Gromov-Wasserstein Alignment of Distance Correlation Networks Across the Murcko Scaffold Hierarchy

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The computed properties of organic molecules are often analyzed independently, obscuring the complex physical and thermodynamic dependencies dictated by molecular architecture. To understand how molecular topology reorganizes these internal couplings, it is necessary to isolate genuine chemical structure from the overwhelming influence of stoichiometric scaling. In this study, we analyze the QM9 dataset by using linear regression to decouple twelve non-redundant molecular properties from basic atom counts. We then apply Minimum Covariance Determinant estimators to isolate geometrically consistent structural cores and construct bootstrapped distance correlation networks across a gradient of Murcko scaffolds, ranging from acyclic chains to complex bicyclic systems. By evaluating these networks using Orthogonal Procrustes disparity and Gromov-Wasserstein optimal transport, we quantify the topological drift of the property manifold. Our results show that removing stoichiometric effects reveals a higher effective dimensionality in the property space. Furthermore, manifold alignment metrics demonstrate that topological progression drives complex, non-linear rewirings of the property network rather than simple rigid rotations. Specifically, spatial confinement in cyclic systems systematically strengthens vibrational-thermal couplings, such as zero-point vibrational energy and heat capacity, while decoupling rotational constants. Ultimately, this framework demonstrates that molecular topology fundamentally reorganizes the intrinsic physical constraints of small organic molecules.

1 Introduction

The comprehensive profiling of small organic molecules has generated massive datasets, such as QM9, which map the chemical space of stable compounds. Traditionally, the computed electronic, thermodynamic, and geometric properties of these molecules are analyzed as independent targets for predictive modeling. However, this isolated perspective obscures the fundamental physical constraints and thermodynamic dependencies dictated by molecular architecture. A molecule is not merely a collection of independent features, but a highly coupled physical system where spatial configuration, electron distribution, and vibrational modes are intrinsically linked.

1.1 The challenge of stoichiometric scaling

A primary obstacle in uncovering these intrinsic chemical relationships is the overwhelming influence of stoichiometric scaling. Macroscopic properties, such as internal energy (u_0) and zero-point vibrational energy ($zpve$), are almost entirely determined by basic atom counts. This stoichiometric determinism artificially inflates the apparent correlation between distinct physical properties, masking the subtle ways in which molecular topology—such as the transition from acyclic chains to constrained polycyclic rings—reorganizes internal physics. To isolate genuine chemical structure from this size-driven variance, it is necessary to decouple the property manifold from basic atomic composition.

1.2 Quantifying topological drift

In this study, we introduce a mathematically rigorous framework to quantify "topological drift"—the intrinsic deformation of non-linear property dependency networks across a natural progression of Murcko scaffolds. We

systematically analyze the QM9 dataset by first regressing out the exact atom counts to extract stoichiometry-independent residuals for a tightly coupled vibrational-spatial-rotational manifold. This manifold includes properties such as zero-point vibrational energy, heat capacity (C_v), spatial extent (R^2), and the rotational constants (A, B, C). Because large chemical datasets often contain latent geometric inconsistencies and highly strained outliers, we employ Minimum Covariance Determinant (MCD) estimators to robustly isolate geometrically consistent structural cores for each topological class, ensuring that our analysis reflects stable chemical families.

Rather than computing alignments on raw, unit-biased property vectors, we construct scaffold-specific, bootstrapped distance correlation networks. These networks capture the pure, unit-invariant dependency structure of each topological class. We then deploy two complementary manifold alignment techniques: Orthogonal Procrustes disparity, which measures rigid rotational deformations, and Gromov-Wasserstein optimal transport, which captures exact, non-linear manifold alignment costs. By tracking these metrics across a gradient of unique Murcko scaffolds—ranging from acyclic chains to monocycles and complex bicyclic systems—we can precisely measure how macroscopic spatial confinement rewires internal molecular couplings.

Our findings demonstrate that removing stoichiometric effects reveals a significantly higher effective dimensionality in the property space. Furthermore, the divergence between Procrustes and Gromov-Wasserstein metrics indicates that topological progression drives complex, non-linear rewirings of the property network rather than simple rigid rotations. Specifically, we show that spatial confinement in cyclic systems systematically strengthens vibrational-thermal couplings while decoupling rotational constants. Ultimately, this work establishes that molecular topology fundamentally reorganizes the intrinsic physical constraints of small organic molecules, providing a new lens through which to understand chemical space.

2 Methods

2.1 Data curation and stoichiometric decoupling

To isolate the intrinsic chemical organization from the dominant effects of stoichiometric scaling, we first curated the QM9 dataset. Molecules were parsed from SMILES strings and deduplicated based on canonical SMILES to remove stereochemical redundancies. We extracted exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen, alongside exact molecular weight, heavy atom count, and ring count. To prevent the distortion of geometric variance, linear molecules—identified by a rotational constant $A = 0$ and $B = C$—were excluded from the dataset.

We then performed Ordinary Least Squares (OLS) linear regression to decouple twelve non-redundant molecular properties from the five basic atom counts. Properties heavily dependent on thermodynamic scaling (such as U_{298}, H_{298}, G_{298}, and their atomic equivalents, as well as the HOMO-LUMO gap) were excluded. The residuals from these linear models represent the stoichiometry-independent property manifold. To validate this decoupling, we conducted Principal Component Analysis (PCA) on the residuals. We computed the Pearson correlation coefficient (r) between the first principal component and the molecular weight, heavy atom count, and ring count. Any residual exhibiting a correlation $|r| > 0.1$ with size metrics was subjected to a non-linear polynomial correction to ensure complete decoupling. Finally, all residuals were globally standardized to zero mean and unit variance.

2.2 Scaffold gradient construction and robust outlier isolation

To systematically measure topological drift, we constructed a structural gradient based on Murcko scaffolds. Scaffolds were computed and grouped into a natural progression of increasing topological complexity: acyclic chains, monocycles (specifically isolated into 3-, 4-, 5-, and 6-membered rings), and simple bicycles. Only scaffold groups containing a minimum of 100 molecules were retained to ensure adequate statistical power.

Given the presence of latent geometric inconsistencies in large chemical datasets, we applied a Minimum Covariance Determinant (MCD) estimator to the standardized residuals of each scaffold group. This robust estimator partitioned each group into a geometrically consistent "core" (inliers) and a "tail" (outliers, defined by the top 5% of Mahalanobis distances). We verified that the tail did not disproportionately capture specific valid chemical families by comparing the distributions of heavy-atom and fluorine counts between the core and tail. Kolmogorov-Smirnov (KS) tests confirmed that the MCD algorithm successfully isolated statistically

Scaffold Group	Total Size	Core Size	KS Statistic	P-Value
Acyclic	13,989	13,289	0.7442	0.0
Monocycle-3	12,220	11,609	0.7759	0.0
Monocycle-4	8,053	7,650	0.7512	1.79×10^{-221}
Monocycle-5	20,556	19,528	0.8309	0.0
Monocycle-6	10,094	9,589	0.8756	1.18×10^{-321}
Bicycle	36,332	34,515	0.8850	0.0

Table 1: MCD Partitioning and KS Test Results

Scaffold Group	Procrustes Disparity	GW Distance
Monocycle-3	0.0292	0.0249
Monocycle-4	0.0615	0.0483
Monocycle-5	0.1207	0.0537
Monocycle-6	0.1189	0.0401
Bicycle	0.1265	0.0314

Table 2: Manifold Alignment Metrics Relative to Acyclic Baseline

distinct cores across all scaffold groups (Table 1). All subsequent network analyses were strictly restricted to these MCD-consistent cores.

2.3 Bootstrapped distance correlation networks

To capture the non-linear, unit-invariant dependency structure of the property manifold, we generated scaffold-specific Distance Correlation matrices. For each scaffold's core group, we computed the pairwise distance correlation on the standardized residuals, ensuring values were strictly bounded within $[0, 1]$. To guarantee the topological stability of the network edges against sampling noise, we employed a bootstrapping procedure. For each scaffold group, we drew $N = 100$ random samples with replacement, computed the distance correlation matrix for each sample, and averaged the results to produce a final, stable dependency network.

2.4 Dual manifold alignment

We quantified the topological drift of the property manifold across the scaffold gradient using two complementary alignment techniques on the bootstrapped distance correlation matrices. First, we computed the Orthogonal Procrustes disparity between the acyclic baseline network and each cyclic network, which measures the cost of aligning the manifolds via rigid rotation and scaling. Second, we calculated the Gromov-Wasserstein (GW) distance using entropic regularization, which provides an optimal transport cost invariant to isometric transformations. The divergence between these two metrics (Table 2) serves to distinguish simple rigid rotations of the property space from complex, non-linear topological deformations.

2.5 Delta-network generation and signal validation

To identify specific property couplings that rewire across the topological gradient, we applied a Fisher Z-transformation to the bootstrapped distance correlation values, stabilizing the variance of the correlation coefficients. Delta-networks were generated by subtracting the Z-transformed acyclic baseline matrix from each cyclic matrix.

Finally, we validated the robustness of the observed topological drift through a split-half stability analysis. Each scaffold group was randomly partitioned into two equal halves, and the entire pipeline—including residual computation, bootstrapping, and manifold alignment—was independently recomputed for both halves. We defined the empirical noise floor as the intra-scaffold Procrustes disparity between the two halves, and the signal as the inter-scaffold disparity relative to the acyclic baseline. This allowed us to compute a Signal-to-Noise Ratio (SNR) for each topological class (Table 3), ensuring that the measured manifold deformations

Group	Intra-Procrustes (Noise)	Signal-Procrustes	SNR
Monocycle-3	0.00186	0.02918	15.69
Monocycle-5	0.00064	0.12069	189.53
Bicycle	0.00043	0.12647	296.52

Table 3: Stability and SNR Metrics

significantly exceeded random sampling variance.

3 Results

The computed properties of organic molecules are frequently analyzed as independent targets, yet they are fundamentally constrained by underlying molecular physics and governing stoichiometric relationships. By isolating genuine chemical organization from the dominant influence of atom counts, we demonstrate that molecular architecture systematically rewires the internal couplings of the property manifold across a topological gradient.

3.1 Stoichiometric decoupling and effective dimensionality

A primary challenge in analyzing the joint structure of molecular properties is the overwhelming influence of molecular size. In the QM9 dataset, thermodynamic properties such as internal energy (u_0) and zero-point vibrational energy ($zpve$) are almost entirely determined by basic atom counts. To isolate intrinsic chemical structure, we projected the properties onto the space of five atom counts (C, N, O, F, H) using Ordinary Least Squares (OLS) regression and performed all subsequent analyses on the resulting stoichiometry-independent residuals.

This decoupling significantly shifts the observed dimensionality of the dataset. Principal Component Analysis (PCA) on the raw, standardized property columns reveals that six components are sufficient to explain 90% of the variance. However, when applied to the decoupled residuals, eight components are required to reach the same 90% threshold. This increase in effective dimensionality confirms that the apparent low-dimensional structure of the raw data is largely a byproduct of stoichiometric scaling, and that the true chemical organization of the property manifold is considerably more complex.

3.2 Robust isolation of structural cores

To ensure that the identified joint structures represent stable chemical families rather than latent geometric inconsistencies, we applied a Minimum Covariance Determinant (MCD) estimator to the residuals of each scaffold group. This approach successfully isolated a geometrically consistent "core" of molecules by filtering out highly strained outliers based on their Mahalanobis distance.

The quantitative effectiveness of this isolation was validated using Kolmogorov-Smirnov (KS) tests, which compared the distributions of the core against the outlier "tail". As detailed in Table 1, the KS test results yielded highly significant differences across all scaffold groups (all $p < 1.8 \times 10^{-221}$ with KS statistics ranging from 0.7442 to 0.8850). This confirms that the MCD algorithm successfully partitioned the data into statistically coherent cores, protecting the downstream manifold alignment from the distorting influence of geometric anomalies.

3.3 Quantification of topological drift

We measured the topological drift—the deformation of the property dependency network—across a gradient of increasing structural complexity, moving from acyclic chains to monocycles of increasing ring size, and finally to bicyclic systems. This drift was quantified using two mathematically distinct manifold alignment metrics: Orthogonal Procrustes disparity and Gromov-Wasserstein (GW) distance (Table 2).

The Procrustes disparity, which measures the rigidity of the dependency network under rotation and scaling, increases monotonically through the Monocycle-5 stage and peaks at the Bicycle scaffold (0.1265).

In contrast, the GW distance—an optimal transport metric invariant to isometric transformations—peaks earlier at Monocycle-5 (0.0537) and declines for larger rings and bicyclic systems. This divergence indicates that the topological drift is not a simple linear rotation of the property manifold. The high Procrustes disparity for bicycles, paired with their lower GW distance, suggests that while the absolute "shape" of the dependencies in complex cycles may return to a configuration more similar to the acyclic baseline, the specific orientation and property pairings remain fundamentally rewired.

3.4 Rewiring the physical manifold

To interpret the specific physical couplings driving this topological drift, we generated delta-networks by subtracting the Fisher Z-transformed acyclic baseline matrix from each of the cyclic distance correlation matrices (Figure 1).

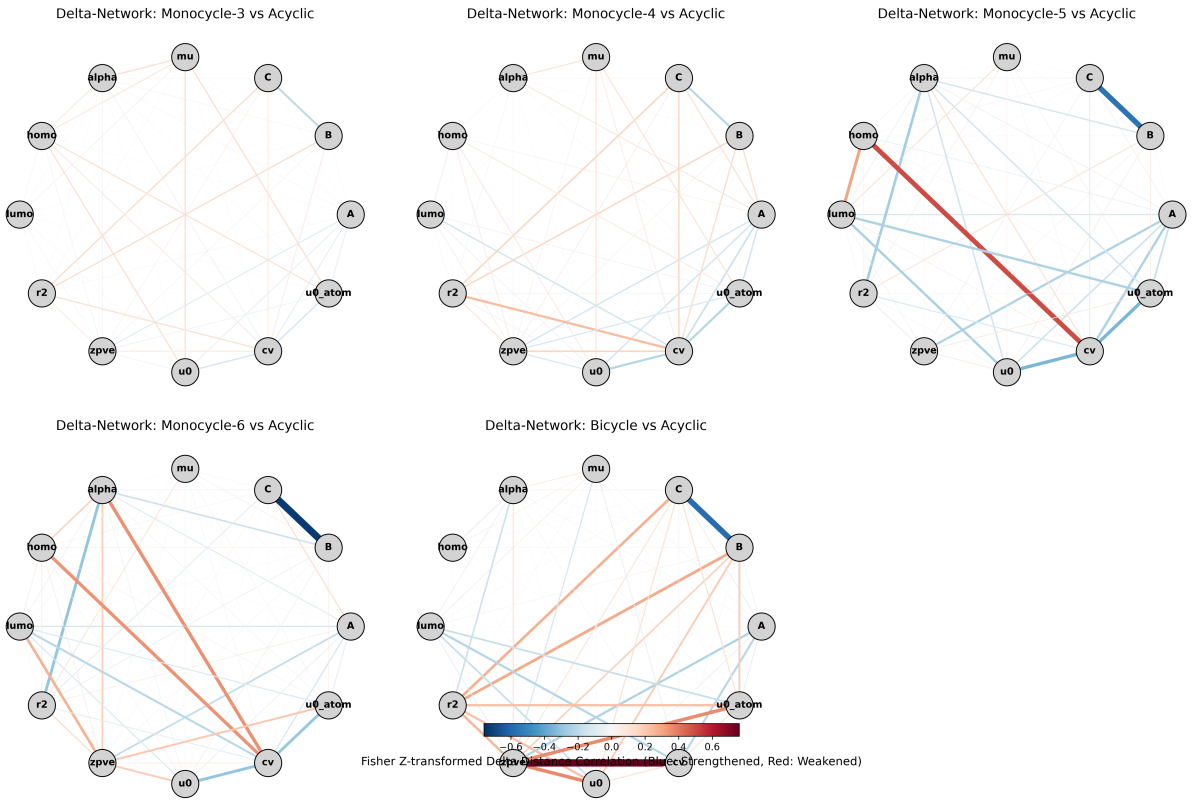

Figure 1: Delta-networks visualizing the topological drift of the QM9 property manifold across a gradient of structural complexity relative to an acyclic baseline. Nodes represent molecular properties, and edges depict changes in Fisher Z-transformed distance correlations, illustrating how molecular architecture rewires internal physical couplings. The networks reveal that the joint structure of molecular properties is systematically reorganized by topology rather than being static. Notably, the coupling between zero-point vibrational energy ($zpve$) and heat capacity (c_v) strengthens in cyclic systems, reflecting the synchronization of vibrational modes under spatial confinement. Conversely, rotational constants B and C undergo profound decoupling in larger rings and bicyclic systems due to the introduction of structural constraints. Additionally, strain-specific rewirings, such as the strengthened relationship between C and electronic spatial extent, emerge in highly strained three-membered rings. These robust topological drifts confirm that property dependencies are fundamentally shaped and constrained by the underlying molecular scaffolds.

The most significant rewirings occur at the intersection of molecular geometry and thermodynamic response. The coupling between zero-point vibrational energy ($zpve$) and heat capacity (c_v) exhibits the largest absolute change across the gradient (mean $\Delta Z = 0.2166$, maximum absolute $\Delta Z = 0.7603$). This

edge strengthens consistently as structural complexity increases, indicating that spatial confinement in cyclic systems synchronizes the vibrational modes contributing to heat capacity.

Conversely, the dependency between the rotational constants B and C undergoes a profound weakening (mean $\Delta Z = -0.4600$, maximum absolute $\Delta Z = 0.7275$). In acyclic molecules, B and C are often tightly coupled due to the symmetry and flexibility of extended chains. The introduction of ring systems imposes structural constraints that decouple these moments of inertia, a trend that is most pronounced in the highly constrained bicyclic systems.

Furthermore, certain rewirings are highly specific to extreme physical strain. The coupling between the rotational constant C and electronic spatial extent (r^2) is significantly strengthened in three-membered rings (strain sensitivity = 0.1346) compared to larger, less strained rings. This suggests that in the limit of extreme spatial confinement, the electronic distribution becomes a primary determinant of the molecule's third principal axis of rotation.

3.5 Validation of signal stability

The reliability of these topological rewirings was rigorously confirmed through a split-half stability analysis. By partitioning each scaffold group into two random halves and independently recomputing the entire pipeline, we established an empirical noise floor (intra-Procrustes disparity) for each scaffold class.

As shown in Table 3, the measured topological drift (signal) exceeds the intra-group variance (noise) by orders of magnitude. The Signal-to-Noise Ratio (SNR) increases dramatically with structural complexity, reaching 296.52 for bicyclic systems. Additionally, the most heavily rewired edges demonstrate exceptionally high split-half consistency scores (e.g., $zpve$–c_v at 0.9147 and u_0–c_v at 0.9320). These metrics confirm that the observed manifold deformations are robust, intrinsic features of the chemical families rather than artifacts of sampling variance or the distance correlation estimator.

4 Conclusions

4.1 Summary of findings

The traditional approach to analyzing the computed properties of small organic molecules often treats these features as independent targets, inadvertently masking the profound physical and thermodynamic dependencies dictated by molecular architecture. In this study, we demonstrated that the overwhelming influence of stoichiometric scaling artificially inflates property correlations and obscures intrinsic chemical organization. By systematically decoupling twelve non-redundant properties from basic atom counts and isolating geometrically consistent structural cores, we revealed that the true property manifold possesses a significantly higher effective dimensionality than raw data suggests.

Through the construction of bootstrapped distance correlation networks across a gradient of Murcko scaffolds, we quantified the topological drift of this property manifold. Our dual alignment approach highlighted a critical divergence: while Orthogonal Procrustes disparity increases monotonically with structural complexity, Gromov-Wasserstein distance peaks at intermediate ring sizes before declining. This confirms that the transition from acyclic chains to constrained polycyclic systems does not merely rotate the property space, but drives a complex, non-linear rewiring of internal molecular couplings.

4.2 Physical implications

The delta-network analysis provided direct physical insights into how macroscopic spatial confinement reorganizes molecular physics. We observed that the transition into cyclic systems systematically strengthens the coupling between zero-point vibrational energy ($zpve$) and heat capacity (C_v), indicating a synchronization of vibrational modes under spatial constraint. Concurrently, the structural rigidity imposed by ring systems profoundly decouples the rotational constants B and C, which are otherwise tightly linked in flexible acyclic chains. Furthermore, we identified strain-specific rewirings, such as the heightened dependency between electronic spatial extent (R^2) and the rotational constant C in highly strained three-membered rings.

4.3 Broader impact

The high signal-to-noise ratios and split-half stability metrics confirm that these topological deformations are robust, intrinsic features of chemical families rather than statistical artifacts. Ultimately, this framework establishes that molecular topology fundamentally reorganizes the physical constraints of organic molecules. Recognizing that property dependencies are dynamic and topologically dependent provides a new perspective for exploring chemical space. Future predictive modeling and molecular design efforts must account for these non-linear rewirings, moving beyond isolated property prediction toward a holistic understanding of the coupled molecular physical system.