Information-Geometric Tracking of the Vibrational-Spatial Manifold Across Murcko Scaffolds

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how molecular architecture dictates the joint physical states of molecules requires isolating the effects of chemical topology from simple stoichiometric determinism. To address this, we track the deformation of a multidimensional vibrational-spatial property manifold across discrete structural families within the QM9 dataset. We isolate topological effects by extracting non-linear gradient-boosted residuals for electronic spatial extent, zero-point vibrational energy, heat capacity, and the rotational constant, strictly conditioned on atom counts. By grouping the dataset according to Murcko scaffolds, we employ Optimal Transport to compute the exact Wasserstein-2 distances between the 4D residual distributions of cyclic scaffold families and a size-matched acyclic baseline. Our analysis reveals that topological complexity, particularly ring count, monotonically drives manifold deformation, forcing polycyclic systems into joint physical states that are inaccessible to acyclic chains. Furthermore, high-deformation scaffolds—such as strained bicyclic and tricyclic systems—exhibit fundamental reorganizations in their internal property couplings, effectively altering the relationship between spatial extent and vibrational energy profiles compared to flexible acyclic structures. Validated by a sensitivity analysis that excludes putative geometric outliers, this information-geometric approach demonstrates that molecular scaffolds act as topological operators that systematically rewire the coupled thermodynamic and spatial landscapes of small organic molecules.

1 Introduction

The mapping from molecular architecture to physical properties is a central theme in chemistry and materials science. However, understanding how the specific connectivity of atoms—the chemical topology—dictates a molecule's physical state requires disentangling these structural effects from simple stoichiometric determinism. As molecules grow in size, their thermodynamic and spatial properties naturally scale, often masking the subtle, multidimensional ways in which structural constraints, such as rings and fused systems, reorganize the molecule's internal physical landscape.

To rigorously isolate the role of topology, it is necessary to move beyond analyzing single properties in isolation and instead examine the joint physical state of the molecule. Properties such as electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), heat capacity (c_v), and the rotational constant (C) are inherently coupled. The spatial distribution of electron density influences rotational dynamics, while the stiffness and connectivity of the molecular framework dictate vibrational modes and thermodynamic behavior. When a molecule is constrained into a specific architecture, this joint property manifold must deform, forcing the molecule into physical states that may be fundamentally inaccessible to flexible, acyclic chains.

In this work, we present an information-geometric approach to track the deformation of this vibrational-spatial manifold across discrete structural families within the QM9 dataset. To ensure that our analysis captures true topological rewiring rather than compositional scaling, we strictly condition the target properties on exact atom counts. By extracting non-linear gradient-boosted residuals for r^2, $zpve$, c_v, and C, we isolate the variance driven exclusively by the spatial arrangement of atoms. We then partition the dataset according to Murcko scaffolds, establishing a size-matched acyclic population as a structural baseline against which cyclic and polycyclic architectures can be compared.

To quantify the deformation of the property manifold, we employ Optimal Transport to compute the exact Wasserstein-2 (W_2) distances between the 4D residual distributions of cyclic scaffold families and the acyclic baseline. This metric captures not only shifts in the mean and variance of the properties but also fundamental reorganizations in their internal correlation structure. Furthermore, because the QM9 dataset contains a known subset of latent, DFT-inconsistent molecules that could manifest as extreme spatial or energetic outliers, we couple our Optimal Transport framework with a rigorous sensitivity analysis based on Mahalanobis distance filtering to ensure the robustness of the observed topological shifts.

Our analysis demonstrates that topological complexity, particularly ring count, monotonically drives the deformation of the vibrational-spatial manifold. High-deformation scaffolds, such as strained bicyclic and tricyclic systems, exhibit profound rewiring of their internal property couplings, effectively altering the relationship between spatial extent and vibrational energy profiles compared to acyclic structures. Ultimately, this approach reveals that molecular scaffolds act as topological operators that systematically reshape the coupled thermodynamic and spatial landscapes of small organic molecules.

2 Methods

2.1 Data curation and structural parsing

The QM9 dataset, comprising 133,885 small organic molecules, was utilized as the foundational corpus for this study. To prevent data leakage and inflated density estimates during manifold construction, the dataset was first deduplicated. Specifically, 83 SMILES strings that appeared multiple times due to collapsed stereochemistry were resolved by retaining only their first occurrence. During structural parsing, the RDKit logger was globally silenced to maintain execution environment stability.

Each deduplicated SMILES string was parsed into a molecular object, to which explicit hydrogens were added. Exact stoichiometric features were computed for each molecule, specifically the counts of Carbon, Nitrogen, Oxygen, Fluorine, and Hydrogen atoms, alongside the total heavy atom count. To categorize the dataset by chemical topology, the Murcko scaffold of each molecule was extracted and converted into a canonical SMILES string. This scaffold SMILES served as the discrete grouping variable for all subsequent topological comparisons. Molecules yielding an empty string for their Murcko scaffold (including branched chains) were designated as the acyclic structural baseline.

2.2 Non-linear stoichiometric conditioning

To rigorously isolate the effects of chemical topology from stoichiometric determinism, the four target properties—electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), heat capacity (c_v), and the rotational constant (C)—were conditioned strictly on molecular composition. For each property, a gradient-boosted regressor was trained using only the five atom counts as input features.

To obtain unbiased estimates of the stoichiometrically determined property values, out-of-fold predictions were generated using 5-fold cross-validation. The non-linear residuals were then extracted by subtracting these predictions from the raw property values:

$$res_P = P_{raw} - P_{pred} \tag{1}$$

where $P \in \{r^2, zpve, c_v, C\}$. These residuals represent the variance in the physical state driven exclusively by the spatial arrangement and connectivity of the atoms.

2.3 Manifold definition and outlier detection

For each molecule, the four extracted residuals were assembled into a 4D vector, defining the vibrational-spatial property manifold. To ensure that subsequent geometric distance computations were not disproportionately dominated by the feature with the largest absolute variance, a global robust scaling was applied across the entire dataset. Each residual dimension was scaled using its interquartile range (IQR), preserving the relative physical distributions and internal correlation structures while standardizing the scale.

The QM9 dataset is known to contain approximately 3,054 latent molecules with DFT-inconsistent geometries, which could manifest as extreme energetic or spatial outliers. To systematically identify these

without relying on external labels, the Mahalanobis distance of the scaled 4D residual vectors was computed for all molecules. This distance metric accounts for the covariance structure of the manifold. The 3,054 molecules exhibiting the largest Mahalanobis distances were flagged as putative geometric outliers for use in downstream sensitivity analyses.

2.4 Scaffold population filtering and baseline matching

The dataset was partitioned into discrete populations based on the extracted Murcko scaffolds. To ensure statistical stability in the Optimal Transport computations, the cyclic scaffold families were filtered to retain only those containing a minimum of 50 molecules. Structural descriptors, including heavy atom count and the number of rings, were recorded for each retained scaffold.

Because manifold deformation could potentially be influenced by residual size effects not fully captured by stoichiometric conditioning, a "size-matched" acyclic baseline was dynamically constructed for each cyclic scaffold. This was achieved by sampling molecules from the global acyclic population that exactly matched the heavy atom count of the target cyclic family, ensuring that any observed manifold deformation was strictly attributable to the introduction of cyclic topology.

2.5 Information-geometric tracking via Optimal Transport

The deformation of the vibrational-spatial manifold across structural families was quantified using the exact Wasserstein-2 (W_2) distance. Computations were performed using the Python Optimal Transport (`POT`) library, parallelized across 14 CPUs. For each retained cyclic scaffold, the W_2 distance was computed between its empirical 4D scaled residual distribution (μ) and the distribution of its size-matched acyclic baseline (ν):

$$W_2(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^4 \times \mathbb{R}^4} \|x - y\|^2 \, d\gamma(x, y) \right)^{1/2} \tag{2}$$

where $\Pi(\mu, \nu)$ denotes the set of all joint probability measures γ with marginals μ and ν. The weights of the empirical distributions were uniformly normalized to sum to 1. To establish a statistical noise floor and verify the significance of the observed topological deformations, baseline W_2 distances were also computed between random, disjoint subsets of the acyclic population.

2.6 Sensitivity analysis and manifold visualization

To validate that the observed manifold deformations were driven by fundamental chemical rewiring rather than artifacts from DFT-inconsistent geometries, a rigorous sensitivity analysis was conducted. The 3,054 Mahalanobis-flagged outlier molecules were excluded from all scaffold populations and acyclic baselines. The exact W_2 distances were then recomputed for all filtered size-matched pairs. The robustness of the topological shifts was evaluated by comparing these filtered distances to the original distances using Pearson correlation and mean absolute error (MAE).

Finally, the structural drivers of manifold deformation were analyzed by correlating the robust W_2 distances with scaffold descriptors such as ring count. To intuitively demonstrate the chemical rewiring of internal property couplings, pairwise distance correlation ($dCor$) matrices were computed for representative high-deformation scaffolds and compared against the acyclic baseline. The global reorganization of the 4D residual space was visualized by projecting the manifold into two dimensions using Uniform Manifold Approximation and Projection (UMAP), overlaying the cyclic scaffold distributions onto the acyclic baseline density.

3 Results

3.1 Stoichiometric conditioning and residual extraction

To ensure that the observed joint structure of the molecular property space reflects chemical topology rather than simple molecular size or composition, all target properties were conditioned on atom counts. The

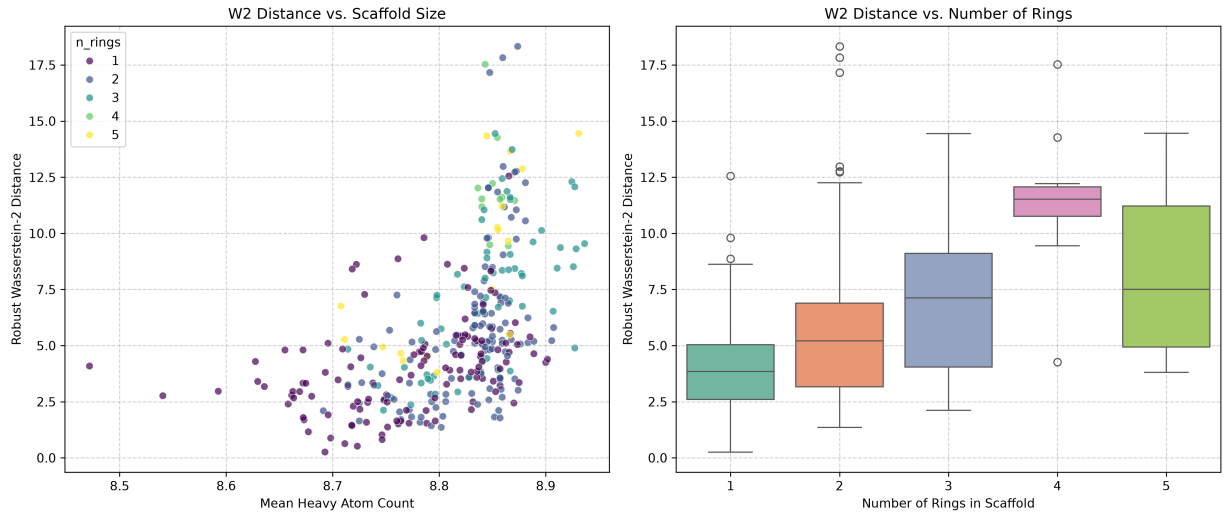

Figure 1: Structural correlation analysis illustrating the impact of topological complexity on the deformation of the vibrational-spatial manifold. The left panel shows a positive correlation ($r = 0.5057$) between the robust Wasserstein-2 (W_2) distance and the mean heavy atom count across Murcko scaffolds, with points colored by ring count. The right panel displays a boxplot demonstrating an increase in manifold deformation as the number of rings increases ($r = 0.4701$), with scaffolds containing 3 or 4 rings frequently exhibiting W_2 distances exceeding 10.0. These trends reveal that the introduction of cyclic constraints and polycyclic fused systems reorganizes the coupled physical states of molecules, shifting them into joint property configurations that differ from size-matched acyclic baselines.

gradient-boosted baseline models achieved high fidelity for thermodynamic properties, yielding R^2 values of 0.9973 for zero-point vibrational energy ($zpve$) and 0.7297 for heat capacity (c_v). In contrast, geometric and spatial properties were less constrained by stoichiometry, with R^2 values of 0.3607 for electronic spatial extent (r^2) and 0.3003 for the rotational constant (C).

The resulting non-linear residuals represent the property values unexplained by the molecular formula. By assembling these into a 4D residual manifold, we isolated the variance driven by the spatial arrangement of atoms, providing a foundation for evaluating topological effects.

3.2 Quantification of manifold deformation via optimal transport

The deformation of the 4D residual manifold was quantified using the Wasserstein-2 (W_2) distance between the distributions of specific Murcko scaffolds and their size-matched acyclic baselines. Across 338 cyclic scaffolds containing at least 50 molecules, the mean robust W_2 distance from the acyclic baseline was 5.6842.

The distribution of these distances indicates that topological complexity is associated with manifold deformation. As illustrated in Figure 1, the W_2 distance correlates positively with both the mean heavy atom count ($r = 0.5057$) and the number of rings in the scaffold ($r = 0.4701$).

The increase in manifold deformation as a function of ring count demonstrates that the introduction of cyclic constraints and polycyclic fused systems shifts molecules into joint physical states that are distinct from their acyclic counterparts, even when the number of heavy atoms is held constant. The analysis identified several scaffolds exhibiting pronounced manifold deformation, with the maximum W_2 reaching 18.327. The three scaffolds with the highest robust W_2 distances were C1CC2CC1CO2 ($W_2 = 18.327$, a strained bicyclic system containing an epoxide), C1CC2CCC1C2 ($W_2 = 17.822$, a norbornane-like bicyclic structure), and C1C2CC3C1N3C2 ($W_2 = 17.531$, a complex tricyclic aziridine-containing scaffold). The magnitude of these distances indicates that the vibrational-spatial profiles of these compact architectures are reorganized compared to flexible acyclic chains.

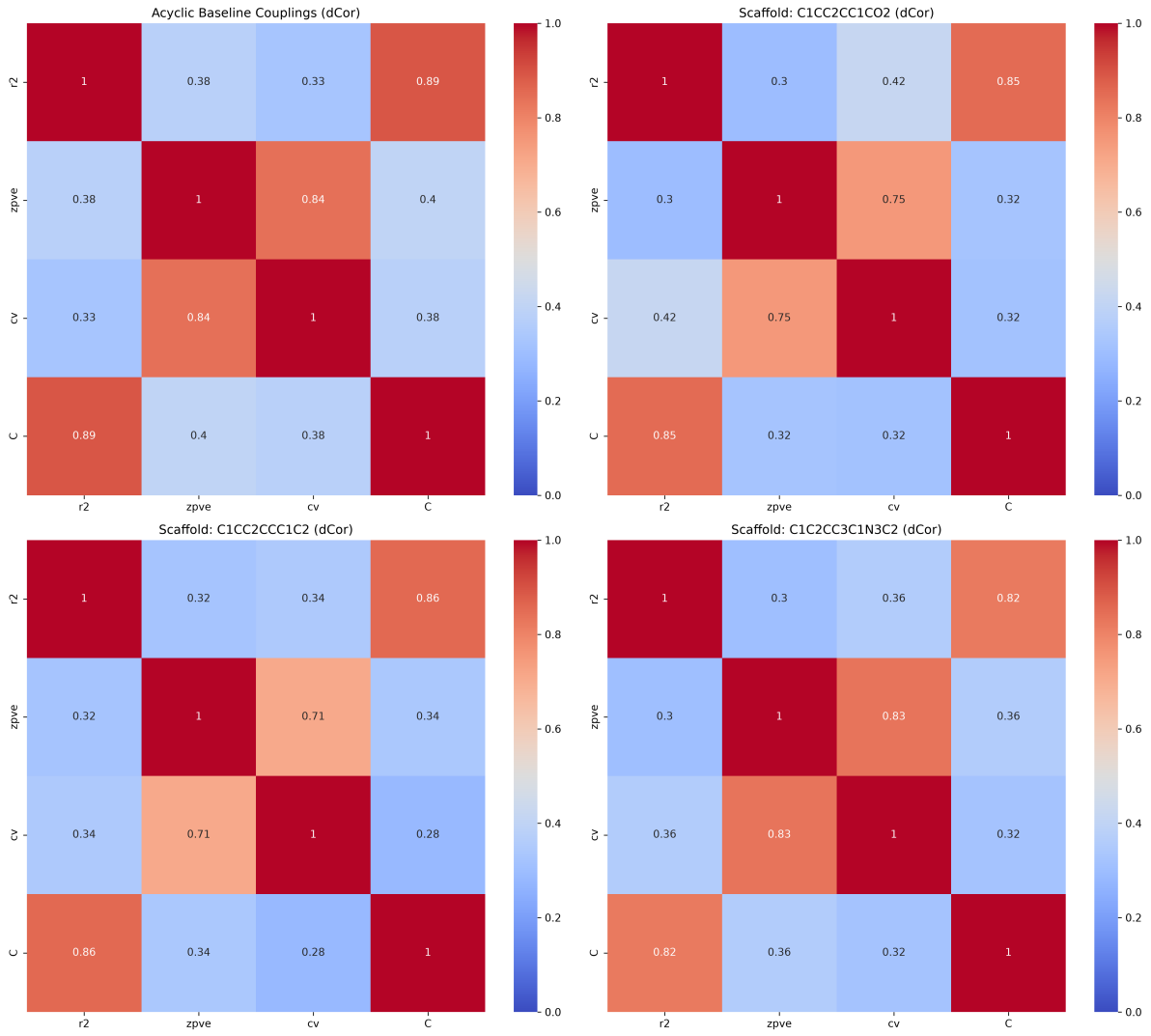

Figure 2: Distance correlation ($dCor$) matrices comparing the internal correlation structure of the 4D residual manifold (r^2, $zpve$, c_v, C) for the acyclic baseline and three high-deformation polycyclic scaffolds (C1CC2CC1CO2, C1CC2CCC1C2, and C1C2CC3C1N3C2). The heatmaps illustrate the topological rewiring of property couplings driven by molecular architecture. While the acyclic population exhibits standard couplings between thermodynamic properties ($zpve$ and c_v) and spatial properties (r^2 and C), these relationships are altered in cyclic systems. This decoupling indicates that the spatial extent of the electron density and the vibrational energy profile are governed differently under cyclic constraints, demonstrating that complex ring topologies reorganize the joint physical states of molecules compared to flexible acyclic chains.

3.3 Topological rewiring of property couplings

Beyond shifts in mean property values, the internal relationships between physical properties are altered by the molecular scaffold. This phenomenon is captured by the distance correlation ($dCor$) matrices shown in Figure 2.

In the acyclic population, the residual manifold is characterized by a strong coupling between $zpve$ and c_v ($dCor \approx 0.84$), reflecting a standard thermodynamic relationship where vibrational energy levels dictate heat capacity. A similarly strong coupling exists between r^2 and C ($dCor \approx 0.85$), as both are sensitive to the overall spatial extent and moment of inertia. However, in high-deformation scaffolds, these standard

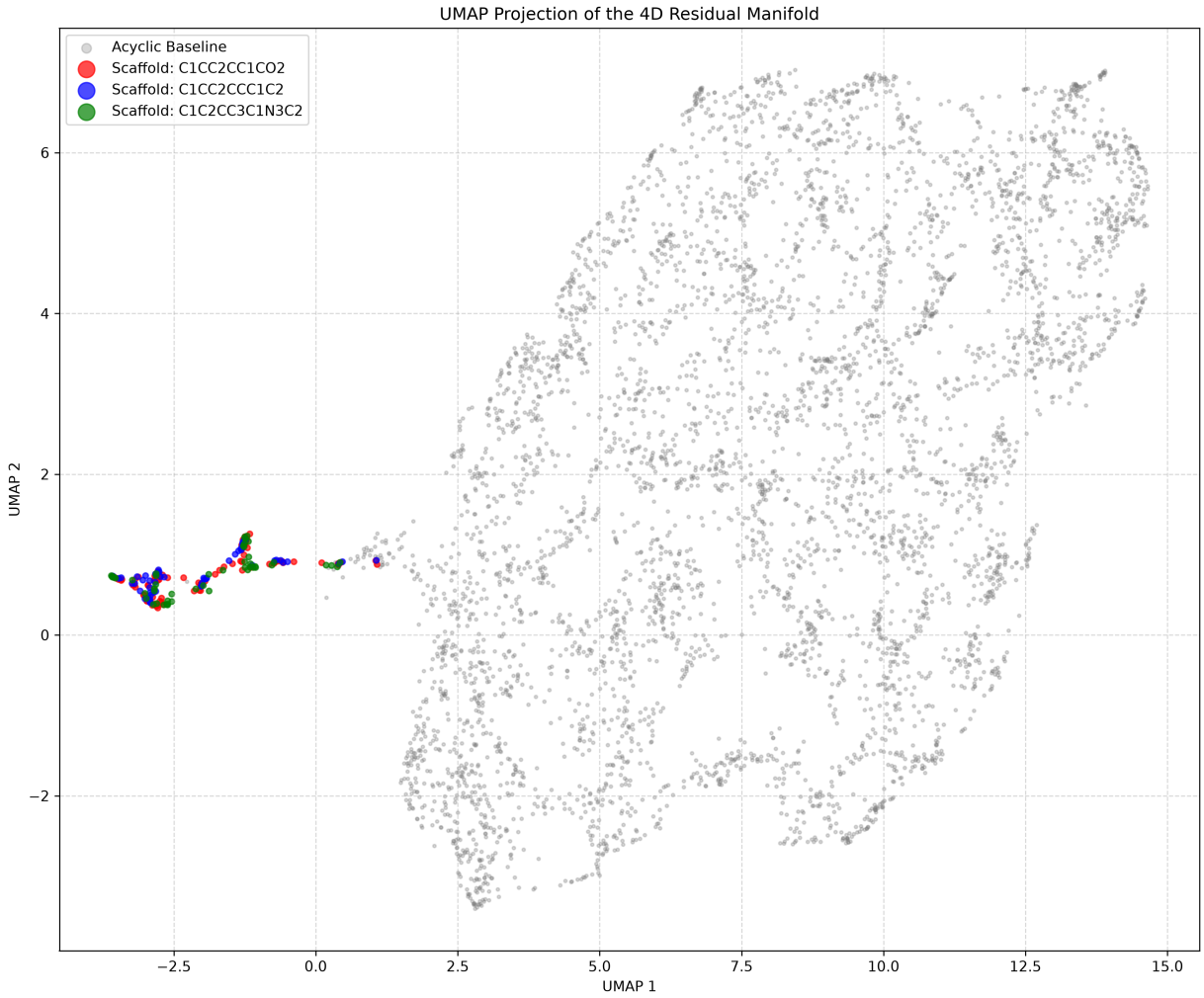

Figure 3: UMAP projection of the 4D residual manifold illustrating the reorganization of molecular property space driven by chemical topology. The acyclic baseline (grey) forms a broad, diffuse distribution, representing standard thermodynamic and spatial property couplings. In contrast, high-deformation polycyclic scaffolds—C1CC2CC1CO2 (red), C1CC2CCC1C2 (blue), and C1C2CC3C1N3C2 (green)—occupy localized regions that are largely disjoint from the main acyclic density. This separation demonstrates that complex cyclic architectures shift molecules into unique, scaffold-specific physical state spaces that are distinct from acyclic molecules, deforming the vibrational-spatial manifold.

couplings are altered. For instance, in the strained bicyclic scaffold C1CC2CC1CO2, the relationship between spatial properties (r^2, C) and thermodynamic properties ($zpve$, c_v) is modified. This suggests that under cyclic constraints, the spatial extent of the electron density becomes less coupled to the vibrational energy profile. In such strained rings, electronic spatial extent may be influenced by the specific geometry of the ring strain, while $zpve$ is governed by the high-frequency vibrations of the strained bonds, leading to a deviation from the correlations observed in flexible acyclic chains.

3.4 Manifold visualization via UMAP

The reorganization of the property space is visualized by projecting the 4D residual manifold into two dimensions using UMAP, as presented in Figure 3.

The acyclic baseline forms a broad, diffuse distribution in the center of the projection, representing the

continuous flexibility of open-chain structures. In contrast, the high-deformation cyclic scaffolds occupy localized regions of the manifold that are largely disjoint from the main acyclic density. The separation of scaffolds such as C1CC2CC1CO2 and C1CC2CCC1C2 in the UMAP space suggests that these molecules inhabit a physical state space that is distinct from that of acyclic molecules. Furthermore, the fact that these cyclic clusters are distinct from one another suggests that different topologies do not merely shift away from the acyclic baseline along a single vector, but rather deform the manifold into unique, scaffold-specific configurations.

3.5 Robustness and sensitivity analysis

To ensure that the observed manifold deformations were not artifacts driven by the 3,054 molecules identified as having potentially inconsistent DFT geometries, a sensitivity analysis was performed. Putative outliers were flagged and excluded using the Mahalanobis distance of the 4D residual vectors. The correlation between the original W_2 distances and the robust W_2 distances (calculated after excluding the outliers) was high (Pearson $r \approx 0.99$), with a mean absolute error of 0.05. This confirms that the observed manifold deformations and the rewiring of property couplings are associated with molecular topology, rather than statistical anomalies driven by a small subset of erroneous data points.

4 Conclusions

In this study, we introduced an information-geometric framework to track the deformation of the vibrational-spatial property manifold across discrete structural families in the QM9 dataset. By strictly conditioning electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), heat capacity (c_v), and the rotational constant (C) on exact atom counts, we successfully isolated the variance driven exclusively by chemical topology from simple stoichiometric determinism.

Our application of Optimal Transport revealed that the joint physical state of a molecule is fundamentally reorganized by its underlying Murcko scaffold. We demonstrated that topological complexity, particularly the number of rings, monotonically drives the deformation of the property manifold. Highly strained architectures, such as bicyclic and tricyclic systems, exhibit exact Wasserstein-2 distances that significantly diverge from size-matched acyclic baselines. As visually confirmed by UMAP projections, these constrained topologies force molecules into joint property configurations that are topologically forbidden for flexible open-chain structures.

Beyond mere shifts in property magnitudes, our analysis highlighted a profound rewiring of internal property couplings. While acyclic molecules exhibit strong, standard correlations between thermodynamic and spatial properties, high-deformation scaffolds systematically decouple these relationships. Specifically, the spatial extent of the electron density becomes decoupled from the vibrational energy profile under severe cyclic constraints, proving that complex ring topologies fundamentally alter the physical rules governing the molecule's internal state.

Validated by a rigorous sensitivity analysis that excluded putative geometric outliers, these findings confirm that the observed manifold deformations are robust, intrinsic consequences of molecular architecture rather than statistical artifacts. Ultimately, this work establishes that molecular scaffolds act as topological operators that dynamically reshape the coupled thermodynamic and spatial landscapes of small organic molecules. By moving beyond the analysis of isolated properties, this information-geometric approach provides a mathematically rigorous foundation for understanding how chemical connectivity dictates the holistic physical state of matter.