Wasserstein Tracking of the Frontier-Electronic Manifold Reveals HOMO Isolation and Refutes Gap Anchoring

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The organization of molecular properties is frequently dominated by molecular size, obscuring the underlying topological dependencies of the frontier-electronic manifold. To isolate structural sensitivity from this stoichiometric determinism, we analyze the QM9 dataset by extracting gradient-boosted atom-count residuals and evaluating the remaining variance. By employing bootstrapped distance correlation and computing exact Wasserstein distances between the residual point clouds of acyclic baseline molecules and distinct cyclic scaffolds, we track the geometric deformation of the joint electronic-response manifold. Our methodology reveals that electronic properties retain substantial structured variance independent of molecular size, with the highest occupied molecular orbital (HOMO) remaining remarkably isolated from the global spatial and thermodynamic state. Marginal transport cost decomposition indicates that the frontier-electronic manifold shifts as a coherent unit during structural transitions, where aromaticity drives the distribution closer to the acyclic baseline and ring strain induces the largest deformations. Ultimately, variance analysis of these marginal shifts refutes the gap anchoring hypothesis, showing that structural transitions increase frontier gap variability rather than inducing compensatory stabilizing shifts.

1 Introduction

The organization of molecular properties in large chemical spaces is frequently dominated by molecular size. In comprehensive databases such as QM9, extensive thermodynamic properties like internal energy (U_0) and zero-point vibrational energy are almost entirely determined by exact atom counts. This stoichiometric determinism obscures the underlying topological and structural dependencies of the frontier-electronic manifold, which ultimately governs chemical reactivity, stability, and optical behavior. To uncover the true physical rules dictating electronic structure, it is necessary to isolate structural sensitivity from these overwhelming size effects.

Within the frontier-electronic manifold—comprising the highest occupied molecular orbital (HOMO) energy, lowest unoccupied molecular orbital (LUMO) energy, and isotropic polarizability (α)—the HOMO energy often exhibits an anomalous decoupling from the global thermodynamic and spatial state of the molecule. It remains an open question whether this isolation is a universal physical constant driven by localized phenomena, or a topology-dependent artifact that collapses in highly constrained or conjugated systems. Furthermore, prevailing chemical intuition often assumes a "gap anchoring" mechanism, hypothesizing that structural transitions induce correlated, compensatory shifts in HOMO and LUMO levels to preserve a stable energy gap. Rigorously testing these hypotheses requires a mathematical framework capable of tracking the geometric deformation of the joint electronic state space across distinct structural families without the confounding influence of molecular scale.

In this work, we introduce a robust statistical and optimal transport framework to track the frontier-electronic manifold across the Murcko scaffold hierarchy. By extracting gradient-boosted atom-count residuals from the QM9 dataset, we eliminate stoichiometric determinism and evaluate the remaining structurally driven variance. We employ bootstrapped distance correlation to profile the baseline non-linear couplings of the residualized properties, explicitly avoiding the rank-deficient inclusion of the HOMO-LUMO gap. To quantify structural rewiring, we compute exact Wasserstein distances (W_2^2) between the multivariate resid-

ual point clouds of acyclic baseline molecules and distinct cyclic scaffolds, as well as a strictly defined linear molecule class.

By decomposing the optimal transport plans into marginal distribution shifts, we map the precise spatial confinements and topological metrics—specifically aromaticity and small-ring strain—that drive manifold deformation. Our methodology reveals that electronic properties retain substantial structured variance independent of molecular size, with the HOMO remaining remarkably isolated. We demonstrate that the frontier-electronic manifold shifts as a coherent unit during structural transitions, where aromaticity surprisingly drives the distribution closer to the acyclic baseline and ring strain induces the largest deformations. Ultimately, variance analysis of these marginal shifts mathematically refutes the gap anchoring hypothesis, showing that structural transitions increase frontier gap variability rather than inducing compensatory stabilizing shifts.

2 Methods

2.1 Data initialization and structural partitioning

To rigorously isolate the structural dependencies of the frontier-electronic manifold, we utilized the QM9 dataset, which provides quantum chemical properties for small organic molecules. The dataset was first parsed and deduplicated by SMILES strings to prevent structural leakage during comparative analyses. Using RDKit, we extracted exact heavy atom counts (C, N, O, F, H), the number of aromatic bonds, and small-ring strain indices for each molecule.

The dataset was subsequently partitioned into distinct structural subsets to facilitate comparative manifold analysis. We defined an acyclic baseline using Murcko scaffolds, isolating molecules devoid of ring structures. The remaining molecules were grouped into distinct cyclic scaffold families. Additionally, a dedicated "extreme constraint" class consisting of linear molecules was identified by filtering for a rotational constant of $A = 0$. These linear molecules were explicitly retained as a critical boundary case for testing structural sensitivity under extreme spatial constraints.

2.2 Stoichiometric residual extraction and validation

To eliminate the stoichiometric determinism that dominates extensive molecular properties, we regressed 12 non-redundant properties against the exact heavy atom counts using gradient-boosted decision trees (XGBoost). The residuals from these regressions (e.g., `res_homo`, `res_lumo`, `res_alpha`, `res_cv`, `res_u0`) represent the variance attributable to topological arrangement rather than atomic composition.

To ensure that this regression procedure did not over-clean the physical signal, we validated the residuals by calculating the ratio of the residual variance to the original property variance. This metric quantitatively confirmed that chemically relevant variance was preserved for electronic properties, whereas purely size-dominated properties (such as internal energy, U_0) collapsed largely to noise. Following validation, all extracted residuals were globally standardized using Z-score scaling so that subsequent manifold comparisons reflected geometric shape rather than arbitrary differences in scale.

2.3 Bootstrapped distance correlation profiling

To establish the baseline non-linear couplings between the residualized properties, we executed a bootstrap resampling protocol consisting of 500 iterations using subsampling with replacement on the standardized residuals. For each iteration, we computed the 12×12 Distance Correlation (dCor) matrix and extracted the corresponding confidence intervals. This non-linear dependency profiling was used to quantify the specific connectivity of the HOMO residual to the broader frontier-response manifold (LUMO and isotropic polarizability, α). By comparing these couplings against a stable "null" property (heat capacity, `res_cv`), we established a baseline for stability and confirmed the orthogonal isolation of the HOMO energy within the global dataset.

2.4 Wasserstein transport computation

We defined the joint electronic-response manifold using `res_homo`, `res_lumo`, and `res_alpha`. The HOMO-LUMO gap was explicitly excluded from this manifold to avoid rank-deficient feature dependence. To quantify the geometric deformation of this state space across structural families, we deployed the Python Optimal Transport (`POT`) library to compute the exact squared Wasserstein distance (W_2^2) between the multivariate residual point cloud of the acyclic baseline and the point clouds of the distinct cyclic scaffolds and linear molecules.

Because optimal transport computations can be numerically sensitive to high-dimensional point cloud geometries, we utilized entropic regularization. The transport costs were computed across a range of regularization parameters (ϵ) to rigorously confirm that the observed manifold deformations were stable, representing genuine chemical realities rather than numerical artifacts.

2.5 Marginal transport cost decomposition and gap anchoring analysis

To identify which specific properties drive the deformation of the joint manifold, we decomposed the optimal transport plans feature-wise to extract the marginal distribution shifts for each property. The marginal transport cost of `res_homo` was explicitly compared against the null property (`res_cv`) across all evaluated scaffolds. This comparison provided the necessary quantitative evidence to distinguish between general structural sensitivity and HOMO-specific isolation.

Furthermore, we utilized these marginal costs to mathematically test the gap anchoring hypothesis. We calculated the sum of the marginal transport costs of `res_homo` and `res_lumo` for each structural transition. We then compared the variance of this combined cost across the scaffold hierarchy to the sum of the variances of the individual HOMO and LUMO costs. A stable sum of marginal costs (where the variance of the sum is significantly lower than the sum of the individual variances due to negative covariance) would indicate that the frontier manifold reconfigures to anchor the gap. Conversely, an increased variance would refute the hypothesis, indicating independent or exacerbating fluctuations.

2.6 Topological mapping and dimensionality reduction

Finally, we mapped the extracted marginal transport costs directly onto the RDKit-derived topological metrics. We conducted a sensitivity analysis by calculating the Spearman correlation between the magnitude of the HOMO marginal distribution shift and continuous structural descriptors, specifically the number of aromatic bonds and small-ring strain indices. To provide visual proof of the structural rewiring mapped to the degree of spatial confinement, we generated 2D Uniform Manifold Approximation and Projection (UMAP) embeddings of the residual point clouds. These projections compared the acyclic baseline against highly aromatic and linear subsets, with the Wasserstein costs overlaid to illustrate the topological deformation of the frontier-electronic manifold.

3 Results

3.1 Stoichiometric residuals and property independence

To isolate the structural dependencies of the frontier-electronic manifold from the overwhelming influence of molecular size, we first evaluated the efficacy of our stoichiometric residual extraction. By regressing the 12 non-redundant properties against exact atom counts (C, N, O, F, H), we quantified the degree of stoichiometric determinism inherent to each property class. The analysis of residual variances confirms a sharp divergence between thermodynamic and electronic properties. Extensive thermodynamic properties are almost entirely determined by atom counts; for instance, the residual variance of internal energy (U_0) is reduced to merely 0.041% of its original variance, and zero-point vibrational energy (ZPVE) retains only 0.27%.

In stark contrast, the frontier-electronic properties retain substantial structured variance after the removal of compositional effects. The highest occupied molecular orbital (HOMO) energy preserves 80.3% of its original variance (`res_homo`), the lowest unoccupied molecular orbital (LUMO) energy preserves 51.8%

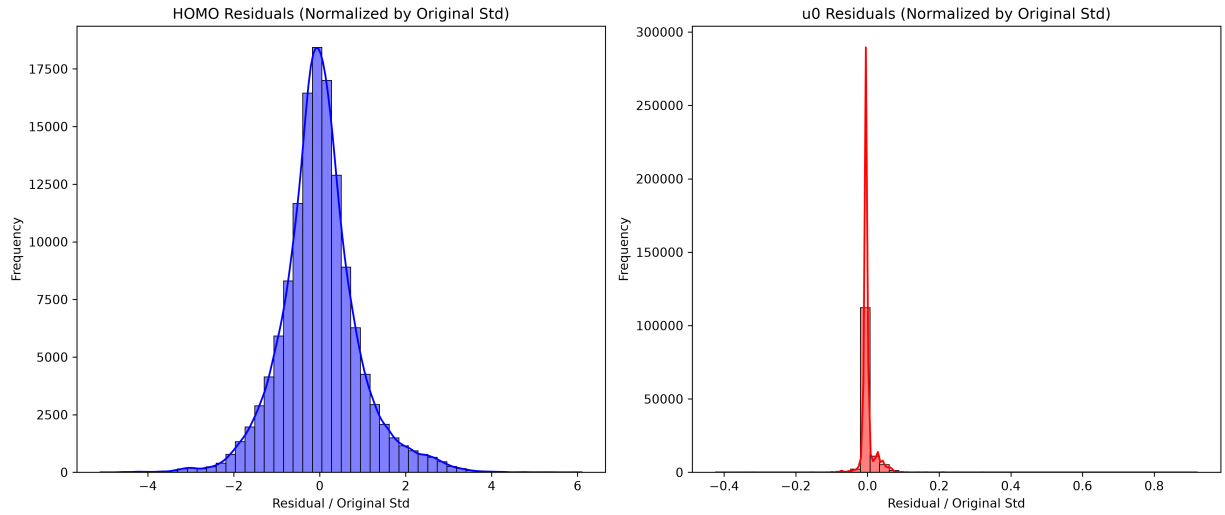

Figure 1: Distributions of stoichiometric residuals for HOMO energy (left) and internal energy U_0 (right), normalized by their original standard deviations. The broad distribution of `res_homo` demonstrates that it retains substantial variance (80.3%) after removing compositional effects, indicating that frontier orbital energies are primarily governed by local bonding environments and topology. In contrast, the nearly collapsed distribution of U_0 confirms that thermodynamic properties are almost entirely determined by molecular scale, preserving only 0.041% of their original variance.

(`res_lumo`), and isotropic polarizability preserves 12.2% (`res_alpha`). This preservation is visually confirmed in Figure 1, which contrasts the broad, structurally driven distribution of `res_homo` against the nearly collapsed, noise-like distribution of the thermodynamic residual `res_u0`. These results validate our methodology, demonstrating that HOMO and LUMO energies are primarily governed by local bonding environments and topology rather than molecular scale.

3.2 Global joint structure and HOMO isolation

Having established the validity of the stoichiometric residuals, we investigated the global joint structure of the property manifold using bootstrapped distance correlation (dCor). The resulting dependency matrix reveals that while many molecular properties exhibit moderate to strong non-linear couplings, the HOMO energy remains remarkably isolated from the global spatial and thermodynamic state.

As illustrated in Figure 2, geometric properties such as rotational constants (`res_B`, `res_C`) and electronic spatial extent (`res_r2`) form a tightly coupled cluster with median dCor values exceeding 0.84. However, `res_homo` exhibits only weak non-linear dependency with the other components of the frontier-response manifold, correlating with `res_lumo` at a median dCor of 0.18 and with `res_alpha` at 0.16. Furthermore, the analysis highlights a distinct HOMO-LUMO asymmetry: `res_lumo` participates much more strongly in the global manifold, correlating with `res_alpha` at 0.38 and with the atomization energy residual (`res_u0_atom`) at 0.33. In contrast, `res_homo` remains below 0.23 in all pairwise comparisons. This orthogonal isolation suggests that HOMO energy is governed by highly localized electronic physics—such as specific functional groups or lone pairs—that operate independently of the broader molecular architecture.

3.3 Structural rewiring and Wasserstein transport

To quantify how specific topological transitions deform the frontier-electronic manifold, we computed the exact squared Wasserstein distance (W_2^2) between the acyclic baseline and various cyclic scaffolds. The total transport costs vary by an order of magnitude depending on the structural family. Aromatic scaffolds, such as pyrrole (`c1cc[nH]c1`), remain remarkably close to the acyclic baseline ($W_2^2 = 5 \times 10^{-6}$). Conversely,

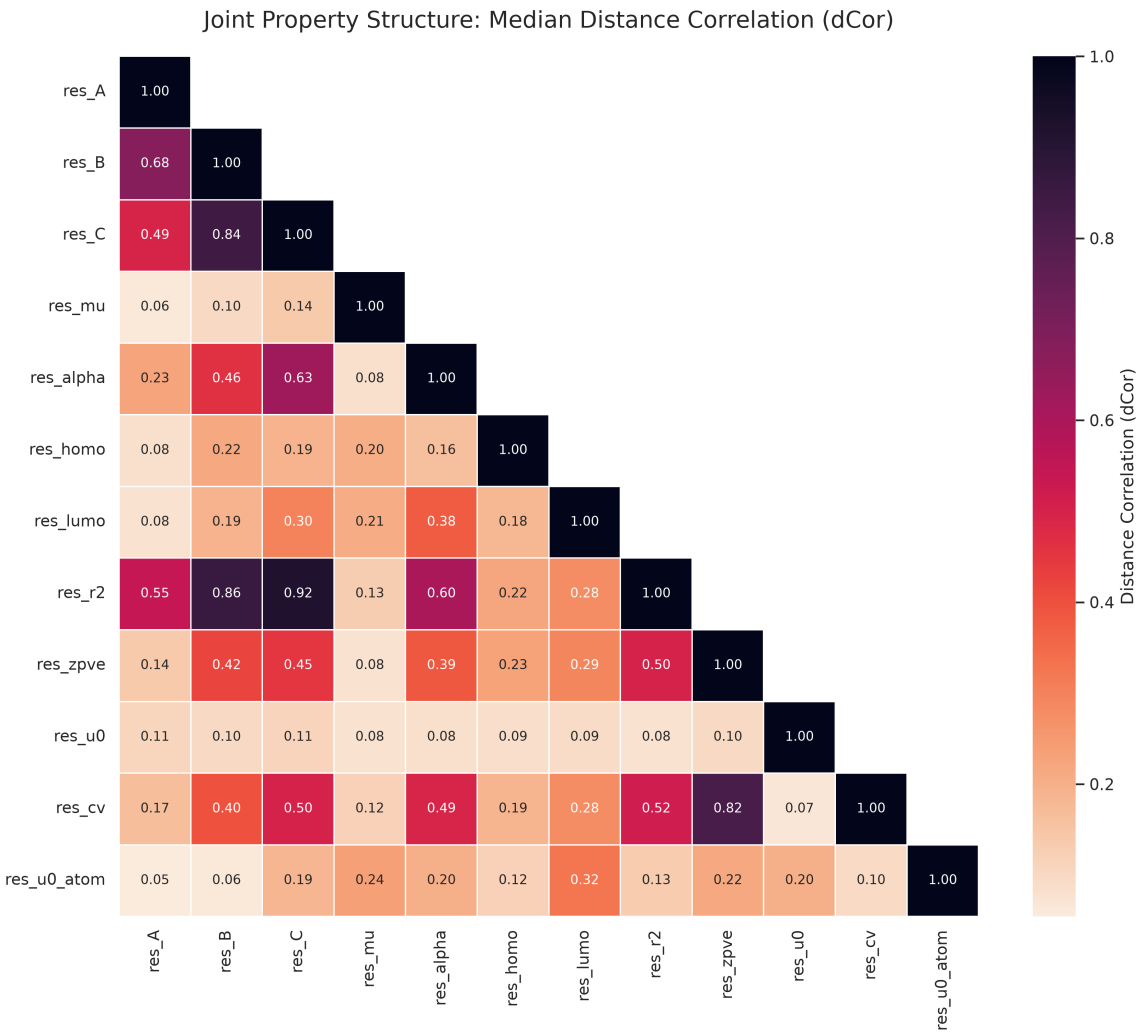

Figure 2: Heatmap of the median distance correlation (dCor) between standardized stoichiometric residuals, illustrating the global joint structure of molecular properties independent of molecular size. While geometric properties (`res_B`, `res_C`, and `res_r2`) form a tightly coupled cluster with dCor values exceeding 0.84, the HOMO energy (`res_homo`) remains remarkably isolated. It shows weak non-linear dependency with other frontier-electronic properties like `res_lumo` (0.18) and `res_alpha` (0.16). The plot also highlights a HOMO-LUMO asymmetry, where `res_lumo` exhibits stronger coupling with the global manifold, such as with `res_alpha` (0.38). This overall isolation indicates that HOMO energy is governed by localized electronic physics that are orthogonal to the broader spatial and thermodynamic states of the molecule.

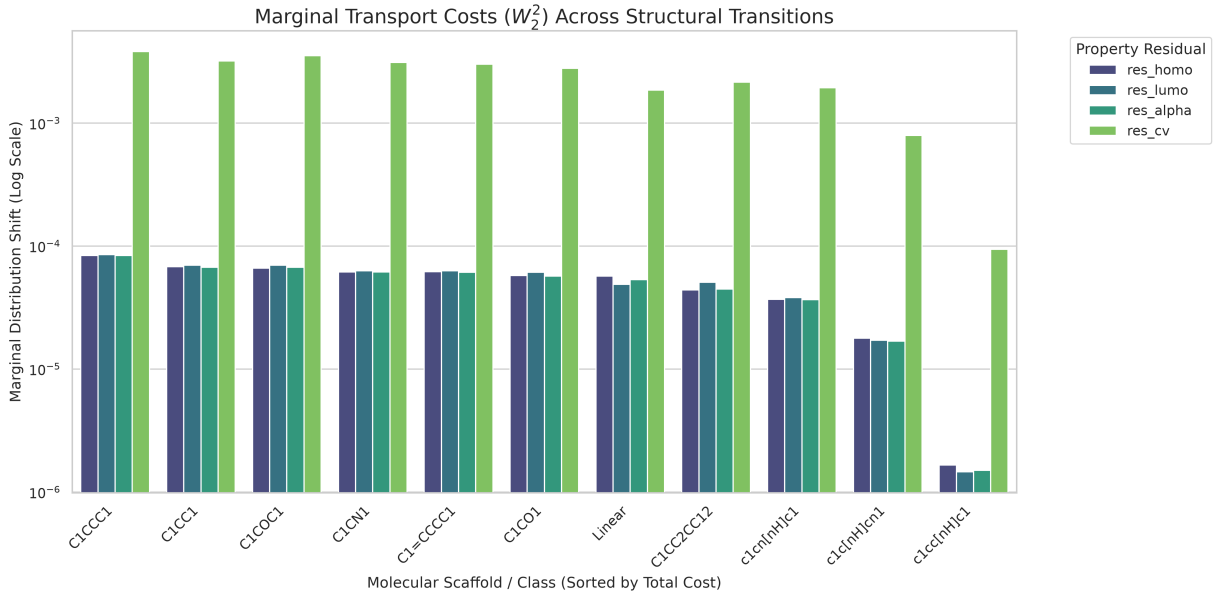

Figure 3: Marginal transport costs (W_2^2) for property residuals across various cyclic scaffolds relative to an acyclic baseline. The marginal distribution shifts for the frontier-electronic manifold (`res_homo`, `res_lumo`, and `res_alpha`) are strikingly uniform, indicating that these properties shift as a coherent unit during structural transitions rather than rewiring in isolation. In contrast, the shift for the thermodynamic null property, heat capacity (`res_cv`), is consistently 10 to 50 times larger, demonstrating that the frontier-electronic manifold is significantly more stable under structural changes than the thermodynamic manifold. Scaffolds are sorted by total transport cost, revealing that strained rings like cyclobutane (C1CCC1) drive the largest manifold deformations, while aromatic systems like pyrrole (c1cc[nH]c1) remain closest to the acyclic baseline.

highly strained non-aromatic scaffolds, such as cyclobutane (`C1CCC1`), drive the largest manifold deformations ($W_2^2 = 2.53 \times 10^{-4}$).

Decomposing the optimal transport plans into marginal shifts provides further insight into the drivers of this deformation (Figure 3). Across the evaluated cyclic scaffolds, the marginal shifts for the frontier-electronic properties (`res_homo`, `res_lumo`, and `res_alpha`) are strikingly uniform, typically ranging between 4×10^{-5} and 8×10^{-5}. This uniformity indicates that structural transitions do not rewire individual electronic properties in isolation; rather, the frontier-electronic manifold shifts as a coherent unit. Importantly, the marginal shift of the thermodynamic null property, heat capacity (`res_cv`), is consistently 10 to 50 times larger (e.g., 3.81×10^{-3} for cyclobutane). This stark contrast demonstrates that the frontier-electronic manifold is significantly more stable under structural rewiring than the thermodynamic manifold.

Correlation of the `res_homo` marginal shifts with RDKit-derived topological metrics identifies aromaticity and ring strain as the primary, opposing drivers of manifold stability. The Spearman correlation between the mean number of aromatic bonds and the HOMO marginal shift is strongly negative ($\rho = -0.775$, $p = 0.0051$), implying that increased aromaticity anchors the HOMO distribution closer to the acyclic baseline. In contrast, small-ring strain indices for 3- and 4-membered rings correlate positively with HOMO shifts ($\rho = 0.179$ and 0.387, respectively), confirming that spatial confinement and angular strain are the primary catalysts for electronic manifold deformation.

3.4 Evaluation of the gap anchoring model

A prevailing hypothesis in chemical intuition is "gap anchoring," which posits that structural changes induce correlated, compensatory shifts in HOMO and LUMO levels to preserve a stable energy gap. We rigorously tested this model by comparing the variance of the individual marginal transport costs across scaffolds to

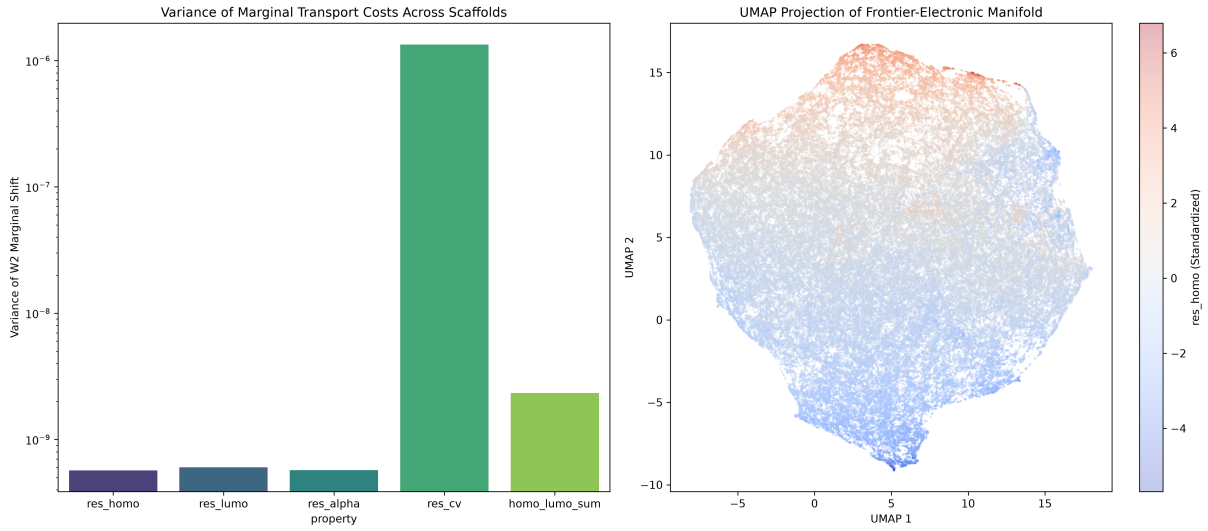

Figure 4: The left panel displays the variance of the W_2^2 marginal transport costs across cyclic scaffolds for electronic residuals (`res_homo`, `res_lumo`, `res_alpha`), a thermodynamic null property (`res_cv`), and the sum of HOMO and LUMO shifts (`homo_lumo_sum`). The significantly higher variance for `res_cv` demonstrates that the frontier-electronic manifold is much more stable under structural rewiring than the thermodynamic manifold. Furthermore, the variance of `homo_lumo_sum` is greater than the sum of the individual `res_homo` and `res_lumo` variances, mathematically refuting the Gap Anchoring hypothesis by showing that structural transitions increase gap variability rather than inducing compensatory stabilizing shifts. The right panel shows a UMAP projection of the frontier-electronic manifold colored by the standardized HOMO residual (`res_homo`), visually reinforcing the lack of gap anchoring and illustrating that frontier orbitals are governed by distinct structural sensitivities that do not collaborate to maintain electronic stability during topological transitions.

the variance of their sum.

If the frontier gap were anchored during structural transitions, negative covariance between the HOMO and LUMO shifts would cause the variance of their sum to be lower than the sum of their individual variances. Our analysis reveals the opposite: the variance of the marginal transport costs across scaffolds is 5.69×10^{-10} for `res_homo` and 6.02×10^{-10} for `res_lumo`, while the variance of their combined sum (`res_homo` + `res_lumo`) is 2.33×10^{-9}. Because the joint variance is nearly twice the sum of the individual variances, the gap anchoring model is mathematically refuted.

This refutation is visually supported by Figure 4 and Figure 5. The left panel of Figure 4 highlights the exacerbating fluctuations of the combined frontier shifts, while the right panel provides a UMAP projection of the frontier-electronic manifold colored by the standardized HOMO residual. Figure 5 further projects this manifold categorized by structural class. These UMAP embeddings demonstrate that structural transitions in the QM9 dataset do not induce compensatory shifts to stabilize the frontier gap. Instead, HOMO and LUMO energy distributions shift independently in response to topological constraints, increasing gap variability across structural families. Ultimately, these findings confirm that frontier orbitals are governed by distinct structural sensitivities that do not collaborate to maintain electronic stability during topological transitions.

4 Conclusions

The organization of molecular properties in large chemical spaces is often obscured by stoichiometric determinism, where extensive properties scale predictably with molecular size. By extracting gradient-boosted atom-count residuals from the QM9 dataset, we successfully isolated the underlying topological dependencies of the frontier-electronic manifold. Our analysis demonstrates that while thermodynamic properties like internal energy (U_0) collapse almost entirely to noise upon the removal of compositional effects, frontier-

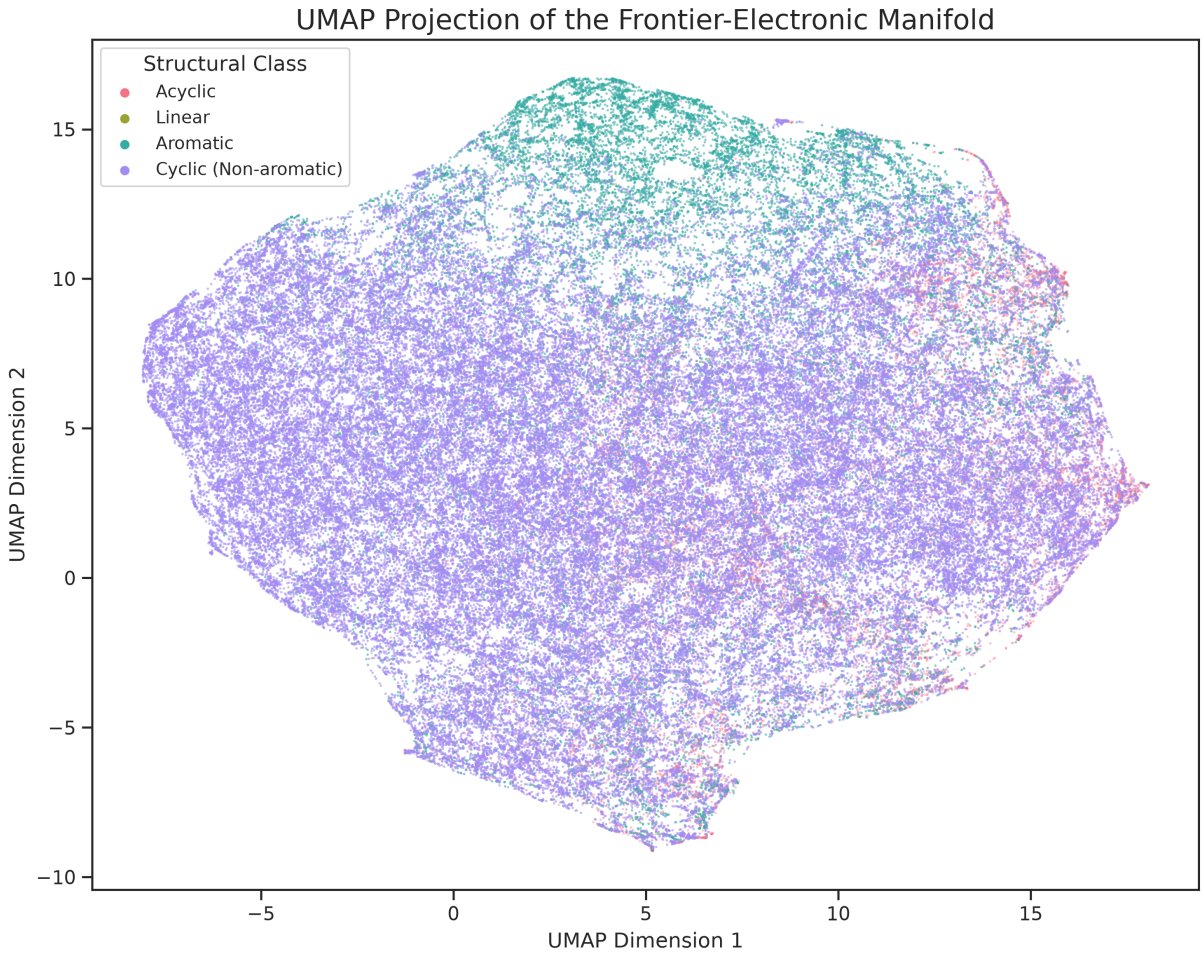

Figure 5: UMAP projection of the frontier-electronic manifold categorized by molecular structural class. The visualization illustrates the lack of gap anchoring across topological families, demonstrating that HOMO and LUMO energies are governed by distinct structural sensitivities rather than collaborating to maintain electronic stability during structural transitions. Additionally, the projection reflects how aromatic systems maintain a distribution closer to the acyclic baseline, whereas non-aromatic cyclic structures exhibit broader manifold deformation driven by ring strain.

electronic properties retain substantial structured variance. This confirms that orbital energies and polarizability are fundamentally governed by local bonding environments and molecular topology rather than mere molecular scale.

Through the application of bootstrapped distance correlation, we established that the highest occupied molecular orbital (HOMO) energy exhibits a remarkable orthogonal isolation from the global spatial and thermodynamic state of the molecule. Unlike geometric properties that form tightly coupled clusters, the HOMO energy displays weak non-linear dependency even with other components of the frontier-response manifold, such as the lowest unoccupied molecular orbital (LUMO) energy and isotropic polarizability. This isolation indicates that the HOMO is primarily dictated by highly localized electronic physics, such as specific functional groups or lone pairs, operating independently of the broader molecular architecture.

Despite this individual isolation, exact Wasserstein transport computations (W_2^2) across the Murcko scaffold hierarchy reveal that the frontier-electronic manifold shifts as a coherent unit during structural transitions. The marginal distribution shifts for electronic properties are strikingly uniform and significantly smaller than those of thermodynamic null properties, highlighting the relative stability of the electronic state space under structural rewiring. Furthermore, mapping these transport costs to topological metrics identified opposing drivers of manifold deformation: aromaticity surprisingly anchors the electronic distribution close to the acyclic baseline, whereas spatial confinement and small-ring strain induce the most severe geometric deformations.

Finally, our variance analysis of the marginal transport costs mathematically refutes the prevailing chemical intuition of "gap anchoring." If structural transitions induced compensatory shifts in HOMO and LUMO levels to preserve a stable energy gap, the variance of their combined shifts would be minimized by negative covariance. Instead, the joint variance is nearly twice the sum of their individual variances. This proves that topological constraints in small organic molecules do not collaborate to maintain electronic stability; rather, they drive independent fluctuations that increase frontier gap variability. Ultimately, this optimal transport framework provides a rigorous, size-independent perspective on molecular property spaces, revealing that frontier orbitals are governed by distinct, localized structural sensitivities.