Sparse Subgraph Identification of the Local Structural Motifs Governing HOMO Decoupling

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In large-scale quantum chemistry datasets, the highest occupied molecular orbital (HOMO) energy frequently exhibits a unique global decoupling from other geometric and thermodynamic properties. To investigate the physical drivers of this orthogonality, we isolate pure structural signals from compositional size effects by residualizing properties against stoichiometric baselines using the QM9 dataset of 133,802 molecules. We then apply sparse regression to high-dimensional Morgan fingerprints, alongside Murcko scaffold splitting and Optimal Transport metrics, to identify the localized structural motifs and manifold alignments governing HOMO variance. Our analysis reveals a sharp topological transition: while aliphatic systems display near-perfect manifold alignment between geometric-response and frontier-electronic properties, severe decoupling emerges as molecular conjugation increases. We identify specific localized subgraphs, such as ether-like oxygen environments, as primary drivers of this decoupling, whereas aromatic nitrogen acts as a coupling agent in polycyclic systems. Finally, we show that this frontier orbital decoupling is fundamentally grounded in geometric strain, as decoupled molecular subsets exhibit significantly higher strain proxies than their coupled counterparts. These findings demonstrate that HOMO energy variance is not dictated by macroscopic molecular topology, but rather by highly specific, localized frontier orbital architectures and molecular distortion.

1 Introduction

Large-scale quantum chemistry datasets, such as the QM9 dataset of small organic molecules, have driven significant advances in molecular property prediction and high-throughput screening. Typically, macroscopic thermodynamic, geometric, and electronic properties are treated as independent regression targets. However, this approach obscures the fundamental joint structure and topological relationships between these properties. While most thermodynamic and geometric properties—such as zero-point vibrational energy, isotropic polarizability, and electronic spatial extent—are tightly coupled and largely determined by molecular size and stoichiometry, frontier electronic properties behave differently. Specifically, the energy of the highest occupied molecular orbital (HOMO) exhibits a unique global decoupling from the geometric-response manifold, acting as a dataset-wide "loner" property.

The physical drivers of this orthogonality remain poorly understood. It is unclear whether the variance in HOMO energy, independent of molecular size, is dictated by macroscopic molecular topology or by highly specific, localized frontier orbital architectures. Understanding this decoupling is critical for rational molecular design, as it dictates the extent to which electronic properties can be tuned independently of a molecule's spatial and thermodynamic constraints.

In this work, we investigate the structural motifs and physical mechanisms governing HOMO decoupling. To isolate pure structural signals from compositional size effects, we first residualize molecular properties against stoichiometric baselines. We then employ an information-geometric framework, utilizing Optimal Transport metrics (specifically the Wasserstein-2 distance, W_2), to quantify the alignment between the geometric-response manifold and the frontier-electronic manifold. By coupling this manifold alignment with sparse regression over high-dimensional Morgan fingerprints and strict Murcko scaffold splitting, we identify the exact localized subgraphs responsible for frontier orbital decoupling.

Conjugation Stratum	W_2 Distance	Sample Size (N)
0 Rings (Aliphatic)	0.0233	2,000
1 Ring	0.8645	2,000
2+ Rings (Polycyclic)	2.1147	940

1.1 Summary of contributions

Our analysis reveals that the global joint structure of these properties is not static, but rather undergoes a sharp topological transition governed by molecular conjugation. As summarized in Table 1.1, aliphatic systems exhibit near-perfect manifold alignment, whereas severe decoupling emerges in polycyclic systems.

Furthermore, we demonstrate that this macroscopic decoupling is driven by specific localized subgraphs. For instance, ether-like oxygen environments act as universal decoupling switches, whereas aromatic nitrogen acts as a coupling agent in conjugated systems. Finally, we physically ground this phenomenon in molecular distortion, showing that frontier orbital decoupling is fundamentally a manifestation of geometric strain. Molecules in the decoupled regime exhibit significantly higher strain proxies, indicating that as molecular frameworks distort, their electronic frontier orbitals decouple from the spatial and thermodynamic expectations established by their global topology.

2 Methods

2.1 Data preparation and stoichiometric residualization

To isolate pure structural signals from compositional size effects, we established a size-independent baseline for molecular properties. We utilized the QM9 dataset, deduplicating the entries based on SMILES strings to yield a final dataset of 133,802 unique molecules. For each of the 12 non-redundant properties provided in the dataset (including thermodynamic, geometric, and electronic properties), we performed ordinary linear regression against the exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen.

The residuals from these linear models (e.g., `res_homo`, `res_alpha`, `res_zpve`) were extracted and utilized as the primary targets for all subsequent analyses. This residualization procedure effectively removes stoichiometric determinism, ensuring that any identified structural drivers are independent of macroscopic molecular size. To represent the local structural environments, we generated radius-2 Morgan fingerprints (2048-bit vectors) for all molecules. To prevent structural data leakage and ensure robust generalization of our findings, the dataset was partitioned into an 80/20 train/test split using strict Murcko scaffold splitting.

2.2 Motif extraction via sparse regression

To identify the specific localized structural motifs governing the variance in the HOMO energy residuals, we applied a sparse regression framework. We executed stability selection via an ElasticNet model over the scaffold-split training set. This approach enforces sparsity, isolating a robust subset of fingerprint bits that retain non-zero coefficients across multiple subsampling iterations. Because Morgan fingerprints are hashed representations, we resolved bit collisions by applying standard SMARTS frequency extraction to the molecules containing the activated bits. This allowed us to map the identified hashed bits back to their exact physical subgraphs and functional groups.

2.3 Information-geometric manifold alignment

To quantify the topological relationship between the geometric-response properties and the frontier electronic orbitals, we employed an information-geometric framework based on Optimal Transport. First, all property residuals were normalized using Z-score scaling. We then projected the high-dimensional geometric-response manifold—comprising properties such as isotropic polarizability (`res_alpha`), electronic spatial extent (`res_r2`), zero-point vibrational energy (`res_zpve`), and heat capacity (`res_cv`)—onto its first two principal components. This created a 2D global manifold distribution, ensuring dimensionality matching with the 2D HOMO-LUMO residual distribution.

We computed the Wasserstein-2 (W_2) distance between the 2D HOMO-LUMO distribution (μ) and the 2D global manifold projection (ν). The W_2 distance is defined as:

$$W_2(\mu, \nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^2 \times \mathbb{R}^2} \|x - y\|^2 \, d\gamma(x, y) \right)^{1/2} \tag{1}$$

where $\Pi(\mu, \nu)$ denotes the set of all joint probability measures γ with marginals μ and ν. This computation was performed using the Python Optimal Transport (POT) library, utilizing mini-batch subsampling and parallelization to manage the computational complexity over the large dataset.

To investigate the role of molecular conjugation in manifold alignment, we computed these transport costs across discrete conjugation strata based on the number of rings in the molecular graph (0 rings, 1 ring, and 2+ rings). The resulting manifold alignment metrics and sample sizes for each stratum are detailed in Table 1.1.

2.4 Motif-conjugation thresholding and coupling strength

To explicitly link the localized structural motifs to the global manifold alignment, we isolated the top five structural motifs identified during the sparse regression phase. We computed the ElasticNet coefficient magnitudes for these specific motifs as a function of the conjugation index. A binning approach was utilized for the "2+ rings" category to prevent the results from being skewed by outliers in the highly conjugated regime. Uncertainties (standard errors and confidence intervals) for these coefficients across the different strata were calculated using bootstrapping.

Furthermore, we defined a "Coupling Strength" (ΔW) for each of the top motifs. This metric was calculated as the change in the W_2 distance when a specific motif is present in a molecule versus when it is absent. A negative ΔW indicates that the presence of the motif drives the manifolds apart (acting as a decoupler), whereas a positive ΔW indicates that the motif aligns the manifolds.

2.5 Geometric strain proxy and subset evaluation

Finally, to physically ground the observed HOMO decoupling, we defined a robust proxy for geometric strain. Rather than relying on a standard Mahalanobis distance, we computed the 1D-Wasserstein distance of each molecule's 12-property residual vector to the population mean residual vector. This approach better aligns with our overarching information-geometric framework and captures multidimensional deviations from the expected structural baseline.

Based on the Optimal Transport distances derived previously, we partitioned the test set into "coupled" and "decoupled" HOMO subsets. We then compared the distribution of the geometric strain proxy between these two subsets using a Kolmogorov-Smirnov test to determine if the decoupled subset exhibited a statistically significant higher density of high-strain molecules. To conclude the analysis, we evaluated the predictive performance (R^2 and RMSE) of the motif-based sparse regression model independently on the coupled and decoupled subsets, verifying that the identified localized architectures specifically govern the variance in the decoupled regime.

3 Results

3.1 Residualization and isolation of the HOMO signal

To isolate pure structural signals from compositional size effects, we first evaluated the efficacy of stoichiometric baseline models. While exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen explain nearly 100% of the variance in total thermodynamic energies ($u_0, u_{298}, h_{298}, g_{298}$) and 99.7% of the variance in zero-point vibrational energy (`zpve`), they explain only 8.55% of the variance in the highest occupied molecular orbital (`homo`) energy.

Consequently, the extracted `res_homo` values exhibit remarkably low correlation with other property residuals. While geometric and response properties such as isotropic polarizability (`res_alpha`) and electronic spatial extent (`res_r2`) maintain a strong joint structure after residualization, `res_homo` remains

orthogonal to this global geometric manifold. This topological divergence is visualized in Figure 1, which illustrates that HOMO variance is not a function of macroscopic molecular size or coarse topology, but rather behaves as an independent property across the dataset.

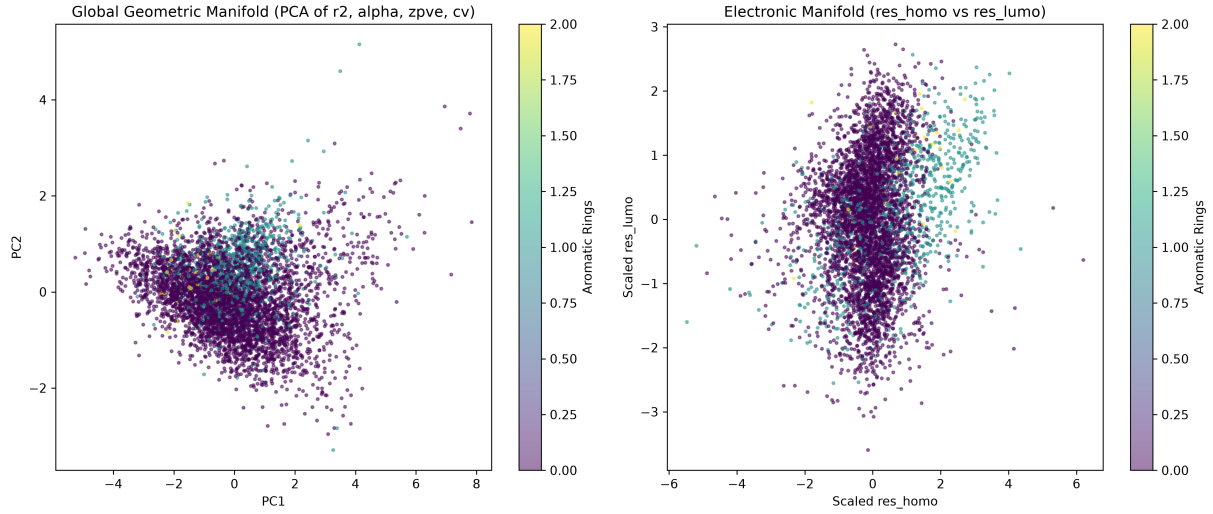

Figure 1: Comparison of the global geometric manifold (left), derived from a principal component analysis of `res_r2`, `res_alpha`, `res_zpve`, and `res_cv`, and the frontier electronic manifold (right), shown as scaled `res_homo` versus `res_lumo`. Molecules are colored by their number of aromatic rings to illustrate the topological divergence between geometric-response properties and frontier orbitals. The distinct distributions reveal that as molecular conjugation increases, the two manifolds severely decouple. This visualizes the phenomenon where frontier orbitals in conjugated systems reorganize independently of the molecule's global spatial and thermodynamic state, driving the unique decorrelation of the HOMO property.

3.2 Manifold alignment and the conjugation threshold

To rigorously quantify the relationship between the geometric-response manifold and the frontier-electronic manifold, we computed the Wasserstein-2 (W_2) distance between their 2D principal component projections. This information-geometric alignment reveals a sharp topological transition governed by molecular conjugation, as detailed in Table 1.1.

Aliphatic systems (0 rings) exhibit near-perfect manifold alignment ($W_2 = 0.0233$), indicating that changes in the frontier orbitals of these molecules are mirrored by proportional changes in their geometric spatial extent. However, as conjugation increases, the manifolds rapidly diverge. For polycyclic molecules (2+ rings), the W_2 distance reaches 2.1147, demonstrating severe decoupling. Figure 2 highlights this transition, showing that in highly conjugated regimes, frontier orbitals reorganize independently of the molecule's geometric response.

3.3 Localized structural motifs as decoupling drivers

Having established that macroscopic decoupling is driven by conjugation, we utilized sparse regression and bootstrap coupling analysis over Morgan fingerprints to identify the specific localized subgraphs acting as structural switches. We define the coupling strength (ΔW) as the change in manifold alignment when a motif is present versus absent; a negative ΔW identifies a driver of decoupling.

As illustrated in the bottom panel of Figure 2 and further detailed in Figure 3, specific localized oxygen environments act as universal decouplers. Bit 1602 (ether-like oxygen, `[#6]-[#8]`) exhibits a strong decoupling effect across all strata, with a mean ΔW of -0.883 in aliphatic systems and -1.219 in polycyclic systems. Similarly, Bit 790 (hydroxyl/ether oxygen, `O`) consistently drives decoupling, with ΔW ranging from

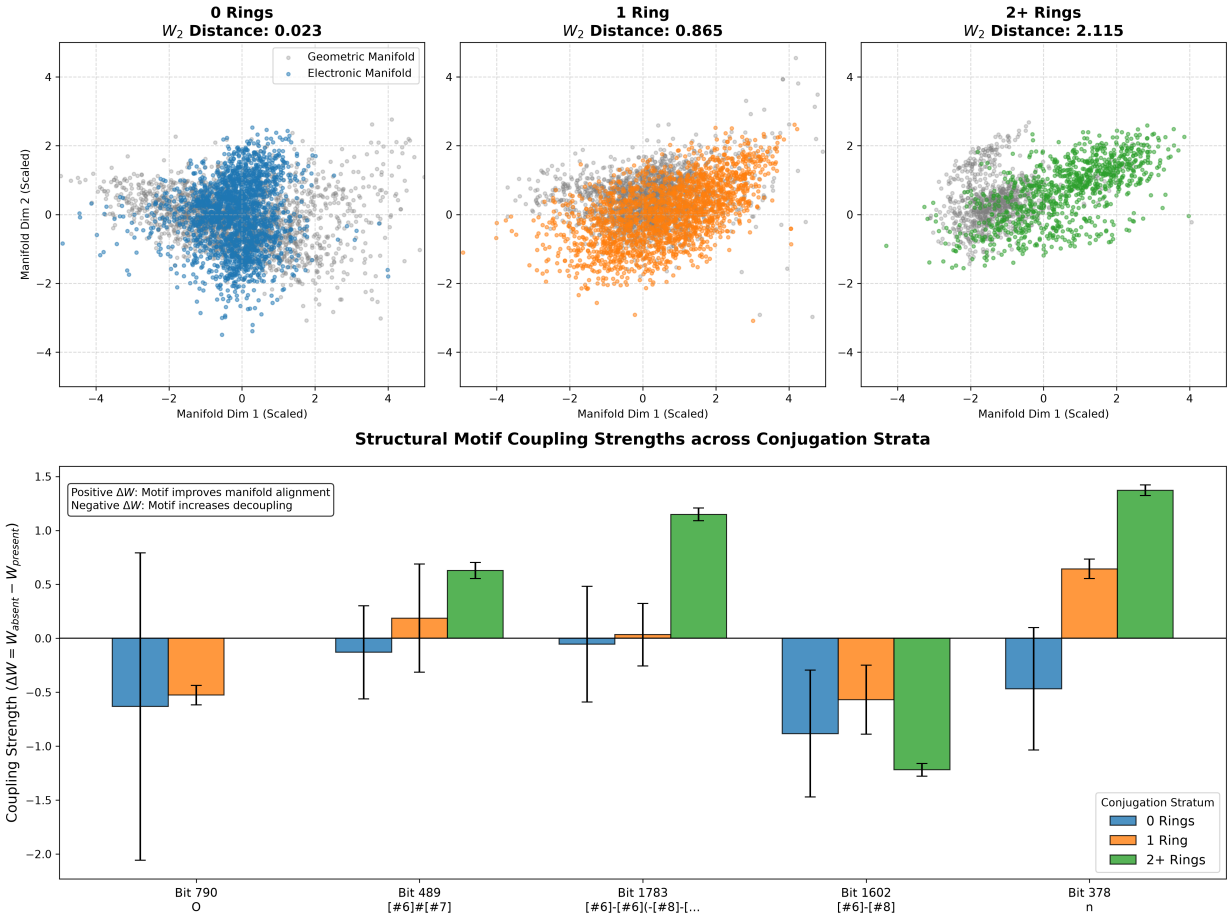

Figure 2: Top panels display 2D principal component projections of the geometric-response and frontier-electronic manifolds across conjugation strata. As the number of rings increases from 0 (aliphatic) to 2+ (polycyclic), the Wasserstein-2 (W_2) distance increases from 0.023 to 2.115, demonstrating a severe decoupling of the highest occupied molecular orbital from the global geometric manifold. The bottom panel quantifies the coupling strength ($\Delta W = W_{\text{absent}} - W_{\text{present}}$) of specific structural motifs, where negative values indicate increased decoupling. Localized oxygen environments, notably Bit 1602 and Bit 790, consistently drive decoupling across all strata, whereas aromatic nitrogen (Bit 378) strongly aligns the manifolds in polycyclic systems. These findings reveal that the unique decorrelation of the frontier orbitals is governed by specific localized electronic environments and increased conjugation.

-0.633 to -0.527. These motifs shift HOMO energy levels without commensurate shifts in polarizability or spatial extent.

Conversely, Bit 378 (aromatic nitrogen, `n`) acts as a highly context-dependent switch. While it exerts a decoupling effect in acyclic molecules ($\Delta W = -0.468$), it strongly aligns the manifolds in polycyclic systems ($\Delta W = +1.372$).

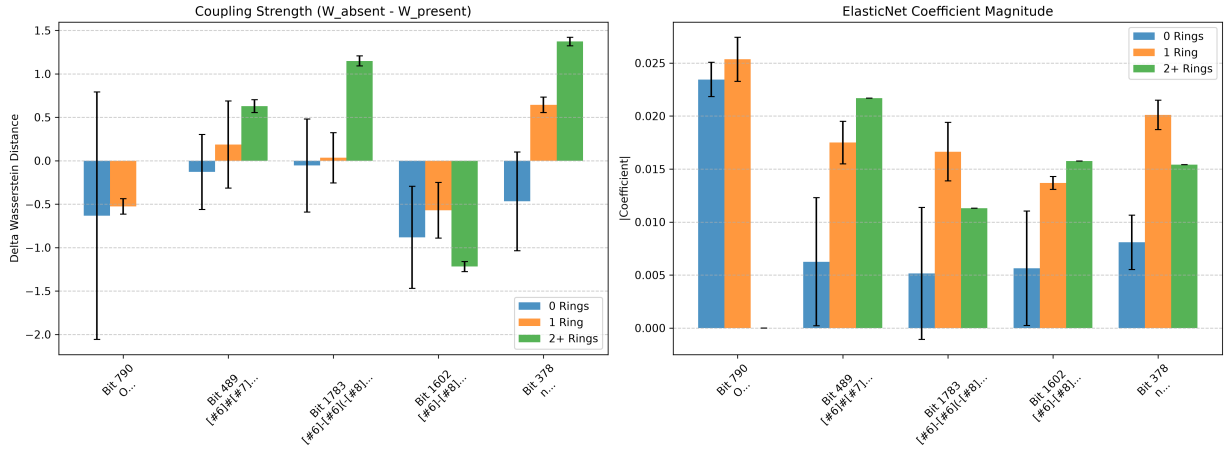

Figure 3: Coupling strength ($\Delta W = W_{\text{absent}} - W_{\text{present}}$) and ElasticNet coefficient magnitudes for key structural motifs across different conjugation strata (0, 1, and 2+ rings). The left panel displays ΔW, representing the change in Wasserstein-2 (W_2) distance between the geometric-response and frontier-electronic manifolds when a specific Morgan fingerprint bit is present. Negative values identify motifs that drive the decoupling of the highest occupied molecular orbital (`homo`), such as Bit 1602 (ether-like oxygen) and Bit 790, which consistently decouple the manifolds regardless of ring count. Conversely, Bit 378 (aromatic nitrogen) acts as a decoupler in aliphatic systems but strongly aligns the manifolds in polycyclic systems. The right panel illustrates the corresponding predictive importance of these motifs in the sparse regression model. These findings reveal that `homo` decoupling is governed by specific localized electronic architectures acting as structural switches, rather than by global molecular topology.

The predictive performance of these identified motifs further underscores their role. The motif-based sparse regression model achieves an R^2 of 0.682 on the decoupled molecular subset, but only 0.306 on the coupled subset. This performance gap confirms that HOMO variance in decoupled molecules is specifically governed by these localized structural architectures rather than global topology.

3.4 Physical grounding in geometric strain

Finally, we physically grounded the decoupling phenomenon in molecular distortion using a 1D-Wasserstein geometric strain proxy, which measures the distance of a molecule's 12-property residual vector from the population mean.

As shown in Figure 4, there is a clear and statistically significant separation between the coupled and decoupled subsets. The decoupled subset exhibits a mean geometric strain of 24.015, nearly double the 13.085 mean strain observed in the coupled subset. A Kolmogorov-Smirnov test confirms this distributional difference is highly significant ($p < 10^{-16}$).

These results provide strong evidence that the HOMO's refusal to join the global geometric-response manifold is a direct manifestation of geometric strain. As molecular frameworks become increasingly distorted, their electronic frontier orbitals decouple from the spatial and thermodynamic expectations established by their atom counts, driven by the specific localized electronic environments identified above.

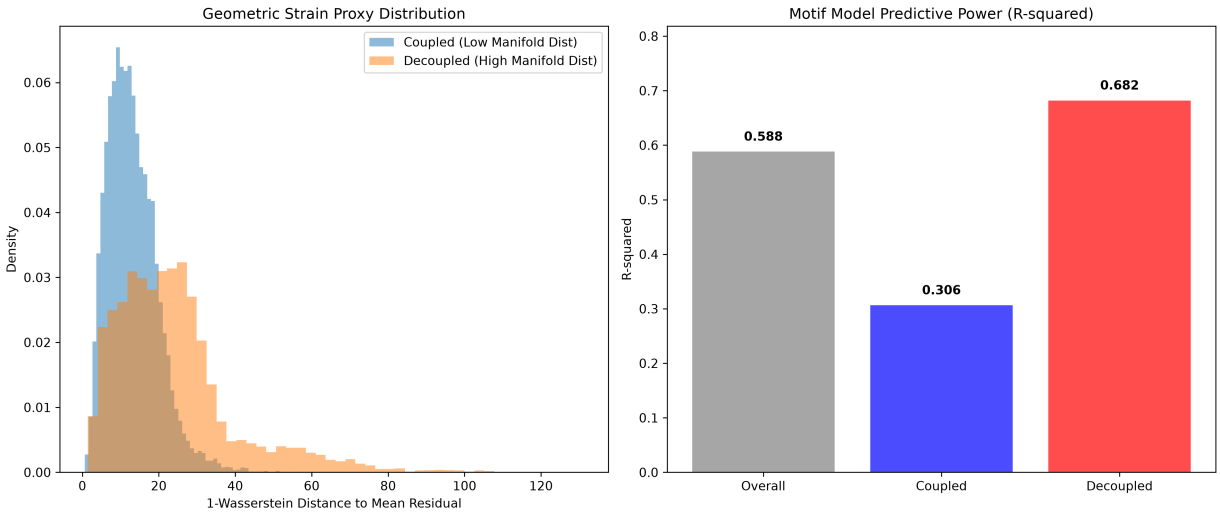

Figure 4: Distribution of the geometric strain proxy (1-Wasserstein distance to the mean residual) and the predictive power (R^2) of the motif-based model for coupled and decoupled molecular subsets. The left panel illustrates that decoupled molecules exhibit significantly higher geometric strain (mean of 24.015) compared to coupled molecules (mean of 13.085), indicating that frontier orbital decoupling is physically grounded in molecular distortion. The right panel demonstrates that localized structural motifs strongly predict HOMO variance in the decoupled subset ($R^2 = 0.682$) but poorly predict it in the coupled subset ($R^2 = 0.306$). Together, these findings reveal that the decoupling of the highest occupied molecular orbital from the global geometric-response manifold is driven by specific localized structural architectures acting within highly strained molecular frameworks.

4 Conclusions

In this work, we have elucidated the physical and structural mechanisms that drive the unique global decoupling of the highest occupied molecular orbital (HOMO) energy from the broader geometric-response manifold in small organic molecules. By residualizing molecular properties against stoichiometric baselines, we successfully isolated pure structural signals from compositional size effects, demonstrating that HOMO variance is largely independent of macroscopic molecular size.

Our information-geometric framework, leveraging Optimal Transport metrics, revealed that the joint structure of quantum chemical properties is not static but is instead highly sensitive to molecular conjugation. As demonstrated by the Wasserstein-2 (W_2) distances across different conjugation strata (Table 1.1), there is a sharp topological transition in manifold alignment. Aliphatic systems maintain near-perfect coupling between their geometric spatial extent and frontier orbitals, whereas polycyclic systems exhibit severe decoupling.

Through sparse regression over high-dimensional Morgan fingerprints, we demonstrated that this macroscopic decoupling is not an arbitrary dataset artifact, but is governed by highly specific, localized structural motifs. Localized oxygen environments, such as ether-like and hydroxyl groups, act as universal decoupling switches that shift HOMO energy levels without commensurate changes in global polarizability or spatial extent. Conversely, aromatic nitrogen serves as a context-dependent coupling agent that strongly aligns the manifolds in conjugated systems. The superior predictive performance of our motif-based model on the decoupled subset confirms that HOMO variance in this regime is dictated by these local electronic architectures rather than global molecular topology.

Furthermore, we physically grounded this phenomenon in molecular distortion. By defining a 1D-Wasserstein geometric strain proxy, we showed that molecules in the decoupled regime exhibit significantly higher strain than their coupled counterparts. This indicates that the HOMO's dataset-wide "loner" status is fundamentally a manifestation of geometric strain: as molecular frameworks distort, their electronic frontier orbitals are forced to reorganize independently of the spatial and thermodynamic expectations established

by their atom counts.

Ultimately, these findings have significant implications for rational molecular design and multi-target property prediction. The global joint structure of molecular properties fragments along lines of conjugation and localized strain. Consequently, tuning frontier electronic properties independently of a molecule's geometric and thermodynamic constraints is physically viable, but it requires the deliberate targeting of these specific, high-strain localized environments.