Topological Isolation of the HOMO Anomaly in the QM9 Property Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy often exhibits an anomalous decoupling from broader thermodynamic and geometric properties, complicating our understanding of global molecular physics. To identify the structural conditions that drive this decoupling, we analyze the QM9 dataset by constructing a 12-dimensional manifold of property residuals, conditioned on atom counts to remove stoichiometric determinism. By applying persistent homology and constructing Vietoris-Rips complexes on this residual point cloud, we detect topological voids that represent sparsely populated regions where standard property correlations collapse. Validating these features against a scaffold-shuffled null model, we isolate 96 void-forming molecules defined by suppressed isotropic polarizability and expanded frontier orbital gaps. These results confirm that topological isolation stems from specific local electronic environments rather than linear geometric artifacts. Finally, through the analysis of minimal structural mutation trajectories, we show that moderate chemical edits—averaging a Tanimoto similarity of 0.506—are sufficient to transition molecules out of these voids, successfully re-coupling their frontier orbitals to the global chemical manifold.

1 Introduction

The exploration of chemical space often relies on the assumption that molecular properties vary smoothly and coherently with structural changes. In large-scale quantum chemistry datasets such as QM9, macroscopic properties like zero-point vibrational energy and internal energy are heavily dominated by stoichiometric determinism. However, once these baseline atom-count dependencies are regressed out, the underlying joint structure of the property manifold reveals complex, non-linear behaviors that challenge standard predictive paradigms.

A prominent anomaly within this residual space is the behavior of the highest occupied molecular orbital (HOMO) energy. While thermodynamic and geometric properties—such as isotropic polarizability, heat capacity, and electronic spatial extent—exhibit strong inter-correlations and form a coherent global manifold, the HOMO energy frequently decouples from this structure. This orthogonal behavior complicates the development of unified physical models and obscures the relationship between local electronic environments and macroscopic spatial organization. Understanding the structural conditions that drive this decoupling, and conversely, the conditions that force frontier orbitals to re-couple with global molecular physics, remains a significant challenge.

To systematically isolate and characterize this anomaly, we analyze the complete 12-dimensional non-redundant property space of the QM9 dataset. We first extract atom-count residuals to eliminate stoichiometric determinism, isolating the intrinsic variations in molecular properties. Rather than relying on coarse scalar descriptors or artificially filtering latent density functional theory (DFT) artifacts, we employ persistent homology—a robust topological data analysis technique. By constructing Vietoris-Rips complexes on the 12-dimensional residual point cloud, we detect topological voids (features of the first homology group, H_1). In the context of a property manifold, these voids represent sparsely populated, physically reachable regions where standard property correlations collapse, effectively acting as boundaries or bottlenecks in chemical space.

In this work, we demonstrate that the topological isolation of the HOMO anomaly is driven by specific local electronic environments rather than linear geometric artifacts. By isolating the highest-persistence

topological features, we identify a distinct subset of 96 void-forming molecules characterized by suppressed isotropic polarizability and expanded frontier orbital gaps. We validate the statistical significance of this decoupling using a scaffold-shuffled null model, confirming that the topological isolation stems from precise chemical motifs governing lone-pair density and electronic saturation. Finally, through the analysis of minimal structural mutation trajectories, we show that moderate chemical edits are sufficient to transition molecules out of these voids, successfully re-coupling their frontier orbitals to the global chemical manifold.

2 Methods

2.1 Data ingestion and stoichiometric conditioning

To isolate the intrinsic variations in molecular properties from baseline size dependencies, we first processed the QM9 dataset to remove stoichiometric determinism. Following initial data ingestion and deduplication, we isolated 12 non-redundant property columns. Because macroscopic properties in QM9 are heavily dominated by the linear influence of atom counts, we regressed out the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. The resulting non-linear atom-count residuals (e.g., *res_homo*, *res_lumo*, *res_alpha*) formed the basis of our 12-dimensional residual manifold. To stabilize the variance across properties with disparate scales prior to topological analysis, a $\log(x + 1)$ transformation was applied where appropriate.

2.2 Topological data analysis and void identification

To detect regions where standard property correlations collapse, we employed persistent homology on the conditioned property space. To ensure uniform density and computational tractability, we initialized a topological subsampling routine using stratified selection across the primary structural axes. We then constructed Vietoris-Rips complexes directly on the 12-dimensional residual point cloud using the `gudhi` library.

By computing the persistence diagrams for these complexes, we extracted features of the first homology group (H_1), which correspond to one-dimensional topological voids or loops in the property manifold. We isolated the highest-persistence H_1 features and mapped the simplices forming these anomalous topological structures back to their RDKit SMILES representations. This purely shape-based approach allowed us to identify the specific subset of 96 void-forming molecules without relying on coarse scalar descriptors or artificially filtering latent density functional theory (DFT) artifacts.

2.3 Null modeling and statistical validation

To rigorously evaluate the statistical significance of the HOMO energy decoupling, we implemented a scaffold-shuffled null model alongside a Copula-based marginal distribution model. The null model was constructed by permuting the HOMO energy residual (*res_homo*) strictly within identical Murcko scaffolds. This localized permutation strategy preserves the local structural density and baseline geometric constraints of the molecules, ensuring that the null distribution accurately reflects the expected property coupling for structurally analogous compounds.

Furthermore, we conducted a sensitivity analysis on the simplices forming the primary voids to rule out geometric outliers. As a negative control, we explicitly checked the behavior of perfectly linear molecules within the dataset to confirm that the identified topological voids are intrinsic features of polyatomic, non-linear chemical space rather than geometric artifacts. Bootstrapped confidence intervals were computed for all baseline HOMO-LUMO conditional correlations.

2.4 Structural feature extraction and mutation trajectories

To bridge the local electronic environments with the macroscopic spatial organization observed in the residual manifold, we executed standard structural feature extraction and applied Principal Component Analysis (PCA) to establish a baseline structural manifold. We calculated exact lone-pair counts using explicit valence-state logic, distinguishing specific electronic environments such as amide versus amine nitrogens based on default valences.

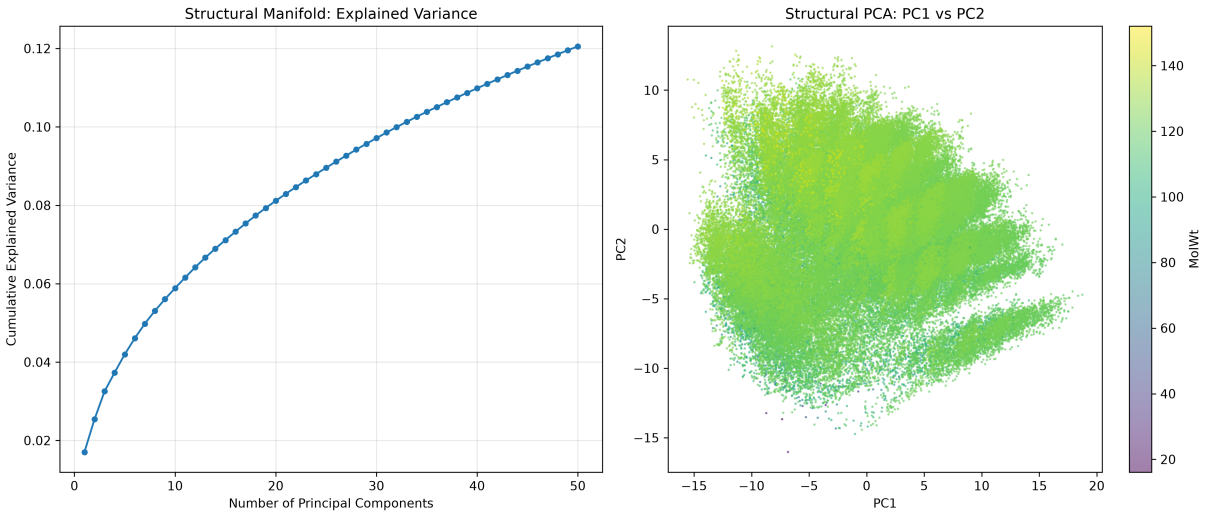

Figure 1: Principal component analysis of the structural descriptors for the QM9 dataset. The left panel displays the cumulative explained variance, showing that the first 50 components capture only $\approx 12\%$ of the variance, which highlights the extreme high dimensionality of the structural space. The right panel projects the molecules onto the first two principal components (PC1 and PC2), with points colored by molecular weight. This projection illustrates the distribution of molecular complexity, where PC1 serves as the primary measure of complexity used to track fluctuations in the distance correlation between res_{homo} and res_{lumo}. The clear gradient of molecular weight across the structural manifold visually demonstrates how standard representations conflate chemical organization with molecular size, reinforcing the necessity of using a size-decoupled residual manifold to successfully isolate topological voids and the HOMO anomaly.

To validate the H_1 voids and determine the structural conditions required for frontier orbitals to re-couple with the global manifold, we analyzed minimal structural mutation trajectories. For each void-forming molecule, we generated systematic minimal structural edits (e.g., single atom or bond swaps) and identified the nearest neighbor in chemical space residing outside the void, measured via Morgan fingerprint Tanimoto similarity. We then compared their property-space trajectories against a control path of non-void-separated molecules with equivalent structural edit distances. This allowed us to distinguish true topological bottlenecks from discrete chemical jumps. Finally, robust Partial Distance Correlation ($pdCor$) and standard distance correlation ($dCor$) metrics were computed across quantile-binned structural axes using a 14-core parallelization protocol to quantify the effect sizes of HOMO re-coupling.

3 Results

3.1 The residual manifold and the HOMO anomaly

Standard analysis of large-scale quantum chemistry datasets frequently conflates intrinsic chemical organization with molecular size. As established by our stoichiometric baseline, macroscopic properties such as zero-point vibrational energy ($zpve$) and internal energy at 0 K (u_0) are almost entirely determined by atom counts ($R^2 > 0.99$). To isolate the intrinsic variations in molecular properties, we operate exclusively on the 12-dimensional residual manifold, where the linear influence of carbon, nitrogen, oxygen, fluorine, and hydrogen counts has been regressed out. The necessity of this size-decoupled approach is highlighted by a Principal Component Analysis (PCA) of the raw structural descriptors (Figure 1), which demonstrates how standard representations are dominated by a gradient of molecular weight.

Within this 12-dimensional residual manifold, a striking decoupling emerges. While properties such as isotropic polarizability (res_{alpha}), heat capacity (res_{cv}), and electronic spatial extent (res_{r2}) participate in a coherent joint structure (distance correlations $dCor > 0.8$ for triads involving rotational constants, heat

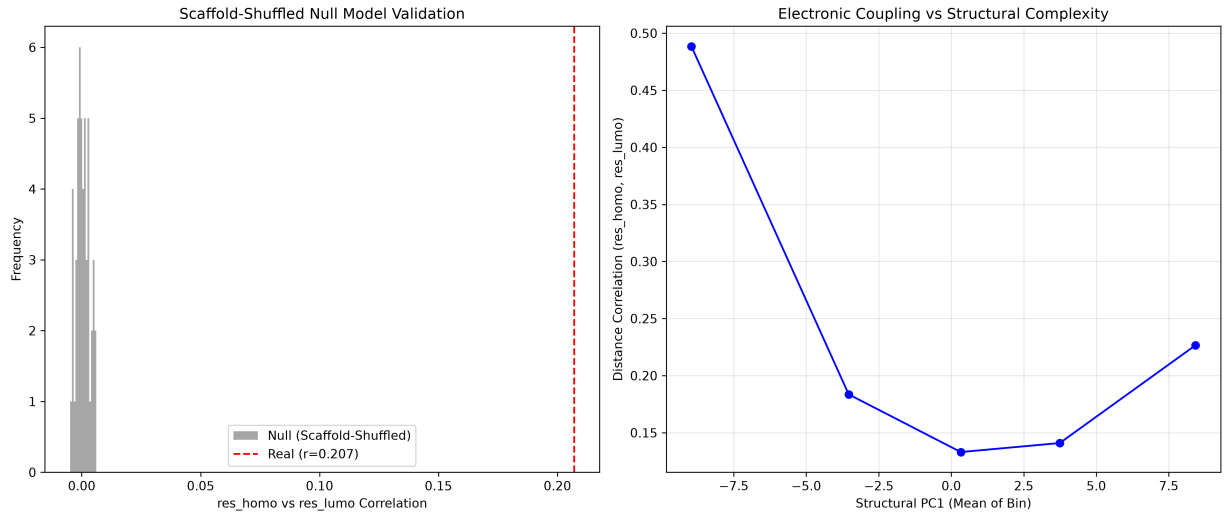

Figure 2: Validation of the topological isolation between the HOMO and LUMO energy residuals (res_{homo} and res_{lumo}). (Left) A scaffold-shuffled null model demonstrates that the observed global Pearson correlation of 0.207 (red dashed line) is significantly suppressed yet statistically non-zero compared to the null distribution, yielding a Z-score of 77.17. This confirms the weak coupling of the HOMO anomaly relative to the broader thermodynamic and geometric manifold. (Right) The distance correlation between res_{homo} and res_{lumo} fluctuates as a function of structural complexity, measured by the first principal component (PC1) of structural descriptors. The coupling peaks at 0.488 for the simplest structures and drops to a minimum of 0.133 for mid-complexity molecules, revealing that local bonding environments governing lone-pair density and electronic saturation effectively decouple frontier orbitals from the global spatial and thermodynamic organization.

capacity, and spatial extent), the HOMO energy residual (res_{homo}) remains globally orthogonal to almost all other properties. The global Pearson correlation between res_{homo} and the LUMO energy residual (res_{lumo}) is a modest 0.2071.

To rigorously validate the significance of this weak coupling, we evaluated the observed correlation against our scaffold-shuffled null model, which permutes res_{homo} strictly within identical Murcko scaffolds to preserve local structural density. The null distribution yielded a mean correlation of 0.0004 ± 0.0027, placing the observed global coupling at a Z-score of 77.17. This confirms that while the coupling is statistically non-zero, it is remarkably suppressed relative to the tight organization of the thermodynamic and geometric blocks. Furthermore, the distance correlation between res_{homo} and res_{lumo} is not static; it fluctuates as a function of structural complexity (measured by PC1 of the structural descriptors), peaking at 0.488 for the simplest structures and dropping to a minimum of 0.133 for mid-complexity molecules (Figure 2).

3.2 Topological structure and void identification

To investigate the breakdown of this coupling and identify the boundaries of property-space predictability, we constructed Vietoris-Rips complexes on the 12-dimensional residual point cloud and computed persistence diagrams for the first homology group (H_1). The resulting H_1 persistence barcode (Figure 3) reveals a series of prominent features representing one-dimensional topological voids, or loops, in the property manifold. These voids represent configurations of properties that are physically reachable but sparsely populated, indicating a bottleneck in chemical space where standard property correlations collapse.

To ensure uniform density and computational tractability during complex construction, we utilized a topological subsampling routine (Figure 4). This subsampling densely covers the core region of the manifold while avoiding extreme spatial outliers.

By extracting the simplices of the highest-persistence H_1 features, we isolated 96 void-forming molecules. These molecules reside in a highly anomalous region of property space where the residual HOMO-LUMO

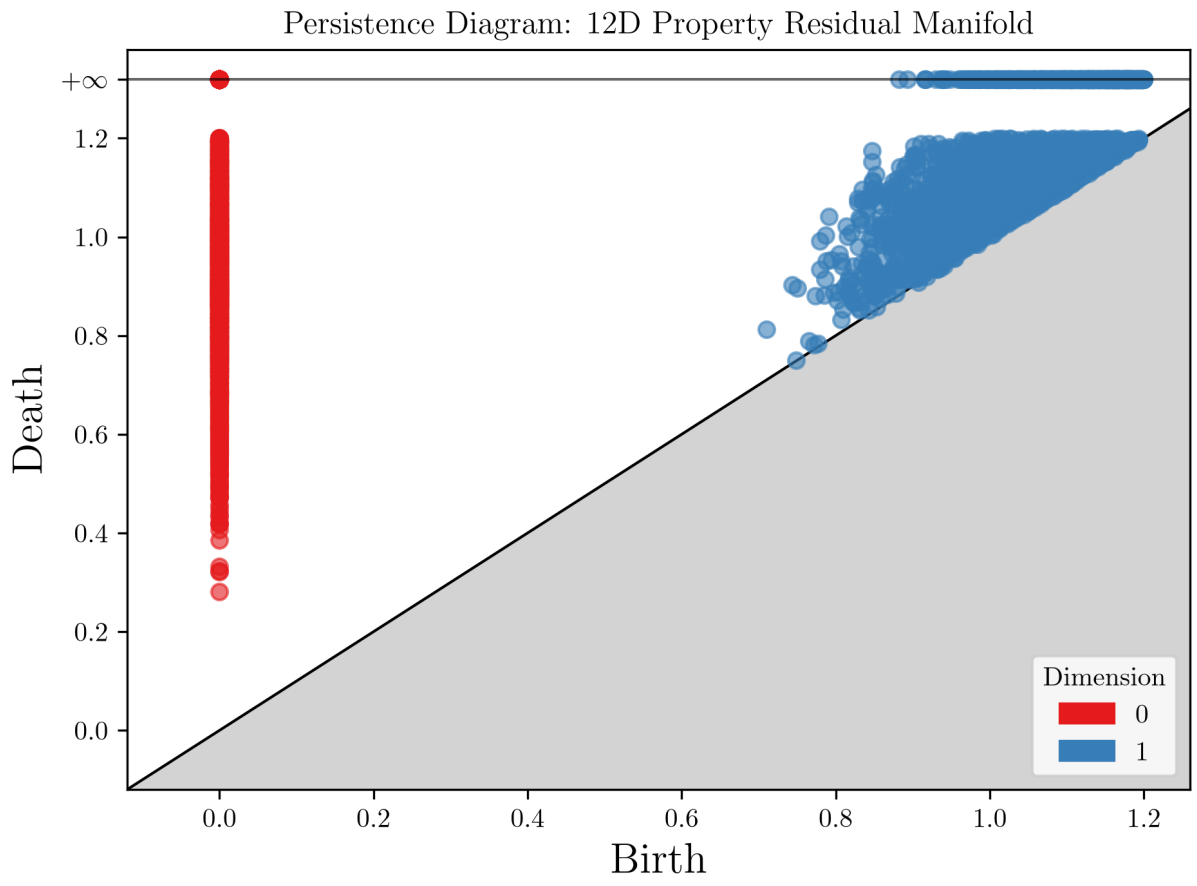

Figure 3: Persistence diagram of the 12-dimensional property residual manifold, displaying 0-dimensional connected components (red) and 1-dimensional topological voids (blue, H_1). The prominent H_1 features above the diagonal represent sparsely populated regions, or bottlenecks, in chemical space where the typical correlation structure of organic molecules breaks down. These voids specifically capture the topological isolation and near-total decoupling of the HOMO energy residual from the broader geometric and thermodynamic manifold. The presence of these distinct topological boundaries demonstrates that the joint structure of molecular properties is not uniform; rather, specific structural motifs involving electronic saturation and suppressed polarizability can decouple frontier orbitals, requiring targeted structural mutations to re-couple them to the global chemical space.

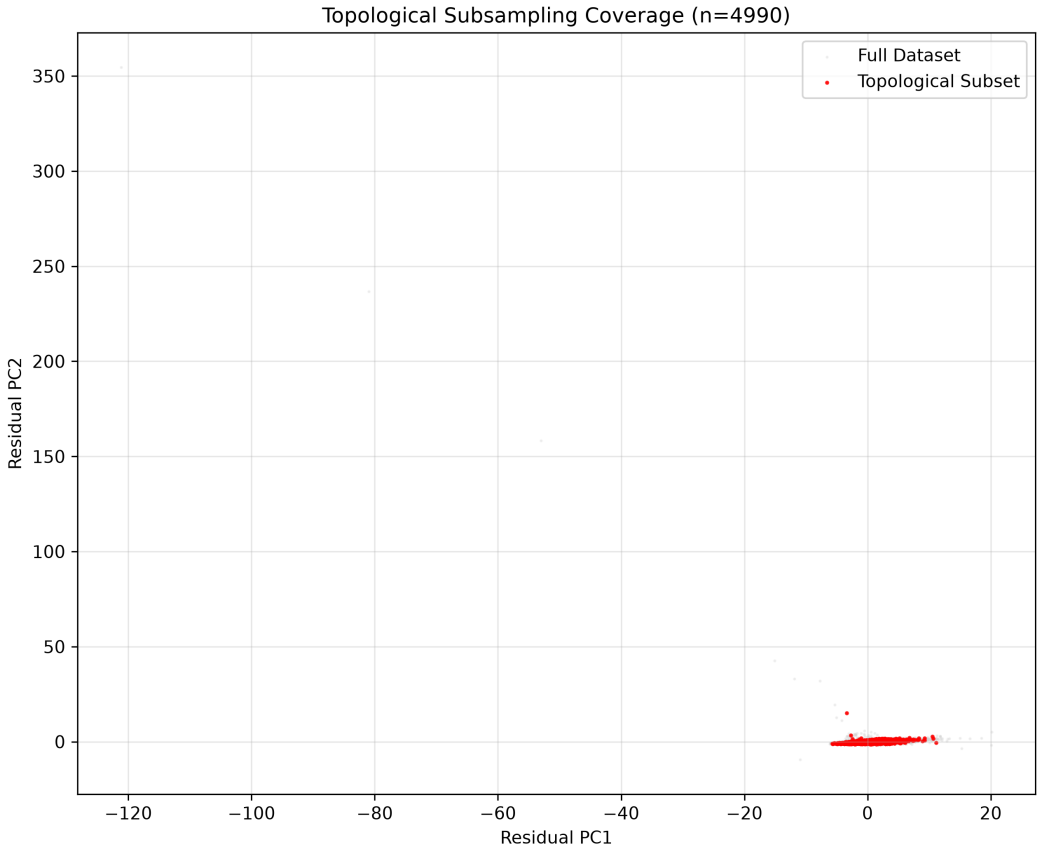

Figure 4: Projection of the 12-dimensional residual manifold onto its first two principal components (Residual PC1 and PC2), illustrating the topological subsampling coverage ($n = 4990$). The full dataset (grey points) includes extreme spatial outliers corresponding to the nine linear molecules, which exhibit massive residual values for properties such as the electronic spatial extent (`res_r2`). In contrast, the topological subset (red points) densely samples the core region of the manifold and contains zero linear molecules. This confirms that the identified topological voids and the associated decoupling of the HOMO energy residual (`res_homo`) are intrinsic features of the polyatomic, non-linear chemical space, rather than artifacts driven by extreme linear geometries.

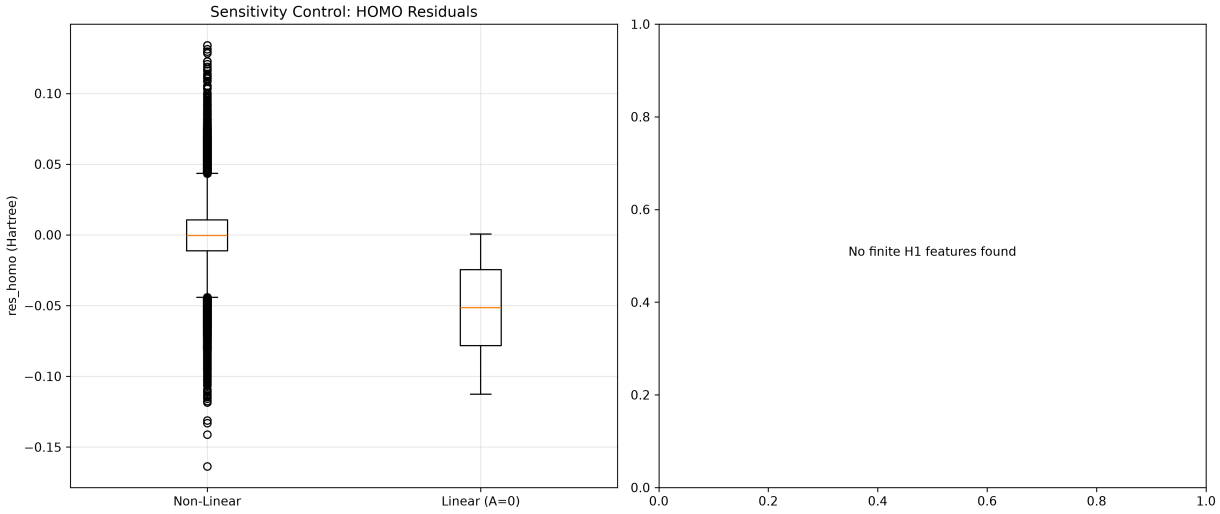

Figure 5: Sensitivity control analysis comparing HOMO energy residuals (res_{homo}) and topological features between non-linear and linear molecules. The left panel demonstrates that linear molecules exhibit distinct, extreme res_{homo} distributions compared to the global non-linear population. The right panel confirms the absence of finite H_1 topological voids for these linear structures. Together, these results verify that the 96 identified void-forming molecules—which define the topological isolation of the HOMO anomaly—are not artifacts of linear geometry, establishing that the observed decoupling is an intrinsic feature of polyatomic, non-linear chemical space.

correlation collapses entirely to 0.0019, representing a near-total decoupling from the global electronic manifold.

3.3 Structural motifs and sensitivity control

Analysis of the 96 void-forming molecules reveals distinct chemical signatures that distinguish them from the global population. First, the void subset exhibits an expansion of the frontier orbital energies beyond the standard manifold, with a mean shift in res_{homo} of +0.0051 Hartree and in res_{lumo} of +0.0120 Hartree relative to the background. Second, these molecules display significant geometric contraction. Isotropic polarizability residuals (res_{alpha}) in the void subset are shifted by -0.481 standard deviations (mean = -1.950 Bohr^3) compared to the global population. This indicates that topological isolation is strongly associated with molecules that are significantly less polarizable than their stoichiometric peers, pointing toward highly compact or electronically saturated structures.

To verify that these topological voids are not artifacts of linear geometry, we conducted a sensitivity control analysis (Figure 5). While the nine perfectly linear molecules in the QM9 dataset (e.g., acetylene, hydrogen cyanide) show extreme residual values—such as a mean res_{r2} of approximately 613.8 compared to -0.007 for non-linear molecules—zero linear molecules were found within the topological void subset.

3.4 Re-coupling via minimal mutations

To understand the structural conditions that force the HOMO anomaly to re-couple with the global manifold, we analyzed minimal mutation trajectories. For each of the 96 void-forming molecules, we identified its nearest neighbor in chemical space that resides outside the void, measured via Morgan fingerprint Tanimoto similarity.

The mean Tanimoto similarity for these mutation paths was 0.5063. This indicates that relatively moderate structural changes—such as the repositioning of a heteroatom, the alteration of lone-pair density, or the saturation of a double bond—are sufficient to transition a molecule from a state of topological isolation to one of global coupling (Figure 6).

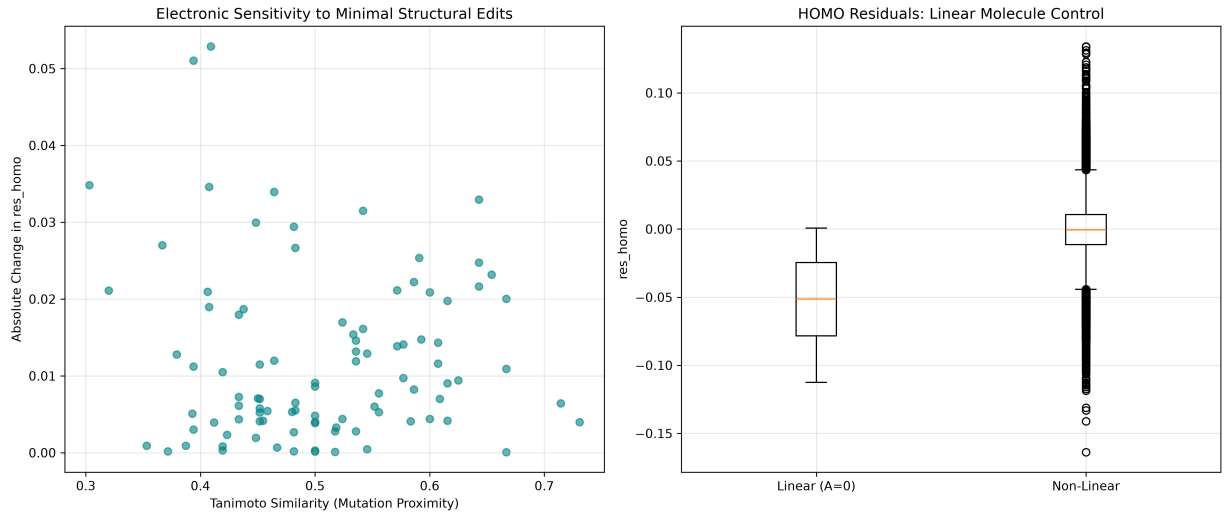

Figure 6: The figure illustrates the structural factors influencing the HOMO energy residual (res_{homo}). The left panel displays the absolute change in res_{homo} against Tanimoto similarity for minimal mutation trajectories. These moderate structural edits (mean similarity ≈ 0.506) induce a jump in HOMO residuals that successfully re-couples isolated void-forming molecules back into the global property manifold. The right panel contrasts the res_{homo} distributions of linear and non-linear molecules as a sensitivity control. Linear molecules exhibit distinctly shifted residual values compared to the non-linear population. Since zero linear molecules are present in the topological void subset, this control confirms that the observed topological isolation is a genuine feature of polyatomic, non-linear chemical space rather than an artifact of linear geometry.

Visualizing these trajectories in the three-dimensional res_{homo}, res_{lumo}, and res_{alpha} subspace (Figure 7) demonstrates the physical manifestation of this re-coupling. The minimal structural mutations typically induce a discrete "jump" in HOMO residuals, with a mean absolute jump of 0.01196 Hartree. This discrete transition aligns the previously isolated molecule with the dominant electronic-geometric manifold.

Collectively, these results demonstrate that the joint structure of the QM9 dataset is not a single, uniform manifold, but rather a composite of a highly organized thermodynamic core and a loosely coupled electronic shell. The topological isolation of the HOMO energy is the defining feature of this shell, concentrated in molecules with suppressed polarizability and expanded frontier orbital gaps. In small organic molecules, specific local bonding environments effectively decouple the frontier orbitals from the molecule's global spatial organization. Transitioning across these topological boundaries requires structural edits that reorganize the local electronic environment, effectively plugging the molecule back into the global manifold of organic physics.

4 Conclusions

In this study, we addressed the challenge of stoichiometric determinism in large-scale quantum chemistry datasets by constructing a 12-dimensional manifold of property residuals. By regressing out the dominant linear effects of atom counts, we exposed the intrinsic topological structure of the QM9 dataset. Our analysis revealed a profound anomaly: the highest occupied molecular orbital (HOMO) energy is globally decoupled from the otherwise coherent thermodynamic and geometric property manifold.

Utilizing persistent homology, we identified 96 distinct topological voids (features of the first homology group, H_1) that represent sparsely populated bottlenecks in chemical space. Within these isolated regions, the standard correlation between HOMO and LUMO energy residuals collapses almost entirely. We demonstrated that this topological isolation is not a geometric artifact of linear molecules, but rather a consequence of specific local electronic environments. The void-forming molecules are characterized by sig-

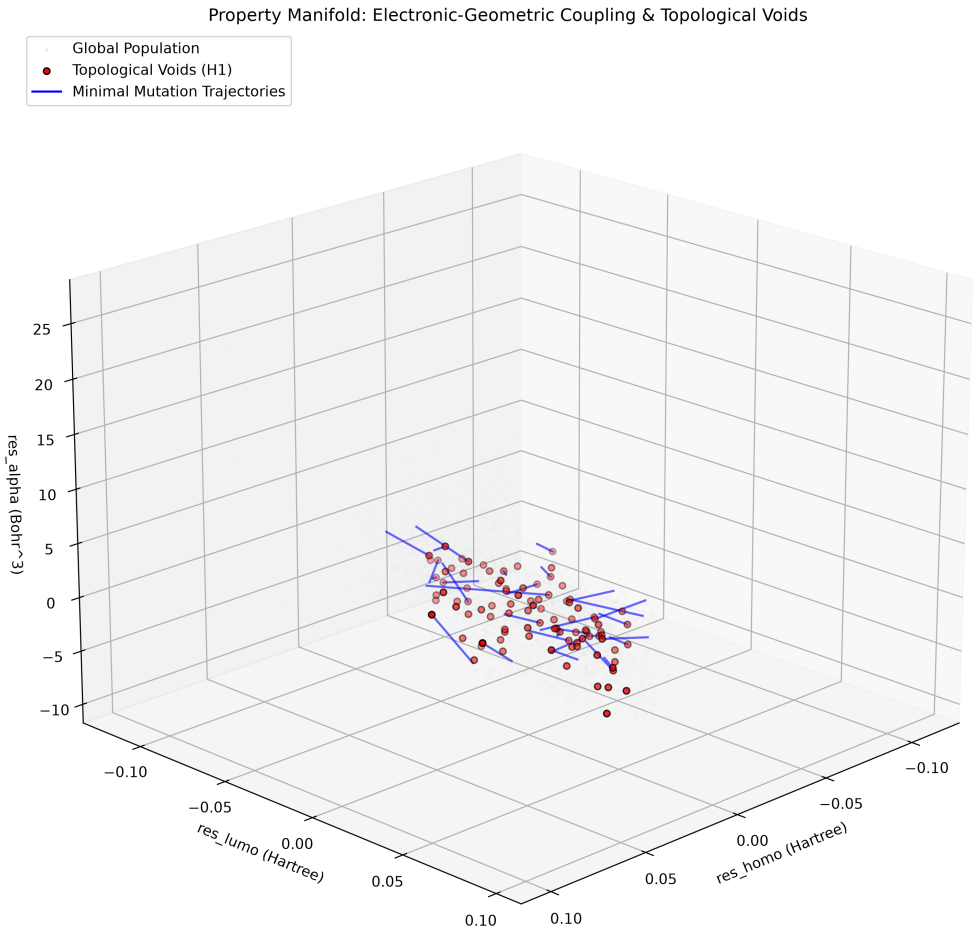

Figure 7: Three-dimensional visualization of the property residual manifold in the res_{homo}, res_{lumo}, and res_{alpha} subspace. The global molecular population (grey) is contrasted with 96 void-forming molecules (red) that define high-persistence H_1 topological features. These topological voids represent sparsely populated regions characterized by expanded frontier orbital energies and suppressed isotropic polarizability (res_{alpha} shifted by -0.481 sigma). Blue lines trace minimal mutation trajectories (mean Tanimoto similarity of 0.5063) connecting these isolated molecules to their nearest neighbors outside the voids. These trajectories reveal that moderate structural edits induce a discrete jump in HOMO residuals (mean absolute jump of 0.01196 Hartree), effectively re-coupling the topologically isolated molecules back into the dominant global manifold.

nificantly suppressed isotropic polarizability and expanded frontier orbital gaps, pointing to highly compact and electronically saturated structures where lone-pair density dictates anomalous orbital behavior.

Furthermore, our analysis of minimal structural mutation trajectories elucidated the mechanism by which molecules transition across these topological boundaries. We found that moderate chemical edits—averaging a Tanimoto similarity of 0.506—are sufficient to induce a discrete shift in the HOMO energy residual (averaging 0.01196 Hartree). These structural modifications, such as the repositioning of a heteroatom or the alteration of a bond order, effectively re-couple the isolated frontier orbitals back into the dominant global manifold.

Ultimately, these findings indicate that the joint structure of molecular properties is not a single, uniform continuum. Instead, it comprises a highly organized thermodynamic core and a loosely coupled electronic shell. The topological boundaries mapped in this work define the limits of property-space predictability for small organic molecules. By identifying the precise structural conditions that drive the decoupling and recoupling of the HOMO anomaly, this shape-based approach provides a robust framework for understanding the complex interplay between local electronic environments and global macroscopic physics.