Gromov-Wasserstein Alignment of Scaffold-Specific Distance Correlation Networks to Quantify Property Rewiring

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The macroscopic organization of molecular properties is heavily influenced by molecular size and composition, which often obscures the intrinsic topological dependencies defining distinct chemical families. To isolate genuine chemical architecture from these stoichiometric effects, we analyze 133,802 unique molecules from the QM9 dataset. We first residualize twelve non-redundant properties using gradient boosting models trained strictly on atom counts, successfully decoupling size-dependent physics to reveal a persistent vibrational-spatial coupling. Next, we partition the dataset into Murcko scaffolds and size-matched acyclic baselines to construct scaffold-specific distance correlation networks. By applying Gromov-Wasserstein optimal transport to these networks, we compute exact manifold alignment costs that quantify the magnitude of property rewiring between structural families. Our results demonstrate that rigid and unsaturated polycyclic systems undergo the most significant network rewiring, whereas highly saturated molecules retain property structures similar to acyclic baselines. Hierarchical clustering of the rewiring profiles identifies four distinct property blocks that deform coherently, highlighting how structural confinement alters the relationship between electronic occupancy and geometric response. Correlation analyses further establish that rewiring magnitude is primarily driven by molecular complexity and maximum ring size, and is inversely proportional to saturation. Finally, ablation studies confirm that these topological signatures are robust against heteroatom substitutions, offering a mathematically rigorous framework for mapping how spatial confinement reorganizes a molecule's thermodynamic and electronic state space.

1 Introduction

The macroscopic organization of molecular properties is fundamentally driven by molecular size and composition. In large-scale chemical databases, this stoichiometric determinism often dominates the variance of thermodynamic and electronic features, masking the intrinsic topological dependencies that define distinct chemical families. For instance, total internal energies and zero-point vibrational energies scale almost perfectly with atom counts, obscuring how spatial confinement and structural architecture independently modulate a molecule's physical state. To understand the genuine impact of chemical topology—such as ring formation, rigidity, and unsaturation—it is necessary to decouple these trivial size effects and isolate the underlying structural signal.

1.1 Quantifying topological transitions

When a molecule undergoes a structural transition from an acyclic chain to a constrained polycyclic framework, the relationships between its thermodynamic, electronic, and geometric properties are fundamentally reorganized. Quantifying this property rewiring presents a significant mathematical challenge. Traditional comparative analyses often rely on global coordinate spaces, which are highly susceptible to latent geometric artifacts and size confounding. Furthermore, standard linear correlation metrics fail to capture the complex, non-linear dependencies that couple vibrational states, spatial extents, and electronic occupancies. A rigorous quantification of structural confinement requires a relational, unit-invariant approach that operates strictly on intra-group dependency structures rather than absolute magnitudes.

1.2 Gromov-Wasserstein manifold alignment

To address this challenge, we propose a framework based on Gromov-Wasserstein (GW) optimal transport applied to scaffold-specific distance correlation networks. Distance correlation provides a powerful, unit-invariant measure capable of capturing complete, non-linear joint property structures. By representing distinct chemical architectures as pairwise distance correlation matrices, we can deploy GW optimal transport to compute exact manifold alignment costs between them. Because the GW distance evaluates the cost of aligning intra-space relational structures rather than absolute coordinates, it inherently neutralizes coordinate-space distortions and provides a pure measure of network deformation.

1.3 Scope of the study

In this study, we analyze 133,802 unique molecules from the QM9 dataset to map how spatial confinement reorganizes molecular state spaces. We first isolate genuine chemical architecture by extracting gradient-boosted atom-count residuals for twelve non-redundant properties. This stoichiometric residualization successfully decouples size-dependent physics, revealing a persistent vibrational-spatial coupling that is intrinsic to the molecular topology. We then partition the dataset into statistically dense Murcko scaffolds and strictly size-matched acyclic baselines. By computing the GW distance between the distance correlation networks of cyclic scaffolds and their acyclic counterparts, we extract a mathematically rigorous, stoichiometry-independent measure of property rewiring.

Our results demonstrate that rigid and unsaturated polycyclic systems undergo the most significant network rewiring, whereas highly saturated molecules retain property structures closely resembling acyclic baselines. Through hierarchical clustering of the rewiring profiles, we identify distinct property blocks that deform coherently, highlighting how structural confinement alters the relationship between electronic occupancy and geometric response. Finally, we establish the structural drivers of this reorganization—specifically molecular complexity, maximum ring size, and saturation—and confirm through ablation studies that these topological signatures are robust against heteroatom substitutions. This approach offers a comprehensive and mathematically exact mapping of intrinsic molecular deformation across chemical space.

2 Methods

2.1 Data curation and preprocessing

To investigate the intrinsic topological dependencies of molecular properties, we utilized the QM9 dataset. To prevent data leakage and ensure structural uniqueness, the dataset was first deduplicated based on SMILES strings, yielding a final working set of 133,802 unique molecules. We isolated 12 non-redundant properties by removing thermodynamic and atomization restatements (e.g., zero-point energy, enthalpy, and free energy at 298 K, as well as their atomization equivalents) and the perfectly collinear HOMO-LUMO gap.

The rotational constant A exhibits an extreme heavy-tailed distribution that would otherwise dominate variance-based metrics; therefore, a $\log(A+1)$ transformation was applied. Additionally, the nine perfectly linear molecules in the dataset (characterized by $A = 0$ and $B = C$) were explicitly identified and tracked to ensure they did not introduce artifactual correlations during downstream structural grouping. For each molecule, exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were parsed, alongside the heavy-atom count (n_{heavy}) and the corresponding Murcko scaffold.

2.2 Stoichiometric decoupling via gradient boosting

Because macroscopic molecular properties in QM9 are heavily determined by simple stoichiometry, we isolated the variance driven strictly by chemical architecture. Independent gradient boosting regressors were trained to predict each of the 12 selected properties using only the five elemental atom counts as features. The residuals from these models were extracted to represent the property variance attributable to structural topology rather than size or composition.

To validate that this residualization process did not destroy underlying physical signals, we computed the global pairwise distance correlation ($dCor$) matrix for both the raw and residualized properties. Distance

correlation was chosen over Pearson correlation due to its ability to capture non-linear dependencies and its unit-invariance. The persistence of core joint structures—such as the coupling between vibrational states and spatial extents—in the residualized *dCor* matrix confirmed that the extracted signals reflect genuine topological factors rather than size-dependent scaling. All subsequent analyses were performed exclusively on these stoichiometric residuals.

2.3 Topological partitioning and baseline matching

Molecules were partitioned into distinct structural families based on their Murcko scaffolds. To establish a baseline for property networks in the absence of spatial confinement, an "acyclic population" was defined by filtering for molecules with a ring count of zero and an empty Murcko scaffold. Rather than treating this acyclic group as a monolithic baseline, it was strictly binned by heavy-atom count (n_{heavy}).

For the cyclic molecules, we filtered for statistically dense scaffold classes containing at least 500 molecules. Each dense cyclic scaffold was then assigned to its corresponding size-matched acyclic baseline based on its exact n_{heavy} value. This direct matching procedure eliminates size confounding without requiring complex downstream regression, ensuring that any observed network deformation is purely a consequence of ring formation and structural confinement.

2.4 Distance correlation networks and Gromov-Wasserstein alignment

For each dense cyclic scaffold class and its corresponding size-matched acyclic bin, a 12×12 pairwise distance correlation matrix was computed across the property residuals. Each *dCor* matrix serves as a unit-invariant, topological representation of the joint property structure specific to that chemical architecture.

To quantify the magnitude of property rewiring between structural families, we deployed Gromov-Wasserstein (GW) optimal transport. The *dCor* matrices were transformed into cost matrices C according to:

$$C_{i,j} = 1 - dCor_{i,j} \tag{1}$$

The diagonal elements were set to zero ($C_{i,i} = 0$) to prevent the optimal transport plan from being biased by self-correlations. The GW distance was then computed between the cost matrix of each cyclic scaffold and the cost matrix of its strictly size-matched acyclic baseline. This computation yields both a scalar GW distance, representing the total magnitude of network rewiring, and an optimal transport coupling plan (π) that maps the alignment of properties between the two spaces.

2.5 Rewiring signatures and structural drivers

To characterize the physical nature of the property shifts, we analyzed the coupling plans of the most highly rewired scaffolds. A difference matrix, defined as $\Delta dCor = dCor_{\text{cyclic}} - dCor_{\text{acyclic}}$, was computed to pinpoint exactly which property-pair correlations were forced to swap or decouple relative to the acyclic baseline. Hierarchical clustering was applied to these difference matrices to visually and mathematically group properties that deform coherently, allowing us to identify distinct blocks of properties (e.g., thermodynamic, electronic, geometric-response) that rewire in unison under structural confinement.

To determine the structural features driving this reorganization, simple topological descriptors were computed for each dense scaffold class, including the average fraction of sp^3 hybridized carbons (F_{sp3}), the maximum ring size, and the scaffold heavy-atom count. Spearman rank correlation analyses were performed between these structural descriptors and the computed GW distances to map how spatial confinement, saturation, and molecular complexity fundamentally deform the joint property network.

2.6 Robustness and sensitivity analysis

Finally, to ensure that the discovered rewiring signatures were global characteristics of the scaffold topologies rather than artifacts driven by localized chemical substitutions, a leave-one-sub-family-out sensitivity analysis was conducted. For the most highly rewired scaffold classes, specific sub-families defined by heteroatom substitutions (e.g., molecules containing specific functional groups like hydroxyl or fluoro substituents) were iteratively removed. The *dCor* difference matrices and GW distances were recomputed at each iteration.

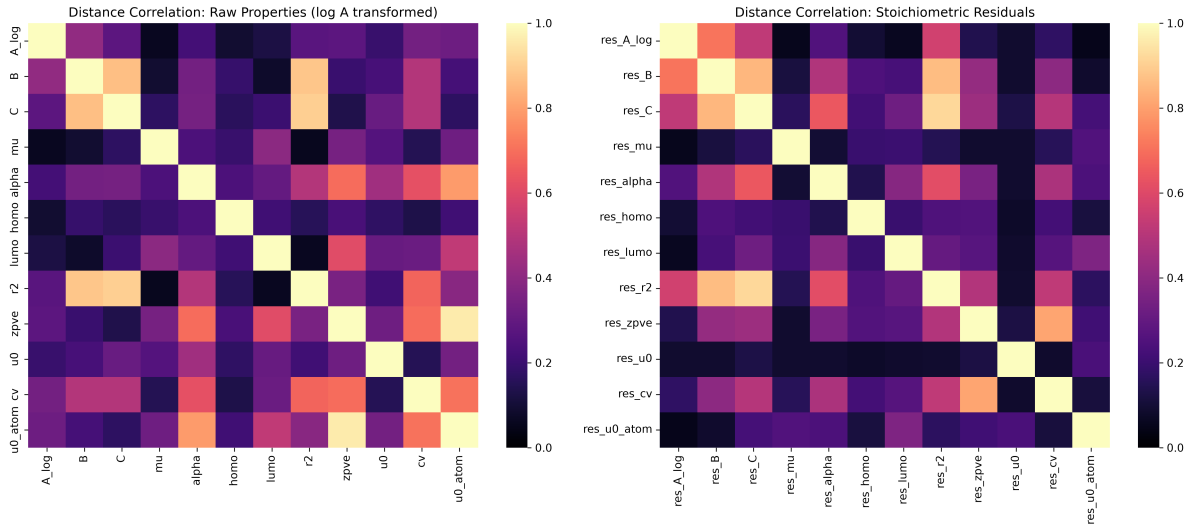

Figure 1: Distance correlation matrices comparing raw molecular properties (left) and stoichiometric residuals (right) for the QM9 dataset. The raw properties exhibit widespread high correlations driven by extreme stoichiometric determinism and molecular size effects. By residualizing these 12 properties based on atom counts, the right panel isolates the latent joint structure intrinsic to molecular topology. The residualized space reveals a persistent vibrational-spatial coupling that survives the removal of stoichiometric scaling, most notably the strong correlations between the residuals of zero-point vibrational energy (res_zpve), heat capacity (res_cv), and electronic spatial extent (res_r2). This demonstrates that these specific spatial and vibrational relationships are coherent structural features rather than artifacts of molecular composition.

The stability of the GW distances and the rank preservation of the difference matrices across these ablations were evaluated to confirm the robustness of the structural claims.

3 Results

3.1 Stoichiometric residualization and latent joint structure

The organization of molecular properties in the QM9 dataset is strongly driven by molecular size and composition. Linear regression based solely on the five elemental atom counts (C, N, O, F, H) achieves R^2 values of 1.0000 for total internal energies (u_0) and 0.9971 for zero-point vibrational energy ($zpve$). To uncover structural dependencies beyond these size effects, twelve non-redundant properties were residualized using gradient boosting models trained strictly on atom counts.

The residualized property space reveals a persistent joint structure that survives the removal of stoichiometric scaling (Figure 1). In the global population, the distance correlation ($dCor$) between the residuals of zero-point vibrational energy (res_zpve) and heat capacity (res_cv) is 0.8365. Furthermore, the electronic spatial extent residual (res_r2) maintains significant coupling with both res_cv ($dCor = 0.5627$) and res_zpve ($dCor = 0.5389$). These relationships indicate a coherent organization of spatial and vibrational states intrinsic to molecular topology, rather than an artifact of atom counts.

3.2 Quantifying network rewiring via Gromov-Wasserstein alignment

To quantify how specific chemical architectures reorganize these intrinsic property relationships, molecules were partitioned into Murcko scaffolds. Gromov-Wasserstein (GW) distances were computed between the $dCor$ matrices of cyclic scaffolds and their strictly size-matched acyclic baselines (molecules with zero rings and identical heavy-atom counts). The GW distance serves as a unit-invariant measure of how much the internal dependency network must be deformed to align one structural class with another.

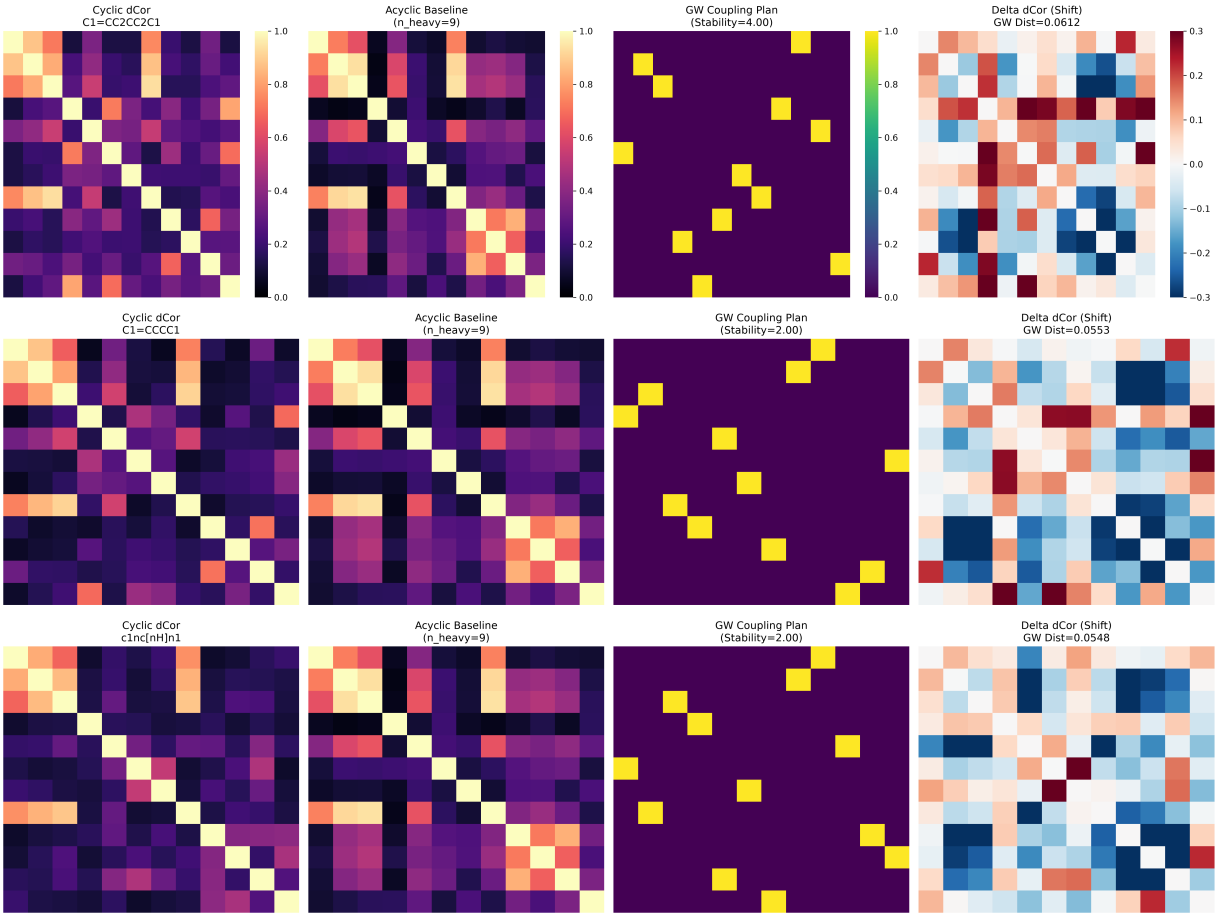

Figure 2: Visualization of property network rewiring for three highly deformed scaffolds: bicyclo[3.1.0]hex-2-ene (C1=CC2CC2C1), cyclopentene (C1=CCCC1), and 1,2,4-triazole (c1nc[nH]n1). The columns display the distance correlation ($dCor$) matrices for the cyclic scaffolds, their size-matched acyclic baselines ($n_{\text{heavy}} = 9$), the Gromov-Wasserstein (GW) coupling plans, and the resulting $\Delta dCor$ shift signatures. The GW distance quantifies the unit-invariant deformation required to align the internal dependency networks of different structural classes. The prominent shifts in the $\Delta dCor$ heatmaps reveal that rigid and unsaturated polycyclic systems experience the highest magnitude of property rewiring, demonstrating how specific cyclic architectures significantly reorganize joint molecular properties compared to acyclic structures.

The magnitude of property rewiring is highest in rigid or unsaturated polycyclic systems. As illustrated in Figure 2, the highest GW distances are observed for scaffolds such as bicyclo[3.1.0]hex-2-ene (GW distance = 0.0612), cyclopentene (0.0553), and 1,2,4-triazole (0.0548). In contrast, simple small rings such as cyclopropane (GW distance = 0.0312) and oxirane (0.0461) exhibit lower rewiring. This suggests their joint property structure remains closer to the acyclic baseline.

3.3 Decoupling signatures and property clustering

To understand the physical nature of this network deformation, we analyzed the average property shift profiles ($\Delta dCor$) across all scaffolds. Hierarchical clustering of the average property shift profiles across all scaffolds identifies four blocks that deform coherently (Figure 3). These blocks only partially align with conventional thermodynamic, electronic, and geometric categories:

- **Cluster 1 (Vibrational-Spatial Core):** Consists of res_α, res_zpve, and res_cv. This block represents the most stable joint structure in the dataset.
- **Cluster 2 (Electronic-Geometric Hybrid):** Groups res_r2, res_homo, res_u0, res_u0_atom, res_B, and res_C. The grouping of rotational constants with the HOMO energy and total internal energy residuals suggests that structural confinement rewires the relationship between electronic occupancy and the molecular moment of inertia.
- **Cluster 3 (Frontier-Response):** Links the dipole moment (res_μ) and LUMO energy (res_lumo).
- **Cluster 4 (Rotational Asymmetry):** Isolates res_A_log, which serves as a unique axis of axial elongation.

The average magnitude of these shifts shows how the vibrational-spatial core often decouples from the electronic-geometric hybrid block during ring formation.

3.4 Structural drivers of property reorganization

The magnitude of network rewiring is governed by specific topological features. Spearman rank correlation analysis identifies three primary drivers (Figure 4):

- **Molecular complexity:** Scaffold heavy-atom count ($n_{\text{heavy_scaffold}}$) is the strongest predictor of rewiring ($\rho = 0.481$, $p = 0.008$).
- **Ring strain/size:** Maximum ring size in the scaffold correlates significantly ($\rho = 0.445$, $p = 0.016$).
- **Saturation (F_{sp3}):** Increased saturation is associated with lower rewiring ($\rho = -0.442$, $p = 0.016$). Highly saturated molecules maintain property networks similar to acyclic alkanes, while sp^2 hybridization and aromaticity drive the network away from the baseline.

3.5 Robustness and cross-family stability

Ablation studies demonstrate that these rewiring signatures are intrinsic to the scaffold topology rather than artifacts of specific heteroatom substitutions. Sub-families defined by specific heteroatoms (e.g., nitrogen, oxygen, or fluorine) were iteratively removed from the most highly rewired scaffold classes, and the network deformations were recomputed.

The results (Figure 5) confirm the stability of the topological signatures:

- **Rank stability:** Across all sub-family ablations (e.g., removing all nitrogen-containing molecules from a scaffold family), the average Kendall's Tau between the full scaffold property network and the ablated version was 0.8697.
- **GW stability:** The standard deviation of the GW distance across these ablations is low (maximum $\sigma = 0.0104$ for bicyclo[3.1.0]hex-2-ene).

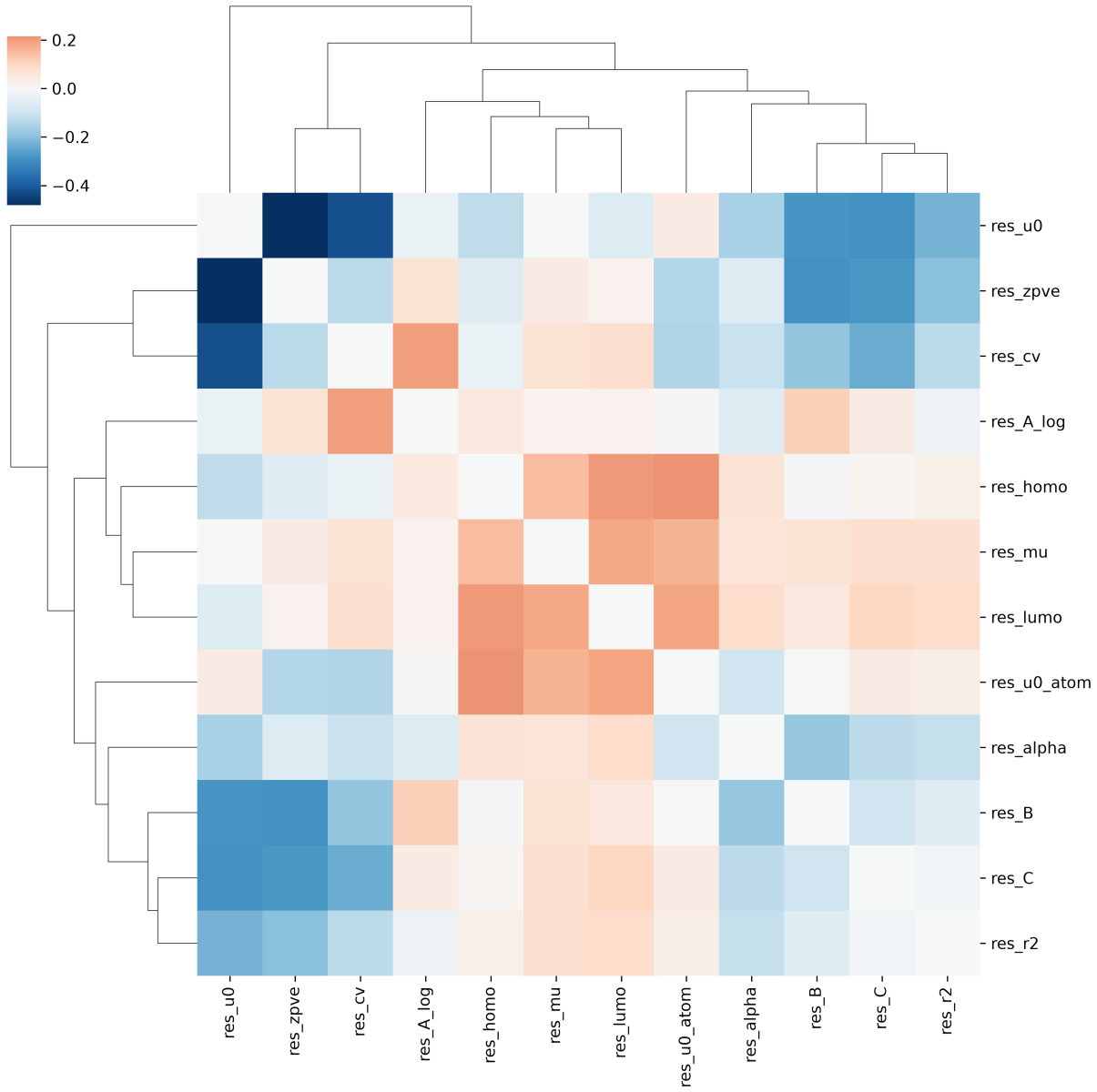

Figure 3: Hierarchical clustering of average property shift profiles ($\Delta dCor$) across Murcko scaffolds, illustrating the coherent deformation of residualized molecular properties during structural transitions. The heatmap identifies distinct blocks of properties that rewire together, demonstrating that these shifts only partially align with conventional thermodynamic, electronic, and geometric categories. Notably, the clustering reveals how the vibrational-spatial core (including *res_zpve* and *res_cv*) decouples from the electronic-geometric hybrid block (such as *res_r2*, *res_B*, and *res_C*) during ring formation. This visualizes the extent to which structural confinement reorganizes the internal dependency network of the molecules, altering fundamental relationships such as those between electronic occupancy and the molecular moment of inertia.

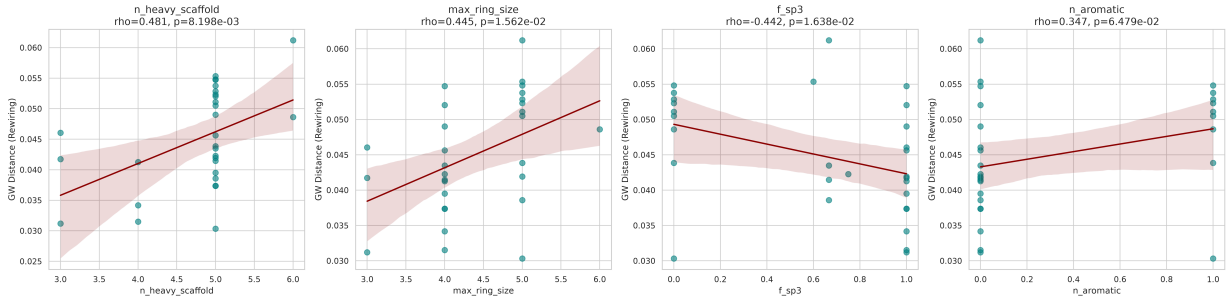

Figure 4: Scatter plots illustrating the structural drivers of property network reorganization, quantified by Gromov-Wasserstein (GW) distance. The magnitude of network rewiring is plotted against four topological features: scaffold heavy-atom count (`n_heavy_scaffold`), maximum ring size (`max_ring_size`), saturation fraction (F_{sp3}), and aromatic ring count (`n_aromatic`). Spearman rank correlation (ρ) analysis reveals that molecular complexity and ring size are significant positive predictors of rewiring, indicating that larger and more complex ring systems deviate further from size-matched acyclic baselines. Conversely, increased saturation exhibits a significant negative correlation, demonstrating that highly saturated molecules maintain joint property structures similar to acyclic alkanes, while sp^2 hybridization and aromaticity drive the internal dependency network away from the baseline.

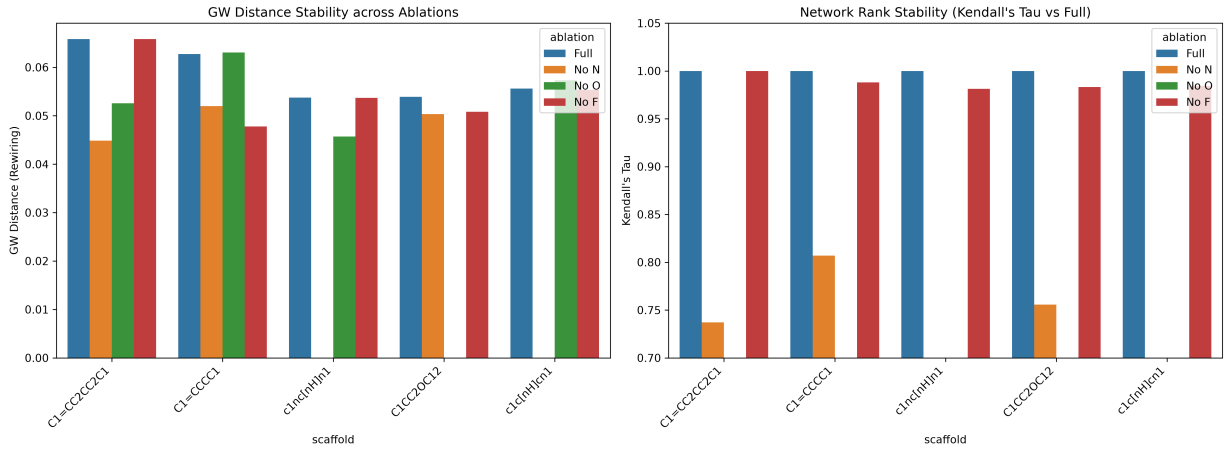

Figure 5: Robustness of property network rewiring signatures across heteroatom ablations for selected Murcko scaffolds. The left panel illustrates the stability of the Gromov-Wasserstein (GW) distance, which quantifies the global magnitude of property network rewiring. The GW distance remains highly consistent despite the removal of specific heteroatoms (nitrogen, oxygen, or fluorine) from the scaffold families. The right panel displays the network rank stability, measured by Kendall's Tau between the ablated and full scaffold networks. Although the removal of certain heteroatoms can shift local property correlations, the overall stability of the GW distances demonstrates that the global rewiring magnitude is a repeatable signature intrinsic to the underlying carbocyclic or heterocyclic scaffold topology, rather than an artifact of specific atomic substitutions.

While the removal of specific heteroatoms (particularly oxygen) can shift local correlations (minimum Kendall's Tau of 0.548 for the triazole scaffold), the global rewiring magnitude (GW distance) remains a stable and repeatable signature of the underlying carbocyclic or heterocyclic framework.

4 Conclusions

In this study, we introduced a mathematically rigorous framework to quantify how spatial confinement and structural topology fundamentally reorganize molecular properties. By analyzing 133,802 unique molecules from the QM9 dataset, we demonstrated that the extreme stoichiometric determinism dominating macroscopic molecular properties can be successfully decoupled. Through the extraction of gradient-boosted atom-count residuals, we isolated the genuine chemical architecture, revealing a persistent latent joint structure—most notably a robust vibrational-spatial coupling—that survives the removal of size-dependent scaling.

By representing distinct chemical architectures as scaffold-specific distance correlation networks, we deployed Gromov-Wasserstein optimal transport to compute exact, unit-invariant manifold alignment costs between cyclic scaffolds and strictly size-matched acyclic baselines. This relational approach successfully neutralized coordinate-space distortions, providing a pure measure of network deformation. Our results establish that rigid and unsaturated polycyclic systems experience the most severe property rewiring. In contrast, highly saturated molecules and simple small rings maintain internal dependency networks that closely resemble those of unconstrained acyclic chains.

Furthermore, hierarchical clustering of the property shift profiles revealed that molecular properties do not deform along conventional thermodynamic, electronic, or geometric lines. Instead, they reorganize into four distinct blocks that shift coherently under structural confinement. Most prominently, ring formation forces a decoupling between the vibrational-spatial core and an electronic-geometric hybrid block, highlighting how spatial constraints fundamentally alter the relationship between electronic occupancy and the molecular moment of inertia.

Correlation analyses confirmed that the magnitude of this network rewiring is systematically driven by molecular complexity and maximum ring size, while being inversely proportional to sp^3 saturation. Crucially, extensive ablation studies demonstrated that these topological signatures are highly robust against heteroatom substitutions. The global magnitude of property rewiring remains a stable, repeatable signature of the underlying carbocyclic or heterocyclic framework, rather than an artifact of localized chemical functionalization.

Ultimately, this work provides a comprehensive, stoichiometry-independent mapping of intrinsic molecular deformation. By shifting the analytical focus from absolute property magnitudes to intra-group relational structures, this Gromov-Wasserstein alignment framework offers a powerful new lens for understanding how topological transitions reorganize a molecule's thermodynamic and electronic state space.