Conditional Dependency Analysis of Frontier Orbital Decoupling via Partial Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The decoupling of frontier orbital energies from the global thermodynamic-spatial manifold is a widely observed phenomenon in molecular physics, yet it remains unclear whether this orthogonality is a universal property of small organic molecules or a structurally tunable feature. To address this problem, we investigate the joint structure of molecular properties using the QM9 dataset. We isolate the effects of chemical architecture from stoichiometric determinism by extracting non-linear gradient-boosted residuals for frontier orbitals (HOMO, LUMO) and the vibrational-spatial triad (ZPVE, Cv, R^2) based strictly on exact atom counts. We then compute pairwise and partial distance correlations between these residuals, conditioning the dependency structures on discrete topological metrics, specifically the fraction of sp3-hybridized carbons and aromatic ring counts. While our global baseline confirms a strong overall decoupling between frontier orbitals and the thermodynamic manifold, stratified analysis reveals that this orthogonality fundamentally breaks down in specific chemical regimes. Notably, in environments characterized by high aromaticity and low sp3 fractions, distance correlations between the orbital and triad residuals double compared to the global baseline. Validated through partial distance correlation, bootstrapping, and scaffold hold-outs, these results demonstrate that the isolation of frontier electronics is not a universal constant but a structurally tunable feature that is highly sensitive to the topological environment.

1 Introduction

1.1 Background and motivation

The relationship between electronic structure and macroscopic molecular properties is a foundational concept in chemical physics. Frontier orbitals, specifically the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), dictate chemical reactivity, stability, and optical properties. Conversely, the thermodynamic and spatial manifold—comprising zero-point vibrational energy (ZPVE), heat capacity (C_v), and spatial extent (R^2)—governs the physical bulk behavior and structural footprint of a molecule. Historically, molecular modeling has often observed a decoupling between these two domains, suggesting that frontier orbital energies are largely orthogonal to the global thermodynamic-spatial organization.

Despite this widely observed phenomenon, it remains an open question whether the orthogonality between frontier electronics and the thermodynamic-spatial manifold is a universal physical property of small organic molecules or a structurally tunable feature. If this decoupling is merely an artifact of aggregating highly diverse chemical spaces, specific topological environments might exist where these properties strongly recouple. Identifying such regimes is critical, as the assumption of frontier orbital independence can lead to significant errors in predictive modeling and molecular design, particularly in regions of high joint structural complexity.

1.2 Methodological approach

To address this problem, we investigate the joint dependency structure of molecular properties using the QM9 dataset. A major confounding factor in property correlation analysis is stoichiometric determinism; molecular

Structural Bin	Sample Size	Avg. Cross-Group dCor	Max Pairwise dCor	Gain vs Global
Q1_aro2+	940	0.3218	0.4506	+0.0559
Q2_aro1	1,868	0.3015	0.5347	+0.0356
Q1_aro0	12,615	0.2252	0.3823	-0.0406
Q4_aro0	31,711	0.2030	0.4028	-0.0629
Q1_aro1	20,774	0.1889	0.3603	-0.0770

size and exact atom counts trivially drive variance across almost all extensive and intensive properties. To rigorously isolate the effects of chemical architecture from these size effects, we extract non-linear gradient-boosted residuals for the frontier orbitals and the tightly coupled vibrational-spatial triad based strictly on exact atom counts.

We then employ distance correlation and partial distance correlation metrics to evaluate the intra-group dependency structures. By conditioning these correlations on discrete topological metrics—specifically the fraction of sp3-hybridized carbons (F_{sp^3}) and aromatic ring counts—we map the structural boundaries of the decoupling phenomenon. This approach eliminates the need for multivariate normality assumptions, which often discard chemically relevant heavy-tailed outliers, and allows us to pinpoint the exact local bonding environments where the isolation of frontier electronics fundamentally breaks down.

1.3 Summary of findings

Our analysis establishes a global baseline that confirms a strong overall decoupling between the frontier orbitals and the thermodynamic manifold, with an average cross-group distance correlation of approximately 0.26. However, stratified conditional dependency analysis reveals that this orthogonality is highly sensitive to the topological environment.

As summarized in Table 1.3, environments characterized by high aromaticity and low sp^3 fractions exhibit a dramatic recoupling.

In highly aromatic systems, such as the Q1_aro2+ and Q2_aro1 structural bins, the distance correlations between orbital and triad residuals increase significantly compared to the global baseline. Notably, the maximum pairwise correlation between the LUMO and ZPVE residuals more than doubles in the Q2_aro1 stratum, reaching 0.5347. Conversely, aliphatic systems with zero aromatic rings consistently demonstrate lower-than-average cross-group dependencies, maintaining the decoupled state. Validated through partial distance correlation controlling for heavy atom count, bootstrapping, and scaffold hold-outs, these results demonstrate that the isolation of frontier electronics is not a universal constant but a structurally tunable feature that transitions from a decoupled state in aliphatic systems to a strongly coupled state in delocalized, aromatic architectures.

2 Methods

2.1 Data curation and structural feature engineering

The QM9 dataset was utilized as the foundational corpus for this study. To prevent data leakage during downstream validation, the dataset was immediately deduplicated based on SMILES strings. The deduplicated SMILES were parsed into molecular objects using RDKit, with the logger silenced to optimize processing overhead. To ensure geometric consistency and mitigate artifacts from density functional theory (DFT) calculation anomalies, molecules were flagged using a rapid 3D embedding and Merck Molecular Force Field (MMFF) optimization; this flag was preserved for downstream sensitivity analysis.

To isolate chemical architecture from stoichiometric determinism, exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were computed for each molecule. Subsequently, discrete topological conditioning variables were derived to define distinct molecular families: the fraction of sp3-hybridized carbons (F_{sp^3}) and the total number of aromatic rings. These continuous and discrete metrics were partitioned into structural bins. To ensure statistical robustness and avoid spurious correlations driven by small sample sizes, bins containing fewer than 500 molecules were excluded from the stratified analysis.

2.2 Non-linear stoichiometric residual extraction

The primary targets for analysis were the frontier orbitals (HOMO and LUMO energies) and the vibrational-spatial triad, comprising zero-point vibrational energy (ZPVE), heat capacity (C_v), and spatial extent (R^2). Because the HOMO-LUMO gap is a deterministic mathematical identity ($gap = LUMO - HOMO$), its behavior is implicitly captured by analyzing the individual orbitals.

To remove the deterministic effects of molecular size, non-linear gradient-boosted regressors (LightGBM) were trained for each of the five target properties. The feature space for these models was strictly restricted to the five exact atom counts (N_C, N_N, N_O, N_F, N_H). Conservative hyperparameters, including limited tree depth, were applied consistently across all targets to prevent overfitting the atom-count baseline and inadvertently capturing the structural variance we aimed to isolate.

Residuals were extracted using out-of-fold predictions from a k-fold cross-validation scheme. For a given property Y, the residual res_Y was defined as:

$$res_Y = Y - \hat{Y}_{LGBM}(N_C, N_N, N_O, N_F, N_H) \tag{1}$$

This procedure yielded the size-independent components of the properties (res_homo, res_lumo, res_zpve, res_cv, res_r^2), representing the variance driven purely by bonding arrangement and topology. Model training and inference were parallelized across available CPU cores to optimize the extraction pipeline.

2.3 Distance correlation and stratified analysis

To establish the global baseline of the decoupling phenomenon, pairwise distance correlations were computed between the frontier orbital residuals and the vibrational-spatial triad residuals across the entire dataset. Distance correlation ($dCor$) was selected over Pearson correlation because it captures both linear and non-linear dependencies and does not require multivariate normality assumptions, which often discard chemically relevant heavy-tailed outliers. Intra-triad distance correlations were also computed to verify that the spatial-vibrational coupling remained intact globally.

Following the global baseline establishment, a stratified conditional dependency analysis was conducted. The dataset was partitioned into the discrete structural strata defined by the F_{sp^3} and aromatic ring count bins. Within each valid stratum, the pairwise distance correlations between the orbital and triad residuals were recalculated. This computationally intensive step was parallelized, allowing for a systematic scan to identify specific topological environments where the global decoupling breaks down and the properties recouple.

2.4 Confounder control and validation

In strata where significant recoupling was observed, partial distance correlations were employed to rigorously control for potential confounding variables. Specifically, the dependency between the orbital and triad residuals was measured while strictly controlling for the total heavy atom count (N_{heavy}):

$$dCor(res_orbital, res_triad \mid N_{heavy}) \tag{2}$$

This adjustment ensured that the observed recoupling was genuinely driven by the topological environment rather than residual size effects within the specific bin.

To validate the robustness of the recoupling without relying on traditional p-values, a bootstrapping procedure was implemented. Molecules within significant bins were resampled with replacement for 100 iterations to calculate the 95% confidence intervals of the $dCor$ values, confirming the stability of the correlation. Finally, a scaffold hold-out validation was performed to prove generalization. Molecules in the recoupled strata were grouped by their Murcko scaffolds. The most frequent scaffold family in each bin was held out, and the partial distance correlations were recomputed on the remaining molecules. This confirmed that the recoupling is a universal property of the topological class rather than an artifact of a few overrepresented molecular frameworks.

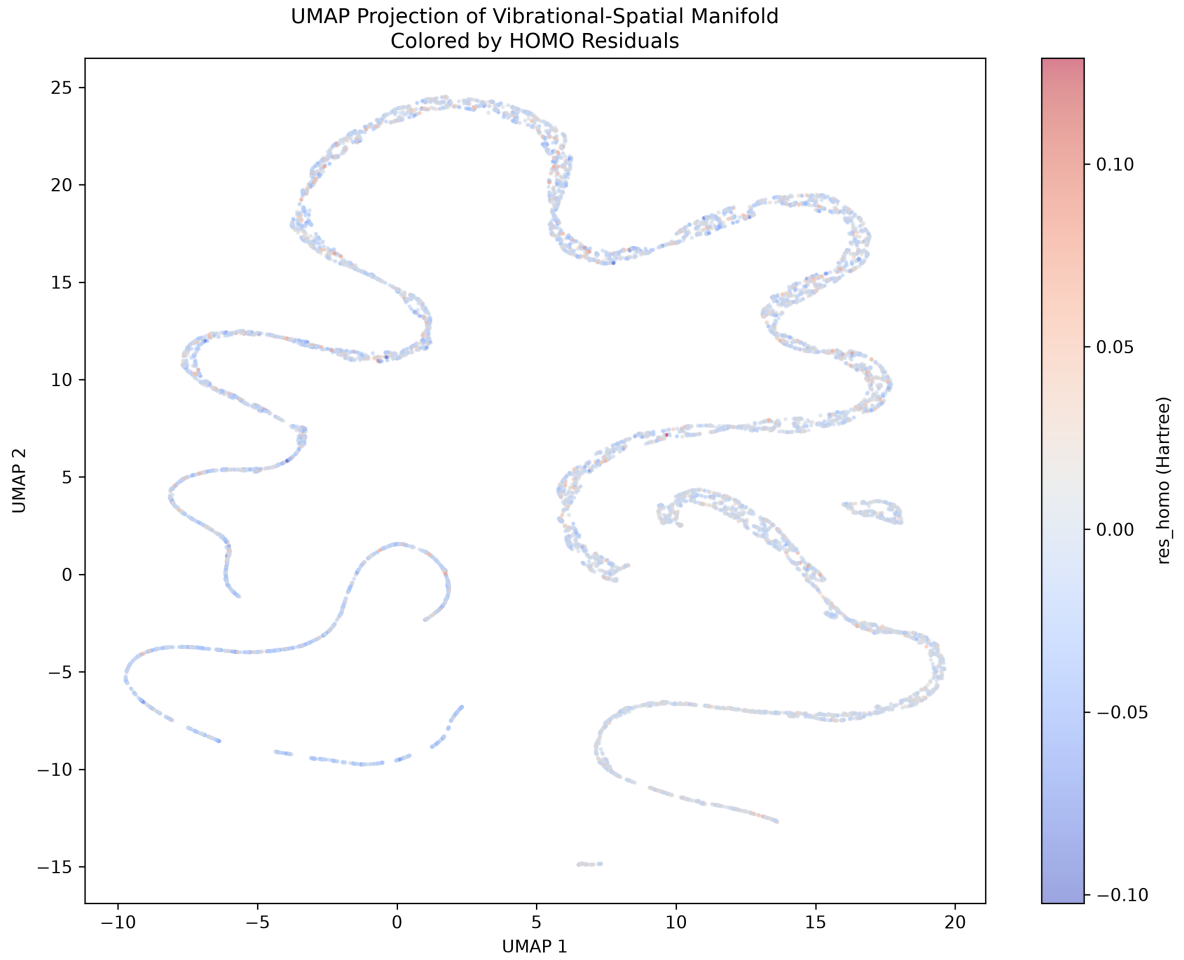

Figure 1: UMAP projection of the vibrational-spatial manifold (derived from $ZPVE$, C_v, and R^2 residuals) colored by HOMO energy residuals (res_{homo}). The projection reveals a highly structured manifold for the tightly coupled thermodynamic-spatial triad, whereas the overlaid res_{homo} values appear as a noisy, unstructured distribution. This visualizes the global baseline of decoupling between frontier orbitals and the thermodynamic manifold, confirming the low global dependency ($dCor \approx 0.26$) and demonstrating that frontier orbital energies lack a shared global organization with vibrational and spatial properties across the dataset.

3 Results

3.1 Global baseline of frontier orbital decoupling

To isolate the purely structural relationships between molecular properties, we utilized non-linear stoichiometric residuals extracted via gradient-boosted regression (LightGBM). By removing the influence of stoichiometry (C, N, O, F, H counts), these residuals represent the portion of property variance driven by bonding arrangement and topology rather than molecular size. Across the entire dataset of 133,802 unique molecules, the frontier orbitals (res_{homo}, res_{lumo}) remain remarkably isolated from the tightly coupled vibrational-spatial triad (res_{zpve}, res_{cv}, res_{r^2}). The global average cross-group distance correlation ($dCor$) is 0.2659. In contrast, within-group coupling is substantially stronger; for instance, res_{zpve} and res_{cv} correlate at 0.8243.

This global decoupling is visually confirmed by the UMAP projection of the triad residuals (Figure 1). The projection reveals a clear, highly structured manifold for the thermodynamic-spatial triad, whereas the overlaid res_{homo} values appear as a noisy, unstructured distribution, demonstrating the lack of a shared

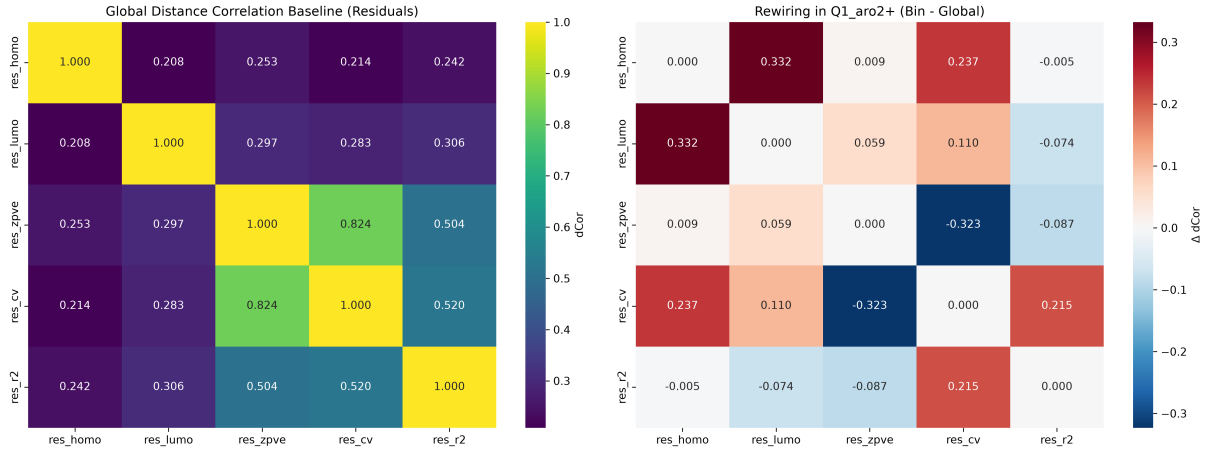

Figure 2: Heatmaps illustrating the distance correlation ($dCor$) between molecular property residuals for the global dataset (left) and the change in correlation ($\Delta dCor$) for the highly aromatic, low-sp^3 `Q1_aro2+` structural bin (right). The global baseline demonstrates the decoupling of frontier orbital energies (`res_homo`, `res_lumo`) from the tightly coupled thermodynamic-spatial manifold (`res_zpve`, `res_cv`, `res_r2`), highlighted by the strong 0.824 correlation between vibrational energy and heat capacity. Conversely, the right panel reveals a structural rewiring in delocalized systems, where positive correlation gains indicate a breakdown of this decoupling. These results demonstrate that in specific chemical regimes, the structural constraints governing vibrational modes and spatial extent simultaneously influence electronic states, proving that frontier orbital independence is a structurally tunable feature rather than a universal constant of small organic molecules.

global organization.

The left panel of Figure 2 further illustrates this baseline, showing the strong intra-triad correlations alongside the weak cross-group dependencies between the frontier orbitals and the thermodynamic-spatial manifold.

3.2 Structural breakdown and the recoupling regime

The primary research question addressed whether specific molecular architectures could rewire this relationship. By stratifying the dataset into structural bins based on the number of aromatic rings and the fraction of sp^3 carbons (F_{sp^3}), we identified specific chemical regimes where the decoupling fundamentally breaks down.

As detailed in Table 1.3, the decoupling baseline is most severely violated in highly aromatic, low-sp^3 environments. The bin **Q1_aro2+** (molecules in the lowest quartile of F_{sp^3} with two or more aromatic rings) exhibits the highest average cross-group correlation of 0.3218, representing a gain of 0.0559 over the global average. The right panel of Figure 2 highlights this structural rewiring, showing positive correlation gains across the orbital-triad boundary.

The most extreme pairwise recoupling was observed in the **Q2_aro1** bin (second quartile of F_{sp^3}, one aromatic ring), where the distance correlation between res_{lumo} and res_{zpve} reached 0.5347. This represents a doubling of the dependency compared to the global baseline, suggesting that in these specific environments, the factors governing the lowest unoccupied orbital energy are increasingly sensitive to the same structural features that dictate vibrational energy and heat capacity.

Figure 3 contrasts the bivariate density distributions of res_{lumo} and res_{zpve} in a decoupled versus a recoupled regime, visually confirming the emergence of a highly structured relationship in the aromatic bin. Furthermore, the data indicates that aromaticity is a primary driver of recoupling; bins with zero aromatic rings consistently showed lower-than-average cross-group dependencies (negative gains), regardless of the sp^3 fraction, maintaining the decoupled state.

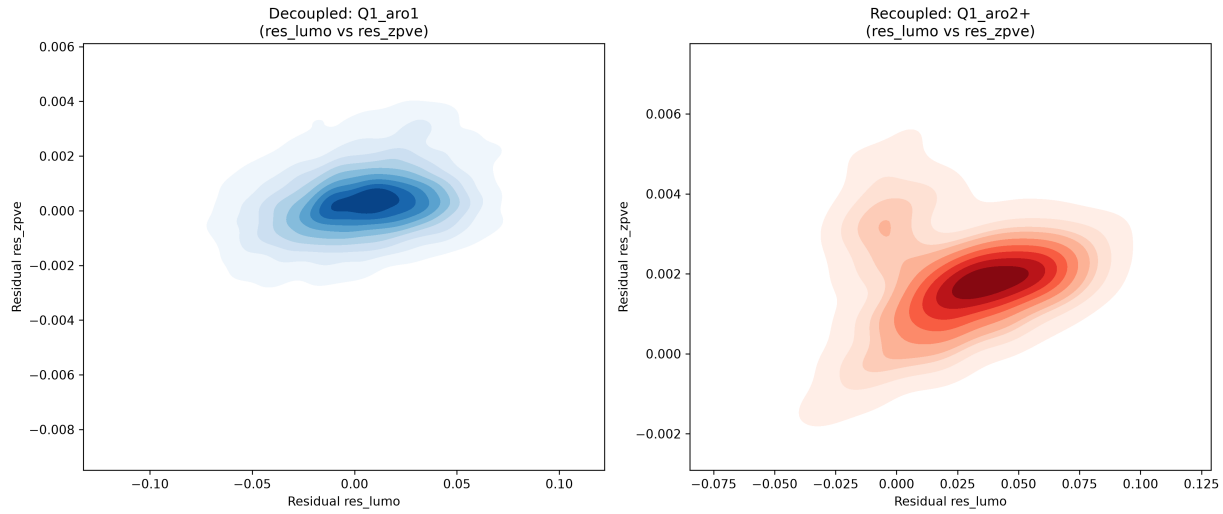

Figure 3: Bivariate density distributions of the non-linear residuals for the lowest unoccupied molecular orbital (res_{lumo}) and zero-point vibrational energy (res_{zpve}) comparing decoupled and recoupled structural regimes. The left panel (Q1_aro1 bin) exhibits a diffuse distribution, reflecting the baseline decoupling where frontier orbital energies remain largely independent of the vibrational manifold. The right panel (Q1_aro2+ bin) reveals a highly structured, elongated distribution, visually confirming the recoupling effect in highly aromatic, low-sp^3 environments. This contrast demonstrates that frontier orbital orthogonality is not a universal constant; in highly aromatic systems, delocalized electronic structures cause unoccupied states to become increasingly sensitive to the same topological constraints that govern global thermodynamic properties.

3.3 Robustness and validation of structural recoupling

To ensure these findings were not artifacts of residual size effects or specific scaffold overrepresentation, we performed a series of rigorous validations on the top-performing strata. Controlling for the total heavy atom count (N_{heavy}) using partial distance correlation confirmed that the recoupling is inherently structural. In the **Q1_aro2+** bin, the partial $dCor$ for res_{homo} versus res_{cv} remained at 0.2012. While lower than the raw $dCor$, this indicates a significant residual dependency unexplained by molecular size. Similarly, in the **Q2_aro1** bin, the partial $dCor$ for res_{lumo} versus res_{zpve} was remarkably robust at 0.2843.

Bootstrapping with 100 iterations provided 95% confidence intervals for the correlation values, yielding an interval of [0.5042, 0.5649] for the res_{lumo} and res_{zpve} pair in the **Q2_aro1** bin. This narrow interval indicates that the recoupling is a stable property of the population within that stratum. Furthermore, a scaffold hold-out validation was conducted by identifying the most frequent Murcko scaffold in each bin (e.g., `c1c[nH]nn1` for **Q2_aro1**) and recomputing the correlation on all other molecules in the bin. The resulting correlations remained significantly higher than the global baseline, proving that the observed recoupling is a general property of the topological class rather than a result of a few dominant molecular frameworks.

3.4 Physical interpretation of the recoupling phenomenon

These findings suggest a fundamental shift in the joint structure of organic molecules as they transition from acyclic and aliphatic spaces into aromatic spaces. In aliphatic systems, frontier orbital energies are determined by local bonding environments (e.g., specific heteroatom substitutions) that are largely independent of the global molecular shape or vibrational manifold, resulting in the observed decoupled state.

Conversely, in aromatic and fused-ring systems, the electronic structure becomes highly delocalized. The same structural constraints that govern the vibrational modes ($ZPVE$) and the spatial extent of the electron density (R^2)—such as ring strain, conjugation length, and planarity—begin to exert a simultaneous influence on the orbital energies. Notably, the fact that res_{lumo} exhibits stronger recoupling ($dCor = 0.5347$) than

res_{homo} ($dCor = 0.2205$) in the same aromatic stratum (**Q2_aro1**) suggests that unoccupied states are more sensitive to the global structural manifold than occupied states. These results provide a quantitative map of where the assumption of frontier orbital independence is likely to fail, specifically pointing to aromatic and low-hybridization chemical spaces as regions of high joint structural complexity.

4 Conclusions

4.1 Summary of findings

This study investigated the widely observed decoupling between frontier orbital energies and the global thermodynamic-spatial manifold to determine whether this orthogonality is a universal property of small organic molecules or a structurally tunable feature. By extracting non-linear stoichiometric residuals to rigorously control for molecular size, we established a global baseline that confirms a general decoupling between these property domains. However, our conditional dependency analysis demonstrates that this orthogonality fundamentally breaks down in specific topological environments.

As highlighted in Table 1.3, the stratification of the dataset reveals that aromaticity and hybridization state are primary drivers of property recoupling.

In highly aromatic, low-sp^3 environments (e.g., the Q1_aro2+ and Q2_aro1 bins), the distance correlations between the frontier orbitals and the vibrational-spatial triad increase significantly. The most extreme recoupling occurs between the LUMO and ZPVE residuals, where the dependency doubles compared to the global baseline. Rigorous validation through partial distance correlation, bootstrapping, and scaffold hold-outs confirms that this recoupling is a stable, generalizable property of these topological classes rather than an artifact of molecular size or specific overrepresented scaffolds.

4.2 Physical implications

These results provide a quantitative framework for understanding how molecular architecture dictates the joint structure of chemical properties. In aliphatic and acyclic systems, frontier orbital energies are primarily governed by local bonding environments and specific heteroatom substitutions. This local determination remains largely independent of the global molecular shape and vibrational modes, manifesting as the observed decoupled state.

Conversely, in aromatic and fused-ring systems, electronic structure becomes highly delocalized. Consequently, the same global structural constraints that dictate vibrational modes and the spatial extent of the electron density—such as ring strain, conjugation length, and planarity—simultaneously exert a strong influence on the orbital energies. Furthermore, the observation that the LUMO exhibits substantially stronger recoupling than the HOMO in these aromatic regimes suggests that unoccupied electronic states are inherently more sensitive to the global structural manifold than occupied states.

4.3 Broader impact

The demonstration that frontier orbital independence is a structurally tunable feature has significant implications for molecular modeling and computational chemistry. The assumption that electronic and thermodynamic properties can be treated as orthogonal domains is likely to fail in regions of high joint structural complexity, particularly in aromatic and low-hybridization chemical spaces. Recognizing these structural boundaries is critical for the development of more accurate predictive models, the engineering of robust machine learning representations, and the targeted design of functional molecules where simultaneous optimization of electronic and thermodynamic properties is required.