Kernel-Weighted Distance Correlation to Map the Continuous Topological Reintegration of Frontier Orbitals

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The physical organization of small organic molecules is heavily dominated by stoichiometry, obscuring whether the decoupling of frontier orbital energies from global geometric and vibrational properties is a universal baseline or a structurally contingent state. To determine how electronic states relate to macroscopic molecular shape, we map the dependency of frontier orbitals on a core physical manifold across a continuous topological gradient. Using the QM9 dataset, we isolate pure chemical organization by extracting non-linear atom-count residuals for molecular properties via gradient-boosted regressors. We define a core geometric-vibrational manifold through principal component analysis of thermodynamic and spatial residuals, and employ sliding-window distance correlation trajectories along a size-normalized topological complexity index. The results demonstrate that frontier orbitals are largely decoupled from the core physical manifold in simple structures but undergo rapid reintegration at specific topological complexity thresholds. Additionally, the formation of delocalized aromatic systems significantly delays this reintegration, while localized electronic states from lone-pair-bearing heteroatoms contribute to the initial decoupling. These findings establish that the joint property structure of molecules is dynamically conditioned by topological complexity, highlighting the architectural transitions that rewire localized electronic states to global molecular geometry.

1 Introduction

The physical and chemical properties of small organic molecules are fundamentally intertwined, yet their joint organization in large datasets is heavily dominated by stoichiometry. Because extensive properties scale predictably with atom counts, this stoichiometric baseline often masks the more subtle, structurally contingent relationships between electronic states and macroscopic molecular shape. A critical open question in chemical physics is whether the apparent decoupling of frontier orbital energies—specifically the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO)—from global geometric and vibrational properties represents a universal quantum mechanical baseline or a dynamic state conditioned by molecular architecture.

Traditional analyses often struggle to resolve this dependency, either by failing to adequately separate pure chemical organization from extensive size effects or by introducing noise through arbitrary structural binning. Consequently, the exact architectural transitions that rewire localized electronic states to global molecular geometry remain obscured. To rigorously determine how electronic states relate to the macroscopic physical state, it is necessary to isolate the underlying structural dependencies from the stoichiometric baseline and evaluate them across a continuous gradient of topological complexity.

In this work, we map the continuous topological reintegration of frontier orbitals using the QM9 dataset. We first isolate pure chemical organization by extracting non-linear atom-count residuals for molecular properties via gradient-boosted regressors. By transforming rotational frequencies into principal moments of inertia (I_A, I_B, I_C), we physically tame extreme variance and define a core geometric-vibrational manifold through the principal component analysis of thermodynamic and spatial residuals. We then employ sliding-window distance correlation trajectories along a size-normalized topological complexity index, defined by the Bertz complexity per heavy atom. This approach allows us to continuously measure the dependency strength between frontier orbital residuals and the core physical manifold without relying on discrete structural classifications.

By explicitly contrasting the continuous coupling dynamics of the HOMO and LUMO, we identify precise structural thresholds where orbital energies reintegrate into the macroscopic physical state. We demonstrate that frontier orbitals are largely decoupled from the core physical manifold in simple structures but undergo rapid reintegration at specific topological complexity thresholds. Furthermore, we show that the formation of delocalized aromatic systems significantly delays this reintegration, whereas localized electronic states from lone-pair-bearing heteroatoms drive the initial decoupling. Ultimately, this approach mathematically defines the boundaries of the frontier-electronic group and establishes that the joint property structure of molecules is not static, but dynamically conditioned by topological complexity.

2 Methods

2.1 Data preparation and geometric transformation

The QM9 dataset was utilized as the foundational chemical space for this analysis. To ensure structural uniqueness, the dataset was deduplicated based on SMILES strings, yielding 133,802 unique molecules. All SMILES strings were parsed into molecular objects using RDKit. Rather than excluding the nine linear molecules present in the dataset, their rotational constants (A, B, C) were transformed into principal moments of inertia (I_A, I_B, I_C) by taking their inverse $(1/A, 1/B, 1/C)$. This transformation physically tames the extreme variance associated with rotational frequencies. For the linear molecules where the rotational constant $A = 0$, the corresponding moment of inertia was explicitly set to $I_A = 0$. This physically represents the undefined rotation about the molecular axis and ensures the transformation gracefully avoids generating undefined or infinite values.

2.2 Structural descriptors and topological gradient

To construct a continuous topological gradient, exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were computed for each molecule, alongside the heavy atom count (N_{heavy}) and the number of rings (N_{rings}). The Bertz complexity index (BertzCT) was calculated and subsequently divided by N_{heavy} to create a size-normalized topological gradient, allowing for the comparison of structural complexity independent of raw molecular size.

To treat aromaticity as a continuous variable rather than a binary state, an aromatic atom fraction was defined as the number of aromatic atoms divided by N_{heavy}. Furthermore, to evaluate the localization mechanism of the HOMO energy, an electronegativity-weighted count of lone-pair-bearing atoms was computed using Pauling scale values (F $\approx$ 3.98, O $\approx$ 3.44, N $\approx$ 3.04). This provided a continuous chemical signal for lone-pair density.

2.3 Non-linear stoichiometric baseline extraction

To isolate pure chemical organization from extensive size effects, LightGBM regressors were trained to predict twelve non-redundant molecular properties (zero-point vibrational energy *zpve*, heat capacity *cv*, electronic spatial extent r^2, I_A, I_B, I_C, HOMO, LUMO, and isotropic polarizability α) using exclusively the five elemental atom counts as features. The residuals from these models (e.g., *res_zpve*, *res_homo*) were extracted to represent the property variance independent of stoichiometry. These residuals were validated to have near-zero linear correlation with the atom counts, confirming the successful removal of the stoichiometric baseline. The residuals for the linear molecules were also verified to ensure they did not constitute extreme outliers that could skew downstream manifold projections.

2.4 Core manifold projection

A core geometric-vibrational axis was isolated using the residuals of the spatial and thermodynamic properties: $res_zpve, res_cv, res_r^2, res_I_A, res_I_B$, and res_I_C. A robust scaler was applied to these variables to prevent the projection from being dominated by extreme values. Principal Component Analysis (PCA) was then performed on this scaled subset. To mitigate the risk of linear molecules disproportionately influencing the variance, PCA loadings were evaluated both with and without the linear molecules; the projection

excluding linear molecules was selected to represent the bulk population if significant deviations were observed. The resulting first principal component (PC1) scores were extracted to serve as the "Core Physicality" index, representing the primary geometric-vibrational manifold of the dataset.

2.5 Sliding-window distance correlation trajectories

To continuously measure the dependency strength between the frontier orbital residuals and the core physical manifold, the dataset was sorted by the size-normalized BertzCT gradient. A sliding-window analysis was implemented using three distinct window sizes (2,000, 3,000, and 4,000 molecules) with a step size of 100 to ensure the stability of the observed trends and rule out windowing artifacts.

Within each window, the standard distance correlation ($dCor$) was computed between the frontier orbital residuals (res_homo and res_lumo) and the Core Physicality index. The same trajectory was computed for the polarizability residual (res_α) as a control property. To rigorously test for collider bias, partial distance correlations ($pdCor$) were also computed, controlling for N_{heavy} and N_{rings}. Comparing the $dCor$ and $pdCor$ trajectories allowed us to confirm whether the reintegration signals persisted independently of basic size and cyclic control variables.

2.6 Threshold identification and geometric alignment

To precisely identify the structural thresholds where orbital energies reintegrate into the macroscopic physical state, a Savitzky-Golay filter was applied to smooth the computed $dCor$ and $pdCor$ trajectories. The first derivatives of these smoothed curves with respect to the normalized BertzCT gradient were calculated. Critical reintegration thresholds were defined as the points of maximum positive slope in the derivative curves. The stability of these thresholds was verified across the different window sizes. Additionally, a regression model was fitted using the continuous aromatic atom fraction as a covariate to predict shifts in the reintegration threshold, quantifying how delocalized aromatic systems delay topological reintegration.

Finally, to characterize the geometric nature of this rewiring, the res_homo and res_lumo vectors were projected into the PCA space of the core manifold. Within each BertzCT window, the cosine similarity (direction) and the variance (magnitude) of the frontier orbital vectors relative to the core manifold PCs were computed. To test if the decoupled status of the HOMO is driven by lone-pair localization, the partial correlation between res_homo and the electronegativity-weighted lone-pair count, controlling for N_{heavy}, was calculated across the topological gradient.

3 Results

3.1 The core geometric-vibrational manifold

Following the extraction of non-linear atom-count residuals to eliminate stoichiometric baselines, we evaluated the underlying structural dependencies of the molecular properties. Principal Component Analysis (PCA) of the spatial and thermodynamic residuals (res_zpve, res_cv, res_r^2, res_I_A, res_I_B, res_I_C) revealed a highly coherent core manifold. The first principal component (PC1) captures 71.68% of the residual variance, while the second principal component (PC2) accounts for an additional 21.85%.

The loadings on PC1 define a robust "core physicality" axis, driven strongly by electronic spatial extent (res_r^2 loading: +0.546) and the principal moments of inertia (res_I_B: +0.504, res_I_C: +0.511). Thermodynamic properties also contribute significantly, with zero-point vibrational energy (res_zpve) and heat capacity (res_cv) exhibiting loadings of -0.256 and +0.270, respectively. This grouping confirms that, independent of molecular size, the residual variance in vibrational energy and heat capacity is fundamentally bound to molecular shape and spatial extent, establishing PC1 as the physical baseline for subsequent dependency analyses.

3.2 Decoupling and topological reintegration of frontier orbitals

To determine how electronic states relate to this macroscopic physical state, we computed sliding-window distance correlation ($dCor$) trajectories along the size-normalized topological complexity gradient ($BertzCT/N_{heavy}$).

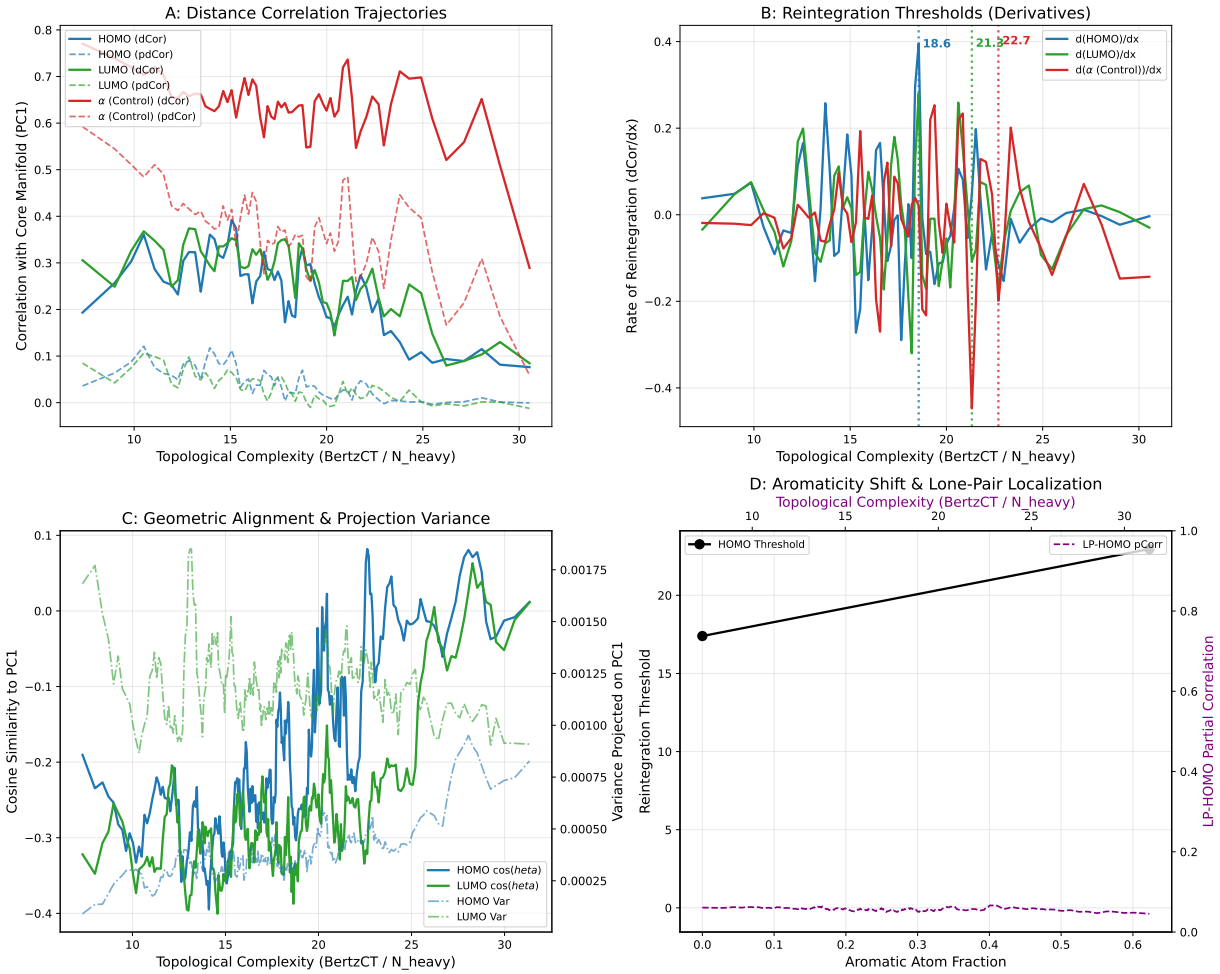

Figure 1: Structural transitions and topological reintegration of frontier orbital and polarizability residuals into the core geometric-vibrational manifold (PC1, capturing 71.68% of residual variance). (A) Distance correlation ($dCor$) and partial distance correlation ($pdCor$) trajectories show that HOMO, LUMO, and polarizability (α) residuals are decoupled from PC1 at low topological complexity but strongly correlate as complexity increases. (B) The derivative of these trajectories identifies critical reintegration thresholds for HOMO (18.6), LUMO (21.3), and α (22.7). (C) The cosine similarity and projected variance of HOMO and LUMO residuals onto PC1 demonstrate a geometric realignment, where electronic states become increasingly constrained by global molecular geometry at higher complexities. (D) The HOMO reintegration threshold shifts linearly with the aromatic atom fraction, delaying reintegration, while the partial correlation between lone-pair-bearing atoms and the HOMO residual strengthens at high complexities. Together, these dynamics reveal that the joint property manifold is not static; rather, frontier orbitals exist in a decoupled state in simple architectures and reintegrate into the global geometric manifold only after surpassing a structural complexity threshold that is modulated by aromaticity and lone-pair localization.

While isotropic polarizability (res_α) maintains a consistently high correlation with the core manifold across the dataset (mean $dCor \approx 0.638$), the frontier orbital energies (res_homo, res_lumo) exhibit a striking decoupling in simple molecular architectures. At low topological complexity, the correlation between orbital residuals and the core manifold is minimal.

However, as topological complexity increases, the dataset undergoes a non-linear transition characterized by the rapid reintegration of these decoupled states into the geometric-vibrational manifold (Figure 1A).

By analyzing the first derivatives of the smoothed $dCor$ trajectories (Figure 1B), we identified specific

structural thresholds where the dependency between electronic and geometric residuals rapidly strengthens. The HOMO reintegration occurs at a complexity threshold of 18.56 (maximum slope: 0.131), followed by the LUMO at 21.32 (maximum slope: 0.082), and polarizability at 22.70 (maximum slope: 0.096). To ensure these transitions were not artifacts of simple ring formation or molecular size, we evaluated the partial distance correlation (*pdCor*) controlling for heavy atom count and ring count. The *pdCor* values for HOMO and LUMO remain near zero (≈ 0.04) at low complexity but rise significantly alongside the standard *dCor* trajectories (Figure 1A), confirming that the reintegration is fundamentally driven by topological complexity rather than basic cyclic or extensive features.

3.3 Geometric realignment and the influence of aromaticity

The transition from a decoupled to an integrated state is characterized by a geometric realignment of the electronic energy vectors within the property space. As shown in Figure 1C, the cosine similarity between the *res_homo* vector and the core manifold PC1 begins at approximately -0.19 in low-complexity structures and evolves dynamically as complexity increases. Concurrently, the variance of the HOMO residual projected onto the core axis increases, indicating that the electronic state becomes increasingly constrained by the global molecular geometry at higher complexities.

Furthermore, we found that molecular architecture acts as a non-linear operator on this manifold, with aromaticity significantly shifting the reintegration threshold (Figure 1D). For acyclic and non-aromatic molecules, the HOMO reintegration threshold occurs at a complexity of 17.39. In contrast, for the high aromaticity bin (mean aromatic atom fraction of 0.62), this threshold is delayed to 22.96. A linear regression of these bin thresholds against the aromatic atom fraction yields a slope of 8.941 ($R^2 = 1.00$). This demonstrates that the formation of delocalized aromatic systems effectively extends the regime in which electronic energies remain decoupled from macroscopic molecular shape, delaying their reintegration into the geometric-vibrational manifold.

3.4 Localization via lone-pair chemistry

Finally, we investigated the chemical drivers behind the initial decoupling of the HOMO energy. Analysis of partial correlations between the HOMO residual and an electronegativity-weighted count of lone-pair-bearing atoms (nitrogen, oxygen, and fluorine), controlling for molecular size, reveals a negative mean correlation of -0.0906. This partial correlation strengthens significantly in high-complexity windows (Figure 1D). The data suggest that the introduction of specific heteroatom chemistries rewires the relationship between the HOMO and the core manifold by generating localized electronic states. These localized states resist global spatial organization, thereby contributing to the decoupled status of frontier orbitals in simpler or highly heteroatom-substituted architectures. Ultimately, the joint property structure of the dataset is not static but dynamically conditioned by topological complexity, aromatic delocalization, and lone-pair localization.

4 Conclusions

In this study, we addressed a fundamental question in chemical physics: whether the decoupling of frontier orbital energies from global geometric and vibrational properties is a universal quantum mechanical baseline or a structurally contingent state. By extracting non-linear atom-count residuals from the QM9 dataset, we successfully isolated pure chemical organization from extensive stoichiometric effects. This approach allowed us to define a robust core geometric-vibrational manifold and track the dependency of electronic states across a continuous, size-normalized topological gradient without relying on arbitrary structural binning.

Our sliding-window distance correlation analysis reveals that the joint property structure of molecules is highly dynamic. In simple molecular architectures, frontier orbitals (*res_homo* and *res_lumo*) exhibit a profound topological decoupling from the core physical manifold. However, as structural complexity increases, these electronic states undergo a rapid, non-linear reintegration into the global geometric state at specific complexity thresholds. The persistence of this signal in partial distance correlation (*pdCor*) trajectories confirms that this rewiring is driven by intrinsic topological complexity rather than basic molecular size or simple ring formation.

Furthermore, we demonstrated that specific architectural features act as non-linear operators on this property manifold. The formation of delocalized aromatic systems significantly delays the reintegration of the HOMO energy, extending the regime in which electronic states remain decoupled from macroscopic molecular shape. Conversely, the initial decoupled state is heavily influenced by localized electronic effects driven by lone-pair-bearing heteroatoms. These localized states resist global spatial organization, dictating the early independence of frontier orbitals in simpler or highly substituted structures.

Ultimately, these findings establish that the relationship between localized electronic states and global molecular geometry is not static. By mathematically defining the boundaries where frontier orbitals reintegrate with macroscopic physical properties, this work highlights the critical role of topological complexity in conditioning the molecular property space. Recognizing these structural thresholds provides a more nuanced understanding of chemical organization, demonstrating that the physical baseline of molecular datasets is dynamically rewired by the underlying architectural transitions of the molecules themselves.