Topological Reorganization of the Molecular Property Manifold: The Decoupling of HOMO Energy in Highly Strained Systems

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Determining how macroscopic molecular architecture—such as ring strain and structural complexity—dictates the relationships between microscopic electronic and geometric properties remains a significant computational challenge. We address this by mapping the topological reorganization of a 12-dimensional molecular property manifold across varying levels of structural confinement. Using the QM9 dataset, we first neutralize compositional size effects by extracting standardized residuals from atom-count regressions. We then stratify the molecules into complexity bins based on ring strain and fusion density, employing Gaussian Graphical Models and persistent homology to quantify global manifold drift. Our analysis uncovers a systematic rewiring of property dependencies as molecules transition from simple acyclic chains to highly strained polycyclic systems. Notably, we test the hypothesis that HOMO energy integrates into a structural hub under extreme spatial confinement, finding instead that it progressively decouples. The partial correlation strength of the HOMO residual peaks at intermediate complexity before sharply declining, as its couplings to global features like isotropic polarizability and spatial extent nearly vanish in highly complex systems. Validated against scaffold hold-outs and geometric outliers, our findings demonstrate that extreme structural strain forces HOMO energy to isolate from the global geometric-vibrational network, becoming an orthogonal feature driven by localized effects.

1 Introduction

Determining how macroscopic molecular architecture dictates the relationships between microscopic electronic and geometric properties remains a significant computational challenge. In large chemical spaces, the covariance among molecular properties is often dominated by simple stoichiometric scaling and compositional size effects. Consequently, the intrinsic influence of structural topology—such as ring strain, spatial confinement, and fusion density—on the joint organization of these properties is frequently obscured. Understanding whether extreme structural strain forces localized electronic states to couple with global molecular physics is critical for mapping the boundaries of chemical property spaces and designing molecules with targeted multidimensional profiles.

A central question in this domain is how the Highest Occupied Molecular Orbital (HOMO) energy behaves under extreme spatial confinement. In simple, acyclic molecules, electronic states often exhibit well-defined couplings to global geometric and vibrational features. It is a plausible hypothesis that as molecules become highly strained and polycyclic, the HOMO energy might be forced to integrate into a global geometric-vibrational axis, effectively acting as a highly connected structural hub within the property manifold. Rigorously testing this hypothesis, however, requires a framework capable of isolating topological effects from compositional variance and quantifying the local dimensionality of the property space across continuous structural boundaries.

In this work, we address this challenge by mapping the topological reorganization of a 12-dimensional molecular property manifold across varying levels of structural confinement. Using the QM9 dataset, we first neutralize compositional size effects by regressing properties against exact atom counts and extracting standardized residuals. This stoichiometric baseline, combined with transformations to handle extreme rotational variance, ensures that the resulting manifold captures purely structural and topological dependencies. We then stratify the dataset into complexity bins based on RDKit-derived metrics of ring strain and fusion density.

To quantify the macroscopic reorganization of the property space, we employ Gaussian Graphical Models to estimate precision matrices across the complexity gradient, capturing the evolving partial correlations among the residual properties. We complement this network approach with persistent homology, analyzing the Bottleneck distances of Vietoris-Rips complexes to measure global manifold drift. This dual approach allows us to track the exact macroscopic architectures that force localized electronic states to rewire their dependencies, inherently absorbing latent geometric artifacts as valid high-strain features rather than discarding them as regression outliers.

Our analysis uncovers a systematic rewiring of property dependencies as molecules transition from simple acyclic chains to highly strained polycyclic systems. Crucially, we find that the HOMO energy does not integrate into a structural hub under extreme spatial confinement. Instead, it progressively decouples. The partial correlation strength of the HOMO residual peaks at intermediate complexity before sharply declining, as its couplings to global features—such as isotropic polarizability and spatial extent—nearly vanish in highly complex systems. Validated against scaffold hold-outs and geometric sensitivity checks, our findings demonstrate that extreme structural strain forces the HOMO energy to isolate from the global geometric-vibrational network, transitioning into an orthogonal feature driven by localized effects.

2 Methods

2.1 Data curation and redundancy reduction

To isolate the intrinsic influence of structural topology from compositional variance, we first processed the QM9 dataset to remove redundant and non-physical representations. We deduplicated 83 SMILES strings that appeared multiple times in the dataset. The 3,054 molecules previously flagged in QM9 for geometric artifacts were intentionally retained; without 3D coordinates to definitively classify them as errors, we hypothesized that these artifacts might represent valid high-strain features rather than mere regression outliers.

To construct a non-redundant property space, we removed thermodynamically redundant features (internal energy at 298 K, enthalpy at 298 K, free energy at 298 K, and their atomization equivalents) as well as the HOMO-LUMO gap, which is a definitional identity. The LUMO energy was retained to preserve the electronic boundary state. To handle the extreme heavy-tailed distribution of the rotational constant A, we applied a logarithmic transformation, $A' = \log(A + 1)$. Finally, we defined a binary indicator variable for linear molecules, capturing the specific physical state where $A = 0$ and $B = C$.

2.2 Establishing the stoichiometric baseline

Because macroscopic molecular architecture is often confounded by simple compositional size effects, we established a rigorous stoichiometric baseline. Each of the 12 remaining numeric properties was regressed against the exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen, alongside the linear molecule indicator. For a given property y and molecule i, the regression model is defined as:

$$y_i = \beta_0 + \beta_C n_{C,i} + \beta_N n_{N,i} + \beta_O n_{O,i} + \beta_F n_{F,i} + \beta_H n_{H,i} + \gamma L_i + \epsilon_i \tag{1}$$

where $n_{X,i}$ represents the count of atom type X, L_i is the linear indicator, and ϵ_i is the residual.

We extracted the residuals (ϵ_i) from these regressions to form a stoichiometry-independent representation of the molecular properties. To mitigate the disproportionate impact of extreme outliers, the residuals were winsorized at the 1st and 99th percentiles and subsequently standardized to zero mean and unit variance. These standardized residuals constitute the 12-dimensional property manifold analyzed in all subsequent steps.

2.3 Structural complexity stratification

To map the manifold across continuous structural boundaries, we computed semi-continuous complexity and strain metrics for each molecule using RDKit. Rather than relying on simple ring counts, we quantified structural confinement using the average ring size, the fraction of atoms participating in rings, the ring

fusion density (number of bridged or fused ring atoms), and a specific strain proxy defined as the fraction of heavy atoms in 3- or 4-membered rings. We also extracted the Murcko scaffold for every molecule to track topological families.

The dataset was then stratified along these metrics into eight distinct bins of increasing structural complexity, ranging from simple acyclic chains (Bin 0) to highly strained polycyclic systems (Bin 7). To ensure statistical stability for downstream covariance estimation, we enforced a minimum threshold of 1,000 molecules per bin, merging adjacent strata where necessary.

2.4 Gaussian graphical models and network analysis

To quantify the local dimensionality and internal dependencies of the property manifold within each complexity bin, we constructed Gaussian Graphical Models (GGMs). For each bin, we estimated the precision matrix (inverse covariance matrix, $\Theta = \Sigma^{-1}$), which captures the conditional dependencies (partial correlations) among the 12 residual properties. We employed the Graphical Lasso algorithm, utilizing a fixed regularization parameter (λ) across all bins. Maintaining a constant λ ensures that any observed differences in the precision matrices reflect true structural reorganization rather than variations in the sparsity penalty.

From the estimated precision matrices, we extracted undirected partial correlation networks. To test the hypothesis that the HOMO energy integrates into a structural hub under extreme spatial confinement, we computed the partial correlation strength (the absolute sum of edge weights) and degree centrality for the HOMO residual (res_homo). We specifically monitored the evolution of its partial correlations with global geometric and vibrational features, such as isotropic polarizability (res_α) and electronic spatial extent (res_r^2).

2.5 Topological analysis via persistent homology

To measure the macroscopic drift of the property manifold across structural boundaries, we computed the Frobenius norm of the difference between the precision matrices of different complexity bins. We generated a pairwise distance matrix to identify potential phase transitions in molecular organization.

We complemented this network approach with persistent homology to capture higher-order topological reorganization. Using the 12-dimensional standardized residuals, we constructed Vietoris-Rips complexes for each complexity bin. We computed the persistent homology for 0-dimensional (H0, connected components) and 1-dimensional (H1, cycles) features. To quantify the topological divergence between acyclic and highly strained systems, we calculated the Bottleneck distance (W_∞) between their respective persistence diagrams:

$$W_\infty(D_1, D_2) = \inf_{\gamma} \sup_{x \in D_1} ||x - \gamma(x)||_\infty \tag{2}$$

where γ ranges over all bijections between the persistence diagrams D_1 and D_2.

2.6 Validation and sensitivity analysis

To ensure that the observed manifold reorganization and the decoupling of the HOMO energy were not artifacts of specific molecular families or geometric outliers, we performed two rigorous validation checks. First, we conducted a leave-molecular-family-out validation on the highest complexity bin (Bin 7). We identified the top 10 most frequent Murcko scaffolds within this bin, removed all molecules sharing these scaffolds, and recomputed the GGMs, centrality metrics, and Frobenius norms.

Second, we performed a sensitivity analysis targeting the 3,054 potential geometric artifacts. We removed the top 1% to 5% of molecules exhibiting the highest residual variance in the regression space and re-ran the entire pipeline. This step verified that the transition of the HOMO energy toward isolation and the global manifold drift remained stable and were not disproportionately driven by extreme geometric outliers.

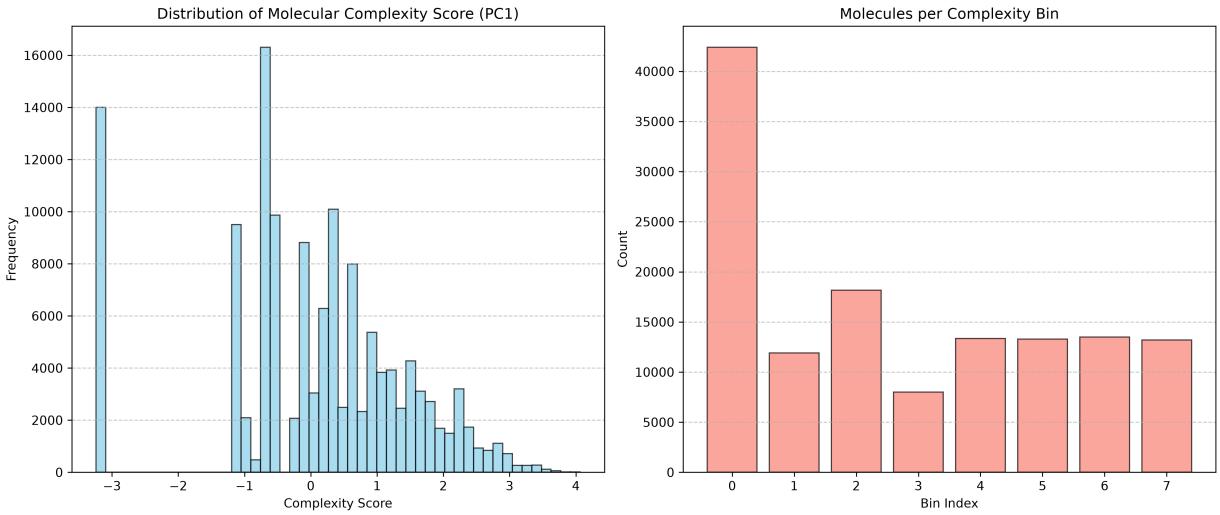

Figure 1: Distribution and stratification of the QM9 dataset based on structural complexity, defined by ring count, ring size, and fusion density. The left panel displays the continuous distribution of the molecular complexity score (PC1), while the right panel illustrates the molecule count across eight discrete complexity bins. Bin 0 represents the lowest-complexity, acyclic baseline and contains the largest population, whereas Bins 1 through 7 capture systems with increasing polycyclic strain and structural confinement. This binning strategy establishes the population framework required to analyze the topological reorganization of the property manifold, demonstrating how increasing structural complexity drives the decoupling of features like the HOMO energy residual from the global property network.

3 Results

3.1 Global manifold drift and topological reorganization

The joint structure of molecular properties in the QM9 dataset undergoes significant reorganization as molecules transition from simple, acyclic chains to highly strained, multi-cyclic systems. Following the extraction of stoichiometry-independent residuals to neutralize compositional size effects, we stratified the population into eight bins of increasing structural complexity, defined by ring count, ring size, and fusion density (Figure 1).

The overall organization of the property manifold exhibits a monotonic divergence from the low-complexity baseline. The Frobenius distance between the precision matrices of the lowest-complexity (Bin 0, acyclic) and highest-complexity (Bin 7, polycyclic/strained) systems reaches 9.80, indicating a substantial rewiring of the underlying partial correlation network (Figure 2, left).

This macroscopic drift is mirrored in the topological structure of the manifold as measured by Vietoris-Rips persistent homology. The Bottleneck distance for 0-dimensional persistence (H0), which captures the scales at which properties cluster together, shows a maximum distance of 0.307 between acyclic molecules and high-strain systems (Figure 3, left). This suggests that the grouping behavior of properties changes significantly; properties that are tightly coupled in simple molecules decouple or re-associate with different partners as structural confinement increases. In contrast, the 1-dimensional persistent homology (H1), which tracks circular dependencies among properties, maintains a steady state. The Bottleneck distance for H1 between the extremes of the complexity gradient is notably lower (0.031, Figure 3, right), suggesting that while the strengths of specific property couplings drift, the fundamental circular logic of the electronic-geometric relationship remains relatively stable.

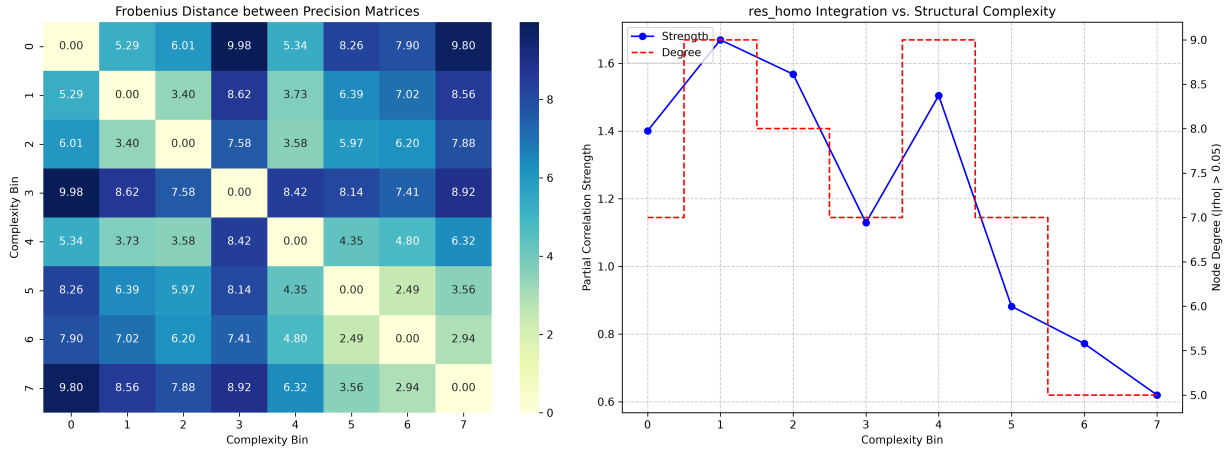

Figure 2: Topological reorganization of the molecular property manifold across structural complexity. (Left) The Frobenius distance between precision matrices across eight complexity bins reveals a monotonic divergence from acyclic (Bin 0) to highly strained polycyclic systems (Bin 7), reaching a maximum distance of 9.80. This indicates a substantial rewiring of the underlying partial correlation network. (Right) The network integration of the HOMO energy residual (`res_homo`) peaks at low complexity (Bin 1, partial correlation strength of 1.67, degree centrality of 9) and declines by over 60% in highly strained systems (Bin 7, strength of 0.62, degree of 5). Together, these trends demonstrate that increasing structural confinement fundamentally reorganizes the property manifold, forcing the HOMO energy to decouple from the global network and become an isolated, orthogonal feature rather than a structural hub.

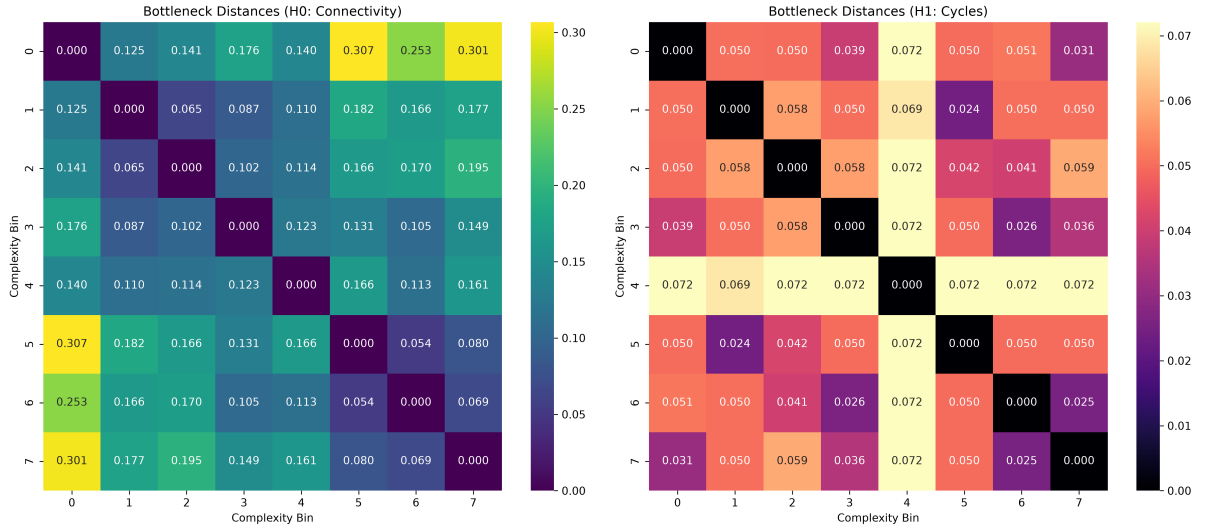

Figure 3: Heatmaps of Bottleneck distances for 0-dimensional (H0, left) and 1-dimensional (H1, right) persistent homology across eight molecular complexity bins. The H0 distances, which capture the scales at which properties cluster together, reveal a substantial topological drift, reaching a maximum distance of 0.307 between simple acyclic molecules (Bin 0) and highly strained systems. This demonstrates that the grouping behavior of molecular properties changes significantly with structural confinement, causing tightly coupled properties to decouple or re-associate. In contrast, the H1 distances, which track circular dependencies among properties, remain notably lower, with a distance of only 0.031 between the extremes of the complexity gradient (Bin 0 and Bin 7). Together, these panels indicate that while the strengths of specific property couplings drift considerably across the complexity gradient, the fundamental circular logic of the electronic-geometric relationship remains relatively stable.

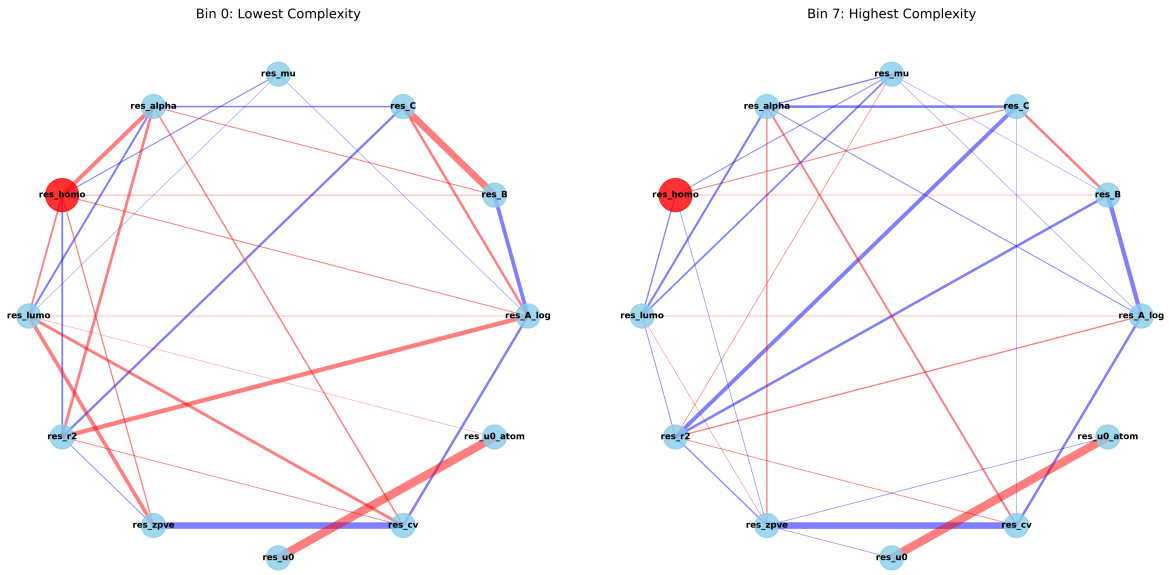

Figure 4: Partial correlation networks of the 12-dimensional property manifold comparing the lowest-complexity acyclic molecules (Bin 0, left) to the highest-complexity polycyclic systems (Bin 7, right). The graphs illustrate the topological reorganization of molecular properties across structural complexity. The HOMO energy residual (`res_homo`, highlighted in red) acts as an integrated hub in simple molecules, exhibiting strong couplings to properties such as isotropic polarizability (`res_alpha`) and electronic spatial extent (`res_r2`). As structural confinement and ring strain increase, the partial correlation strength of `res_homo` drops by more than 60%, and its connections to the global network substantially weaken. This visually demonstrates the decoupling of `res_homo`, revealing that in highly strained systems, the HOMO energy becomes an orthogonal feature governed by localized effects rather than global geometric-vibrational properties.

3.2 The HOMO residual: from integration to isolation

A primary discovery of this analysis is the trajectory of the HOMO energy residual (`res_homo`). Contrary to the hypothesis that the HOMO energy might integrate into a structural hub under extreme spatial confinement, the data reveals that its integration into the global property network peaks at intermediate levels of complexity and subsequently declines. In low-complexity molecules (Bin 1), `res_homo` exhibits its highest partial correlation strength (1.67) and degree centrality (9), indicating it is a well-integrated component of the molecular physics (Figure 2, right). As ring strain and fusion density increase toward Bin 7, the partial correlation strength of `res_homo` drops by more than 60% to 0.62, and its degree centrality falls to 5. This topological decoupling is visually evident in the evolution of the partial correlation networks (Figure 4).

Specific edge-level drift clarifies the nature of this decoupling (Figure 5). The coupling between `res_homo` and isotropic polarizability (`res_alpha`) is strong in acyclic molecules ($|\rho| = 0.471$) but essentially vanishes in the highest-complexity systems ($|\rho| = 0.028$). Similarly, the connection to electronic spatial extent (`res_r2`) weakens from $|\rho| = 0.208$ in Bin 0 to $|\rho| = 0.019$ in Bin 7. This evidence refutes the structural hub hypothesis, demonstrating instead that as molecules become more constrained and strained, the HOMO energy is governed increasingly by localized bonding environments or strain-specific electronic effects that are orthogonal to the global geometric-vibrational properties.

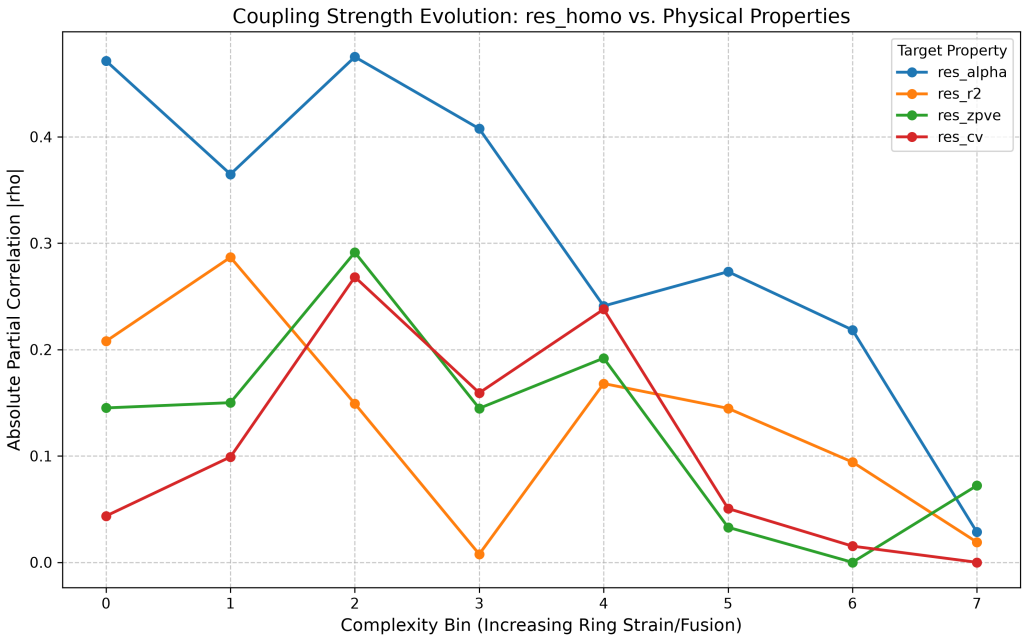

Figure 5: Evolution of the absolute partial correlation ($|\rho|$) between the HOMO energy residual (res_homo) and other physical property residuals (res_alpha, res_r2, res_zpve, res_cv) across eight bins of increasing structural complexity. The plot illustrates the progressive decoupling of res_homo from the global property network as molecules transition from simple acyclic chains (Bin 0) to highly strained polycyclic systems (Bin 7). Specifically, the strong initial couplings with isotropic polarizability (res_alpha) and electronic spatial extent (res_r2) weaken significantly, dropping from $|\rho| = 0.471$ and 0.208 in Bin 0 to 0.028 and 0.019 in Bin 7, respectively. This edge-level drift refutes the hypothesis that HOMO acts as a structural hub under spatial confinement; instead, it demonstrates that HOMO energy becomes an orthogonal feature, increasingly governed by localized or strain-specific effects rather than global geometric-vibrational properties.

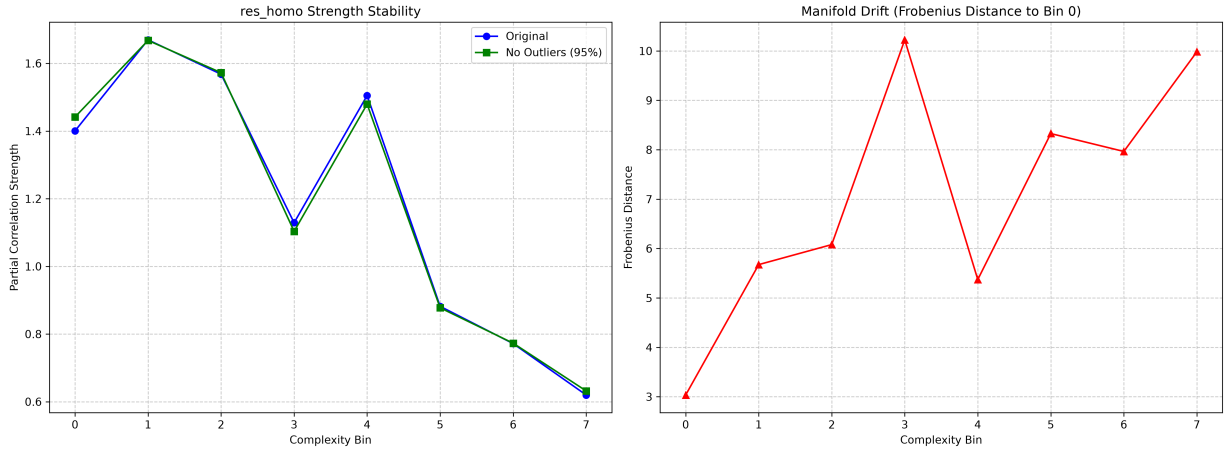

Figure 6: Robustness of manifold drift and HOMO residual isolation across structural complexity bins. The left panel shows the partial correlation strength of the HOMO residual (`res_homo`) peaking at low complexity (Bin 1) and declining toward highly strained systems (Bin 7). This decoupling trend is stable even after removing the 5% of molecules with the highest residual variance, confirming that HOMO isolation is a robust feature of the chemical space rather than a geometric artifact. The right panel illustrates the global manifold drift via the Frobenius distance of precision matrices relative to the acyclic baseline (Bin 0). The general increase in distance, reaching approximately 10 in Bin 7, highlights a substantial and systematic topological reorganization of the property manifold as molecular complexity increases.

3.3 Robustness and validation

The observed reorganization and the isolation of the HOMO energy are not artifacts of specific molecular families or geometric outliers (Figure 6). A scaffold hold-out validation performed on Bin 7—removing all molecules belonging to the ten most frequent Murcko scaffolds (e.g., bicyclo[2.2.1]heptane systems)—produced a `res_homo` centrality strength of 0.651, compared to the original 0.620. The Frobenius distance from the hold-out Bin 7 to the original Bin 0 remained high (10.03), confirming that the manifold drift is a general structural phenomenon rather than a bias introduced by overrepresented topologies.

Furthermore, sensitivity checks removing the 5% of molecules with the highest residual variance—representing potential geometric artifacts—showed that the centrality metrics remained highly stable. The `res_homo` strength in the cleaned Bin 7 was 0.632, nearly identical to the baseline result. This indicates that the transition of the HOMO energy toward isolation is a robust feature of the QM9 chemical space, independent of the 3,054 molecules previously flagged for geometric inconsistency.

4 Conclusions

In this study, we mapped the topological reorganization of a 12-dimensional molecular property manifold to determine how macroscopic structural architecture dictates the relationships between microscopic electronic and geometric properties. By establishing a rigorous stoichiometric baseline to neutralize compositional size effects, we isolated the intrinsic influence of ring strain and spatial confinement across the QM9 dataset. Our dual approach, utilizing Gaussian Graphical Models and persistent homology, reveals that the joint structure of molecular properties is fundamentally dynamic. As molecules transition from simple acyclic chains to highly strained polycyclic systems, the property manifold undergoes a systematic macroscopic drift, characterized by a substantial rewiring of partial correlation networks and a divergence in clustering behavior.

A central focus of this work was to test the hypothesis that the HOMO energy integrates into a global geometric-vibrational hub under extreme spatial confinement. Our findings definitively refute this assumption. Rather than acting as a highly connected structural node, the HOMO energy progressively decouples from the global property network as structural complexity increases. While simple molecules exhibit an

integrated manifold where electronic states are strongly coupled to global spatial extent and isotropic polarizability, these dependencies nearly vanish in highly strained architectures. The partial correlation strength of the HOMO residual peaks at intermediate complexity before sharply declining, forcing the HOMO energy to isolate into an orthogonal feature.

Rigorous validation through scaffold hold-outs and sensitivity checks confirms that this topological decoupling is a robust, intrinsic feature of the chemical space, independent of overrepresented molecular families or latent geometric artifacts. Ultimately, this work provides a quantitative map of how macroscopic molecular architecture modifies the joint organization of microscopic properties. By demonstrating that extreme structural strain forces localized electronic states to break away from global molecular physics and become governed by localized effects, these insights offer a critical foundation for navigating complex chemical property spaces and designing molecules with targeted, multidimensional profiles.