Conditional Re-Coupling of Frontier Orbital Energies to the Geometric-Thermodynamic Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy typically exhibits a unique orthogonal decoupling from global geometric and thermodynamic molecular properties. It remains unclear whether this independence is a universal physical baseline or a conditionally mediated state. To address this, we investigate the structural conditions under which frontier orbital energies integrate with the macroscopic molecular state using the QM9 dataset. We isolate topology-driven variance by extracting non-linear stoichiometric residuals via gradient-boosted trees and define a geometric-thermodynamic manifold using principal component analysis. By stratifying the dataset based on aromaticity, saturation, and heteroatom presence, we employ partial distance correlation to quantify the architectural conditions that drive the re-coupling of the HOMO energy to this manifold. While global baseline analysis confirms that the HOMO energy is largely decoupled from the geometric manifold, structural stratification reveals that this decoupling breaks down in highly saturated systems. Specifically, the HOMO energy significantly re-couples to the manifold in saturated molecules, particularly those containing heteroatoms, whereas the presence of aromaticity reinforces electronic-geometric orthogonality. Furthermore, mediation analysis demonstrates that this re-coupling is driven by polarizability in saturated hydrocarbons, but manifests as a direct topological integration in heteroatomic systems. Validated through a robust leave-one-scaffold-out approach, our findings demonstrate that the independence between frontier electronic and geometric properties is not absolute, but conditionally dependent on the degree of saturation and π-conjugation.

1 Introduction

The mapping of molecular structures to their physicochemical properties is a central pursuit in chemical physics and informatics. Within the multidimensional space of molecular properties, distinct clusters or "manifolds" of correlated features often emerge. Geometric and thermodynamic properties, such as rotational constants, spatial extent, and heat capacity, frequently covary as they are fundamentally tied to the macroscopic mass distribution and vibrational modes of the molecule. In contrast, frontier electronic properties—most notably the highest occupied molecular orbital (HOMO) energy—typically exhibit a unique orthogonal decoupling from this global geometric-thermodynamic manifold.

This apparent independence of the HOMO energy raises a fundamental question: is this orthogonal decoupling a universal physical baseline inherent to all molecular systems, or is it a conditionally mediated state that can be broken by specific architectural motifs? While bulk dataset analyses often suggest a global decoupling, such macroscopic averaging may obscure localized regions of the chemical space where frontier orbital topology is forced to integrate with the molecule's global geometric state. Understanding the boundaries of this decoupling is critical for refining our physical intuition of structure-property relationships and for understanding how localized electronic states interact with global molecular architecture.

To address this gap, we systematically investigate the structural conditions under which the HOMO energy re-couples to the geometric-thermodynamic manifold using the QM9 dataset. Because molecular size and stoichiometry dominate raw property variance, we first isolate pure topology-driven variance by extracting non-linear stoichiometric residuals via gradient-boosted trees. We then define a unified geometric-thermodynamic manifold using principal component analysis on the residuals of geometric response properties and heat capacity. By explicitly stratifying the dataset based on interpretable chemical descriptors—specifically aromaticity, degree of saturation, and heteroatom presence—we deploy partial distance

correlation to mathematically quantify the exact architectural conditions that drive the re-coupling of the HOMO energy to this manifold.

Our analysis reveals that the independence between frontier electronic and geometric properties is not absolute. While a global baseline analysis confirms that the HOMO energy is largely decoupled from the geometric manifold, structural stratification demonstrates that this decoupling breaks down significantly in highly saturated systems, particularly those containing heteroatoms. Conversely, the presence of aromaticity strongly reinforces electronic-geometric orthogonality. Through mediation analysis, we further show that this re-coupling is driven by polarizability in saturated hydrocarbons, whereas it manifests as a direct topological integration in heteroatomic systems. Validated through a robust leave-one-scaffold-out approach, these findings provide a nuanced understanding of how the degree of saturation and π-conjugation conditionally dictate the integration of frontier orbitals into the macroscopic molecular state.

2 Methods

2.1 Data preparation and stoichiometric residual extraction

To isolate the variance in molecular properties driven purely by structural topology, we first removed the dominant scaling effects of molecular size and stoichiometry. We utilized the QM9 dataset, immediately excluding 9 linear molecules characterized by a rotational constant $A = 0$ to prevent geometric singularities that could distort variance-based manifold definitions. We retained properties for which structural topology plays a significant role: the rotational constants A (transformed as $\log(A + 1)$ to compress its dynamic range), B, and C; dipole moment μ; isotropic polarizability α; electronic spatial extent r^2; heat capacity c_v; and the frontier orbital energies, HOMO and LUMO. Thermodynamic and atomization properties that are near-perfectly predicted by simple atom counts were excluded from this analysis.

We extracted the non-linear stoichiometric residuals for each retained property by training a LightGBM regressor on the filtered dataset. The sole predictors for these models were the exact atom counts (C, N, O, F, H) and the total heavy atom count. The residuals from these regressions (e.g., res_homo, res_r^2) represent the property variations attributable to molecular architecture rather than bulk composition. To prevent heavy-tailed distributions from skewing downstream distance metrics, the resulting residual matrix was standardized using a robust scaler based on the median and interquartile range.

2.2 Explicit structural stratification

To determine whether the decoupling of the HOMO energy is a universal baseline or a conditionally mediated state, we partitioned the dataset into discrete, interpretable chemical families. Using RDKit, we computed three explicit structural descriptors for each molecule:

- **Aromaticity:** A boolean flag indicating the presence of at least one aromatic ring.
- **Heteroatom density:** A categorical classification separating pure hydrocarbons from molecules containing heteroatoms (N, O, or F).
- **Saturation:** The fraction of sp^3-hybridized carbons, used to bin the molecules into low, medium, and high saturation regimes.

Combinations of these descriptors were used to stratify the dataset into distinct structural bins (e.g., highly saturated heteroatomic systems, low-saturation aromatic hydrocarbons). This explicit stratification avoids the opacity of unsupervised clustering and allows for a direct chemical interpretation of the boundaries where property relationships shift.

2.3 Defining the geometric-thermodynamic manifold

We defined a unified geometric-thermodynamic manifold to serve as the macroscopic baseline against which frontier orbital integration could be measured. We isolated the residual matrix of the geometric and thermodynamic properties ($\log(A + 1)$, B, C, μ, α, r^2, c_v) by explicitly excluding the frontier electronic residuals (res_homo and res_lumo).

Principal Component Analysis (PCA) was performed on this reduced, robustly scaled residual matrix across the entire dataset. By fitting the PCA globally rather than per-bin, we ensured that the primary axis of variance—the first principal component (PC1)—was defined consistently across all structural partitions. PC1 was extracted to serve as the single latent variable representing the geometric-thermodynamic manifold, and its feature loadings were recorded to interpret the physical composition of this axis.

2.4 Conditional re-coupling and mediation analysis

To mathematically quantify the integration of the HOMO energy into the manifold, we employed partial distance correlation, a metric capable of capturing non-linear dependencies between variables while controlling for confounding factors. We computed a global baseline partial distance correlation between *res_homo* and PC1 across the entire dataset, strictly conditioning on *res_lumo* to isolate HOMO-specific topological effects. A consistent Euclidean distance metric was applied to the standardized residuals. As a high-correlation null control, we also calculated the partial distance correlation between res_r^2 and PC1.

We then iterated through each discrete structural partition, computing the same partial distance correlation between *res_homo* and PC1 (conditioned on *res_lumo*) within each bin. The delta between these bin-specific correlations and the global baseline was used to identify the architectural conditions that force the HOMO energy to re-couple to the macroscopic manifold.

Furthermore, we conducted a mediation analysis to determine if isotropic polarizability acts as a "hub" property bridging the electronic structure and the geometric manifold. For structural partitions exhibiting significant re-coupling, we recomputed the partial distance correlation between *res_homo* and PC1, this time conditioning on both *res_lumo* and res_α. A substantial reduction (or "drop") in the correlation magnitude upon the inclusion of res_α indicates that the re-coupling is mediated by the molecule's polarizability, whereas a negligible drop suggests a direct topological integration.

2.5 Leave-one-scaffold-out validation

To ensure that the observed re-coupling phenomena were fundamental properties of the chemical families' topologies rather than artifacts driven by highly populated, specific molecular scaffolds, we implemented a rigorous leave-one-scaffold-out validation strategy. For the target partitions, we computed the Murcko scaffold for each molecule. We iteratively held out entire scaffold families, recomputed the partial distance correlation on the remaining molecules within the partition, and tracked the variance and stability of the effect size across the folds. This confirmed the robustness of the structural claims regarding saturation and π-conjugation.

3 Results

3.1 Composition of the geometric-thermodynamic manifold

To establish a macroscopic baseline for evaluating frontier orbital integration, we first characterized the geometric-thermodynamic manifold using principal component analysis (PCA) on the non-linear stoichiometric residuals of the geometric and thermodynamic properties. As illustrated in Figures 1 and 2, the primary axis of variance (PC1) is fundamentally a rotational axis. The manifold is overwhelmingly dominated by the residuals of the rotational constants B and C, which exhibit large positive loadings of 0.6685 and 0.7324, respectively. In contrast, other properties contribute minimally to this primary axis: the electronic spatial extent r^2 carries a loading of -0.0907, isotropic polarizability α -0.0642, and heat capacity c_v -0.0474. The dipole moment μ (-0.0113) and the log-transformed rotational constant A (-0.0443) are essentially orthogonal to this axis. Consequently, PC1 effectively tracks the molecular moment of inertia independent of stoichiometric scaling, providing a robust geometric baseline.

3.2 Global decoupling and conditional re-coupling

With the manifold defined, we evaluated the global baseline integration of the HOMO energy. Across the entire dataset of 133,793 molecules, the HOMO energy is almost entirely decoupled from the geometric

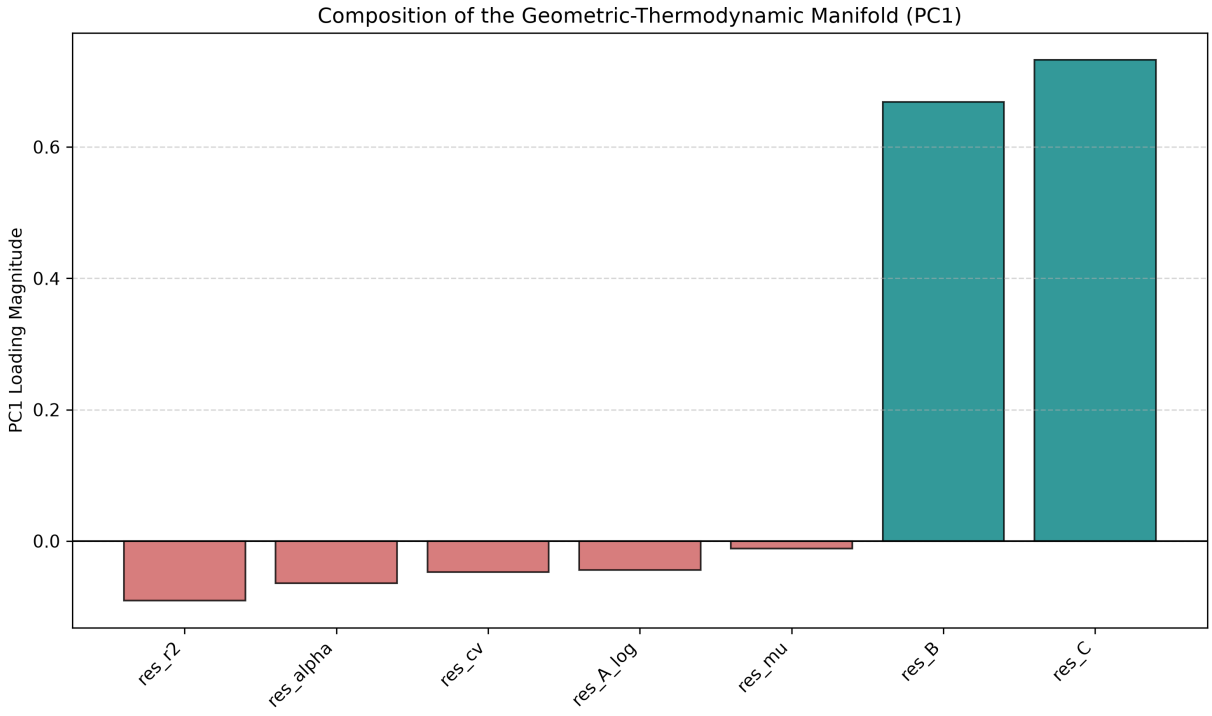

Figure 1: Principal component analysis loadings defining the geometric-thermodynamic manifold (PC1). The plot illustrates the contributions of various property residuals, specifically `res_r2`, `res_alpha`, `res_cv`, `res_A_log`, `res_mu`, `res_B`, and `res_C`. The manifold is overwhelmingly dominated by the residuals of the rotational constants B and C, which exhibit large positive loadings of 0.6685 and 0.7324, respectively. In contrast, other properties such as the electronic spatial extent (r^2), polarizability (α), and heat capacity (c_v) contribute minimally. This loading profile demonstrates that the primary axis of variance is fundamentally a rotational axis, revealing that the manifold effectively tracks the molecular moment of inertia independent of stoichiometric scaling.

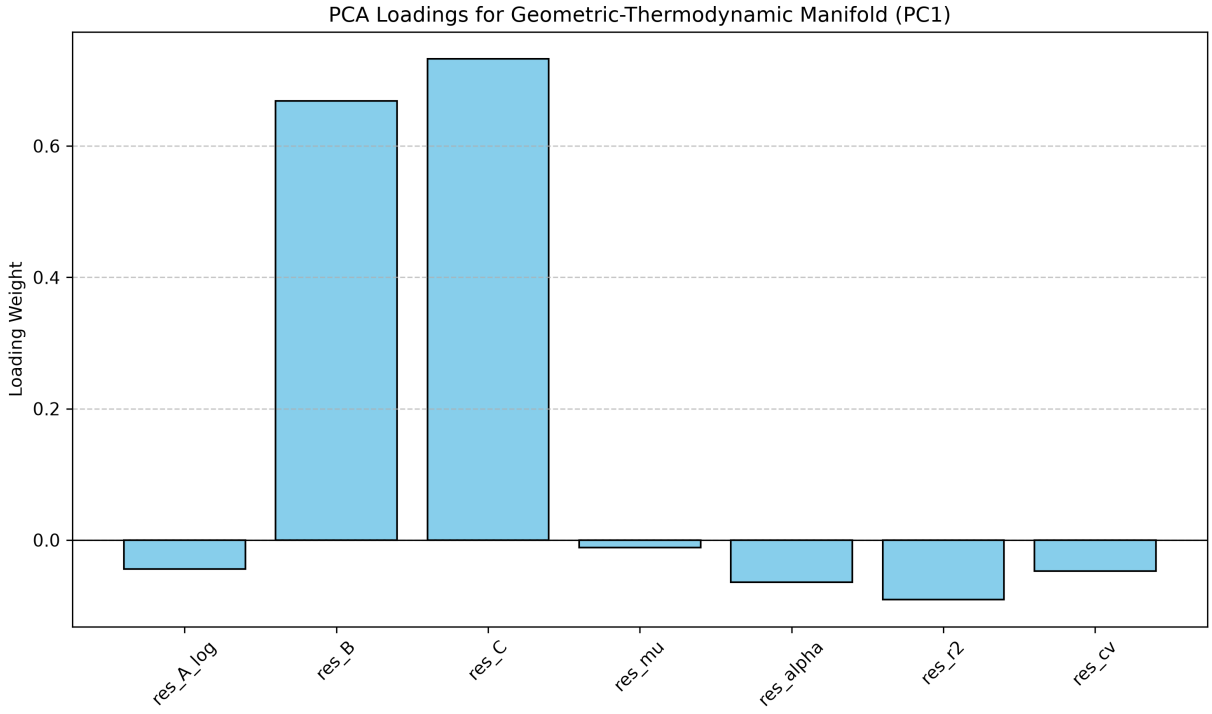

Figure 2: PCA loadings for the first principal component (PC1) characterizing the geometric-thermodynamic manifold. The plot displays the loading weights for the residuals of various molecular properties, revealing that the manifold is predominantly driven by the rotational constants B and C. Other properties, including the log-transformed constant A, dipole moment μ, polarizability α, electronic spatial extent r^2, and heat capacity c_v, show minimal contributions. This dominant weighting of B and C demonstrates that the primary axis of variance functions fundamentally as a rotational axis, capturing the molecular moment of inertia independently of stoichiometric scaling.

manifold. The partial distance correlation between the HOMO residual (res_{homo}) and PC1, conditioned on the LUMO residual (res_{lumo}) to isolate HOMO-specific effects, is merely 0.0338. For comparison, a highly integrated property such as res_{r^2} exhibits a strong partial distance correlation of 0.8825. However, stratifying the dataset into discrete structural bins (Figure 3) reveals that this decoupling is not a universal physical constant but is highly sensitive to molecular architecture.

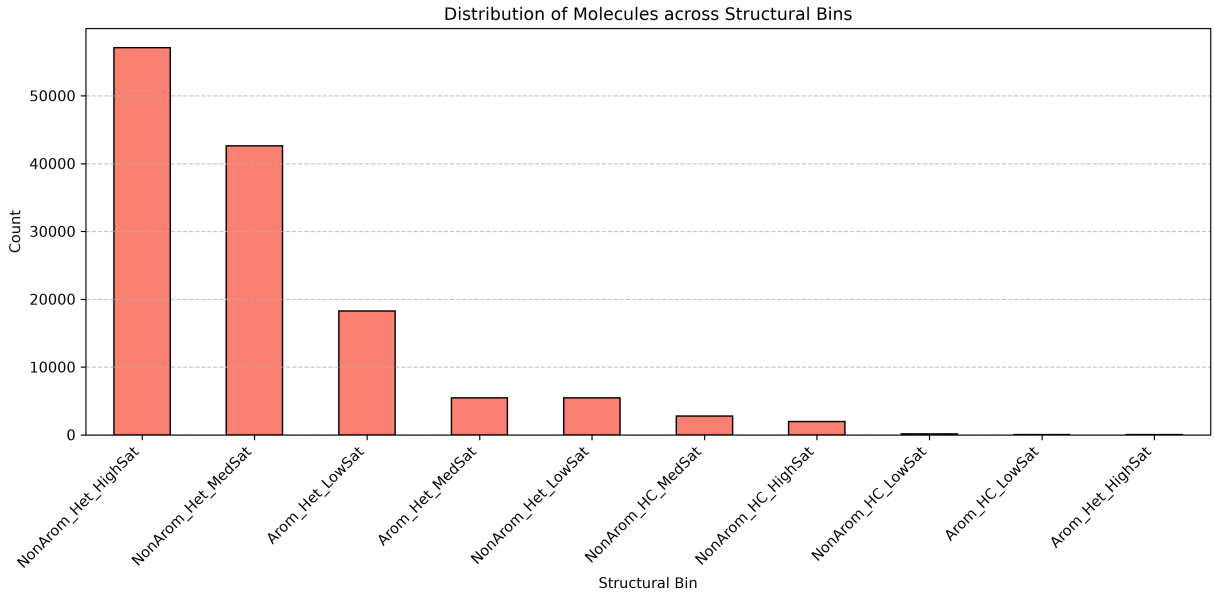

Figure 3: Distribution of the 133,793 molecules from the QM9 dataset across distinct structural bins based on aromaticity, heteroatom composition, and saturation level. The dataset is predominantly composed of saturated heteroatomic systems, specifically the `NonArom_Het_HighSat` ($N = 57,070$) and `NonArom_Het_MedSat` ($N = 42,619$) regimes. These highly populated bins correspond to architectures exhibiting the strongest conditional re-coupling between the HOMO energy and the geometric-thermodynamic manifold. Conversely, the `Arom_Het_LowSat` bin ($N = 18,260$), where aromaticity enforces decoupling and maintains electronic-geometric orthogonality, represents a smaller fraction of the dataset. This distribution reveals that the observed breakdown of independence between frontier electronic and geometric properties is characteristic of the most abundant molecular families in the bulk dataset, whereas purely orthogonal π-systems are significantly less frequent.

As shown in Figure 4, the relationship between res_{homo} and the manifold varies significantly across structural regimes. The strongest re-coupling occurs in highly saturated molecules containing heteroatoms (N, O, or F). In the `NonArom_Het_HighSat` bin ($N = 57,070$), the partial distance correlation rises to 0.0618, representing a substantial increase ($\Delta = +0.0280$) from the global baseline. Similar re-coupling is observed in hydrocarbon systems, with the `NonArom_HC_MedSat` ($N = 2,780$) and `NonArom_Het_MedSat` ($N = 42,619$) regimes yielding correlations of 0.0616 and 0.0606, respectively.

Conversely, the presence of aromaticity strongly reinforces the decoupling of the frontier orbitals from the geometric state. In the `Arom_Het_LowSat` bin ($N = 18,260$), the correlation collapses to 0.0006, a decrease of $\Delta = -0.0332$ relative to the global baseline. The differences in partial distance correlation across these regimes are summarized in Figures 5 and 6. These results demonstrate that as molecular saturation increases and aromaticity is removed, the frontier orbital topology is forced to integrate with the global geometric state, whereas aromatic π-systems robustly maintain electronic-geometric orthogonality.

3.3 Mediation via polarizability

To understand the physical mechanism driving this conditional re-coupling, we conducted a mediation analysis to test whether isotropic polarizability (α) acts as a hub property bridging the electronic structure and

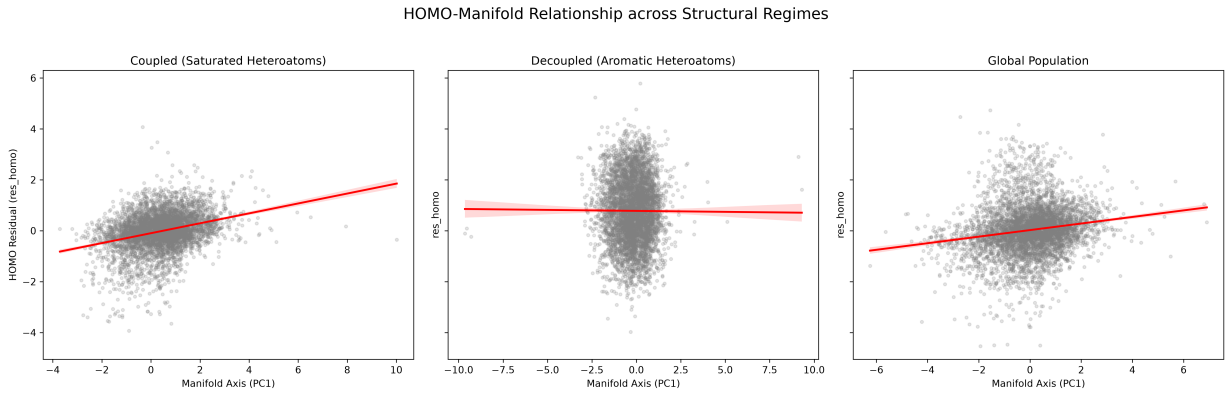

Figure 4: Scatter plots illustrating the relationship between the HOMO residual and the primary geometric-thermodynamic manifold axis (PC1) across distinct structural regimes. The left panel demonstrates the conditional re-coupling observed in saturated heteroatomic systems, where frontier orbital topology integrates with the global geometric state. In contrast, the middle panel shows near-complete decoupling in aromatic heteroatomic systems, revealing that aromatic π-systems maintain electronic-geometric orthogonality. The right panel displays the global population baseline, which exhibits minimal overall coupling. These comparisons highlight that the independence of frontier electronic and geometric properties is not a universal constant, but rather breaks down in highly saturated molecular architectures.

the geometric manifold. We defined a mediation effect as the reduction in partial distance correlation when the polarizability residual (res_α) is added to the conditioning set.

The mediation effects reveal a stark contrast between chemical families. In moderately saturated hydrocarbons (`NonArom_HC_MedSat`), the re-coupling is almost entirely mediated by polarizability. The base correlation of 0.0634 drops to -0.0018 when conditioning on res_α (a drop of 0.0652). This indicates that for pure hydrocarbons, the apparent link between the HOMO energy and the geometric manifold is a secondary effect of the manifold's influence on the deformability of the electron cloud. In contrast, for highly saturated heteroatomic systems (`NonArom_Het_HighSat`), the mediation is negligible. The correlation only drops from 0.0678 to 0.0642 (a drop of 0.0036). In these systems, the re-coupling of the HOMO energy to the manifold manifests as a direct topological integration that does not depend on global polarizability.

3.4 Stability and robustness of the re-coupling phenomenon

Finally, to ensure that the observed re-coupling phenomena are fundamental properties of the structural regimes rather than artifacts of specific, highly populated molecular scaffolds within the QM9 dataset, we validated our findings using a leave-one-scaffold-out approach. As illustrated in Figure 7, the re-coupling effect in the `NonArom_Het_HighSat` regime is highly stable across 10 scaffold-holdout folds, yielding a mean partial distance correlation of 0.0772 ($\sigma = 0.0091$). The correlation remained positive and significant in every fold, ranging from 0.0634 to 0.0930.

Similarly, the near-zero correlation observed in aromatic heteroatomic systems (`Arom_Het_LowSat`) was robustly recovered, with a mean of 0.0004 ($\sigma = 0.0006$) across all folds. This rigorous validation confirms that the conditional re-coupling of the HOMO energy is an intrinsic feature of saturated architectures. The frontier electronic and geometric property blocks are only truly independent in the presence of π-conjugation; in saturated systems, this orthogonal decoupling partially breaks down.

4 Conclusions

In this study, we addressed a fundamental question in chemical physics: whether the apparent orthogonal decoupling of the highest occupied molecular orbital (HOMO) energy from macroscopic geometric and thermodynamic properties is a universal physical constant or a conditionally mediated state. By extracting non-

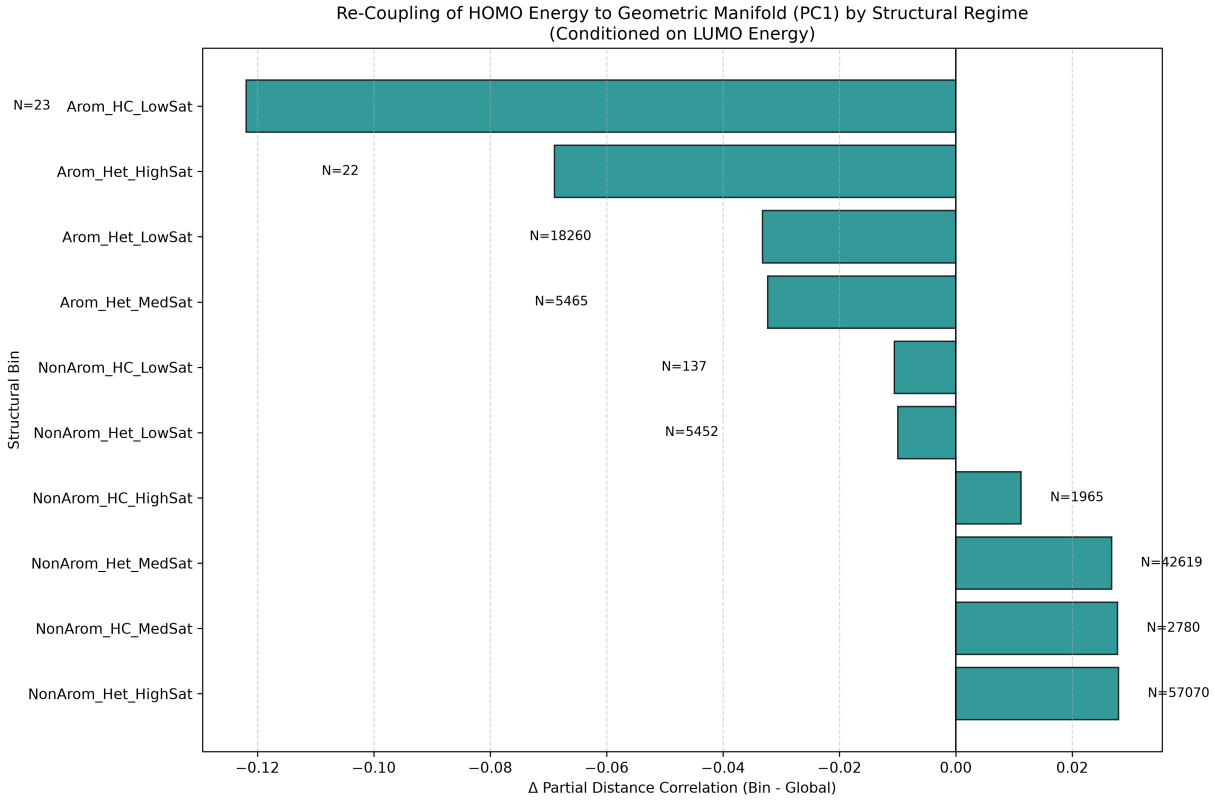

Figure 5: Difference in partial distance correlation (Δ) between the HOMO energy and the geometric-thermodynamic manifold (PC1) across distinct structural regimes relative to the global baseline, conditioned on LUMO energy. The chart illustrates that the decoupling of frontier orbital energies from global geometric properties is not a universal constant but is highly sensitive to molecular architecture. Highly saturated, non-aromatic systems (e.g., `NonArom_Het_HighSat` and `NonArom_HC_MedSat`) exhibit positive Δ values, indicating a conditional re-coupling where the frontier orbital topology integrates with the geometric state. In contrast, aromatic systems (e.g., `Arom_Het_LowSat`) show negative Δ values, demonstrating that the presence of π-conjugation reinforces the orthogonality between electronic and geometric properties. Sample sizes (N) for each structural bin are annotated alongside the bars.

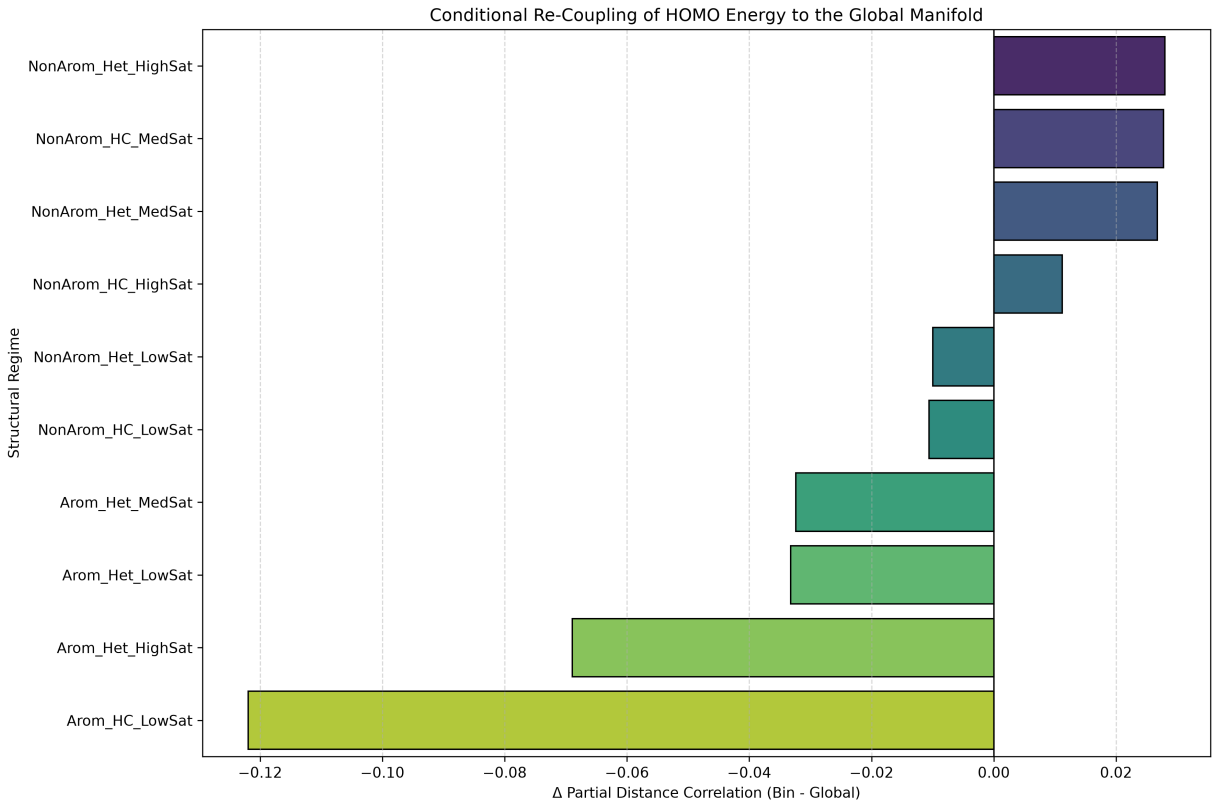

Figure 6: The change in partial distance correlation (Δ Partial Distance Correlation) between the HOMO energy and the geometric-thermodynamic manifold across distinct molecular structural regimes, relative to the global baseline. Positive values indicate a conditional re-coupling effect, which is strongest in highly saturated, non-aromatic systems (e.g., `NonArom_Het_HighSat`). Conversely, negative values demonstrate that the presence of aromaticity (e.g., `Arom_HC_LowSat` and `Arom_Het_LowSat`) reinforces the decoupling of frontier orbital energies from the global geometric state. This reveals that the orthogonality between electronic and geometric properties is not a universal constant, but rather a feature maintained in aromatic π-systems that partially breaks down as molecular saturation increases.

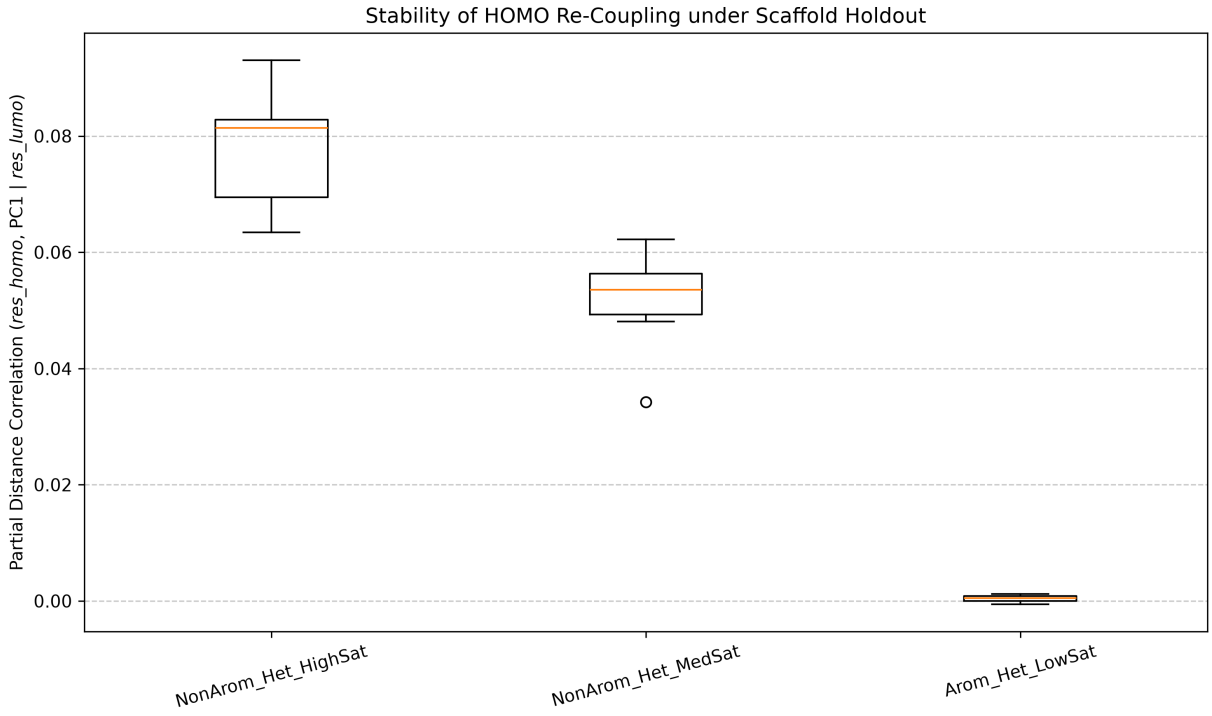

Figure 7: Box plots illustrating the stability of the partial distance correlation between the residual HOMO energy (res_{homo}) and the geometric-thermodynamic manifold (PC1), conditioned on res_{lumo}, evaluated across 10 scaffold-holdout folds. The distributions compare three distinct structural regimes: highly saturated heteroatomic systems (`NonArom_Het_HighSat`), moderately saturated systems (`NonArom_Het_MedSat`), and aromatic systems (`Arom_Het_LowSat`). The data reveal that the re-coupling of the frontier orbital energy to the geometric state is a robust, fundamental property of saturated molecules, which maintain a consistently positive correlation. In contrast, aromatic π-systems exhibit near-zero correlation across all folds, confirming that they robustly maintain electronic-geometric orthogonality independent of specific chemical scaffolds.

linear stoichiometric residuals and defining a unified geometric-thermodynamic manifold—predominantly characterized by the rotational constants B and C—we established a robust macroscopic baseline. Utilizing partial distance correlation across explicitly stratified structural regimes in the QM9 dataset, we quantified the precise architectural conditions that govern the integration of frontier orbital topology with the global molecular state.

Our findings demonstrate that the independence between frontier electronic and geometric properties is not absolute. While a global baseline analysis confirms a general decoupling, this orthogonality breaks down significantly in highly saturated molecular architectures. Specifically, saturated systems, particularly those containing heteroatoms (N, O, or F), exhibit a pronounced re-coupling of the HOMO energy to the geometric manifold. Conversely, the presence of aromaticity strongly reinforces the decoupling, indicating that π-conjugation is the primary structural driver maintaining electronic-geometric orthogonality.

Furthermore, mediation analysis revealed that the physical mechanism underlying this re-coupling diverges based on chemical composition. In moderately saturated hydrocarbons, the integration of the HOMO energy into the geometric manifold is almost entirely mediated by isotropic polarizability (α), suggesting a secondary effect driven by the deformability of the electron cloud. In contrast, highly saturated heteroatomic systems exhibit a direct topological integration that is independent of global polarizability.

The robustness of these structural dynamics was rigorously confirmed through a leave-one-scaffold-out validation, ensuring that the observed conditional re-coupling is an intrinsic property of the molecular architecture rather than an artifact of specific chemical scaffolds. Ultimately, this work refines our physical intuition of structure-property relationships, establishing that the frontier electronic and geometric property blocks are only truly independent in the presence of extended π-systems. In saturated environments, this fundamental independence partially, yet systematically, breaks down, highlighting the necessity of context-aware approaches when mapping the multidimensional property space of molecules.