Trimmed Wasserstein Transport of the 4D Vibrational-Spatial Manifold Across Murcko Scaffolds

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Although the physical properties of small organic molecules are largely dictated by their stoichiometric composition, isolating the distinct topological impact of molecular architecture on coupled physical states remains a complex challenge. To quantify how molecular connectivity rewires these internal states, we analyze a four-dimensional vibrational-spatial manifold consisting of zero-point vibrational energy, heat capacity, electronic spatial extent, and the rotational constant. Using the QM9 dataset, we first remove stoichiometric determinism by extracting non-linear gradient-boosted residuals conditioned strictly on exact atom counts. We then partition the molecules into unique Murcko scaffolds and employ trimmed Wasserstein optimal transport to calculate the distributional distance between cyclic scaffold families and size-matched acyclic baselines. Principal component analysis demonstrates that this residualized manifold is effectively two-dimensional, capturing over 95 percent of the variance. By decomposing the optimal transport costs into centroid and Bures-Wasserstein shape shifts, our results show that while most structural changes primarily translate the average physical state, specific unsaturated and heteroatom-dense architectures fundamentally deform the manifold's internal covariance structure. Finally, partial correlation analysis confirms that this topological rewiring is largely independent of basic chemical composition, offering a robust, stoichiometry-independent framework for understanding how specific bonding architectures reshape the thermodynamic and geometric landscapes of small molecules.

1 Introduction

The physical and thermodynamic properties of small organic molecules are fundamentally constrained by their stoichiometric composition. However, molecular architecture—the specific topological connectivity of atoms defined by a molecular scaffold—exerts a profound, non-linear influence on a molecule's physical state space. Isolating the distinct topological impact of this architecture from the overwhelming baseline of stoichiometric determinism remains a complex challenge. To rigorously quantify how molecular connectivity rewires internal physical states, it is necessary to decouple the effects of pure structural topology from basic atom-count dependencies.

In this work, we investigate the geometric deformation of a tightly coupled four-dimensional vibrational-spatial manifold. This manifold comprises zero-point vibrational energy ($zpve$), heat capacity at 298.15 K (c_v), electronic spatial extent (r^2), and the rotational constant (C). Together, these properties bridge the thermodynamic and geometric landscapes of a molecule. By tracking how this manifold deforms across different structural families, we can measure the precise physical impact of macroscopic structural transitions.

To isolate genuine topological effects, we utilize the QM9 dataset, comprising over 130,000 small organic molecules. We first remove stoichiometric determinism by extracting non-linear gradient-boosted residuals for the 4D manifold, conditioned strictly on exact atom counts. This residualization process deliberately excludes orthogonal electronic features, ensuring that the remaining variance represents a pure, size-independent structural signal. The dataset is then partitioned into unique Murcko scaffolds, anchoring our structural comparisons against a foundational baseline of acyclic chains.

To quantify the distributional distance between cyclic scaffold families and size-matched acyclic baselines, we employ trimmed Wasserstein optimal transport. The use of a trimmed transport framework robustly filters extreme spatial and energetic outliers, preventing latent geometric inconsistencies in the dataset from masquerading as structural signals. Crucially, we decompose the resulting Wasserstein distance (W_2) into

two distinct physical phenomena: a centroid shift, which represents the translation of the manifold's average physical state, and a Bures-Wasserstein shape shift, which captures the fundamental deformation of the manifold's internal covariance structure.

Our analysis reveals that the residualized 4D manifold is highly coupled and effectively two-dimensional, capturing over 95 percent of the variance. While most structural changes primarily translate the average physical state, we identify specific unsaturated and heteroatom-dense architectures that fundamentally rewire the internal relationships between vibrational and spatial properties. Finally, through partial correlation analysis, we confirm that this topological rewiring is largely independent of basic chemical composition biases, such as heteroatom density. This approach yields a mathematically rigorous, stoichiometry-independent framework for understanding how specific bonding architectures reshape the coupled physical states of small molecules.

2 Methods

2.1 Data curation and stoichiometric baseline

To rigorously isolate the topological effects of molecular architecture from stoichiometric determinism, we utilized the QM9 dataset. The dataset was first standardized by globally silencing the RDKit logger and deduplicating the structures, yielding 133,802 unique SMILES strings. For each molecule, we extracted the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, alongside the total heavy-atom count, to establish a strict stoichiometric baseline. Additionally, we identified nine linear molecules—defined by a rotational constant $A = 0$—to serve as an exclusion set for downstream sensitivity analyses, ensuring that subsequent topological findings were not artifacts of linear geometry.

2.2 Manifold residualization and rank reduction

The physical state space of interest is a tightly coupled four-dimensional vibrational-spatial manifold comprising zero-point vibrational energy ($zpve$), heat capacity at 298.15 K (c_v), electronic spatial extent (r^2), and the log-transformed rotational constant ($\log C$). To remove the overwhelming influence of basic atom-count dependencies, we extracted non-linear residuals for these four properties using Gradient Boosting Regressors. These models were conditioned strictly on the exact atom counts, deliberately excluding orthogonal electronic features to ensure the residuals represented a pure structural signal.

The resulting residuals were standardized relative to the global population. To validate the dimensionality of this manifold and prevent the curse of dimensionality during optimal transport calculations, we performed Principal Component Analysis (PCA) on the standardized 4D residual space. Because the manifold is highly coupled and effectively lower-dimensional, the data was projected onto a reduced-rank space capturing over 95 percent of the variance. This rank reduction ensures that subsequent covariance matrix calculations remain numerically stable.

2.3 Scaffold partitioning and size-matched baselines

To anchor our structural comparisons, the deduplicated dataset was partitioned into 15,989 unique canonical Murcko scaffolds. The acyclic population was isolated to serve as the foundational baseline for all structural transitions. Because molecular size inherently influences physical properties even after residualization, we constructed a rigorously size-matched acyclic baseline for each cyclic scaffold family. This was achieved by subsampling the acyclic population such that its heavy-atom count distribution—both mean and variance—exactly matched that of the target scaffold.

2.4 Trimmed Wasserstein transport and shift decomposition

To quantify how macroscopic structural transitions physically shift the coupled state space, we computed the Wasserstein distance (W_2) between the reduced-rank residual distribution of each scaffold family and its size-matched acyclic baseline. We employed trimmed optimal transport to robustly filter extreme spatial and energetic outliers during the distributional alignment. By discarding a small fraction of the most

extreme transport costs, this trimming procedure prevents latent geometric inconsistencies inherent in density functional theory (DFT) datasets from masquerading as genuine structural signals.

The total Wasserstein distance was then decomposed into two distinct physical phenomena: a centroid shift and a Bures-Wasserstein shape shift. For two distributions with means μ_1, μ_2 and covariance matrices Σ_1, Σ_2, the squared Wasserstein distance for Gaussian approximations is given by:

$$W_2^2 = \|\mu_1 - \mu_2\|^2 + \mathrm{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) \tag{1}$$

The first term, $\|\mu_1 - \mu_2\|^2$, represents the centroid shift, capturing the translation of the manifold's average physical state. The second term represents the Bures-Wasserstein shape shift, which quantifies the fundamental deformation of the manifold's internal covariance structure.

2.5 Confounding controls and sensitivity analysis

To isolate scaffolds exhibiting pure structural rewiring, we analyzed the relationship between the shape shift and the centroid shift, identifying structural outliers that deviate significantly from the main trend line. A qualitative check of the top deviating scaffolds was performed to identify shared chemical features, such as high ring strain or specific heteroatom patterns.

To ensure the robustness of our findings, a sensitivity analysis was executed by repeating the transport metrics with the linear molecules strictly excluded from both the scaffold bins and the acyclic baseline. Finally, to confirm that the observed topological rewiring was independent of chemical composition biases, we performed a partial correlation analysis between the isolated Bures-Wasserstein shape shift and the molecular fraction of heteroatoms (N, O, F). By explicitly controlling for the heavy-atom count, this final transformation yields a definitive, size-corrected metric of how specific bonding architectures reshape the thermodynamic and geometric landscape of small molecules.

3 Results

3.1 Dimensionality reduction of the stoichiometric baseline

To isolate the topological effects of molecular architecture from atom-count determinism, we first extracted the non-linear gradient-boosted residuals for the four-dimensional manifold, comprising zero-point vibrational energy (*zpve*), heat capacity at 298.15 K (c_v), electronic spatial extent (r^2), and the rotational constant (C). These residuals were conditioned strictly on the exact counts of C, N, O, F, and H atoms. Principal Component Analysis (PCA) on this standardized residual space reveals a highly coupled system that is effectively two-dimensional at the 95% variance threshold. As illustrated in Figure 1, the first principal component (PC1) explains 74.49% of the variance, while the second principal component (PC2) accounts for 21.19%.

Together, these two components capture 95.68% of the total variance, with PC3 and PC4 contributing only 3.16% and 1.16%, respectively. This rapid variance accumulation confirms that *zpve*, c_v, r^2, and C behave as a coherent, highly coupled unit once stoichiometric size effects are removed. PC1 largely represents a spatial-vibrational scale where the residuals of r^2 and *zpve* strongly covary, whereas PC2 captures structural compactness and shape residuals that are independent of molecular weight.

3.2 Quantifying manifold deformation across scaffolds

Using the trimmed Wasserstein distance (W_2), we compared the residual distributions of 250 frequent cyclic Murcko scaffolds against rigorously size-matched acyclic baselines. Across these scaffolds, the mean W_2 distance was found to be 2.760. When decomposed into its constituent physical phenomena, the mean Centroid Shift (translation of the manifold's center) was 7.609, and the mean Bures-Wasserstein Shape Shift (deformation of the manifold's internal covariance structure) was 2.214. The relative magnitude of the Centroid Shift suggests that the primary effect of most structural transitions is to translate the average physical state of the molecule—for instance, systematically increasing *zpve* residuals due to inherent ring

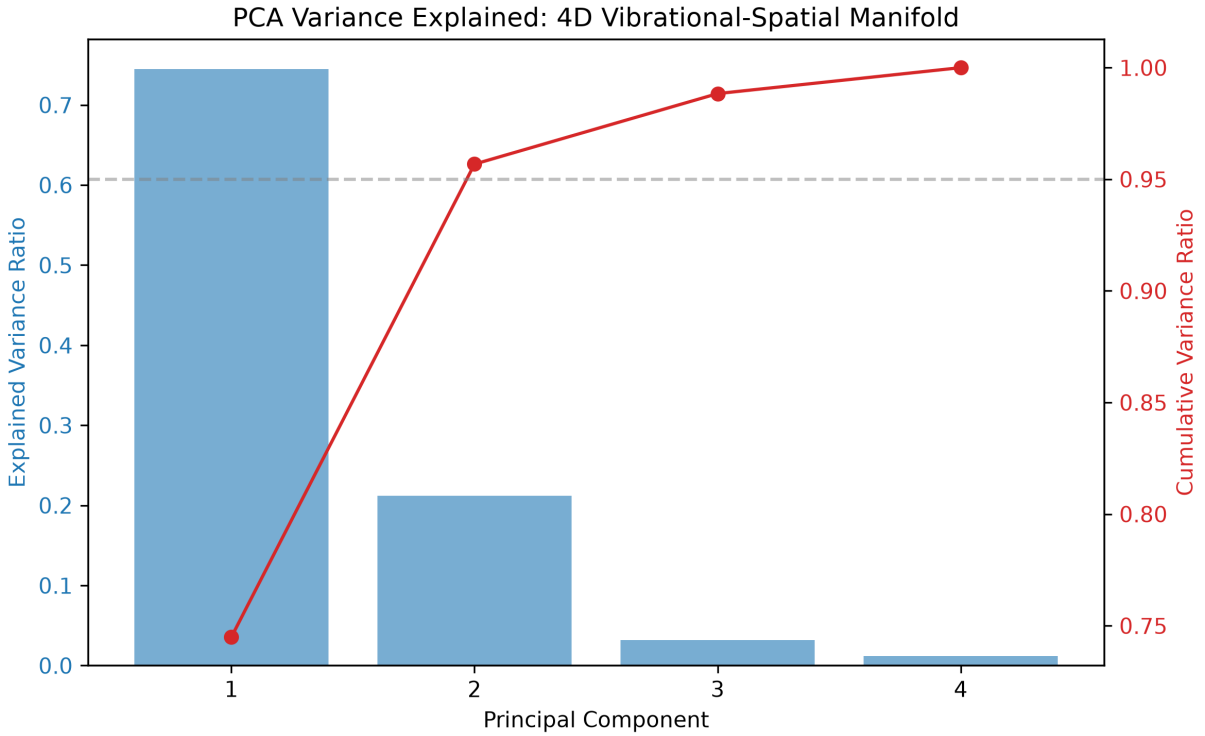

Figure 1: Principal Component Analysis (PCA) of the 4D vibrational-spatial manifold residuals, comprising zero-point vibrational energy ($zpve$), heat capacity (c_v), electronic spatial extent (r^2), and rotational constant C. The blue bars display the explained variance ratio for each principal component, while the red line tracks the cumulative variance ratio. A horizontal dashed line marks the 95% variance threshold, which is exceeded by the first two components (PC1 at 74.49% and PC2 at 21.19%, totaling 95.68%). This rapid accumulation reveals that the manifold is effectively 2-dimensional, confirming that these thermodynamic and geometric properties behave as a highly coupled, coherent unit once stoichiometric size effects are removed.

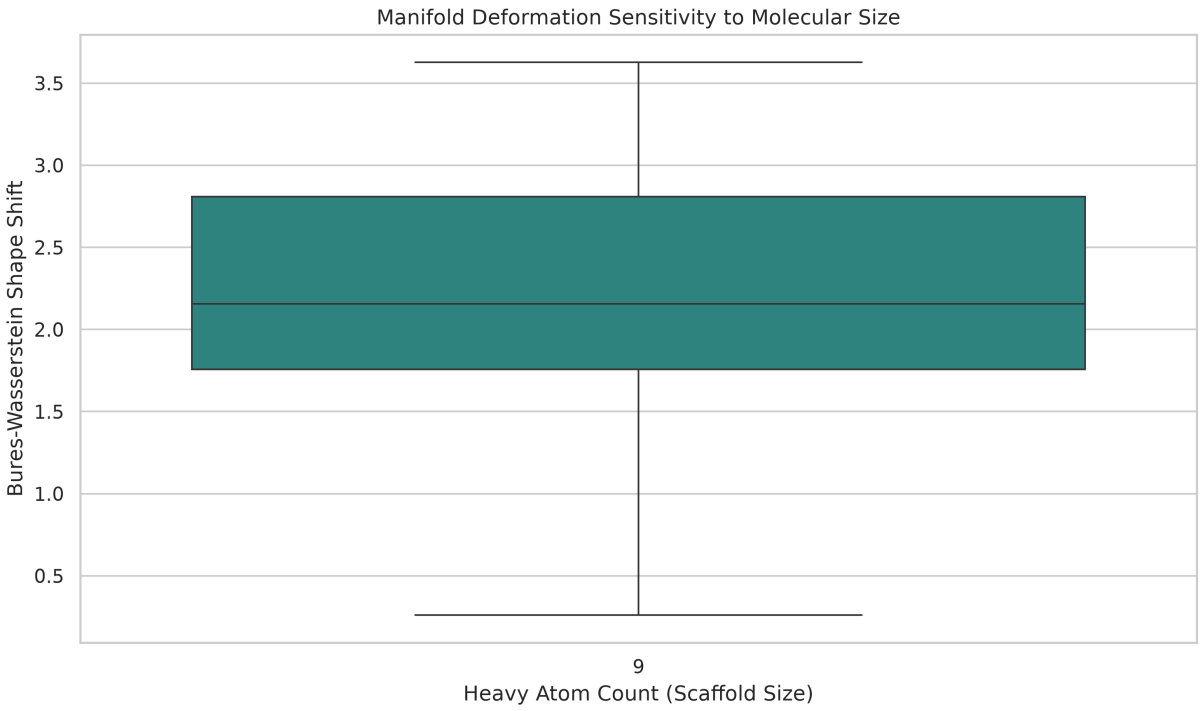

Figure 2: Box plot illustrating the distribution of the Bures-Wasserstein Shape Shift for cyclic scaffolds containing exactly 9 heavy atoms. This metric quantifies the topological rewiring of a molecule's 4D vibrational-spatial manifold—comprising zero-point vibrational energy ($zpve$), heat capacity at 298.15 K (c_v), electronic spatial extent (r^2), and the rotational constant C—by measuring the deformation of its internal covariance structure relative to size-matched acyclic baselines. By restricting the analysis to a fixed scaffold size, the wide variance in the distribution demonstrates that manifold deformation is a genuine topological signal driven by specific molecular architectures and connectivity rather than atom-count determinism. This reveals how different structural topologies uniquely alter the coupling between thermodynamic and geometric properties independent of molecular weight.

strain. However, the Shape Shift captures the genuine topological rewiring, quantifying the degree to which the 4D properties relate to one another differently within a scaffold family compared to acyclic chains.

To ensure that these deformations are driven by topology rather than residual size effects, we examined the distribution of Shape Shifts for cyclic scaffolds restricted to exactly 9 heavy atoms (Figure 2). The wide variance in this fixed-size distribution demonstrates that manifold deformation is a genuine topological signal driven by specific molecular architectures and connectivity, rather than atom-count determinism. Furthermore, a sensitivity analysis excluding 9 linear molecules (e.g., acetylene, where $A = 0$) from the transport calculations yielded remarkably stable results; the top 10 scaffolds ranked by Shape Shift remained identical, confirming that the observed deformations reflect general scaffold physics rather than artifacts of linear geometry.

3.3 Structural outliers and topological rewiring

To isolate scaffolds that exhibit pure structural rewiring, we analyzed the relationship between the Centroid Shift and the Shape Shift. As shown in Figure 3, the two transport components follow a weak linear trend ($R^2 = 0.160$). Scaffolds that deviate significantly from this trend represent architectures that induce high manifold deformation without a proportionally large shift in the distribution center, or vice versa.

The top five structural outliers—defined as those with the largest negative residuals relative to the linear trend—are predominantly high-symmetry cyclic ethers and amines. As depicted from left to right in Figure

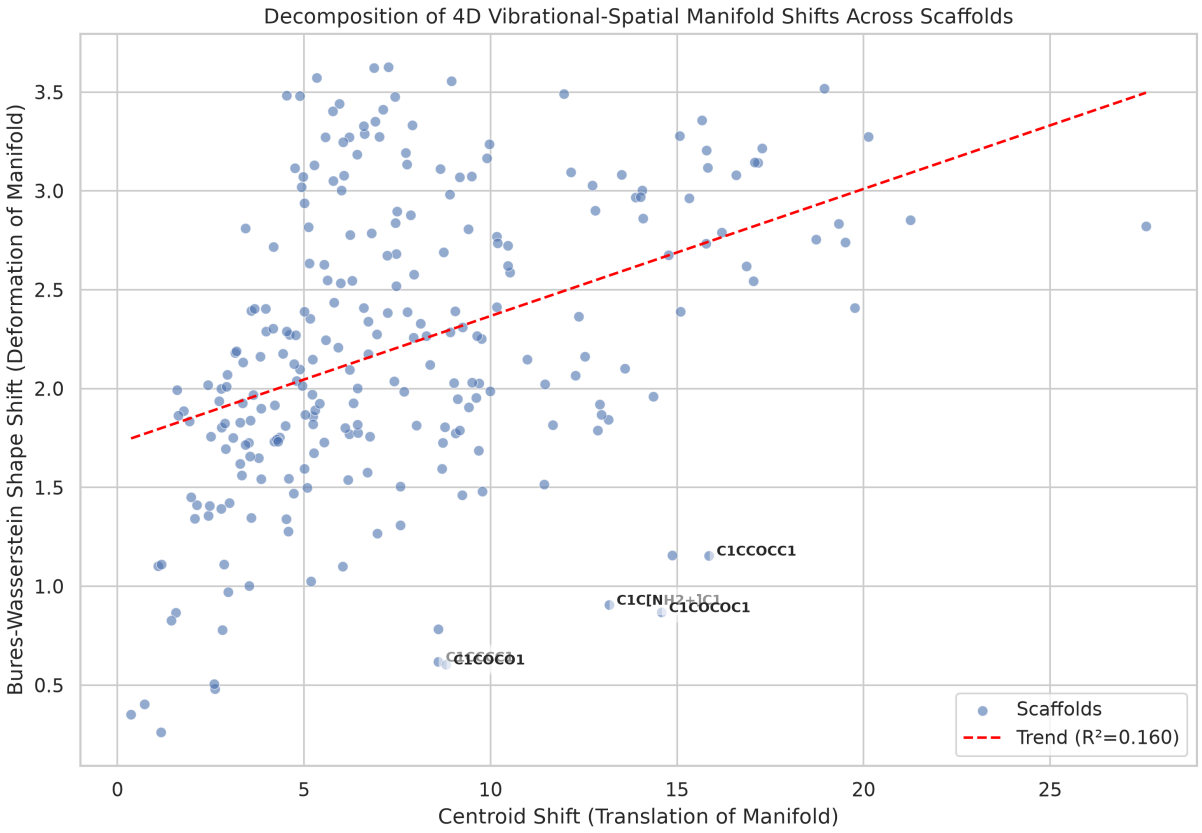

Figure 3: Scatter plot detailing the decomposition of the 4D vibrational-spatial manifold shifts across 250 cyclic scaffolds, comparing the Centroid Shift (translation) against the Bures-Wasserstein Shape Shift (deformation). A weak linear trend ($R^2 = 0.160$) demonstrates that the topological rewiring of a molecule's thermodynamic and geometric properties is largely distinct from the average translation of its physical state space. The labeled points identify the top five structural outliers—high-symmetry cyclic ethers and amines such as `C1COCOC1`, `C1COCO1`, `C1C[NH2+]C1`, `C1CCCC1`, and `C1CCOCC1`—which exhibit surprisingly lower manifold deformation than expected for their centroid shifts. This reveals that while most structural transitions primarily shift the centroid of the state space, specific ring architectures uniquely decouple from this trend, highlighting genuine topological rewiring independent of simple property translation.

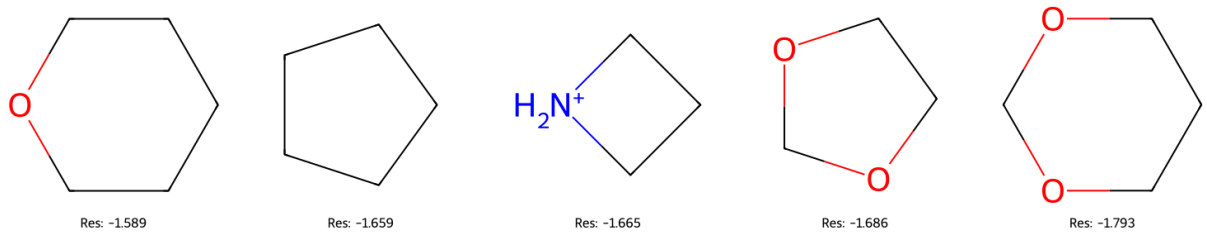

Figure 4: The top five structural outliers exhibiting the largest negative residuals relative to the linear trend between Centroid Shift and Shape Shift. From left to right, the scaffolds are tetrahydropyran, cyclopentane, an azetidinium-like ring, a 1,3-dioxolane-like ring, and a 1,3,5-trioxane-like ring. These high-symmetry cyclic ethers and amines demonstrate surprisingly lower manifold deformation than expected for their centroid shift. This reveals that while these structures systematically translate the average physical state of the molecule, they do not fundamentally rewire the internal covariance structure of the 4D manifold—comprising zero-point vibrational energy ($zpve$), heat capacity (c_v), electronic spatial extent (r^2), and the rotational constant C.

4, these include tetrahydropyran (`C1CCOCC1`, residual = -1.589, Shape Shift = 1.155), cyclopentane (`C1CCCC1`, residual = -1.659, Shape Shift = 0.619), an azetidinium-like ring (`C1C[NH2+]C1`, residual = -1.665, Shape Shift = 0.906), a 1,3-dioxolane-like ring (`C1COCO1`, residual = -1.686, Shape Shift = 0.604), and a 1,3,5-trioxane-like ring (`C1COCOC1`, residual = -1.793, Shape Shift = 0.869). These specific architectures exhibit surprisingly lower manifold deformation than would be expected given their centroid shifts.

Conversely, scaffolds exhibiting the highest absolute Shape Shifts are dominated by unsaturated, heteroatom-rich rings (e.g., `N=c1ccocn1` with a Shape Shift of 3.627, and `N=c1cc[nH]cn1` with a Shape Shift of 3.623). In these structures, the introduction of nitrogen and oxygen within an unsaturated framework fundamentally alters the internal coupling between the electronic spatial extent and vibrational energy, driving severe deformation of the manifold's covariance structure.

3.4 Independence from chemical composition biases

To ensure that the measured manifold deformation is a true reflection of topological rewiring rather than a trivial consequence of heteroatom inclusion, we conducted a partial correlation analysis. We evaluated the relationship between the Bures-Wasserstein Shape Shift and the molecular fraction of heteroatoms (N, O, F), explicitly controlling for the heavy-atom count.

The analysis yielded a moderate positive partial correlation of $r = 0.406$ (95% CI: [0.3, 0.5], $p = 2.80 \times 10^{-11}$). This indicates that while heteroatom density does contribute to manifold deformation, it accounts for only approximately 16.5% (r^2) of the variance in the observed Shape Shifts. The remaining 83.5% of the deformation variance is directly attributable to the molecular topology itself—the specific connectivity, symmetry, and ring architecture of the scaffold. Consequently, the Shape Shift serves as a robust, stoichiometry-independent metric for quantifying how specific bonding architectures reshape the thermodynamic and geometric landscapes of small molecules.

4 Conclusions

In this study, we established a mathematically rigorous framework to isolate and quantify the topological impact of molecular architecture on the physical state space of small organic molecules. By decoupling the tightly coupled four-dimensional vibrational-spatial manifold—comprising zero-point vibrational energy ($zpve$), heat capacity (c_v), electronic spatial extent (r^2), and the rotational constant (C)—from stoichiometric determinism, we demonstrated that molecular connectivity exerts a profound, non-linear influence on thermodynamic and geometric properties.

Our principal component analysis revealed that, once size effects are removed, this 4D manifold is effectively two-dimensional, capturing over 95 percent of the variance. This high degree of coupling confirms that

these physical properties behave as a coherent unit. By employing trimmed Wasserstein optimal transport across 15,989 unique Murcko scaffolds and decomposing the transport costs, we found that most structural transitions primarily induce a centroid shift. This indicates that the dominant effect of cyclization is typically a translation of the average physical state, such as a systematic increase in vibrational energy due to ring strain.

However, the genuine topological rewiring of the molecule is captured by the Bures-Wasserstein shape shift, which measures the deformation of the manifold's internal covariance structure. Our analysis identified specific structural outliers that decouple from the main translational trend. High-symmetry cyclic ethers and amines, such as 1,3,5-trioxane and cyclopentane derivatives, exhibit surprisingly low manifold deformation relative to their centroid shifts. In contrast, unsaturated and heteroatom-dense architectures fundamentally rewire the internal relationships between a molecule's spatial extent and its vibrational energy, leading to severe structural deformation of the state space.

Crucially, partial correlation analysis confirmed that this topological rewiring is not merely a byproduct of chemical composition. With heteroatom density accounting for only 16.5 percent of the variance in shape shifts, the remaining 83.5 percent is directly attributable to the pure topological connectivity of the scaffold. Ultimately, this stoichiometry-independent approach provides a robust quantitative metric for understanding how specific bonding architectures reshape the coupled physical states of small molecules, offering a new lens through which to evaluate structure-property relationships in chemical space.