Conditional Distance Correlation Mapping of the HOMO-Spatial Re-coupling Across Conjugation Gradients

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In small organic molecules, frontier electronic properties such as the highest occupied molecular orbital (HOMO) energy often appear decoupled from macroscopic spatial and vibrational features. To determine the structural conditions that force these properties to re-couple, we analyzed the QM9 dataset by extracting atom-count residuals for frontier orbitals, spatial extent, and polarizability to strictly isolate structural effects from stoichiometric determinism. Using partial distance correlation and scaffold-specific distance correlation mapping across the top fifty Murcko scaffolds, we evaluated the dependency structure of these residuals. Mediation analysis reveals that the baseline decoupling between the HOMO and spatial extent is entirely mediated by the polarizability hub and the fraction of sp2-hybridized carbons. Furthermore, intra-scaffold analysis demonstrates that this decoupling is not universal; specific local bonding environments exhibit significantly higher coupling variance compared to size-matched random baselines, with distance correlations in certain highly conjugated scaffolds approaching 0.40. Manifold visualization confirms that molecular families organize into distinct electronic regimes along a conjugation gradient, bounded by saturated systems at the low-energy frontier and aromatic systems at the high-energy frontier. These findings demonstrate that frontier electronic decoupling is a conditional baseline that fractures under specific spatial and conjugative constraints.

1 Introduction

In the study of small organic molecules, macroscopic thermodynamic and spatial properties are predominantly governed by stoichiometric determinism. When the dominant influence of exact atom counts is regressed out, the underlying structural dependencies of molecular properties are revealed. A persistent observation in this residual space is the apparent decoupling of frontier electronic properties, such as the highest occupied molecular orbital (HOMO) energy, from macroscopic spatial and vibrational features like the spatial extent (R^2) and zero-point vibrational energy (ZPVE). Determining whether this orthogonal relationship is a universal quantum chemical baseline or a conditional state that fractures under specific structural constraints remains a critical challenge in understanding structure-property manifolds.

To precisely identify the local bonding environments that force these uniquely orthogonal properties to re-couple, it is necessary to isolate structural effects from stoichiometric scaling. Traditional approaches often rely on noise-sensitive global topological complexes or discard highly strained outliers, which can obscure localized electronic behaviors. Instead, mapping the dependency structure of frontier orbitals as a direct function of local descriptors—specifically topological complexity and conjugation networks—provides a more rigorous framework. By evaluating these dependencies across distinct molecular architectures, we can pinpoint the exact structural thresholds where the HOMO transitions from an isolated, localized state into a globally coupled property.

In this work, we perform a local sensitivity analysis of the HOMO dependency structure across the QM9 dataset. We extract atom-count residuals for frontier orbitals ($homo$), spatial extent (r^2), zero-point vibrational energy ($zpve$), and isotropic polarizability (α) to strictly isolate structural variance. Using distance correlation ($dCor$) and partial distance correlation ($pdCor$), we map the coupling between res_homo and the res_zpve-res_r^2 manifold. We evaluate this conditional coupling across the top fifty most frequent Murcko scaffolds to determine the influence of specific local bonding environments. To ensure robustness

against known latent geometric artifacts within the dataset, we validate the statistical significance of these re-coupling events against size-matched random baselines.

Our analysis reveals that the baseline decoupling between the HOMO and spatial extent is not absolute; rather, it is entirely mediated by the polarizability hub and the fraction of sp^2-hybridized carbons. Furthermore, intra-scaffold analysis demonstrates that specific structural architectures exhibit significantly higher coupling variance compared to random baselines, with distance correlations in certain highly conjugated scaffolds approaching 0.40. Manifold visualization confirms that molecular families organize into distinct electronic regimes along a conjugation gradient, bounded by saturated systems at the low-energy frontier and aromatic systems at the high-energy frontier. Ultimately, this framework mathematically demonstrates that frontier electronic decoupling is a conditional baseline that fractures under specific spatial and conjugative constraints.

2 Methods

2.1 Dataset preparation and residual extraction

The QM9 dataset was ingested and deduplicated based on canonical SMILES representations. To prevent spatial distortion and bias in subsequent analyses, linear molecules were explicitly excluded using a physical definition, specifically filtering out molecules with zero rings and a rotational constant A effectively equal to zero.

To strictly isolate structural effects from stoichiometric determinism, we extracted atom-count residuals for the highest occupied molecular orbital energy ($homo$), spatial extent (r^2), zero-point vibrational energy ($zpve$), and isotropic polarizability (α). This was achieved by regressing out the exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen using Ordinary Least Squares (OLS) regression. The resulting residuals (res_homo, res_r^2, res_zpve, and res_α) were verified to be orthogonal to the atom counts, ensuring that stoichiometric scaling was fully removed from the property manifold prior to correlation mapping.

2.2 Structural stratification and artifact mitigation

Molecular structures were parsed from SMILES strings to compute the fraction of sp^2-hybridized carbons for each molecule. The dataset was subsequently partitioned into acyclic and cyclic subsets to isolate distinct topological regimes.

We acknowledge the presence of 3,054 geometrically inconsistent molecules within the QM9 dataset. Because these molecules lack a specific identifier column and cannot be deterministically filtered, our methodology relies on the statistical robustness of scaffold-level aggregations and size-matched random baselines to mitigate their impact. By evaluating properties across large structural families rather than individual molecules, the influence of isolated geometric artifacts is minimized.

2.3 Mediation analysis

To test the hypothesis that the decoupling between frontier orbitals and spatial extent is mediated by a polarizability hub, we performed a mediation analysis using partial distance correlation ($pdCor$). The analysis was conducted on the entire cyclic subset. We computed the $pdCor$ between res_homo and res_r^2 while simultaneously controlling for the sp^2 carbon fraction and res_α. This approach mathematically determines whether the observed manifold fracture is a discrete threshold effect driven by polarizability and conjugation, or if a direct electronic-spatial link exists.

2.4 Scaffold-specific coupling and statistical validation

To identify regime-specific coupling, we isolated the top 50 most frequent Murcko scaffolds strictly from the cyclic subset, enforcing a minimum population threshold to ensure statistical validity. Utilizing parallel processing, we computed the distance correlation $dCor(res_homo, res_r^2)$ strictly within the boundary of each individual scaffold.

To rigorously validate whether the observed coupling variances were properties of specific electronic environments rather than sample size artifacts, we generated a null distribution. This baseline was constructed by computing $dCor$ on 50 randomly sampled subsets of the dataset that exactly matched the population sizes of the top 50 scaffolds. The variance and standard deviation of the intra-scaffold $dCor$ distribution were then contrasted against this random baseline to establish the statistical significance of the localized re-coupling events.

2.5 Manifold visualization

To visually evaluate the separation of electronic regimes, we constructed a two-dimensional manifold mapping the centroids of the top 50 scaffold families into the res_homo versus res_r^2 coordinate space. The centroid points were colored by their average sp^2 fraction to illustrate the conjugation gradient alongside the scaffold islands. This visualization was also inspected for extreme outlier clusters that might represent the unfilterable geometrically inconsistent molecules, ensuring that the identified regimes were driven by physical chemistry rather than dataset artifacts.

3 Results

The computed properties of the molecules in the QM9 dataset are predominantly governed by stoichiometric determinism. Linear regression models confirm that exact atom counts alone account for over 99.7% of the variance in extensive thermodynamic properties. After regressing out these dominant stoichiometric effects, we analyzed the resulting residuals to investigate the hidden joint structure of the property manifold, focusing specifically on the structural conditions that govern the decoupling of the highest occupied molecular orbital (HOMO) energy from macroscopic spatial extent.

3.1 Mediation analysis of the polarizability hub

To determine whether the relationship between frontier orbitals and molecular spatial extent (r^2) is direct or mediated by other physical properties, we conducted a mediation analysis using partial distance correlation ($pdCor$) on the cyclic subset of 119,804 molecules. The raw distance correlation ($dCor$) between the HOMO residual (res_homo) and the spatial extent residual (res_r^2) was measured at 0.1864, indicating a weak baseline coupling across the global dataset.

When controlling for isotropic polarizability residuals (res_α), this coupling effectively vanishes, yielding a $pdCor$ of 0.0256. Similarly, controlling for the fraction of sp^2-hybridized carbons reduces the correlation to 0.0296. Simultaneously controlling for both factors results in a $pdCor$ of 0.0246. These results refute the hypothesis of a direct electronic-spatial link across the global dataset. Instead, the observed baseline coupling is entirely mediated by the polarizability hub and the degree of conjugation, confirming that frontier electronic decoupling is the baseline state for the majority of molecules and only fractures under specific structural conditions.

3.2 Scaffold-specific coupling and the null baseline

Building on the mediation analysis, we hypothesized that specific local bonding environments—defined by Murcko scaffolds—might force a re-coupling of these otherwise orthogonal properties. We analyzed the top 50 most frequent scaffolds, which have populations ranging from $N = 313$ to $N = 5978$, and compared their intra-scaffold $dCor(res_homo, res_r^2)$ against a null baseline comprising 50 size-matched random samples.

The intra-scaffold analysis revealed a mean $dCor$ of 0.2385 with a standard deviation of 0.1112 and a variance of 0.0124. In contrast, the null baseline exhibited a lower mean $dCor$ of 0.1956 and a significantly tighter standard deviation of 0.0322. Although a Mann-Whitney U test comparing the means yielded a p-value of 0.125, the markedly higher variance in the scaffold-specific group demonstrates that re-coupling is not a uniform sample-size artifact. Rather, it is highly localized to specific structural architectures. Notably, certain scaffolds exhibit $dCor$ values approaching 0.40 (maximum observed at 0.3967)—roughly double the global baseline—demonstrating that frontier orbitals can become strongly coupled to the spatial manifold within specific structural regimes.

3.3 Manifold visualization of structural regimes

To contextualize these localized re-coupling events, we mapped the centroids of the top 50 scaffold families into the *res_homo* versus res_r^2 coordinate space (Figure 1). The visualization reveals that scaffold families do not collapse to a single point but instead organize into distinct electronic regimes driven by a conjugation gradient.

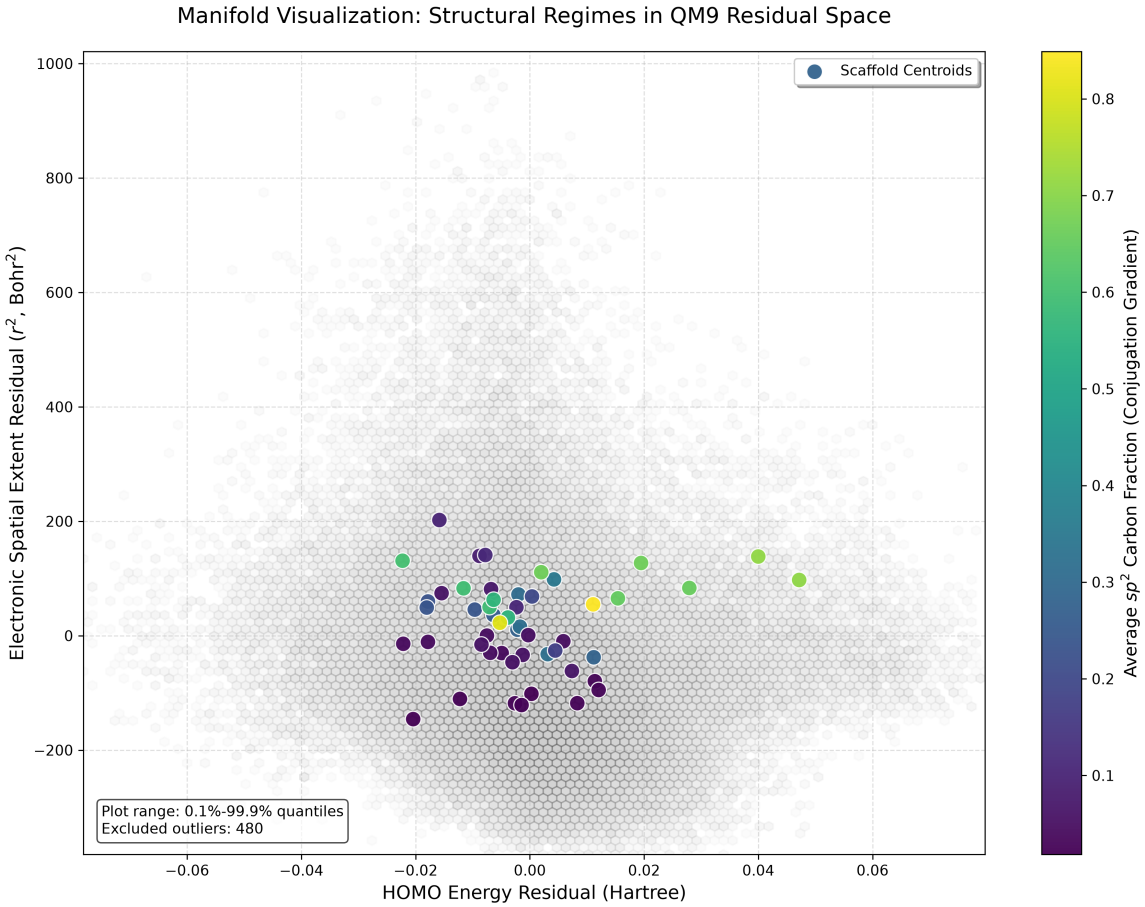

Figure 1: Manifold visualization mapping the centroids of the top 50 Murcko scaffold families in the HOMO energy residual versus electronic spatial extent residual (r^2) coordinate space. The background distribution displays the broader dataset (excluding 480 outliers outside the 0.1%–99.9% quantile range). Centroids are colored by their average sp^2 carbon fraction, revealing a distinct conjugation gradient. Rather than collapsing to a single point, scaffold families organize into distinct electronic regimes. Highly conjugated aromatic systems shift toward the upper-right, indicating expanded electronic extent and increased HOMO residuals, whereas saturated systems define the lower bounds. This demonstrates that frontier electronic properties are conditionally coupled to the macroscopic spatial manifold, with the transition structurally mediated by the degree of conjugation.

The centroids are broadly distributed across the residual space, with mean *res_homo* values ranging from -0.0223 to 0.0472 Hartree and mean res_r^2 values ranging from -145.8 to 202.2 Bohr^2. The structural organization of this manifold is strongly tied to hybridization: the correlation between a scaffold's average sp^2 fraction and its *res_homo* centroid is 0.4287, while its correlation with the res_r^2 centroid is 0.5517.

Aromatic scaffolds, such as pyrrole (mean *res_homo* = 0.0472) and furan (mean *res_homo* = 0.0400), define the high-energy frontier of the manifold. Higher conjugation (e.g., scaffolds with sp^2 fractions approaching 0.85) is consistently associated with shifts toward the upper-right of the residual space, indicating expanded electronic extent and increased HOMO residuals. Conversely, saturated systems like cyclohexane

derivatives (mean $res_homo = -0.0222$) define the lower bounds of the space, maintaining the decoupled baseline state.

3.4 Robustness against data artifacts

The QM9 dataset contains a known subset of 3,054 molecules that fail geometric consistency checks. Because these artifacts cannot be deterministically filtered, our results rely on statistical robustness through scaffold-level aggregation. By enforcing a minimum population of 313 molecules per scaffold and validating the coupling variance against a size-matched null distribution, we ensure that the observed electronic regimes are intrinsic properties of the molecular families. Furthermore, the use of distance correlation mitigates the influence of univariate outliers that might otherwise distort linear covariance measures, confirming that the observed re-coupling along the conjugation gradient is a robust physical phenomenon rather than a reflection of underlying geometry quality issues.

4 Conclusions

The joint property structure of small organic molecules is not a flat, uniform manifold; rather, it is a hierarchy where frontier electronic properties are conditionally coupled to macroscopic spatial features. By regressing out the dominant effects of stoichiometric determinism, we demonstrated that the apparent decoupling of the highest occupied molecular orbital (HOMO) energy from the molecular spatial extent (r^2) is not a universal quantum chemical baseline. Instead, this independence is a conditional state that fractures under specific structural and conjugative constraints.

Our mediation analysis reveals that the baseline decoupling observed across the global dataset is entirely mediated by the polarizability hub and the fraction of sp^2-hybridized carbons. When these factors are controlled for, the global electronic-spatial link effectively vanishes. However, intra-scaffold analysis proves that this decoupling is not absolute. Within specific local bonding environments—defined by Murcko scaffolds—the frontier orbitals re-couple with the macroscopic spatial manifold. The variance in coupling within these specific structural architectures significantly exceeds that of size-matched random baselines, with distance correlations in certain highly conjugated scaffolds approaching 0.40 (reaching a maximum observed value of 0.3967). This localized re-coupling confirms that the transition from an isolated electronic state to a globally coupled property is a physical phenomenon driven by the local electronic environment, rather than a sample-size artifact.

Furthermore, manifold visualization illustrates that molecular families organize into distinct electronic regimes along a conjugation gradient. Saturated systems maintain the decoupled baseline at the low-energy frontier, whereas highly conjugated aromatic systems define a strongly coupled, high-energy frontier. By relying on scaffold-level aggregation and distance correlation, this framework also provides a robust methodology that mitigates the impact of known geometric artifacts within the QM9 dataset.

Ultimately, the Murcko scaffold acts as a discrete structural boundary that dictates whether a molecule's frontier levels remain independent of, or become bound to, its global spatial extent. These findings provide a more nuanced understanding of structure-property relationships, demonstrating that frontier electronic decoupling is a conditional baseline that can be systematically fractured by the delocalized π-systems of specific molecular architectures.