Quantifying Frontier Orbital Decoupling via Partial Distance Correlation of Local Structural Descriptors

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how frontier molecular orbitals integrate into global macroscopic molecular properties is frequently obscured by baseline stoichiometric effects. To address this problem, we investigate the structural conditions governing the orthogonality of the highest occupied molecular orbital (HOMO) against a global spatial-thermodynamic manifold comprising zero-point vibrational energy, heat capacity, and electronic spatial extent. Utilizing the QM9 dataset, we isolate chemical residuals by removing compositional variance via Ridge regression and compute distance correlations between frontier orbital residuals, macroscopic properties, and local topological features such as conjugation index and heteroatom density. Our analysis reveals a fundamental disparity: while the lowest unoccupied molecular orbital (LUMO) demonstrates moderate integration into the global manifold and strong sensitivity to broader conjugation topology, the HOMO remains persistently decoupled and tied to localized electronic pathways. This structural isolation of the HOMO is robust across diverse organic scaffolds, breaking down only in the geometric limit of linear molecules. Additionally, in highly conjugated outliers, the HOMO-LUMO gap exhibits near-total independence from global thermodynamic states, driven instead by heteroatom density. Ultimately, these results indicate that the variance of the HOMO-LUMO gap is predominantly driven by the structural sensitivity of the LUMO, whereas the HOMO acts as an isolated, localized feature.

1 Introduction

Understanding how microscopic electronic features dictate macroscopic molecular properties is a central challenge in physical chemistry. Frontier molecular orbitals, specifically the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), are fundamental to chemical reactivity and electronic behavior. However, mapping these localized quantum states to global spatial and thermodynamic properties—such as zero-point vibrational energy, heat capacity, and electronic spatial extent—is frequently obscured by baseline stoichiometric effects. The sheer variance in molecular composition often masks the underlying topological rules governing property integration, making it difficult to determine whether frontier orbitals participate in the global spatial-thermodynamic manifold or remain isolated.

1.1 Methodological framework

To uncover the true structural dependencies of these properties, it is necessary to bypass standard stoichiometric baselines. Previous approaches often conflate bulk volume effects with local electronic pathways. In this work, we introduce a framework to mathematically formalize the structural conditions that govern the orthogonality or integration of frontier orbital energies against macroscopic states.

Utilizing the QM9 dataset, we isolate chemical residuals by removing compositional variance (the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen) via Ridge regression. This yields stoichiometric-free residual properties for the HOMO (res_homo), LUMO (res_lumo), zero-point vibrational energy (res_zpve), heat capacity (res_c_v), and electronic spatial extent (res_r^2). We then compute the distance correlation ($dCor$)—a metric capable of capturing both linear and non-linear dependencies—between these frontier orbital residuals, the macroscopic property triad, and local topological features such as the conjugation index and heteroatom density.

Relationship	Mean $dCor$	95% CI (Low)	95% CI (High)
res_lumo vs. Conjugation Index	0.3596	0.2741	0.4452
res_lumo vs. $triad_pc1$	0.2854	0.1845	0.3864
res_homo vs. $triad_pc1$	0.2252	0.0769	0.3735
res_homo vs. res_r^2	0.2063	0.1014	0.3113

1.2 Summary of findings

Our analysis reveals a fundamental disparity in frontier orbital behavior. We demonstrate that while the LUMO exhibits moderate integration into the global manifold and strong sensitivity to broader conjugation topology, the HOMO remains persistently decoupled. Instead, the HOMO acts as an isolated feature tied strictly to localized electronic pathways.

To ensure these observations are stable chemical properties rather than dataset-specific artifacts, we performed a rigorous scaffold-based stability analysis using 5-fold cross-validation split by Murcko scaffolds. As summarized in Table 1.2, the decoupling of the HOMO from the global manifold (represented by the first principal component of the triad, $triad_pc1$) and its spatial extent (res_r^2) is a robust trait across thousands of unique structural families, contrasting with the LUMO's stronger structural sensitivity.

This structural isolation of the HOMO breaks down only in the geometric limit of linear molecules, where electronic and spatial extents become coupled by geometric necessity. Furthermore, we explore highly conjugated outliers, revealing that the HOMO-LUMO gap exhibits near-total independence from global thermodynamic states in these regimes, driven instead by heteroatom density. Ultimately, this framework establishes that the variance of the HOMO-LUMO gap is predominantly driven by the structural sensitivity of the LUMO, whereas the HOMO refuses to join the global property network.

2 Methods

2.1 Data curation and stoichiometric baseline derivation

To isolate the topological rules governing property integration and build upon the framework introduced previously, we first established a stoichiometric-free baseline using the QM9 dataset. Following the removal of 83 duplicate entries based on SMILES strings, we explicitly filtered out 9 linear molecules (rotational constant $A = 0$) and 45 molecules with a zero dipole moment ($\mu = 0$). These were reserved for subsequent geometric sensitivity checks to ensure that any observed decoupling was not an artifact of geometric singularities.

For the remaining molecules, exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted. We then fitted a cross-validated Ridge regression model using degree-2 polynomial features of these atom counts to predict the raw properties: HOMO, LUMO, zero-point vibrational energy (ZPVE), heat capacity (C_v), and electronic spatial extent (R^2). The residuals extracted from these models (res_homo, res_lumo, res_zpve, res_c_v, res_r^2) represent the variance strictly independent of baseline composition. A robust scaler was applied to res_r^2 to mitigate variance distortion from any remaining geometric outliers.

2.2 Structural descriptor computation

To formalize the distinction between localized electronic pathways and bulk volume effects, we computed two distinct classes of structural descriptors using a parallelized routine. Global descriptors included the total heavy atom count and exact molecular weight. Local topological features included the maximum path length, the longest conjugated path (determined via the longest shortest-path on the subgraph of aromatic and conjugated atoms), the conjugation index (defined as the fraction of sp^2 and sp hybridized carbons), and heteroatom density (the sum of N, O, and F atoms divided by the heavy atom count). All computed features were standardized prior to downstream analysis.

2.3 Distance correlation and manifold integration

We quantified the integration of frontier orbitals into the global spatial-thermodynamic manifold using distance correlation ($dCor$), a metric chosen for its ability to capture both linear and non-linear dependencies.

A representative subsample of 25,000 molecules was drawn, strictly stratified by Murcko scaffold to maintain structural diversity, while verifying that the heavy atom count distribution matched the full dataset.

We computed the first principal component ($triad_pc1$) of the macroscopic triad (res_zpve, res_c_v, res_r^2). Utilizing parallelized distance matrix computations to manage the $O(N^2)$ complexity, we calculated the unconditional $dCor$ of res_homo and res_lumo against the individual manifold components and $triad_pc1$. Furthermore, we computed the $dCor$ of all residuals against the local and global structural descriptors to establish the scale-mismatch between local electronic excitations and global thermodynamic states.

2.4 Conditional stratification and outlier analysis

To investigate how structural constraints modulate these relationships, we shifted from unconditional correlations to a direct interaction analysis. The 25,000-molecule subsample was stratified into deciles based on the conjugation index and heteroatom density. Within each bin, the distance correlation between the frontier orbital residuals and the individual manifold components was recomputed.

Additionally, we isolated the top decile of the conjugation index to investigate the behavior of the HOMO-LUMO gap in highly conjugated outliers. The $dCor$ of the raw gap and its stoichiometric residual against the manifold components and local descriptors was computed and compared to the bottom decile. This determined whether the gap exhibits distinct structural dependencies driven by the LUMO's topological integration. Finally, the previously filtered high-symmetry and linear molecules were analyzed to test if the HOMO decoupling breaks down in restricted one-dimensional geometries.

2.5 Scaffold-based stability analysis

To confirm that the observed scale-mismatch and decoupling are stable chemical properties rather than dataset-specific artifacts, we performed a rigorous stability analysis. The entire filtered dataset was grouped by Murcko scaffolds, and a 5-fold cross-validation was executed where splits were strictly defined by scaffold. This ensured that structurally similar molecules were kept within the same fold, testing the generalizability of the relationships across unique structural families.

Within each fold, up to 5,000 molecules were sampled, and the unconditional distance correlations and local-descriptor correlations were recomputed. The mean, variance, and 95% confidence intervals of these effect sizes were recorded across the folds (Table 1.2), confirming the robustness of the HOMO's structural isolation against the global manifold.

3 Results

3.1 Abstract structure of the property manifold

Building upon the stoichiometric-free baseline established in the methods, we first examined the abstract structure of the QM9 property set. The macroscopic properties—zero-point vibrational energy (res_zpve), heat capacity (res_c_v), and electronic spatial extent (res_r^2)—form a highly coherent global manifold. In contrast, the highest occupied molecular orbital (HOMO) exhibits persistent decoupling from this macroscopic block, a phenomenon that dictates the variance of the HOMO-LUMO gap.

3.2 Unconditional integration and scale sensitivity

To quantify the integration of frontier orbitals into the global manifold, we evaluated their distance correlation ($dCor$) against the first principal component of the thermodynamic-spatial triad ($triad_pc1$). As illustrated in Figure 1, the residuals of the global manifold demonstrate a sharp divergence in how they incorporate frontier orbital information.

Specifically, the lowest unoccupied molecular orbital (res_lumo) exhibits a moderate integration into the global manifold ($dCor = 0.3046$), which is significantly higher than that of the HOMO ($dCor = 0.2343$), as shown in the top-right panel of Figure 1. This divergence is mirrored in individual property pairings; for instance, the correlation between res_lumo and res_zpve is 0.2841, whereas for res_homo and res_zpve,

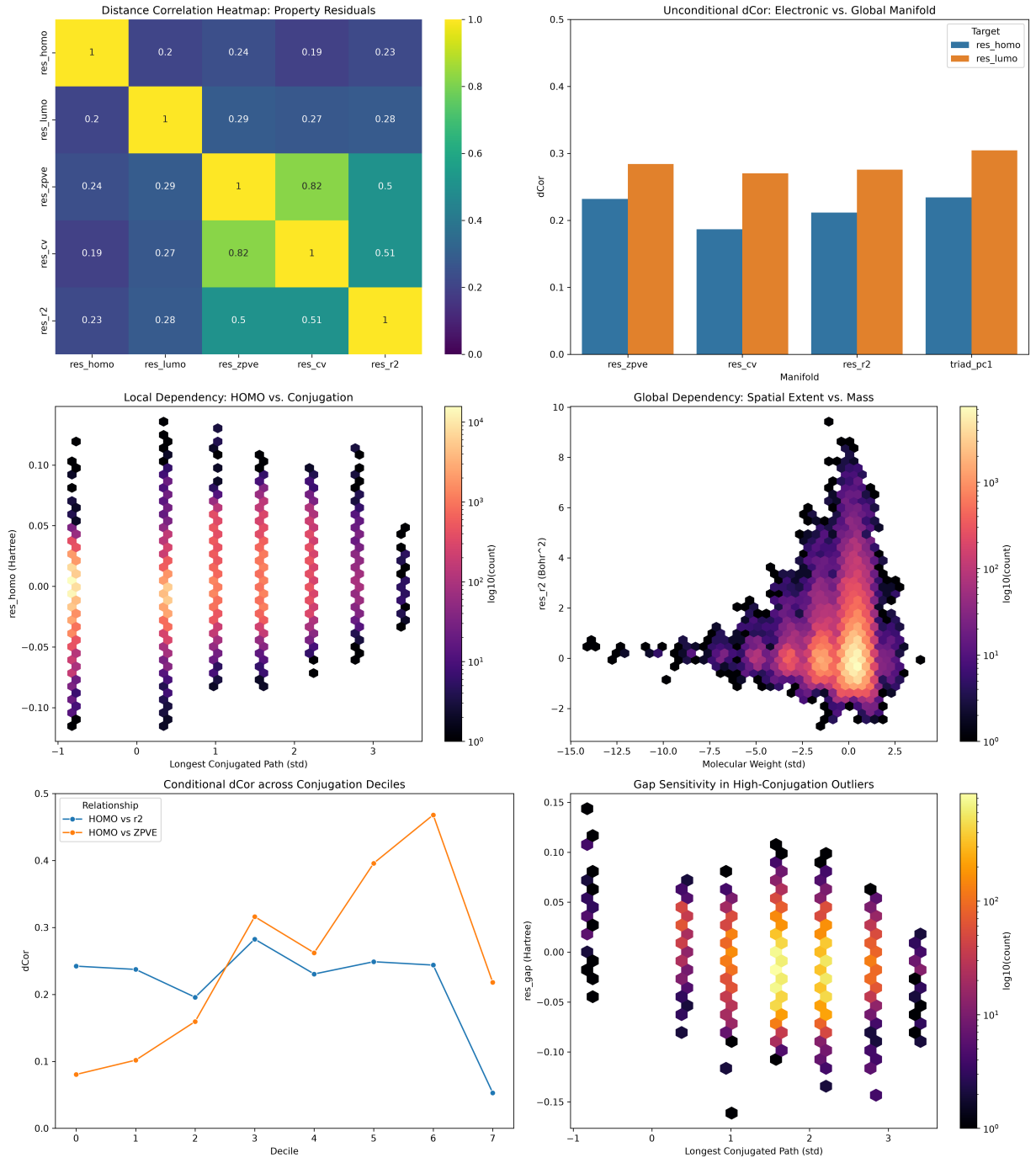

Figure 1: This figure illustrates the structural decoupling of the highest occupied molecular orbital (HOMO) from the global property manifold in the QM9 dataset. The top-left distance correlation heatmap demonstrates that the thermodynamic-spatial triad (res_{zpve}, res_{c_v}, res_{r^2}) forms a highly coherent macroscopic block, whereas res_{homo} remains persistently isolated. The top-right bar chart quantifies this divergence, showing that the lowest unoccupied molecular orbital (res_{lumo}) exhibits significantly higher unconditional integration into the global manifold's principal component ($triad_{pc1}$) than HOMO. Middle panels contrast these dependencies, revealing that while global spatial extent (res_{r^2}) scales with macroscopic features like molecular mass, frontier orbitals such as HOMO are governed by localized topological features like the longest conjugated path. The bottom panels display conditional dependencies across conjugation deciles; although the decoupling of HOMO is generally stable, extreme structural conditions in highly conjugated outliers cause the HOMO-LUMO gap (res_{gap}) to decouple entirely from the global thermodynamic manifold. Together, these visualizations demonstrate that the variance of the HOMO-LUMO gap is predominantly driven by LUMO's structural sensitivity, while HOMO acts as a localized electronic feature orthogonal to the broader macroscopic property network.

it drops to 0.2320. The distance correlation heatmap (Figure 1, top-left) visually confirms that while the thermodynamic-spatial triad forms a cohesive macroscopic block, *res_homo* remains an isolated feature.

This decoupling is fundamentally characterized by a mismatch in structural scale. Global spatial extent (res_r^2) is strongly governed by macroscopic geometric features, such as the maximum path length ($dCor = 0.7232$). Conversely, frontier orbitals respond to localized topological features (Figure 1, middle panels). The LUMO shows a markedly stronger sensitivity to the conjugation index ($dCor = 0.3547$) than the HOMO ($dCor = 0.2779$). However, the HOMO is more specifically tied to the longest conjugated path ($dCor = 0.2298$) compared to the LUMO ($dCor = 0.1880$), indicating that the HOMO's variance is governed by the length of the primary electronic pathway rather than broader conjugation density.

3.3 Scaffold-based stability and invariance

To ensure that the observed decoupling is a stable feature of organic chemistry rather than an artifact of the QM9 dataset's composition, we evaluated these relationships across 15,989 unique Murcko scaffolds using 5-fold cross-validation.

As detailed in Table 1.2, the structural isolation of the HOMO is a robust and invariant trait. The mean $dCor$ of the HOMO against the global manifold ($triad_pc1$) is 0.2252, consistently lower than the LUMO's integration of 0.2854. Furthermore, the narrow 95% confidence intervals for local descriptors, such as the relationship between *res_lumo* and the conjugation index, confirm that frontier orbital energies are reliably governed by local topology rather than the global spatial state.

3.4 Conditional reorganization in structural outliers

While the decoupling of the HOMO persists across the vast majority of structural scaffolds, extreme structural conditions reveal significant shifts in the electronic property network. Stratifying the dataset by conjugation index (Figure 1, bottom panels) demonstrates that for most molecules, the $dCor$ between the HOMO and the global manifold remains stable. However, in the top decile of conjugated molecules, the HOMO-LUMO gap (*res_gap*) exhibits near-total independence from the global thermodynamic manifold. In this highly conjugated regime, the correlation of *res_gap* against *res_zpve* and res_r^2 plummets to 0.0366 and 0.0639, respectively. Instead, the variance of the gap is increasingly driven by heteroatom density ($dCor = 0.2705$), a value nearly double its baseline in the general population. This suggests that extensive conjugation rewires the electronic gap to be sensitive to the chemical identity of the atoms within the conjugated path, rather than the overall spatial extent of the molecule.

Finally, sensitivity checks on restricted geometries reveal that the structural isolation of the HOMO partially breaks down in the physical limit of linear molecules. In these one-dimensional geometries, the correlation between *res_homo* and res_r^2 rises sharply to 0.4942 (compared to the global average of 0.2116), and the *res_lumo* versus *res_zpve* correlation reaches 0.7665. This indicates that the decoupling of the HOMO is an inherent property of three-dimensional, branched, or cyclic organic scaffolds, whereas in linear molecules, electronic and spatial extents become coupled by geometric necessity.

4 Conclusions

In this study, we addressed the challenge of mapping localized quantum states to global macroscopic properties by removing baseline stoichiometric effects from the QM9 dataset. Our partial distance correlation framework reveals a fundamental disparity in how frontier molecular orbitals integrate into the global spatial-thermodynamic manifold, which is defined by zero-point vibrational energy (*res_zpve*), heat capacity (res_c_v), and electronic spatial extent (res_r^2). While the LUMO demonstrates moderate integration into this macroscopic block and strong sensitivity to broader conjugation topology, the HOMO remains persistently decoupled, acting as an isolated electronic feature.

This decoupling is characterized by a distinct structural scale mismatch. The variance of the HOMO is strictly tied to localized electronic pathways, such as the longest conjugated path, whereas the LUMO is highly sensitive to the overall conjugation index. As demonstrated by our scaffold-based stability analysis (Table 1.2), the structural isolation of the HOMO from the global manifold's principal component ($triad_pc1$)

is not an artifact of dataset composition. Rather, it is a robust, invariant property across thousands of diverse organic scaffolds.

Despite this widespread stability, our analysis identifies specific structural regimes where the electronic property network undergoes significant reorganization. In the geometric limit of linear molecules, the structural isolation of the HOMO breaks down, and electronic and spatial extents become coupled by geometric necessity. Furthermore, in highly conjugated outliers, the HOMO-LUMO gap exhibits near-total independence from the global thermodynamic state. In these extended systems, the variance of the gap is rewired to depend predominantly on heteroatom density rather than spatial extent or bulk volume.

Ultimately, these findings mathematically formalize the structural conditions governing frontier orbital behavior. The variance of the HOMO-LUMO gap is primarily driven by the topological integration and structural sensitivity of the LUMO. In contrast, the HOMO fundamentally resists integration into the global property network, maintaining its role as a localized, structurally isolated quantum state across the vast majority of organic chemical space.