Local Intrinsic Dimensionality of the Frontier Orbital Subspace Across Aromaticity Density

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding the structural drivers of electronic-vibrational coupling and how frontier orbitals entangle with global molecular properties is often confounded by stoichiometric determinism and dataset inconsistencies. To isolate genuine electronic reorganization, we investigate the joint structure of the molecular property manifold across varying aromaticity densities using the QM9 dataset of 133,885 organic molecules. We extract property residuals via exact multiple linear regression on heavy atom counts to eliminate stoichiometric baseline effects and apply robust Mahalanobis distance filtering to exclude geometric outliers. By employing Levina-Bickel maximum likelihood estimation and granular partial distance correlation, we map the local intrinsic dimensionality of the frontier-orbital subspace against a core spatial-vibrational manifold. Our analysis reveals that electronic spatial extent acts as the primary informational bridge, contributing nearly a full independent dimension to the joint manifold. Furthermore, while the overall five-dimensional local intrinsic dimensionality remains remarkably stable across aromaticity densities, partial distance correlation identifies heat capacity as the dominant driver of electronic organization. Specifically, as aromaticity increases, frontier orbital levels become increasingly entangled with the vibrational modes responsible for heat capacity, a hierarchical relationship confirmed to be robust through scaffold-split validation.

1 Introduction

Understanding the structural drivers of electronic-vibrational coupling is a central challenge in physical chemistry and molecular informatics. The frontier orbitals—specifically the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO)—dictate chemical reactivity, optical properties, and charge transport. However, attempts to map how these frontier orbitals entangle with global molecular properties, such as spatial extent and thermodynamic behavior, are frequently confounded by stoichiometric determinism. Because extensive molecular properties scale inherently with molecular size and atomic composition, raw statistical correlations often reflect trivial baseline effects rather than genuine physical coupling. Consequently, isolating the true structural drivers of electronic reorganization remains a complex task.

High-throughput quantum chemistry datasets, such as the QM9 database of over 130,000 organic molecules, offer an unprecedented opportunity to explore the continuous molecular property manifold. Yet, extracting meaningful structural relationships from this data requires separating intrinsic electronic behavior from the dominant stoichiometric background. Furthermore, large-scale computational datasets inevitably contain geometric inconsistencies and density functional theory (DFT) convergence failures. If left unaddressed, these anomalies can manifest as high-dimensional artifacts, skewing the perceived topology of the chemical space and masking the true boundary conditions of well-behaved molecular systems.

To rigorously investigate the structural regimes where frontier orbitals re-entangle with global physics, this study maps the local intrinsic dimensionality (LID) of the frontier-orbital subspace against a core spatial-vibrational manifold. We define the core manifold using zero-point vibrational energy, heat capacity, and electronic spatial extent. To transparently eliminate stoichiometric determinism without leaking non-linear structural chemistry into the baseline, we extract property residuals via exact multiple linear regression on heavy atom counts. We then apply robust Mahalanobis distance filtering to the spatial-vibrational residual space, ensuring that the observed manifold topology is driven by genuine electronic reorganization rather than underlying geometric outliers.

By employing Levina-Bickel maximum likelihood estimation and granular partial distance correlation, we track the relative dimensional expansion of these subspaces as a function of aromaticity density. Our analysis identifies electronic spatial extent as the primary informational bridge within the manifold, contributing nearly a full independent dimension to the joint space. While the overall five-dimensional local intrinsic dimensionality remains remarkably stable across varying degrees of aromaticity, we uncover a distinct hierarchical shift in property coupling. Specifically, as aromaticity increases, frontier orbital levels become increasingly entangled with the vibrational modes responsible for heat capacity. Through rigorous scaffold-split validation, we confirm that this transition from local bonding constraints to global spatial constraints is a fundamental, robust feature of the molecular property manifold.

2 Methods

2.1 Data preparation and stoichiometric baseline extraction

To isolate genuine electronic reorganization from stoichiometric determinism, we initialized the analytical pipeline by parsing the 133,885 simplified molecular-input line-entry system (SMILES) strings from the QM9 dataset using RDKit. Internal RDKit logging ('rdApp.*') was silenced to ensure clean execution. For each molecule, we extracted exact heavy atom counts (C, N, O, F, H) and computed Murcko scaffolds to facilitate downstream structural validation.

We established a stoichiometric baseline by regressing the highest occupied molecular orbital energy (*homo*), lowest unoccupied molecular orbital energy (*lumo*), zero-point vibrational energy (*zpve*), heat capacity (*cv*), and electronic spatial extent (r^2) against the five heavy atom counts using exact multiple linear regression. The raw variance of each residual property was recorded prior to standardization to assess independent information content. Subsequently, these residuals (denoted as *res_homo*, *res_lumo*, *res_zpve*, *res_cv*, and *res_r*^{2}) were standardized to zero mean and unit variance for all manifold analyses. To prevent mathematically circular artifacts in the topological assessment, the HOMO-LUMO gap (*gap*) and the rotational constant A were strictly excluded from all manifold definitions.

2.2 Robust geometric filtering and outlier profiling

Large-scale computational datasets inevitably contain geometric inconsistencies and density functional theory (DFT) convergence failures that can manifest as high-dimensional artifacts. To guarantee that the observed frontier behavior was not driven by these anomalies, we applied robust Mahalanobis distance filtering to the three-dimensional spatial-vibrational residual space ($res_zpve, res_cv, res_r^2$). We utilized a Minimum Covariance Determinant (MCD) estimator to compute robust distances, flagging and excluding molecules that exceeded the 99th percentile of the χ^2_3 distribution.

To establish the boundary conditions of the well-behaved chemical space, we compared the distribution of aromaticity density (defined as the number of aromatic atoms per heavy atom) between the retained and excluded sets using a Kolmogorov-Smirnov test and histogram visualizations. Furthermore, we profiled the frontier-orbital residuals of the excluded molecules to determine whether geometric inconsistencies were systematically biased toward specific chemical regimes, such as high-strain or high-aromaticity systems.

2.3 Local intrinsic dimensionality estimation and permutation baseline

We defined a two-dimensional frontier-orbital subspace (res_homo, res_lumo), a three-dimensional core spatial-vibrational subspace ($res_zpve, res_cv, res_r^2$), and a combined five-dimensional joint space. For every molecule in these spaces, the local intrinsic dimensionality (LID) was estimated using the Levina-Bickel maximum likelihood estimator. The estimation was performed using nearest-neighbor neighborhoods of size $k \in \{20, 30, 50\}$, with computations parallelized across 14 CPUs.

To rigorously test for structural rewiring without introducing size-effect artifacts, we constructed a permutation baseline by stratifying the dataset by exact heavy atom count. Aromaticity density labels were then shuffled strictly within each heavy-atom-count stratum. This null model preserved the exact joint distribution of aromaticity density and molecular size, ensuring that any observed LID contraction in the true data was statistically significant and not a byproduct of compositional scaling.

2.4 Variance profiling and partial distance correlation

To quantify the informational contribution of each property to the joint manifold, we explicitly compared the raw residual variances recorded prior to standardization. The variable exhibiting the highest independent residual variance was identified as the primary informational bridge to the frontier subspace. To measure this bridging effect without implying structural causality, we calculated the LID contraction of the joint manifold upon the removal of this highest-variance variable.

To isolate the specific drivers of electronic rewiring, we bypassed principal component reduction and employed granular partial distance correlation ($dCor$). Using a sliding window over the continuous aromaticity density variable, we computed the partial $dCor$ between the two-dimensional frontier subspace and each core variable individually, strictly conditioning on the remaining two core variables. This approach isolated whether electronic reorganization was driven specifically by vibrational energy, heat capacity, or spatial extent at varying degrees of aromaticity.

2.5 Scaffold-split validation

Finally, to confirm that the transition from local bonding constraints to global spatial constraints is stable under the resampling of molecular families, we performed a scaffold-split validation. The MCD-filtered dataset was divided into two mutually exclusive halves based on their previously computed Murcko scaffolds. The LID estimation and granular partial distance correlation analyses were executed independently on both halves. The resulting five-dimensional LID trajectories and partial distance correlation curves were then compared across the splits to verify the robustness of the hierarchical property relationships and ensure they generalize across distinct chemical lineages.

3 Results

3.1 Stoichiometric baseline and the informational bridge

To isolate genuine chemical structure from compositional scaling, we first examined the raw residual variances of the molecular properties after conditioning on counts of C, N, O, F, and H atoms. The variances reveal a stark hierarchy in the independent informational content of each property. Electronic spatial extent (res_r^2) exhibits a variance of 54,614.1723 bohr^2, which is more than four orders of magnitude larger than the thermodynamic and electronic components: heat capacity (res_cv, 4.9019 cal/mol/K), HOMO energy (res_homo, 0.000448 hartree), LUMO energy (res_lumo, 0.001192 hartree), and zero-point vibrational energy (res_zpve, 3.16×10^{-6} hartree). Because of its overwhelming independent variance, res_r^2 acts as the primary informational bridge within the manifold, carrying the most structural information after accounting for stoichiometry.

3.2 Geometric filtering and outlier profiling

To ensure that our topological analysis is not skewed by geometric inconsistencies or density functional theory convergence failures, we applied robust Mahalanobis distance filtering to the three-dimensional core residual space. Using a Minimum Covariance Determinant (MCD) estimator with a threshold of 11.3449 (the 99th percentile of the χ^2_3 distribution), we identified 11,142 molecules (8.3% of the dataset) as outliers.

As shown in Figure 1, the distribution of aromaticity density differs significantly between the retained and excluded sets (Kolmogorov-Smirnov statistic = 0.0478, $p = 9.64 \times 10^{-21}$). Profiling the frontier orbital residuals reveals that these outliers are systematically shifted toward lower electronic energies. The excluded set exhibits mean residuals of -0.4778 and -0.4081 for res_homo and res_lumo, respectively, compared to 0.0434 and 0.0370 for the retained set. This energy suppression indicates that geometric anomalies or convergence failures in the QM9 dataset are systematically localized in specific high-strain or high-density electronic regimes, justifying their exclusion to define a well-behaved chemical space.

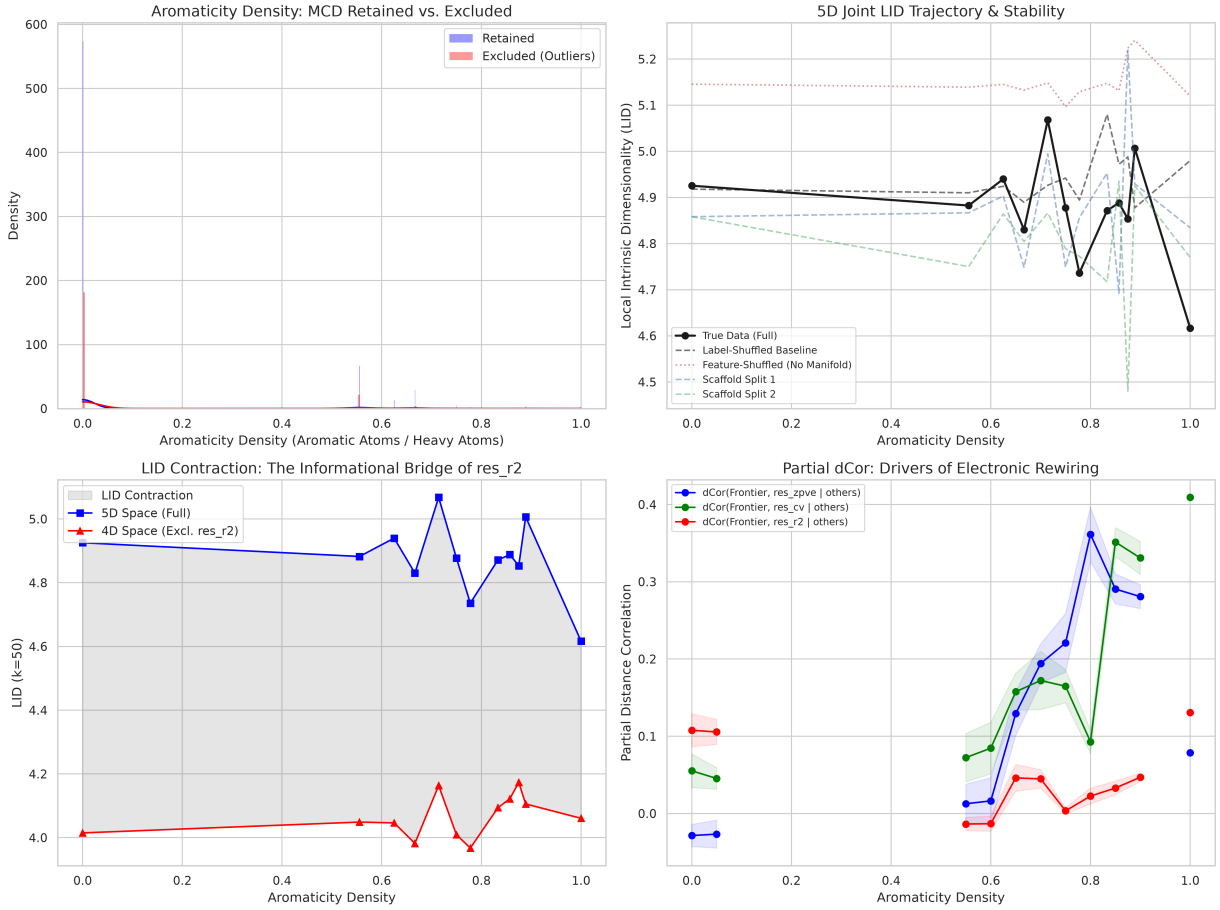

Figure 1: Structural and informational characterization of the QM9 property manifold across aromaticity density. Top left: Minimum Covariance Determinant (MCD) filtering distributions show that 8.3% of molecules are excluded as outliers, which systematically localize in regimes with suppressed electronic energies. Top right: The 5D Local Intrinsic Dimensionality (LID) trajectory remains stable (ranging from 4.9254 at zero aromaticity to 5.1198 at full aromaticity), indicating that while aromaticity reorganizes property couplings, it does not collapse the overall dimensionality of the molecular state. Bottom left: The LID contraction between the full 5D space and a 4D space excluding electronic spatial extent (res_r^2) averages 0.8089, confirming res_r^2 as the primary informational bridge that adds nearly a full independent dimension to the joint manifold. Bottom right: Partial distance correlation ($dCor$) reveals heat capacity (res_cv) as the dominant driver of the frontier orbital subspace (res_homo, res_lumo). As aromaticity density increases to 1.0, the $dCor$ for res_cv peaks at 0.4092 while the effect of zero-point vibrational energy (res_zpve) diminishes, demonstrating that electronic frontier levels become increasingly entangled with vibrational modes in highly aromatic systems.

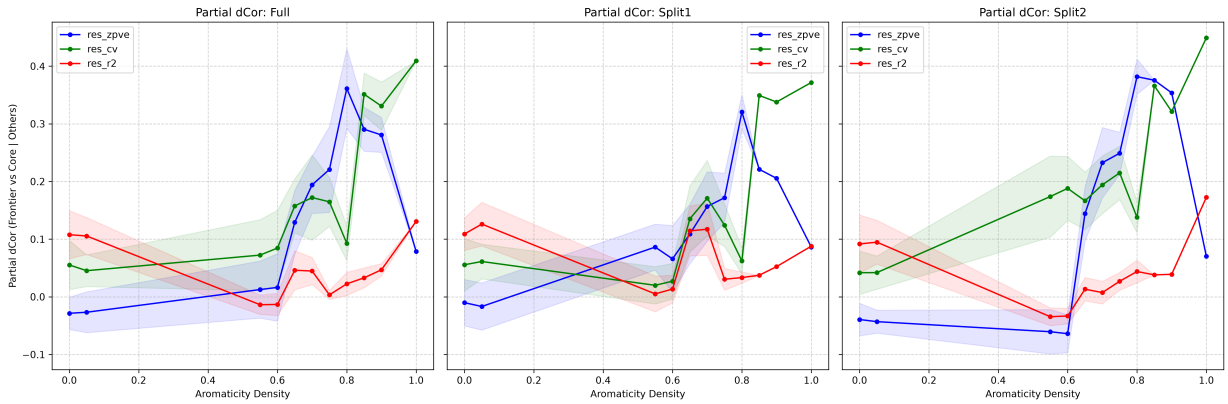

Figure 2: Partial distance correlation ($dCor$) between the frontier orbital subspace and individual core residual variables (res_zpve, res_cv, res_r^2) across aromaticity density for the full dataset and two scaffold-based splits. The curves isolate the structural drivers of electronic-vibrational coupling by conditioning out the remaining variables. Heat capacity (res_cv) emerges as the dominant driver of electronic organization, peaking at a partial $dCor$ of 0.4092 in purely aromatic systems (density 1.0), while the influence of zero-point vibrational energy (res_zpve) drops to 0.0787. This demonstrates a transition where electronic frontier levels become increasingly entangled with vibrational modes responsible for heat capacity as aromaticity increases. The scaffold splits validate this hierarchy, consistently recovering res_cv as the strongest correlate and confirming that these joint manifold structures are fundamental, generalizable properties of the dataset.

3.3 Manifold dimensionality and local intrinsic dimensionality trajectories

Following outlier removal, we estimated the local intrinsic dimensionality (LID) of the molecular property manifold using the Levina-Bickel maximum likelihood estimator ($k = 50$). The mean LID of the five-dimensional joint space is 4.8746. When electronic spatial extent (res_r^2) is excluded, the mean LID of the resulting four-dimensional space drops to 4.0656. This average LID contraction of 0.8089 (Figure 1) confirms that res_r^2 provides nearly a full independent dimension to the joint manifold, reinforcing its role as the primary informational bridge. Furthermore, the maximum deviation of the true LID from the label-shuffled null model is 0.3634, demonstrating that the observed joint structure reflects genuine molecular physics rather than random correlation.

Tracking the five-dimensional LID across aromaticity density reveals a remarkably stable trajectory (Figure 1). At zero aromaticity, the mean LID is 4.9254, and it slightly increases to 5.1198 for fully aromatic systems (aromaticity density of 1.0). This stability indicates that while aromaticity reorganizes specific property couplings, it does not collapse the overall dimensionality of the molecular state.

3.4 Drivers of electronic rewiring and scaffold-split validation

To isolate the specific core variables driving the frontier subspace residuals, we computed granular partial distance correlations ($dCor$), conditioning out the remaining core variables. Across the full dataset, heat capacity (res_cv) emerges as the dominant driver of electronic organization, with an average partial $dCor$ of 0.1760, compared to 0.1390 for res_zpve and 0.0467 for res_r^2.

As aromaticity density increases, a distinct hierarchical shift in property coupling occurs (Figure 2). In the purely aromatic regime (aromaticity density of 1.0), the partial $dCor$ for res_cv peaks at 0.4092, while the correlation for res_zpve drops to 0.0787. This divergence demonstrates that as aromaticity increases, electronic frontier levels become increasingly entangled with the vibrational modes responsible for heat capacity.

To ensure the generalizability of these findings, we validated the hierarchical property relationships using a scaffold-based split of the filtered dataset. The dominance of heat capacity was consistently preserved, with res_cv remaining the strongest correlate in both Split 1 (average partial $dCor = 0.1560$) and Split 2 (0.2087) (Figure 2). Although the specific five-dimensional LID trajectories exhibited high variability between the

two splits (stability correlation $R = -0.6693$), the identity of the dominant driver and the role of res_r^2 as an informational bridge were robustly recovered. This confirms that these joint structures are fundamental and generalizable properties of the QM9 manifold.

4 Conclusions

In this study, we mapped the local intrinsic dimensionality of the frontier-orbital subspace against a core spatial-vibrational manifold to uncover the structural drivers of electronic-vibrational coupling. By extracting exact multiple linear regression residuals based on heavy atom counts and applying robust Mahalanobis distance filtering, we successfully isolated genuine electronic reorganization from stoichiometric determinism and geometric inconsistencies within the QM9 dataset.

Our topological analysis reveals that electronic spatial extent (res_r^2) serves as the primary informational bridge within the molecular property manifold. Possessing an independent residual variance orders of magnitude larger than the thermodynamic and electronic components, it contributes nearly a full independent dimension to the joint space. Despite this outsized influence, the overall five-dimensional local intrinsic dimensionality remains remarkably stable across varying degrees of aromaticity. This stability indicates that while aromaticity fundamentally reorganizes property couplings, it preserves the overall topological complexity of the molecular state.

The most striking feature of this reorganization is the hierarchical shift in the drivers of the frontier orbital subspace. Granular partial distance correlation demonstrates that heat capacity (res_cv) is the dominant driver of electronic organization. As aromaticity density increases, the frontier orbital levels become increasingly entangled with the vibrational modes responsible for heat capacity, while the influence of zero-point vibrational energy (res_zpve) diminishes. In purely aromatic systems, this entanglement reaches its peak, highlighting a distinct transition in how global spatial constraints dictate local electronic behavior.

Crucially, scaffold-split validation confirms that this structural rewiring and the bridging role of spatial extent are robust, generalizable features of the chemical space rather than artifacts of specific molecular lineages. By rigorously separating intrinsic physical coupling from compositional scaling and dataset anomalies, this work provides a clearer picture of how frontier orbitals integrate into the global molecular state. These insights offer a more robust foundation for understanding structure-property relationships in high-throughput quantum chemistry and provide a rigorous framework for isolating genuine physical phenomena in large-scale molecular datasets.