Persistent Homology of Electronic-Thermodynamic Decoupling as an Intrinsic Stress Test for Molecular Property Manifolds

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding whether the decoupling between frontier electronic states and global thermodynamic properties is a fundamental chemical rule or a conditional artifact of structural strain remains a significant challenge in computational chemistry. To address this, we investigate the joint property structure of 133,802 small organic molecules from the QM9 dataset. By training gradient boosting regressors on exact atom counts, we extract stoichiometric residuals for frontier orbitals and thermodynamic properties to isolate intrinsic chemical variance. We then map the property manifold across a continuous structural strain coordinate using Principal Component Analysis, distance correlation metrics, and Topological Data Analysis with Vietoris-Rips complexes. Our results demonstrate a baseline decoupling where localized electronic states are driven by saturation and hydrogen count, while the global thermodynamic-spatial manifold is governed by skeletal architecture. However, this isolation breaks down under extreme geometric stress and at the limits of electronic stability. Anomaly detection identifies highly strained molecular populations that exhibit significant topological deformations in the property-dependency graph, evidenced by highly persistent homology cycles. Furthermore, stratification by lowest unoccupied molecular orbital residuals reveals a phase transition where the electronic and thermodynamic manifolds systematically recouple at stability extremes. These findings establish that electronic-thermodynamic isolation is a conditional state modulated by molecular strain rather than an absolute physical rule.

1 Introduction

The mapping of molecular property manifolds is a central pursuit in computational chemistry, providing the foundation for rational molecular design and high-throughput materials discovery. A recurring feature within large-scale quantum chemical datasets is the apparent orthogonality between localized frontier electronic states, such as the highest occupied molecular orbital (*homo*), and global thermodynamic or spatial properties, including heat capacity (c_v), zero-point vibrational energy (*zpve*), and electronic spatial extent (r^2). This electronic-thermodynamic decoupling is frequently interpreted as a fundamental chemical rule, ostensibly arising from the distinct physical origins of localized orbital energies versus extensive skeletal vibrations.

However, it remains an open question whether this isolation is an absolute physical reality or a conditional artifact. Specifically, it is unclear if the decoupling persists uniformly across chemical space or if it breaks down under extreme structural strain and at the boundaries of density functional theory (DFT) approximations, such as the rigid-rotor harmonic-oscillator limit. Understanding the fragility of this property-dependency graph is critical for determining the limits of predictive chemical models.

To address this challenge, we investigate the joint property structure of 133,802 small organic molecules from the QM9 dataset. Rather than discarding geometrically strained or anomalous molecules as computational failures, we exploit them as an intrinsic stress test for the property manifold. Because raw molecular properties are overwhelmingly dominated by trivial stoichiometric baselines, we first isolate the intrinsic chemical variance by training non-linear regressors on exact atom counts. By extracting these stoichiometric residuals for *homo*, *lumo*, *zpve*, c_v, and r^2, we project the molecules into a normalized multidimensional space that captures purely structural and electronic fluctuations.

To rigorously evaluate the stability of the property manifold, we map these residuals across a continuous structural strain coordinate using Principal Component Analysis and distance correlation metrics. Further-

more, to capture the global shape of the property-dependency graph beyond simple pairwise correlations, we employ Topological Data Analysis (TDA). By constructing Vietoris-Rips complexes on the residual point clouds, we extract persistent homology signatures. This topological framework allows us to quantify deformations in the property manifold by computing the Bottleneck distance between the persistence diagrams of geometrically consistent molecules and highly strained anomalous populations.

In this work, we demonstrate that while a baseline electronic-thermodynamic decoupling exists, it is fundamentally conditional. We show that localized electronic states are primarily driven by saturation and hydrogen count, whereas the global thermodynamic-spatial manifold is governed by skeletal architecture. Crucially, we identify that this isolation breaks down under extreme geometric stress, evidenced by highly persistent homology cycles in strained populations, and undergoes a systematic phase transition at the limits of electronic stability. These findings establish that the orthogonality of frontier orbitals to the thermodynamic manifold is not an absolute physical rule, but a dynamic state modulated by molecular strain.

2 Methods

2.1 Data preparation and structural descriptors

To establish a robust foundation for mapping the property manifold, we utilized the QM9 dataset. The dataset was first deduplicated based on canonical SMILES strings, retaining the first instance of each to yield 133,802 unique molecules. Molecular structures were parsed, and exact stoichiometric counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted.

To quantify molecular architecture and geometric stress, we computed a suite of structural descriptors. These included the Bertz complexity index, the fraction of sp^3 hybridized carbons (F_{sp^3}), heteroatom counts, and the number of aromatic, aliphatic, three-membered, and four-membered rings. Linear molecules were explicitly identified using a binary flag corresponding to a rotational constant of $A = 0$, and a log-transformation was applied to non-zero rotational constants to compress their scale.

We defined a continuous "Strain Coordinate" to serve as a proxy for geometric stress. This was constructed by applying a robust scaler to the ring strain proxies, F_{sp^3}, and the linear molecule flag, followed by a Principal Component Analysis (PCA). The first principal component of this standardized feature space was extracted and utilized as the continuous Strain Coordinate for downstream stratification.

2.2 Isolation of intrinsic chemical variance

Raw molecular properties in large datasets are overwhelmingly dominated by trivial stoichiometric baselines. To isolate the intrinsic chemical variance, we trained cross-validated Gradient Boosting Regressors to predict the highest occupied molecular orbital (*homo*), lowest unoccupied molecular orbital (*lumo*), zero-point vibrational energy (*zpve*), heat capacity (c_v), and electronic spatial extent (r^2) using strictly the five exact atom counts as input features.

The residuals from these predictions were computed and globally standardized to zero mean and unit variance. These stoichiometric residuals represent the variance in the quantum chemical properties that cannot be explained by atomic composition alone. To orthogonalize the global thermodynamic-spatial manifold, we performed PCA on the triad of residuals (*zpve*, c_v, r^2), extracting three principal components (PC1, PC2, and PC3). To evaluate the baseline decoupling between the electronic and thermodynamic manifolds, we subsequently trained non-linear Gradient Boosting Regressors to predict the *homo* and *lumo* residuals using these three thermodynamic principal components as features.

2.3 Anomaly detection and topological data analysis

To test whether the decoupling of frontier orbitals breaks down under extreme structural strain, we employed an Isolation Forest algorithm on the 4D stoichiometric residual space (*homo*, *zpve*, c_v, r^2). This identified a highly strained, "deviant" population of 3,054 molecules. To ensure rigorous comparative analysis, we constructed a strain-matched control group of 3,054 molecules from the inlier population using a nearest-neighbor matching algorithm based on the continuous Strain Coordinate.

We applied Topological Data Analysis (TDA) to capture the global shape of the property-dependency graph beyond pairwise correlations. Using the `gudhi` library, we constructed Vietoris-Rips complexes on the 4D residual point clouds for both the deviant and control groups. We extracted persistent homology signatures, specifically focusing on connected components (H_0) and one-dimensional loops (H_1). To mathematically quantify the deformation of the property manifold under geometric stress, we computed the Bottleneck distance between the H_1 persistence diagrams of the two populations.

2.4 Continuous mapping of the decoupling transition

To determine if the recoupling of electronic and thermodynamic states is a gradual effect or a sudden phase transition, we treated the degree of decoupling as a continuous phenomenon. The dataset was stratified into 10 deciles based on the Strain Coordinate. Within each decile, we computed the distance correlation ($dCor$) between the *homo* residual and the primary thermodynamic principal component (PC1). A null baseline was established by randomly shuffling the *homo* residuals within each decile and recomputing the distance correlation, ensuring statistical rigor.

Furthermore, we investigated the recoupling mechanism as a function of electronic state stability. The dataset was stratified based on the *lumo* residual into three subsets: the top 10% (highest LUMO), the bottom 10% (lowest LUMO), and the middle 80%. We generated 2D density distributions of the *homo* residual versus the thermodynamic PC1 for these subsets and applied locally weighted scatterplot smoothing (LOWESS) regression to visually and quantitatively assess changes in the coupling slope at the limits of electronic stability.

2.5 Structural drivers of the property manifold

To uncover the physical basis of the observed decoupling and recoupling behaviors, we trained parallel Gradient Boosting models to predict the *homo* residual, the *lumo* residual, and the primary thermodynamic PC1 using the extended suite of structural descriptors. We extracted feature importances using permutation importance analysis to identify whether frontier orbitals and thermodynamic components are driven by local/intensive features or global/extensive descriptors. Finally, we computed the Spearman rank correlation between the feature importance rankings of the three models to quantify the structural divergence between the electronic and thermodynamic manifolds.

3 Results

The joint property structure of the QM9 dataset, comprising 133,802 unique molecules, is characterized by a fundamental decoupling between frontier electronic states and the global thermodynamic-spatial manifold. Building upon the stoichiometric residuals extracted to isolate intrinsic chemical variance, our analysis reveals that this decoupling is not an absolute physical rule, but rather a conditional state regulated by molecular structure, geometric strain, and saturation.

3.1 Orthogonalization of the property manifold

Because raw molecular properties are overwhelmingly dominated by trivial stoichiometric baselines—where atom counts account for up to 99.99% of variance in thermodynamic properties—we first examined the residuals of the thermodynamic and spatial properties to map the intrinsic chemical space. Principal Component Analysis (PCA) on the triad of electronic spatial extent (r^2), heat capacity (c_v), and zero-point vibrational energy ($zpve$) yielded an orthogonalized thermodynamic-spatial manifold defined by three principal modes. The first principal component (PC1) accounts for 57.3% of the variance and represents a global response mode with high loadings from r^2 (0.671) and c_v (0.666). The second component (PC2, 30.5% variance) is a vibrational mode dominated by $zpve$ residuals (loading 0.945), while the third (PC3, 12.3% variance) captures a trade-off between spatial extent (0.710) and heat capacity (-0.704).

To evaluate the baseline relationship between this global manifold and localized electronic states, we trained non-linear regressors to predict the *homo* and *lumo* residuals using these three thermodynamic principal components. The resulting cross-validated R^2 values were negative (-0.435 for *homo* and -0.422

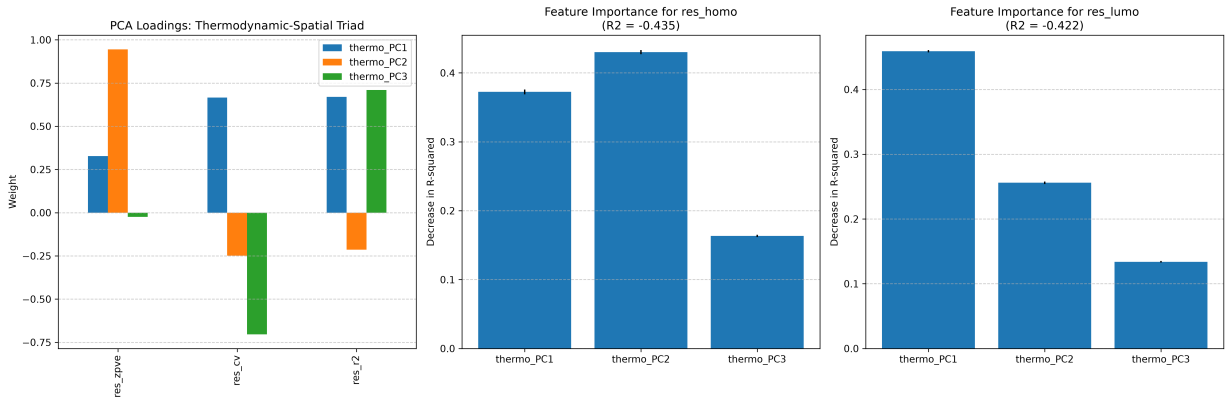

Figure 1: Principal component analysis loadings for the thermodynamic-spatial triad (left) and feature importance for predicting frontier orbital residuals (middle and right). The left panel displays the orthogonalized manifold defined by three principal components derived from the residuals of zero-point vibrational energy ($zpve$), heat capacity (c_v), and spatial extent (r^2). The middle and right panels illustrate the inability of these thermodynamic components to predict the residuals of the highest occupied ($homo$) and lowest unoccupied ($lumo$) molecular orbitals. The resulting negative cross-validated R-squared values (-0.435 and -0.422, respectively) reveal a fundamental decoupling between the global thermodynamic-spatial manifold and localized frontier electronic states, establishing that electronic fluctuations are independent of global thermodynamic organization.

for $lumo$), demonstrating that the thermodynamic components are entirely incapable of predicting frontier orbital fluctuations. This establishes a persistent baseline decoupling where localized electronic states are independent of global thermodynamic organization in the bulk molecular population (Figure 1).

3.2 Structural strain and the reorganization of joint space

While the bulk population exhibits strict electronic-thermodynamic orthogonality, we hypothesized that this isolation is modulated by structural strain. We defined a continuous Strain Coordinate derived from the first principal component of structural metrics, including Bertz complexity, ring counts, and the fraction of sp^3 hybridized carbons (F_{sp^3}). The sensitivity of this coordinate to molecular complexity and geometric stress is validated by its distinct distribution across linear and non-linear molecules (Figure 2).

3.2.1 Anomaly detection and geometric stress

To identify regions where the property manifold might deform, we applied an Isolation Forest to the 4D stoichiometric residual space ($homo$, $zpve$, c_v, r^2), isolating 3,054 "deviant" molecules. The anomaly scores of these molecules correlate significantly with the continuous Strain Coordinate (Pearson $r = 0.221$), indicating that statistical deviance in the property space is physically driven by geometric stress. Notably, this deviant population is highly enriched for structural extremes: while linear molecules constitute only 0.0067% of the global dataset, they make up 0.262% of the deviant group, representing a nearly 40-fold enrichment.

3.2.2 Topological persistence of property cycles

To capture the global shape of the property-dependency graph beyond pairwise correlations, we employed Topological Data Analysis (TDA). Vietoris-Rips complexes were constructed on the 4D residual point clouds for both the deviant group and a strain-matched control group. The resulting persistence diagrams (Figure 3) reveal that structural strain fundamentally deforms the topology of the joint property space. Although the control group exhibits a higher density of one-dimensional loops (H_1 features, 371 vs. 283), the cycles within the geometrically stressed deviant group are markedly more persistent. The maximum lifetime of an H_1 feature in the deviant group reaches 1.328, compared to only 0.534 in the control group. This

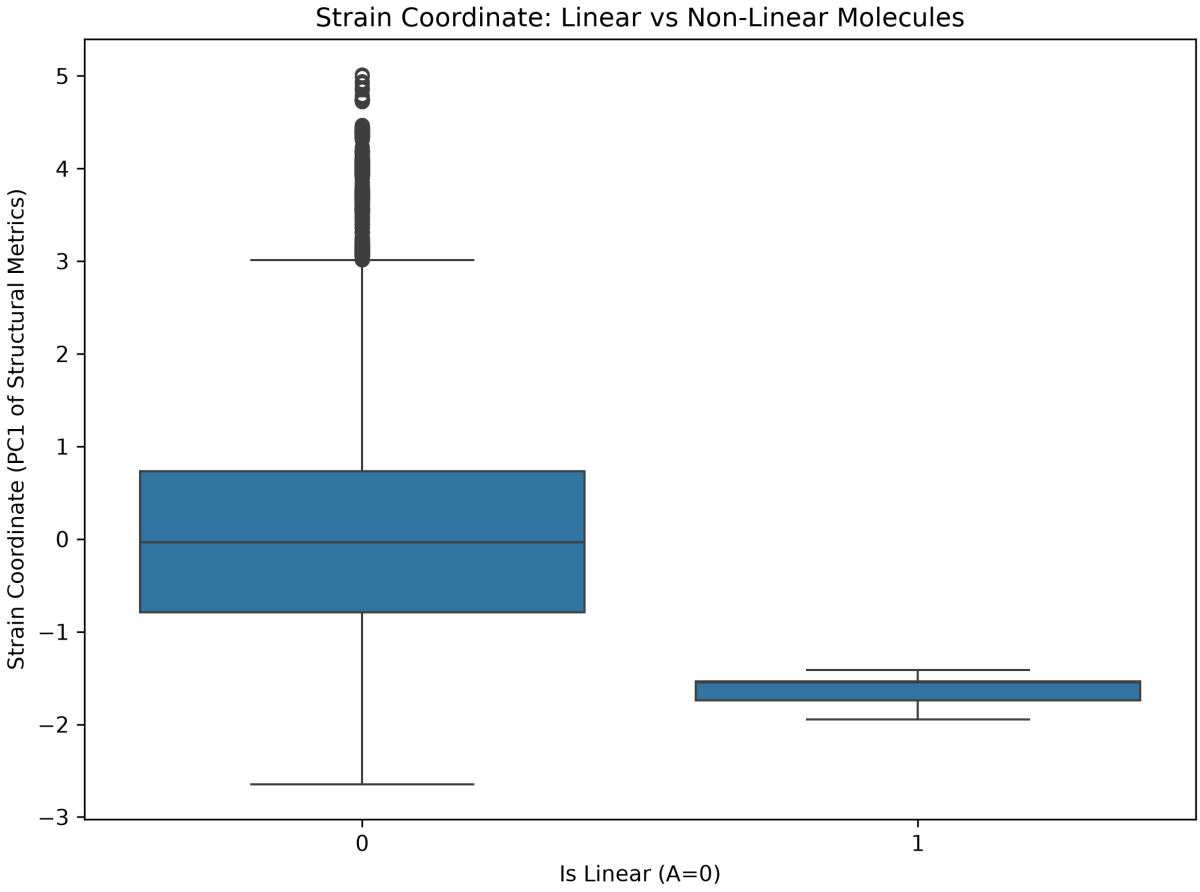

Figure 2: Distribution of the Strain Coordinate, derived from the first principal component of structural metrics including Bertz complexity, ring counts, and F_{sp^3}, for non-linear (0) and linear (1) molecules. The plot demonstrates that structurally simple linear molecules possess a significantly lower and more constrained strain coordinate than non-linear molecules. This confirms the sensitivity of the coordinate to molecular complexity, validating its application in assessing how geometric stress modulates the fundamental decoupling between frontier electronic states and the global thermodynamic-spatial manifold.

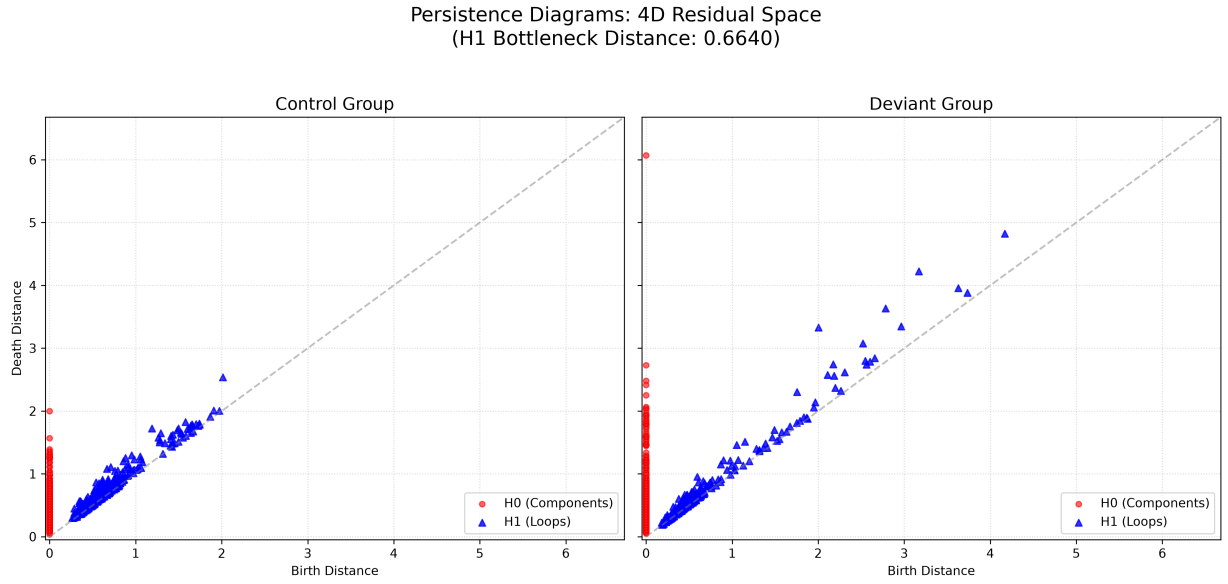

Figure 3: Persistence diagrams of the 4D residual space (*homo*, *zpve*, c_v, r^2) comparing a strain-matched control group (left) and a geometrically stressed deviant group (right). Red circles denote H0 connected components and blue triangles denote H1 loops. The diagrams illustrate that structural strain deforms the global shape of the property-dependency graph; although the control group contains a higher density of H1 features (371 versus 283), the cycles within the deviant group are markedly more persistent, exhibiting a maximum lifetime of 1.328 compared to 0.534 in the control. This significant topological shift is further quantified by an H1 Bottleneck distance of 0.664, revealing how geometric stress reorganizes the joint property space.

significant topological shift is further quantified by an H_1 Bottleneck distance of 0.664, confirming that extreme geometric stress reorganizes the fundamental dependencies between electronic and thermodynamic properties.

3.3 The recoupling transition

To determine whether the breakdown of electronic-thermodynamic isolation is a gradual effect or a sudden phase transition, we mapped the degree of decoupling continuously across strain deciles. Distance correlation ($dCor$) analysis between the *homo* residual and the primary thermodynamic component (PC1) shows that true correlations remain consistently above a shuffled null baseline across all deciles, with $dCor$ values ranging from 0.05 to 0.21 (Figure 4). This indicates a persistent, non-random coupling that fluctuates with the degree of molecular strain.

Crucially, the decoupling of *homo* is most fragile at the limits of electronic stability. By stratifying the dataset based on *lumo* residuals, we observed a distinct phase shift in coupling behavior. For the middle 80% of molecules (representing stable electronic states), there is negligible correlation between *homo* and the thermodynamic PC1 ($r = -0.029$). However, at the electronic extremes, the manifolds systematically recouple: molecules in the top 10% of *lumo* residuals exhibit significant positive coupling ($r = 0.327$), while those in the bottom 10% show significant negative coupling ($r = -0.238$). This phase transition, visualized in Figure 5, provides clear evidence that electronic and thermodynamic manifolds converge as molecules approach the boundaries of electronic stability.

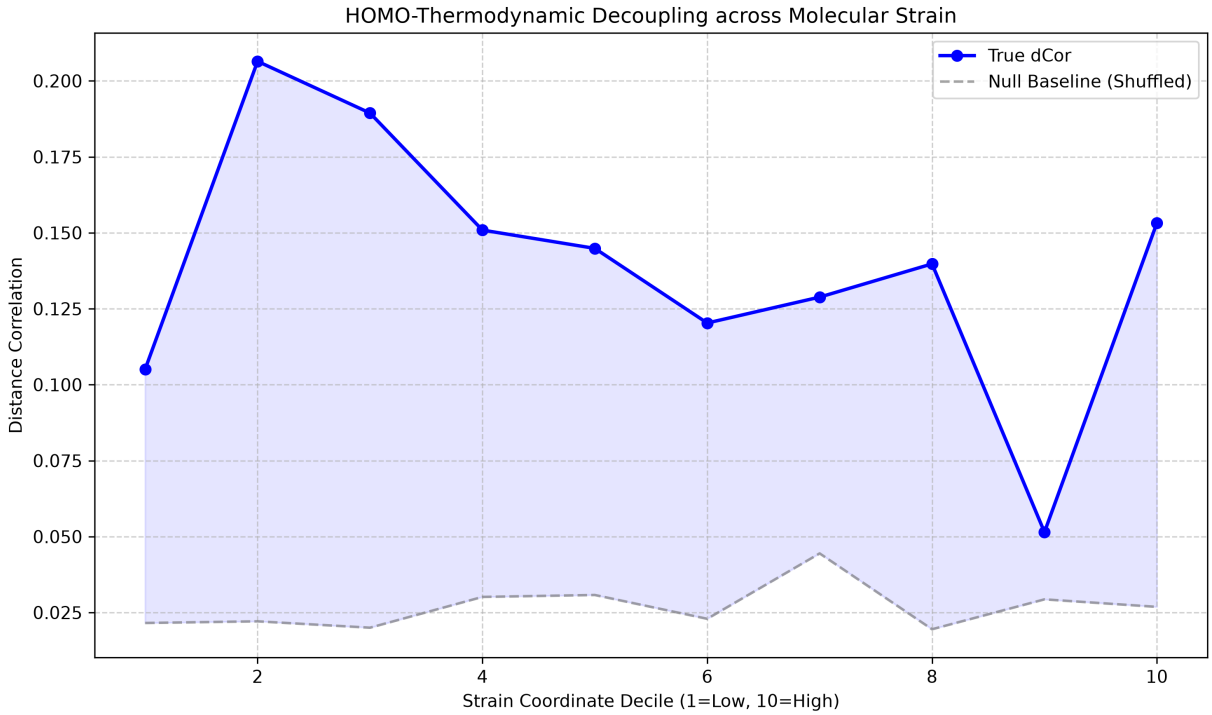

Figure 4: Distance correlation ($dCor$) between *homo* and the thermodynamic manifold across molecular strain deciles. The true $dCor$ (solid blue line) is compared against a shuffled null baseline (dashed grey line). True correlations remain consistently above the null baseline, with $dCor$ values ranging from 0.05 to 0.21. This demonstrates that the degree of decoupling between frontier electronic states and global thermodynamic properties varies with molecular state, indicating that structural strain modulates this relationship while maintaining a persistent, non-random coupling across all strain levels.

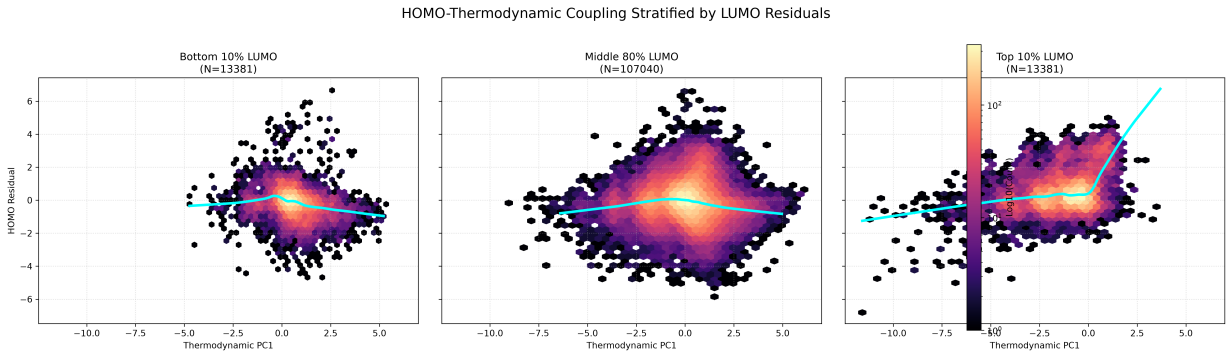

Figure 5: Density plots illustrating the coupling between *homo* residuals and the first principal component of the thermodynamic manifold (PC1), stratified by *lumo* residuals. While the middle 80% of molecules exhibit negligible correlation ($r = -0.029$), demonstrating a fundamental decoupling between frontier electronic states and global thermodynamic properties, a phase shift occurs at electronic extremes. Molecules in the bottom 10% of *lumo* residuals show significant negative coupling ($r = -0.238$), whereas those in the top 10% display significant positive coupling ($r = 0.327$). This reveals that the decoupling of *homo* is most fragile at these extremes, with the electronic and thermodynamic manifolds converging as molecules approach the limits of electronic stability.

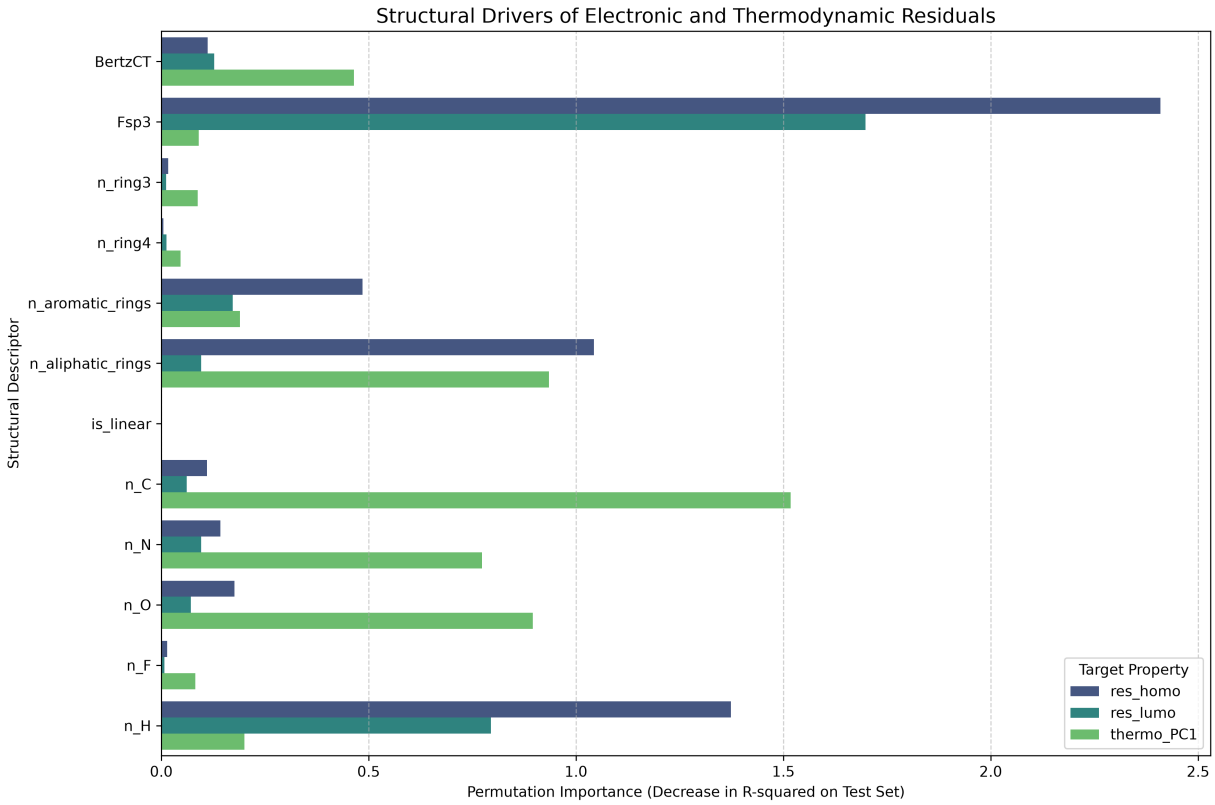

Figure 6: Permutation importance analysis of structural descriptors for electronic (*homo* and *lumo*) and thermodynamic (PC1) residuals. The plot demonstrates that electronic properties are primarily driven by saturation (F_{sp^3}, importance 2.41 for *homo* and 1.70 for *lumo*) and hydrogen count. In contrast, the thermodynamic manifold is driven by skeletal architecture, specifically carbon count (n_C, importance 1.52) and the number of aliphatic rings (0.935). This distinct structural dependence reveals the mechanism behind property decoupling: frontier orbitals are governed by local bonding and saturation, whereas the thermodynamic manifold tracks the global carbon skeleton.

3.4 Structural mediators of decoupling

To uncover the physical basis for the observed decoupling and subsequent recoupling, we conducted a permutation importance analysis to identify the structural drivers of each property manifold (Figure 6). The analysis reveals a stark divergence in structural dependence. The electronic residuals (*homo* and *lumo*) are primarily driven by local bonding characteristics, specifically saturation (F_{sp^3}, with an importance of 2.41 for *homo* and 1.70 for *lumo*) and hydrogen count. In contrast, the thermodynamic manifold (PC1) is governed by global skeletal architecture, predominantly carbon count (n_C, importance 1.52) and the number of aliphatic rings (0.935).

This structural divergence is quantitatively supported by the Spearman rank correlations between the feature importance profiles. The structural driver rankings for *homo* and *lumo* are highly correlated (0.916), reflecting their shared dependence on the local electronic environment. Conversely, the correlation between the drivers of *homo* and the thermodynamic PC1 is significantly lower (0.524). This confirms the mechanism behind the baseline property decoupling: frontier orbitals track local saturation and bonding, whereas the thermodynamic manifold tracks the extensive carbon skeleton. It is only when extreme geometric stress or electronic instability forces these local and global structural features into conflict that the baseline decoupling breaks down.

4 Conclusions

In this study, we investigated the fundamental assumption of electronic-thermodynamic decoupling within the QM9 dataset to determine whether the orthogonality between localized frontier orbitals and global thermodynamic properties is an absolute chemical rule or a conditional artifact. By isolating intrinsic chemical variance through stoichiometric residuals, we established that the baseline decoupling observed in the bulk molecular population is driven by a stark divergence in structural dependencies. Frontier electronic states (*homo* and *lumo*) are governed by local bonding characteristics, specifically saturation (F_{sp^3}) and hydrogen count, whereas the thermodynamic-spatial manifold ($zpve$, c_v, r^2) is dictated by the global skeletal architecture.

Crucially, our results demonstrate that this isolation is highly fragile and breaks down under geometric stress and electronic instability. Topological Data Analysis revealed that highly strained molecular populations undergo a fundamental reorganization of their joint property space. The emergence of highly persistent H_1 homology cycles in these geometrically stressed populations, quantified by a significant Bottleneck distance relative to a strain-matched control, provides rigorous mathematical evidence of a topological deformation in the property-dependency graph.

Furthermore, we identified a systematic phase transition at the limits of electronic stability. While molecules with stable *lumo* residuals maintain strict electronic-thermodynamic orthogonality, those at the extremes of the *lumo* distribution exhibit significant recoupling. This convergence of the electronic and thermodynamic manifolds indicates that the independence of localized orbitals cannot be guaranteed at the boundaries of chemical stability.

Ultimately, these findings establish that the decoupling of frontier orbitals from the thermodynamic manifold is a dynamic, conditional state rather than an absolute physical law. The recoupling observed under extreme structural strain likely reflects the breakdown of underlying quantum chemical approximations, such as the rigid-rotor harmonic-oscillator limit, rather than a genuine physical phenomenon. Consequently, this persistent homology framework serves as a rigorous intrinsic stress test for molecular property manifolds. Recognizing the conditional nature of this decoupling is essential for the development of robust machine learning models and high-throughput screening pipelines, ensuring that predictive failures are anticipated when navigating highly strained or electronically extreme regions of chemical space.