Isolating Local Structural Drivers of the HOMO Decoupling Anomaly via Partial Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy persistently decouples from global thermodynamic and spatial manifolds, presenting an anomaly in molecular property modeling. To determine whether this decoupling is governed by localized chemical environments rather than global topology, we investigate the structural drivers of HOMO variance using the QM9 dataset. We extract atom-count residuals to isolate chemical structure from size effects and apply Partial Distance Correlation to quantify non-linear dependencies between HOMO residuals and local structural features, explicitly conditioning on spatial baselines to control for global molecular shape. Additionally, we employ bidirectional mediation models and Z-score boundary testing across acyclic, cyclic, and linear molecular partitions. Our results reveal that HOMO decoupling is a fundamental electronic property driven by local heteroatom environments—specifically nitrogen content and topological polar surface area—operating almost entirely independently of the thermodynamic state. While acyclic molecules demonstrate sensitivity to specific atom counts and conformational flexibility, cyclic structures exhibit near-total decoupling. Finally, this electronic isolation partially breaks down at the extreme geometric boundary of linear molecules, indicating that the joint structure of molecular properties depends on both local chemical identity and global geometric constraints.

1 Introduction

The highest occupied molecular orbital (HOMO) energy plays a critical role in determining molecular reactivity, stability, and electronic behavior. In large-scale quantum chemical datasets such as QM9, most macroscopic and spatial molecular properties—including heat capacity (C_v), zero-point vibrational energy (ZPVE), and electronic spatial extent (r^2)—organize into tightly bound joint manifolds dominated by molecular size and global topology. However, the HOMO energy exhibits a persistent and anomalous decoupling from these global thermodynamic and spatial manifolds. Preliminary exploratory analyses reveal that the distance correlation ($dCor$) between HOMO and these spatial-thermodynamic properties remains strikingly low ($dCor < 0.3$). This orthogonal trajectory presents a significant challenge for molecular property modeling, raising a fundamental question: is this decoupling an artifact of geometric outliers, or is it a distinct physical phenomenon governed by localized chemical environments?

To resolve this anomaly, we hypothesize that the variance in HOMO energy is primarily driven by local structural features—such as specific heteroatom content and topological polar surface area (TPSA)—rather than the global geometric constraints that dictate the rest of the property manifold. Testing this hypothesis requires a mathematical framework capable of isolating local chemical effects from the overwhelming influence of molecular size and shape.

1.1 Methodological approach

In this work, we introduce a robust analytical pipeline to isolate the structural drivers of HOMO decoupling. We first extract non-linear gradient-boosted atom-count residuals for properties in the QM9 dataset, effectively removing stoichiometric size effects. To quantify the non-linear dependencies between the HOMO residuals (res_homo) and local structural features, we apply Partial Distance Correlation (pdCor). By explicitly conditioning this correlation on a tightly bound spatial-vibrational baseline (res_zpve, res_c_v,

Partition	Total Effect (Path c)	Direct Effect (Path c')	Indirect Effect ($a \times b$)	Decoupling Index
Acyclic	-0.2589	-0.2468	-0.0121	**0.9531**
Cyclic	0.0203	0.0203	-0.0000	**1.0000**

res_r^2), we strictly control for global molecular shape, ensuring that the measured dependencies reflect purely local chemical environments.

Furthermore, we recognize that molecular flexibility and ring strain may fundamentally alter electronic behavior. Therefore, we partition the dataset into mutually exclusive structural categories: highly flexible acyclic chains, constrained cyclic populations, and extreme linear molecules. This partitioning allows us to evaluate whether the decoupling behavior is a universal electronic property or a context-dependent phenomenon.

1.2 Overview of key findings

Our investigation reveals that the HOMO decoupling anomaly is indeed a fundamental electronic property driven by local heteroatom environments, operating almost entirely independently of the thermodynamic state. To formally quantify this independence, we employ bidirectional mediation models comparing the structural influence on HOMO through the thermodynamic manifold versus the direct structural effect. We define a Decoupling Index as the proportion of the total structural effect that acts directly on the HOMO residual, bypassing thermodynamic mediators.

As demonstrated in Table 1.2, the structural influence on HOMO is overwhelmingly direct. In acyclic molecules, over 95% of the effect bypasses the thermodynamic manifold, while in cyclic structures, this decoupling is absolute.

Beyond establishing the magnitude of this decoupling, our partial distance correlation analysis identifies nitrogen content and TPSA as the primary local drivers of HOMO variance. We also observe that this electronic isolation is highly sensitive to structural boundaries. While cyclic structures exhibit near-total decoupling, acyclic molecules demonstrate a nuanced sensitivity to conformational flexibility. Finally, by applying Z-score boundary testing to linear molecules, we find that extreme geometric constraints force a partial re-coupling of the HOMO energy to the spatial manifold. Together, these results provide a chemically interpretable explanation for the orthogonal trajectory of HOMO, demonstrating that the joint structure of molecular properties depends intricately on both local chemical identity and global geometric constraints.

2 Methods

2.1 Data processing and residual extraction

To isolate intrinsic chemical variations from the dominant effects of molecular size and stoichiometry, we first established a baseline using atom-count features (C, N, O, F, H). For each of the 12 non-redundant properties in the QM9 dataset, we trained non-linear gradient-boosted decision trees to predict the property solely from this stoichiometric baseline. The residuals from these models (e.g., res_homo, res_zpve, res_c_v, res_r^2) were extracted to represent the size-independent variance of each property. Additionally, to tame extreme values associated with rotational constants, a $\log(A+1)$ transformation was applied prior to residual extraction.

2.2 Feature engineering and structural partitioning

To capture the local chemical environments hypothesized to drive HOMO variance, we computed a suite of structural descriptors using RDKit. These included topological polar surface area (TPSA), specific heteroatom counts (e.g., nitrogen and oxygen), the number of rotatable bonds, and ring strain metrics. Following feature standardization, the dataset was partitioned into three mutually exclusive structural categories based on empirically justified density thresholds: highly flexible acyclic chains ($N = 13,998$), constrained small-ring cyclic populations, and a boundary group of linear molecules. This partitioning enabled the comparative evaluation of local versus global structural hypotheses across distinct geometric regimes.

2.3 Partial distance correlation analysis

To quantify the non-linear dependencies between the HOMO residuals and the engineered local structural features, we employed Partial Distance Correlation (pdCor). To ensure that the measured correlations reflected purely local chemical effects rather than latent global shape dependencies, the pdCor calculations were explicitly conditioned on a tightly bound spatial-vibrational baseline comprising res_zpve, res_c_v, and res_r^2.

Statistical significance was established by comparing the observed pdCor values against a robust permuted null distribution. This baseline was generated by permuting the structural feature columns 1,000 times, with computations accelerated via parallelization across subsampling iterations. The results were subsequently aggregated and visualized using correlation heatmaps and hexbin scatter plots with overlaid global density contours.

2.4 Bidirectional mediation modeling

To distinguish true physical orthogonality from weak statistical mediation, we implemented a bidirectional mediation framework. We evaluated competing residual-based pathways: Model A, where local structure influences the thermodynamic state via the HOMO energy (Structure $\rightarrow res_homo \rightarrow res_u_0_atom$), and Model B, where structure influences HOMO via the thermodynamic state (Structure $\rightarrow res_u_0_atom \rightarrow res_homo$).

Focusing on Model B, we isolated the direct effect (Path c') of the structural driver on res_homo from the total effect (Path c) and the indirect effect ($a \times b$). To formally quantify this relationship, we defined a Decoupling Index as the proportion of the total structural effect that acts directly on the HOMO residual, bypassing the thermodynamic mediator:

$$\text{Decoupling Index} = \frac{c'}{|c'| + |a \times b|} \tag{1}$$

This metric isolates the degree to which HOMO variance is governed by physics entirely distinct from the thermodynamic manifold.

2.5 Boundary condition Z-score analysis

Standard correlation metrics are statistically unreliable for extremely small sample sizes. Therefore, to evaluate the decoupling behavior at the extreme geometric boundary of the dataset—specifically the nine linear molecules where rotation about the primary axis is undefined ($A = 0$)—we bypassed small-sample correlation entirely. Instead, we computed the Z-scores of res_homo and the spatial properties for these linear molecules relative to the global dataset distribution. By analyzing the mean absolute Z-score differences and the re-coupling of HOMO Z-scores with spatial Z-scores, we tested whether extreme geometric constraints force a breakdown of the electronic isolation observed in the broader dataset.

3 Results

3.1 Structural decoupling and the global manifold

The joint structure of molecular properties in the QM9 dataset is heavily dominated by molecular size effects. Once these macroscopic effects are removed by computing residuals against an atom-count baseline, a clear partition emerges between the spatial-vibrational block and the frontier-electronic sector.

As illustrated in Figure 1, properties such as polarizability (`res_alpha`), electronic spatial extent (`res_r2`), zero-point vibrational energy (`res_zpve`), and heat capacity (`res_cv`) exhibit moderate to strong mutual correlations. For instance, in acyclic molecules, `res_zpve` and `res_cv` maintain a correlation of $+0.47$, while `res_alpha` and `res_r2` correlate at $+0.76$. In stark contrast, the HOMO residual (`res_homo`) consistently exhibits near-zero correlations with these spatial and thermodynamic variables. Within the acyclic partition, the correlation of `res_homo` with atomization energy (`res_u0_atom`) is only -0.07, and its correlation with

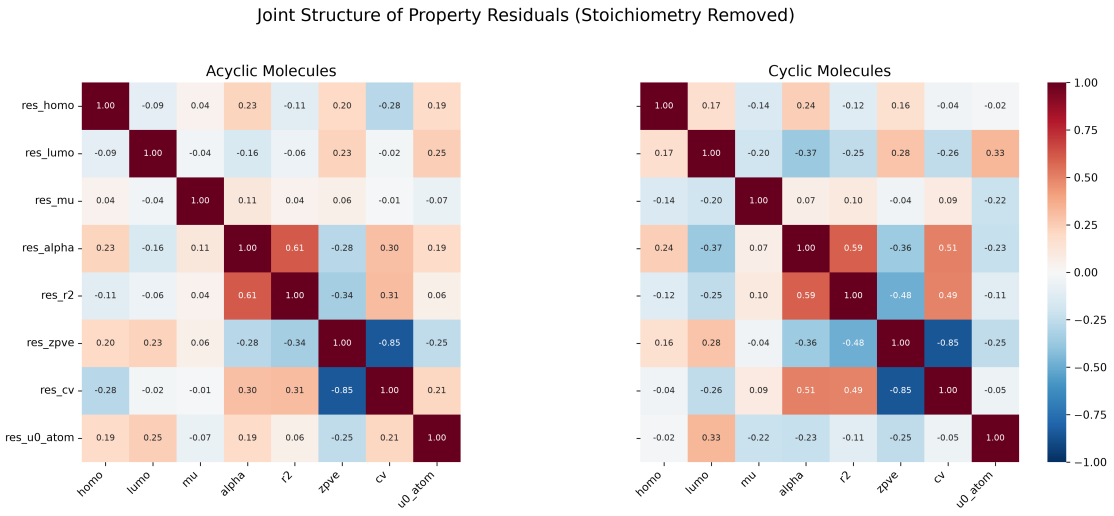

Figure 1: Correlation heatmaps of molecular property residuals for acyclic and cyclic partitions, demonstrating the joint structure of the dataset after removing stoichiometric size effects. The plot illustrates the pronounced decoupling of the frontier-electronic sector, specifically the HOMO energy (`res_homo`), from the spatial and thermodynamic manifolds. While spatial and vibrational properties such as `res_alpha`, `res_r2`, `res_zpve`, and `res_cv` exhibit moderate to strong mutual correlations, `res_homo` maintains near-zero correlations with these variables. This isolation is evident in acyclic chains and becomes nearly total in rigid cyclic molecules, revealing that HOMO variance is governed by local chemical environments rather than the global spatial-thermodynamic core.

`res_r2` is -0.15. In cyclic molecules, this electronic isolation is even more pronounced, with all correlations between `res_homo` and other properties falling below an absolute value of 0.10.

To formally test whether the structural influence on the HOMO energy is independent of the thermodynamic manifold, we conducted a bidirectional mediation analysis. Using nitrogen count as a primary structural driver and `res_u0_atom` as the mediator, we calculated the Decoupling Index, defined as the proportion of the total effect that is direct ($c'/(|c'| + |a \times b|)$). The results (Table 1.2) confirm that the structural driver's influence on the HOMO energy is almost entirely independent of the thermodynamic state. In acyclic molecules, 95.31% of the structural influence is direct, bypassing the thermodynamic manifold entirely. In cyclic molecules, this decoupling reaches 100.0%, indicating that HOMO variance is governed by physical mechanisms distinct from those dictating atomization energy or spatial extent.

3.2 Local structural drivers of HOMO variance

Given that the HOMO energy is decoupled from global size and shape, its variance must be driven by local chemical environments. We utilized Partial Distance Correlation (pdCor) to measure the non-linear dependency of `res_homo` on specific local descriptors, explicitly conditioning on the spatial manifold (`res_zpve`, `res_cv`, `res_r2`) to ensure the results reflect purely local effects rather than latent shape correlations.

The pdCor analysis identifies nitrogen count (`n_count`) and Topological Polar Surface Area (TPSA) as the primary drivers of HOMO residuals, with distinct behaviors observed across molecular partitions (Figure 2). For acyclic molecules, the strongest driver is `n_count` (pdCor = 0.0862, 95% CI: $[0.0727, 0.1074]$), followed by oxygen count (pdCor = 0.0594). In cyclic molecules, the primary drivers shift: TPSA exhibits the highest dependency (pdCor = 0.0309, 95% CI: $[0.0228, 0.0444]$), while `n_count` remains significant (pdCor = 0.0251).

Notably, the count of rotatable bonds (`rot_bonds`) shows a significant dependency in acyclic chains (pdCor = 0.0276), which drops substantially in the more rigid cyclic partition (pdCor = 0.0089). This contrast suggests that in flexible molecules, conformational degrees of freedom—likely those affecting the lone-pair orientations of nitrogen and oxygen atoms—interact with the HOMO level in ways that are sup-

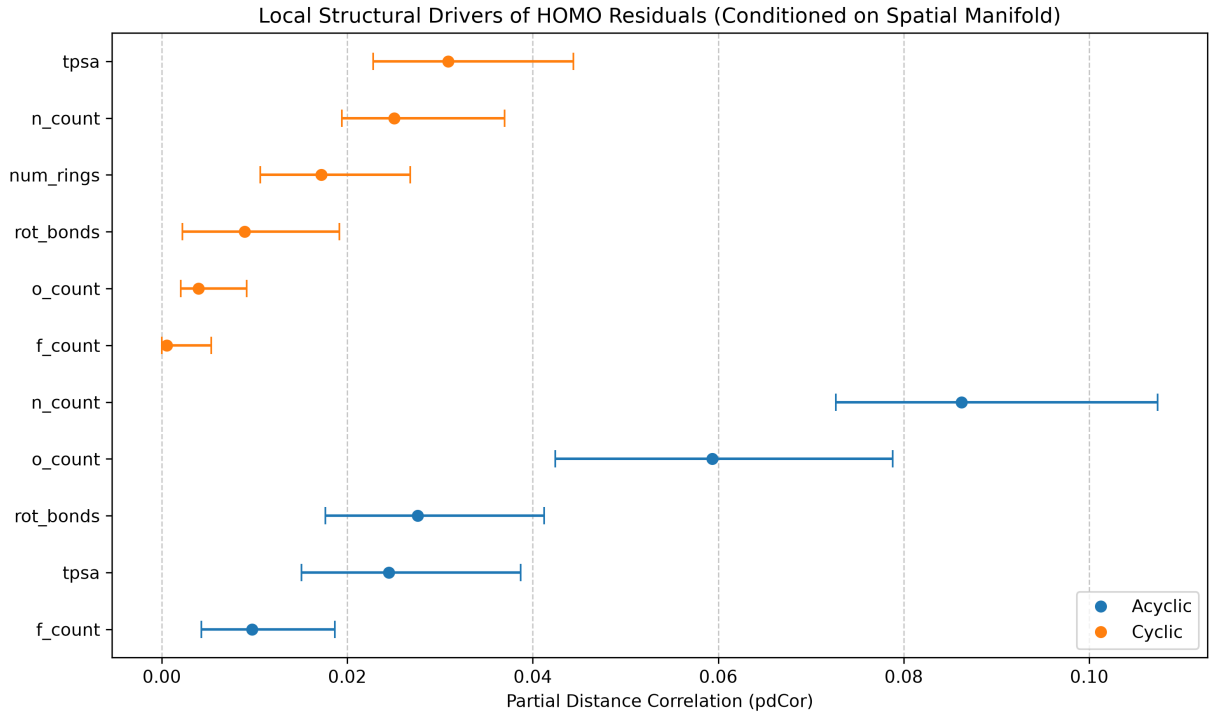

Figure 2: Forest plot of Partial Distance Correlation (pdCor) illustrating the local structural drivers of HOMO residuals, explicitly conditioned on the spatial manifold to isolate local chemical effects. The analysis compares acyclic (blue) and cyclic (orange) molecular partitions, with error bars representing 95% confidence intervals. We observe a distinct re-organization of dependencies between the two families: in acyclic molecules, nitrogen count (`n_count`) and oxygen count (`o_count`) are the dominant drivers, whereas in cyclic molecules, Topological Polar Surface Area (TPSA) exhibits the highest dependency. Furthermore, the influence of rotatable bonds (`rot_bonds`) is significantly stronger in acyclic chains than in rigid cyclic structures, indicating that conformational degrees of freedom interact with the HOMO level in flexible molecules but are suppressed by ring constraints.

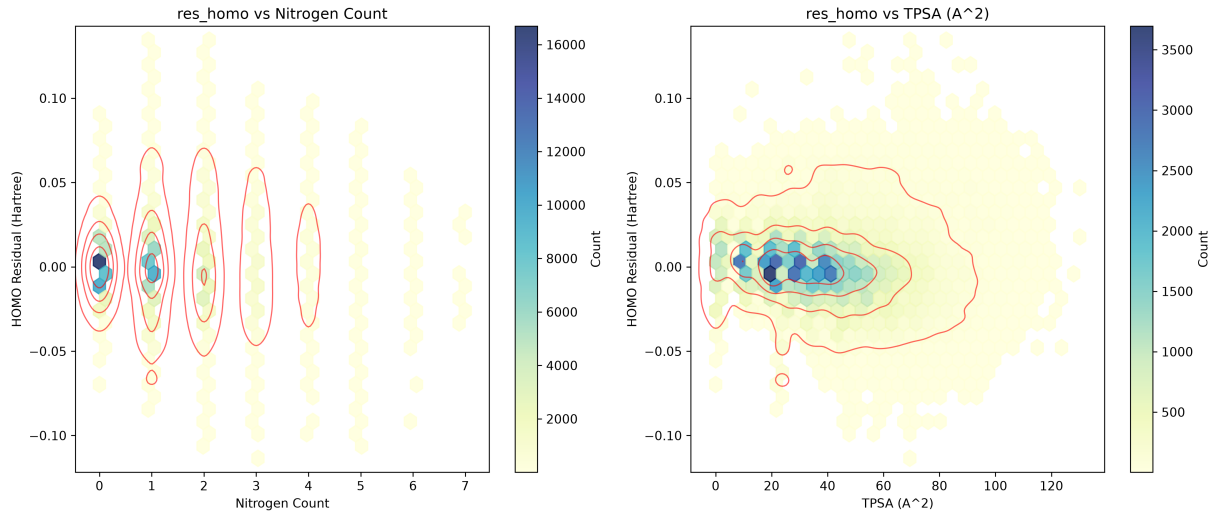

Figure 3: Hexbin scatter plots illustrating the distribution of the HOMO residual (`res_homo`) against its primary local structural drivers: nitrogen count (`n_count`) and Topological Polar Surface Area (TPSA). The left panel reveals a clear downward shift in the `res_homo` mean as nitrogen atoms are added, despite maintaining high variance. The right panel demonstrates a continuous relationship where increased TPSA, associated with heteroatoms, correlates with specific shifts in the electronic residual independent of the overall molecular size. Together, these distributions highlight that HOMO variance is governed by local chemical environments rather than the global spatial or thermodynamic manifolds.

pressed by ring constraints.

To visualize these dependencies, hexbin scatter plots were generated to illustrate the distribution of `res_homo` against its top drivers (Figure 3). The relationship between `res_homo` and `n_count` reveals a clear downward shift in the HOMO residual mean as nitrogen atoms are added, despite maintaining a high overall variance. Conversely, the TPSA plot demonstrates a more continuous relationship, where an increased polar surface area correlates with specific shifts in the electronic residual, independent of the molecule's overall size.

3.3 Boundary behavior in linear molecules

To determine whether this decoupling is a universal property or subject to geometric limits, we evaluated the extreme boundary of the dataset: the nine linear molecules (e.g., acetylene, hydrogen cyanide), where rotation about the primary axis is undefined ($A = 0$).

An analysis of Z-scores relative to the global population reveals that these linear molecules act as significant outliers in the electronic domain. For example, hydrogen cyanide exhibits a `res_homo` Z-score of -1.0135, whereas its spatial properties remain much closer to the global mean (spatial Z-avg $= -0.0444$). Across this boundary group, the mean absolute Z-score difference between the HOMO energy and the spatial manifold is 0.6268 ± 0.4212.

Crucially, in this extreme geometric regime, we observe a moderate re-coupling effect. The correlation between the HOMO Z-scores and the spatial Z-scores rises to $r = 0.5314$. This indicates that while the HOMO energy is decoupled in the vast majority of organic structures, extreme geometric constraints—which maximize the overlap between electronic states and the primary molecular axis—force the electronic residuals to partially align with the spatial manifold.

3.4 Summary of the decoupling anomaly

The "HOMO decoupling anomaly" emerges as a fundamental feature of the QM9 property manifold. It is not an artifact of molecular size, as the isolation persists in the residuals after removing stoichiometric ef-

fects. Instead, it represents an electronic dimension governed by local heteroatom environments—specifically nitrogen content and polar surface area—that operates independently of the spatial-thermodynamic core of the molecule. This re-organization is family-dependent: acyclic molecules show higher sensitivity to specific atom counts and conformational flexibility, while cyclic molecules exhibit a more uniform decoupling. Only at the extreme boundary of linear geometry does this electronic isolation begin to break down, suggesting that the "jointness" of molecular properties is a function of both chemical identity and the geometric constraints of the structure.

4 Conclusions

The persistent decoupling of the highest occupied molecular orbital (HOMO) energy from the global thermodynamic and spatial manifolds is a fundamental feature of the QM9 dataset, rather than an artifact of molecular size or geometric outliers. By extracting stoichiometric residuals to isolate intrinsic chemical variations, we have demonstrated that this anomaly represents a distinct electronic dimension governed by local heteroatom environments—specifically nitrogen content and topological polar surface area—that operates independently of the spatial-thermodynamic core of the molecule.

Our bidirectional mediation analysis formally quantifies this independence, confirming that the structural influence on the HOMO energy almost entirely bypasses the thermodynamic state. As detailed in Table 1.2, the Decoupling Index reveals that in acyclic molecules, over 95% of the structural effect is direct. In cyclic structures, this decoupling is absolute, indicating that HOMO variance is dictated by physical mechanisms entirely distinct from those governing atomization energy or spatial extent.

Furthermore, the re-organization of these local drivers is highly dependent on the molecular family. Partial distance correlation analysis highlights that acyclic chains exhibit a pronounced sensitivity to specific atom counts and conformational flexibility. In these flexible molecules, degrees of freedom such as rotatable bonds interact with the HOMO level, likely by altering lone-pair orientations. In contrast, rigid cyclic molecules suppress these conformational effects, resulting in a more uniform decoupling driven predominantly by polar surface area.

Finally, we established that this electronic isolation is not absolute across all possible geometries. At the extreme boundary of linear molecules, where rotation about the primary axis is undefined, the decoupling partially breaks down. The moderate re-coupling of HOMO and spatial Z-scores in this regime suggests that extreme geometric constraints force electronic states to align with the spatial manifold. Ultimately, these findings provide a chemically interpretable resolution to the HOMO anomaly, demonstrating that the joint structure of molecular properties is an intricate function of both local chemical identity and global geometric constraints.