Persistent Homology of Distance Correlation Networks to Map Thermodynamic-Geometric Rewiring

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The joint structure of molecular properties is often obscured by stoichiometric scaling, making it difficult to understand how molecular architecture reorganizes global physical states independently of molecular size. To isolate this underlying chemical organization, we systematically quantify the topological deformation of non-linear property-dependency networks across distinct structural families. Using the QM9 dataset, we train XGBoost regressors on basic atom counts to extract size-neutralized residuals for twelve non-redundant quantum chemical properties. We partition the molecules into acyclic, monocyclic, and polycyclic families based on their structural scaffolds and compute distance correlation matrices for these residuals. Treating these dependency matrices as metric spaces, we apply persistent homology to extract topological features and quantify network rewiring using Bottleneck distances. Our analysis evaluates how topological complexity adapts to structural constraints and measures the extent of network rewiring between families. Additionally, we assess the robustness of these topological features against latent geometric artifacts and examine the mechanistic role of specific structural descriptors in driving shifts in property couplings.

1 Introduction

Understanding the mapping between molecular architecture and physical properties is a central goal in chemical physics and materials design. While extensive efforts have been dedicated to predicting individual quantum chemical properties from molecular structure, the joint distribution of these properties remains less explored. Physical properties do not exist in isolation; they form a complex dependency network where shifts in electronic structure inherently couple with thermodynamic, spatial, and rotational characteristics. Mapping this network is crucial for understanding how global physical states reorganize in response to structural constraints.

1.1 The challenge of stoichiometric scaling

A major obstacle in analyzing the joint structure of molecular properties is the overwhelming influence of stoichiometric scaling. In comprehensive chemical spaces such as the QM9 dataset, extensive properties like total energy, zero-point vibrational energy, and atomization energy scale almost deterministically with molecular size and atom counts. This stoichiometric baseline obscures the underlying chemical organization, making it difficult to isolate how specific architectural features—such as ring fusion, unsaturation, or branching—modulate property relationships independently of simple mass or composition effects.

Furthermore, large-scale computational datasets often harbor latent geometric artifacts or consistency failures that can distort traditional linear correlation analyses. To accurately capture how molecular architecture rewires physical states, it is necessary to employ a robust, coordinate-free approach that can quantify non-linear property dependencies while remaining insensitive to localized geometric noise.

1.2 Topological mapping of property networks

To systematically isolate and quantify this underlying chemical organization, we introduce a framework based on the persistent homology of distance correlation networks. By training non-linear regressors solely

on basic atom counts and linearity indicators, we extract size-neutralized property residuals. These residuals define a normalized property space where any observed coupling represents a genuine physical dependency rather than a trivial consequence of molecular scaling.

We partition the chemical space into distinct structural families based on their molecular scaffolds—ranging from acyclic molecules to complex polycyclic systems. For each family, we compute distance correlation matrices of the property residuals to capture the exact strengths of non-linear couplings. By treating these dependency matrices as metric spaces, we apply persistent homology, a mathematical tool from algebraic topology, to extract robust topological features such as multi-property loops (H_1 features). This approach allows us to track the deformation of the property network itself as structural constraints are applied.

1.3 Overview of the study

In this work, we track the topological deformation of non-linear property-dependency networks across structural families in the QM9 dataset. We demonstrate that topological complexity increases significantly with structural constraint, with polycyclic architectures exhibiting highly persistent multi-property couplings compared to their acyclic and monocyclic counterparts.

By computing Bottleneck distances between persistence diagrams, we directly measure the extent of network rewiring. We validate these topological shifts against intra-family statistical noise to ensure significance, and we demonstrate that the extracted homological features are robust against the dataset's latent geometric artifacts. Finally, we ground these abstract topological features in physical chemistry by identifying specific mechanistic drivers. We show how structural descriptors, such as carbon hybridization and ring counts, directly reorganize key axes of the property network, such as the fundamental coupling between zero-point vibrational energy and heat capacity.

2 Methods

2.1 Data preparation and structural stratification

The QM9 dataset was initially deduplicated based on SMILES strings, removing 83 duplicate structures to prevent data leakage during subsequent family comparisons. The analysis focused on twelve non-redundant quantum chemical properties: 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 ($ZPVE$), total energy (U_0), heat capacity (C_v), and atomization energy ($U_{0,atom}$). To address the extreme heavy tail of the rotational constant A, a $\log(A+1)$ transformation was applied. Additionally, a binary indicator variable was created for the nine linear molecules (where $A = 0$ and $B = C$) to explicitly capture their distinct physical state and ensure models could distinguish transformed zeros from near-zero values of non-linear molecules.

Molecules were parsed into RDKit objects to extract exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. To systematically evaluate the effect of molecular architecture, the dataset was stratified into four chemically meaningful families based on their Murcko Scaffolds: Acyclic (no ring scaffold), Saturated Monocycle, Unsaturated Monocycle, and Polycyclic (fused or bridged ring scaffolds).

2.2 Stoichiometric neutralization

To eliminate the dominant effect of stoichiometric scaling and isolate the underlying chemical organization, XGBoost regressors were trained to predict the twelve properties using only the five atom counts and the linearity indicator as features. Out-of-fold predictions from a k-fold cross-validation scheme were used to compute residuals for every molecule. These residuals were subsequently standardized (z-scored) using the mean and standard deviation calculated from the entire dataset. This global standardization ensures that downstream analyses reflect true shifts in variance and scale across structural families, defining a size-neutralized residual property space where correlations represent genuine physical dependencies.

2.3 Distance correlation network construction

Because distance correlation ($dCor$) is sensitive to sample size, equal-sized random sub-samples were drawn from each structural family to ensure unbiased comparisons. For each family's sub-sample, the pairwise distance correlation matrix of the twelve standardized property residuals was computed. These matrices capture the exact strengths of non-linear couplings between properties. Each distance correlation matrix was then converted into a metric distance matrix using the transformation $D = 1 - dCor$, treating the property-dependency network as a formal metric space.

2.4 Persistent homology and topological rewiring

To extract the topological features of the property networks, the 12×12 metric distance matrices were used to construct Vietoris-Rips complexes via the `gudhi` library. Persistent homology was computed to track the evolution of connected components (H_0) and multi-property loops (H_1) across filtration values. The persistence diagrams for each structural family were extracted, with a specific focus on H_1 features that represent robust, closed cycles of property dependencies. The extent of topological rewiring between structural families was quantified by computing the Bottleneck distance between their respective H_1 persistence diagrams, providing a direct measure of how the global physical state deforms under structural constraints.

2.5 Stability validation and mechanistic analysis

To ensure that the observed topological rewiring represents a genuine chemical phenomenon rather than statistical noise, an intra-family stability check was performed. Equal-sized random sub-samples were repeatedly drawn from within the same structural family, and the Bottleneck distances between their persistence diagrams were computed to establish a null distribution (noise floor). Inter-family distances were then compared against this maximum noise floor to determine statistical significance.

Furthermore, to verify that the topological features were not driven by latent geometric artifacts—such as the known geometric consistency failures in the QM9 dataset—an Isolation Forest algorithm was applied to the residuals to identify extreme outliers. Persistence diagrams were recomputed after excluding these outliers to assess the robustness of the homological features. Finally, to ground the abstract topology in physical chemistry, the molecules' property residuals were correlated with specific structural descriptors, including ring counts, rotatable bonds, and the fraction of sp^3 hybridized carbons (F_{sp^3}). This mechanistic investigation aimed to elucidate the physical drivers behind the rewiring of key network axes, such as the coupling between zero-point vibrational energy and heat capacity.

3 Results

3.1 Stoichiometric neutralization and the residual property space

The primary challenge in analyzing the joint structure of molecular properties is the overwhelming dominance of molecular size, which drives the variance of extensive properties. To isolate the underlying chemical organization, we first evaluated the performance of XGBoost regressors trained solely on basic atom counts (C, N, O, F, and H) and a linearity indicator. As expected, the baseline stoichiometric model achieved near-perfect reconstruction for extensive thermodynamic and energetic properties, including total energy (u_0, $R^2 = 0.9970$), zero-point vibrational energy ($zpve$, $R^2 = 0.9972$), and atomization energy ($u_{0,atom}$, $R^2 = 0.9886$). Isotropic polarizability (α) also exhibited high stoichiometric dependence ($R^2 = 0.8755$). In contrast, properties heavily dependent on 3D geometry and bond topology, such as the rotational constants (A, B, C) and dipole moment (μ), showed significantly lower stoichiometric dependence, with R^2 values ranging from 0.11 to 0.30.

By extracting and standardizing the out-of-fold residuals from these models, we defined a 12-dimensional residual property space. This space effectively neutralizes the stoichiometric baseline, ensuring that any observed correlations represent genuine physical couplings conditioned on molecular size rather than trivial scaling artifacts.

3.2 Topological rewiring across structural families

To understand how molecular architecture reorganizes these physical states, we partitioned the dataset into four structural families based on their Murcko Scaffolds: Acyclic, Saturated Monocycle, Unsaturated Monocycle, and Polycyclic. For each family, we constructed a property-dependency network using the pairwise distance correlation ($dCor$) of the residuals. As illustrated in Figure 1, the correlation patterns shift markedly across the families, indicating a fundamental reorganization of the property space as structural constraints increase.

Treating these dependency matrices as metric spaces, we applied persistent homology to identify H_1 loops—sets of properties that are mutually coupled, forming closed topological features. The topological complexity of the property network was found to increase significantly with structural constraint (Figure 2). The Acyclic and Unsaturated Monocycle families exhibited lower topological complexity, featuring two H_1 loops each with maximum persistences of 0.0640 and 0.0267, respectively. The Saturated Monocycle family contained a single loop with a maximum persistence of 0.0872. In stark contrast, the Polycyclic family exhibited three distinct H_1 loops and a maximum persistence of 0.1341. This substantial increase in persistence suggests that the fusion and bridging of rings create much stronger and more robust multi-property couplings than those found in simpler architectures.

To quantify the extent of this network rewiring, we computed the Bottleneck distance between the H_1 persistence diagrams of the different families. The topological distance between the Acyclic and Unsaturated Monocycle families was relatively small (0.0320). However, comparisons involving the Polycyclic family yielded much larger distances: 0.0670 for both Acyclic vs. Polycyclic and Unsaturated Monocycle vs. Polycyclic. These larger distances confirm a substantial deformation of the property network induced by polycyclic constraints.

3.3 Stability and robustness of topological features

To ensure that the observed topological rewiring is a genuine chemical phenomenon and not an artifact of sampling variance or latent geometric noise, we performed rigorous stability validations. Bootstrap analysis ($N = 1000$, 50 iterations) was used to establish an intra-family noise floor. The mean intra-family Bottleneck distance ranged from 0.0247 for Unsaturated Monocycles to a maximum of 0.0422 for the Polycyclic family.

As shown in Figure 3, the inter-family distances comparing the Polycyclic family to the Acyclic and Unsaturated Monocycle families (0.0670) significantly exceed this maximum noise floor, yielding a Significance Ratio of 1.59. This confirms that the topological rewiring is statistically significant relative to the internal variance of the structural families.

Furthermore, we assessed the sensitivity of these topological features to the known geometric consistency failures in the QM9 dataset. Using an Isolation Forest algorithm (with a 2.3% contamination proxy), we identified 3,078 potential geometric outliers. Recomputing the persistence diagrams after removing these outliers resulted in minimal shifts: the Bottleneck distance between the original and inlier-only diagrams was as low as 0.0098 for Unsaturated Monocycles and remained below 0.028 across all families. This demonstrates that the extracted H_1 loops are robust features of the bulk molecular population and are not driven by extreme geometric artifacts.

3.4 Mechanistic drivers of property couplings

To ground the abstract topological features in physical chemistry, we investigated the mechanistic drivers behind the rewiring of the property networks. Correlation analysis between structural descriptors and property residuals revealed several key relationships (Figure 4). For instance, the number of rotatable bonds correlates positively with spatial extent residuals (res_r2, $r = 0.5311$), while ring count correlates negatively with res_r2 ($r = -0.4883$) and positively with LUMO energy residuals (res_lumo, $r = 0.4220$). Most notably, ring count exhibits a strong negative correlation with heat capacity residuals (res_cv, $r = -0.5814$). This suggests that ring-closure restricts vibrational modes beyond what is accounted for by the stoichiometric baseline, thereby reorganizing the thermodynamic sector of the property network.

A central axis of the rewired property network is the coupling between zero-point vibrational energy ($zpve$) and heat capacity (c_v). We found that the strength of this coupling is highly sensitive to molecular saturation, parameterized by the fraction of sp^3 hybridized carbons (F_{sp^3}). For molecules with low saturation

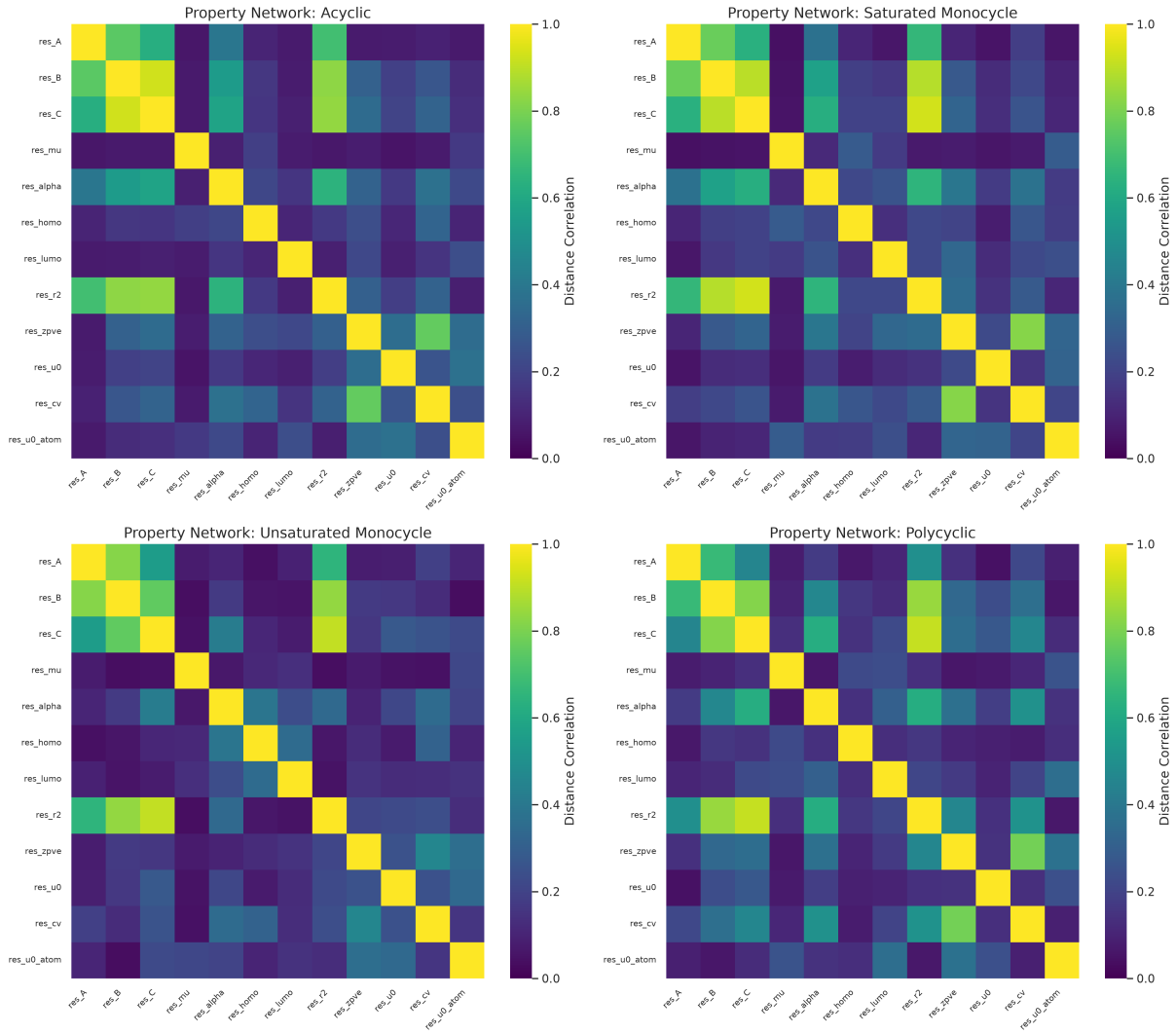

Figure 1: Distance correlation ($dCor$) matrices of twelve standardized property residuals across four molecular structural families (Acyclic, Saturated Monocycle, Unsaturated Monocycle, and Polycyclic). By extracting residuals for properties such as total energy (u_0) and atomization energy ($u_{0,atom}$) from stoichiometric baselines, these matrices capture the genuine physical state of the molecules conditioned on size. The heatmaps define the property-dependency networks whose varying correlation patterns illustrate the topological rewiring driven by molecular architecture. Moving from acyclic to highly constrained polycyclic scaffolds reveals a fundamental reorganization of the property space, characterized by the emergence of stronger multi-property couplings. This structural dependence is exemplified by the shifting correlations between thermodynamic and vibrational regimes, such as the coupling between heat capacity (`res_cv`) and zero-point vibrational energy (`res_zpve`), demonstrating how ring fusion and saturation fundamentally alter the joint structure of molecular properties.

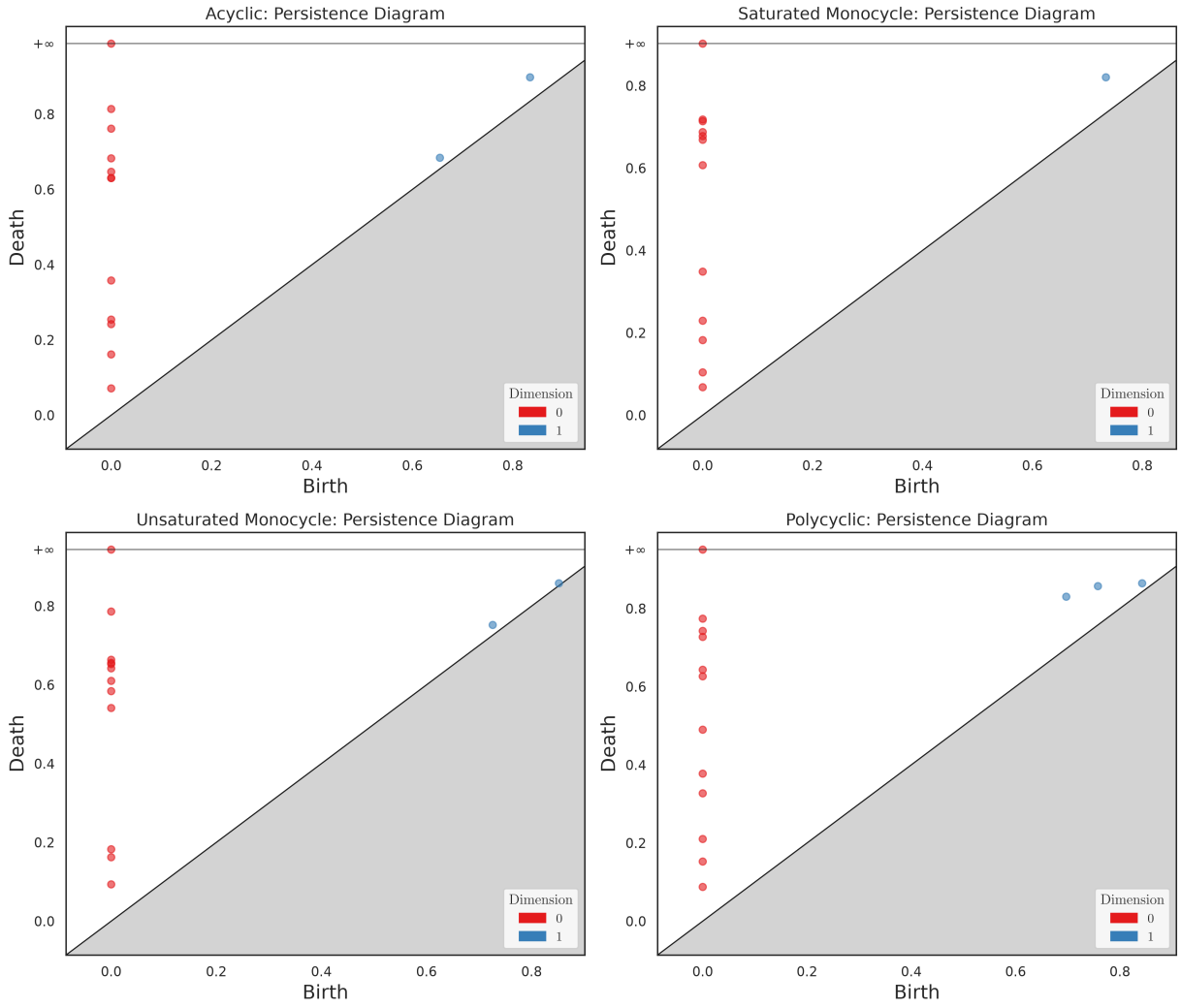

Figure 2: Persistence diagrams of property-dependency networks for four molecular structural families, displaying topological features in dimension 0 (red components) and dimension 1 (blue H_1 loops). The networks are constructed from pairwise distance correlations of stoichiometric property residuals. The Polycyclic family exhibits the highest topological complexity, featuring three H_1 loops and a maximum persistence of 0.1341, compared to fewer loops and lower persistence in the other families (such as 0.0267 for Unsaturated Monocycles). This increased complexity and higher maximum persistence demonstrate that the fusion and bridging of rings in polycyclic architectures create significantly stronger and more robust multi-property couplings than those found in acyclic or simple monocycle systems.

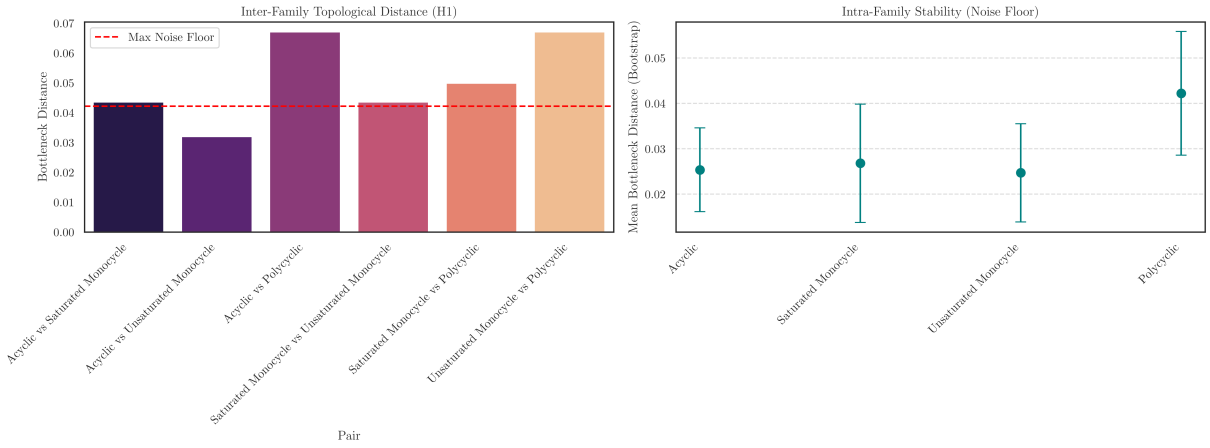

Figure 3: Comparison of inter-family topological distances and intra-family stability for property-dependency networks across molecular structural families. The left panel displays the Bottleneck distance between H_1 persistence diagrams for pairs of families, with a red dashed line marking the maximum intra-family noise floor. The right panel illustrates this noise floor via bootstrap sampling, showing the mean intra-family Bottleneck distance, which reaches a maximum of 0.0422 for the Polycyclic family. Distances for pairs comparing the Polycyclic family to Acyclic and Unsaturated Monocycle families (0.0670) significantly exceed this maximum noise floor. This yields a significance ratio of 1.59, confirming that the transition to polycyclic architectures induces a statistically significant topological rewiring and a fundamental reorganization of property residual couplings that cannot be attributed to internal family variance.

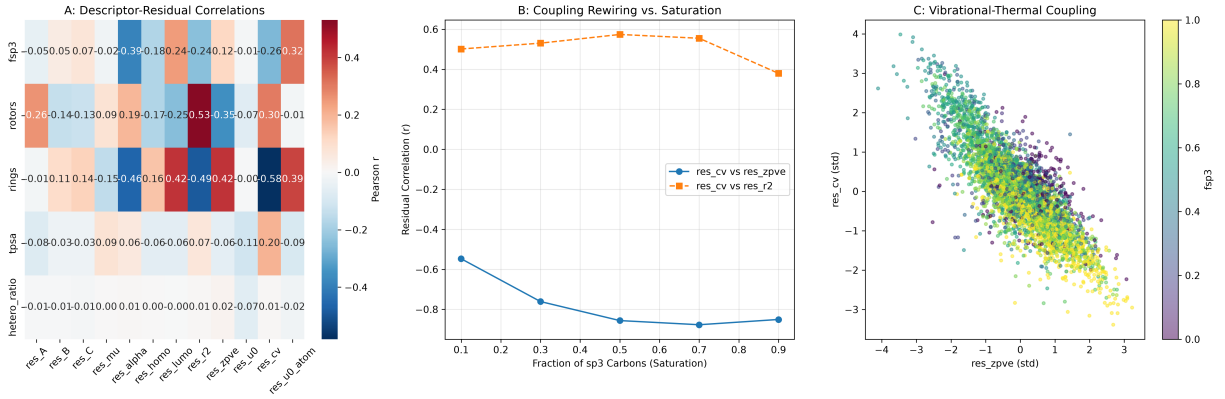

Figure 4: Mechanistic drivers of property couplings in the residual space. (A) Pearson correlations between chemical descriptors and stoichiometric property residuals. The heatmap reveals that structural features like ring count strongly drive thermodynamic and electronic residuals, such as the strong negative correlation with heat capacity residuals (`res_cv`) which suggests ring-closure restricts vibrational modes beyond stoichiometric scaling. (B) The dependence of specific residual couplings on molecular saturation (fraction of sp^3 carbons, F_{sp^3}). The coupling between `res_cv` and zero-point vibrational energy (`res_zpve`) transitions from a moderate to a very strong negative correlation as carbon hybridization increases. (C) Scatter plot of standardized `res_zpve` and `res_cv` residuals colored by F_{sp^3}, visually confirming the saturation-dependent tightening of this vibrational-thermal coupling. Together, these panels demonstrate how underlying bond topology and molecular architecture mechanistically rewire the relationships between fundamental energy levels and thermal responses within the property-dependency networks.

($F_{sp^3} \in [0, 0.2]$), the correlation between *res_cv* and *res_zpve* is moderate ($r = -0.5466$). However, for highly saturated molecules ($F_{sp^3} \in [0.6, 0.8]$), this coupling tightens dramatically into a very strong negative correlation ($r = -0.8769$).

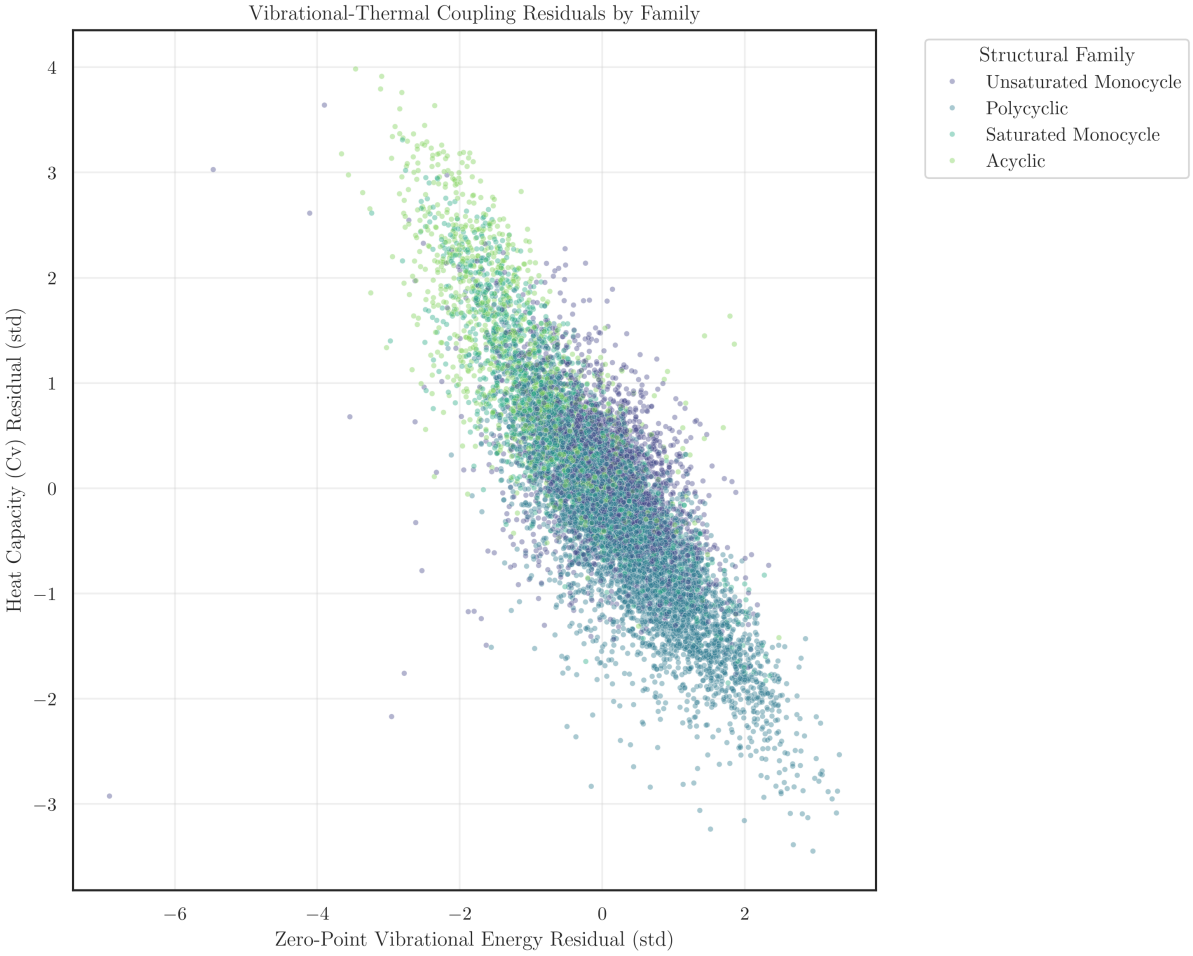

Figure 5: Scatter plot of standardized zero-point vibrational energy (*zpve*) and heat capacity (c_v) residuals, categorized by structural family. The strong negative correlation illustrates a central axis of the property-dependency network after removing stoichiometric baseline effects. The distribution shifts distinctly across structural families: acyclic molecules predominantly exhibit positive c_v and negative *zpve* residuals, while polycyclic architectures occupy the opposite extreme. This separation demonstrates that ring-closure restricts vibrational modes beyond stoichiometric expectations, fundamentally rewiring the coupling between fundamental energy levels and thermal response.

This transition, visualized across structural families in Figure 5, demonstrates how underlying bond topology dictates the relationship between fundamental energy levels and thermal response. Acyclic molecules predominantly exhibit positive c_v and negative *zpve* residuals, whereas polycyclic architectures occupy the opposite extreme. The polycyclic scaffold provides the most constrained environment for these couplings, directly leading to the highly persistent, complex topological loops observed in the persistent homology analysis.

4 Conclusions

In this work, we introduced a coordinate-free topological framework to systematically map how molecular architecture reorganizes the joint structure of physical properties. By neutralizing the dominant effects of stoichiometric scaling through gradient-boosted atom-count residuals, we successfully isolated the underlying chemical organization of the QM9 dataset. This residual property space allowed us to evaluate genuine physical dependencies that are otherwise obscured by trivial molecular scaling.

Our analysis of distance correlation networks revealed that the topological complexity of property dependencies inherently scales with structural constraint. Using persistent homology, we demonstrated that polycyclic architectures exhibit significantly stronger and more robust multi-property couplings—captured as highly persistent H_1 loops—compared to acyclic or simple monocyclic systems. The computed Bottleneck distances confirmed a fundamental rewiring of the property space as molecular scaffolds become more constrained. Rigorous stability validations established that this topological deformation is statistically significant relative to intra-family noise and remains highly robust against the dataset's latent geometric artifacts.

Crucially, we grounded these abstract topological shifts in concrete physical chemistry. By correlating property residuals with structural descriptors, we identified the mechanistic drivers of the network rewiring. We showed that structural features such as ring closure and carbon hybridization (F_{sp^3}) directly modulate the thermodynamic and vibrational sectors of the property space. Most notably, the transition from acyclic to polycyclic architectures dramatically tightens the coupling between zero-point vibrational energy ($zpve$) and heat capacity (c_v), demonstrating how underlying bond topology restricts vibrational modes beyond stoichiometric expectations.

Ultimately, this topological approach provides a robust methodology for characterizing global physical states independently of molecular size. By translating non-linear property dependencies into metric spaces, we offer a new lens for understanding how chemical architecture dictates physical behavior. This framework not only enhances our fundamental understanding of structure-property relationships but also provides a rigorous tool for evaluating the physical consistency of complex chemical spaces and generative molecular models.