non-parametric test used to evaluate the equality of continuous, one-dimensional probability distributions that are derived from two independent samples. The null hypothesis (H_0) asserts that the samples are drawn from the same distribution, while the alternative hypothesis (H_1) states that the samples are drawn from different distributions. The KS test statistic (D) is used to evaluate which of the two hypothesis is statistically correct and it is defined as the maximum absolute difference between the empirical cumulative distribution functions (ECDF) of the two samples. Mathematically:

$$D = \max |F_1(x) - F_2(x)| \tag{6}$$

where $F_1(x)$ and $F_2(x)$ represent the ECDFs of the two samples coming from the empirical f_i distribution functions. Upon calculation of the test statistic, we can compare it to the critical value from the Kolmogorov distribution (D_α), given the desired significance level (commonly $\alpha = 0.05$) and the sample sizes. The value D_α can be calculated as:

$$D_\alpha = c_\alpha \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \tag{7}$$

where N_1 and N_2 are the sample sizes, and c_α is a constant depending on the significance level that takes the approximate value of 1.36 for a significance level of $\alpha = 0.05$.

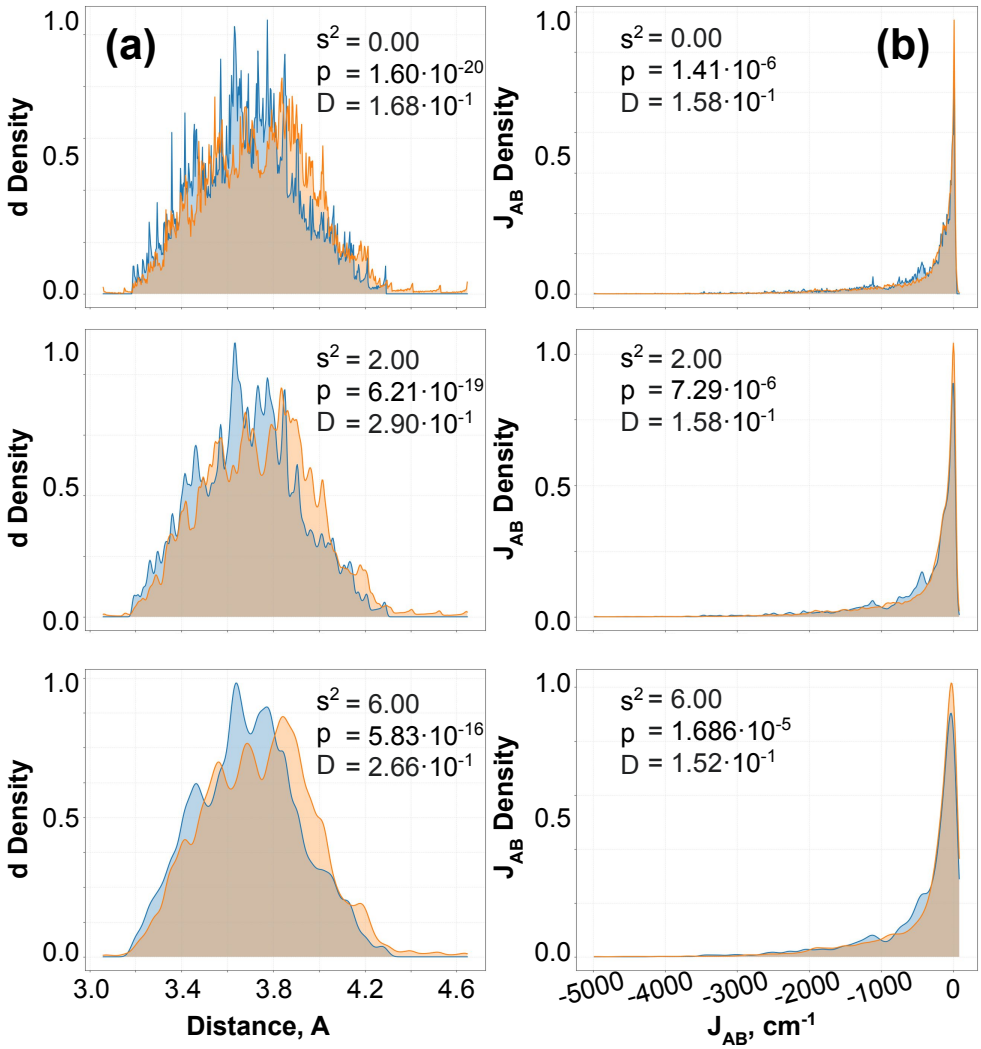

Figure 4: Two-samples Kolmogorov-Smirnov test results of HT-300K (orange) and HT-250K (blue) samples of TTTA datasets. (a) Distribution ($f^{d}_{\text{HT-300K}}$ and $f^{d}_{\text{HT-250K}}$) functions of centroid-centroid distance (d) using Gaussian smoothing filters with different variances (s^2) indicated in the inset of each plot. (b) Analogous results when considering J_{AB} distribution ($f^{J_{AB}}_{\text{HT-300K}}$ and $f^{J_{AB}}_{\text{HT-250K}}$). The insets of each plot also specify the p-value and D for the KS tests.

If the test statistic is greater than the critical value, we reject H_0 in favor of H_1, providing evidence that the two samples are drawn from different distributions. Conversely, if the test statistic is less than or equal to the critical value, we fail to reject the null hypothesis, indicating insufficient evidence to claim that the distributions are different. A p-value can also be computed to summarize the strength of the evidence against H_0. A small p-value (≤ 0.05) provides strong evidence against H_0, suggesting the two samples are likely drawn from different distributions.

We employed the two-sample KS test to confirm that the HT-300K and HT-250K samples derive from distinct distributions, thus indicating a substantial exploration of different regions of the

potential energy landscape. Specifically, we applied the test to both the centroid-centroid distance (denoted as d in Fig. 4b, with corresponding distributions $f^{d}_{\text{HT-300K}}$ and $f^{d}_{\text{HT-250K}}$) and the J_{AB} distributions ($f^{J_{AB}}_{\text{HT-300K}}$ and $f^{J_{AB}}_{\text{HT-250K}}$) from the HT-300K and HT-250K samples, respectively. The panels in Fig. 4 display the distance (a) and J_{AB} (b) distributions of HT-300K and HT-250K together, with a chosen significance level of $\alpha = 0.05$. While the KS test is typically applied to continuous distributions, our distributions were generated by binning or discretizing the space. To account for potential artifacts due to this discretization, we applied a Gaussian filter with a variable smoothing factor controlled by the variance (s^2) to both d and J_{AB} distributions. Our analysis revealed that for any chosen value of s^2, the critical D_α value and p-value consistently suggest a significant difference between the distributions. This observation implies that a certain degree of extrapolation is likely to be present when modeling the data.

It is crucial to emphasize that the KS test assumes the independence of samples, whereas our data stem from an AIMD simulation, inherently producing highly correlated data points. Nonetheless, the total simulation time (10 ps) ensures ergodicity in our simulations, thereby facilitating a robust and accurate sampling of d and J_{AB} distributions. The principle of ergodicity allows us to consider these time-dependent, correlated samples as effectively independent over the long term, allowing the application of KS test in practice.

4 Further Agglomerative Clustering Results

4.1 Variation of the number of clusters

To delve deeper into the Agglomerative Clustering (AC) analysis, we present the clustering results for MODA, BoB, and structure-average SOAP representations with varied cluster numbers in the PHYL and THIL datasets. As discussed in the main text, AC models informed by the MODA representation accurately capture the electronic structure information, which is crucial to classify structures based on J_{AB}. For the PHYL dataset, as demonstrated in Fig. 5, MODA maintains the anticipated clustering symmetry across different cluster counts (n_c) beyond the specific $n_c = 10$ shown in the main text. In contrast, structure-average SOAP and BoB display clustering patterns

where the symmetry of J_{AB} *vs.* θ is not preserved for any n_c values, thereby reinforcing the conclusions developed in the main text.

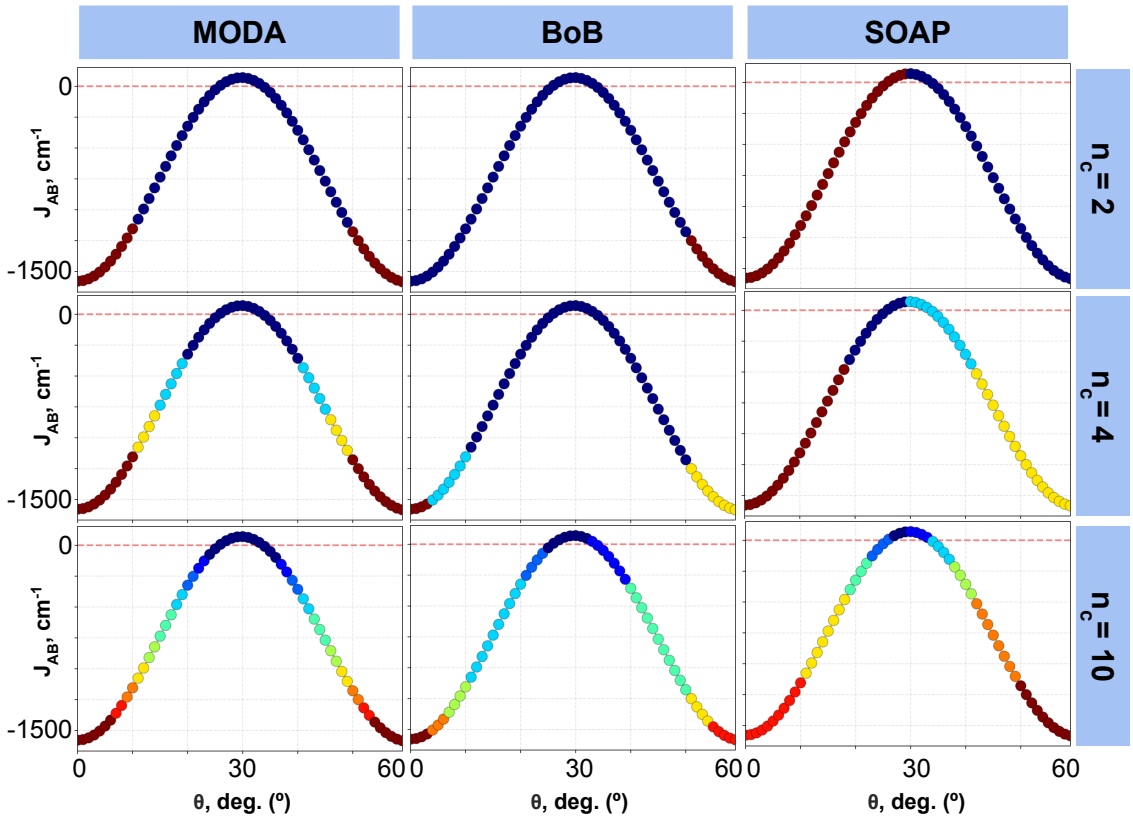

Figure 5: Agglomerative Clustering outcomes for the PHYL dataset. The labels at the top specify the descriptor in use, while the labels on the right indicate the number of clusters (n_c) represented in each plot.

In regard to the AC results of the THIL dataset, subtle differences in the clustering patterns can be observed, especially in the $n_c = 2$ and $n_c = 4$ cases. While MODA results display a balanced distribution across the samples, indicating a more coherent grouping of conformers with similar J_{AB} values, SOAP and BoB tend to generate slightly uneven distribution of clusters that occasionally group structures with significantly different J_{AB} values within the same cluster (as seen in the blue and yellow clusters of Fig. 6 when $n_c = 4$). However, it is important to note that these differences are relatively minor and the overall performance of all descriptors remains comparable. As illustrated in the main text, this can be attributed to the fact that J_{AB} in the THIL dataset is well-defined by structure-based descriptors.

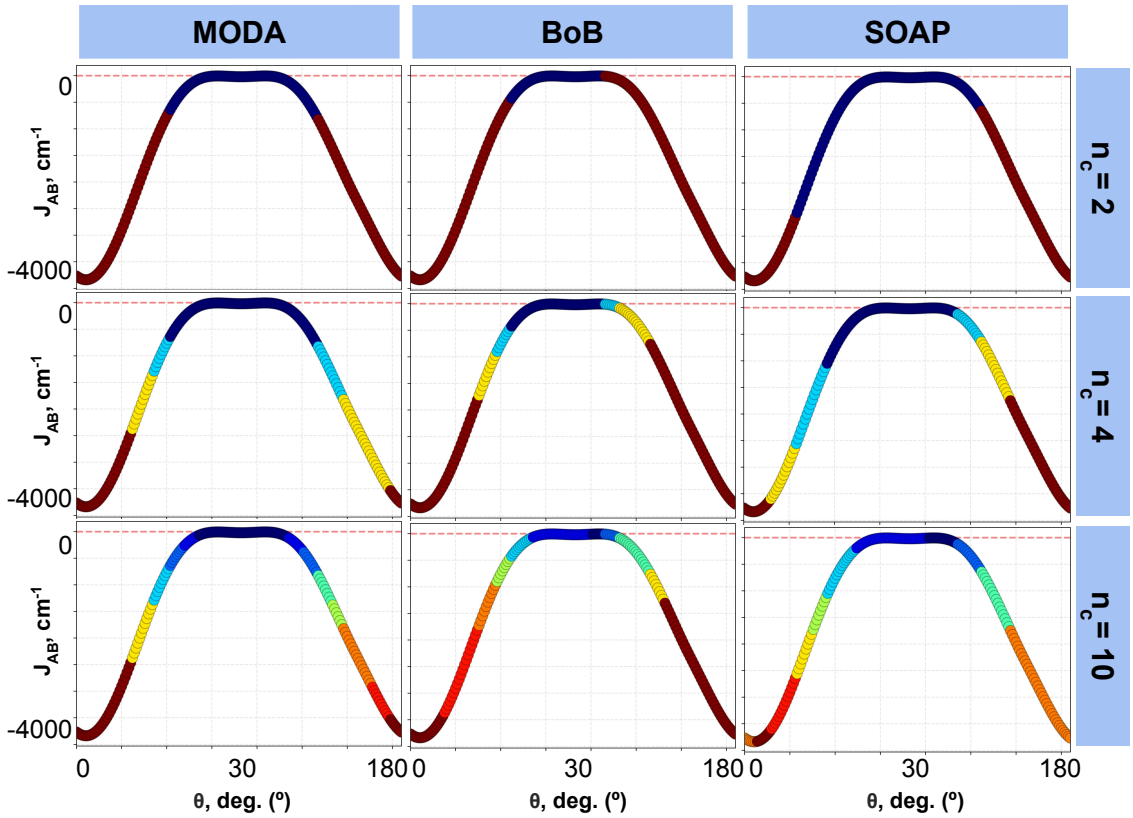

Figure 6: Agglomerative Clustering outcomes for the THIL dataset. The labels at the top specify the descriptor in use, while the labels on the right indicate the number of clusters (n_c) represented in each plot.

4.2 Local vs. structure-average SOAP Agglomerative Clustering Results

As we continue to explore the performances of various descriptors, it is crucial to note that an ideal descriptor should map PHYL's conformations with identical J_{AB} values to similar representations at each branch of the pivotal point at $\theta = 30°$. In other words, we expect the dot product between representations of data samples with the same J_{AB} to approach 1. More generally, a given non-linear similarity metric should approach 1. With this fundamental principle in mind, we turned our attention to the non-linear RBF metric that euclideate pairwise similarity relations and quantifies the correlation between data samples. This can provide qualitative insights about the performance of a descriptor without using a predictive model. This metric allows us to produce a similarity map for a sample (see Fig. 7). Our findings indicated that MODA displays a symmetric similarity map around $\theta = 30°$, while the structure-average SOAP presents a gradual change throughout the

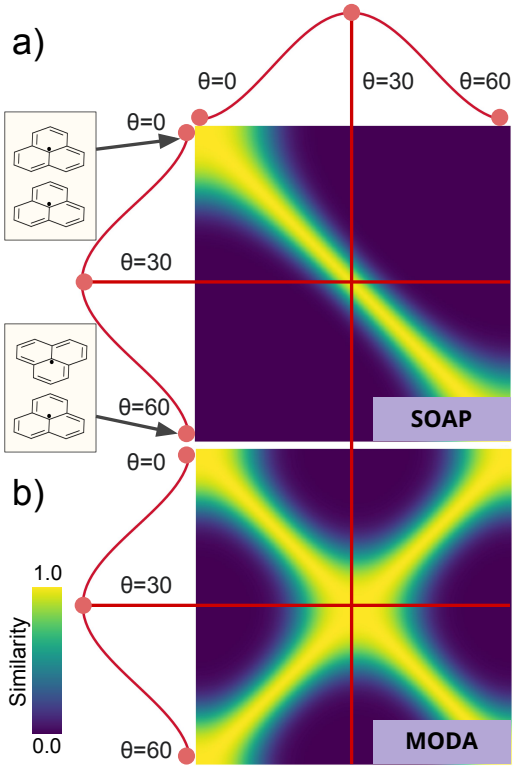

Figure 7: Similarity maps associated with the (a) structure-average SOAP and (b) MODA representations of the PHYL dimer along a change of θ. The color code indicates the similarity between two samples, ranging from yellow (1, equal samples) to dark blue (0, orthogonal samples).

θ range. This gradual change suggests that equivalent representations for PHYL conformations with identical J_{AB} are not as clearly reflected by the structure-average SOAP.

However, as mentioned in the main text, structure-average SOAP may not leverage the full potential of the descriptor. For this reason we employed the Regularized Entropy Match (REMatch) method, in conjunction with the RBF kernel, to derive global similarity measures from the local SOAP spectrum. This process involves assessing the impact of the entropy penalty parameter (α), with distinct values of α employed for kernel computation. These specific α values were chosen to evaluate local similarities in best match (α =0.01), intermediate (α =1.00), and average-like regimes (α =10.0). Utilizing these similarity maps, we conducted the Agglomerative Clustering analysis and contrasted the results with those obtained from structure-average SOAP and MODA. Our qualitative assessment revealed a minor difference between the similarity maps produced by the RBF-REMatch kernel (visible in the bottom panels of Fig. 8) and the RBF kernel used for structure-average SOAP (see Fig. 7a). The local SOAP version succeeded in capturing relevant

similarities among PHYL conformers that its structure-average counterpart could not discern. Nevertheless, a persistent asymmetry remains in both the kernel and the resulting clustering pattern independently of the α parameter of choice (see the top panels of Fig. 8), showing similar limitations compared to the structure-average version.

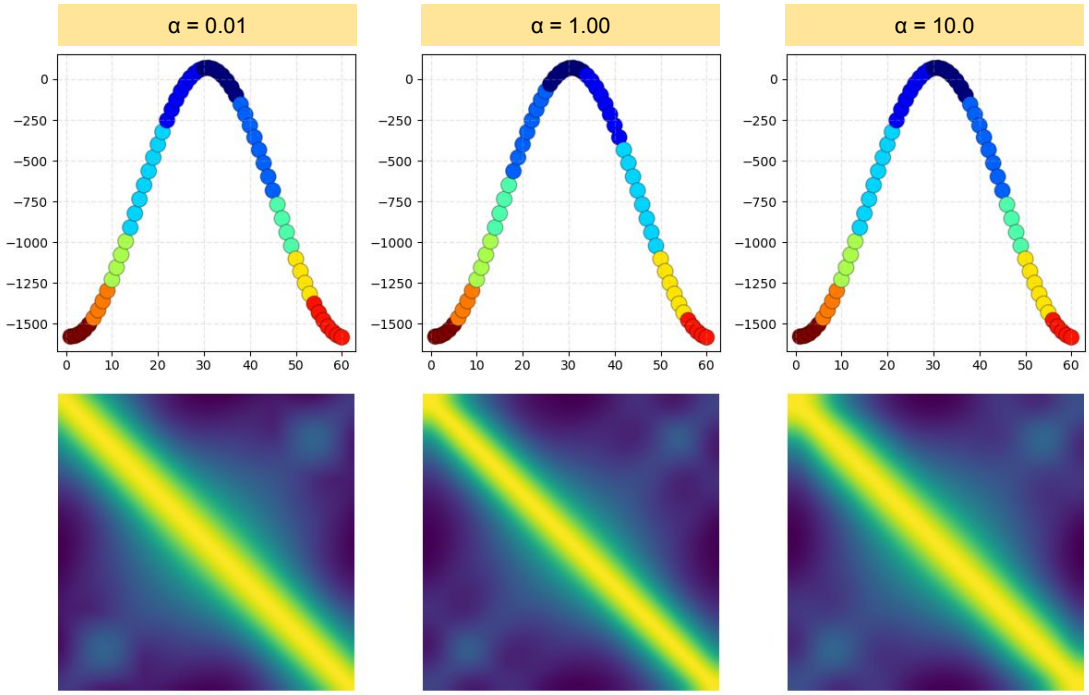

Figure 8: Agglomerative Clustering results with $n_c = 10$ of PHYL dataset for the local SOAP. Each column shows the results for a specific value of the entropy penalty parameter (α) of the RBF-REMatch kernel.

5 Calculation of Molecular Orbitals for MODA

5.1 Natural Orbitals Extraction from the Density Matrix

As mentioned in the main text, our implementation of MODA uses the pySCF package to estimate the density matrix using the Superposition of Atomic Densities (SAD) guess and subsequently employs Natural Orbitals (NOs). However, pySCF does not provide the required functionality to extract the NOs. Instead, a custom routine must be established. In this section, we present the general formalism required to extract the NOs from the 1-body reduced density matrix and discuss how to operate when dealing with open-shell systems such as radicals. To this end, let D be the

1-body reduced density matrix obtained, for instance, via the SAD "guess". D can be defined as:

$$D = CnC^T \tag{8}$$

Here, C corresponds to the unknown and target matrix containing the atomic orbital coefficients of the NOs in columns, C^T is its transpose, and n is a diagonal matrix with the occupation number of each NO. Since C consists of an orthonormal set in a non-orthogonal Hilbert space, the following equality must hold:

$$I = C^T SC \tag{9}$$

where I is the identity matrix and S is the overlap matrix associated with the non-orthogonality of the basis. Multiplying equation 8 by SC on the right on both sides of the equation and using the identity of equation 9 results in

$$\begin{aligned} DSC &= CnC^T SC \\ DSC &= Cn \end{aligned} \tag{10}$$

where $C^T SC$ simplifies according to equation 9. Now, we can modify equation 10 to obtain an eigenvalue equation. Before that, it is essential to note that for a positive semidefinite matrix (*i.e.*, its eigenvalues, λ_i, satisfy $\lambda_i \geq 0$) such as S, it can be decomposed as $S = S^{1/2}S^{1/2}$,[6] thus, we can multiply equation 10 by $S^{1/2}$ and proceed as follows:

$$\begin{aligned} S^{1/2}DSC &= S^{1/2}Cn \\ S^{1/2}DS^{1/2}S^{1/2}C &= S^{1/2}Cn \\ D'K &= Kn \end{aligned} \tag{11}$$

to obtain an ordinary eigenvalue equation in the last step,[7] where the variable changes $K = S^{1/2}C$ and $D' = S^{1/2}DS^{1/2}$ are set for clear identification of the eigenvalue problem elements. From this point, K and n can be simultaneously solved using standard numerical recipes to diagonalize D'. Subsequently, the target NOs can be obtained via $C = S^{-1/2}K$.

It is worth noting that $S^{\pm 1/2}$ matrices can be obtained by 1) diagonalizing S, 2) raising all the eigenvalues of S to $\pm 1/2$, and 3) undoing the diagonalization of S, or mathematically:

$$S^{\pm 1/2} = U\Lambda^{\pm 1/2}U^T \tag{12}$$

where U corresponds to the "maximum overlap orbitals"[8] and Λ is the eigenvalues matrix associated with S. Overall, the process of obtaining the NOs from the 1-body reduced density matrix requires the numerical resolution of two eigenvalue problems.

Finally, it is important to note that the density matrix D is formed from the contributions of both α and β electrons. Therefore, the density matrix can be expressed as $D = D^{\alpha} + D^{\beta}$. In the case of closed-shell systems, $D^{\alpha} = D^{\beta}$ holds true. However, this equality is not valid for spin-polarized systems. Instead, each contribution to the density matrix is calculated individually and then combined. This condition has been taken into account for all calculations involving D, as this work primarily focuses on open-shell systems.

5.1.1 Sum of Doublets for Multi-moiety Systems

Kahn's model[9] for multi-moiety systems establishes a clear relationship between the singly-occupied natural orbital (SONO) of each moiety and J_{AB}. Consequently, our approach to the electronic structure of dimers (e.g., PHYL and TTTA) involves calculating the SONO of each monomer prior to obtaining the dimer's SONOs as a linear combination of the monomers' SONOs.

Consider a multi-moiety system defined by the set $\mathcal{S} = \mathcal{M}_A \cup \mathcal{M}_B$, where $\mathcal{M}_i$ represents the subset of parameters ascribed to the i-th moiety. Let $D_{\mathcal{M}_i}$ denote the 1-body reduced density matrix of $\mathcal{M}_i$. As discussed in the previous section, the natural orbitals (NOs), $\{\psi_{ij}\}$, and their occupations, $\{n_{ij}\}$, of each moiety can be derived from their respective density matrix expressions. Consequently, the SONO of each moiety, $\psi_{i,SONO}$, can be identified as the only NO that satisfies $n_{ij} = n_{i,SONO} = 1$. Expanding on this, we can approximate the pair of SONOs for the multi-moiety system using valence bond theory, as the positive (Φ^+) and the negative (Φ^-) linear combinations of the monomer's SONO. More precisely, the orthonormal SONOs of the dimer can be approximated

as

$$\Phi^{\pm} = \frac{1}{\sqrt{2(1 \pm S_{AB})}} (\psi_{A,SONO} \pm \psi_{B,SONO}) \tag{13}$$

where $S_{AB} = \langle \psi_{A,SONO} | \psi_{B,SONO} \rangle$ denotes the overlap integral between the SONO of each moiety. These are the sub-space of selected orbitals to construct MODA representations of the TTTA and PHYL conformers.

5.2 TTTA Singly-occupied Natural Orbitals

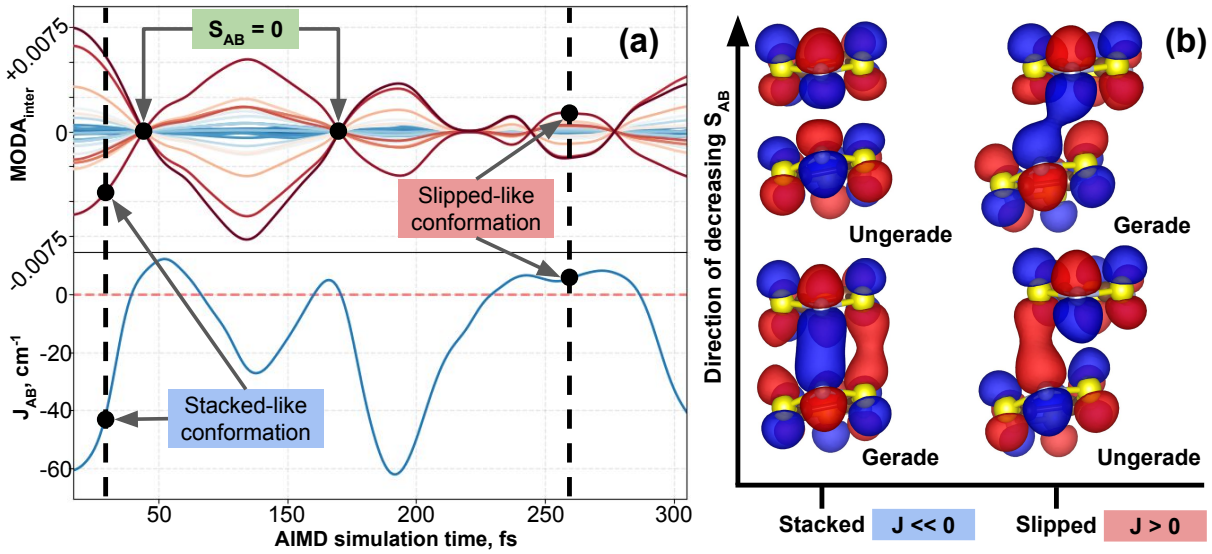

Figure 9: (a) Evolution of MODA components within a specific region of the AIMD where transitions from AFM to FM are prevalent. Labels denote particular time-steps likely to host stacked, slipped, or intermediary configurations. (b) Schematic depiction of the Singly-occupied Natural Orbitals (SONOs) of a TTTA dimer with isosurface value at 0.025 $e/\text{Å}^3$, organized according to the conformational structure and overlap (S_{AB}).

Along the AIMD simulation, π-stacked TTTA dimers explore the potential energy surface due to thermal fluctuations. By examining the various conformations along the AIMD, we can identify two primary factors significantly influencing the J_{AB} between adjacent stacked TTTA units: (a) the distance between TTTA monomers, and (b) the slippage of one TTTA unit with respect to the other. The effect of monomers drawing closer is well-known: it maximizes the overlap between the TTTA's SONOs, leading to a strong antiferromagnetic (AFM) coupling between the TTTA units. In contrast, slippage results in a drastic reduction in overlap, dropping towards zero and causing $J_{AB} \geq 0$. Indeed, when slippage is sufficiently large, the ungerade linear combination of SONOs can achieve a higher overlap compared to the gerade combination, as exemplified in Fig.

9b. Moreover, we showcase the evolution of MODA components during a fragment of the AIMD simulation where J_{AB} oscillates between antiferromagnetic (AFM) and ferromagnetic (FM) values multiple times. In response to this, the MODA components also change sign (see Fig. 9a).

5.3 THIL & PHYL Singly-occupied Natural Orbitals

Fig. 10 displays the pair of SONOs of THIL, which correspond to the ungerade and gerade linear combination of Atomic Orbitals.

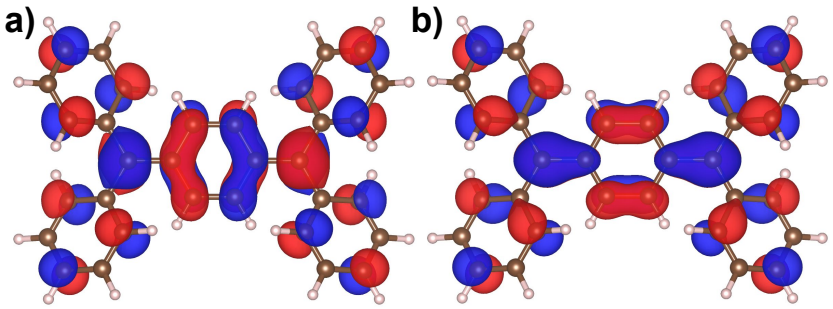

Figure 10: Pair of SONOs included in the MODA representation for THIL with isosurface value of 0.025 $e/\text{Å}^3$. a) corresponds to the ungerade and b) to the gerade orbitals.

Fig. 10 displays the pair of SONOs of PHIL, which correspond to the bonding and anti-bonding inter-radical linear combination of SONOs of extreme $\theta = 0°$ and $\theta = 60°$ conformers. The SONO of the PHYL monomer is also displayed.

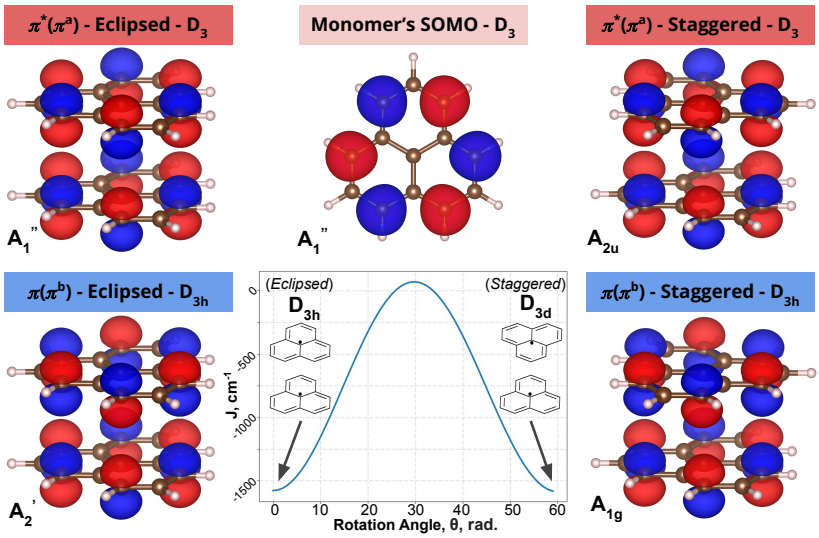

Figure 11: Illustration of the THIL's structure and SONOs with isosurface value of 0.025 $e/\text{Å}^3$ at eclipsed (left) and staggered conformations (right). The Molecular Orbitals are accompanied by a blue/red label indicating bonding/antibonding linear combinations, both included in MODA representation. Note that both eclipsed and staggered conformers present bonding and antibonding orbitals with the same point group symmetry.

6 Analysis of intra- & intermolecular decoupling of MODA and BoB

This section elaborates on the evolution of decoupled versions of both BoB (see Fig. 12) and MODA (see Fig. 13). Similar to the discussion on SOAP in the main manuscript, the intramolecular components of BoB and MODA do not evolve coherently with the time-resolved J_{AB} values. Conversely, the intermolecular components exhibit an apparent correlation with J_{AB}. In line with the examination of SOAP and MODA intermolecular components, BoB components present a markedly lower degree of correlation compared to MODA. Analogous to SOAP, BoB can also exhibit similar values across its components when the corresponding J_{AB} values are drastically different, as illustrated by comparing the green and purple highlighted regions in Fig. 14. On the other hand, MODA exhibits a coherent representation with respect to J_{AB} changes and thus proves to be a more reliable descriptor.

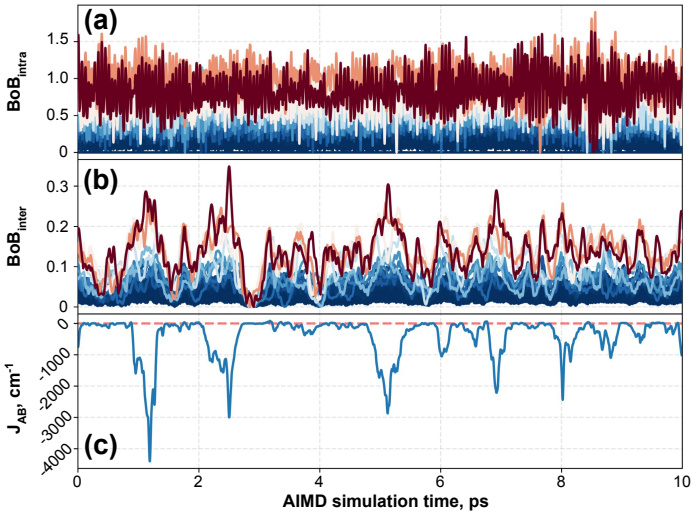

Figure 12: Decoupled intra- (a) and intermolecular (b) components of BoB along 10 ps of AIMD simulation together with (c) the time-resolved J_{AB} values for comparison. Each curve in (a) and (b) corresponds to a feature of BoB, where the color scheme ranges from high variance (red) to low variance (blue).

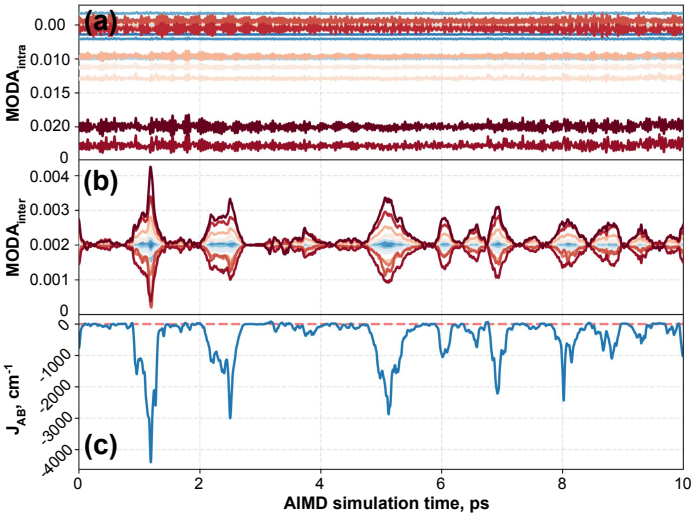

Figure 13: Decoupled intra- (a) and intermolecular (b) components of MODA along 10 ps of AIMD simulation together with (c) the time-resolved J_{AB} values for comparison. Each curve in (a) and (b) corresponds to a feature of MODA, where the color scheme ranges from high variance (red) to low variance (blue).

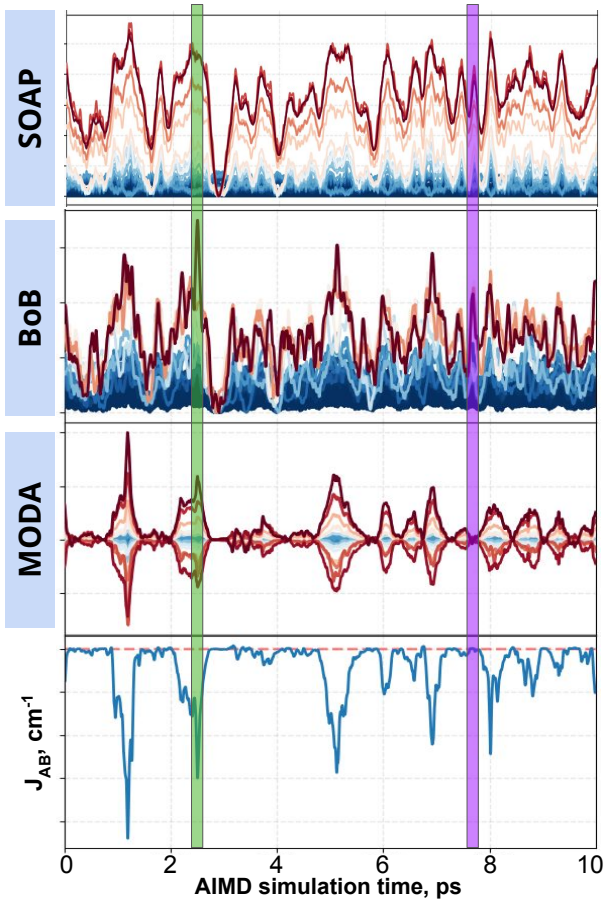

Figure 14: Comparison of intermolecular components of SOAP, BoB, and MODA. The time-resolved J_{AB} evolution is included for comparison. Green and purple boxes indicate regions with extreme J_{AB} values (largely AFM and slightly FM, respectively). Note that MODA is the unique descriptor able to sharply differentiate the highlighted regions.

7 Explanation of the Cross-Validation Scheme

Leave-P-Out (LPO) and Leave-P-Groups-Out (LPGO) are cross-validation schemes frequently used in model evaluation and selection. These are common tools in machine learning, statistical analysis, and predictive modeling. These validation methods can be crucial in determining the performance of a model, preventing overfitting, and enhancing generalizability.

The LPO method involves generating all possible combinations from the total number of samples N, taking p samples out at a time (see scheme in Fig 15a). The number of combinations (or folds) can be calculated using the binomial coefficient, $N!/(p!(N-p)!)$. For each fold, the model is trained on the remaining $(N-p)$ samples, and hyperparameters are optimized before testing on the selected p samples. The error metric is computed for each split, and the overall performance

is obtained by averaging these metrics to obtain the cross-validated error.

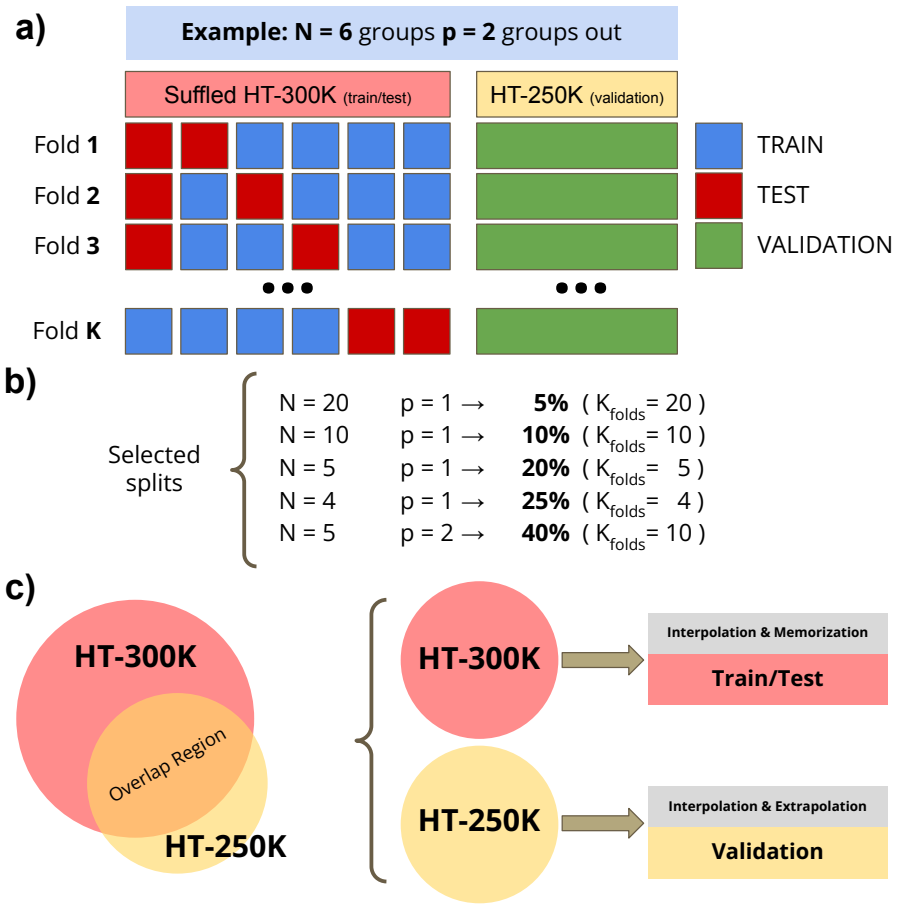

Figure 15: (a) Schematic representation of the possible combinations to generate all the cross-validation folds within the leave-p-groups out procedure. The scheme indicates that train (blue) and test (red) groups correspond to HT-300K dataset while validation (green) correspond to the HT-250K. (b) Specific values for total groups and groups out (N and p, respectively) and the percentage of training and number of folds that implies each combination. (c) Schematic representation of the overlap between HT-300K and HT-250K distributions.

However, the number of folds is factorial in nature, which can quickly increase with large datasets. This escalates computational cost and time complexity, making LPO practically unfeasible when dealing with big data. This is where the LPGO cross-validation scheme offers a solution. Instead of individual samples, LPGO treats groups of samples as the fundamental unit. In this context, N refers to the total number of these groups. The technique partitions the data into N groups, leaving p groups out for testing the model. By varying the total number of groups and the groups left out, we can calculate the training size as the ratio $(N - p)/N$. Adjusting these two values allows for the creation of a learning curve, offering insights into model performance as the amount of training data changes (see Fig 15b).

In cross-validation schemes, samples included in the test (p) and training ($N-p$) sets are routinely exchanged. Moreover, the hyperparameter optimization process involves evaluation of the model specifically on the test set, effectively using the test set as part of the training process. To further prevent overfitting, a third set (validation set) can be added to the cross-validation process, thereby adding another layer to model evaluation and tuning.

In this work, we employ an LPGO cross-validation strategy where we exploit the different origins of our TTTA datasets, which come from different temperatures (HT-300K and HT-250K). Here, samples from HT-300K are used within the LPGO scheme, while HT-250K samples are employed as a separate validation set. As illustrated in Fig. 15c, the HT-300K and HT-250K datasets contain samples that mutually overlap but also include samples unseen by the other. This is why the evaluation of HT-300K error can indicate the model's ability to interpolate, while the HT-250K validation set can indicate both interpolation and extrapolation capabilities.

8 Notes on the Performance of Descriptors

Table 2 presents the computational cost and the number of features for each descriptor. In terms of representation time (Rep. Time), BoB and SOAP require fractions of a second per sample, while MODA (with 6-31+G* basis set) takes approximately half a second. For SOAP, the time to compute a single element of the RBF kernel matrix (Kernel Time) is about twice as long than for MODA and approximately 3.5 times longer than for BoB. The number of features (n$^{\underline{o}}$ Features) varies among the descriptors, with SOAP having the highest number.

The disparities in representation time and computational cost can be attributed to the intrinsic characteristics of the descriptors (*e.g.*, hyperparameters). MODA, despite requiring more time for representation generation, provides crucial electronic structure information. On the other hand, the higher cost for SOAP's kernel computation can be attributed to the larger number of features that it incorporates. This is due to the dot product calculation needed for the RBF similarity, a process with computational cost proportional to the number of features. Notably, the kernel matrix is real, symmetric and large for sizable datasets, resulting in a significant computational

Descriptor	Rep. Time (s)	Kernel Time (s)	nº Features
BoB	0.008	0.070	128
SOAP	0.007	0.243	1710
MODA	0.452	0.101	600

Table 2: Comparison of time requirements for generating representations (Rep. Time), computing kernels (Kernel Time), and the number of features of the descriptor (nº Features). Times are given per sample, while number of features is dimensionless.

toll during the training step (as it takes $n(n-1)/2$ dot product evaluations for a dataset of size n). However, its impact can be mitigated by dimensionality reduction techniques such as Principal Component Analysis (PCA).

Regarding the impact of the representation length in the performance of the descriptor, we believe that it is more related to the variance of each component (and the information that these carry, as indicated by the covariance with the target property) rather than to the number of components. This can be illustrated as follows: let S be the distance between two data samples, which – in methods based on the "kernel trick" – is the centerpiece for the vast majority of kernels. If ρ_k and and ρ_l are the vector representations of two samples of an arbitrarily large descriptor, S is simply:

$$S = ||\rho_k - \rho_l|| = ||\Delta\rho|| = \sqrt{\sum_{\forall i} \Delta\rho_i^2} \tag{14}$$

where $\Delta\rho_i^2$ corresponds to the variation associated with the i-th component of $\Delta\rho$. Then, the variation of the distance S (or the ordinary derivative of S, *i.e.* dS) can be recasted using differential forms:

$$dS = \sum_{\forall i} d\Delta\rho_i^2 \left(\frac{\partial S}{\partial \Delta\rho_i^2}\right)_{\Delta\rho_{j\neq i}^2} = \frac{1}{2||\Delta\rho||} \sum_{\forall i} d\Delta\rho_i^2 \tag{15}$$

where the partial derivative of S with respect to the i-th variation, $\Delta\rho_i^2$, is

$$\left(\frac{\partial S}{\partial \Delta\rho_i^2}\right)_{\Delta\rho_{j\neq i}^2} = \frac{1}{2\sqrt{\sum_{\forall i} \Delta\rho_i^2}} = \frac{1}{2||\Delta\rho||} \tag{16}$$

The result of eq. 15 shows that the determining effect on the distance and, thus, on the differential performance of a regression model is determined by the variation associated with each component

($\Delta\rho_i^2$). In summary, the effect of having both relevant components presenting strong variations ($\Delta\rho_i^2 \gg 0$) and uninformative features that are hardly varying ($\Delta\rho_i^2 \approx 0$) is more important for the model performance than the total number of components itself. However, when one, or a significant set, of uninformative components substantially contribute ($\Delta\rho_i^2 \geq 0$) to the total variation ($||\Delta\rho||$), then the descriptor is not appropriate for the task entailed (as $dS/d\Delta\rho_i^2 \propto ||\Delta\rho||^{-1} \propto \Delta\rho_i$).

An indicative metric of the quality of the information carried by a component is its covariance with respect to the target property, J_{AB} in our case. As we demonstrate in our variance-covariance analysis (discussed in the main text), MODA is the most appropriate descriptor for this purpose.

References

(1) K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

(2) R. K. Cersonsky, M. Pakhnova, E. A. Engel and M. Ceriotti, *Chem. Sci.*, 2023, **14**, 1272–1285.

(3) S. Vela, F. Mota, M. Deumal, R. Suizu, Y. Shuku, A. Mizuno, K. Awaga, M. Shiga, J. J. Novoa and J. Ribas-Arino, *Nat. Commun.*, 2014, **5**, 4411.

(4) S. Vela, M. Deumal, M. Shiga, J. J. Novoa and J. Ribas-Arino, *Chem. Sci.*, 2015, **6**, 2371–2381.

(5) F. James, *Statistical Methods in Experimental Physics*, WORLD SCIENTIFIC, 2006.

(6) P.-O. Löwdin and H. Shull, *Phys. Rev.*, 1956, **101**, 1730–1739.

(7) P. Pulay and T. P. Hamilton, *J. Chem. Phys.*, 1988, **88**, 4926–4933.

(8) I.-M. Høyvik, *Mol. Phys.*, 2020, **118**, e1765034.

(9) J. J. Girerd, Y. Journaux and O. Kahn, *Chem. Phys. Lett.*, 1981, **82**, 534–538.