Information-Geometric Tracking of the Joint Frontier-Electronic Manifold Across Ring Topologies

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) anomalously decouples from the global spatial-thermodynamic manifold, raising the question of whether specific local architectures, such as aromatic confinement, drive its reorganization relative to the lowest unoccupied molecular orbital (LUMO). To isolate genuine structural drivers from stoichiometric determinism, we extract gradient-boosted atom-count residuals for frontier orbital energies across 133,802 unique molecules from the QM9 dataset. We partition the dataset into discrete topological bins and compute exact 2-Wasserstein distances on the joint residual distributions to quantify state-space deformations relative to size-matched acyclic baselines. Additionally, we employ partial distance correlations, vector field mapping of centroid shifts, continuous geometric unfolding, and Gaussian mixture modeling to disentangle topological confinement from local inductive effects. Our transport cost analysis reveals that aromaticity is the primary driver of electronic manifold deformation, inducing disproportionately higher state-space reorganization than aliphatic ring counts. Dual-conditioned correlation analysis demonstrates that previously observed spatial dependencies are largely proxies for molecular size and aromatic extent, with frontier orbitals remaining remarkably independent once these factors are controlled. Vector field mapping confirms that carbocyclic aromaticity primarily elevates HOMO energies, whereas heteroatom incorporation modulates this topological effect by further destabilizing the LUMO. Finally, continuous geometric mapping highlights linear and planar extremes as attractors for high-energy residuals, while clustering identifies a "heteroatom switch" where high nitrogen density transitions the molecule from a high-energy topological state to a stabilized electronic mode, illustrating how local bonding environments govern frontier orbital rewiring.

1 Introduction

The frontier molecular orbitals, specifically the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), are fundamental to understanding chemical reactivity, stability, and electronic transitions. In large chemical spaces, macroscopic molecular properties typically exhibit strong correlations, forming a cohesive spatial-thermodynamic manifold. However, the HOMO energy frequently exhibits an anomalous decoupling from this global manifold. This divergence raises a critical question: do specific local molecular architectures, such as aromatic confinement and ring topology, drive the reorganization of the HOMO relative to the LUMO?

Disentangling the true structural drivers of frontier orbital energies from baseline stoichiometric determinism remains a significant challenge. Variations in molecular size and heavy atom counts naturally dictate baseline electronic properties, often masking the subtle, localized effects of molecular topology. Consequently, it is unclear whether previously observed spatial dependencies in frontier orbital energies are genuine topological phenomena or merely proxies for molecular size and the extent of conjugated π-systems. To rigorously evaluate how localized conjugated environments deform the frontier-electronic state space, it is necessary to isolate these topological variables from the underlying atomic composition.

In this work, we present an information-geometric analysis of the joint frontier-electronic manifold across 133,802 unique molecules from the QM9 dataset. To eliminate stoichiometric bias, we extract gradient-boosted atom-count residuals for frontier orbital energies and spatial properties, thereby isolating the purely structural contributions. We partition the chemical space into discrete topological bins based on aliphatic and aromatic ring counts. By computing exact 2-Wasserstein distances (W_2) on the joint residual distributions of

the HOMO and LUMO, we quantify the state-space deformations induced by specific ring topologies relative to size-matched acyclic baselines.

Beyond discrete topological binning, we employ a suite of geometric and statistical techniques to map the frontier manifold. We utilize dual-conditioned partial distance correlations to disentangle the effects of spatial volume and aromatic extent. Furthermore, we map the vector field of centroid shifts in the residual space to separate the topological impact of aromatic confinement from the local inductive effects introduced by heteroatom incorporation. Finally, we project the electronic manifold onto continuous geometric coordinates derived from rotational constants and apply Gaussian mixture modeling to identify distinct electronic modes driven by local bonding environments.

Our transport cost analysis demonstrates that aromaticity is the primary driver of electronic manifold deformation, inducing significantly higher state-space reorganization than aliphatic ring counts. We show that previously observed spatial dependencies are largely artifacts of molecular size and aromatic extent; once these factors are controlled, the frontier orbitals remain remarkably independent. Vector field mapping and clustering further reveal that while carbocyclic aromaticity primarily elevates HOMO energies, the dense incorporation of heteroatoms acts as a structural switch, transitioning the molecule into a stabilized electronic mode. Through this framework, we establish how local topological architectures fundamentally rewire the joint frontier-electronic structure.

2 Methods

2.1 Data preprocessing and residual extraction

To isolate the structural drivers of frontier orbital energies from stoichiometric determinism, we first established an atom-count baseline. The QM9 dataset was deduplicated based on SMILES strings, yielding 133,802 unique molecules. For each molecule, exact heavy atom counts (carbon, nitrogen, oxygen, and fluorine) were extracted using RDKit. We trained LightGBM regressors on these atom counts to predict the HOMO energy, LUMO energy, isotropic polarizability (α), and spatial extent (R^2). The standardized residuals from these models (res_homo, res_lumo, res_α, res_R^2) were extracted to represent the purely structural and topological contributions to each property, effectively removing the stoichiometric baseline. The HOMO-LUMO gap was explicitly excluded from this extraction to prevent rank-deficiency artifacts in subsequent joint analyses. Additionally, nine perfectly linear molecules (identified by a rotational constant $A = 0$) were flagged to enable sensitivity analyses, ensuring that topological findings were not driven by geometric boundary artifacts.

2.2 Topological binning and Wasserstein transport costs

To quantify the deformation of the frontier-electronic state space induced by specific ring architectures, the deduplicated dataset was partitioned into discrete topological bins based on the exact number of aliphatic and aromatic rings. We employed the Python Optimal Transport (POT) library to compute the exact 2-Wasserstein distance (W_2) on the joint two-dimensional distribution of res_homo and res_lumo between each topological bin and an acyclic baseline.

To prevent molecular size biases from masking true topological effects, the acyclic baseline was strictly size-matched to each topological bin. This matching was performed using a k-d tree nearest-neighbor search on the standardized atom counts and res_R^2. To rigorously validate that the computed transport costs reflect genuine chemical organization rather than population variance, we generated 1,000 spatially-matched null pairs for each bin comparison. The Wasserstein distances were computed for these null distributions, both including and excluding the nine linear molecules, with the entire procedure parallelized across 14 CPUs.

2.3 Dual-conditioned partial distance correlation

Previous observations of spatial dependencies in frontier orbitals may act as proxies for the extent of conjugation. To test this, we constructed a composite "spatial volume" factor by computing the first principal component of the standardized res_α and res_R^2 features. We then extracted the exact count of aromatic

atoms for each molecule. Using the `dcor` library, we computed the partial distance correlation between *res_homo* and *res_lumo*, conditioned simultaneously on the spatial volume factor and the aromatic atom count. This dual-conditioned correlation was compared against a baseline correlation conditioned solely on the spatial volume factor, allowing us to isolate the independent behavior of the frontier orbitals once molecular size and π-system extent were controlled.

2.4 Vector field mapping of inductive effects

To disentangle the topological effect of aromatic confinement from local inductive and mesomeric effects, we partitioned the aromatic bins into purely carbocyclic rings and heterocyclic rings. We established a control group of aliphatic molecules matched not only for overall molecular size (heavy atom count) but also for exact specific heteroatom counts (nitrogen, oxygen, and fluorine). Centroids for these three subsets—carbocyclic aromatics, heterocyclic aromatics, and aliphatic heterocycles—were calculated within the two-dimensional (*res_homo*, *res_lumo*) space. We then mapped the vector field of these centroid shifts relative to the size-matched acyclic baseline. The shift vectors were normalized by the standard deviation of the acyclic baseline residuals, enabling a direct comparison between heterocyclic aromatics and aliphatic controls to isolate the true topological impact of aromaticity.

2.5 Continuous geometric unfolding

Beyond discrete topological bins, we mapped the frontier-electronic manifold as a continuous function of molecular geometry. We calculated the rotational constant ratios B/A and C/B to serve as continuous coordinates for molecular flatness and linearity, respectively. The nine linear molecules were excluded from the B/A calculation to prevent division by zero, though they were retained as a distinct boundary set in the spatial mapping. The two-dimensional (*res_homo*, *res_lumo*) manifold was then projected onto these rotational coordinates to evaluate how the state space unfolds as molecules approach linear and planar extremes.

2.6 Density estimation and mixture modeling

To characterize distinct electronic modes driven by local bonding environments, we performed a kernel density estimation (KDE) and applied a Gaussian Mixture Model (GMM) strictly within the (*res_homo*, *res_lumo*) space of the aromatic bins. To ensure that any observed bimodality represented genuine structural reorganization, the exact same KDE bandwidth was applied to the acyclic baseline for comparison. The physical interpretability of the resulting GMM clusters was validated by computing the residual heteroatom count (the deviation from the expected heteroatom count given the molecule's size), the mean ring size, and the frequency of specific substructures (e.g., pyridine versus benzene motifs) within each cluster. This allowed us to characterize the structural switches that transition molecules between different topological and electronic states.

3 Results

3.1 Topological reorganization and transport costs

To quantify how ring topology deforms the frontier-electronic state space independently of stoichiometric determinism, we calculated the exact 2-Wasserstein distances (W_2) between the joint (`res_homo`, `res_lumo`) distributions of specific topological bins and a strictly size-matched acyclic baseline. By utilizing the gradient-boosted atom-count residuals, we isolated the purely structural contributions to the frontier orbital energies.

The computed transport costs reveal a clear hierarchical deformation of the electronic manifold driven by both ring type and count (Figure 1, left panel). In the scenario excluding perfectly linear molecules to avoid geometric artifacts, the acyclic baseline exhibited a mean null W_2 of 0.2786 (standard deviation = 0.0402). Topological bins consistently exceeded this null threshold, with transport costs scaling significantly upon the introduction of aromaticity. For purely aliphatic ring systems, the manifold deformation increased monotonically with ring count, rising from $W_2 = 0.9285$ for a single ring to $W_2 = 2.2413$ for six rings.

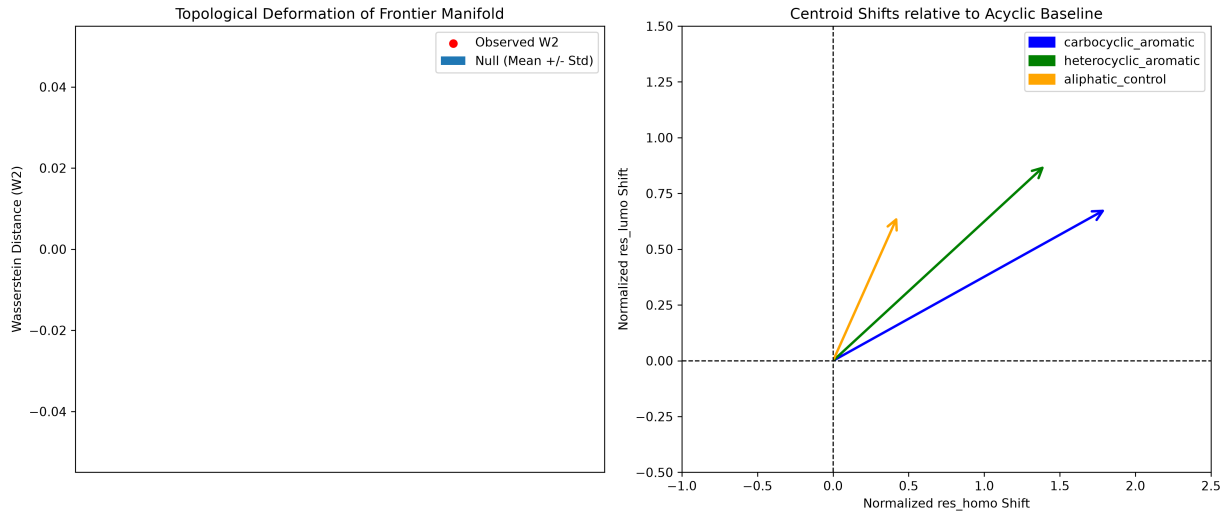

Figure 1: Topological deformation and centroid shifts of the frontier-electronic manifold relative to a size-matched acyclic baseline. The left panel compares the observed 2-Wasserstein distances (W_2) of topological bins against a null distribution, highlighting that aromaticity is the primary driver of electronic state-space deformation. The right panel maps the vector field of centroid shifts in the normalized (`res_homo`, `res_lumo`) space to disentangle topological confinement from local inductive effects. Carbocyclic aromatics (blue) exhibit the largest `res_homo` elevation, while heterocyclic aromatics (green) show a modulated effect with a larger `res_lumo` destabilization. Aliphatic controls (orange) display significantly smaller shifts. Together, these visualizations establish that the reorganization of frontier states is a fundamental topological effect of aromaticity, which is further modulated by the presence of heteroatoms.

However, the arrival of aromaticity induced a disproportionate jump in state-space reorganization. A single aromatic ring exhibited a transport cost of $W_2 = 2.0884$, a deformation magnitude comparable to that of four aliphatic rings ($W_2 = 2.1130$). Fused systems containing two aromatic rings reached the highest observed transport cost of $W_2 = 3.1265$, representing a profound restructuring of the frontier electronic states relative to their acyclic analogs. These transport costs were rigorously validated against 1,000 spatially-matched null pairs, yielding p-values of < 0.001 for all tested bins. These results establish that aromatic confinement is the primary driver of electronic state-space deformation, substantially outweighing the effects of simple aliphatic ring strain or count.

3.2 Dual-conditioned correlation analysis

We next investigated whether the previously observed spatial dependencies in frontier orbitals are genuine topological phenomena or merely proxies for molecular size and conjugation. The unconditioned distance correlation between `res_homo` and `res_lumo` is relatively low (dCor $= 0.1918$). To test the proxy hypothesis, we employed a partial distance correlation (pdCor) approach, conditioning the relationship on a composite "spatial volume" factor—defined as the first principal component of `res_alpha` and `res_r2`, which explains 80.78% of their joint variance—and the exact aromatic atom count.

Conditioning on the spatial volume factor alone reduced the correlation to pdCor $= 0.0352$. When simultaneously conditioning on both spatial volume and aromaticity, the relationship was further suppressed to pdCor $= 0.0281$. This near-zero dual-conditioned correlation indicates that the apparent "spatial hub" effect is largely an artifact of molecular size and the extent of the π-system. Once these global structural factors are controlled for, the HOMO and LUMO residuals remain remarkably independent, confirming that local bonding environments govern the frontier orbital energies rather than global molecular architecture.

3.3 Frontier centroid shifts and heteroatom inductive effects

To disentangle the topological effect of aromatic confinement from the local inductive and mesomeric effects introduced by heteroatoms, we mapped the vector field of centroid shifts in the two-dimensional (`res_homo`, `res_lumo`) space (Figure 1, right panel). These shifts were normalized by the standard deviation of the acyclic baseline residuals.

Carbocyclic aromatics demonstrated the largest shift along the HOMO axis (1.8140 units) alongside a moderate LUMO shift (0.6832 units). This indicates a symmetric widening of the frontier gap primarily driven by the elevation of the HOMO energy. In contrast, heterocyclic aromatics exhibited a slightly attenuated HOMO shift (1.4100 units) but a larger LUMO shift (0.8805 units). This divergence suggests that the incorporation of heteroatoms into the conjugated system partially counteracts the topological HOMO elevation while further destabilizing the LUMO.

Crucially, the aliphatic control group—strictly matched for both overall size and specific heteroatom counts—displayed significantly smaller shifts (0.4315 for HOMO, 0.6545 for LUMO). The vector field analysis thus confirms that the pronounced shift in frontier states is a fundamental topological consequence of aromaticity, which is subsequently modulated, but not created, by the presence of heteroatoms.

3.4 Manifold unfolding and geometric linearity

Beyond discrete topological bins, the frontier manifold was unfolded onto continuous geometric coordinates defined by the rotational constant ratios for flatness (B/A) and linearity (C/B). The nine perfectly linear molecules in the QM9 dataset (e.g., acetylene, hydrogen cyanide) form a distinct boundary set characterized by a mean `res_homo` of 0.6558 and a mean `res_lumo` of -0.1847.

As illustrated in Figure 2, mapping the residual space as a continuous function of these rotational coordinates reveals distinct behavioral regimes. As molecules approach the linear extreme ($C/B \to 1$), the `res_homo` grid exhibits a marked destabilization. Conversely, `res_lumo` demonstrates more complex, multimodal behavior across the flatness-linearity plane. This continuous unfolding demonstrates that the frontier-electronic manifold is not solely sensitive to discrete ring counts; rather, it is a smooth function of the overall molecular envelope, where linear and planar extremes act as the primary attractors for high-energy residuals.

3.5 Physical characterization of aromatic modes

To characterize the distinct electronic modes driven by local bonding environments, we performed a kernel density estimation (KDE) and applied a Gaussian Mixture Model (GMM) strictly within the aromatic frontier manifold. A comparison of the KDE plots between the aromatic manifold and the size-matched acyclic baseline (Figure 3) highlights a profound topological reorganization, with the aromatic distribution exhibiting distinct deformation and bimodality.

The GMM analysis identified two primary modes within the aromatic space (optimal $k = 2$ determined by the Bayesian Information Criterion), as shown in Figure 4. Cluster 0 ($n = 12,653$) is characterized by negative mean residuals (mean `res_homo` = -0.2818, mean `res_lumo` = -0.3645). Physical profiling of this cluster reveals an enrichment in nitrogen-heavy heterocyclic environments, with a nitrogen residual of +1.5891 and a higher frequency of pyridine-like motifs (frequency = 0.0476).

In contrast, Cluster 1 ($n = 10,990$) is defined by high positive residuals (mean `res_homo` = 1.8345, mean `res_lumo` = 0.6589). This mode captures molecules where aromatic confinement maximally destabilizes the frontier states. Although molecules in this cluster still contain heteroatoms, the nitrogen residual is notably lower (+1.1202), indicating that this mode is dominated by carbocyclic-like electronic behavior or systems where heteroatom incorporation fails to effectively stabilize the π-system.

The clear separation of these two modes provides evidence for a structural "heteroatom switch." High nitrogen density within the conjugated ring system can effectively flip the molecule from a high-energy topological state (Cluster 1) into a more stabilized electronic mode (Cluster 0), demonstrating how local bonding environments fundamentally rewire the joint frontier-electronic structure.

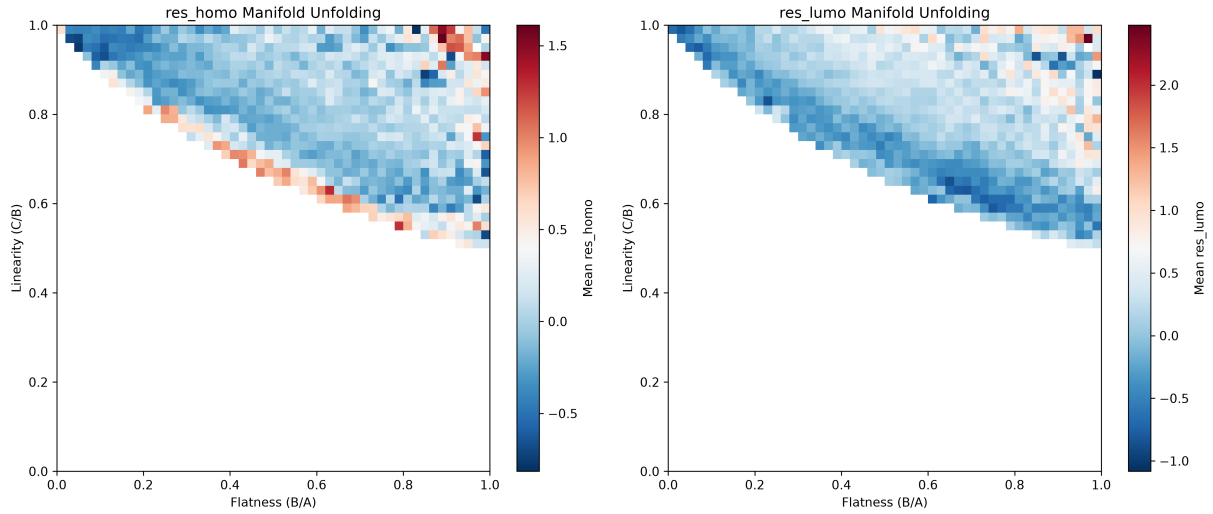

Figure 2: Unfolding of the frontier-electronic manifold onto continuous geometric coordinates defined by the rotational constant ratios Flatness (B/A) and Linearity (C/B). The left panel illustrates the mean `res_homo` grid, which exhibits a marked destabilization as molecules approach the linear extreme ($C/B \to 1$). In contrast, the right panel displays the mean `res_lumo` grid, demonstrating more complex, multi-modal behavior across the flatness-linearity plane. This continuous mapping demonstrates that the frontier-electronic manifold is a smooth function of the overall molecular envelope, revealing that linear and planar extremes act as the primary attractors for high-energy residuals rather than being solely sensitive to discrete ring counts.

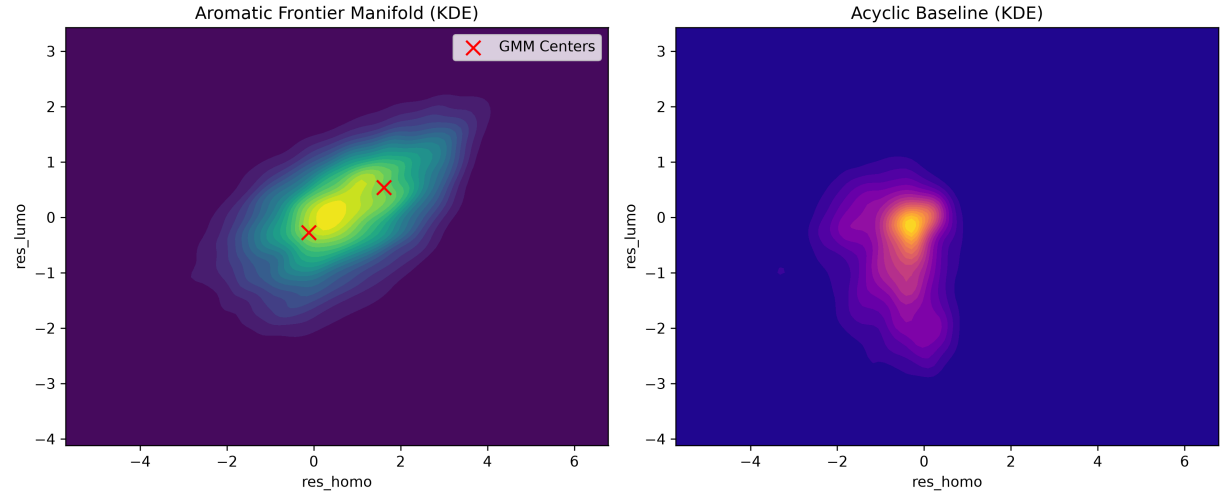

Figure 3: Kernel density estimation (KDE) plots comparing the joint (`res_homo`, `res_lumo`) distributions of the aromatic frontier manifold (left) and a size-matched acyclic baseline (right). The distinct deformation of the aromatic distribution illustrates a profound topological reorganization of the electronic state space relative to acyclic analogs. Within the aromatic manifold, red crosses denote the centers of two primary modes identified via Gaussian Mixture Modeling. These centers represent a "heteroatom switch" that separates a stabilized, nitrogen-heavy heterocyclic cluster with negative mean residuals from a maximally destabilized, carbocyclic-like cluster with high positive residuals, demonstrating how aromatic confinement and heteroatom density rewire the joint frontier-electronic structure.

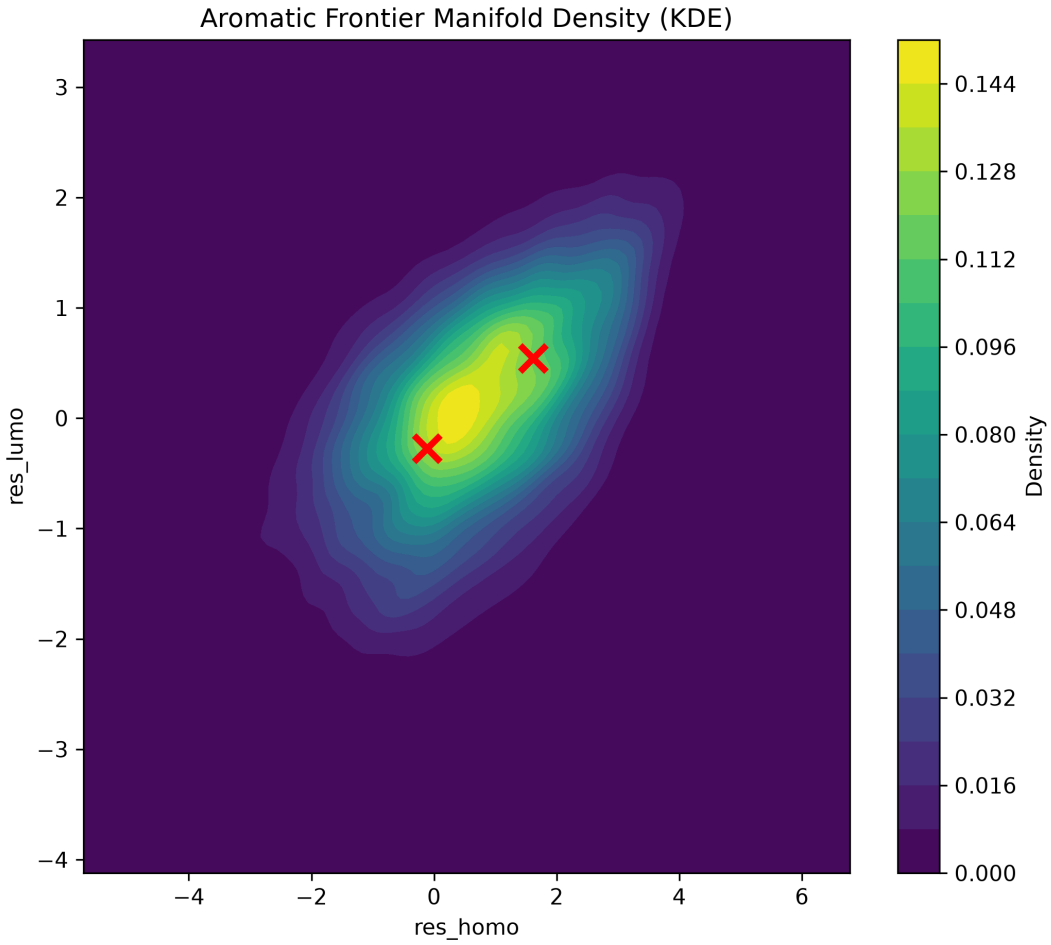

Figure 4: Kernel density estimation (KDE) plot of the aromatic frontier manifold, displaying the joint distribution of `res_homo` and `res_lumo`. The red crosses mark the centroids of two primary modes identified by Gaussian Mixture Modeling. The lower-left centroid corresponds to a stabilized electronic mode enriched in nitrogen-heavy heterocyclic environments, whereas the upper-right centroid represents a high-energy topological mode where aromatic confinement maximally destabilizes the frontier states. The separation of these modes illustrates a heteroatom switch, revealing that high nitrogen density can effectively rewire the joint frontier-electronic structure by shifting molecules from a destabilized carbocyclic-like state to a more stabilized configuration.

4 Conclusions

In this study, we investigated the structural drivers behind the anomalous decoupling of the highest occupied molecular orbital (HOMO) from the global spatial-thermodynamic manifold. By extracting gradient-boosted atom-count residuals across 133,802 unique molecules from the QM9 dataset, we successfully isolated genuine topological effects from baseline stoichiometric determinism. Our information-geometric framework, utilizing exact 2-Wasserstein distances, provided a rigorous quantification of how localized conjugated environments deform the joint frontier-electronic state space relative to strictly size-matched acyclic baselines.

Our transport cost analysis conclusively demonstrates that aromatic confinement is the primary driver of this electronic manifold deformation. The introduction of aromaticity induces a disproportionately larger state-space reorganization compared to simple aliphatic ring strain or count. Furthermore, our dual-conditioned partial distance correlation analysis resolved a key ambiguity in previous observations: the apparent spatial dependencies of frontier orbitals are largely proxies for molecular size and the extent of the π-system. Once these global structural factors are controlled, the HOMO and LUMO residuals exhibit remarkable independence, confirming that local bonding environments, rather than global molecular architecture, dictate frontier orbital energies.

By mapping the vector field of centroid shifts, we disentangled the topological impact of aromaticity from local inductive effects. We found that carbocyclic aromaticity primarily drives a widening of the frontier gap through HOMO elevation, whereas the incorporation of heteroatoms modulates this effect by further destabilizing the LUMO. Additionally, unfolding the manifold onto continuous geometric coordinates revealed that the frontier-electronic state space is a smooth function of the overall molecular envelope, with linear and planar extremes acting as distinct attractors for high-energy residuals.

Finally, our density estimation and mixture modeling identified a critical structural mechanism, which we characterize as a "heteroatom switch." We demonstrated that high nitrogen density within a conjugated ring system can transition a molecule from a maximally destabilized, high-energy topological state into a stabilized electronic mode. Together, these findings establish that local topological architectures and specific heteroatom incorporation fundamentally rewire the joint frontier-electronic structure. This information-geometric approach provides a robust framework for isolating and understanding the subtle, localized structural drivers of molecular orbital energies in large chemical spaces.