Topological Rewiring of Molecular Property Networks Across the QM9 Chemical Space

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how molecular topology and spatial confinement influence interdependent physical properties is challenging because dominant stoichiometric size-scaling often masks intrinsic structural effects. To isolate genuine structural dependencies, we analyze the QM9 dataset by extracting atom-count residuals for twelve electronic, thermodynamic, and geometric properties. We then partition this chemical space into statistically robust topological manifold bins based on molecular complexity indices and ring counts. Within each bin, we construct property dependency networks using the Hilbert-Schmidt Independence Criterion and quantify structural rewiring across complexity regimes using Gromov-Wasserstein optimal transport relative to an acyclic baseline. Our results show that increasing molecular complexity systematically deforms the joint property manifold, with specific couplings between spatial, thermal, and rotational properties either strengthening or weakening depending on the topological regime. These findings demonstrate that fundamental relationships between thermodynamic and geometric properties dynamically reorganize with molecular topology, highlighting the need to account for topological regime shifts in multi-target molecular modeling.

1 Introduction

1.1 Motivation and background

Understanding the fundamental relationships between the physical, thermodynamic, and electronic properties of molecules is a central goal in computational chemistry and molecular design. In multi-target molecular modeling, properties are rarely independent; they form a deeply interconnected joint physical manifold. For instance, the spatial extent of a molecule's electron cloud is naturally coupled to its rotational constants and vibrational modes. However, characterizing how molecular topology and spatial confinement influence these interdependent properties remains highly challenging.

A primary obstacle in isolating these intrinsic structural effects is the dominant confounding role of stoichiometry. Simple atom counts dictate the vast majority of the variance for extensive properties such as internal energy, zero-point vibrational energy, and isotropic polarizability. This stoichiometric size-scaling often masks the more subtle, genuine effects of molecular topology, such as ring closures and spatial rigidification. Consequently, it is often implicitly assumed that the residual correlations between physical properties remain static across different classes of molecules. Yet, it is physically plausible that the relationship between spatial extent and heat capacity in a flexible, acyclic alkane differs fundamentally from that in a highly constrained, polycyclic framework.

1.2 Methodological approach

To isolate and quantify this topological rewiring, we present a framework capable of comparing complex property dependency networks across different structural regimes. We analyze the QM9 dataset, explicitly excluding linear molecules to prevent extreme rotational variance from distorting structural signals. To decouple stoichiometry from structure, we extract atom-count residuals for twelve non-redundant electronic, thermodynamic, and geometric properties. This residualization ensures that the observed correlations reflect how atoms are arranged rather than merely which atoms are present.

Property Pair	Baseline Distance	Bin 26 Distance	Δ Distance (Shift)
res_r^2 vs res_C_v	1.4007	1.2238	-0.1769
res_C vs res_r^2	1.0236	1.1863	+0.1628
res_B vs res_r^2	1.0345	1.1930	+0.1585

We then partition this deduplicated chemical space into statistically robust topological manifold bins based on molecular complexity indices (such as the Bertz complexity index) and exact ring counts. Within each topological bin, we construct property dependency networks using the Hilbert-Schmidt Independence Criterion (HSIC) to capture the intrinsic joint topology of the property space.

To rigorously quantify the structural rewiring across complexity regimes, we deploy Gromov-Wasserstein (GW) optimal transport. By calculating the GW transport costs between the dependency network of each complex structural bin and an acyclic baseline (molecules with zero rings), we directly measure the mathematical deformation of the property manifold as spatial confinement increases.

1.3 Summary of contributions

Our findings demonstrate that the fundamental relationships between thermodynamic and geometric properties are not static, but dynamically reorganize with molecular topology. We show that increasing molecular complexity systematically deforms the joint property manifold.

Specifically, the transition from acyclic to polycyclic systems is characterized by a strengthening of spatial-thermal couplings and a concomitant weakening of rotational-spatial dependencies. Table 1.3 highlights these specific topological shifts, detailing the most significant changes in the HSIC-derived distance metric between the acyclic baseline and the most highly deformed structural regime (Bin 26). A negative shift in distance indicates a strengthening of the dependency, whereas a positive shift indicates a decoupling.

By providing a statistically validated, stoichiometry-independent mapping of this network deformation, this work highlights the critical need to account for topological regime shifts in multi-target molecular modeling. Relationships learned in one region of the chemical manifold may fundamentally break down or reverse in another, driven by the macroscopic spatial confinement of the molecular scaffold.

2 Methods

2.1 Data preprocessing and feature isolation

The QM9 dataset was initially deduplicated based on SMILES strings, retaining only the first occurrence to prevent data leakage. Linear molecules were explicitly filtered out by removing records where the rotational constant A was near zero ($< 10^{-3}$) or where rotational constants B and C were numerically equivalent. This exclusion is critical, as the extreme rotational variance of linear molecules would otherwise distort the structural signals of interest. Twelve non-redundant physical properties were isolated for analysis: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), HOMO and LUMO energies, electronic spatial extent (R^2), zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity at 298.15 K (C_v), and atomization energy $(U_{0,\text{atom}})$. To mitigate the heavy-tailed distribution of the rotational constant A, a log-transformation was applied prior to downstream modeling.

2.2 Stoichiometric residualization

To isolate the variance driven purely by molecular structure and topology, we established a linear stoichiometric baseline. For each molecule, the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were extracted. A Ridge regression model was trained for each of the twelve isolated properties using only these five atom counts as features. The residuals from these models were extracted for all subsequent analyses. By operating exclusively on these residuals, we effectively remove the dominant stoichiometric size-scaling effects, ensuring that the topological analysis captures genuine structural dependencies rather than compositional artifacts.

2.3 Topological manifold embedding and binning

To partition the chemical space into statistically robust structural regimes, three descriptors were computed for all molecules: the Bertz complexity index, the fraction of sp^3 hybridized carbons, and the Hall-Kier α shape index. These descriptors were standardized and projected into a lower-dimensional topological manifold using Uniform Manifold Approximation and Projection (UMAP).

We applied k-means clustering on the UMAP embeddings to partition the dataset into distinct structural bins, tuning k to ensure every bin contained a minimum of 1,000 molecules. An acyclic baseline subset was defined, consisting exclusively of molecules with zero rings, to serve as a topological ground state. To ensure this baseline was representative and to control for residual size effects, we generated a stratified sample of the acyclic population that matched the heavy-atom count distribution of the overall dataset. Multiple bootstrap iterations of this stratified acyclic baseline were drawn to compute confidence intervals for the optimal transport metrics.

2.4 Kernel-based dependency networks

Within each structural bin and across the acyclic baseline bootstrap iterations, the extracted property residuals were independently standardized (z-scored). To track shifts in multivariate binding, the Total Correlation was approximated by the sum of pairwise dependencies within each bin. Furthermore, the R^2 of the stoichiometric Ridge models was calculated within each bin to monitor for potential decoupling paradoxes, where models might explain away more signal in complex molecules.

To construct the property dependency networks, we computed the normalized Hilbert-Schmidt Independence Criterion (HSIC) matrix (12×12) for the standardized residuals. A Gaussian Radial Basis Function (RBF) kernel was employed, utilizing a global median heuristic derived from a pooled sample of the dataset to set the bandwidth σ. This ensures comparability across different bins. The HSIC values were consistently normalized by the product of the standard deviations of the kernels so that the resulting values reflect genuine dependency structures rather than local scale differences. Finally, the normalized HSIC similarity matrix, K, was converted into a distance matrix, D, suitable for optimal transport using the transformation:

$$D_{ij} = \sqrt{K_{ii} + K_{jj} - 2K_{ij}} \tag{1}$$

2.5 Gromov-Wasserstein optimal transport

To quantify the mathematical deformation of the property dependency networks as spatial confinement increases, we calculated the Gromov-Wasserstein (GW) transport cost between the HSIC-derived distance matrices of the acyclic baseline iterations and each complex structural bin. This was implemented using entropic regularization. The GW ground metric was symmetrically weighted by the inverse variance of the raw residuals of the involved properties, and the resulting GW distance was normalized by the average internal distance of the matrices.

To rigorously validate that the observed network rewiring reflects genuine chemical organization rather than stoichiometry-driven variance fluctuations, the binning and GW distance calculations were repeated strictly within well-populated strata of constant heavy-atom counts to maintain statistical power while guaranteeing stoichiometry independence.

2.6 Differential network analysis

Rather than collapsing the structural rewiring to a single GW distance scalar, a differential analysis of the HSIC matrices was performed to extract the physical meaning of the deformations. The structural bin exhibiting the highest mean GW distance from the acyclic baseline was identified. We then computed the element-wise difference (ΔHSIC) between the HSIC matrix of this highly deformed bin and the mean HSIC matrix of the acyclic baseline. This allowed us to isolate the specific property pairs with the largest absolute changes in dependency, explicitly recording the sign of the change to distinguish between physical couplings that strengthen versus those that weaken under topological confinement.

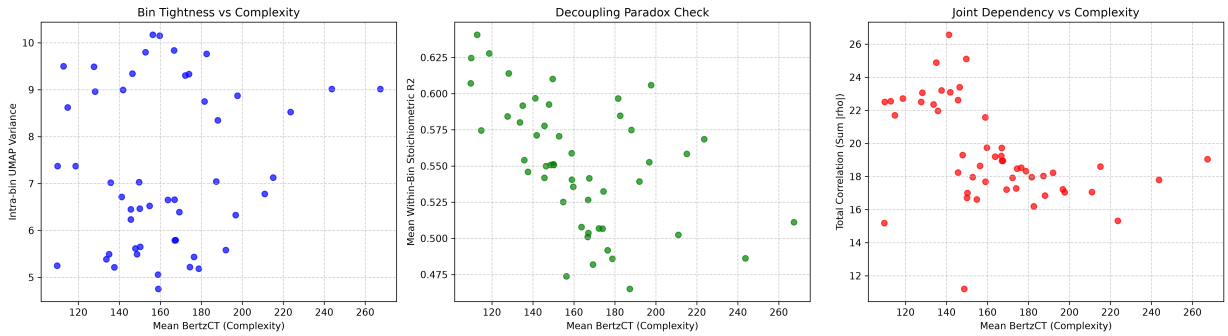

Figure 1: Characteristics of the 50 topological manifold bins as a function of molecular complexity, quantified by the mean Bertz complexity index (`BertzCT`). The left panel displays the intra-bin UMAP variance, illustrating the structural tightness of each cluster. The middle panel reveals that the mean within-bin stoichiometric R^2 decreases as complexity increases, demonstrating that simple atom counts lose their predictive dominance in more complex, ring-dense molecules. The right panel shows that the total joint dependency, measured by the sum of absolute correlations ($\sum |\rho|$), similarly diminishes with rising complexity. Together, these trends confirm that increasing structural confinement and topological rewiring fundamentally reorganize the physical state space, breaking down the simple property dependencies observed in the acyclic baseline.

3 Results

3.1 Global joint structure of property residuals

Following stoichiometric residualization, we isolated the variance driven primarily by molecular structure. Analysis of the 12×12 normalized Hilbert-Schmidt Independence Criterion (HSIC) matrix across the dataset reveals three primary axes of joint structure that persist independently of molecular size.

First, a vibrational-spatial axis emerges, characterized by a tight coupling between the residual zero-point vibrational energy (res_zpve) and heat capacity (res_C_v) with an HSIC of 0.517. This axis also links to the residual electronic spatial extent (res_r^2, HSIC = 0.164). Second, a rotational-spatial axis is defined by strong dependencies between the residual rotational constants B and C (HSIC = 0.558) and their joint relationship with the spatial extent (res_r^2 vs res_C, HSIC = 0.628; res_r^2 vs res_B, HSIC = 0.525). Finally, a thermodynamic identity block is observed, where residual internal energy (res_u_0) and atomization energy ($res_u_{0,\mathrm{atom}}$) remain almost perfectly correlated (HSIC = 0.947), confirming they behave as a single structural block. Notably, purely electronic properties, such as the residual HOMO energy (res_homo), remain largely decoupled from this global organization, exhibiting HSIC values below 0.05 with almost all thermodynamic and spatial properties.

3.2 Topological manifold binning and artifact diagnostics

Before analyzing the optimal transport costs, we evaluated the structural integrity of the 50 topological manifold bins. Figure 1 illustrates the bin characteristics as a function of mean Bertz complexity index (`BertzCT`).

As shown in Figure 1, the intra-bin UMAP variance confirms the structural tightness of the clusters. Crucially, the mean within-bin stoichiometric R^2 decreases as complexity increases. This indicates that simple atom counts lose their predictive dominance in more complex, ring-dense molecules, supporting our residualization approach. Furthermore, the total joint dependency diminishes with rising complexity, indicating that the physical state space undergoes a fundamental reorganization as molecules become more topologically constrained.

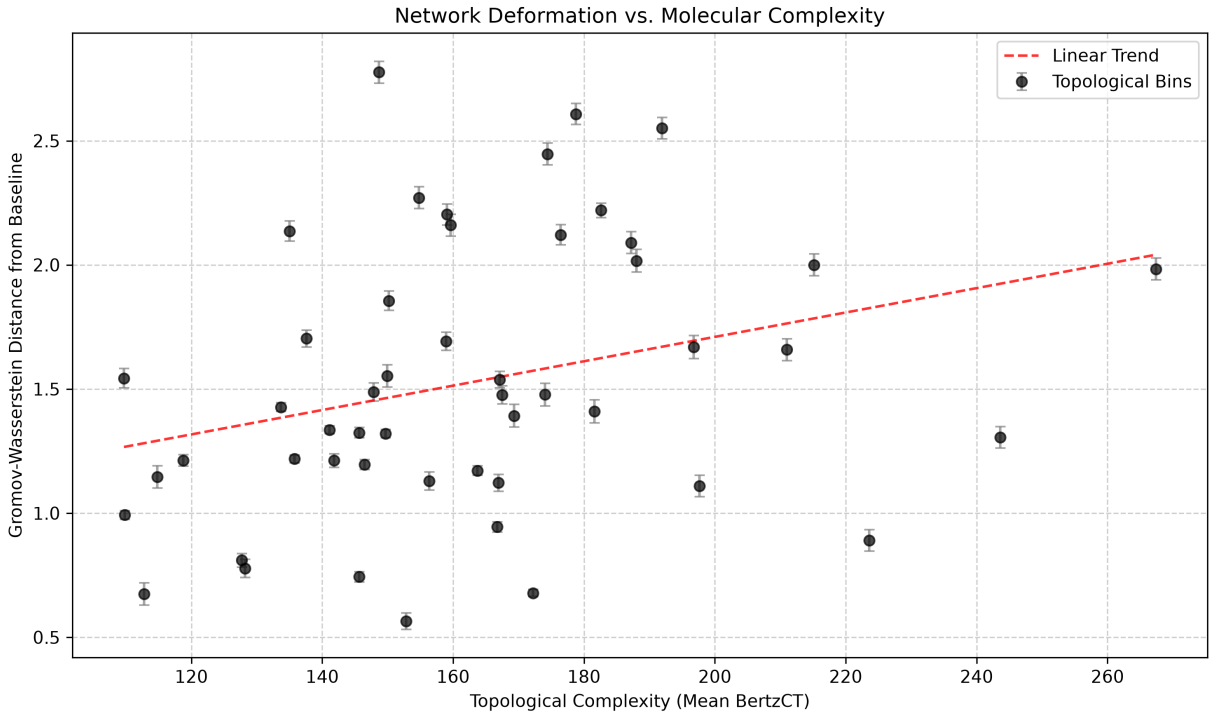

Figure 2: Gromov-Wasserstein (GW) distance from an acyclic baseline as a function of topological complexity (mean `BertzCT`) across 50 molecular bins. The positive linear trend (Pearson $r = 0.2929$) illustrates that network deformation systematically increases with molecular complexity. The maximum deformation reaches a GW distance of 2.776, corresponding to highly complex cyclic systems where the acyclic baseline's property dependencies fundamentally break down. This demonstrates that increasing structural confinement and ring density significantly rewire the physical state space of the molecules, specifically altering the couplings between electronic spatial extent (r^2), rotational constants (B, C), and heat capacity (C_v).

3.3 Network deformation across topological complexity

To quantify this reorganization, we computed the Gromov-Wasserstein (GW) optimal transport distances between the acyclic baseline and each of the 50 structural bins. As shown in Figure 2, network deformation systematically increases with molecular complexity, exhibiting a positive linear trend (Pearson $r = 0.2929$).

The global mean GW distance across all bins is 1.527. The maximum deformation is observed in Cluster 26 (mean `BertzCT` = 148.65), which reaches a GW distance of 2.776 ± 0.044. This peak indicates a substantial deviation from the acyclic baseline's property dependencies. Similarly, the most complex bin, Cluster 2 (mean `BertzCT` = 191.93), exhibits a high deformation (GW = 2.551), suggesting that the coupling of molecular properties becomes increasingly complex as ring density and spatial confinement increase. The minimum deformation is found in Cluster 33 (GW = 0.565), which structurally remains closer to the acyclic baseline.

3.4 Physical interpretation of network shifts

To extract the physical meaning of this topological rewiring, we performed a differential analysis of the HSIC-derived distance matrices between the acyclic baseline and the maximum-deformation bin (Cluster 26). Figure 3 visualizes these distance matrices and their element-wise difference (Δ distance).

Table 1.3 details the top property pairs exhibiting the largest absolute shifts in the HSIC-derived distance metric.

The most significant shift is the strengthening of the spatial-thermal coupling. The distance between the electronic spatial extent (res_r^2) and heat capacity (res_C_v) decreases by 0.1769. Since a decrease in distance in the HSIC-derived metric space corresponds to an increase in dependency, this indicates that in

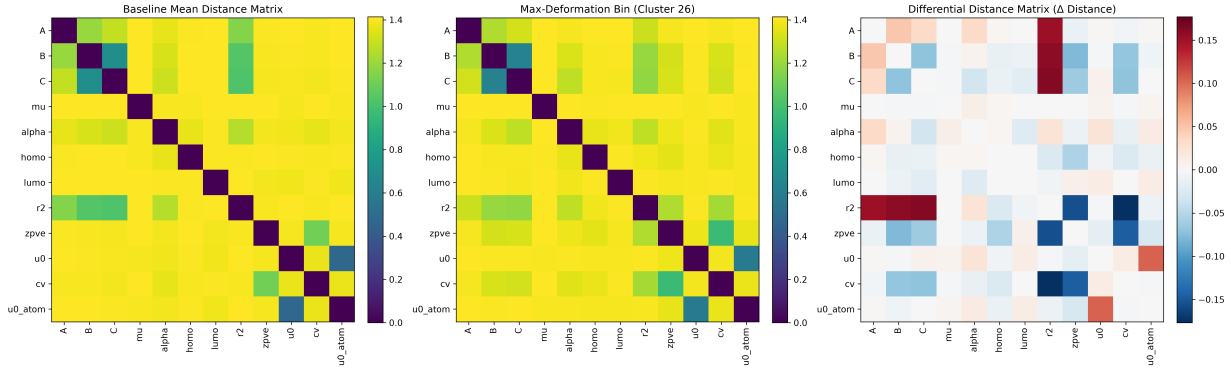

Figure 3: Distance matrices comparing the acyclic baseline (left) to the maximum-deformation bin, Cluster 26 (center), alongside their differential distance matrix (right). The differential matrix illustrates the structural rewiring of molecular property networks as topological complexity increases. A negative shift between the electronic spatial extent (r^2) and heat capacity (C_v) indicates a strengthened spatial-thermal coupling, suggesting that ring closures synchronize the electron cloud's spatial distribution with available vibrational modes. Conversely, positive shifts between r^2 and the rotational constants (B and C) reveal a weakened rotational-spatial coupling, demonstrating that internal structural constraints in complex cyclic systems uncouple the global electronic extent from the specific mass distribution.

highly complex, cyclic molecules, r^2 becomes more tightly coupled to C_v than in acyclic molecules. Physically, this suggests that the rigidification of the molecular scaffold through ring closures synchronizes the spatial distribution of the electron cloud with the available vibrational modes.

Conversely, we observe a pronounced weakening of the rotational-spatial coupling. The distances between res_r^2 and the rotational constants (res_B and res_C) increase by 0.1585 and 0.1628, respectively. In acyclic molecules, r^2 serves as a dominant predictor of the moments of inertia. However, in complex cyclic systems, internal structural constraints uncouple the global electronic extent from the specific mass distribution that dictates rotational constants. This represents a deviation from the simple geometric-electronic relationships found in more flexible, linear-like structures, underscoring the importance of accounting for topological regime shifts in multi-target molecular modeling.

4 Conclusions

The joint structure of molecular properties within the QM9 chemical space is not a static entity, but a dynamic manifold that fundamentally reorganizes with molecular topology. By rigorously decoupling dominant stoichiometric size-scaling from intrinsic structural variance, we have demonstrated that the underlying dependencies between thermodynamic, geometric, and electronic properties are systematically rewired by structural complexity.

Using Gromov-Wasserstein optimal transport to quantify this deformation, our analysis reveals that the transition from flexible, acyclic structures to highly constrained, polycyclic systems induces significant topological regime shifts. As highlighted in Table 1.3, the most profound alterations occur in the couplings between electronic spatial extent (res_r^2), heat capacity (res_C_v), and rotational constants (res_B and res_C).

Physically, the rigidification of the molecular scaffold through ring closures strengthens the spatial-thermal coupling, synchronizing the spatial distribution of the electron cloud with available vibrational modes. Concurrently, this spatial confinement uncouples the global electronic extent from the specific mass distribution, leading to a pronounced weakening of the rotational-spatial dependencies that typically dominate acyclic molecules.

These findings have critical implications for computational chemistry and molecular design. They establish that the fundamental "glue" of molecular physics is highly context-dependent; a property relationship learned in one region of the chemical manifold may break down or entirely reverse in another. Consequently,

future multi-target prediction models and generative frameworks must explicitly account for these topological regime shifts to accurately capture the coupled physical state space of complex molecules.