Information-Geometric Decoupling of Frontier Orbitals Across Topological Bins

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The fundamental relationship between frontier orbitals (HOMO and LUMO) and molecular topology is often obscured by dominant stoichiometric and basic structural effects. To uncover the true topological influence on electronic freedom, we isolate the non-linear electronic signals of 133,802 molecules from the QM9 dataset. We achieve this by extracting HOMO and LUMO residuals using gradient-boosted trees conditioned on an augmented structural baseline. The molecules are then partitioned into acyclic, compact polycyclic, and extended polycyclic aromatic bins. Using covariance matrix eigenvalue analysis and two-dimensional Wasserstein optimal transport distances, we quantify the geometric distribution shifts of these residuals across topological boundaries. The results indicate that while acyclic molecules possess broad, independent frontier orbital distributions, polycyclic aromatic systems exhibit a significant contraction in total electronic variance. Furthermore, extended polycyclic aromatics demonstrate a robust structural locking effect, where the coupling index between orbital residuals more than doubles compared to acyclic baselines. Wasserstein distance decompositions reveal that the transition into aromatic topologies is predominantly driven by shape shifts in the joint distribution rather than simple mean displacements. Ultimately, these findings establish that while acyclic frontier orbitals are governed by independent local bonding environments, global delocalization in extended conjugated systems restricts these orbitals to a highly coupled, primary electronic manifold.

1 Introduction

The frontier orbitals of a molecule—specifically the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO)—are fundamental to understanding its chemical reactivity, stability, and electronic properties. However, the intrinsic relationship between these frontier orbitals and the underlying molecular topology is frequently obscured by dominant stoichiometric and basic structural effects. In large chemical spaces, simple atom counts and basic structural indices account for the vast majority of variance in macroscopic thermodynamic properties. Consequently, the subtle, non-linear electronic signals that arise from complex topological features, such as extended conjugation and polycyclic aromaticity, are difficult to isolate and quantify.

1.1 The challenge of isolating electronic signals

To uncover the true topological influence on electronic freedom, it is necessary to decouple the frontier orbitals from their baseline thermodynamic and spatial context. Traditional approaches often analyze absolute orbital energies, which inherently preserves the rigid physical identity of the energy gap, defined as $gap = lumo - homo$, but fails to separate local bonding effects from macroscopic structural shifts. Because local properties should theoretically remain mathematically invariant to macroscopic structural changes, any significant deformation in the joint distribution of HOMO and LUMO energies across different topological classes provides direct evidence of global structural influence. Standard extraction of non-linear structural residuals is therefore required to establish a baseline thermodynamic-spatial context before jumping directly to macroscopic manifold comparisons.

1.2 Information-geometric approach

In this work, we present an information-geometric framework to quantify the decoupling and re-organization of frontier orbitals across distinct topological boundaries. Utilizing a deduplicated set of 133,802 molecules from the QM9 dataset, we isolate the non-linear electronic signals by extracting HOMO and LUMO residuals. These residuals are generated using gradient-boosted trees conditioned on an augmented structural baseline, effectively stripping away the variance explained by basic stoichiometry and hybridization. The molecular space is then partitioned into discrete topological bins: an acyclic baseline, compact polycyclic aromatic systems, and extended polycyclic aromatic systems.

By analyzing the joint distribution of these isolated residuals, we employ covariance matrix eigenvalue analysis to measure the structural invariance and coupling of the electronic states. Furthermore, we utilize two-dimensional Wasserstein optimal transport distances (W_2) to evaluate the geometric distribution shifts between topological classes. Our analysis reveals an asymmetric Wasserstein cost profile and a pronounced "locking" effect in extended conjugated systems. While acyclic molecules possess broad, independent frontier orbital distributions governed by local bonding environments, the transition into extended polycyclic topologies induces a significant contraction in total electronic variance. Ultimately, we demonstrate that global delocalization restricts these frontier orbitals to a highly coupled, primary electronic manifold, driven predominantly by shape shifts in the joint distribution rather than simple mean displacements.

2 Methods

2.1 Data ingestion and baseline initialization

To establish the baseline thermodynamic and spatial context discussed previously, we utilized the QM9 dataset. The dataset was loaded and deduplicated by SMILES strings, yielding 133,802 unique molecular entries. During ingestion, the RDKit logger was globally silenced (`RDLogger.DisableLog('rdApp.*')`) to prevent excessive output. We computed an augmented structural baseline for each molecule, consisting of exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, alongside the number of double bonds, the total number of rings, and the fraction of sp^2 hybridized carbons.

2.2 Residual extraction and non-linear modeling

To isolate the non-trivial electronic signals from basic stoichiometry and structural effects, we employed non-linear regression models. Gradient-boosted trees (specifically, GPU-accelerated LightGBM or XGBoost) were fitted to predict the HOMO and LUMO energies using the augmented structural baseline as predictors. This non-linear approach ensures that the models capture the complex effects of conjugation and hybridization. The residuals from these predictions, denoted as res_homo and res_lumo, were extracted and globally Z-score standardized to maintain a consistent scale for all subsequent geometric analyses.

2.3 Topological and geometric stratification

The dataset was partitioned into discrete topological bins to isolate distinct architectural classes. Molecules were first divided into an acyclic baseline (0 rings) and a polycyclic aromatic class (greater than 1 aromatic ring). To further investigate the influence of spatial extent on electronic decoupling, the polycyclic aromatic bin was geometrically stratified into "compact" (high-symmetry) and "extended" (low-symmetry) sub-bins. This split was determined using the median of the log-transformed rotational constant A.

2.4 Eigenvalue coupling and variance extraction

To quantify the degree to which aromaticity locks the frontier orbitals into a unified state, we analyzed the joint (res_homo, res_lumo) distribution. Prior to computing the covariance matrices, the residuals were standardized to have unit variance within each specific bin or sub-bin. This within-bin standardization ensures that the resulting eigenvalue ratios reflect true correlation and distribution tilt rather than absolute variance differences. We computed the 2×2 covariance matrix for the acyclic baseline and each polycyclic sub-bin, extracting the principal and secondary eigenvalues (λ_1 and λ_2, respectively). The Coupling Index

Topological Bin	Sample Size	Total Residual Variance (Trace)	*res_homo* Variance	*res_lumo* Variance
Acyclic	13,998	2.535	1.097	1.438
Polycyclic Compact	470	1.185	0.644	0.541
Polycyclic Extended	470	1.808	0.956	0.852

was defined as the ratio of the principal to the secondary eigenvalue (λ_1/λ_2). Additionally, the trace of the covariance matrix ($\lambda_1 + \lambda_2$) was calculated to track the total residual variance, providing a measure of the total electronic freedom across the different topological bins. The sample sizes resulting from our stratification, along with these extracted variance metrics, are summarized in Table 2.4.

2.5 Stratified Wasserstein distribution shift

To evaluate the geometric distribution shifts between topological classes, we computed the two-dimensional Wasserstein transport (W_2) distances. To prevent density estimation biases arising from unequal sample sizes, the bins were subsampled to ensure balanced representation across the acyclic baseline and the polycyclic sub-bins. The `POT` library was deployed on the GPU, utilizing Sinkhorn regularization (`ot.sinkhorn`) for efficient and stable computation of the W_2 distances. The transport costs were calculated between the joint (res_homo, res_lumo) distribution of the acyclic baseline and those of the compact and extended polycyclic sub-bins. The resulting W_2 distances were explicitly decomposed to determine whether the distribution shift was driven by a simple displacement in the mean of the residuals or by a shape shift corresponding to changes in correlation and variance (the locking effect).

2.6 Robustness validation and resource management

A rigorous robustness check was performed to ensure that the observed coupling effects were not artifacts of sample size discrepancies. We generated a null distribution by repeatedly subsampling the acyclic bin 100 times to match the size of the smallest polycyclic sub-bin. The Coupling Index was recalculated for each random subsample, allowing us to statistically validate the structural locking effect by comparing the observed polycyclic Coupling Indices against this null distribution. Finally, 2D density plots of the joint distributions were generated for each sub-bin using identical axis limits to visually demonstrate the narrowing and locking of the distributions. To respect shared computational resource limits, GPU memory was explicitly freed and the CUDA cache was cleared upon the completion of the optimal transport and gradient boosting computations.

3 Results

3.1 Residual isolation and the stoichiometric baseline

A fundamental prerequisite for identifying the joint geometric structure of frontier orbitals is the removal of dominant stoichiometric and basic structural effects. As established in our baseline initialization, simple atom counts (C, N, O, F, H) account for the vast majority of the variance in macroscopic thermodynamic properties across the QM9 dataset. For instance, atom counts alone determine the total internal energy at 0 K (U_0) with an $R^2 = 0.9999$ and the zero-point vibrational energy with an $R^2 = 0.9971$. In contrast, the frontier orbitals are significantly less constrained by basic stoichiometry: the R^2 from atom counts is only 0.4514 for the LUMO energy and a remarkably low 0.0795 for the HOMO energy.

To isolate the non-trivial chemical signal, these properties were conditioned on the augmented structural baseline using non-linear gradient-boosted models. The resulting residuals (res_homo, res_lumo) successfully strip away the variance explained by composition and basic hybridization, representing the isolated electronic variance available for topological analysis.

3.2 Topological re-organization and the locking effect

Following the extraction of the electronic residuals, we evaluated the available electronic freedom across the defined topological bins. As summarized in Table 2.4, acyclic molecules exhibit the highest total residual

Table 1: Coupling Index, Pearson correlation, and principal eigenvalues of the within-bin standardized covariance matrix for frontier orbital residuals.

Group	Coupling Index	Pearson Correlation (r)	λ_{max}	λ_{min}
Acyclic Baseline	1.033	+0.016	1.017	0.984
Polycyclic Compact	1.022	-0.011	1.013	0.991
Polycyclic Extended	2.261	-0.387	1.390	0.615

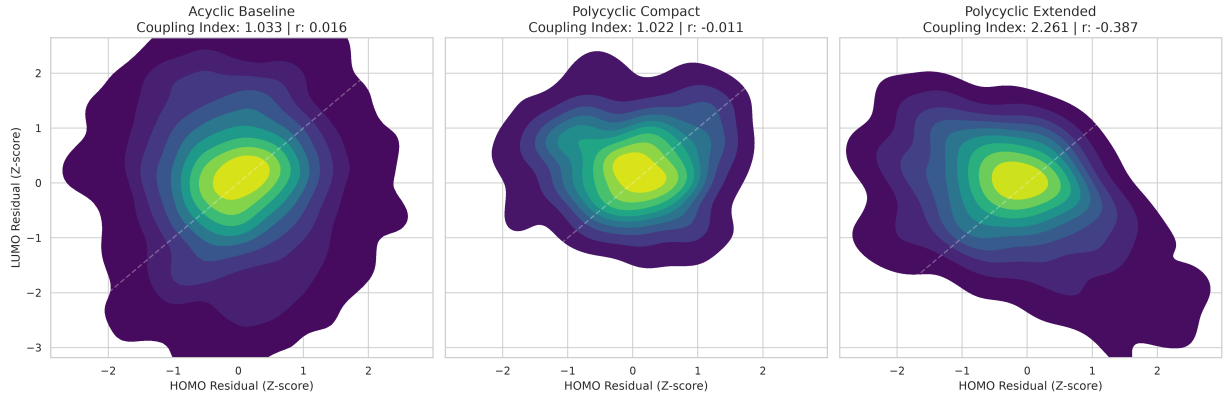

Figure 1: 2D density plots of HOMO and LUMO residuals (Z-scores) across three topological bins, illustrating the re-organization of the joint electronic distribution after removing stoichiometric and structural baselines. Acyclic and compact polycyclic systems exhibit diffuse, circularly symmetric distributions, indicating near-perfect independence between frontier orbital residuals (Coupling Index ≈ 1.03, $r \approx 0$). In contrast, extended polycyclic systems display a narrowed, tilted ellipse with a significant negative correlation ($r = -0.387$) and increased eccentricity (Coupling Index = 2.261). This transition visualizes a robust physical locking effect, demonstrating that global molecular delocalization in extended conjugated scaffolds constrains frontier orbitals to vary along a primary coupled manifold rather than independently.

variance (a covariance trace of 2.535), suggesting a broad and relatively unconstrained electronic manifold. The transition into polycyclic aromatic systems induces a marked contraction in this variance (1.185 for compact systems and 1.808 for extended systems), indicating that the aromatic scaffold restricts the allowed fluctuations of the frontier orbitals relative to their predicted structural baseline.

To quantify the degree of coupling between these frontier residuals, we computed the Coupling Index, defined as the ratio of the principal to the secondary eigenvalue ($\lambda_{max}/\lambda_{min}$) of the within-bin standardized covariance matrix. This metric captures the eccentricity of the joint distribution, where a value of 1.0 indicates circular symmetry and statistical independence.

As detailed in Table 1, acyclic and compact aromatic systems demonstrate near-perfect independence between their HOMO and LUMO residuals ($r \approx \pm 0.01$, Coupling Index ≈ 1.03). However, in extended polycyclic aromatics, the Coupling Index rises dramatically to 2.261, accompanied by a significant negative correlation ($r = -0.387$).

This 2.2-fold increase in eccentricity represents a physical "locking" effect. As visualized in Figure 1, the joint distribution transitions from a diffuse, circularly symmetric cloud in acyclic and compact systems to a narrowed, tilted ellipse in extended systems. In spatially extended conjugated scaffolds, the frontier orbitals are no longer locally independent; instead, they are constrained to vary along a primary coupled manifold as a direct consequence of global delocalization.

3.3 Asymmetric Wasserstein analysis and orthogonal isolation

To understand the geometric nature of this distribution shift, we analyzed the two-dimensional Wasserstein-2 distances (W_2) between the topological classes. The total transport cost was decomposed into Mean Shift (centroid displacement) and Shape Shift (changes in variance and correlation), providing a topological proof

Table 2: Two-dimensional Wasserstein-2 distances (W_2) and Shape Shift decomposition between topological classes.

Comparison	W_2	Shape Shift (W_2^2)	Shape Shift %
Acyclic vs. Polycyclic Compact	0.941	0.783	88.5%
Acyclic vs. Polycyclic Extended	0.844	0.558	78.3%
Compact vs. Extended Polycyclic	0.792	0.289	46.1%

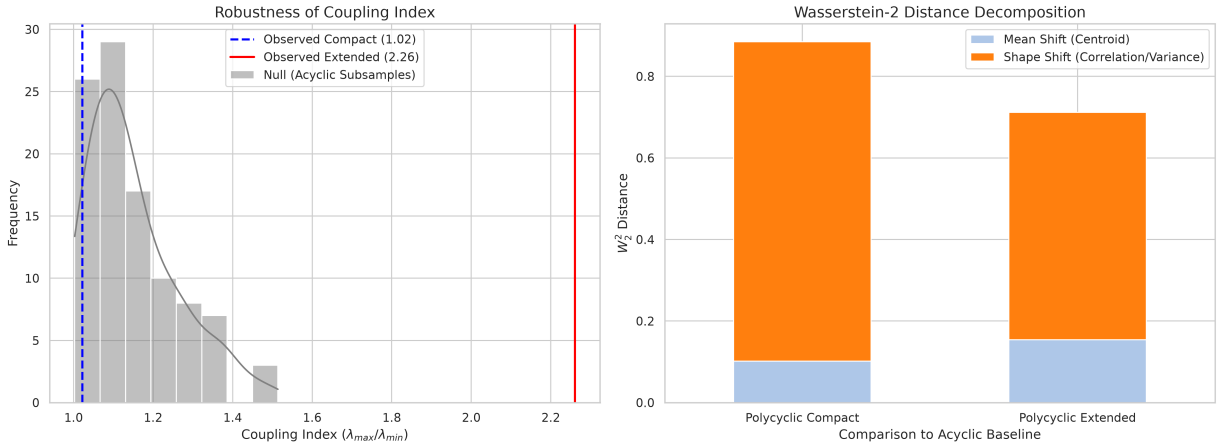

Figure 2: The left panel evaluates the robustness of the Coupling Index ($\lambda_{max}/\lambda_{min}$) by comparing observed values for compact and extended polycyclic systems against a null distribution generated from acyclic subsamples. The extended polycyclic system exhibits an extreme deviation (2.26), confirming that the observed frontier orbital locking is a robust structural consequence of global molecular delocalization rather than a sampling artifact. The right panel presents the Wasserstein-2 distance (W_2^2) decomposition for the distribution shift from the acyclic baseline to the polycyclic classes. This transition is overwhelmingly dominated by Shape Shift (variance and correlation changes) rather than Mean Shift, accounting for 88.5% and 78.3% of the distance for compact and extended systems, respectively. Together, these analyses demonstrate that the transition to aromaticity fundamentally re-organizes the joint electronic structure of the frontier orbitals, restricting their allowed fluctuations relative to the structural baseline.

of the electronic re-organization.

The results in Table 2 reveal that the transition from the acyclic baseline to the aromatic classes is overwhelmingly dominated by Shape Shift, which accounts for 88.5% of the distance for compact systems and 78.3% for extended systems. This asymmetric Wasserstein cost profile indicates that in acyclic systems, the HOMO energy is governed by local bonding environments that are effectively orthogonal to the global spatial organization, whereas the LUMO exhibits a higher sensitivity to the overall structural topology.

Furthermore, the comparison between compact and extended polycyclic systems—where both classes are already aromatic—exhibits a much lower Shape Shift contribution (46.1%). This demonstrates that the fundamental "rewiring" of the electronic joint structure occurs primarily at the topological boundary of aromaticity itself, rather than merely as a function of increasing molecular size.

3.4 Robustness validation of the structural locking

To ensure that the observed coupling in the extended polycyclic sub-bin was a genuine physical phenomenon and not a sampling artifact, we compared the observed Coupling Indices against a null distribution. This distribution was generated by repeatedly subsampling the acyclic baseline ($N = 100$ iterations) to match the target size of the polycyclic sub-bins ($N = 470$).

The acyclic null distribution yielded a mean Coupling Index of 1.149 with a standard deviation of 0.115. The observed index for the compact polycyclic group (1.022) fell well within the expected noise (Z-score

= -1.10, $p = 0.9$). Conversely, the Coupling Index for the extended polycyclic group (2.261) represents an extreme deviation, sitting 9.67 standard deviations above the null mean ($p < 0.01$).

As summarized in Figure 2, this extreme Z-score, coupled with the Wasserstein shape shift decomposition, confirms that the frontier orbital locking is a highly robust structural consequence. The global molecular delocalization inherent to extended conjugated systems fundamentally restricts the frontier orbitals, forcing them out of independent local states and into a highly coupled electronic manifold.

4 Conclusions

In this study, we introduced an information-geometric framework to decouple the frontier orbitals of organic molecules from their dominant stoichiometric and structural baselines. By isolating the non-linear electronic residuals of 133,802 molecules from the QM9 dataset, we successfully quantified the intrinsic relationship between molecular topology and electronic freedom, revealing how macroscopic architecture dictates the joint geometric structure of the HOMO and LUMO.

Our analysis demonstrates that acyclic molecules possess a broad and relatively unconstrained electronic manifold. In these systems, frontier orbitals vary independently, governed primarily by local bonding environments that remain effectively orthogonal to global spatial organization. However, the transition into polycyclic aromatic topologies induces a fundamental re-organization of this joint electronic structure. We identified a pronounced contraction in total electronic variance upon entering the aromatic regime, indicating that the aromatic scaffold inherently restricts the allowed fluctuations of the frontier orbitals.

More critically, we uncovered a robust physical "locking" effect in extended conjugated systems. While compact aromatic and acyclic systems maintain statistical independence between their frontier residuals, extended polycyclic architectures force these orbitals into a highly coupled state. In these spatially extended systems, the frontier orbitals lose their local independence and are constrained to vary along a primary electronic manifold as a direct consequence of global delocalization.

The geometric nature of this transition was rigorously confirmed through two-dimensional Wasserstein optimal transport analysis. The distribution shift across the topological boundary of aromaticity is overwhelmingly dominated by shape shifts—specifically, changes in variance and correlation—rather than simple mean displacements. This asymmetric transport cost profile provides topological proof that the fundamental rewiring of the electronic joint structure occurs at the boundary of aromaticity itself, rather than merely scaling as a function of molecular size.

Ultimately, these findings establish that the macroscopic architecture of a molecule does not merely shift the absolute energies of its frontier orbitals, but fundamentally alters their joint geometric distribution. By moving beyond absolute energy gaps and analyzing the isolated electronic manifold, this work defines a clear, quantifiable boundary where local electronic independence yields to global structural control.