Gromov-Wasserstein Alignment of Distance Correlation Networks Across Macroscopic Topological Regimes

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how spatial confinement and structural complexity reorganize the joint relationships between molecular properties remains challenging, as intrinsic topological effects are often masked by extensive compositional scaling. To isolate chemical architecture from stoichiometric determinism, we analyze the QM9 dataset by extracting gradient-boosted atom-count residuals for twelve non-redundant physical properties. We then stratify the dataset into continuous topological regimes based on ring density and sp^3 hybridization. Within each regime, we construct partial distance correlation networks to capture unit-invariant joint structures and deploy Gromov-Wasserstein optimal transport to calculate exact manifold alignment costs against an acyclic baseline. Our results demonstrate that the relational manifold of molecular properties progressively diverges from the flexible chain baseline, peaking at a critical transition point in structural complexity where the network undergoes significant reorganization. Rather than merely scaling, the fundamental connections between properties rewire across regimes. Node-level sensitivity analysis identifies the rotational constant C, atomization energy, and LUMO energy as the primary drivers of this topological rewiring, while the dipole moment remains relationally invariant. Finally, scaffold-based split-half validation confirms that these network deformations are statistically robust across distinct molecular families, establishing that the joint structure of molecular properties is a dynamic manifold that deforms systematically with molecular topology.

1 Introduction

Understanding the intricate relationships between the physical, electronic, and thermodynamic properties of molecules is a central objective in chemical physics and molecular design. Traditionally, molecular properties are analyzed either as independent targets for predictive modeling or through pairwise linear correlations. However, these approaches often fail to capture the complex, joint structure of the molecular property space. A major confounding factor in this analysis is stoichiometric determinism: extensive compositional scaling dominates the variance of many physical observables, masking the intrinsic topological effects that arise from chemical architecture. Consequently, it remains challenging to determine how spatial confinement and structural complexity fundamentally reorganize the joint relationships between molecular properties.

1.1 The challenge of compositional scaling and topological complexity

As molecules transition from flexible, acyclic chains to highly constrained, ring-dense structures, their internal thermodynamic and geometric state spaces are altered. Spatial confinement restricts conformational flexibility, which in turn should theoretically couple or decouple specific physical observables. Yet, observing this phenomenon directly is difficult because the addition of heavy atoms simultaneously drives up extensive properties such as internal energy, zero-point vibrational energy, and polarizability.

To rigorously investigate whether the fundamental connections between properties merely scale with size or fundamentally rewire due to structural constraints, it is necessary to isolate chemical topology from stoichiometry. Furthermore, standard correlation metrics are often insufficient for capturing these relationships, as they are sensitive to the units of measurement and often assume linear dependencies. A framework is needed that can capture unit-invariant, non-linear joint structures and track their evolution across macroscopic topological regimes.

1.2 A relational manifold approach

In this work, we propose a mathematically rigorous framework to quantify how spatial confinement reorganizes molecular physics. We analyze the QM9 dataset by first extracting gradient-boosted atom-count residuals for twelve non-redundant physical properties. This stoichiometric decoupling effectively isolates the intensive signatures of chemical structure from extensive compositional scaling. We then stratify the deduplicated dataset into continuous topological regimes based on a composite metric of ring density and sp^3 hybridization.

Within each topological regime, we construct partial distance correlation networks. Distance correlation provides a unit-invariant measure of dependence that captures both linear and non-linear associations, while partialization isolates the direct relationships between property residuals by regressing out the influence of the remaining network. To measure the deformation of these networks across the structural complexity gradient, we deploy Gromov-Wasserstein (GW) optimal transport. By calculating the exact manifold alignment costs between the distance correlation matrix of each regime and a strictly acyclic baseline, we obtain a stoichiometry-independent measurement of network rewiring.

1.3 Overview of findings

Our analysis reveals that the relational manifold of molecular properties is not static; rather, it progressively diverges from the flexible chain baseline as topological complexity increases. We identify a critical transition point in structural complexity where the property network undergoes its most significant reorganization, demonstrating that properties fundamentally change their connectivity to other physical observables in constrained frameworks.

Through node-level sensitivity analysis, we isolate the specific physical properties driving this topological rewiring. The rotational constant C, atomization energy ($u_{0,atom}$), and LUMO energy emerge as the primary drivers of network deformation, acting as relational hinges across regimes, whereas the dipole moment (μ) remains relationally invariant. Finally, using scaffold-based split-half validation, we confirm that these network deformations are statistically robust across distinct molecular families. This establishes that the joint structure of molecular properties is a dynamic manifold that deforms systematically with molecular topology.

2 Methods

2.1 Data curation and stoichiometric decoupling

To isolate the intrinsic topological effects of chemical architecture from extensive compositional scaling, we utilized the QM9 dataset, deduplicating the entries based on SMILES strings to ensure each unique molecular structure was represented exactly once. We focused on twelve non-redundant physical and electronic properties: the rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), HOMO and LUMO energies, spatial extent ($\langle R^2 \rangle$), zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity (C_v), and atomization energy ($U_{0,\text{atom}}$). Redundant thermodynamic columns were excluded to prevent rank-deficient covariance structures. The rotational constant A was log-transformed to mitigate extreme skewness. Nine strictly linear molecules (characterized by $A = 0$) were isolated and excluded from the continuous manifold analysis, as their undefined rotational axes render them topological outliers.

To decouple the properties from stoichiometric determinism, we fit a gradient boosting regressor to each of the twelve properties using only the exact atom counts of C, N, O, F, and H as independent variables. The residuals from these models—representing the variance in each property that cannot be explained by simple atomic composition—were extracted and standardized (z-scored). All subsequent network constructions and manifold alignments were performed on these standardized residuals.

2.2 Topological stratification

Molecular graphs were parsed from SMILES strings with explicit hydrogens added. For each molecule, we calculated two continuous metrics of topological complexity: the fraction of sp^3 hybridized carbons and

the ring density (defined as the number of rings divided by the number of heavy atoms). The dataset was sorted along a composite gradient of these metrics and partitioned into 12 overlapping sliding windows, ensuring each window contained a statistically robust number of samples. The initial window (Window 0), which consists strictly of acyclic molecules with the highest sp^3 fraction, was designated as the flexible chain reference baseline for all subsequent manifold alignments.

2.3 Partial distance correlation networks

Within each topological window, we constructed a partial distance correlation network to capture the unit-invariant, non-linear joint structures between the twelve property residuals. To isolate the direct relationship between any given pair of properties, we employed a regularized linear model (Ridge regression) to regress out the influence of the remaining ten properties.

The standard distance correlation (dCor) was then computed on the resulting second-order residuals. Distance correlation is uniquely suited for this task as it equals zero if and only if the variables are statistically independent, capturing complex dependencies that standard Pearson correlation might miss. The resulting 12×12 correlation matrices were converted into dissimilarity matrices D by taking:

$$D_{ij} = 1 - \mathrm{dCor}_{\mathrm{partial}}(i, j) \tag{1}$$

The diagonal elements were strictly set to zero. This procedure yielded a fully connected, weighted graph for each topological regime, where edge weights represent the isolated relational distance between physical properties.

2.4 Gromov-Wasserstein manifold alignment

To quantify the deformation of the property networks across the structural complexity gradient, we deployed Gromov-Wasserstein (GW) optimal transport. The GW distance provides a mathematically rigorous method to compare metric measure spaces (in this case, the dissimilarity matrices of different topological windows) by finding an optimal coupling that minimizes the discrepancy between intra-network distances.

For each topological window, we computed the GW distance between its dissimilarity matrix $D^{(w)}$ and the acyclic baseline matrix $D^{(0)}$. The alignment cost is defined by the optimization problem:

$$\mathrm{GW}(D^{(w)}, D^{(0)}) = \min_{T \in \Pi} \sum_{i,j,k,l} \left| D^{(w)}_{i,j} - D^{(0)}_{k,l} \right|^2 T_{i,k} T_{j,l} \tag{2}$$

where T is the optimal transport coupling matrix and Π is the set of all valid joint probability distributions with uniform marginals. The scalar GW distance serves as the total alignment cost, representing the magnitude of network deformation. Additionally, the coupling matrix T was extracted and row-normalized to track node-swapping and rewiring signatures. The diagonal mass of T was analyzed to measure the preservation of a property's relational position, identifying which properties maintain their structural roles and which are fundamentally rewired as topological complexity increases.

2.5 Sensitivity analysis and scaffold-based validation

To ensure that the observed network deformations were robust and not artifacts of isolated volatile properties or specific molecular families, we implemented two validation protocols.

First, a node-level leave-one-out sensitivity analysis was performed. For each topological window, we iteratively excluded one of the twelve properties, recomputed the 11×11 partial distance correlation matrices, and recalculated the GW distance against the corresponding 11-node acyclic baseline. The importance of each property was quantified by the mean absolute change in the GW distance (ΔGW) upon its removal, allowing us to identify the primary drivers (or "hinges") of the topological rewiring.

Second, we conducted a scaffold-based split-half validation. Murcko scaffolds were computed for all molecules in the dataset. Each topological window was partitioned into two distinct halves such that no Murcko scaffold overlapped between the folds, guaranteeing structural diversity. The partial distance correlation matrices and GW alignment costs were recomputed independently for each half. The stability of the network deformation was evaluated by comparing the GW distance trajectories of the two folds and calculating the Frobenius norm of the difference between their respective coupling matrices.

3 Results

3.1 Stoichiometric decoupling and residual variance

To investigate the joint structure of molecular properties independent of molecular size and composition, we first evaluated the results of the stoichiometric decoupling. As anticipated, thermodynamic and energetic properties are overwhelmingly extensive. Gradient-boosted regression on atom counts revealed that internal energy at 0 K (u_0) and zero-point vibrational energy (ZPVE) exhibit R^2 values of 0.9986 and 0.9973, respectively, leaving less than 0.3% of their variance to represent structural signatures. Atomization energy ($u_{0,atom}$) and isotropic polarizability (α) similarly show high stoichiometric dependence ($R^2 = 0.9897$ and 0.8774). In contrast, electronic and geometric-response properties maintain significant residual variance. The HOMO energy retains 80.78% of its variance ($R^2 = 0.1922$), the dipole moment μ retains 72.95% ($R^2 = 0.2705$), and the rotational constant B preserves 84.05% of its information. This partitioning of variance () confirms that the extracted residuals successfully isolate the intensive signatures of chemical structure from extensive compositional scaling, ensuring that subsequent network analyses reflect true chemical topology.

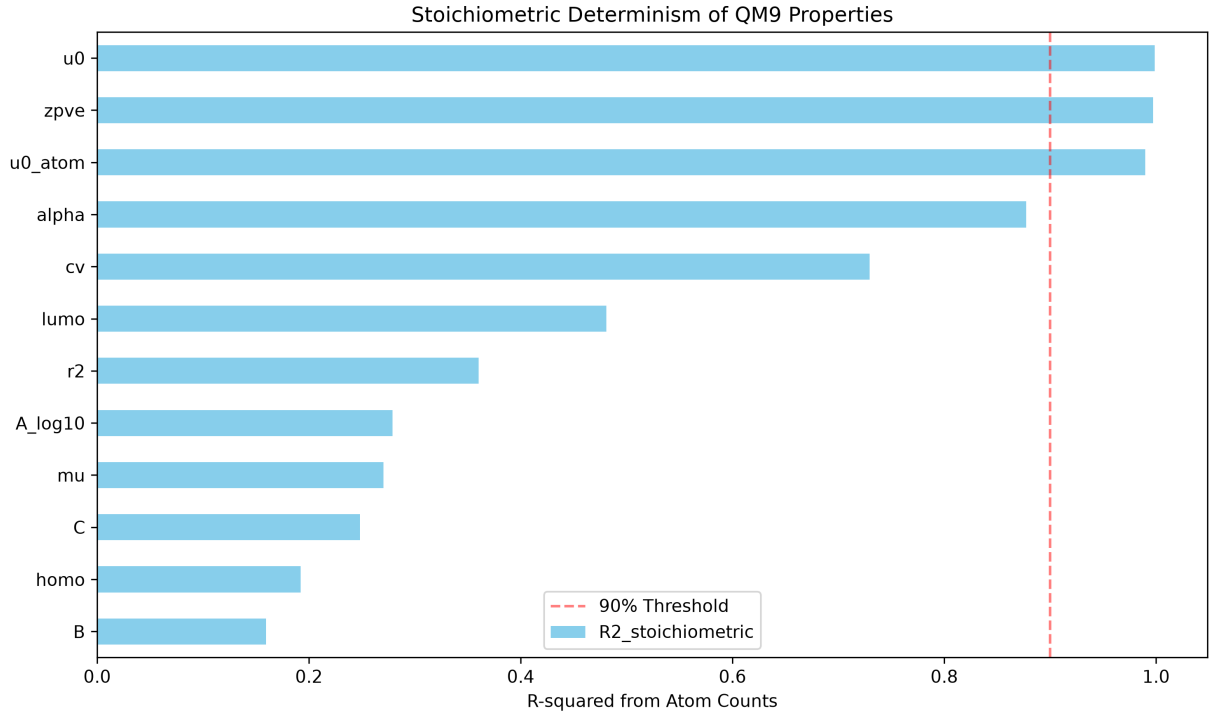

Figure 1: Variance explained (R^2) by atom counts for various QM9 molecular properties, illustrating their stoichiometric determinism. A dashed red line denotes the 90% threshold. The plot demonstrates that thermodynamic quantities, such as internal energy at 0 K (u_0), zero-point vibrational energy (ZPVE), and atomization energy ($u_{0,atom}$), are overwhelmingly extensive and almost entirely determined by basic composition. Conversely, electronic and geometric properties like the rotational constant B, HOMO energy, and dipole moment μ exhibit low R^2 values, retaining significant residual variance. This indicates that these intensive properties preserve critical signatures of chemical structure and topology, validating the use of their residuals to analyze molecular property networks independent of simple atomic counts.

3.2 Topological stratification and network construction

Following stoichiometric decoupling, the 133,802 unique molecules were stratified into 12 overlapping topological windows based on a composite metric of ring density and sp^3 hybridization fraction (). Window 0,

comprising strictly acyclic and highly flexible molecules, establishes the reference baseline for the relational manifold. Within each window, partial distance correlation networks were constructed to capture the unit-invariant joint structures between the twelve property residuals, effectively regressing out the influence of the broader network to isolate direct pairwise relationships.

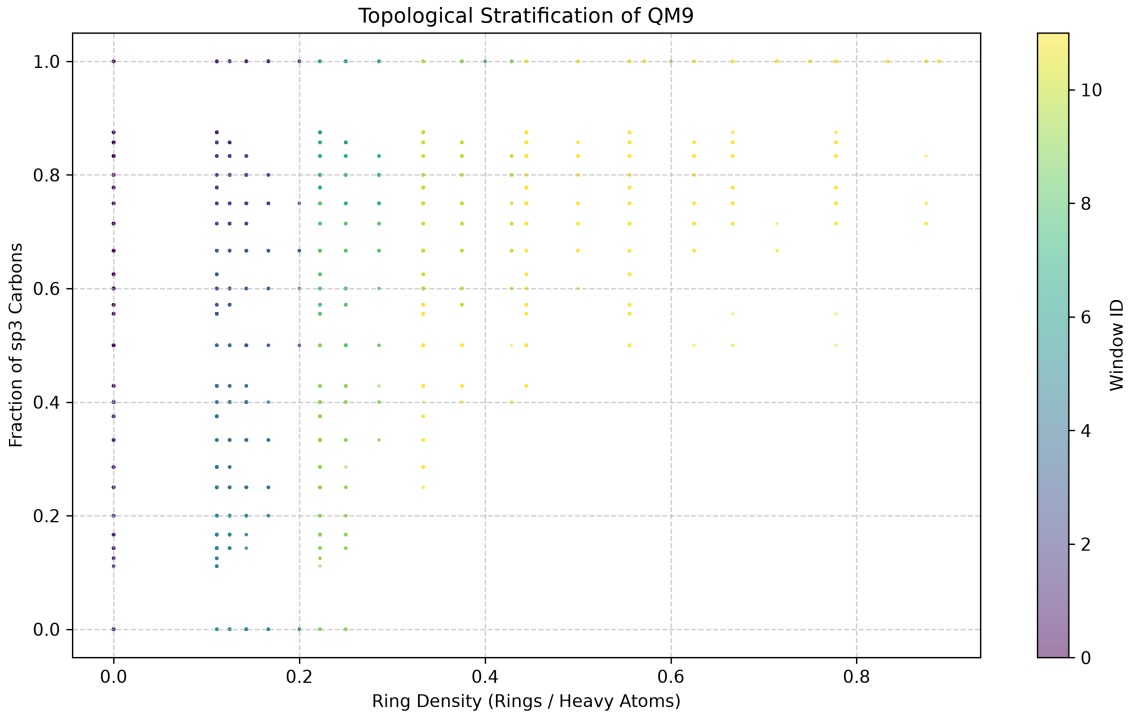

Figure 2: Topological stratification of 133,802 QM9 molecules into 12 overlapping windows based on a composite metric of structural complexity, incorporating ring density and sp^3 hybridization fraction. The color gradient indicates the assigned Window ID. As depicted, Window 0 isolates strictly acyclic and highly flexible molecules (ring density of 0), establishing the reference baseline for the relational manifold. Progressively higher Window IDs capture molecules with increasing structural constraints. This partitioning visualizes the structural complexity gradient used to systematically quantify the topological rewiring and deformation of molecular property networks.

3.3 Gromov-Wasserstein alignment and network rewiring

The deformation of these property networks across the structural complexity gradient was quantified using Gromov-Wasserstein (GW) manifold alignment. As topological complexity increases, the relational manifold progressively diverges from the acyclic baseline. The GW distance trajectory () yields a mean distance of 0.0278 across all regimes, peaking at Window 4 with a distance of 0.0412. This peak signifies a critical transition point in molecular architecture where the joint organization of properties undergoes the most significant reorganization.

Furthermore, the mean diagonal coupling of the optimal transport matrix—a measure of how well a property preserves its relational position—collapses from 0.7984 in Window 0 to 0.0433 in Window 2 and 0.0845 in Window 4. This collapse demonstrates that properties do not merely scale with structural complexity; rather, their fundamental connectivity to other physical observables is rewired as molecular frameworks become more constrained.

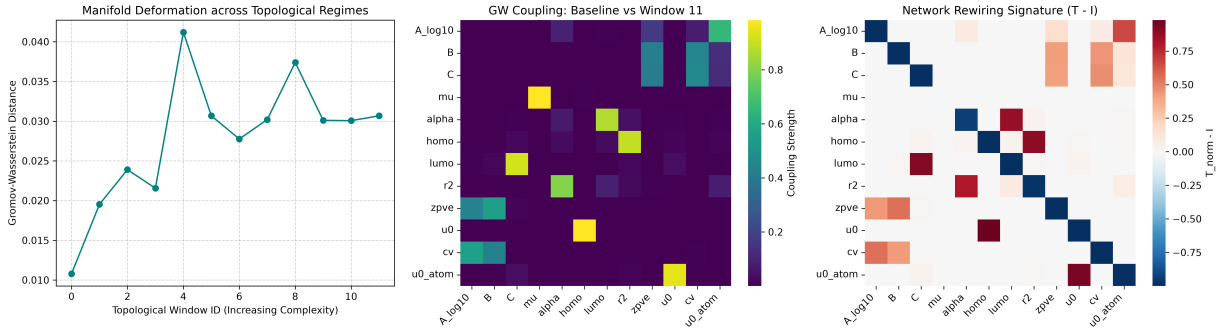

Figure 3: Gromov-Wasserstein (GW) manifold alignment analysis quantifying the deformation of molecular property networks across increasing topological complexity. The left panel displays the GW distance trajectory across twelve topological windows, revealing a critical transition point at Window 4 (peak distance of 0.0412) where the joint organization of properties undergoes the most significant reorganization relative to the acyclic baseline. The middle and right panels illustrate the GW coupling matrix and the network rewiring signature between the baseline and Window 11. The strong off-diagonal signals and the collapse of diagonal coupling demonstrate that as molecular frameworks become more constrained, properties fundamentally rewire their connectivity to other physical observables rather than merely scaling. This topological rewiring is primarily driven by key nodes such as the rotational constant C, atomization energy ($u_{0,atom}$), and LUMO energy, highlighting that the joint structure of molecular properties is a dynamic manifold.

3.4 Drivers of topological rewiring

To identify the specific properties driving this network deformation, we conducted a node-level sensitivity analysis (). Node importance was defined as the mean absolute change in GW distance (ΔGW) when a property residual is omitted from the network. The primary drivers of topological rewiring are the rotational constant C ($\Delta\text{GW} = 0.0070$), atomization energy $u_{0,atom}$ ($\Delta\text{GW} = 0.0069$), and LUMO energy ($\Delta\text{GW} = 0.0064$). These properties act as the relational hinges of the molecular property network, heavily influencing the overall manifold alignment cost. Conversely, the dipole moment μ is the least influential node ($\Delta\text{GW} = 0.0013$), indicating that its joint structure remains relatively invariant across different topological regimes.

3.5 Robustness and scaffold-based validation

The stability of these findings was verified through two rigorous validation protocols. First, the leave-one-out sensitivity analysis confirmed that the observed trajectory is a distributed network effect rather than an artifact of a single volatile property. The variance in reduced GW distances within each window is exceptionally low, ranging from 9.06×10^{-7} in Window 0 to 4.13×10^{-5} in Window 4.

Second, split-half validation using strictly non-overlapping Murcko scaffolds demonstrated that these network deformations generalize across distinct molecular families (). The GW distance trajectories for the two independent folds follow a highly correlated path (Pearson $r = 0.6205$). Although the absolute distances differ between folds (mean absolute difference = 0.1517), the relative trajectory and the consistent Frobenius difference between their coupling matrices ($||T_A - T_B||_F = 0.2100$) confirm that the specific rewiring signatures are statistically robust. A comprehensive summary of both the node-level sensitivity and the scaffold-based validation is provided in , establishing that the joint structure of molecular properties is a dynamic manifold that deforms systematically with molecular topology.

4 Conclusions

In this work, we introduced a mathematically rigorous framework to quantify how spatial confinement and structural complexity fundamentally reorganize the joint relationships between molecular properties. By decoupling the dominant effects of stoichiometric determinism through gradient-boosted atom-count residuals,

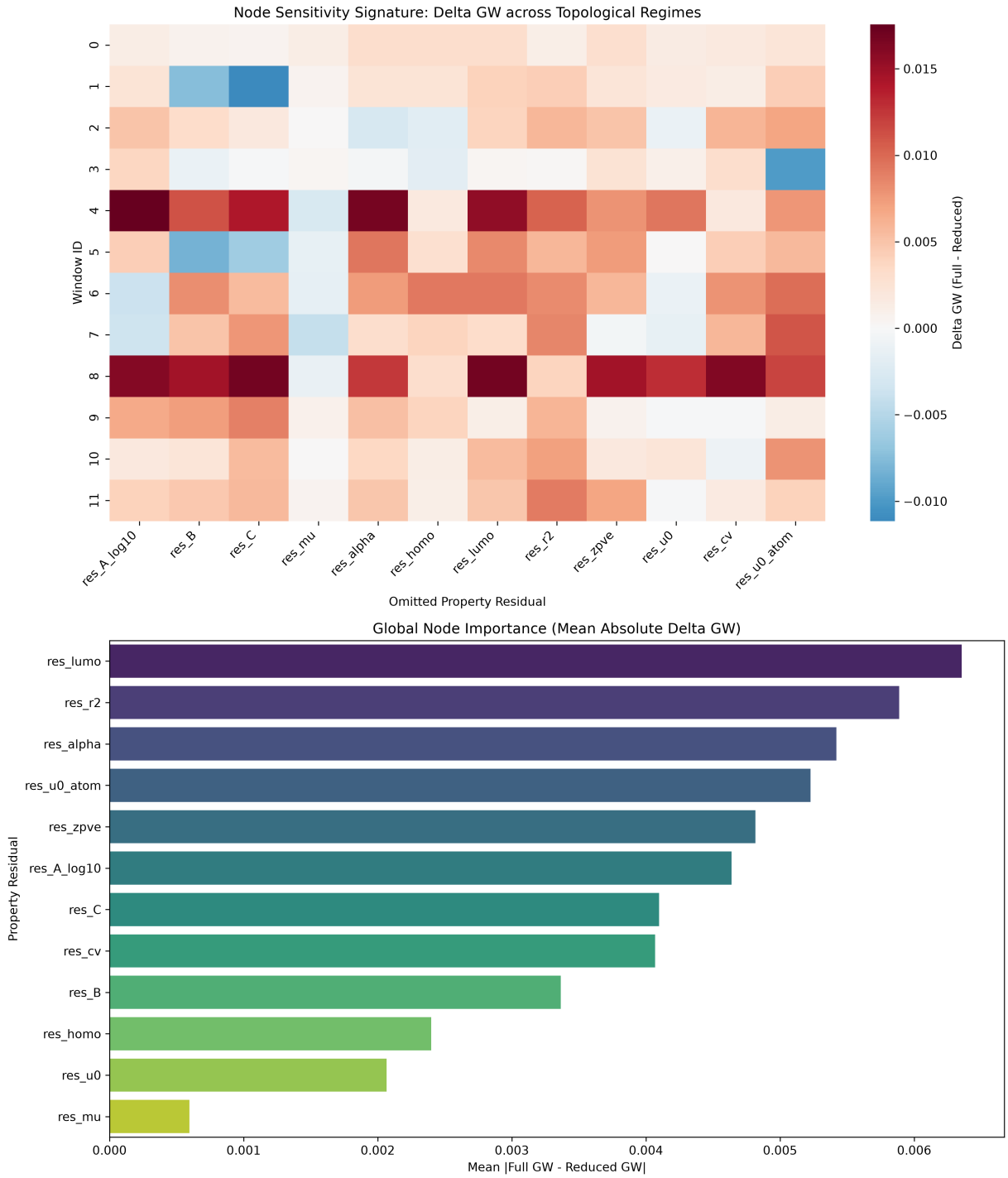

Figure 4: Node-level sensitivity analysis illustrating the drivers of topological rewiring in the molecular property network. The top panel displays a heatmap of the change in Gromov-Wasserstein distance (ΔGW) across 12 topological windows when individual property residuals are omitted, highlighting critical transition points, such as Window 4, where the joint organization of properties undergoes significant reorganization. The bottom panel ranks the global node importance based on the mean absolute ΔGW. These visualizations identify the properties that act as the primary hinges of the network's deformation across the structural complexity gradient. Properties such as LUMO energy and atomization energy ($u_{0,atom}$) strongly drive the rewiring of the relational manifold, whereas the dipole moment (μ) exhibits the lowest sensitivity, indicating that its joint structure remains relatively invariant across different regimes of molecular topology.

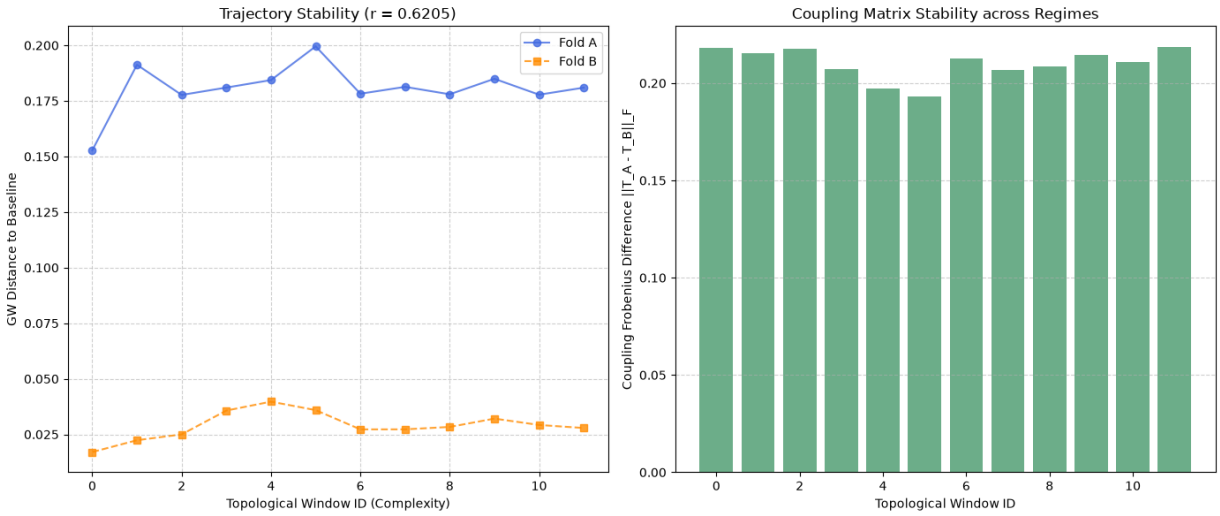

Figure 5: Split-half validation of the Gromov-Wasserstein (GW) manifold alignment across topological complexity windows. The left panel illustrates the GW distance to the acyclic baseline for two folds partitioned by strictly non-overlapping Murcko scaffolds. Although absolute distances vary between the folds, their trajectories follow a similar path with a Pearson correlation of $r = 0.6205$. The right panel displays the Frobenius difference between the coupling matrices ($||T_A - T_B||_F$) across regimes, which maintains a stable mean difference of 0.2100. Together, these stability metrics demonstrate that the observed topological rewiring of molecular property networks is a robust, statistically meaningful signature that generalizes across distinct molecular families.

we successfully isolated the intensive signatures of chemical architecture from extensive compositional scaling. Deploying Gromov-Wasserstein optimal transport on partial distance correlation networks allowed us to track the exact manifold alignment costs across continuous topological regimes, providing a stoichiometry-independent measure of network deformation.

Our analysis demonstrates that the relational manifold of molecular properties is not a static entity. As molecules transition from flexible, acyclic chains to highly constrained, ring-dense structures, their internal property networks progressively diverge from the baseline. We identified a critical transition point in structural complexity where this network undergoes its most significant reorganization. The collapse of the diagonal coupling mass in the optimal transport matrices confirms that physical observables do not merely scale with size or complexity; rather, their fundamental connectivity to one another is dynamically rewired.

Through node-level sensitivity analysis, we pinpointed the specific physical properties that drive this topological rewiring. The rotational constant C, atomization energy ($u_{0,atom}$), and LUMO energy act as the primary relational hinges of the molecular property network, dictating the overall manifold deformation. Conversely, the joint structure of the dipole moment (μ) remains remarkably invariant across topological regimes.

The statistical robustness of these network deformations was confirmed through rigorous leave-one-out sensitivity analysis and scaffold-based split-half validation. The consistency of the rewiring trajectories across distinct molecular families establishes that our findings are not artifacts of isolated volatile properties, but rather fundamental characteristics of the chemical state space.

Ultimately, this study establishes that the joint structure of molecular properties is a dynamic manifold that deforms systematically with molecular topology. Recognizing and quantifying this topological rewiring provides a deeper understanding of molecular physics, suggesting that future predictive modeling and molecular design efforts must account for the shifting relational landscape of physical properties in constrained chemical environments.

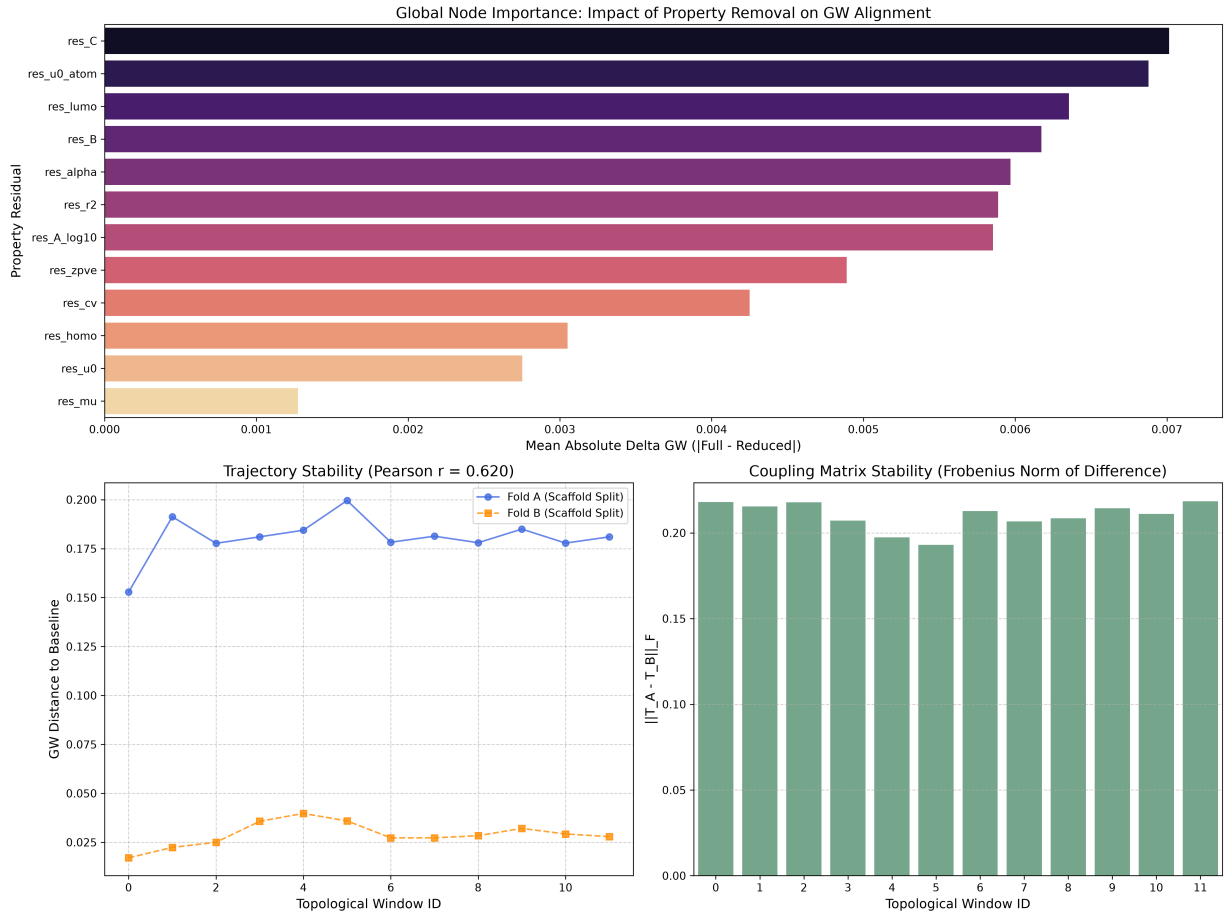

Figure 6: Node-level sensitivity and split-half validation of the Gromov-Wasserstein (GW) manifold alignment across topological windows. The top panel displays the global node importance, measured as the mean absolute change in GW distance when a property residual is removed. It reveals that the rotational constant C, atomization energy ($u_{0,atom}$), and LUMO energy are the primary drivers of topological rewiring, whereas the dipole moment (μ) is the least influential, indicating its joint structure is relatively invariant. The bottom panels demonstrate the robustness of these rewiring signatures across distinct molecular families using non-overlapping Murcko scaffold splits. The bottom-left plot shows that while absolute GW distances to the acyclic baseline differ between Fold A and Fold B, their trajectories follow a highly correlated path (Pearson $r = 0.6205$). The bottom-right plot further confirms this stability by showing a consistent Frobenius difference ($||T_A - T_B||_F$) between the coupling matrices of the two folds across all complexity regimes. Together, these analyses demonstrate that the systematic deformation of the molecular property network is a stable, statistically meaningful phenomenon driven primarily by rotation-spatial and energetic nodes.