Residual Manifold Decomposition Uncovers the Structural Drivers and Complex Rewiring of HOMO Orthogonality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The computed properties of small organic molecules are typically dominated by stoichiometric size effects, yet the energy of the highest occupied molecular orbital (HOMO) exhibits a unique orthogonality to this global spatial-thermodynamic manifold. To uncover the localized structural drivers of this electronic decoupling, we performed a residual manifold decomposition on 133,802 deduplicated molecules from the QM9 dataset. By training Random Forest regressors on heavy atom counts, we extracted non-linear stoichiometric residuals for twelve non-redundant properties to eliminate baseline size effects. We then applied contrastive motif mining on explicit functional group representations to identify structural features enriched in the extreme tails of the HOMO residual distribution, validating our findings through split-half stability testing and distance correlation analysis. Our results reveal that specific nitrogen-containing motifs, notably nitriles and aromatic nitrogens, are the primary drivers of extreme HOMO deviations. Furthermore, these motifs do not merely shift isolated electronic states but actively rewire the joint property manifold, significantly altering the coupling between polarizability and zero-point vibrational energy. Crucially, the presumed orthogonality of the HOMO partially breaks down within these structural families, as the identified motifs simultaneously introduce non-negligible perturbations to thermodynamic baselines such as heat capacity. Ultimately, this study demonstrates that electronic decoupling is not an absolute chemical rule, but rather a population-level feature actively reconfigured by localized functional architectures.

1 Introduction

The rapid expansion of high-throughput computational chemistry has generated massive datasets of molecular properties, providing unprecedented opportunities to map the chemical space of small organic molecules. Within datasets such as QM9, computed properties are typically dominated by stoichiometric size effects. Thermodynamic and geometric properties, including internal energy (u_0) and zero-point vibrational energy ($zpve$), exhibit near-perfect linear correlations with simple heavy atom counts. While this stoichiometric determinism provides a robust baseline for macroscopic molecular behavior, it artificially collapses the dimensionality of the property space, obscuring the subtle, localized structural features that govern complex chemical phenomena.

Amidst this tightly bound spatial-thermodynamic manifold, the energy of the highest occupied molecular orbital (*homo*) presents a striking anomaly. Unlike the majority of molecular properties, *homo* exhibits a unique orthogonality to global molecular size, demonstrating a remarkably low correlation with atom counts ($R^2 = 0.0855$). This electronic decoupling suggests that frontier orbital energies are governed by localized structural architectures rather than global geometric constraints. However, the precise functional groups driving this decoupling, and whether they operate independently of the broader property manifold, remain poorly understood.

To uncover the physical mechanisms behind this phenomenon, we present a residual manifold decomposition of 133,802 deduplicated molecules from the QM9 dataset. By training non-linear Random Forest regressors on heavy atom counts, we extract stoichiometric residuals for twelve non-redundant properties, effectively eliminating baseline size effects. We then employ contrastive motif mining on explicit functional group representations to isolate the rare, localized bonding environments that exclusively govern *homo* variance. This approach bypasses black-box statistical dependencies, yielding highly interpretable structural rules validated through rigorous split-half stability testing and distance correlation analysis.

Property	Motif Mean	Bulk Mean	Cohen's d	p-value
res_homo	-0.182	0.069	-0.253	6.94×10^{-208}
res_zpve	-0.125	0.048	-0.173	2.95×10^{-201}
res_cv	0.228	-0.086	0.317	0.0

Our analysis reveals that specific nitrogen-containing motifs—most notably nitriles and aromatic nitrogens—are the primary drivers of extreme *homo* deviations. Beyond merely shifting isolated electronic states, we demonstrate that these motifs actively rewire the joint property manifold. The presence of these functional groups significantly alters the underlying coupling between seemingly disparate properties, such as polarizability (α) and zero-point vibrational energy, indicating that the global organization of molecular properties is highly sensitive to localized functional architectures.

Crucially, we challenge the assumption that the electronic decoupling of the HOMO is an absolute chemical rule. By evaluating the thermodynamic baselines of molecules containing these extreme *homo*-driving motifs, we observe a complex rewiring effect. As shown in Table 1, the identified motifs simultaneously introduce non-negligible perturbations to thermodynamic baselines, including heat capacity (*res_cv*) and zero-point vibrational energy (*res_zpve*).

The effect size on heat capacity even exceeds the target effect on the HOMO residual, demonstrating that the presumed orthogonality of the HOMO partially breaks down within these specific structural families. Ultimately, this study demonstrates that electronic decoupling is not a static, universal property of chemical space. Rather, it is a population-level feature that is actively reconfigured by localized functional groups. By mapping the residual manifold of the QM9 dataset, we provide a comprehensive framework for understanding how specific structural motifs isolate frontier electronics and rewire the macroscopic geometric and thermodynamic constraints of small organic molecules.

2 Methods

2.1 Data ingestion and preprocessing

The QM9 dataset was utilized as the foundational chemical space for this study. To ensure a unique set of molecular graphs, the dataset was deduplicated based on canonical SMILES strings, retaining the first occurrence of each structure, which yielded a final dataset of 133,802 unique molecules. The analysis focused on twelve non-redundant computed properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), highest occupied molecular orbital energy (*homo*), lowest unoccupied molecular orbital energy (*lumo*), electronic spatial extent (r^2), zero-point vibrational energy (*zpve*), internal energy at 0 K (u_0), heat capacity at 298.15 K (c_v), and atomization energy at 0 K (u_{0_atom}). Redundant thermodynamic properties and the HOMO-LUMO gap were excluded to prevent multicollinearity in downstream manifold analyses.

To mitigate the extreme heavy-tailed distribution of the rotational constant A while gracefully handling zero values, a $\log(1+x)$ transformation was applied. A binary feature, *is_linear*, was engineered to flag the nine perfectly linear molecules in the dataset for which A is exactly zero. Prior to structural parsing, the RDKit logger was silenced to prevent deprecation warnings from interfering with the computational pipeline.

2.2 Non-linear stoichiometric baseline and residual extraction

To eliminate the dominant stoichiometric size effects that artificially collapse the dimensionality of the property space, a residual manifold decomposition was performed. SMILES strings were parsed into RDKit molecule objects to compute the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. For each of the twelve non-redundant properties, a Random Forest regressor was trained using only the five heavy atom counts and the *is_linear* flag as input features. Model training was parallelized across all available CPU cores to manage the computational load.

The predictions from these stoichiometric baseline models were subtracted from the raw property values to isolate the non-linear residuals. These residuals were subsequently standardized using z-score normalization to ensure comparability across properties with different physical units. To confirm the successful removal

of stoichiometric determinism, the Pearson and distance correlations between the resulting residuals (e.g., res_α and res_r^2) and the heavy atom counts were computed, verifying that the correlations were effectively reduced to zero. Finally, the residuals of the nine linear molecules were inspected to ensure they did not manifest as extreme outliers that could skew the manifold topology.

2.3 Structural representation and contrastive motif mining

Rather than relying on high-dimensional and sparse topological fingerprints, the chemical space was characterized using explicit, predefined functional groups to maximize interpretability. The RDKit fragments module was employed to compute the exact counts of standard functional groups (e.g., aliphatic amines, aromatic rings, carbonyls, halogens, nitriles) for all deduplicated molecules.

To identify the localized structural drivers of HOMO orthogonality, contrastive motif mining was performed on the standardized *res_homo* manifold. The nine linear molecules were excluded from this step to prevent geometric anomalies from confounding the chemical motif analysis. The continuous distribution of *res_homo* was partitioned into three distinct cohorts: a lower tail (bottom 5% of values), an upper tail (top 5% of values), and a bulk population defined by a tight central band (the 40th to 60th percentile). This partitioning maximized the contrast between the baseline electronic state and the extreme deviations.

The Mann-Whitney U test was applied to identify functional groups that were statistically enriched in the extreme tails compared to the bulk population. To control for false positives across the multiple functional groups tested, p-values were adjusted using the Benjamini-Hochberg multiple hypothesis correction. Effect sizes were quantified using Cliff's Delta to extract the exact localized bonding environments driving extreme *homo* deviations.

2.4 Split-half stability validation

To rigorously verify that the identified structural motifs represent genuine chemical rules rather than statistical artifacts of the QM9 dataset, a split-half stability test was implemented. The dataset was randomly partitioned into two equal halves, and the contrastive motif mining procedure was repeated independently on both subsets. Only functional groups that demonstrated consistent, statistically significant enrichment (post-correction) and comparable effect sizes across both independent halves were retained as validated structural drivers.

2.5 Manifold rewiring and orthogonality verification

To investigate whether the identified structural features actively rewire the joint organization of the property manifold, the dataset was subset based on the presence or absence of the top validated functional groups. For each subset, the pairwise distance correlation matrix among all twelve property residuals was computed using the `dcor` package. A permutation test was utilized to estimate confidence intervals and statistically evaluate the differences between the correlation matrices of the motif-containing subsets and the bulk population. Difference heatmaps were generated to visually isolate which specific property pairs became coupled or decoupled in the presence of the motifs.

Finally, to test the assumption of HOMO orthogonality against the thermodynamic baseline, the distributions of *res_zpve* and *res_cv* for molecules containing the top *homo*-driving motifs were compared against the bulk population using Welch's ANOVA via the `pingouin` package. A Uniform Manifold Approximation and Projection (UMAP) of the 12-dimensional residual space was also generated, with points colored by the presence of the identified functional groups, to visually highlight the orthogonal displacement of *homo* and the localized gradients within the residual manifold.

3 Results

3.1 Stoichiometric residuals and the joint manifold

The removal of stoichiometric determinism from the QM9 dataset reveals a complex, higher-dimensional property space that is otherwise obscured by macroscopic size effects. While thermodynamic properties such

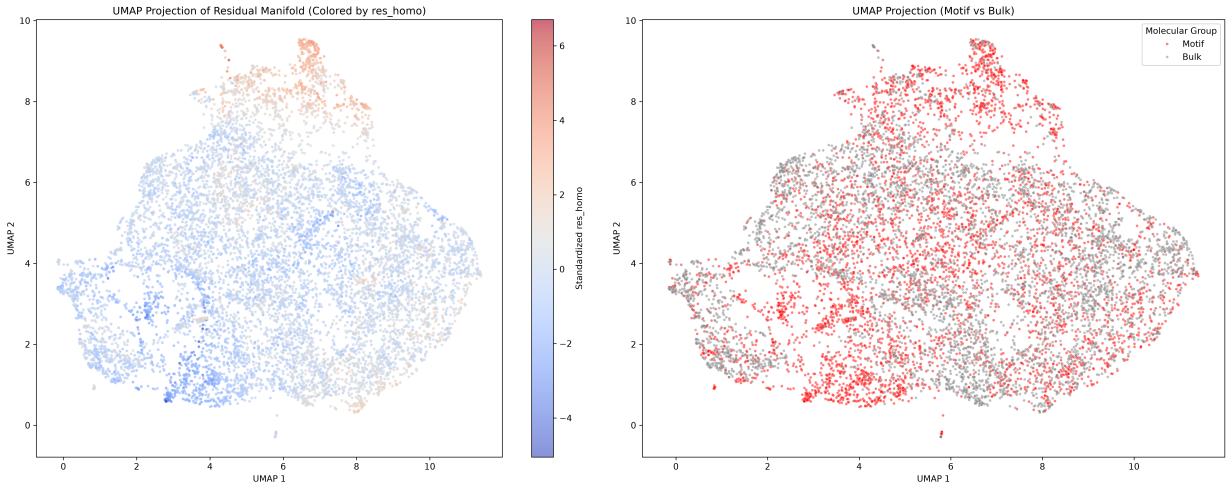

Figure 1: UMAP projection of the 12-dimensional residual property manifold. The left panel is colored by the standardized residual of the highest occupied molecular orbital energy (res_{homo}), displaying localized gradients rather than a global trend. The right panel contrasts the general molecular population (Bulk) with a subset containing specific nitrogen-bonding environments (Motif), such as nitriles and aromatic nitrogens. Together, these projections reveal that *homo* variance is driven by specific structural motifs rather than global molecular features. The nitrogen-containing motifs cluster in distinct sub-regions that correspond to the extreme tails of the res_{homo} distribution, visually confirming their role as the primary agents that decouple *homo* from the global spatial-thermodynamic manifold and rewire the joint property structure.

as internal energy (u_0) and zero-point vibrational energy ($zpve$) exhibit near-perfect linear correlation with heavy atom counts ($R^2 \approx 0.9999$ and 0.9971, respectively), the highest occupied molecular orbital energy ($homo$) is uniquely decoupled from this baseline ($R^2 = 0.0855$).

By standardizing the non-linear residuals derived from a Random Forest regressor trained on atom counts (C, N, O, F, H), we isolated the intrinsic chemical organization of these properties. In this residual space, res_{homo} remains largely orthogonal to the global spatial-thermodynamic manifold. For instance, distance correlation ($dCor$) analysis shows that while res_{zpve} and heat capacity (res_{cv}) maintain a moderate coupling in specific subsets ($dCor = 0.432$), res_{homo} exhibits near-zero distance correlation with the majority of other properties in the bulk population.

To visualize this organization, we generated a Uniform Manifold Approximation and Projection (UMAP) of the 12-dimensional residual space (Figure 1). The projection demonstrates that the manifold is not a single uniform cluster but is instead characterized by distinct sub-regions. When colored by res_{homo}, the manifold displays localized gradients rather than a global, continuous trend. This topology confirms that *homo* variance is not a product of global molecular features, but is instead driven by specific, localized structural motifs that occupy distinct regions of the chemical space.

3.2 Structural drivers of HOMO deviations

Contrastive motif mining across the 133,802 deduplicated molecules successfully identified the explicit functional groups responsible for these localized gradients. The analysis revealed that specific nitrogen-containing bonding environments are the most potent drivers of extreme res_{homo} deviations. These findings were robust across a split-half stability test, confirming them as generalizable chemical rules rather than dataset artifacts.

Nitriles (`fr_nitrile`) emerged as the strongest drivers of the lower tail of the res_{homo} distribution, exhibiting a substantial enrichment with a Cliff's Delta of 0.7525. As strong electron-withdrawing groups, nitriles significantly stabilize the HOMO. Conversely, aromatic nitrogens (`fr_Ar_N`) were highly enriched in the upper tail (Cliff's Delta: 0.5735), driving the HOMO energy upward. Primary amines and anilines (`fr_NH2`, `fr_aniline`) also showed consistent enrichment in the upper tail, with Cliff's Delta values of 0.3046 and 0.4261, respectively.

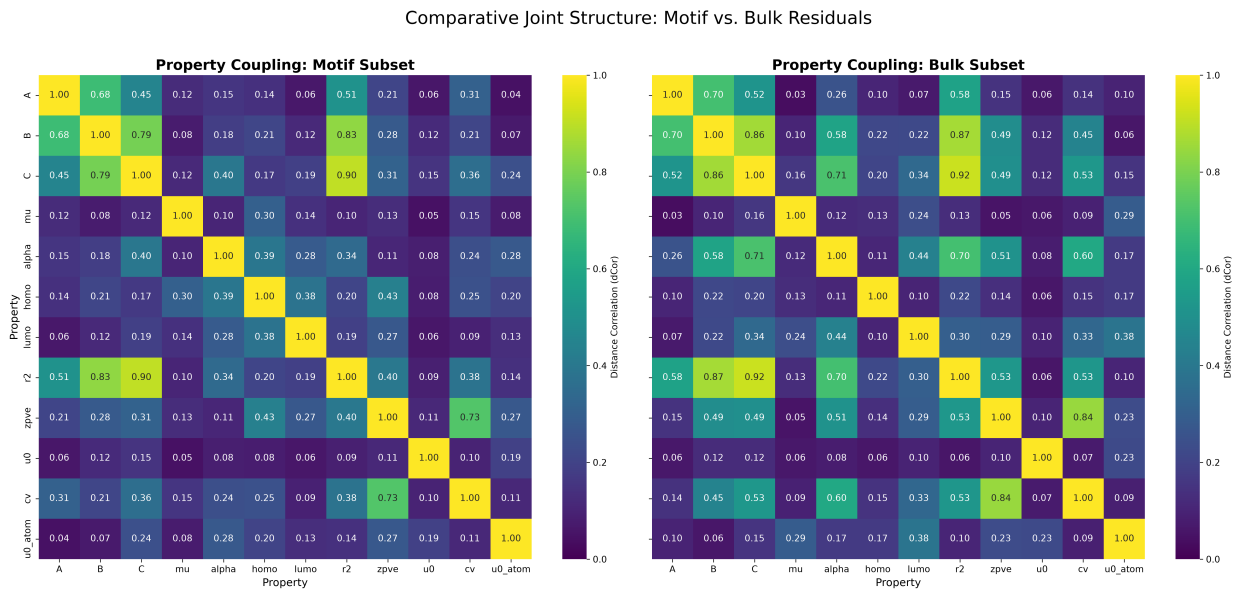

Figure 2: Distance correlation matrices comparing the joint property structure of the Motif subset (molecules containing nitriles or aromatic nitrogen) and the Bulk population. The heatmaps illustrate how specific nitrogen-containing motifs rewire the global property manifold. Overall, the Motif subset exhibits a lower mean distance correlation compared to the Bulk population, indicating a general fragmentation of the tightly bound property space. Notably, the presence of these motifs significantly disrupts the standard spatial-vibrational coupling, highlighted by a sharp decrease in correlation between polarizability (res_{α}) and zero-point vibrational energy (res_{zpve}). Conversely, the motifs increase the electronic-thermodynamic coupling, as seen in the stronger correlations between the residual HOMO energy (res_{homo}) and both res_{lumo} and res_{zpve}. These shifts demonstrate that the decoupling of HOMO is not an absolute chemical rule, but rather a feature of the bulk population that is actively reconfigured by localized functional architectures.

In contrast, oxygen-containing functional groups such as carbonyls (`fr_C_O`) and ethers (`fr_ether`) were significantly depleted in the upper tail (Cliff's Delta: -0.3087 and -0.2952). This suggests that these oxygen motifs either play a stabilizing role or are systematically excluded from the high-energy frontier electronic regimes dominated by nitrogen-rich architectures.

3.3 Rewiring of the joint property structure

The identification of these structural drivers raises a critical question: do these motifs merely shift isolated electronic states, or do they actively reconfigure the joint organization of the property manifold? To answer this, we compared the distance correlation ($dCor$) matrices of a "Motif" subset (molecules containing nitriles or aromatic nitrogen) against the "Bulk" population (Figure 2).

The analysis reveals a significant rewiring of the global structure. In the presence of these nitrogen-containing motifs, the overall sensitivity of the manifold increases, leading to a general fragmentation of the tightly bound property space. The mean distance correlation across all 12 properties drops from 0.2881 in the Bulk population to 0.2430 in the Motif subset.

More importantly, specific pairwise relationships are drastically altered, as highlighted by the difference heatmap (Figure 3). The most profound shift occurs between polarizability (res_{α}) and zero-point vibrational energy (res_{zpve}), which exhibits a correlation decrease of -0.4037. This indicates that nitrogen-driven motifs disrupt the standard physical relationship between molecular volume and vibrational energy. Simultaneously, the motifs induce an anomalous electronic-thermodynamic coupling, increasing the correlation between res_{homo} and both res_{lumo} and res_{zpve}.

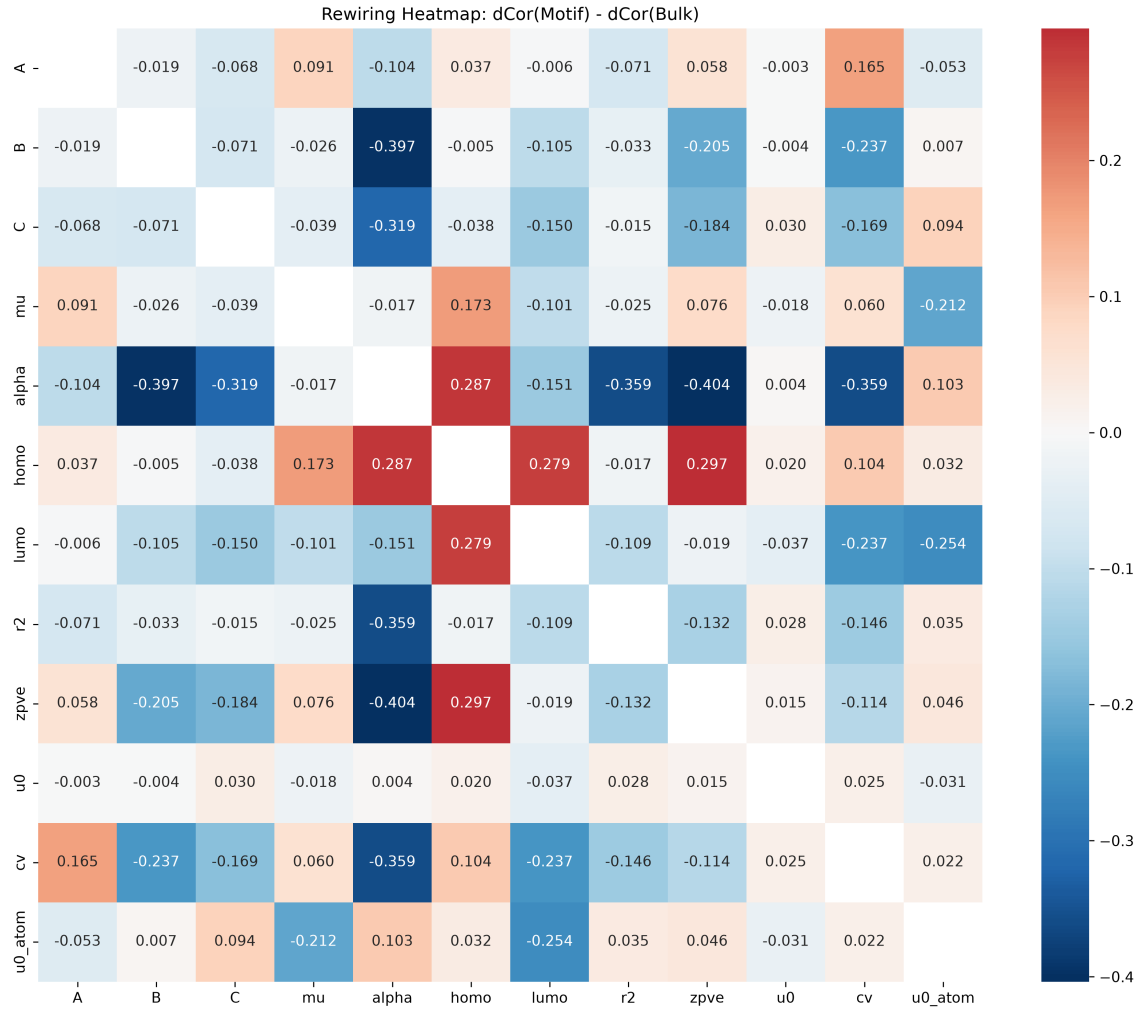

Figure 3: Heatmap illustrating the difference in distance correlation ($dCor$) between the motif-enriched subset (containing nitriles or aromatic nitrogen) and the bulk molecular population across 12 residual properties. This rewiring matrix demonstrates how specific nitrogen-containing motifs reconfigure the joint property manifold. The presence of these motifs increases the coupling of the highest occupied molecular orbital energy (res_{homo}) with both res_{lumo} and zero-point vibrational energy (res_{zpve}). Conversely, the most significant decoupling occurs between polarizability (res_{α}) and res_{zpve}, with a shift of -0.404, disrupting the standard relationship between molecular volume and vibrational energy. The overall prevalence of negative shifts reflects a global sensitivity where these localized functional architectures fragment the tightly bound property manifold, lowering the mean distance correlation compared to the bulk population.

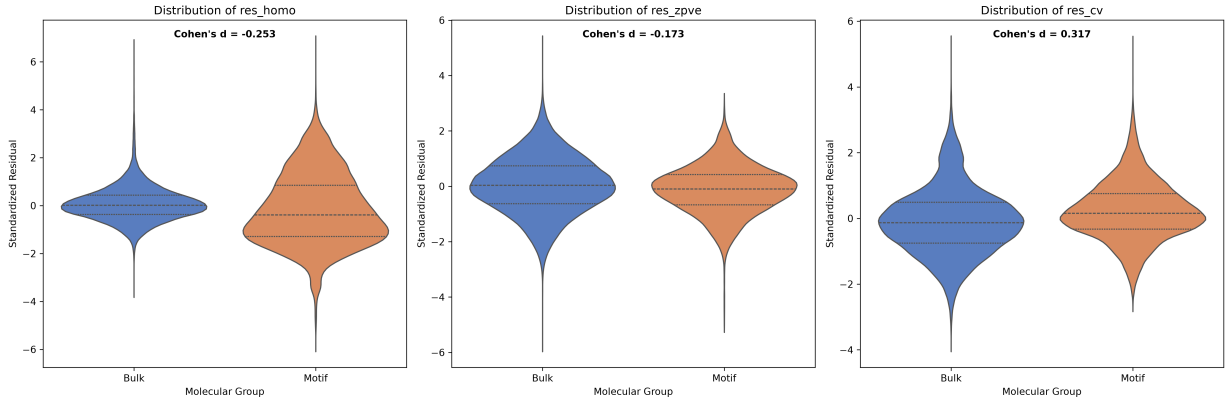

Figure 4: Violin plots comparing the distributions of standardized residuals for the highest occupied molecular orbital energy (res_{homo}), zero-point vibrational energy (res_{zpve}), and heat capacity (res_{cv}) between the bulk molecular population and a subset containing specific nitrogen-driven motifs. While these motifs significantly perturb frontier electronics (Cohen's d = -0.253 for res_{homo}), they are not entirely orthogonal to the thermodynamic baselines. The substantial shifts in res_{zpve} (Cohen's d = -0.173) and res_{cv} (Cohen's d = 0.317) reveal a complex rewiring effect. Rather than exclusively isolating res_{homo}, these nitrogen-containing motifs simultaneously perturb the molecule's heat capacity and vibrational profile, demonstrating that the global orthogonality of HOMO partially breaks down within specific structural architectures.

3.4 Orthogonality verification against the thermodynamic baseline

Finally, we tested the core assumption that *homo* orthogonality holds uniformly across the chemical space. If these structural drivers exclusively perturb frontier electronics, the thermodynamic baselines of the motif-containing molecules should remain indistinguishable from the bulk population. We evaluated this by comparing the distributions of res_{homo}, res_{zpve}, and res_{cv} using Welch's ANOVA (Table 1).

The results quantitatively demonstrate that the presumed orthogonality of the HOMO partially breaks down within these specific structural families. While the motifs significantly perturb res_{homo} (Cohen's d = -0.253), they simultaneously introduce non-negligible perturbations to the thermodynamic baselines. Strikingly, the effect size on heat capacity (res_{cv}, Cohen's d = 0.317) exceeds the target effect on the HOMO residual itself.

The violin plots in Figure 4 further illustrate this complex rewiring effect. The same nitrogen-containing motifs that isolate and shift frontier electronics also systematically alter the molecule's heat capacity and vibrational profile. Consequently, the electronic decoupling of the HOMO is not an absolute chemical rule, but rather a population-level feature of the structural bulk that is actively modified and reconfigured by localized functional architectures.

4 Conclusions

The computed properties of small organic molecules are typically dominated by macroscopic size effects, which obscure the localized structural features governing complex chemical phenomena. In this study, we applied a residual manifold decomposition to the QM9 dataset to strip away stoichiometric determinism and investigate the unique orthogonality of the highest occupied molecular orbital (*homo*) energy. By mapping the non-linear residual space of twelve non-redundant properties, we demonstrated that the decoupling of frontier electronics is not a static, universal law of chemical space, but rather a population-level feature that is actively reconfigured by specific functional architectures.

Through contrastive motif mining, we identified localized nitrogen-bonding environments—specifically nitriles and aromatic nitrogens—as the primary structural drivers of extreme *homo* deviations. However, our analysis reveals that these motifs do not merely shift isolated electronic states. Instead, they induce a complex rewiring of the joint property manifold. The presence of these nitrogen-rich groups fragments the

tightly bound global structure, significantly altering pairwise property relationships. Most notably, these motifs disrupt the standard physical coupling between molecular volume and vibrational energy, evidenced by a sharp decrease in the distance correlation between polarizability (res_α) and zero-point vibrational energy (res_zpve).

Crucially, we established that the presumed orthogonality of the HOMO partially breaks down within these specific structural families. As summarized in Table 1, the functional groups responsible for isolating and shifting frontier electronics simultaneously introduce non-negligible perturbations to the thermodynamic baselines.

The effect size of these motifs on heat capacity (res_cv) actually exceeds their target effect on the HOMO residual, indicating that the structural drivers of electronic decoupling also systematically alter a molecule's macroscopic thermodynamic and vibrational profile.

Ultimately, this work highlights the limitations of treating molecular property manifolds as uniform spaces. By moving beyond global stoichiometric baselines, we provide a highly interpretable framework for understanding how localized functional groups actively rewire the fundamental relationships between electronic, geometric, and thermodynamic variables.