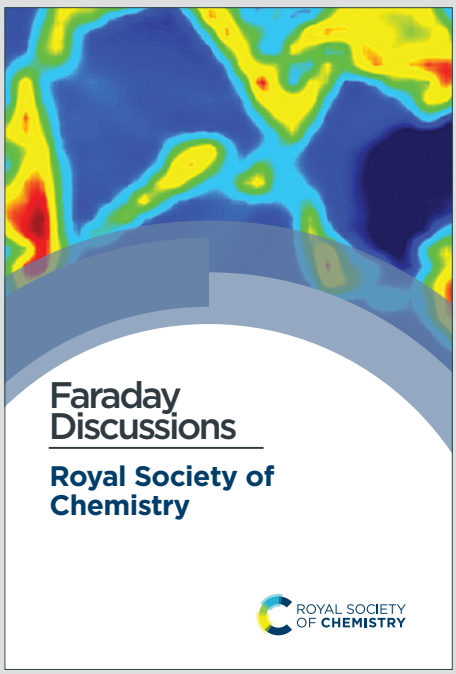
Faraday Discussions

Accepted Manuscript



This article can be cited before page numbers have been issued, to do this please use: C. Baucom, E. Mainas and E. Pieri, *Faraday Discuss.*, 2026, DOI: 10.1039/D6FD00060F.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

View Article Online
DOI: 10.1039/D6FD00060F

Beyond Minimum Energy Conical Intersections: A Data-Driven Reconstruction of the Accessible Intersection Seam

Conner J. Baucom,[a] Eleftherios Mainas[a] and Elisa Pieri*[a]

Received 00th January 20xx, Accepted 00th January 20xx DOI: 10.1039/x0xx00000x

Abstract

Minimum energy conical intersections (MECIs) are widely used to rationalize nonadiabatic relaxation pathways, yet it remains unclear to what extent they provide a representative description of the intersection seam effectively explored during dynamics. In this work, we combine large ensembles of seam points from nonadiabatic simulations with dimensionality reduction and density-based analysis to systematically investigate the structure of the accessed seam regions and its relationship to MECIs. We first establish a robust, physically-meaningful protocol for representing seam morphology across systems. After accounting for rigid-body alignment and permutation symmetry, Cartesian coordinates are found to best preserve structural relationships, while densMAP provides an optimal low-dimensional embedding that captures both geometric variation and sampling density. Applying this framework to ethylene, butadiene, and benzene reveals that the intersection seam is organized into dynamically accessible basins that are not uniformly represented by discrete MECIs. While MECIs associated with highly populated basins are located near density maxima and capture the dominant relaxation pathways, many other MECIs correspond to peripheral or weakly sampled regions. Moreover, neither geometric proximity nor energetic accessibility alone reliably predicts dynamical relevance: geometrically closest MECIs do not always correspond to optimization outcomes, and low-energy MECIs may remain largely unvisited. These results demonstrate that MECIs should be viewed as discrete markers within a broader, continuous, and dynamically weighted seam landscape. A comprehensive understanding of photochemical relaxation therefore requires not only the identification of MECIs, but also an explicit characterization of the regions of the seam that are actually sampled and the pathways connecting them.

Introduction

Photochemical reactions are governed by the breakdown of the Born-Oppenheimer approximation and the ensuing coupling between electronic and nuclear degrees of freedom. Central to this nonadiabatic regime are conical intersections (CIs) - regions of degeneracy between electronic states that enable ultrafast, radiationless population transfer and thereby control excited-state lifetimes, branching ratios, and photochemical outcomes.[1–3] For polyatomic molecules, such degeneracies are not isolated points in nuclear configuration space but instead form high-dimensional (3N-8 locally in internal coordinates, where N is the number of atoms in the system) intersection seams (ISs), reflecting the fact that the degeneracy is lifted only along two specific nuclear directions while persisting in all remaining degrees of freedom.[4–7]

Despite this intrinsic high dimensionality, much of the mechanistic understanding of excited-state dynamics has historically been built around a handful of special points on the seam: the minimum-energy conical intersections (MECIs).[8] As ISs have their own topography, MECIs are defined as minimum energy points on these manifolds, and have long been treated as a structural proxy for photochemical decay pathways.[9,10] This practice is motivated both by computational convenience (MECIs are well-defined optimization targets) and by a simplistic physical analogy to transition states in ground-state reactivity.[11]

Historically, the centrality of the MECIs in excited-state theory reflects not only physical intuition but also practical necessity. For decades, electronic-structure methods capable of treating strong nonadiabatic effects were computationally demanding, and optimization of a single well-defined point on the seam offered a tractable route to mechanistic insight[12]. As a result, large bodies of excited-state literature, particularly in organic photochemistry,[13–15] have relied on comparisons between Franck-Condon geometries, excited-state minima, and MECIs to rationalize reaction pathways and photoproducts. This focus has implicitly elevated the MECIs from a convenient reference structure to a presumed dynamical representative, often without explicit verification that it coincides with the seam regions most frequently accessed during relaxation.

The conceptual justification for this analogy is often framed in terms of a statistical picture of excited-state dynamics. Following photoexcitation, rapid internal conversion and vibrational energy redistribution are assumed to erase memory of the initial Franck–Condon geometry, allowing the nuclear wavepacket to explore the excited-state potential energy surface quasi-thermally.[3,12,16] Within this view, the IS functions as a multidimensional funnel, and the system is expected to relax toward its lowest-energy points, the MECIs, prior to nonadiabatic decay.[17] Under such conditions, MECIs would indeed represent the most probable region of seam access and thus serve as a meaningful structural descriptor of the photoreaction.

However, this reasoning rests on assumptions that are not universally valid. Many photochemical processes occur on ultrafast timescales, comparable to or shorter than vibrational relaxation, and are therefore better described as ballistic rather than statistical.[18] In these regimes, nuclear motion is strongly influenced by the initial excitation conditions and by steep excited-state gradients, and nonadiabatic transitions may occur far from minima on the IS.[19–21] More generally, even when relaxation does occur, regions of distinct geometric and dynamical character may coexist at similar energies.[22–24] As a result, the dynamically accessible portion of the seam (i.e., the subset actually sampled during nonadiabatic dynamics) need not coincide with MECIs.[25]

These considerations motivate the central question addressed in this work: are MECIs genuinely representative of the seam regions accessed during photochemical reactions, or are they convenient but potentially misleading stand-ins for a far richer dynamical landscape? If MECIs do coincide with the most dynamically relevant portions of the seam, then their continued use as mechanistic anchors is well justified. If, however, the effective crossing region is distributed across extended portions of the seam, or depends sensitively on excitation conditions and topography, then analyses based solely on MECIs risk missing key features that control photochemical outcomes. In this sense, the

View Article Online
DOI: 10.1039/D6FD00060F

question of MECI representativeness is inseparable from broader debates about statistical versus non-statistical dynamics and about the extent to which static calculations can stand in for explicitly dynamical descriptions.

Addressing this problem demands a shift in perspective: from identifying individual "important" geometries to learning structure directly from ensembles of dynamical data. Modern nonadiabatic molecular dynamics simulations[19,26,27] can generate thousands of geometries associated with electronic transitions, but the sheer dimensionality of nuclear configuration space[3] obscures patterns that are not obvious in any small set of internal coordinates. In this context, machine-learning and data-driven approaches[28–33] can extract essential information. Dimensionality-reduction techniques[34–36] provide a means to construct low-dimensional embeddings that preserve the intrinsic topography of the seam, while clustering and density-based analyses[37,38] offer a quantitative notion of seam accessibility that goes beyond qualitative inspection. Here, we obtain datasets by running large ensembles of nonadiabatic dynamics for several small molecules and collecting thousands of quasi CI geometries that serve as dynamical proxies for seam access and define relevant regions.

A key methodological challenge in this approach is how to represent, embed, and analyze these high-dimensional data. Specifically, three interrelated questions arise:

- Which molecular representation[39–41] is most appropriate for describing seam geometries, given the need for rotational and translational invariance and the potential presence of molecular symmetry (all limitations of both Cartesian and internal coordinate systems)?
- Which dimensionality-reduction techniques are best suited to capturing the intrinsic topology of the IS, preserving chemically meaningful collective coordinates while avoiding distortions?
- How can the resulting low-dimensional representations be analyzed to identify distinct regions of seam accessibility, particularly in cases where boundaries are diffuse and the dynamical sampling is intrinsically continuous?

To address these questions, we systematically explore combinations of coordinate choices, dimensionality-reduction methods (including both linear and nonlinear embeddings). Rather than imposing a discrete clustering of configurations, we adopt a density-based perspective in which the embedding is interpreted as a continuous landscape of seam accessibility. High-density regions correspond to frequently visited portions of the seam, and local maxima define basins of dynamical accessibility. This framework enables us to assess whether MECIs coincide with the centers of these basins and to establish a general, transferable approach for mapping, comparing, and interpreting ISs using data-driven analysis. By reframing CIs as objects to be statistically learned rather than singularly optimized, this work aims to bridge traditional excited-state theory with emerging data-centric methodologies, and to clarify when - and when not - MECIs remain a faithful representative of photochemical reality.

Methods

This study comprises several stages, each described in detail in the subsections below. A schematic workflow can be found in Figure 1. We first provide a high-level overview of the workflow and its guiding motivations.

Our first step is the construction of large datasets of quasi-CI geometries. These datasets were generated by performing extensive nonadiabatic molecular dynamics simulations on ethylene, butadiene, and benzene and collecting the spawning points (SPs) associated with electronic transitions. Although SPs are not formally CIs - typically exhibiting small residual energy gaps ($\sim$0.05–0.3 eV) between the coupled electronic states - they lie in the vicinity of the IS and therefore serve as practical dynamical proxies for seam access. We use these geometries to reconstruct the dynamically sampled portion of the IS. Our objective is not to accurately reproduce experimental data such as quantum yields or excited state lifetimes, but rather to generate sufficiently large ensembles of seam-access geometries to enable statistically robust analysis. Accordingly, we employ computationally efficient electronic-structure methods that permit broad sampling of the IS.

We next evaluate the ability of several molecular representations, in combination with dimensionality-reduction techniques, to capture seam morphology. A critical preprocessing step is the alignment of SP geometries, particularly when using Cartesian coordinates, where arbitrary rotations and translations can obscure intrinsic structural relationships. Moreover, the three test molecules exhibit high symmetry, and their photochemical pathways frequently involve symmetry- and permutation-equivalent distortions. Thus, accurate reconstruction of the seam requires representations and/or alignment procedures that are invariant with respect to rotation, translation, and atom permutation. Without enforcing these invariances, geometries corresponding to the same physical region of the seam may appear artificially distant in the chosen representation.

Finally, once an appropriate molecular representation and embedding that accurately preserve local and global seam morphology have been identified, we investigate how the MECIs relate to the dynamically sampled seam landscape. Specifically, we analyze the distribution of SPs in the low-dimensional embedding to identify densely populated basins corresponding to frequently accessed regions of the IS. We then compare the positions of the MECIs with the centers of these basins to assess whether they coincide with dynamically relevant regions of the seam or instead lie in less frequently visited areas.

Dataset Creation

The ground state minimum energy structures were optimized using ωB97x-D3/6-311+G*. Frequencies and Hessian matrix were obtained on the optimized structure with the same level of theory. The Hessian matrix was then used to obtain a Wigner distribution at 0 K, yielding 992 initial conditions (positions and momenta) for ethylene, 485 for benzene and 271 for butadiene. *Ab initio* multiple spawning (AIMS)[42] was used to propagate each initial condition starting from S_1 for ethylene and butadiene, and S_2 for benzene. The electronic structure methods used for the nonadiabatic dynamics simulations are SA2-CASSCF(2,2)/6-31G* for ethylene, SA3-CASSCF(4,3)/6-31G* for butadiene and (t_0=0.15)-FOMO-CASCI(6,5)/6-31G for benzene,[43,44] as suggested by Autopylot.[45] The threshold of the derivative coupling vectors for spawning events was set to 10 au. All trajectories were propagated for 1 ps or until no more living trajectory basis

View Article Online
DOI: 10.1039/D6FD00060F

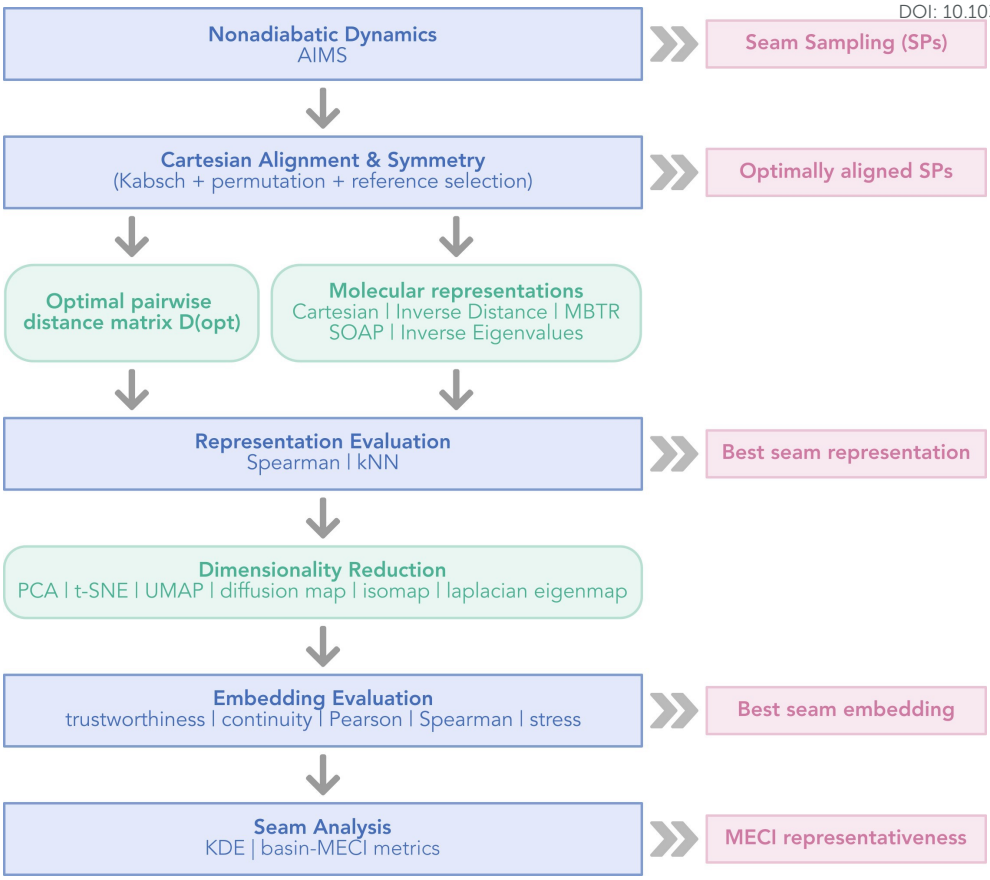


Figure 1 Workflow for the analysis of the seam sampling. Spawning points (SPs) are aligned to remove rotational, translational, and permutational invariances, yielding a reference distance matrix $D^{(\mathrm{opt})}$. Molecular representations are benchmarked against $D^{(\mathrm{opt})}$ using local and global metrics, followed by dimensionality reduction and evaluation of embedding quality. The resulting low-dimensional representations are analyzed to identify populated seam regions and their relationship to MECIs.

functions with a population higher than 5% were found on excited states. Trajectory basis functions evolving on the ground-state surface were removed if their overlap with the remaining functions remained negligible for longer than 5 fs. The SPs collected were then used as starting guess for MECI optimizations. All the presented calculations were performed using TeraChem [46–48] and FMS90. The final numbers of SPs and MECIs gathered per molecule are summarized in Table 1.

Table 1 Total number of SPs and MECIs collected per molecule and IS.

Molecule	Seam	# of SPs	# of MECIs
Ethylene	S_1/S_0	2557	9
Butadiene	S_1/S_2	881	8
	S_0/S_1	1246	33
Benzene	S_1/S_2	1717	3
	S_0/S_1	3657	5

Sampling considerations and statistical robustness Finally, we comment on what constitutes sufficient sampling of the intersection seam for statistically robust analysis. Owing to the high dimensionality and complexity of the seam, sampling can never be formally complete by definition; instead, the objective is to obtain a representative ensemble of geometries that captures the dominant, dynamically accessible regions. In practice, we assess convergence in terms of (i) the stability of global and local structural metrics (e.g., distance matrix correlations, nearest-neighbor overlap) upon addition of new samples, and (ii) the consistency of low-dimensional embeddings and basin populations. The resulting sampling is inherently biased toward specific segments of the seam that are dynamically reachable, and therefore depends strongly on the initial conditions, such as the mode of excitation, as well as on the underlying electronic structure description and the nonadiabatic dynamics methodology. While these latter factors can influence the detailed distribution of sampled geometries, we expect that, when suitably chosen, different electronic structure methods and nonadiabatic dynamics approaches converge toward the same qualitative picture of the seam regions associated with large-amplitude, low-energy distortions. Within this framework, statistical robustness is therefore defined operationally by the reproducibility of structural and topological features of the sampled seam, rather than

View Article Online
DOI: 10.1039/D6FD00060F

by completeness in a formal sense.

Molecular representations and embedding techniques

Dimensionality reduction To identify a low-dimensional representation that faithfully captures the morphology of the IS, we evaluate a set of dimensionality-reduction techniques that differ in how they preserve geometric and statistical structure. In particular, we consider methods that emphasize (i) global variance, (ii) local neighborhood structure, (iii) manifold/geodesic distances, and (iv) sampling density. This allows us to assess which aspects of the seam are most critical for constructing meaningful embeddings.

We first include *Principal Component Analysis* (PCA),[49] a linear baseline that identifies orthogonal directions of maximum variance. PCA provides a useful reference for global structure preservation and offers interpretable axes corresponding to collective deformation modes. However, as a linear method, it cannot capture nonlinear manifold structure, potentially leading to distortions of curved seam regions and mixing of distinct dynamically accessible configurations.

To probe local neighborhood preservation, we consider *t-Distributed Stochastic Neighbor Embedding* (*t*-SNE)[50] and *Uniform Manifold Approximation and Projection* (UMAP).[51] Both methods construct embeddings by preserving local similarities, making them well suited for identifying clusters and locally coherent regions of the seam. While *t*-SNE strongly emphasizes local structure at the expense of global geometry, UMAP attempts to balance local and global relationships through a graph-based construction. Both methods are stochastic and depend on hyperparameters controlling the effective neighborhood size. Because the identification of seam basins depends not only on geometry but also on the distribution of sampled configurations, we include *densMAP*,[52] a density-preserving variant of UMAP. In addition to maintaining local neighborhood relationships, densMAP explicitly constrains the embedding to preserve relative sampling densities, providing a framework to assess whether densely populated regions of the IS are faithfully represented in low-dimensional space.

Finally, to capture intrinsic manifold structure beyond local neighborhoods, we include *Isomap*[53] and *Diffusion Maps*.[54] These methods approximate the geometry of the underlying low-dimensional manifold: Isomap does so by preserving geodesic distances computed from shortest paths on a neighborhood graph, while diffusion maps encode connectivity through a diffusion process that integrates multiple paths between points. These approaches are particularly relevant for smoothly varying reaction coordinates but can be sensitive to sampling irregularities and graph construction parameters.

Together, this set of methods enables a systematic comparison of how different notions of structure (global geometry, local neighborhoods, manifold connectivity, and sampling density) impact the representation of ISs.

Molecular representations The choice of molecular representation determines how geometric relationships between SP structures are encoded prior to dimensionality reduction. In the context of ISs, an effective representation must balance several competing requirements, including invariance to rotations, translations, and permutations, while retaining sufficient information to distinguish structurally and chemically distinct configurations. Cartesian coordinates provide a direct and complete description of molecular geometry, but require explicit preprocessing to enforce these invariances (see next Section). To assess how different representations capture seam morphology, we consider four additional descriptors that encode complementary aspects of molecular structure, ranging from local atomic environments to global geometric relationships.

- **Smooth Overlap of Atomic Positions (SOAP)** descriptors represent each atomic environment as a continuous atomic density expanded in a basis of radial functions and spherical harmonics, and compare environments through the overlap of these densities.[55] SOAP provides a rich, rotation- and permutation-invariant description of local geometry. However, because it operates at the level of atomic environments, global molecular structure must be recovered through aggregation across atoms.

- **Many-Body Tensor Representation (MBTR)** encodes molecular structure through distributions of geometric features such as inverse distances, bond angles, and higher-order many-body terms.[56] By incorporating two- and three-body correlations, MBTR captures both local and medium-range structural information. This flexibility comes at the cost of increased dimensionality and sensitivity to discretization choices.

- **Inverse Distance Matrix** representations encode molecular geometry through pairwise inverse interatomic distances.[57] This provides a direct description of global molecular structure and offers a natural, automatically generated set of internal-coordinate-like features that remain well defined even in the presence of connectivity changes. However, the representation depends on a consistent atom ordering and can mix equivalent structures if permutation symmetry is not properly resolved.

- **Sorted Inverse Eigenvalues** of the distance (or Coulomb-like) matrix provide a compact, permutation-invariant descriptor derived from the matrix spectrum.[57,58] While this representation removes the need for explicit atom indexing, it introduces a loss of information, as distinct molecular geometries may share identical spectra.

These representations span a range of design principles from local environment-based descriptors to global, invariant encodings, allowing us to assess which types of structural information are most critical for accurately describing seam morphology.

Preprocessing of Cartesian coordinates Before dimensionality reduction and any subsequent analysis, all SP geometries expressed in Cartesian coordinates must be brought into a common reference frame so that distances in feature space reflect genuine structural differences. For the highly symmetric systems considered here, this preprocessing step involves three components: (i) rigid-body alignment, (ii) resolution of molecular permutation symmetry, and (iii) selection of appropriate reference structures. A schematic summary of the algorithm is provided in Figure 2A. Rigid-body alignment is performed using the Kabsch algorithm,[59,60] allowing reflections (i.e., optimization over $\mathrm{O}(3)$)[61,62] so that enantiomeric distortions are treated as equivalent (see Figure 2B). To resolve permutation symmetry (see Figure 2C), we employ a brute-force search over element-constrained permutations, which guarantees identification of the global minimum-RMSD correspondence despite its factorial cost for highly symmetric molecules. Alignment can be carried out with respect to

View Article Online
DOI: 10.1039/D6FD00060F

either a single reference structure (e.g., the medoid of the MECI set) or multiple references (e.g., the full set of MECIs). Comparison with the optimal pairwise distance matrix $D^{(\mathrm{opt})}$ indicates that single-reference alignment is sufficient for systems with limited connectivity changes (i.e., benzene), whereas multi-reference alignment more accurately preserves local neighborhood structure in systems exhibiting substantial structural rearrangements (i.e., ethylene and butadiene). Full details of the alignment procedure, alternative permutation strategies, and evaluation of reference-selection methods are provided in the Supporting Information.

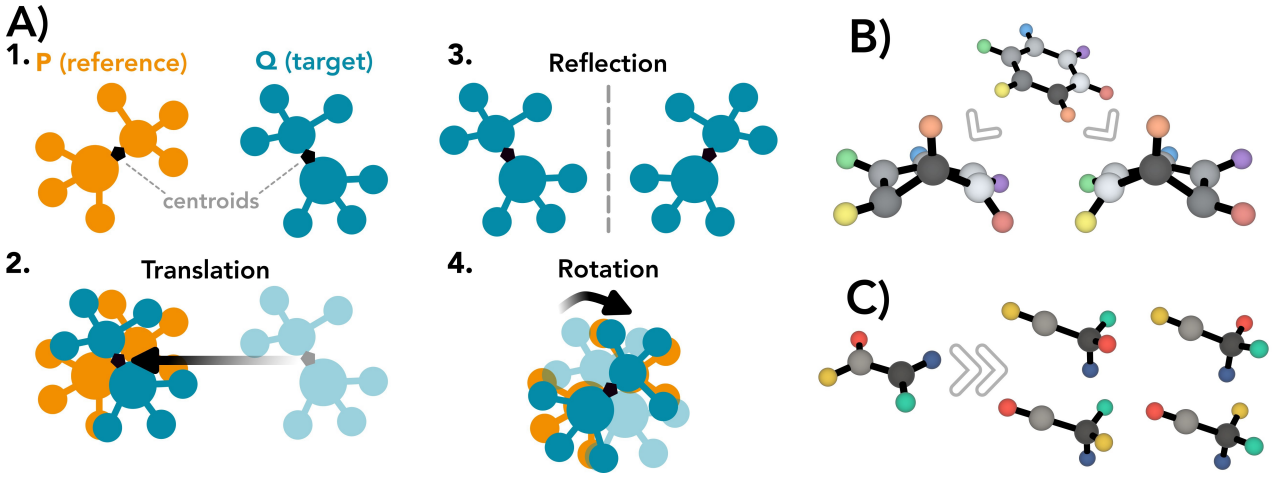


Figure 2 A) Schematic representation of the Kabsch algorithm with reflections. On the right: examples of symmetry- and permutation-related SPs; carbon atoms are shown in different shades of gray and hydrogen atoms in distinct colors to facilitate comparison between equivalent geometries. B) Example of chiral SPs: symmetry-breaking deformations of benzene generate pairs of enantiomeric geometries that are energetically and physically equivalent; these must be identified and treated as the same configuration when constructing the seam representation. C) Example of permutationally identical SPs: hydrogen migration in ethylene yields eight geometrically distinct yet permutation-equivalent SPs (only four are shown), due to the indistinguishability of the transferring hydrogen and the symmetry-related attachment sites on the adjacent carbon.

Representation and embedding quality metrics To quantify how well each molecular representation and embedding combination preserves seam morphology, we evaluate both representation quality and embedding fidelity. Because dimensionality reduction operates on a chosen molecular representation, it is first necessary to assess how well each representation captures the underlying geometric relationships between structures. To this end, we compare distances in representation space D^{repr} to the reference distance matrix $D^{(\mathrm{opt})}$, defined from optimal RMSD alignment. Representation quality is evaluated using both local and global metrics. Local structure preservation is assessed via the *k-nearest-neighbor (kNN) overlap*[63] between D^{repr} and $D^{(\mathrm{opt})}$, defined as the average fraction of shared neighbors:

$$\mathrm{kNN\ overlap} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{N}_k^{\mathrm{repr}}(i)\cap\mathcal{N}_k^{(\mathrm{opt})}(i)|}{k}, \tag{1}$$

where N is the total number of points, and $\mathcal{N}_k^{\mathrm{repr}}(i)$ and $\mathcal{N}_k^{(\mathrm{opt})}(i)$ denote the sets of k nearest neighbors of point i in D^{repr} and $D^{(\mathrm{opt})}$, respectively. Global agreement is quantified using the *Spearman rank correlation coefficient*[64] ρ between pairwise distances in the representation space and those in $D^{(\mathrm{opt})}$, assessing the preservation of their relative ordering:

$$\rho = \mathrm{corr}_{\mathrm{Spearman}}\left(\{D_{ij}^{\mathrm{repr}}\},\{D_{ij}^{(\mathrm{opt})}\}\right), \tag{2}$$

We then evaluate how well each embedding preserves the structure of the chosen representation using complementary metrics that probe local geometry, global relationships, and sampling density. Local neighborhood preservation is assessed using *trustworthiness* and *continuity*.[65] Trustworthiness T measures the extent to which neighbors in the low-dimensional embedding are also neighbors in the original space, thereby penalizing spurious neighbor relationships introduced by the embedding:

$$T(k) = 1 - \frac{2}{Nk(2N-3k-1)}\sum_{i=1}^{N}\sum_{j\in U_k(i)}(r_{ij}-k), \tag{3}$$

where the prefactor normalizes the penalty term such that $T(k)$ ranges between 0 and 1, with $T(k)=1$ corresponding to perfect neighborhood preservation. Here, $U_k(i)$ is the set of points that are among the k nearest neighbors of point i in the embedding but not in the original space, and r_{ij} is the rank of point j in the ordered list of neighbors of i in the original space. Continuity C provides the complementary measure by measuring the fraction of true neighbors in the high-dimensional space that remain neighbors after dimensionality reduction, thus penalizing neighborhoods that are broken during embedding:

$$C(k) = 1 - \frac{2}{Nk(2n-3k-1)}\sum_{i=1}^{N}\sum_{j\in V_k(i)}(r'_{ij}-k), \tag{4}$$

View Article Online
DOI: 10.1039/D6FD00060F

where $V_k(i)$ contains points that are neighbors in the original space but not in the embedding, and r'_{ij} is their rank in the embedding.
Global geometric structure is evaluated using two metrics based on pairwise distances. Pairwise distances in the high- ($\mathbf{P}$) and low- ($\mathbf{z}$) dimensional spaces are defined as

$$D_{ij}^{\mathrm{HD}} = \mathrm{RMSD}(\mathbf{P}_i, \mathbf{P}_j), \quad D_{ij}^{\mathrm{LD}} = \|\mathbf{z}_i - \mathbf{z}_j\|, \tag{5}$$

where $\mathrm{RMSD}(\mathbf{P}_i, \mathbf{P}_j)$ denotes the root-mean-square deviation between geometries i and j after optimal rigid-body alignment. The *Spearman correlation*[64] between the high- and low-dimensional pairwise distance matrices measures how well the embedding preserves the rank ordering of distances and therefore the large-scale arrangement of points on the seam:

$$\rho = \mathrm{corr}_{\mathrm{Spearman}}\left(\{D_{ij}^{\mathrm{HD}}\}, \{D_{ij}^{\mathrm{LD}}\}\right), \tag{6}$$

Stress[66] (σ) quantifies the overall distortion of distances:

$$\sigma = \sqrt{\frac{\sum_{i<j}\left(D_{ij}^{\mathrm{HD}} - D_{ij}^{\mathrm{LD}}\right)^2}{\sum_{i<j}\left(D_{ij}^{\mathrm{HD}}\right)^2}}. \tag{7}$$

Because the identification of seam basins depends not only on geometric relationships but also on the distribution of sampled configurations, we additionally evaluate how well each embedding preserves local sampling density. For each point i, a local density estimate ρ_i is computed in both the original space (using $D^{(\mathrm{opt})}$) and the low-dimensional embedding using a k-nearest-neighbor estimator. Density preservation is then quantified via a rank-based metric ρ_{dens},

$$\rho_{\mathrm{dens}} = \mathrm{corr}_{\mathrm{Spearman}}\left(\{\rho_i^{\mathrm{HD}}\}, \{\rho_i^{\mathrm{LD}}\}\right), \tag{8}$$

which measures whether regions that are densely sampled in the original space remain dense after dimensionality reduction. In summary, Spearman correlation assesses preservation of the rank ordering of distances, stress quantifies absolute distance distortion, and the density rank correlation captures preservation of sampling heterogeneity across the seam. All metrics are normalized such that higher trustworthiness, continuity, Spearman correlation, and density rank correlation indicate better preservation, whereas lower stress corresponds to lower geometric distortion.

Seam basin analysis

To assess whether MECIs lie at the centers of dynamically accessible regions of the IS, we analyze the distribution of SP geometries in the selected low-dimensional embedding. After identifying optimally aligned Cartesian coordinates and densMAP as the combination that best preserves seam morphology (see Results), both SP geometries and the corresponding MECIs are projected into a common embedding space. For this analysis, the first three components of the densMAP embedding are used for basin assignment, while two-dimensional projections are used solely for visualization. Denoting by $\mathbf{z}_i \in \mathbb{R}^3$ the embedded coordinates of the i-th SP geometry and by $\mathbf{z}_m^{\mathrm{MECI}}$ the embedded coordinates of the m-th MECI, the density of SP geometries in the embedding $\hat{f}$ is estimated using kernel density estimation (KDE),[67]

$$\hat{f}(\mathbf{z}) = \frac{1}{Nh^3}\sum_{i=1}^{N} K\left(\frac{\|\mathbf{z} - \mathbf{z}_i\|}{h}\right), \tag{9}$$

where N is the number of SP geometries, h is the kernel bandwidth, and K is a radially symmetric kernel function (here taken as a Gaussian kernel). The resulting density field provides a continuous approximation of the dynamically sampled seam landscape.
Local maxima of $\hat{f}(\mathbf{z})$ are identified by evaluating the KDE on a regular grid of 200 points. A grid cell is classified as a maximum if its density equals the largest value within a cubic neighborhood of half-width δ (in grid cells) and exceeds a threshold $\varepsilon \hat{f}_{\max}$. Here, δ controls the minimum separation between distinct maxima, while $\varepsilon \in (0,1)$ suppresses low-density artifacts at the periphery of the embedding. The parameters δ and ε are selected to ensure robustness of the identified basins. We first vary δ over the range $[1,25]$ at fixed $\varepsilon = 0.01$, and monitor both the number of detected maxima and the mean distance $\bar{d}_m$ between each MECI and its nearest maximum. The resulting sensitivity curves exhibit a step-like behavior: within plateau regions, both the basin count and $\bar{d}_m$ remain stable, whereas at transition points nearby maxima merge and $\bar{d}_m$ increases. The value of δ is chosen within the lowest-$\bar{d}_m$ plateau, ensuring that the detected maxima correspond to stable features of the density landscape. With δ fixed, the threshold ε is then decreased from 0.1 to 0.001 until the basin count ceases to be stable; the final value is taken as the smallest ε for which no additional maxima appear, ensuring that spurious low-density peaks are excluded. The resulting discrete set of maxima is interpreted as basin centers $\{\mathbf{z}_k^*\}$ corresponding to frequently accessed regions of the seam.[68]
We then quantify the relationship between MECIs and these dynamically populated basins using two complementary measures. First, the local density value at the embedded coordinates of each MECI is evaluated,

$$\hat{f}_m = \hat{f}\left(\mathbf{z}_m^{\mathrm{MECI}}\right), \tag{10}$$

which measures how frequently the surrounding region of the seam is accessed by the dynamics. Second, the distance between each MECI and the nearest density maximum is computed as

$$d_m = \min_k \|\mathbf{z}_m^{\mathrm{MECI}} - \mathbf{z}_k^*\|. \tag{11}$$

View Article Online
DOI: 10.1039/D6FD00060F

Small values of d_m indicate that the MECI lies close to the center of a frequently visited basin, whereas larger values suggest that the MECI is located at the periphery of the dynamically sampled region.
To complement this analysis, each i-th SP geometry is assigned to the nearest density basin k according to

$$k(i) = \arg\min_k \|\mathbf{z}_i - \mathbf{z}_k^*\|, \tag{12}$$

which partitions the embedding into regions associated with each density maximum. The population of basin k is then estimated as

$$N_k = |\{i : k(i) = k\}|, \tag{13}$$

providing a measure of how frequently that region of the seam is accessed during the dynamics. The relationship between MECIs and dynamically sampled seam regions is evaluated by measuring (i) the distance between each MECI and the nearest density maximum and (ii) the population of the basin containing the MECI. These quantities indicate whether MECIs lie near the centers of highly populated seam basins or instead occupy peripheral regions of the dynamically sampled landscape.
To quantify the relationship between density-based basin assignments and MECI-based classifications, we evaluate the agreement between different partitions of the SP ensemble using the adjusted Rand index (ARI)[69] and normalized mutual information (NMI).[70] Specifically, we compare (i) basin assignments with alignment-based MECI labels (i.e., which MECI is closest to each SP in terms of RMSD), and (ii) alignment-based labels with the MECIs reached upon optimization. These metrics provide complementary measures of clustering agreement, with ARI quantifying exact label correspondence and NMI capturing shared information between partitions. In this context, high values of ARI and NMI indicate that the two partitions provide a consistent description of the SP ensemble, corresponding to a near one-to-one mapping between basins and MECIs (or between alignment and optimization assignments). Conversely, lower values reflect the presence of merging or splitting between clusters, indicating that multiple MECIs may contribute to the same dynamically sampled region, or that geometrically similar structures evolve toward different MECIs upon optimization.

Results and Discussion

Analysis of MECI types

We note that the assignment of each SP geometry to a specific MECI type via optimization should be interpreted with some caution. The mapping between SPs and MECIs is not strictly unique and may depend on details of the optimization procedure, such as the algorithm employed. More fundamentally, the IS is a continuous, high-dimensional object, and its representation in terms of a discrete set of MECI structures constitutes an approximation. As a consequence, classifying SP geometries solely based on the identity of the optimized MECI may not provide a fully robust partitioning of the seam. This motivates the use of density-based approaches in the low-dimensional embedding presented in a later Section, which allow the identification of dynamically relevant regions of the seam without relying on discrete MECI assignments.

Ethylene Ethylene photodynamics is characterized by a competition between internal conversion pathways involving large-amplitude distortions of the C=C bond and dissociative channels enabled by the high excess energy and gas-phase environment. The non-dissociative SP geometries predominantly sample two classes of nonadiabatic distortions: (i) hydrogen migration from one carbon to the other, leading to ethylidene-like configurations, and (ii) torsion around the C=C bond accompanied by pyramidalization at one or both carbon centers. This picture is in close agreement with previous theoretical studies of ethylene photodynamics, which identify torsion and ethylidene formation as the dominant decay pathways.[20] Dissociative configurations are also observed, including C–C bond cleavage, H-atom loss, and H_2 elimination. MECI optimization of the SP ensemble identifies nine distinct MECI families, for which both the relative energy with respect to the Franck–Condon point and the fraction of SPs converging to each structure are reported in Figure 3. The resulting distribution reveals a clear hierarchy of dynamically accessible relaxation pathways. The largest fraction of SPs converges to torsional MECIs (36.7%), confirming that torsion about the C=C bond is the dominant nonadiabatic decay coordinate. A comparable portion of the population relaxes to ethylidene-like MECIs (25.8% bent and 20.0% linear), indicating that hydrogen migration is a major competing pathway. Together, these two classes account for over 80% of the sampled seam access. Dissociative MECIs contribute more modestly to the overall population. H-loss pathways account for a non-negligible fraction (12.8%), while other channels each contribute less than 5%. Despite their lower statistical weight, all MECIs lie within ∼6 eV of the Franck–Condon point, reflecting the large excess energy available under photoexcitation conditions. Overall, these results indicate that while dissociation pathways are accessible, ethylene photodynamics is dominated by non-dissociative seam regions associated with torsion and hydrogen migration. The relative populations and energies further suggest that the IS is structured into a small number of highly populated basins corresponding to distinct mechanistic classes, with the most frequently accessed regions associated with lower-energy MECIs (albeit not exclusively).

Butadiene Butadiene photodynamics exhibits both similarities to and important differences from ethylene. Sampling of the S_1/S_2 IS reveals eight distinct MECI families (Figure 4, left), with three structures dominating the population. These are primarily characterized by torsion about one of the three C–C bonds. In particular, torsion around the terminal C–C bonds is more frequently accessed than torsion about the central bond, indicating a preference for distortions that preserve conjugation over those that more strongly disrupt the π-system. The corresponding SP geometries display relatively limited structural variability compared to those associated with the lower S_0/S_1 seam, suggesting that the S_1/S_2 seam is accessed within a more confined region of configuration space. In contrast, the S_0/S_1 seam exhibits significantly greater structural diversity, with 33 distinct MECI families identified. Of these, only four account for the majority of the population (Figure 4, bottom left), while each of the remaining structures are accessed by less than 5% of the SPs. The most populated MECIs on this seam are predominantly characterized by pyramidalization at carbon centers, reflecting a shift from torsional to out-of-plane distortion as the dominant decay coordinate. Our results are in good qualitative agreement with prior work,[71] supporting a picture of butadiene photodynamics governed by terminal-bond torsion and distributed relaxation across multiple seam regions. Similarly to ethylene,

View Article Online
DOI: 10.1039/D6FD00060F

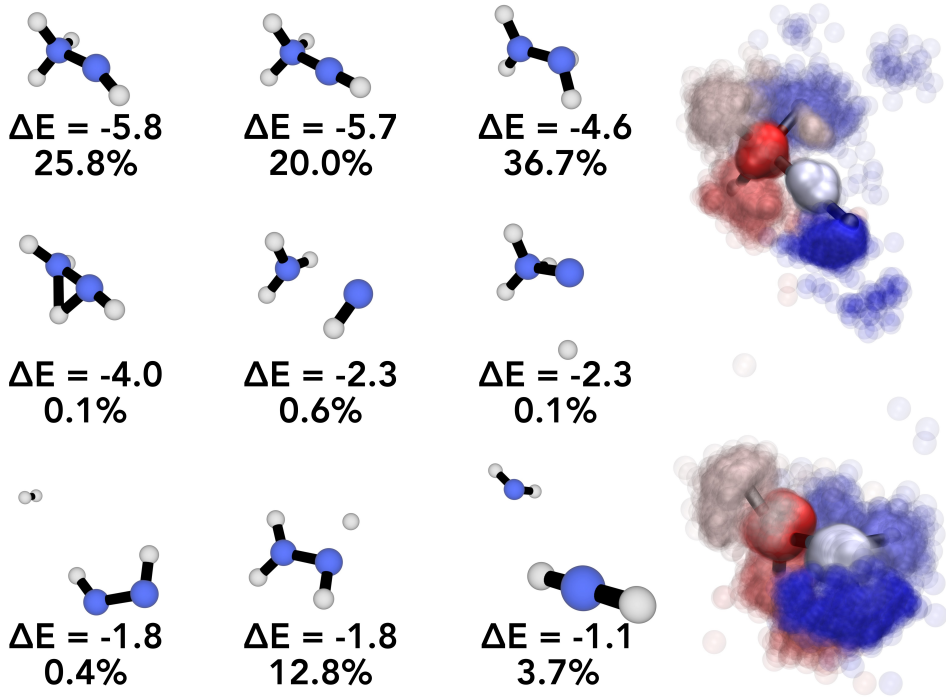


Figure 3 3D structures of the ethylene MECI families (left), with the energy relative to the Franck–Condon point (in eV) and the percentage of SP geometries that converge to each MECI upon optimization; bonds are drawn based on the default distance cutoff implemented in RDKit. Superposition of aligned SP geometries for the dominant deactivation channels, with atoms colored by index (right): ethylidene-like SPs overlaid onto the linear ethylidene MECI (top) and torsional SPs overlaid onto the torsional–pyramidalized MECI (bottom).

the most frequently accessed S_0/S_1 MECIs correspond to low-energy regions of the seam, albeit not exclusively. However, in contrast to ethylene, the higher-energy MECIs form a set of nearly degenerate structures with comparable (lower) populations. The coexistence of many such states indicates that the S_0/S_1 IS in butadiene contains extended regions that are both relatively flat and structurally rugged. At the same time, the dominant decay pathways remain well defined and are governed by a small number of low-energy basins.

Benzene In contrast to ethylene and butadiene, where double-bond isomerization and large-amplitude backbone distortions are accessible, benzene photodynamics is constrained by the rigidity of the aromatic ring. As a result, the dominant distortions associated with both ISs involve out-of-plane deformations, carbon pyramidalization, and symmetry breaking (Figure 4, right) rather than large-scale rearrangements of the carbon framework. Similarly to butadiene, the higher S_1/S_2 seam is accessed through a limited number of configurations, with three distinct MECI families identified. Among these, one pathway clearly dominates the population, while the others contribute only marginally. The associated distortions are relatively modest, with the carbon framework remaining largely intact due to conjugation, and structural deviations primarily involving out-of-plane motion of hydrogen atoms. This behavior suggests a more localized and constrained exploration of configuration space on the upper seam. In contrast, the lower S_0/S_1 seam exhibits greater structural diversity, with five MECI families identified from the SP ensemble, compared to seven obtained from enhanced sampling of the seam.[23] This discrepancy indicates that not all regions of the seam are dynamically accessible, and that dynamical sampling is essential to identify the most relevant MECI types. The accessible MECIs are characterized by more pronounced ring distortions and symmetry breaking, with the relative positioning of hydrogen atoms playing a key role in defining distinct structures. Notably, the most frequently accessed MECIs are not the lowest-energy ones, in contrast to both ethylene and butadiene. In particular, a hydrogen-transfer MECI, although the lowest in energy, is only rarely accessed. This suggests that reaching this region of the seam requires traversal along a distinct reaction coordinate that is not efficiently sampled under the present dynamical conditions. Overall, these results indicate that in benzene, dynamical accessibility is governed more strongly by the topology of the relaxation pathways than by energetic proximity alone.

Selection of molecular representation and embedding for the seam space

After this preliminary characterization of the sampled seams, we identify the representation and dimensionality reduction technique that best preserve seam morphology, with the goal of constructing a robust and transferrable protocol of seam representation for future studies.

Optimal molecular representations

We first evaluate the ability of each molecular representation to reproduce the geometric relationships between SPs, as defined by the optimal pairwise distance matrix $D^{(\mathrm{opt})}$. This $N \times N$ matrix, constructed from fully aligned and permuted pairs of N Cartesian geometries, serves as a reference describing structural similarity on the IS. Representation quality is assessed using both global (Spearman correlation, Eq. 2) and local (k-nearest-neighbor overlap, Eq. 1) metrics. The performance of each molecular representation in the five datasets is

View Article Online
DOI: 10.1039/D6FD00060F

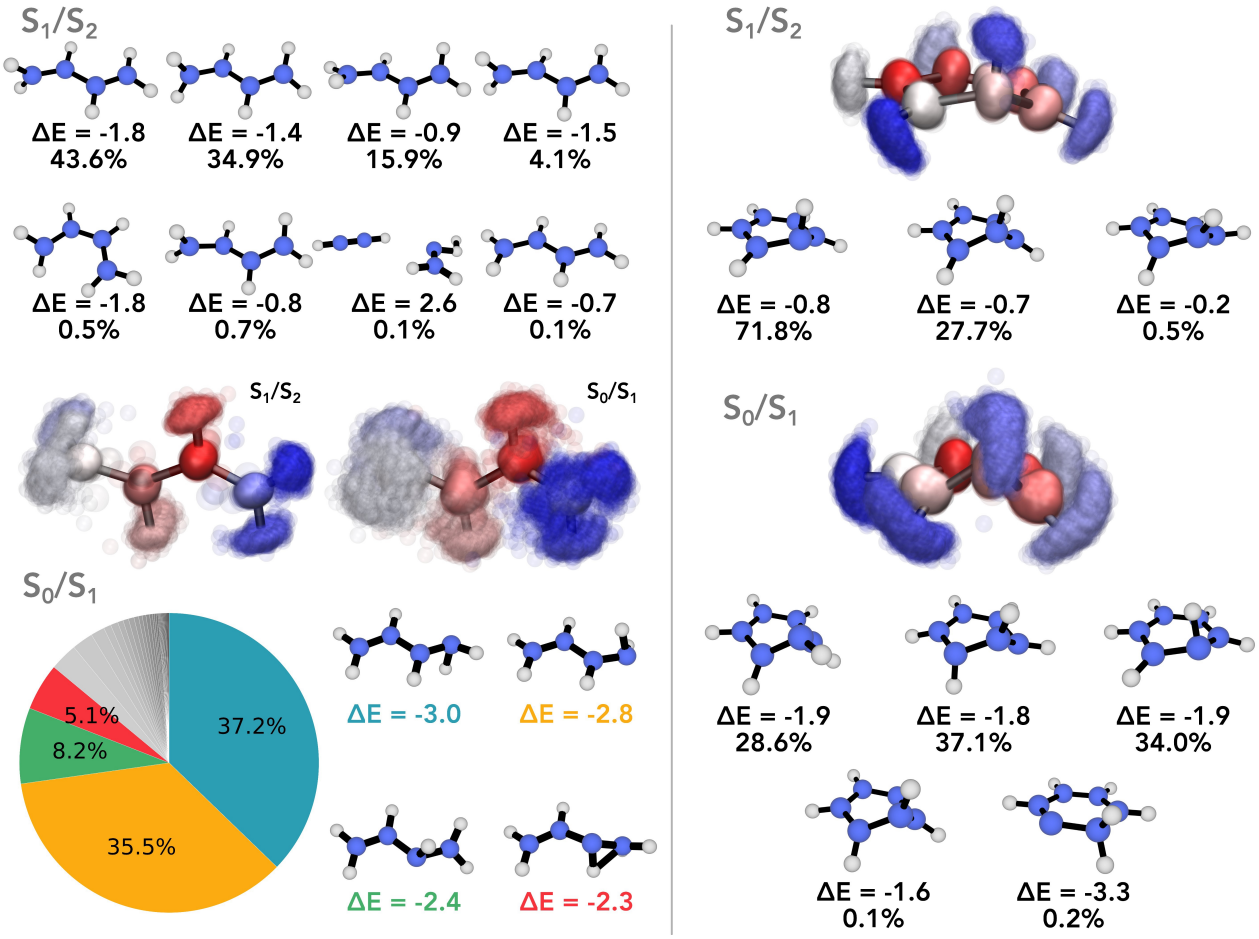


Figure 4 3D structures of the butadiene (left) and benzene (right) MECI families for all sampled ISs, with the energy relative to the Franck–Condon point (in eV) and the percentage of SP geometries that converge to each MECI upon optimization. Bonds are drawn based on the default distance cutoff implemented in RDKit. For the S_0/S_1 IS of butadiene, only the 4 most representative MECIs are shown (see SI for comprehensive data); the pie chart reports in gray shades the MECIs with SP labeling lower than 5%. Superpositions of aligned SP geometries onto the most representative MECI of each seam are shown, with atoms colored by index (right).

summarized in Figure 5.

Across all systems and electronic-state manifolds, aligned Cartesian coordinates consistently yield the highest agreement with $D^{(\mathrm{opt})}$, both in terms of rank correlation and neighborhood preservation. This indicates that, once rotational, translational, and permutational invariances are properly resolved, Cartesian coordinates provide the most faithful representation of molecular geometry. The performance gap is particularly pronounced for the more flexible systems (ethylene and butadiene), where preserving atom-wise correspondence is essential to maintaining meaningful structural relationships.

Among alternative representations, inverse distance matrices perform reasonably well, particularly in terms of global structure, but show a systematic reduction in local neighborhood fidelity. This likely reflects the loss of orientational information and the presence of degeneracies in distance-based encodings. MBTR exhibits intermediate performance, with improved behavior for more rigid systems such as benzene, suggesting that many-body correlations become more informative when structural variability is limited. In contrast, SOAP descriptors show comparatively poor agreement with $D^{(\mathrm{opt})}$ across all datasets. This is consistent with their design as local environment descriptors, which does not explicitly encode global molecular geometry and therefore struggle to preserve the structural relationships relevant for seam analysis. The poorest performance is observed for inverse eigenvalue representations, highlighting the significant information loss associated with spectral compression.

It is important to note that this comparison is performed with respect to a reference metric defined in Cartesian space. As a result, representations that directly encode Cartesian geometry are expected to perform better in this benchmark. The present analysis should therefore be interpreted as an assessment of geometric fidelity relative to RMSD, rather than a general evaluation of representation quality across all possible tasks. In particular, representations such as SOAP or MBTR, which are designed for transferability and learning across chemical space, may not be optimized for reproducing exact pairwise structural distances.

Taken together, these results indicate that the primary challenge in representing seam geometries lies not in the complexity of the descriptor, but in the consistent enforcement of geometric invariances during preprocessing (see Supporting Information for comprehensive data). Once these invariances are properly addressed, Cartesian coordinates provide a robust and high-fidelity representation of seam structure, while alternative descriptors introduce varying degrees of information loss that can impact both local and global structural relationships.

View Article Online
DOI: 10.1039/D6FD00060F

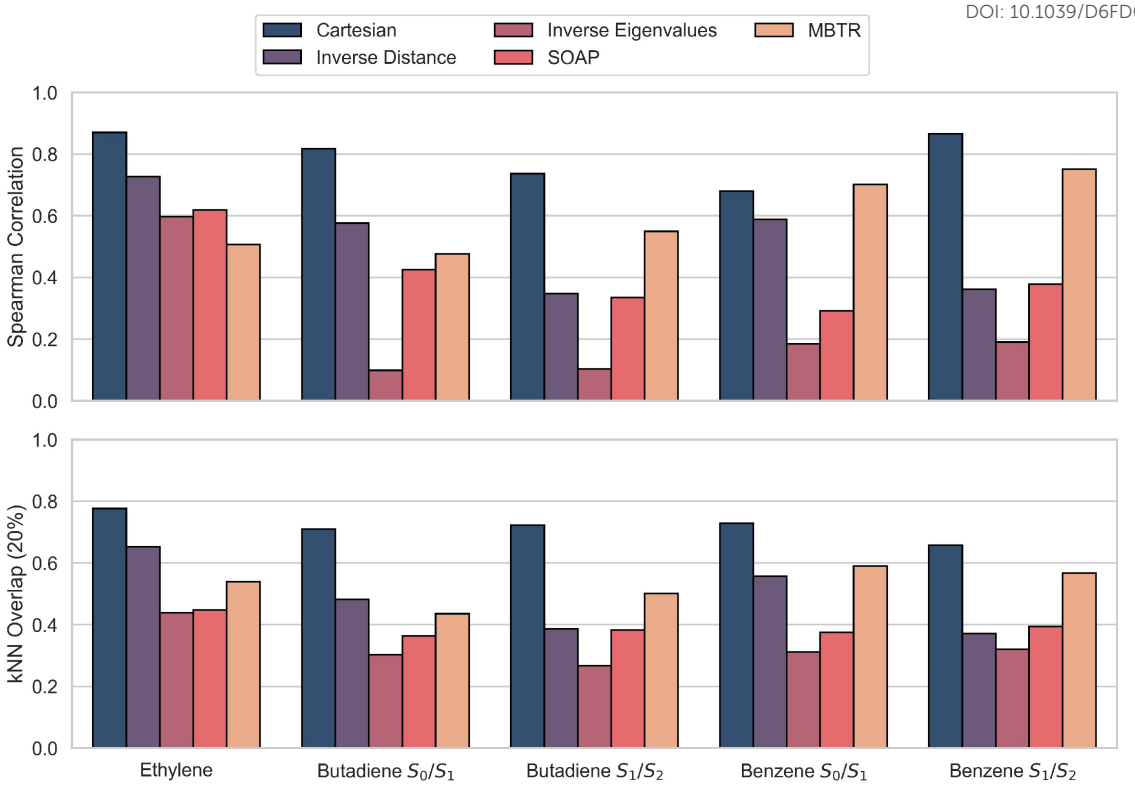


Figure 5 Comparison of molecular representations in reproducing the reference distance matrix $D^{(\mathrm{opt})}$ across all datasets. Top: Spearman correlation between pairwise distances in representation space and $D^{(\mathrm{opt})}$, measuring preservation of global structure. Bottom: k-nearest-neighbor (kNN) overlap ($k = 20\%$) quantifying local neighborhood preservation.

Optimal embedding selection We then evaluate the ability of the different dimensionality reduction methods to preserve the structure of the intersection seam across all datasets using five metrics, defined in the Methods section: trustworthiness, continuity, Spearman distance correlation, density rank correlation, and stress. The results, summarized in Figure 6, reveal consistent trends across systems, while also highlighting important differences depending on the degree of structural flexibility and sampling complexity.

Ethylene represents the most challenging case, as the seam is characterized by large-amplitude distortions and significant connectivity changes. In this regime, methods that emphasize only geometric structure or only density fail to provide a faithful representation. PCA yields low stress but does not capture the nonlinear structure of the seam, while diffusion maps preserve density but distort global geometry. UMAP and t-SNE show poor density preservation and substantial distortion. Among all methods, densMAP provides the most balanced performance, simultaneously maintaining a high level of global structure preservation and capturing the distribution of sampled configurations. Isomap also performs well in terms of global geometry, but lacks comparable density fidelity.

For butadiene, the behavior depends on the seam considered. In the S_0/S_1 case, global geometric structure is relatively well preserved by several methods, with PCA performing particularly strongly. However, density preservation remains a distinguishing factor: densMAP clearly outperforms all other methods in capturing the distribution of sampled configurations, whereas UMAP exhibits severe degradation of density information. In the S_1/S_2 seam, which is more structured and less heterogeneous, most methods recover the global geometry with reasonable accuracy. Nevertheless, densMAP again provides the most faithful representation of density, with diffusion maps also performing well in this respect, albeit with weaker global structure preservation.

Benzene represents the most constrained system, where the seam is dominated by relatively small distortions of a rigid framework. In this case, global geometric relationships are readily captured, and PCA consistently provides an excellent description with minimal distortion. However, preserving the distribution of sampled configurations remains nontrivial. Across both seams, densMAP is the only method that reliably captures density variations, while other approaches yield significantly weaker agreement. This indicates that even in systems where geometry is simple, density preservation is not guaranteed.

These results highlight a clear separation between the preservation of geometric structure and the preservation of sampling density. Methods such as PCA and Isomap excel at maintaining global relationships between configurations but do not capture variations in sampling density. Diffusion maps, in contrast, recover aspects of density but may distort global geometry. UMAP and t-SNE, while effective at identifying local clusters, can significantly distort both global structure and density, particularly in more flexible systems. Across all datasets, densMAP is the only method that consistently provides a balanced representation, simultaneously preserving to a good degree local neighborhoods, global geometry, and the distribution of sampled configurations. This is particularly important for the identification of dynamically accessible seam regions, where both geometric proximity and sampling density play a central role. Thus, we employ densMAP for the following seam basin analysis.

View Article Online
DOI: 10.1039/D6FD00060F

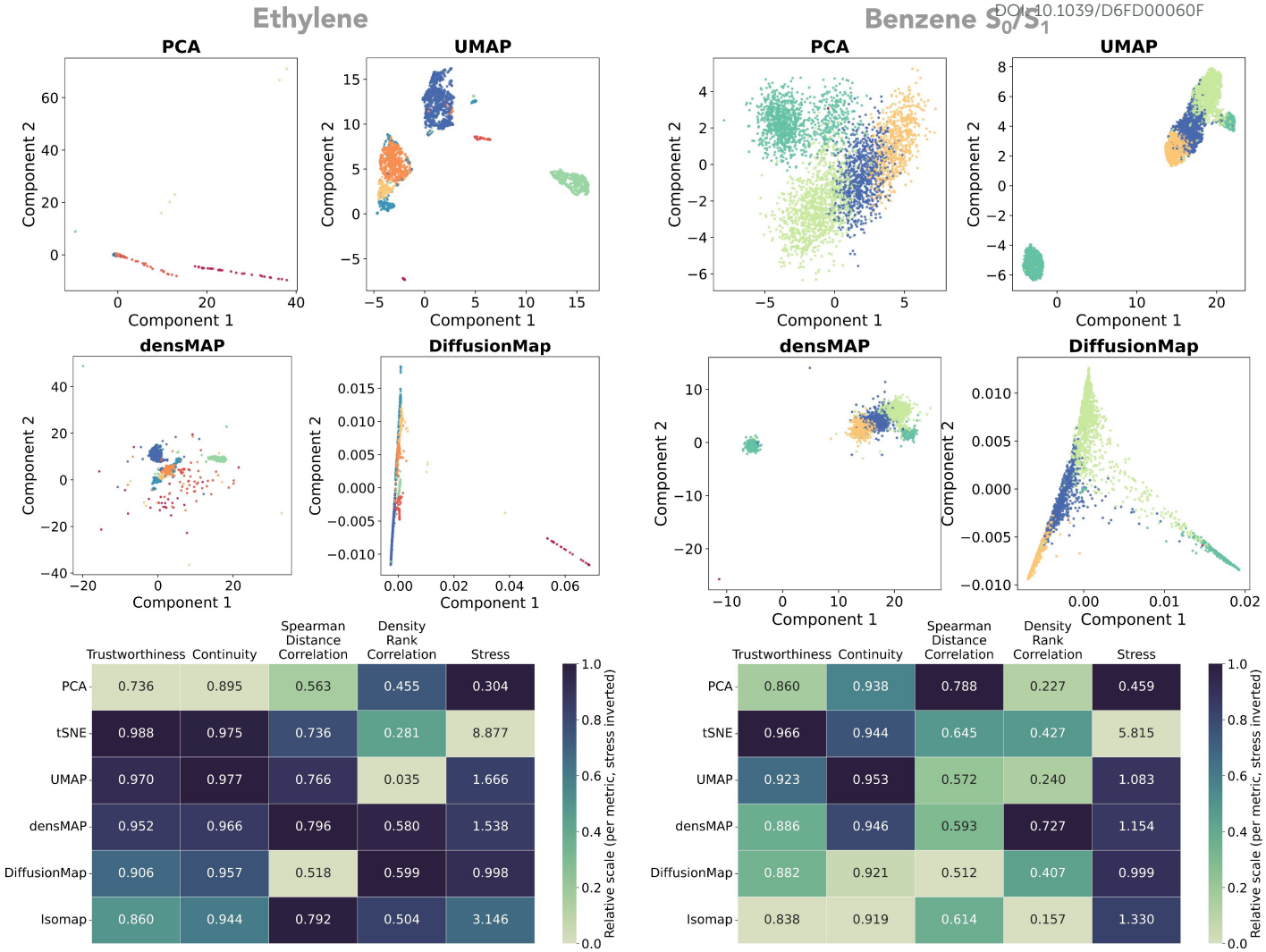


Figure 6 Comparison of dimensionality reduction methods for Cartesian representations of ethylene (left) and benzene S_0/S_1 (right). Top panels show two-dimensional embeddings obtained using PCA, UMAP, densMAP, and diffusion maps, with points colored according to the closest MECI in terms of RMSD. Bottom panels report quantitative metrics evaluating the preservation of local structure (trustworthiness, continuity), global geometry (Spearman distance correlation, stress), and sampling density (density rank correlation). Heatmaps are shown on a normalized scale from 0 (worst) to 1 (best) for all metrics except stress, which is inverted prior to normalization so that lower distortion corresponds to higher scores. Consequently, darker colors indicate better performance, while lighter colors indicate poorer preservation.

Relationship between MECIs and dynamically populated seam basins

We finally analyze the relationship between MECIs and dynamically sampled regions of the seam by comparing their positions in the embedding with the density distribution of SP geometries. The results, summarized in Figure 7A, reveal a clear system-dependent behavior, as well as a consistent trend linking dynamical relevance to proximity to density maxima.

Ethylene exhibits a clear separation between dynamically relevant and peripheral MECIs. The MECIs associated with the largest populations - defined here as the number of SP geometries for which a given MECI yields the lowest RMSD upon alignment - are consistently located very close to the maxima of the density landscape, with small values of d_m. This indicates that the dominant decay pathways correspond to regions of the seam that are both frequently sampled and roughly centered around these structures. Several additional MECIs are also found in close proximity to one another, forming extended high-density regions that can be interpreted as a single basin rather than distinct, well-separated ones. In contrast, less populated MECIs, including dissociative channels, tend to lie further from density maxima, suggesting that they are accessed less frequently or require specific dynamical conditions. Overall, the seam is organized into a small number of highly populated basins that strongly correlate with the most dynamically relevant MECIs.

For butadiene, the relationship between MECIs and basin centers depends on the seam considered. In the S_0/S_1 case, a large number of distinct MECIs is identified, indicative of a highly rugged and potentially flat IS, as previously mentioned. These structures are broadly distributed in the embedding, and many lie far from density maxima, as reflected by the wide spread of d_m values (Figure 7B). However, a clear trend persists: MECIs with higher-than-average populations are preferentially located near density maxima, whereas less populated MECIs are more broadly distributed and often peripheral. In the S_1/S_2 seam, the structure is more organized, with a few dominant basins concentrating most of the population. The MECIs associated with these basins lie close to density maxima, while several other MECIs

View Article Online
DOI: 10.1039/D6FD00060F

exhibit very large values of d_m, indicating that they are dynamically disconnected from the main flow of the trajectory ensemble. As in ethylene, closely related MECI families may cluster within the same high-density region, effectively forming a single basin rather than distinct, well-separated ones.

In benzene, the correspondence between MECIs and basin centers is much tighter than in the more flexible systems, leading to a comparatively clean assignment between dynamically sampled regions and representative structures. For the S_0/S_1 seam, most MECIs are located close to density maxima, and the distribution of d_m values is narrow, indicating that the dynamically accessible regions of the seam are well captured by a small number of representative structures. A similar trend is observed for the S_1/S_2 seam, although with slightly larger variability. An important subtlety, however, is that geometrically distinct regions of the embedding may still converge to the same MECI upon optimization. In these cases, SP geometries that are far apart in the embedding space are mapped to a single MECI, resulting in density maxima that do not have a corresponding MECI located nearby. This highlights that the mapping between embedding space and optimized MECI structures is not strictly one-to-one, even in relatively rigid systems. Nevertheless, in both seams, the most populated basins correspond to MECIs located near density maxima, while isolated exceptions indicate that energetic accessibility alone does not guarantee dynamical relevance.

View Article Online
DOI: 10.1039/D6FD00060F

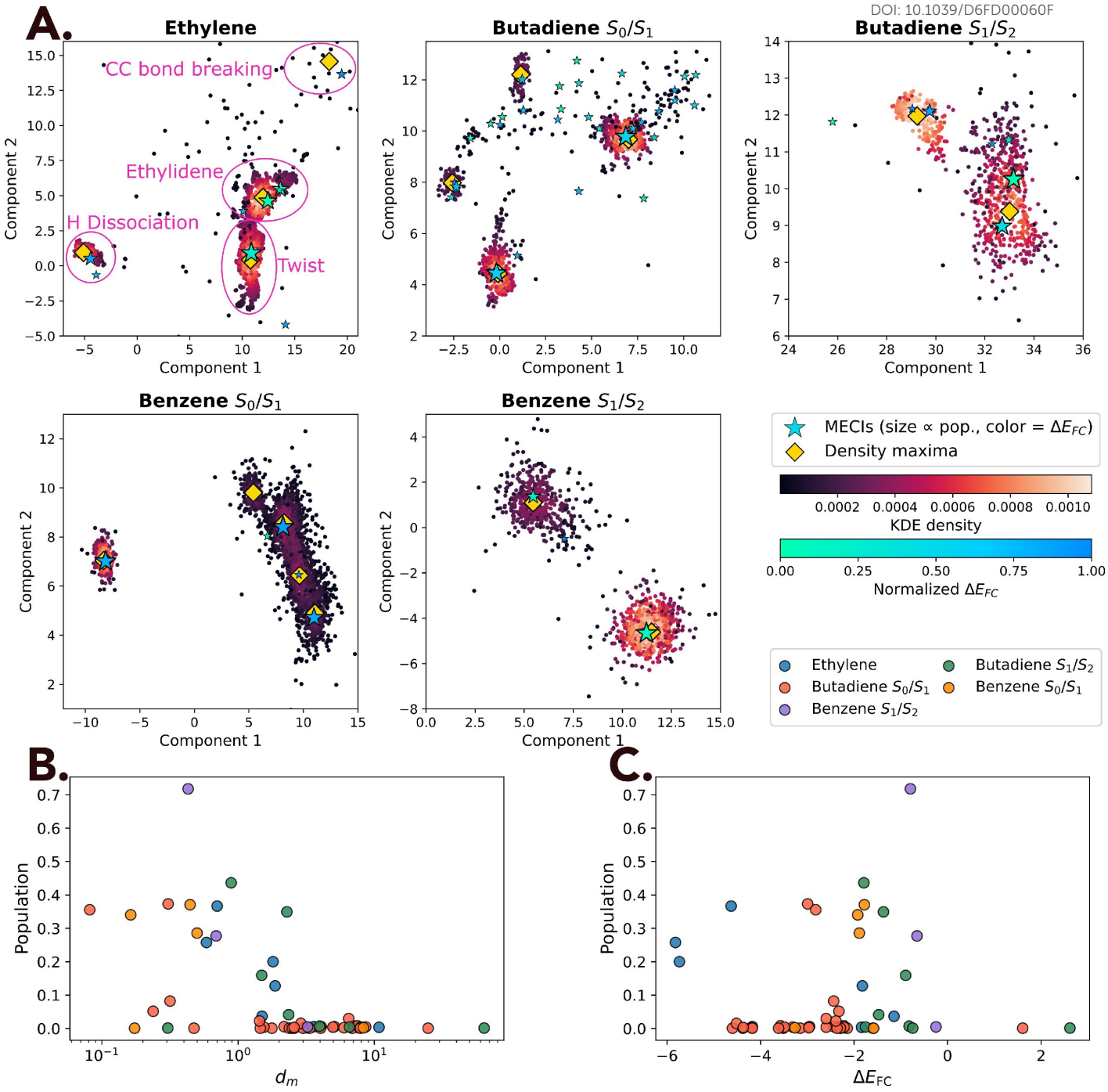


Figure 7 Seam basin analysis across all datasets. (A) Two-dimensional densMAP embeddings (components 1 and 2) of SP geometries, colored according to the local density estimated via KDE. Density maxima are indicated by gold diamonds. MECIs are shown as stars, with marker size proportional to their population (defined as the number of SP geometries for which the MECI yields the lowest RMSD upon alignment) and color dependent on their normalized energy gap with respect to the Franck Condon point (i.e., lowest energy MECI in green and highest energy MECI in blue). Ethylene's plot has been annotated to indicate the character of the main seam basins. For clarity, the plots are cropped to focus on the most densely sampled regions of the embedding; sparsely populated outliers are not shown. (B) Relationship between MECI population and distance d_m from the nearest density maximum (logarithmic scale). (C) Relationship between MECI population and energy gap with respect to the Franck–Condon point (in eV).

Across all systems, MECIs do not uniformly coincide with the centers of dynamically sampled seam regions, indicating that the identification of MECIs alone does not directly reveal the regions of the seam that are most frequently accessed by the dynamics. Flexible systems such as ethylene and butadiene exhibit a broad distribution of distances d_m (Figure 7B), with many MECIs located far from density maxima, reflecting the presence of multiple competing pathways and a fragmented seam topology. In contrast, more rigid systems such as benzene show a tighter correspondence between MECIs and basin centers.

A consistent trend nevertheless emerges: MECIs associated with the most highly populated basins are systematically located near density

View Article Online
DOI: 10.1039/D6FD00060F

maxima, whereas MECIs with low associated populations are more broadly distributed and often lie at the periphery of the sampled landscape. This is evident from the correlation between MECI population and distance d_m (Figure 7B), where highly populated MECIs cluster at small d_m, while low-population MECIs span a wide range of distances. This indicates that, while MECIs as a whole do not provide a complete representation of the IS, a subset of MECIs (those corresponding to high-density basins) captures the dominant dynamical pathways. It must be however noted that, while certain MECIs remain representative, the dynamics access a broader seam region centered around them, underscoring the need for approaches that characterize the full ensemble of dynamically accessible geometries rather than isolated stationary points.

Importantly, the results also show that MECI energy alone is not a reliable indicator of dynamical relevance. As shown in Figure 7C, no clear monotonic relationship is observed between MECI population and energy gap with respect to the Franck–Condon point. Although lower-energy MECIs often correspond to frequently accessed regions, notable exceptions are observed. For example, in benzene, the hydrogen-transfer MECI lies among the lowest-energy structures but is not significantly populated in the dynamics. This demonstrates that energetic accessibility does not guarantee dynamical accessibility, and that identifying representative regions of the seam requires accounting not only for the location and energy of MECIs, but also for the reaction pathways that connect them to the evolving molecular geometry.

Agreement between basin, alignment, and optimization assignments To quantify the relationship between density-defined basins, geometric proximity to MECIs, and dynamical outcomes, we evaluate the agreement between the corresponding partitions using the ARI and NMI. Specifically, we compare (i) basin assignments with alignment-based MECI labels, and (ii) alignment-based labels with the MECIs reached upon optimization. The resulting values are reported in Table 2, and representative mappings are illustrated in the Sankey diagrams of Figure 8 for representative systems (see SI for comprehensive results).

Table 2 Agreement between density-based basin assignments, alignment-based MECI labels, and optimization outcomes, quantified using the adjusted Rand index (ARI) and normalized mutual information (NMI).

Dataset	**Basin ↔ Align.**		**Align. ↔ Opt.**	
	ARI	NMI	ARI	NMI
Ethylene	0.680	0.696	0.499	0.522
Butadiene S_0/S_1	0.744	0.665	0.846	0.743
Butadiene S_1/S_2	0.395	0.472	0.225	0.296
Benzene S_0/S_1	0.703	0.708	0.584	0.618
Benzene S_1/S_2	0.847	0.717	0.842	0.693

View Article Online
DOI: 10.1039/D6FD00060F

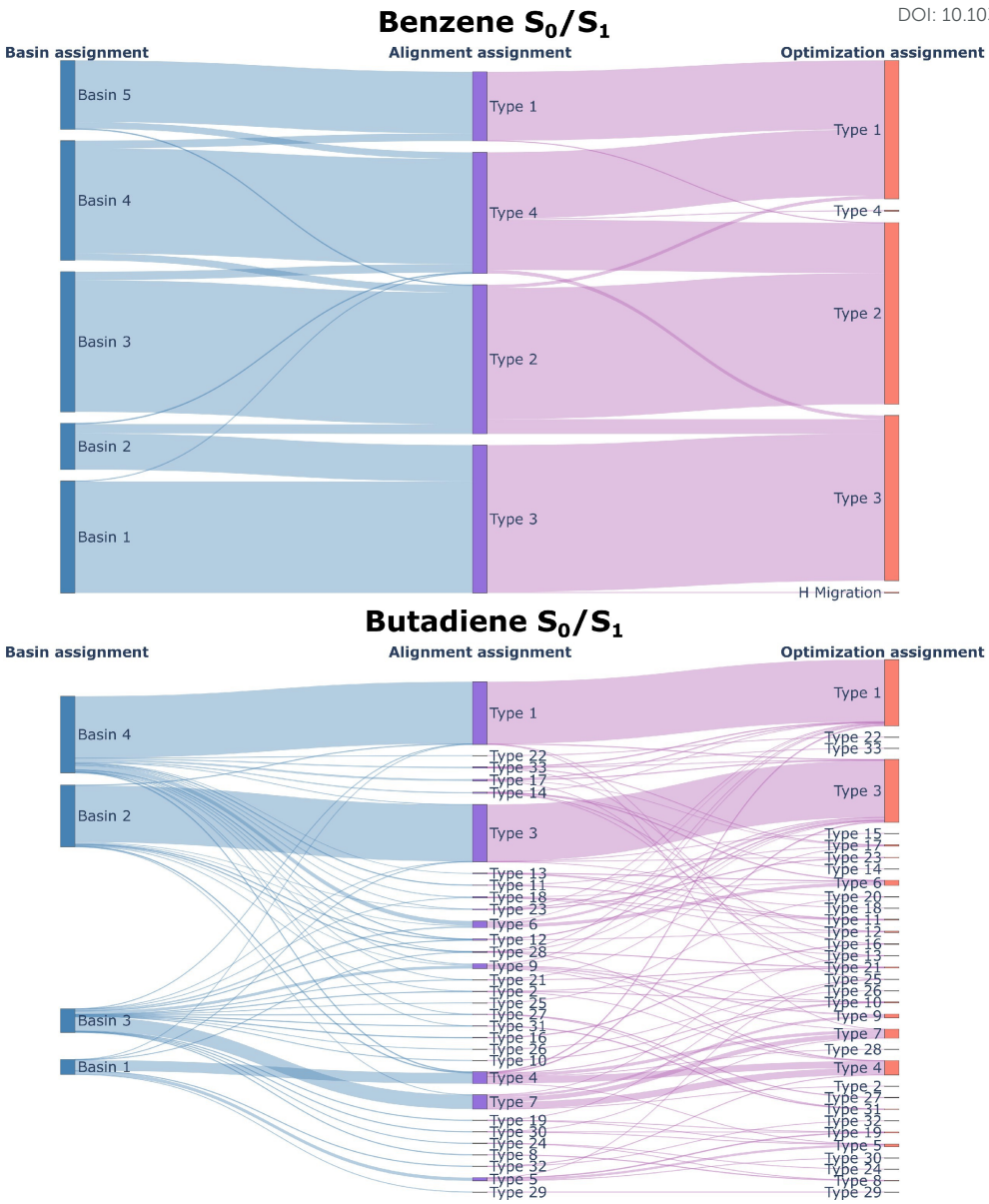


Figure 8 Three-way mapping between density-based basin assignments, alignment-based MECI labels, and optimization outcomes for benzene S_0/S_1 and butadiene S_0/S_1. Flows connect density basins (left) to the MECI to which each SP geometry is closest in RMSD (center), and to the MECI reached upon optimization (right), with widths proportional to the number of SP geometries following each path.

The agreement between basin and alignment assignments is generally moderate to high, with ARI values ranging from 0.395 to 0.847. For benzene, particularly in the S_1/S_2 seam (ARI = 0.847, NMI = 0.717), the high agreement indicates that density basins correspond closely to individual MECIs, consistent with the visually clean one-to-one mappings observed in Figure 7 and in the Supporting Information. A similarly high agreement is observed for butadiene S_0/S_1 (ARI = 0.744), suggesting that, despite the large number of MECIs, the dominant basins are still largely associated with specific structures (see Figure 8). In contrast, butadiene S_1/S_2 exhibits significantly lower agreement (ARI = 0.395, NMI = 0.472), reflecting the merging of multiple MECI families into common dynamically sampled regions, as also evident from the many-to-one flows in the corresponding Sankey diagram. Ethylene and benzene S_0/S_1 show intermediate behavior, indicating partial but not complete correspondence between basins and geometrically closest MECIs (Figure 8).

The agreement between alignment and optimization assignments is more variable across systems, revealing how well geometric proximity predicts the outcome of MECI optimization, which is likely highly sensitive to the specific optimization algorithm and thresholds used. For benzene S_1/S_2 (ARI = 0.842) and butadiene S_0/S_1 (ARI = 0.846), the high agreement indicates that the closest MECI is generally also the one reached upon optimization, consistent with a relatively well-defined local structure of the seam. In contrast, ethylene (ARI = 0.499) and benzene S_0/S_1 (ARI = 0.584) show only moderate agreement, suggesting that dynamical pathways can redirect trajectories toward different MECIs than those identified by geometric proximity alone. In particular, we note that, while many SPs are aligned to MECI Type 4, subsequent optimizations predominantly reassign these geometries to either Type 1 or Type 2. This behavior is consistent with previous seam-mapping studies of benzene,[23] in which Type 4 has been proposed to correspond to an intermediate region of the seam connecting Types 1 and 2. The lowest agreement is observed for butadiene S_1/S_2 (ARI = 0.225, NMI = 0.296), indicating a substantial

View Article Online
DOI: 10.1039/D6FD00060F

mismatch between geometric assignment and optimization outcome, and highlighting the importance of pathway-dependent effects in this system.

Overall, density-defined basins and MECI-based partitions are not strictly equivalent: while they often correlate, particularly in rigid systems, multiple MECIs may contribute to the same dynamically sampled region, or a single MECI may span multiple basins. Second, geometric proximity to a MECI does not universally predict the outcome of optimization, especially in systems with more complex or competing pathways. These observations reinforce the conclusion that MECIs alone do not provide a complete representation of the IS. Instead, both the structure of the density landscape and the connectivity of reaction pathways must be considered to identify the dynamically relevant regions of the seam.

Conclusions

In this work, we investigated the extent to which MECIs provide a representative description of the IS sampled during nonadiabatic dynamics. To this end, we combined large ensembles of SPs with a systematic evaluation of molecular representations and dimensionality reduction techniques, followed by a density-based analysis of the seam structure.

A key methodological result of this study is the identification of an effective protocol for representing and analyzing seam geometries. We find that, once rigid-body alignment and permutation symmetry are properly accounted for, Cartesian coordinates provide the most faithful representation of molecular structure with respect to RMSD-based similarity. Among dimensionality reduction techniques, densMAP consistently offers the best balance between preserving global geometric relationships and capturing variations in sampling density across all systems considered.

Using this framework, we show that the IS is not uniformly represented by discrete MECIs, but instead exhibits a structured organization into dynamically accessible basins. In flexible systems such as ethylene and butadiene, the seam is fragmented into multiple regions, with many MECIs located far from density maxima and therefore not significantly sampled during the dynamics. In contrast, more rigid systems such as benzene display a tighter correspondence between MECIs and dynamically populated regions, although deviations remain.

A consistent trend emerges across all systems: MECIs associated with highly populated basins are located near density maxima and capture the dominant dynamical pathways, whereas many other MECIs correspond to peripheral or weakly sampled regions of the seam. This indicates that while certain MECIs can identify important regions of the seam, they do not, in general, provide a complete or unbiased representation of the dynamically relevant landscape.

Furthermore, our analysis demonstrates that neither geometric proximity nor energetic accessibility alone is sufficient to determine dynamical relevance. In particular, although lower-energy MECIs are frequently associated with highly populated regions, notable exceptions are found. Additionally, while certain MECIs act as centers of dynamically accessed basins, the associated distributions are broad, indicating that excess energy enables extensive exploration of the seam, akin to the broadening of configurational sampling at elevated temperature on a ground-state potential energy surface.

Taken together, these results demonstrate that MECIs should be viewed not as a complete description of the seam, but as discrete markers embedded within a broader, continuous, and dynamically weighted landscape. A comprehensive understanding of photochemical relaxation pathways therefore requires not only the identification of MECIs, but also an explicit characterization of the dynamically sampled regions of the seam and the pathways connecting them.

Data availability

The data supporting this article have been included as part of the Supplementary Information. Supplementary information:

- Detailed procedure for the preprocessing of Cartesian coordinates
- MECI information for butadiene S_0/S_1
- Comprehensive nonlinear dimensionality results and comparison
- Detailed basin analysis
- Sankey diagrams for assignment correspondence of additional datasets

Additional data for this article, including complete nonadiabatic dynamics datasets and MECI and SP information, are available at Zenodo at DOI: 10.5281/zenodo.19462190.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award Number DE-SC0026291. The authors would like to thank the University of North Carolina at Chapel Hill and the Research Computing group for providing computational resources and support that have contributed to these research results. The authors also thank Dr. Alessio Valentini and Dr. Roman Ellerbrock for valuable discussions and suggestions.

Author Contributions

Conceptualization: Elisa Pieri, Conner Baucom
Methodology: All authors
Investigation: Conner J. Baucom
Formal analysis: Conner J. Baucom, Eleftherios Mainas

[0a] *Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. E-mail: elipieri@unc.edu*

View Article Online
DOI: 10.1039/D6FD00060F

Data curation: Conner J. Baucom
Writing - original draft: Elisa Pieri, Conner J. Baucom
Writing - review & editing: All authors
Visualization: Elisa Pieri, Conner J. Baucom
Funding acquisition: Elisa Pieri
Project admninistration: Elisa Pieri
Resources: Elisa Pieri
Software: Conner Baucom
Supervision: Elisa Pieri, Eleftherios Mainas
Validation: All authors

Conflicts of interest

There are no conflicts to declare.

References

[1] W. Domcke, H. Koppel and D. R. Yarkony, *Conical intersections: electronic structure, dynamics & spectroscopy*, World Scientific, 2004, vol. 15.

[2] D. R. Yarkony, *The Journal of Chemical Physics*, 2001, **114**, 2601–2613.

[3] W. Domcke and D. R. Yarkony, *Annual Review of Physical Chemistry*, 2012, **63**, 325–352.

[4] D. R. Yarkony, *Reviews of Modern Physics*, 1996, **68**, 985–1013.

[5] D. R. Yarkony, *Accounts of Chemical Research*, 1998, **31**, 511–518.

[6] D. R. Yarkony, *Journal of Physical Chemistry A*, 2001, **105**, 6277–6293.

[7] D. R. Yarkony, *The Journal of Chemical Physics*, 2005, **123**, 204101.

[8] T. W. Keal, A. Koslowski and W. Thiel, *Theoretical Chemistry Accounts*, 2007, **118**, 837–844.

[9] S. Maeda, Y. Harabuchi, T. Taketsugu and K. Morokuma, *The Journal of Physical Chemistry A*, 2014, **118**, 12050–12058.

[10] G. J. Atchity, S. S. Xantheas and K. Ruedenberg, *The Journal of Chemical Physics*, 1991, **95**, 1862–1876.

[11] M. J. Bearpark and M. A. Robb, in *Conical Intersection Species as Reactive Intermediates*, 2007, pp. 379–414.

[12] T. J. Martínez, *Accounts of Chemical Research*, 2006, **39**, 119–126.

[13] M. A. Robb, M. Garavelli, M. Olivucci and F. Bernardi, in *A Computational Strategy for Organic Photochemistry*, 2000, pp. 87–146.

[14] B. G. Levine and T. J. Martínez, *Annual Review of Physical Chemistry*, 2007, **58**, 613–634.

[15] S. Gozem, H. L. Luk, I. Schapiro and M. Olivucci, *Chemical Reviews*, 2017, **117**, 13502–13565.

[16] H. Köuppel, W. Domcke and L. S. Cederbaum, in *Multimode Molecular Dynamics Beyond the Born-Oppenheimer Approximation*, John Wiley & Sons, Ltd, 1984, pp. 59–246.

[17] Y. Boeije and M. Olivucci, *Chemical Society Reviews*, 2023, **52**, 2643–2687.

[18] A. Migani, M. A. Robb and M. Olivucci, *Journal of the American Chemical Society*, 2003, **125**, 2804–2808.

[19] B. F. E. Curchod and T. J. Martínez, *Chemical Reviews*, 2018, **118**, 3305–3336.

[20] M. Ben-Nun and T. J. Martínez, *Chemical Physics*, 2000, **259**, 237–248.

[21] S. Matsika, *Chemical Reviews*, 2021, **121**, 9407–9449.

[22] E. Pieri, A. R. Walker, M. Zhu and T. J. Martínez, *Journal of the American Chemical Society*, 2024, **146**, 17646–17658.

[23] E. Pieri, D. Lahana, A. M. Chang, C. R. Aldaz, K. C. Thompson and T. J. Martínez, *Chemical Science*, 2021, **12**, 7294–7307.

[24] M. Ben-Nun, M. F., S. K. and M. T. J., *Proceedings of the National Academy of Sciences*, 2002, **99**, 1769–1773.

[25] K. Tong, E. Mainas and E. Pieri, *The Journal of Chemical Physics*, 2026, **164**, 084120.

[26] F. Agostini and B. F. E. Curchod, *WIREs Computational Molecular Science*, 2019, **9**, e1417.

View Article Online
DOI: 10.1039/D6FD00060F

[27] A. Prlj, J. T. Taylor, J. Janoš, E. Lognon, D. Hollas, P. Slavíček, F. Agostini and B. F. E. Curchod, *Living Journal of Computational Molecular Science*, 2026, **7**, 4157.

[28] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang and G.-W. Wei, *Chemical Reviews*, 2023, **123**, 8736–8780.

[29] S. K. Achar and J. A. Keith, *Chemical Reviews*, 2024, **124**, 13571–13573.

[30] C. Ashraf, N. Joshi, D. A. Beck and J. Pfaendtner, *Annual Review of Chemical and Biomolecular Engineering*, 2021, **12**, 15–37.

[31] R. Gómez-Bombarelli and A. Aspuru-Guzik, *Handbook of Materials Modeling: Methods: Theory and Modeling*, Springer, 2020, pp. 1939–1962.

[32] J. M. Crawford, C. Kingston, F. D. Toste and M. S. Sigman, *Accounts of Chemical Research*, 2021, **54**, 3136–3148.

[33] F. Noé, A. Tkatchenko, K.-R. Müller and C. Clementi, *Annual Review of Physical Chemistry*, 2020, **71**, 361–390.

[34] H. Yin, *International Journal of Automation and Computing*, 2007, **4**, 294–303.

[35] L. Van Der Maaten, E. O. Postma, H. J. Van Den Herik *et al.*, *Journal of Machine Learning Research*, 2009, **10**, 1–41.

[36] A. Gisbrecht and B. Hammer, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2015, **5**, 51–73.

[37] P. Bhattacharjee and P. Mitra, *Frontiers of Computer Science*, 2021, **15**, 151308.

[38] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding and C.-T. Lin, *Neurocomputing*, 2017, **267**, 664–681.

[39] D. S. Wigh, J. M. Goodman and A. A. Lapkin, *WIREs Computational Molecular Science*, 2022, **12**, e1603.

[40] S. Raghunathan and U. D. Priyakumar, *International Journal of Quantum Chemistry*, 2022, **122**, e26870.

[41] G. M. Jones, B. Story, V. Maroulas and K. D. Vogiatzis, *Molecular Representations for Machine Learning*, American Chemical Society, Washington, DC, USA, 2023.

[42] M. Ben-Nun, J. Quenneville and T. J. Martínez, *The Journal of Physical Chemistry A*, 2000, **104**, 5161–5175.

[43] G. Granucci and A. Toniolo, *Chemical Physics Letters*, 2000, **325**, 79–85.

[44] P. Slavíček and T. J. Martínez, *The Journal of Chemical Physics*, 2010, **132**, 234102.

[45] G. M. Curtin, M. L. Thomas and E. Pieri, *Journal of Chemical Theory and Computation*, 2025.

[46] I. S. Ufimtsev and T. J. Martínez, *Journal of Chemical Theory and Computation*, 2008, **4**, 222–231.

[47] I. S. Ufimtsev and T. J. Martínez, *Journal of Chemical Theory and Computation*, 2009, **5**, 1004–1015.

[48] I. S. Ufimtsev and T. J. Martínez, *Journal of Chemical Theory and Computation*, 2009, **5**, 2619–2628.

[49] I. T. Jolliffe and J. Cadima, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016, **374**, 20150202.

[50] L. van der Maaten and G. E. Hinton, *Journal of Machine Learning Research*, 2008, **9**, 2579–2605.

[51] L. McInnes, J. Healy and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, 2020, `https://arxiv.org/abs/1802.03426`.

[52] A. Narayan, B. Berger and H. Cho, *Nature Biotechnology*, 2021, **39**, 765–774.

[53] J. Tenenbaum, Advances in Neural Information Processing Systems, 1997.

[54] R. R. Coifman and S. Lafon, *Applied and Computational Harmonic Analysis*, 2006, **21**, 5–30.

[55] A. P. Bartók, R. Kondor and G. Csányi, *Physical Review B*, 2013, **87**, 184115.

[56] H. Huo and M. Rupp, *Machine Learning : Science and Technology*, 2022, **3**, 045017.

[57] M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

[58] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Chemical Reviews*, 2021, **121**, 9759–9815.

[59] W. Kabsch, *Acta Crystallographica Section A*, 1976, **32**, 922–923.

View Article Online
DOI: 10.1039/D6FD00060F

[60] W. Kabsch, *Acta Crystallographica Section A*, 1978, **34**, 827–828.

[61] B. K. P. Horn, *J. Opt. Soc. Am. A*, 1987, **4**, 629–642.

[62] E. A. Coutsias, C. Seok and K. A. Dill, *Journal of Computational Chemistry*, 2004, **25**, 1849–1857.

[63] T. Cover and P. Hart, *IEEE Transactions on Information Theory*, 1967, **13**, 21–27.

[64] C. Spearman, *The American Journal of Psychology*, 1987, **100**, 441–471.

[65] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen and E. Castrén, *BMC Bioinformatics*, 2003, **4**, 48.

[66] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*, Springer, 2005.

[67] B. W. Silverman, *Density estimation for statistics and data analysis*, Routledge, 2018.

[68] A. Rodriguez and A. Laio, *Science*, 2014, **344**, 1492–1496.

[69] M. J. Warrens and H. van der Hoef, *Journal of Classification*, 2022, **39**, 487–509.

[70] A. Strehl and J. Ghosh, *J. Mach. Learn. Res.*, 2003, **3**, 583–617.

[71] B. G. Levine and M. T. J., *The Journal of Physical Chemistry A*, 2009, **113**, 12815—12824.

View Article Online
DOI: 10.1039/D6FD00060F

Data availability

All data supporting the findings of this study are available within the article and its Supplementary Information, or from the Zenodo repository.

The Supplementary Information includes:

- Detailed procedure for the preprocessing of Cartesian coordinates
- MECI information for butadiene S_0/S_1
- Comprehensive nonlinear dimensionality reduction results and comparisons
- Detailed basin analysis
- Sankey diagrams for assignment correspondence of additional datasets

Additional datasets, including complete nonadiabatic dynamics trajectories as well as MECI and spawning point (SP) information, are openly available without restriction at:

`https://doi.org/10.5281/zenodo.19462190`, under a Creative Commons Attribution license (CC BY 4.0), which permits reuse with appropriate credit.