Persistent Homology of 12-Dimensional Property Manifolds to Quantify Topological Rewiring Across Murcko Scaffolds

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The joint structure of molecular properties is fundamentally influenced by structural architecture, yet isolating this geometric effect from stoichiometric composition remains challenging. To rigorously quantify how macroscopic spatial confinement rewires molecular physics, we isolate structural variance by extracting gradient-boosted atom-count residuals for twelve physical properties across the QM9 dataset. Partitioning the deduplicated data into an acyclic baseline and distinct cyclic Murcko scaffolds, we construct Vietoris-Rips complexes on the twelve-dimensional residual point clouds to extract their persistent homology. By computing Wasserstein distances between the persistence diagrams of different structural families, we directly measure the geometric deformation of the property space. Our analysis reveals that rigid spatial constraints, particularly in highly strained bicyclic systems, force the joint property manifold into topologically distinct configurations with transport costs far exceeding a bootstrapped noise floor. Furthermore, correlation analysis with structural descriptors demonstrates that while ring count and structural rigidity drive this manifold reorganization, the presence of heteroatoms and aromaticity mitigates topological divergence, highlighting a fundamental balance between geometric confinement and electronic delocalization in shaping the physical state space of organic molecules.

1 Introduction

The fundamental goal of chemical design is to understand the mapping between molecular architecture and physical properties. While the stoichiometric composition of a molecule dictates the baseline for many thermodynamic and electronic features, the specific geometric arrangement of these atoms—defined by the molecular scaffold—exerts a profound influence on the joint structure of molecular properties. Isolating this purely structural effect from stoichiometric determinism remains a significant challenge. When analyzing large chemical spaces, the dominant variance explained by simple atom counts often obscures the subtle, high-dimensional deformations induced by spatial confinement, ring strain, and structural rigidity.

To rigorously quantify how macroscopic spatial confinement rewires molecular physics, it is necessary to move beyond scalar dependence metrics and examine the joint property space as a high-dimensional manifold. Topological data analysis, particularly persistent homology, provides a mathematically rigorous, coordinate-free framework to characterize the shape and connectivity of such complex data structures. By evaluating the persistent homology of property manifolds, we can extract topological signatures that are robust to latent geometric artifacts and coordinate transformations, allowing for a direct comparison of the physical state spaces occupied by different structural families.

In this study, we introduce a framework to isolate and quantify the topological rewiring of 12-dimensional property manifolds across distinct Murcko scaffolds in the QM9 dataset. We first remove stoichiometric determinism by extracting gradient-boosted atom-count residuals for twelve non-redundant physical properties, thereby isolating the purely structural and geometric variance. Partitioning the deduplicated dataset into an acyclic baseline and highly populated cyclic scaffolds, we construct Vietoris-Rips complexes directly on each family's residual point cloud. By computing the exact Wasserstein distances between the persistence diagrams of distinct structural families, we directly measure the geometric deformation—or topological transport cost—of the property space.

Our results demonstrate that rigid spatial constraints fundamentally reorganize the joint property manifold into topologically distinct configurations, with transport costs far exceeding a bootstrapped noise floor.

Furthermore, we show that this manifold rewiring is driven by a delicate balance between geometric confinement and electronic delocalization. While increased ring count and structural rigidity drive topological divergence from the acyclic baseline, the introduction of heteroatoms and aromaticity mitigates this effect, standardizing the joint property structure. Through this topological lens, we provide a robust quantification of how molecular scaffolds shape the physical state space of organic molecules.

2 Methods

2.1 Data curation and stoichiometric baseline extraction

To isolate the geometric and structural variance of molecular properties from their stoichiometric determinism, we first established a compositional baseline using the QM9 dataset. The dataset was deduplicated based on SMILES strings to prevent data leakage across validation splits. We selected twelve non-redundant physical properties to define the physical state space: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), frontier orbital energies (HOMO, LUMO), electronic spatial extent $(\langle R^2 \rangle)$, zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity at 298.15 K (C_v), and atomization energy at 0 K (U_0^{atom}). Nine linear molecules were explicitly excluded to prevent singularities in the rotational constants from skewing the global coordinate system. Additionally, a $\log(A + 1)$ transformation was applied to the rotational constant A to mitigate its heavy-tailed distribution.

To extract the purely structural variance, we regressed out the atom-count baseline. For each molecule, exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were extracted. A multi-output gradient boosting regressor was trained to predict the twelve physical properties strictly from these five atom counts. Hyperparameters were constrained to prevent overfitting to specific molecular instances. The regression was audited by evaluating the R^2 values; properties not strictly extensive were verified to have $R^2 < 0.99$ to ensure no data leakage. The residuals from this model, representing the variance unexplained by stoichiometry, formed the 12-dimensional property manifold used for subsequent topological analysis.

2.2 Manifold construction and artifact pruning

The 12-dimensional residual dataset was globally standardized using a standard scaler fit across all molecules. This global normalization is critical to preserving the relative geometric deformations across different molecular families. To partition the chemical space by structural architecture, we computed the Murcko scaffold for each molecule. The dataset was divided into an acyclic baseline family (molecules with an empty scaffold) and distinct cyclic scaffold families. To ensure statistical robustness in the topological analysis, only cyclic families containing at least 500 molecules were retained.

Within each retained scaffold family, including the acyclic baseline, we computed the Mahalanobis distance of the scaled residuals using a robust covariance estimator. Molecules exhibiting a Mahalanobis distance greater than three standard deviations from their family centroid were discarded. This step effectively pruned latent geometric artifacts originating from density functional theory (DFT) convergence failures or extreme outliers. Following artifact removal, the local intrinsic dimensionality of each family's residual manifold was estimated using a nearest-neighbors maximum likelihood approach to verify sufficient manifold complexity. Finally, a principal component analysis (PCA) was fit on the globally scaled and pruned dataset to establish a shared, aligned coordinate system. Each scaffold family was projected into this PCA space, retaining the principal components that cumulatively explained 95% of the global variance. The PCA-transformed coordinates were subsequently standardized to ensure an isotropic distance metric for topological extraction.

2.3 Topological data analysis via persistent homology

To extract a coordinate-free representation of the joint physical state space, we applied persistent homology to the PCA-aligned residual manifolds. To manage computational complexity, each scaffold family was randomly subsampled to a maximum of 2,000 points. A global reference point cloud was similarly generated by subsampling 2,000 points from the entire dataset.

Vietoris-Rips complexes were constructed on the point clouds of each family using the `gudhi` library. A consistent maximum edge length (filtration parameter) and a simplex tree filtration limit were applied

uniformly across all families to ensure comparability. We computed the persistence diagrams for connected components (H_0) and one-dimensional topological features (H_1). From these diagrams, we derived scalar topological summaries, specifically the total persistence and persistence entropy, to quantify the fragmentation, density distribution, and structural organization of each family's property manifold.

2.4 Topological transport cost and structural correlation

To directly measure the geometric deformation of the property space induced by spatial confinement, we computed the exact Wasserstein distance between the persistence diagrams of distinct structural families. This metric, termed the topological transport cost, quantifies the minimal deformation required to map one family's joint property distribution onto another.

To distinguish true physical manifold reorganization from sampling variance, we established a topological noise floor via a bootstrap procedure. For each scaffold family, we performed 10 iterations of random 50/50 splits, computing the exact Wasserstein distance between the halves using the `POT` library. The mean and standard deviation of these intra-family distances provided a rigorous baseline; inter-family transport costs were only considered significant if they substantially exceeded this noise floor. Finally, the validated topological transport costs (relative to the acyclic baseline) and scalar summaries were correlated with structural descriptors, including ring count, heteroatom count, and aromaticity indices, to rigorously quantify the chemical mechanisms driving the topological rewiring of the molecular property space.

3 Results

The joint structure of molecular properties is fundamentally influenced by structural architecture, independent of stoichiometric composition. By isolating the variance not explained by simple atom counts and applying persistent homology to the resulting property residuals, this analysis identifies how specific molecular scaffolds reorganize the 12-dimensional physical state space of the QM9 dataset.

3.1 Isolation of structural variance

To ensure the identified structure reflects chemical architecture rather than molecular size, a stoichiometric baseline was established using gradient-boosted regression on counts of C, N, O, F, and H atoms, as detailed in the Methods. This procedure confirmed that thermodynamic properties (U_0, ZPVE, C_v) are highly deterministic based on composition. By extracting residuals from this baseline, a 12-dimensional manifold was constructed from properties including rotational constants (A, B, C), dipole moment (μ), polarizability (α), frontier orbital energies (HOMO, LUMO), and electronic spatial extent ($\langle R^2 \rangle$).

Global standardization and UMAP visualization of these residuals reveal that molecules do not populate the property space uniformly. Instead, they form distinct clusters and filaments corresponding to their Murcko scaffolds (Figure 1). Acyclic molecules occupy a central, diffuse region, while cyclic and polycyclic families deform this distribution into specialized regions of the manifold. This suggests that spatial confinement rewires the joint relationships between electronic and geometric properties.

3.2 Topological manifold analysis via persistent homology

The property manifolds of 30 major structural families (minimum 500 molecules each) were characterized using H_0 persistent homology, which measures the fragmentation and density distribution of the point clouds.

3.2.1 Cluster fragmentation and noise floor

A bootstrapped noise floor was established to distinguish physical manifold structure from sampling variance using split-half Wasserstein distances. The acyclic baseline exhibited a noise floor of 8.03 ± 3.22, while the most populated cyclic scaffold (cyclopropane) showed 7.43 ± 2.50. The average noise floor across these benchmarks was 7.73. This low noise provides a rigorous threshold for identifying meaningful structural reorganization.

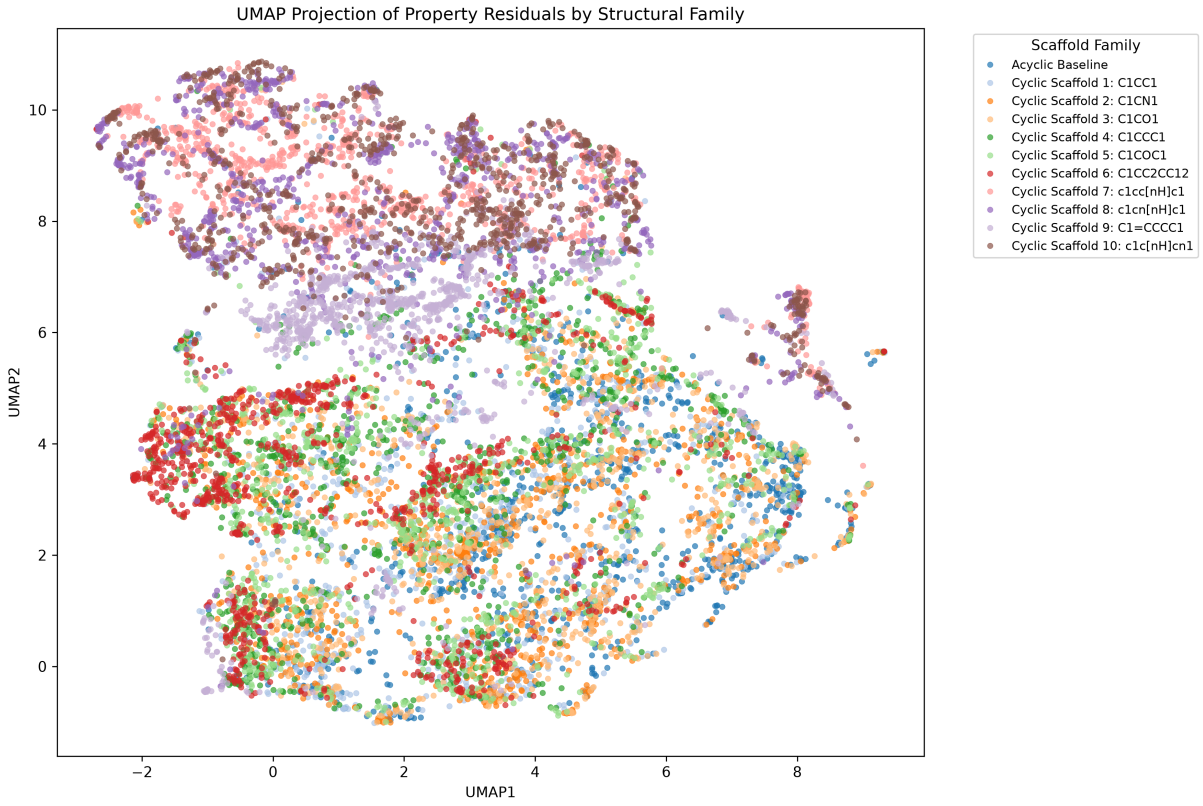

Figure 1: UMAP projection of 12-dimensional property residuals for the acyclic baseline and ten cyclic scaffold families from the QM9 dataset. The residuals, isolated from stoichiometric effects, reveal that molecules populate the property space non-uniformly, forming distinct clusters and filaments based on their Murcko scaffolds. Acyclic molecules occupy a central, diffuse region, whereas cyclic and polycyclic families deform the distribution into specialized, topologically distinct regions of the manifold. This visualizes how spatial confinement and structural rigidity rewire the joint relationships between electronic and geometric properties, overriding the typical correlations observed in flexible acyclic systems.

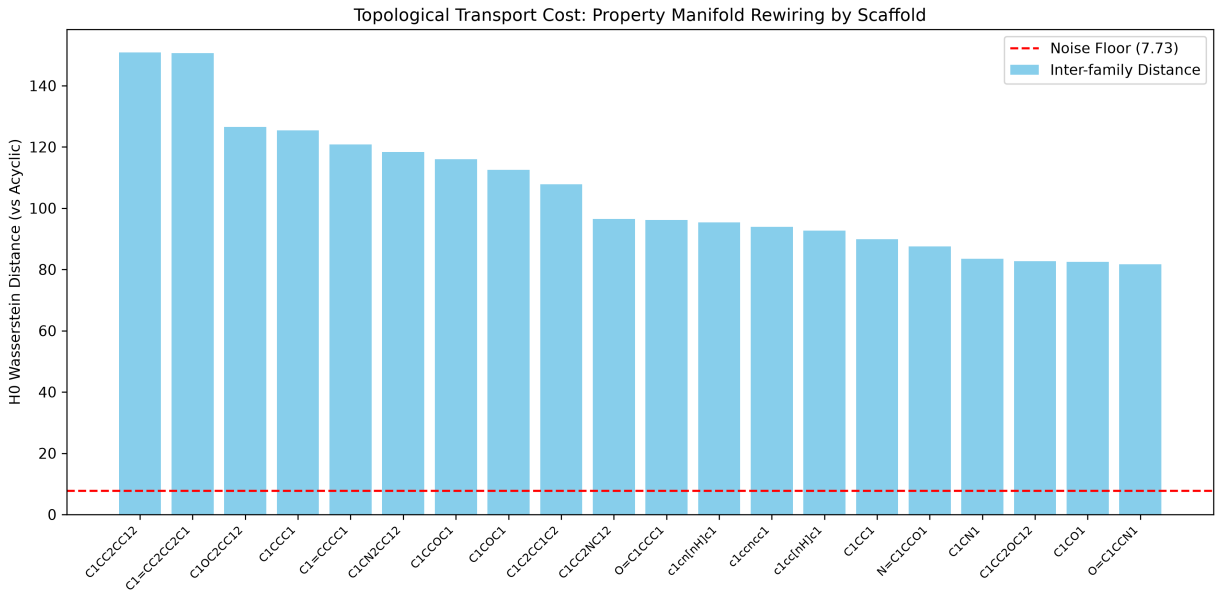

Figure 2: Topological transport cost, measured as the H_0 Wasserstein distance, quantifying the property manifold rewiring of various molecular scaffolds relative to the acyclic baseline. The red dashed line denotes the bootstrapped noise floor of 7.73, confirming that the observed manifold deformations significantly exceed sampling variance and represent meaningful structural reorganization. Highly strained bicyclic systems, such as `C1CC2CC12` and `C1=CC2CC2C1`, display the most extreme topological distances. This reveals that rigid spatial constraints and cyclization fundamentally reorganize the 12-dimensional physical state space, forcing geometric and electronic properties into topologically distinct configurations compared to flexible acyclic molecules.

The scalar topological metrics for H_0 diagrams quantify how clustered or fragmented a family's property distribution is. The acyclic baseline has the highest total persistence at 1466.76, reflecting a more dispersed property distribution. In contrast, the most rewired systems, such as bicyclo[2.1.0]pent-2-ene, show values as low as 219.34. Similarly, the acyclic baseline persistence entropy is 7.56, whereas more constrained cyclic systems generally show lower entropy (e.g., 6.21 for the `c1nc[nH]n1` scaffold). Lower entropy suggests a more tightly organized joint property structure.

3.2.2 Quantifying topological rewiring

The topological transport cost, defined as the H_0 Wasserstein distance between a scaffold family and the acyclic baseline, measures the degree of manifold deformation required to map one family's joint property distribution onto another. The results show massive reorganization far exceeding the noise floor (Figure 2).

The most extreme rewiring is observed in highly strained bicyclic systems (Table 1). These structures force the joint property manifold into configurations that are topologically distinct from the acyclic baseline, indicating that rigid spatial constraints override the typical correlations observed in flexible molecules.

3.3 Structural drivers of manifold reorganization

Correlation analysis between topological metrics and structural descriptors identifies the chemical mechanisms driving this manifold rewiring (Table 2 and Figure 3).

3.3.1 Ring strain and confinement

The number of rings shows a positive correlation with the rewiring distance ($R = 0.454$, Figure 3). Polycyclic systems exhibit the highest distances (Table 1), confirming that cyclization is a primary driver of manifold

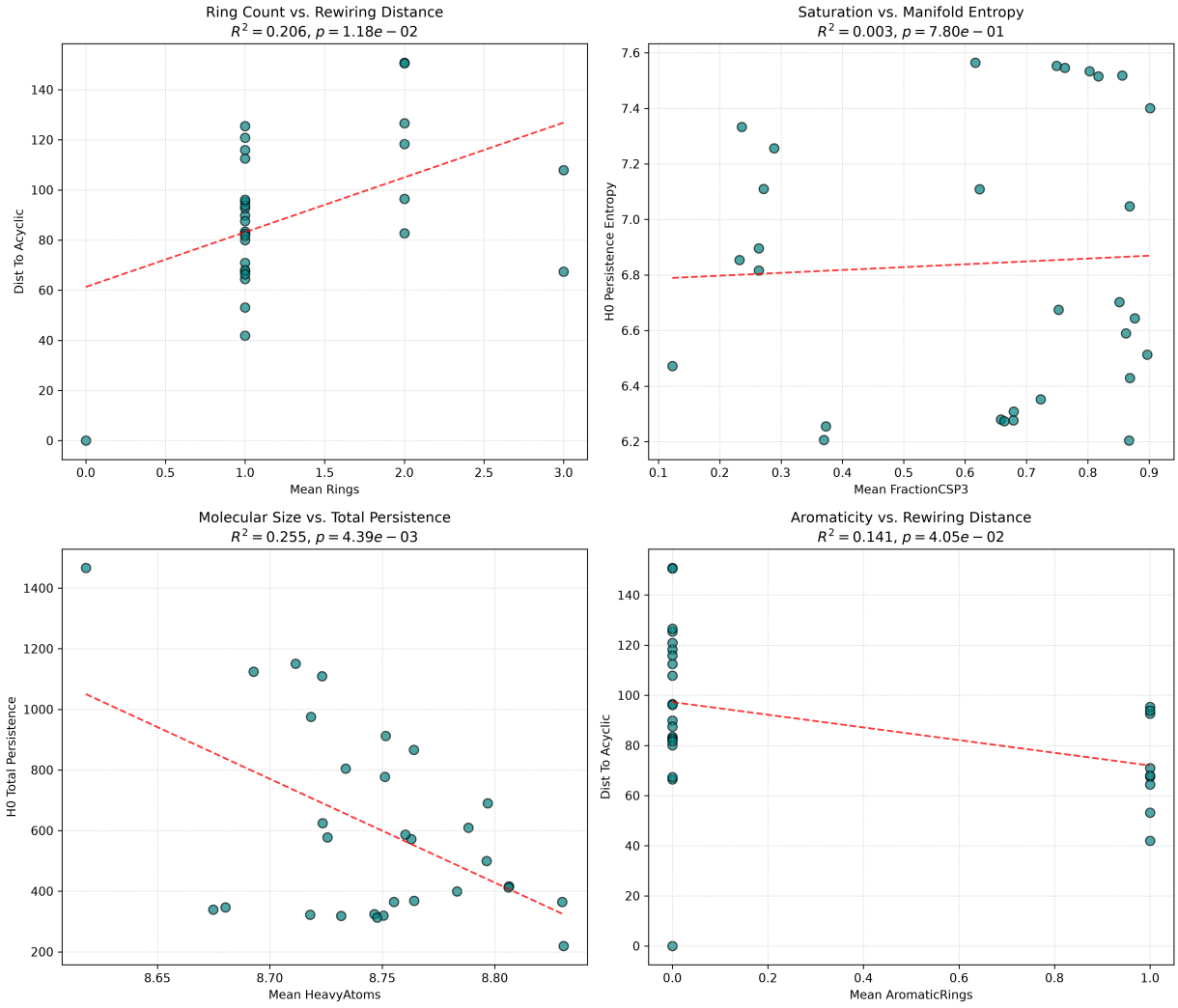

Figure 3: Scatter plots illustrating the relationships between structural descriptors and topological metrics of molecular property manifolds. The top-left panel demonstrates a positive correlation between the mean number of rings and the topological rewiring distance to the acyclic baseline, confirming that cyclization and spatial confinement drive manifold deformation. Conversely, the bottom-right panel shows that aromaticity acts as a counter-force, negatively correlating with rewiring distance because electronic delocalization standardizes the joint property structure. The bottom-left panel reveals a negative correlation between molecular size (mean heavy atoms) and H_0 total persistence, reflecting that more constrained systems exhibit less dispersed property distributions compared to the acyclic baseline. Finally, the top-right panel indicates no significant linear relationship between saturation (mean fraction of sp^3 carbons) and H_0 persistence entropy. Together, these trends demonstrate that structural rigidity and electronic delocalization are primary, opposing drivers in reorganizing the physical state space of organic molecules.

Scaffold SMILES	Structure Description	Dist. to Acyclic
C1CC2CC12	Bicyclo[1.1.1]pentane	150.82
C1=CC2CC2C1	Bicyclo[2.1.0]pent-2-ene	150.57
C1OC2CC12	Oxabicyclo[1.1.1]pentane	126.56
C1CCC1	Cyclobutane	125.42
C1=CCCC1	Cyclopentene	120.84
C1CN2CC12	Azabicyclo[1.1.1]pentane	118.33
C1CCOC1	Tetrahydrofuran	115.95
C1COC1	Oxetane	112.50
C1C2CC1C2	Bicyclobutane derivatives	107.85
C1CC2NC12	Azabicyclo[1.1.1]pentane deriv.	96.48

Table 1: Top 10 scaffolds by topological rewiring distance.

Topological Metric	Structural Descriptor	Pearson R	p-value
Dist. to Acyclic	Mean Heavy Atoms	0.715	9.0×10^{-6}
Dist. to Acyclic	Mean Heteroatoms	-0.709	1.2×10^{-5}
Dist. to Acyclic	Mean Rings	0.454	1.1×10^{-2}
Total Persistence	Mean Heavy Atoms	-0.505	4.4×10^{-3}
Total Persistence	Mean Rings	-0.480	7.2×10^{-3}

Table 2: Key correlations between topology and structure.

deformation. This is due to the restriction of conformational degrees of freedom, which forces geometric and electronic properties into narrow, non-standard couplings that differ from the diffuse acyclic state. Additionally, both the mean number of rings and molecular size (mean heavy atoms) negatively correlate with H_0 total persistence (Table 2, Figure 3), reflecting that more constrained systems exhibit less dispersed property distributions compared to the acyclic baseline.

3.3.2 Heteroatom influence

The number of heteroatoms (N, O, F) is strongly negatively correlated with the rewiring distance ($R = -0.709$, $p = 1.2 \times 10^{-5}$; Table 2). This suggests that as molecules become more heteroatom-rich, their joint property manifolds tend to converge toward a more baseline topological state. Heteroatoms likely provide electronic pathways that mitigate the topological impact of the carbon skeleton's rigidity.

3.3.3 Aromaticity and saturation

Aromaticity acts as a counter-force to rewiring (Figure 3). Aromatic scaffolds like pyridine (distance = 93.88) show lower rewiring distances than saturated counterparts like cyclobutane (distance = 125.42). This indicates that the electronic delocalization inherent in aromatic systems creates a more standardized joint property structure, whereas saturated, strained rings produce more exotic topological configurations. Furthermore, as shown in Figure 3, saturation alone (mean fraction of sp^3 carbons) does not exhibit a significant linear relationship with H_0 persistence entropy, highlighting that it is the specific combination of strain and lack of delocalization that drives the most extreme topological rewiring.

4 Conclusions

In this study, we introduced a mathematically rigorous, coordinate-free framework to quantify how molecular architecture dictates the joint structure of physical properties. By regressing out stoichiometric determinism and applying persistent homology to the resulting 12-dimensional property residuals, we successfully isolated the purely geometric and structural variance across the QM9 dataset. This approach allowed us to move beyond scalar correlations and directly measure the topological deformation of the physical state space induced by distinct Murcko scaffolds.

Our topological transport cost analysis revealed that spatial confinement fundamentally rewires the molecular property manifold. We demonstrated that rigid, highly strained systems—particularly bicyclic architectures—force geometric and electronic properties into exotic, non-standard couplings. These structural families occupy topologically distinct configurations whose transport costs relative to an acyclic baseline far exceed the bootstrapped noise floor. This confirms that cyclization restricts conformational degrees of freedom and severely deforms the property space, overriding the diffuse property distributions typically observed in flexible acyclic molecules.

Crucially, our correlation analysis uncovered a fundamental balance between geometric confinement and electronic delocalization. While increased ring count and structural rigidity drive topological divergence, the presence of heteroatoms and aromaticity acts as a strong mitigating force. Aromatic systems and heteroatom-rich scaffolds provide electronic pathways that counteract the topological impact of a rigid carbon skeleton, causing their joint property manifolds to converge back toward a more standardized, baseline state.

Ultimately, these findings highlight that the physical state space of organic molecules is shaped by a dynamic trade-off between the deforming forces of spatial confinement and the standardizing effects of electronic delocalization. By framing molecular property relationships as high-dimensional topological manifolds, this work provides a robust, outlier-resistant methodology for understanding chemical space. This perspective not only deepens our fundamental understanding of structure-property relationships but also offers a novel geometric tool for characterizing and navigating the complex property landscapes essential to targeted chemical design.