Persistent Homology of Distance Correlation Networks to Map Topological Property Rewiring Across Scaffolds

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how macroscopic molecular architecture influences the internal dependencies of physical properties remains challenging, particularly when attempting to decouple these effects from molecular size. To address this problem, we introduce a size-invariant topological framework that maps the deformation of property-dependency networks across distinct structural families. Using the QM9 dataset, we first eliminate stoichiometric determinism by extracting non-linear gradient-boosted residuals for twelve non-redundant physical properties. The dataset is then partitioned into an acyclic baseline and twenty highly populated cyclic Murcko scaffolds. For each structural family, we compute distance correlation matrices to capture intrinsic joint structures and treat them as adjacency matrices to construct Vietoris-Rips simplicial complexes. By extracting the persistent homology of these complexes and computing Bottleneck distances between scaffold persistence diagrams, we mathematically quantify the deformation of the property networks. Our results demonstrate that structural transitions, specifically cyclization and heteroatom substitution, fundamentally rewire property couplings. While acyclic molecules are characterized by thermodynamic-geometric loops, strained bicyclic systems and heterocycles exhibit significant topological shifts toward complex electronic-vibrational couplings, demonstrating how specific structural motifs govern the coherent physical state space of molecules.

1 Introduction

1.1 Background and motivation

The relationship between a molecule's macroscopic architecture and its intrinsic physical properties is a foundational concept in chemistry. While the prediction of individual molecular properties has seen significant advancements, understanding the joint organization and internal dependencies among multiple physical properties remains a complex challenge. A major confounding factor in this domain is stoichiometric determinism; extensive properties such as internal energy or polarizability are heavily dominated by molecular size and atom counts. Consequently, when analyzing property correlations across large chemical spaces, the subtle, intrinsic effects of structural motifs—such as cyclization or heteroatom substitution—are often masked by the overwhelming signal of molecular weight. To uncover how specific structural scaffolds reorganize a molecule's physical state space, it is essential to decouple these size effects and examine the underlying property dependency networks.

1.2 Topological framework for property networks

Traditional approaches to analyzing property relationships often rely on linear correlation metrics or assume multivariate normality, which can fail to capture complex, non-linear dependencies and are highly sensitive to coordinate-space outliers. Furthermore, comparing these dependency structures across different chemical families requires a mathematical framework capable of quantifying network deformation in a robust manner.

In this work, we introduce a size-invariant topological framework to map the rewiring of property-dependency networks across distinct structural families. By leveraging persistent homology, a tool from topological data analysis, we extract global, coordinate-free signatures of these networks. We specifically focus on higher-order circular constraints, represented by one-dimensional homological features (H_1 cycles), which indicate that several physical properties are mutually bound in a continuous loop of dependencies.

1.3 Scope of the study

We apply this framework to the QM9 dataset, analyzing twelve non-redundant physical properties that encompass thermodynamic, geometric, and electronic states. To eliminate stoichiometric determinism, we first extract non-linear gradient-boosted residuals for these properties, controlling for exact heavy atom counts and basic connectivity. The dataset is partitioned into an acyclic baseline and the twenty most populated cyclic Murcko scaffolds.

For each structural family, we compute distance correlation ($dCor$) matrices to capture the intrinsic joint structures of the size-invariant residuals. Treating these matrices as adjacency structures, we construct Vietoris-Rips simplicial complexes and compute their persistent homology. The deformation of the property networks is then mathematically quantified by calculating the Bottleneck distances between the persistence diagrams of the cyclic scaffolds and the acyclic baseline.

Our analysis reveals that macroscopic spatial confinement and structural transitions fundamentally rewire internal property couplings. We demonstrate that while the state space of acyclic molecules is primarily characterized by thermodynamic-geometric loops, the introduction of specific structural motifs induces significant topological shifts. Strained bicyclic systems and small heterocycles exhibit maximal divergence from the acyclic baseline, transitioning toward complex electronic-vibrational couplings. By quantifying these topological deformations, this work provides a rigorous, size-invariant method for mapping how molecular architecture governs the coherent physical state of molecules.

2 Methods

2.1 Data ingestion and structural featurization

The QM9 dataset was utilized as the primary data source. To prevent data leakage, the dataset was deduplicated based on SMILES strings, yielding 133,802 unique molecules. Molecular graphs were parsed using RDKit, with parallel processing employed to optimize computational efficiency. For each molecule, exact heavy atom counts (C, N, O, F) and the number of hydrogen atoms were extracted. To control for basic connectivity, the fraction of sp^3 hybridized carbons and the total number of rings were computed. The macroscopic structural family of each molecule was defined by extracting its exact Murcko scaffold. Nine linear molecules, identified by a rotational constant A of exactly zero, were explicitly excluded from subsequent training and topological analyses to prevent geometric singularities from skewing the results. A $\log(A+1)$ transformation was applied to the rotational constant A to mitigate its extreme heavy-tailed distribution.

2.2 Stoichiometric decoupling via gradient boosting

To ensure that the observed property dependencies reflect intrinsic structural organization rather than extensive molecular size effects, stoichiometric determinism was rigorously decoupled. The analysis focused on twelve non-redundant physical properties: $\log(A+1)$, B, C, μ, α, $HOMO$, $LUMO$, r^2, $ZPVE$, U_0, C_v, and $U_{0,atom}$. Non-linear gradient-boosted regression trees were trained using LightGBM for each of the twelve properties. The predictive features comprised the five atom counts, the sp^3 carbon fraction, and the ring count. By computing the residuals from these models, the pure structural contributions to each property were isolated. These residuals were subsequently standardized to form the basis for all downstream topological analyses, ensuring that shared linear dependencies driven by molecular size were effectively eliminated.

2.3 Structural family partitioning

The dataset was partitioned into distinct structural families based on the computed Murcko scaffolds. Molecules lacking a ring system (an empty Murcko scaffold) were isolated to serve as the acyclic baseline, comprising 13,989 molecules. Rather than grouping molecules by generic ring counts, the top 20 most populated specific cyclic Murcko scaffolds were selected, with each scaffold containing strictly more than 500 molecules. This specific partitioning strategy ensures that the topological rewiring signals can be attributed to distinct chemical motifs—such as specific heterocycles or strained bicyclic systems—rather than generic spatial confinement.

2.4 Distance correlation networks and metric validation

For the acyclic baseline and each of the 20 cyclic scaffold families, pairwise distance correlation ($dCor$) matrices were computed for the twelve standardized property residuals. To handle potential numerical noise, the $dCor$ values were clipped to the $[0, 1]$ interval. Each correlation matrix was then transformed into a distance matrix using the relation $D = \sqrt{1 - dCor}$.

Because the construction of Vietoris-Rips complexes requires a strict metric space, the triangle inequality was explicitly verified for all triplets in the resulting distance matrices. In cases where violations might occur, a shortest-path graph distance transformation via the Floyd-Warshall algorithm was designated to enforce metric properties. The validity of this metric enforcement was confirmed by calculating the Spearman correlation between the original distances D and the metric-enforced shortest-path distances. For all scaffolds analyzed, this correlation was 1.00, with zero triangle inequality violations, confirming that the topological representations accurately reflect the original $dCor$ manifold.

2.5 Persistent homology and topological rewiring quantification

The `gudhi` library was utilized to construct Vietoris-Rips simplicial complexes from the validated distance matrices of each structural family. Persistent homology was computed up to dimension 1, capturing connected components (H_0) and higher-order circular constraints or cycles (H_1). The persistence diagrams were extracted, and the most persistent H_1 cycles were mapped back to the specific properties and dependencies in the original $dCor$ matrices to identify the chemical properties forming these loops.

To mathematically quantify the deformation of the property-dependency networks, the Bottleneck distance was computed between the persistence diagram of the acyclic baseline and that of each cyclic scaffold. No post-hoc normalization was applied to these Bottleneck distances, as the gradient-boosted residuals were already size-invariant. Pairwise Bottleneck distances among all selected cyclic scaffolds were also computed to summarize the global rewiring landscape.

2.6 Differential dependency analysis and stability validation

Differential dependency graphs were generated for each cyclic scaffold by subtracting the acyclic baseline $dCor$ matrix from the scaffold's $dCor$ matrix ($\Delta dCor = dCor_{scaffold} - dCor_{acyclic}$). Correlation analyses were performed between the magnitude of specific edge shifts and the fraction of heteroatoms (N, O, F) in the scaffold to provide a mechanistic explanation for the observed rewiring.

To ensure the chemical grounding and stability of the topological features, the two cyclic scaffolds exhibiting the highest and lowest Bottleneck distances relative to the baseline were selected for detailed case studies. A bootstrapping validation scheme was implemented ($B = 50$ iterations), wherein the acyclic baseline and the extreme scaffolds were subsampled to match sizes. The $dCor$ matrices, persistence diagrams, and Bottleneck distances were recomputed across these iterations to report the mean and variance, thereby confirming that the observed topological rewiring is a robust and stable chemical phenomenon.

3 Results

3.1 Stoichiometric decoupling and baseline joint structure

To ensure that the observed property dependencies reflect intrinsic structural organization rather than extensive molecular size effects, we first evaluated the success of the stoichiometric decoupling. Prior to residual extraction, total internal energies in the QM9 dataset were highly predictable from atom counts alone ($R^2 = 0.9999$). By extracting the standardized residuals from the LightGBM models, we successfully isolated the pure structural variance for the twelve physical properties. The baseline linear dependencies among these size-invariant thermodynamic, geometric, and electronic states are captured in the residual correlation matrix (Figure 1).

Using the distance correlation ($dCor$) matrix of these residuals, we analyzed the acyclic baseline (`scaffold_0`, comprising 13,989 molecules). The persistent homology of the acyclic dependency network reveals a prominent H_1 cycle with a persistence of 0.0743. This topological feature is born at a filtration distance of 0.8698,

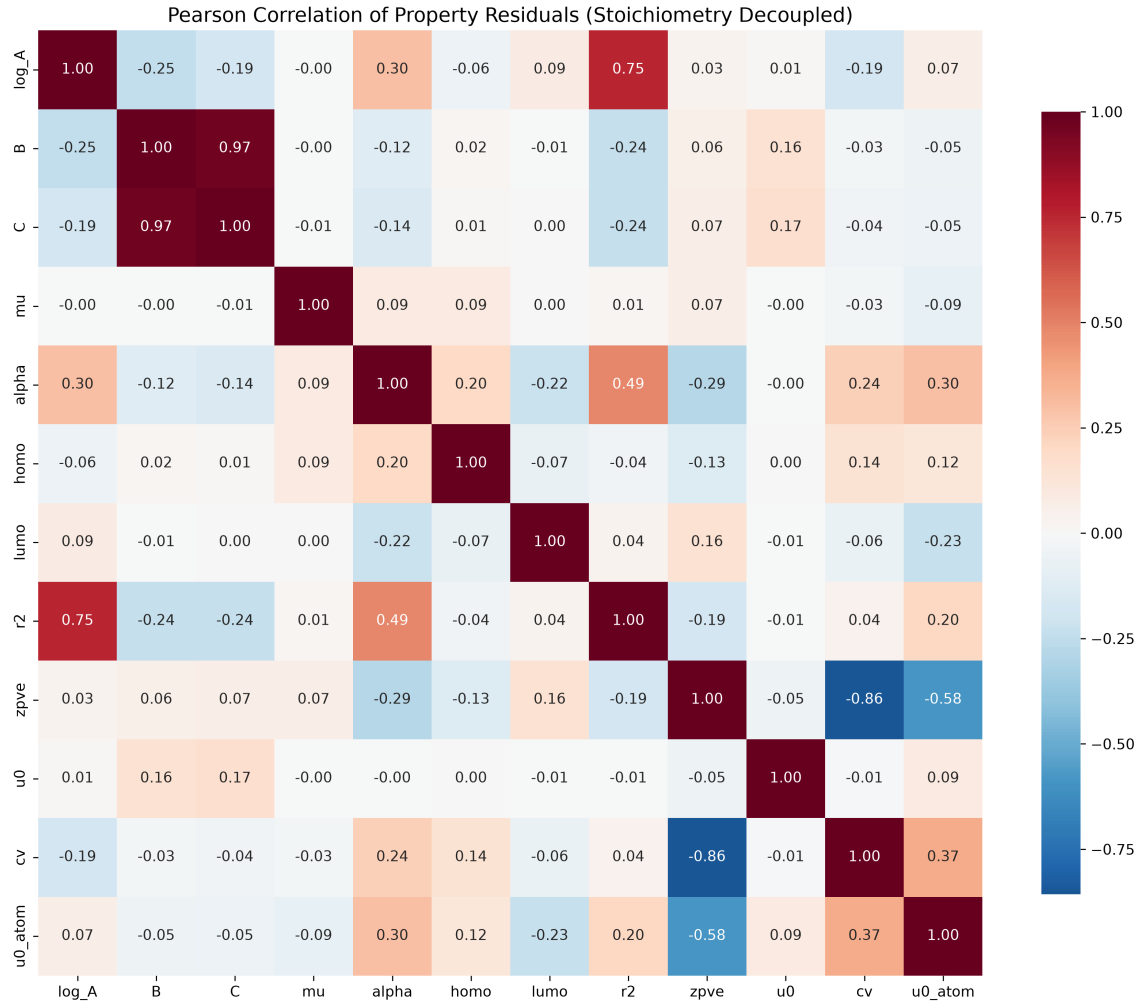

Figure 1: Pearson correlation matrix of the 12 molecular property residuals following stoichiometric decoupling via LightGBM. By eliminating variance explained by atom counts, the heatmap isolates pure structural signals and reveals the baseline linear dependencies among thermodynamic, geometric, and electronic states. We observe strong residual couplings, such as the positive correlation between rotational constants B and C and the negative correlation between zero-point vibrational energy ($ZPVE$) and heat capacity (C_v). These pairwise correlations serve as the foundational metric for constructing Vietoris-Rips simplicial complexes, enabling the subsequent identification of higher-order topological cycles (H_1) and the quantification of network deformation across different cyclic scaffolds.

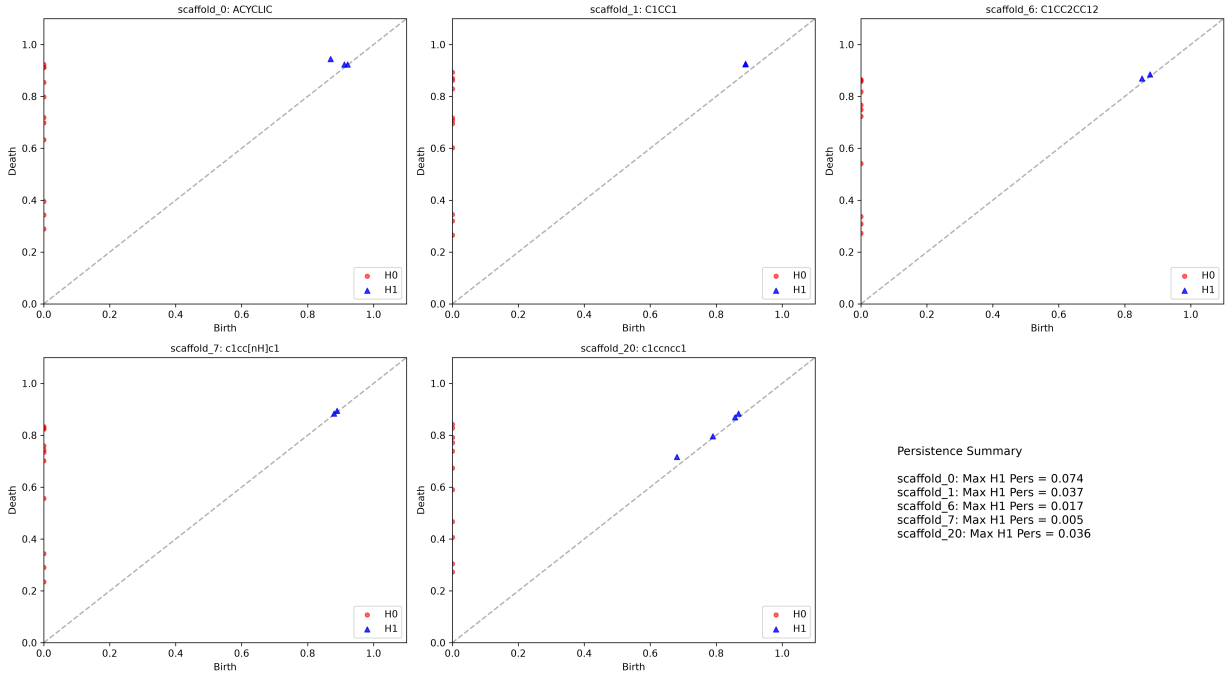

Figure 2: Persistence diagrams illustrating the topological rewiring of molecular property dependency networks across distinct structural scaffolds. The plots display zero-dimensional (H_0, red circles) and one-dimensional (H_1, blue triangles) topological features, with the acyclic baseline (`scaffold_0`) exhibiting a prominent H_1 cycle (persistence 0.074) that represents a thermodynamic-geometric loop. As molecules transition to cyclic and heteroaromatic architectures, such as the strained bicyclo[1.1.0]butane (`scaffold_6`) and pyridine (`scaffold_20`), the dependency networks undergo significant deformation. Notably, pyridine's primary H_1 cycle shifts to a lower birth filtration with a reduced persistence of 0.036, reflecting a fundamental rewiring from thermodynamic constraints to complex electronic-vibrational couplings. The overall variation in H_1 birth, death, and maximum persistence across the scaffolds demonstrates how macroscopic structural motifs govern the coherent physical state and internal dependencies of molecules.

originating from the coupling between the internal energy at 0 K (U_0) and the rotational constant C. The cycle dies at a filtration of 0.9441, closed by the property triangle $\{U_{0,atom}, U_0, ZPVE\}$. This signature indicates that in flexible, acyclic molecules, the primary higher-order constraint is a thermodynamic-geometric loop that links internal energy to mass distribution and the vibration-atomization axis.

3.2 Topological rewiring across structural scaffolds

Transitioning from acyclic chains to cyclic and heteroaromatic architectures induces significant deformation in the property dependency networks. The variation in the birth, death, and persistence of H_1 cycles across different structural families highlights how macroscopic motifs govern the coherent physical state of molecules (Figure 2).

We mathematically quantified this network deformation by computing the H_1 Bottleneck distance between the persistence diagram of the acyclic baseline and those of the 20 most populated cyclic Murcko scaffolds (Figure 3).

The global mean pairwise Bottleneck distance across the structural families is 0.0188. However, specific architectures exhibit maximal divergence from the acyclic baseline (H_1 distance ≈ 0.0372). These highly divergent structures predominantly include small heterocycles, such as oxirane (`scaffold_3`, `C1CO1`) and pyridine (`scaffold_20`, `c1ccncc1`), as well as highly strained bicyclic systems, such as bicyclo[1.1.1]pentane (`scaffold_19`, `C1C2CC1C2`) and bicyclo[1.1.0]butane (`scaffold_6`, `C1CC2CC12`).

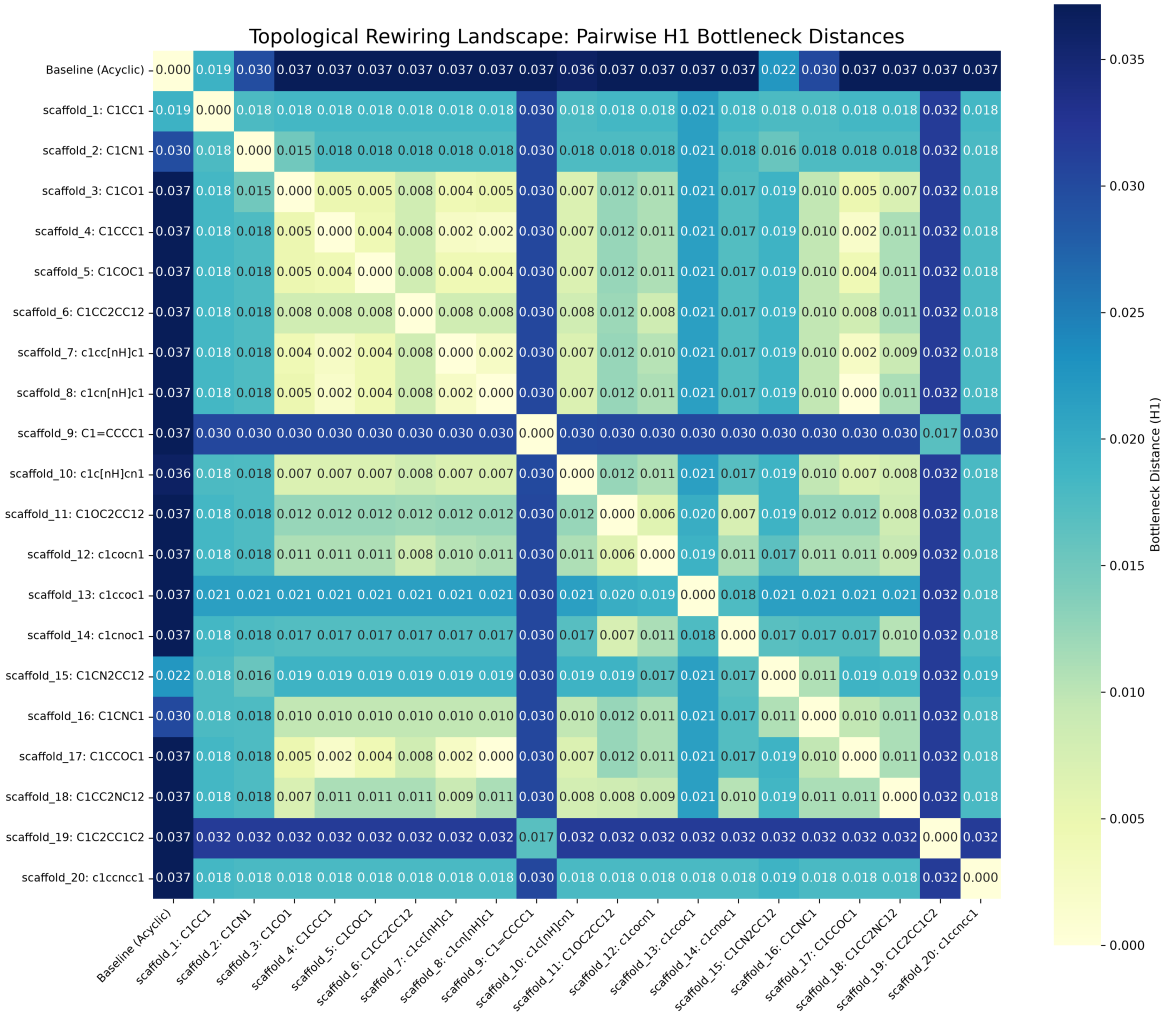

Figure 3: Heatmap of pairwise H_1 Bottleneck distances quantifying the topological deformation of molecular property dependency networks across an acyclic baseline and 20 cyclic Murcko scaffolds. Computed via Vietoris-Rips persistent homology after decoupling stoichiometric determinism, these distances measure how structural transitions rewire internal property couplings. The landscape reveals that small heterocycles, such as oxirane (`scaffold_3`) and pyridine (`scaffold_20`), alongside strained systems like bicyclo[1.1.1]pentane (`scaffold_19`), exhibit maximal divergence (distance ≈ 0.037) from the acyclic baseline. This demonstrates that cyclization and heteroatom substitution fundamentally shift the global constraints of the molecular state space from thermodynamic-geometric loops to complex electronic-vibrational couplings.

3.3 Differential dependencies and chemical insights

To elucidate the specific chemical drivers underlying this topological deformation, we analyzed the differential correlation matrices ($\Delta dCor$) for the cyclic scaffolds relative to the acyclic baseline (Figure 4).

The differential analysis reveals that heteroatom substitution is a primary driver of network rewiring. We found strong, statistically significant correlations between a scaffold's heteroatom ratio and specific shifts in property couplings. Most notably, the C-$HOMO$ coupling correlates with the heteroatom ratio at $r = -0.7816$ ($p < 0.0001$), the $HOMO$-r^2 coupling at $r = -0.7608$ ($p < 0.0001$), and the μ-$ZPVE$ coupling at $r = -0.6760$ ($p = 0.0011$).

Detailed case studies of the maximally divergent scaffolds further ground these topological shifts in chemical intuition:

- **Pyridine (`scaffold_20`):** Pyridine represents a significant departure from the acyclic baseline. Its primary H_1 cycle (persistence 0.0359) is born at a lower filtration of 0.6806 from the C_v-$HOMO$ edge and dies at 0.7165 in the $\{C_v, ZPVE, HOMO\}$ triangle. This indicates a fundamental shift from thermodynamic-geometric loops to complex electronic-vibrational couplings. Influence analysis demonstrates that derivatives such as pyridine-4-acetaldehyde maximize this specific coupling.
- **Bicyclo[1.1.1]pentane (`scaffold_19`):** This strained scaffold exhibits a distinct H_1 cycle (persistence 0.0635) born from the $U_{0,atom}$-α edge. This suggests that in rigid, strained architectures, the energy required to fragment the molecule becomes uniquely coupled to its electronic polarizability—a dependency structure that is entirely absent in flexible acyclic chains.

3.4 Stability and validation of topological features

To confirm that the observed topological rewiring is a stable chemical phenomenon and not an artifact of sample size or dataset partitioning, we performed a bootstrap validation ($B = 50$ iterations) on scaffolds representing both low and high divergence (Figure 5).

The bootstrap analysis demonstrates that the topological shifts are highly robust. The low divergence group (typified by cyclopropane) yielded a mean H_1 distance of 0.0255 ± 0.0116 (95% CI: [0.0221, 0.0288]), while the high divergence group (typified by pyridine) maintained a stable mean H_1 distance of 0.0258 ± 0.0114 (95% CI: [0.0225, 0.0291]).

4 Conclusions

In this study, we addressed the challenge of mapping the internal dependencies of molecular properties by decoupling stoichiometric determinism from intrinsic structural effects. By extracting size-invariant, gradient-boosted residuals and employing persistent homology on distance correlation networks, we developed a robust topological framework to quantify network deformation across distinct chemical families.

Our analysis of the QM9 dataset reveals that macroscopic molecular architecture fundamentally rewires the joint organization of physical properties. We established that the state space of flexible, acyclic molecules is primarily governed by thermodynamic-geometric loops, which link internal energy to mass distribution and vibrational axes. However, structural transitions such as cyclization and heteroatom substitution drive significant topological shifts. Highly strained architectures, like bicyclo[1.1.1]pentane, and small heterocycles, such as pyridine, exhibit maximal network deformation relative to the acyclic baseline, as quantified by their H_1 Bottleneck distances.

Crucially, this topological divergence corresponds to a mechanistic shift in the molecular state space, transitioning from baseline thermodynamic constraints to complex electronic-vibrational couplings. The strong correlation between heteroatom content and specific dependency shifts—such as the C-$HOMO$ and μ-$ZPVE$ couplings—demonstrates that specific chemical motifs dictate how distinct physical domains interact. For instance, the emergence of strong couplings between electronic polarizability and atomization energy in strained systems highlights how spatial confinement uniquely binds properties that remain independent in flexible chains.

Ultimately, this work establishes that while molecular size and stoichiometry set the extensive scale of physical properties, it is the underlying structural scaffold that defines their internal dependency network.

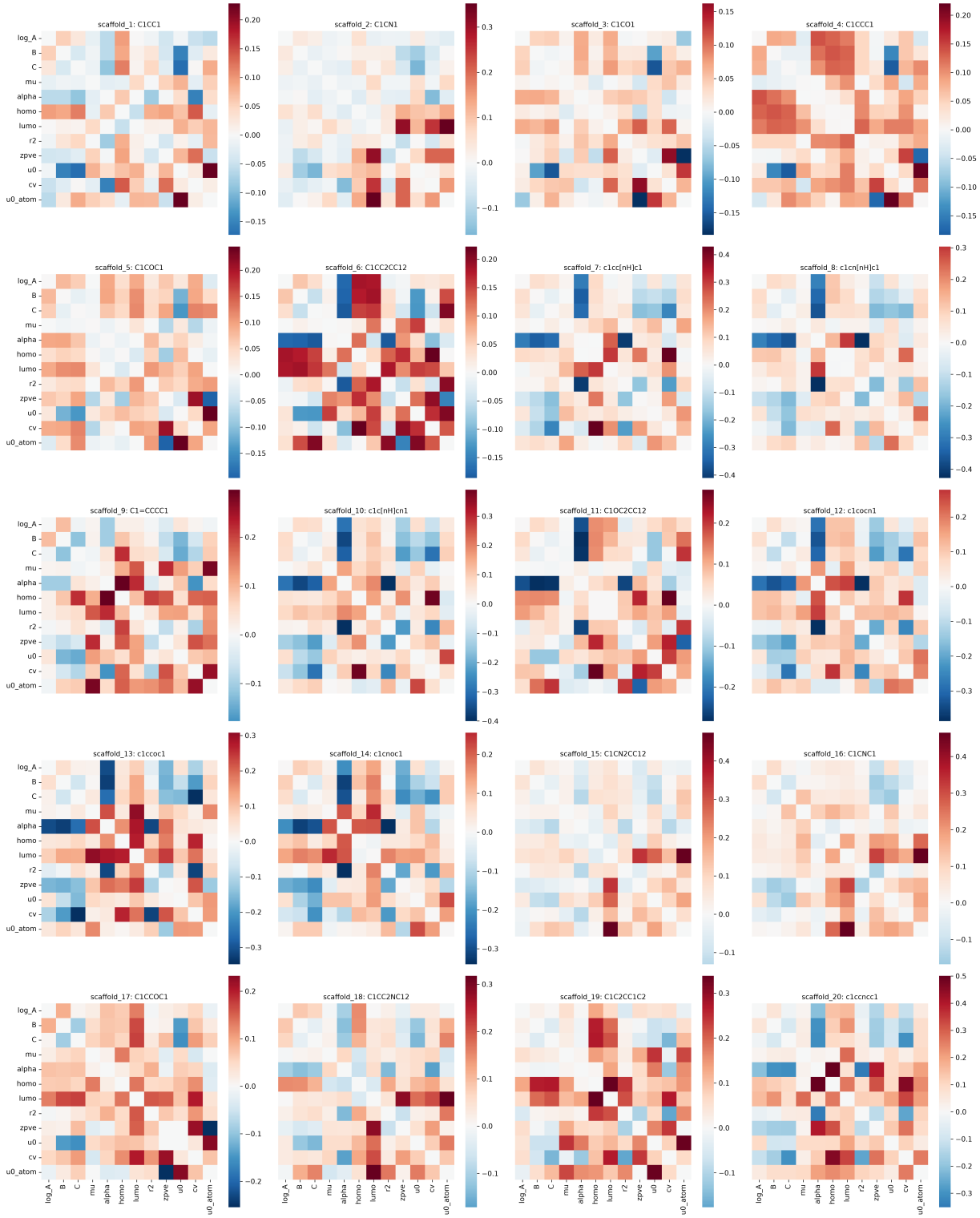

Figure 4: Differential correlation matrices ($\Delta dCor$) illustrating the topological rewiring of 12 physical property networks (including U_0, $U_{0,atom}$, C_v, and $ZPVE$) across the top 20 cyclic Murcko scaffolds relative to an acyclic baseline. Red and blue regions indicate positive and negative deviations in distance correlation, respectively, after decoupling stoichiometric determinism. The heatmaps reveal that structural transitions, specifically cyclization and heteroatom substitution, fundamentally alter internal property couplings. Scaffolds with strained architectures or high heteroatom ratios, such as bicyclo[1.1.1]pentane and pyridine, display maximal divergence. Ultimately, these differential patterns demonstrate how macroscopic molecular architecture shifts global H_1 constraints from thermodynamic-geometric loops to complex electronic-vibrational couplings, highlighted by significant correlation shifts in pairs like C-$HOMO$, $HOMO$-r^2, and μ-$ZPVE$.

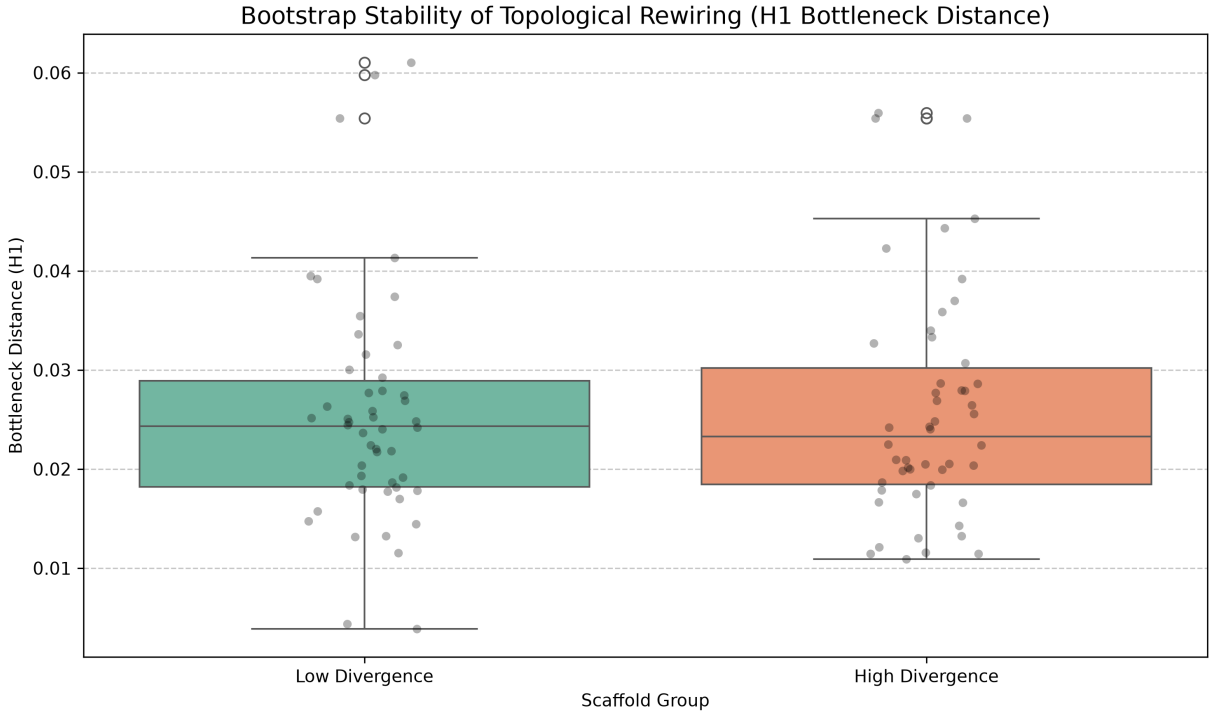

Figure 5: Bootstrap stability of topological rewiring measured by the H_1 Bottleneck distance for low divergence (e.g., cyclopropane) and high divergence (e.g., pyridine) scaffold groups. The boxplots summarize $B = 50$ bootstrap iterations, illustrating that the observed topological shifts in molecular property networks are robust to sample size. Both groups maintain stable mean H_1 distances (0.0255 ± 0.0116 and 0.0258 ± 0.0114, respectively) with consistent 95% confidence intervals, confirming that the quantification of network deformation relative to the acyclic baseline is a reliable indicator of structural organization rather than a sampling artifact.

The application of persistent homology provides a rigorous, coordinate-free method to capture these higher-order constraints, proving highly stable against sample size variations. This framework not only deepens our fundamental understanding of how spatial confinement and chemical composition govern the coherent physical state of molecules, but it also offers a scalable, mathematically grounded approach for analyzing complex property landscapes across broader chemical spaces.