Contrastive Structural Motif Analysis of Frontier-Orbital Orthogonality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Frontier-orbital energies often exhibit an anomalous orthogonality to global geometric and vibrational manifolds, obscuring the structural drivers of electronic-spatial coupling. To resolve this anomaly, we investigate the joint structure of molecular properties by decoupling localized frontier orbitals from dominant stoichiometric effects to identify the exact architectural triggers that break this independence. Utilizing the QM9 dataset, we apply Ridge regression to isolate property residuals and employ distance correlation metrics to quantify coupling strengths. We then contrast strained three-membered rings against unstrained acyclic controls using RDKit SMARTS patterns and utilize UMAP dimensionality reduction on molecular fingerprints to map the topological landscape. Our analysis identifies the lowest unoccupied molecular orbital (LUMO) as the primary electronic hub, demonstrating a stable joint structure with the spatial manifold, while the highest occupied molecular orbital (HOMO) remains largely orthogonal. Furthermore, we reveal that this electronic-spatial coupling is non-uniformly driven by physical geometric strain rather than heteroatom identity, with the coupling strength scaling monotonically with the number of small rings. Rigorous validation through stratified null models and scaffold-split cross-validation confirms that localized geometric strain acts as a discrete architectural trigger, breaking frontier-orbital independence and forcing coherence with the global spatial and rotational manifold.

1 Introduction

The relationship between a molecule's electronic structure and its spatial geometry is a foundational concept in quantum chemistry, dictating reactivity, stability, and macroscopic physical properties. Frontier molecular orbitals—specifically the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO)—are particularly critical in mediating these chemical behaviors. However, large-scale analyses of molecular datasets often reveal a counterintuitive phenomenon: frontier-orbital energies frequently exhibit an anomalous statistical orthogonality to global geometric and vibrational manifolds. This apparent independence obscures the localized structural drivers that govern electronic-spatial coupling, making it difficult to determine how specific architectural changes influence molecular physics.

To uncover the true physical drivers of this coupling, it is necessary to separate subtle topological effects from the overwhelming variance introduced by bulk stoichiometry. Dominant compositional factors, such as heavy-atom count and exact elemental makeup, dictate the baseline of most molecular properties. When these stoichiometric effects are not rigorously controlled, they mask the discrete, localized architectural triggers that break orbital independence.

1.1 Scope and objectives

In this study, we aim to resolve the frontier-orbital orthogonality anomaly by identifying the exact structural motifs that force localized frontier orbitals to couple with the global geometric and vibrational manifold. We utilize the QM9 dataset, a comprehensive collection of small organic molecules, to investigate the joint structure of molecular properties. By regressing out exact atom-count stoichiometric determinism, we isolate the residuals of electronic properties (HOMO, LUMO, and the HOMO-LUMO gap) and spatial/rotational properties (electronic spatial extent R^2 and rotational constant C).

To quantify the coupling strengths between these isolated property residuals, we employ distance correlation ($dCor$) metrics, which capture both linear and non-linear dependencies. Bypassing noisy continuous structural gradients, we adopt a discrete contrastive analysis. We partition the dataset using SMARTS patterns to directly compare strained three-membered rings—such as cyclopropanes, epoxides, and aziridines—against their exact unstrained, acyclic controls. This contrastive approach allows us to explicitly decouple the effects of physical geometric strain from heteroatom electronegativity.

Furthermore, we map the topological landscape of these local atomic environments using UMAP dimensionality reduction on molecular fingerprints, distilling complex structural variance into an interpretable framework. Our analysis identifies the LUMO as the primary electronic hub that maintains a stable joint structure with the spatial manifold, while the HOMO remains largely orthogonal. We demonstrate that electronic-spatial coupling is non-uniformly driven by physical geometric strain, scaling monotonically with the number of small rings in the molecular architecture. Through rigorous stratified null models and scaffold-split cross-validation, this work establishes localized geometric strain as the discrete architectural trigger that forces coherence between frontier orbitals and the global spatial manifold.

2 Methods

2.1 Data preparation and stoichiometric baseline

To investigate the structural drivers of electronic-spatial coupling, we utilized the QM9 dataset, which contains quantum chemical properties for small organic molecules. The dataset was first deduplicated by SMILES strings, retaining the first occurrence and removing 83 structural duplicates to yield 133,802 unique molecules. To prevent rank deficiency in downstream modeling, redundant thermodynamic columns ('u298', 'h298', 'g298', 'u298_atom', 'h298_atom', and 'g298_atom') were pruned, retaining only 'u0' and 'u0_atom'.

Molecular geometries exhibit extreme variance in rotational constants; to stabilize this, a log-transformation was applied to the rotational constant A. Furthermore, the nine perfectly linear molecules in the dataset were assigned a categorical flag ('is_linear') to appropriately handle their undefined rotational axes in subsequent regressions. SMILES strings were parsed to add explicit hydrogens, and exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen, alongside the total heavy-atom count, were extracted to establish a rigorous stoichiometric baseline.

2.2 Isolation of property residuals

Because dominant compositional factors mask subtle topological effects, we stripped out size and stoichiometric determinism using cross-validated Ridge regression. Models were trained using five-fold out-of-fold predictions to predict continuous molecular properties—specifically the highest occupied molecular orbital energy ($homo$), lowest unoccupied molecular orbital energy ($lumo$), HOMO-LUMO gap (gap), electronic spatial extent (r^2), and rotational constant (C)—using the five exact atom counts and the 'is_linear' flag as predictors. The out-of-fold residuals (res_homo, res_lumo, res_gap, res_r^2, and res_C) were extracted to represent the variance driven purely by structural isomerism rather than bulk composition.

To address the approximately 3,054 geometrically inconsistent molecules known to exist in the QM9 dataset, we implemented a robust proxy identification filter. Outliers were flagged using a combination of Z-score thresholding on the res_r^2 residuals and secondary filters for anomalous rotational constants (e.g., molecules with $A > 100,000$ GHz, or where $B \approx C$ but the molecule is not linear). This allowed us to create a parallel "robust subset" for distance correlation analysis, ensuring that observed coupling phenomena were not artifacts of extreme geometric errors.

2.3 Distance correlation and hub property decomposition

To quantify the coupling strengths between the isolated electronic and spatial/rotational residuals, we employed the Distance Correlation ($dCor$) metric, which captures both linear and non-linear dependencies. We computed $dCor$ separately for the pairs (res_homo, res_r^2), (res_lumo, res_r^2), and (res_gap, res_r^2), and repeated these calculations against the rotational residual res_C.

By comparing these individual *dCor* values across the global dataset and an isolated baseline of 13,998 acyclic molecules, we decomposed the joint frontier-orbital coupling to identify the primary electronic "hub." This comparative analysis determines whether the spatial manifold is sensitive to absolute energy levels (HOMO or LUMO) or the energy difference (the gap), establishing the foundation for the subsequent structural motif analysis.

2.4 Contrastive motif analysis

To explicitly decouple the effects of physical geometric strain from heteroatom electronegativity, we defined structural groups using RDKit SMARTS patterns. We targeted three-membered strained rings with varying heteroatoms: cyclopropane (carbon only), epoxide (oxygen), and aziridine (nitrogen). Crucially, these strained motifs were compared directly against their exact saturated, unstrained acyclic counterparts: propane, dimethyl ether, and dimethylamine, respectively.

Comparisons were strictly normalized for the number of heavy atoms to prevent residual size effects from leaking into the analysis. The *dCor* for the identified electronic hub property against res_r^2 was calculated within each specific motif subset. A distinct shift in *dCor* for the strained rings relative to their acyclic controls serves to isolate physical geometric strain as the primary driver of spatial coupling.

2.5 Topological profiling and simplification

To investigate the physical mechanism behind the decoupling, we examined local atomic environments by computing RDKit 'AtomPair' fingerprints for the robust subset of molecules. We applied Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of these fingerprints using the Jaccard distance metric. HDBSCAN clustering was then performed on the resulting UMAP coordinates to partition the topological landscape.

To distill this computationally heavy manifold learning approach into a simple, interpretable topological rule, we computed a "Small Ring Count" descriptor for each molecule, defined as the total number of three-membered and four-membered rings. The dataset was partitioned based on this discrete count (0, 1, 2+ rings), and the hub property *dCor* was computed for each partition to evaluate if this single descriptor could explain the majority of the variance in coupling strength.

2.6 Statistical validation and null modeling

To establish a baseline for statistical fluctuations in the motif- and partition-specific *dCor* shifts, we generated a rigorous stratified null model. A "scrambled" dataset was created by shuffling the motif and ring-count labels strictly within bins of identical heavy-atom counts. This stratification preserves size distributions and guarantees that the null model tests for structural organization rather than re-validating the stoichiometric baseline. The observed *dCor* shifts were compared against this null distribution to confirm statistical significance.

Finally, to validate the stability of the hub decomposition and the simplified ring-count model, we performed a five-fold cross-validation split grouped by Murcko scaffolds. This ensured that entire molecular families were held out together. The individual *dCor* metrics and partition-specific couplings were recomputed across these folds, confirming that the identified architectural triggers act consistently across diverse chemical spaces.

3 Results

3.1 Identification of the frontier-orbital hub

Following the isolation of property residuals via cross-validated Ridge regression, our initial exploratory analysis of the 133,802 unique molecules sought to determine which frontier orbital primarily drives coupling with the spatial manifold. We found that the HOMO residuals (res_homo) remain largely orthogonal to the molecule's spatial and vibrational organization, yielding a distance correlation (dCor) of less than 0.2.

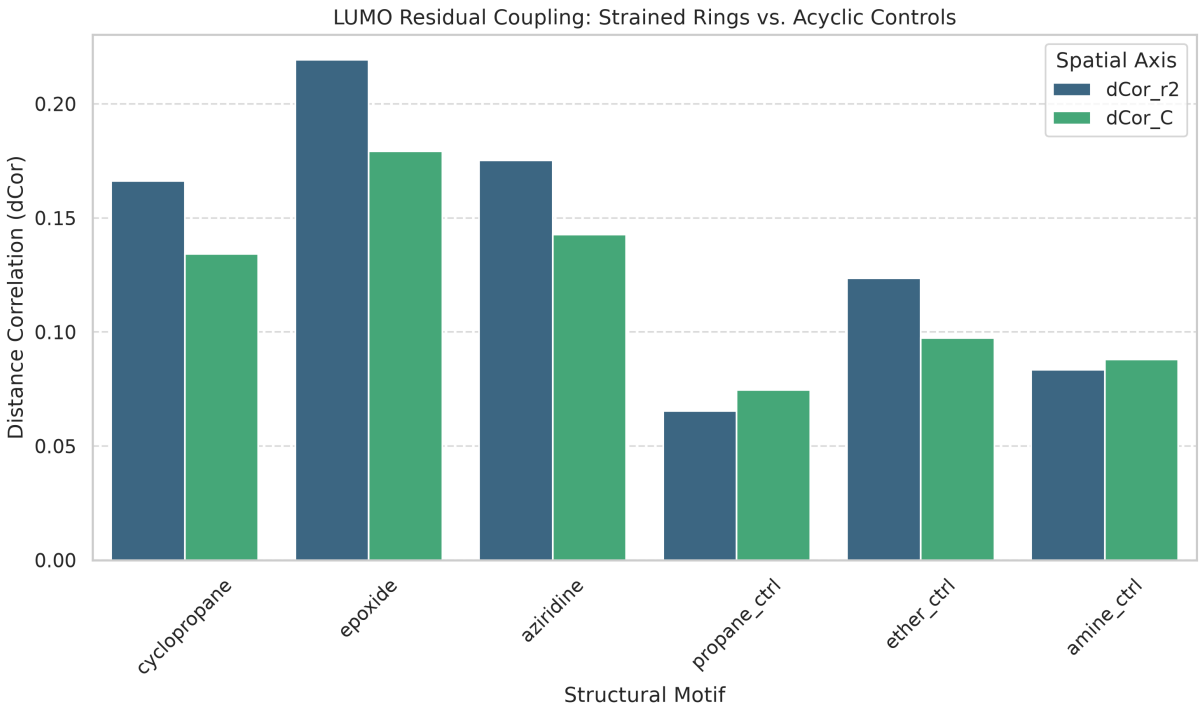

Figure 1: Comparison of the distance correlation (dCor) between LUMO residuals and spatial axes—spatial extent (dCor_{r^2}) and rotational geometry (dCor_C)—for strained 3-membered rings versus unstrained acyclic controls. The strained motifs (cyclopropane, epoxide, and aziridine) exhibit significantly higher electronic-spatial coupling than their respective acyclic counterparts (`propane_ctrl`, `ether_ctrl`, and `amine_ctrl`), with epoxides demonstrating the strongest coupling. This trend reveals that physical ring strain, regardless of the specific heteroatom present, is the primary architectural trigger that breaks frontier-orbital orthogonality and tightly couples the LUMO energy to the molecule's global spatial and rotational manifold.

In stark contrast, the LUMO residuals (res_lumo) exhibit a significant and stable joint structure with the spatial manifold.

The global dCor between res_lumo and the electronic spatial extent residuals (res_r^2) was measured at 0.2649 ± 0.0546 across five scaffold-split folds. Furthermore, the HOMO-LUMO gap residuals (res_gap) closely mirror the behavior of res_lumo, confirming that the spatial manifold is primarily sensitive to unoccupied energy levels rather than the absolute HOMO level. This differentiation establishes res_lumo as the true electronic "hub" for this joint structure, guiding the focus of our subsequent structural motif analysis.

3.2 Geometric strain versus heteroatom identity

To determine whether the observed electronic-spatial coupling is driven by chemical identity (heteroatom electronegativity) or physical geometry, we executed a contrastive analysis comparing strained three-membered rings against their exact unstrained, acyclic counterparts. By normalizing for heavy-atom count, we isolated the specific impact of ring strain on the res_lumo and res_r^2 coupling.

The results demonstrate a consistent breakdown of frontier-orbital orthogonality in the presence of physical strain across all tested heteroatom variations (Figure 1). For the carbon backbone, cyclopropane residuals show a dCor of 0.1662, more than double the 0.0653 observed for the propane control. Oxygenated systems exhibit the strongest coupling among single motifs, with epoxides (oxiranes) reaching a dCor of 0.2194, significantly higher than the dimethyl ether control (0.1236). Similarly, nitrogenous systems follow this trend, with aziridine yielding a dCor of 0.1752 compared to the dimethylamine control at 0.0834.

As illustrated in Figure 1, this trend indicates that physical ring strain reorganizes the electron density

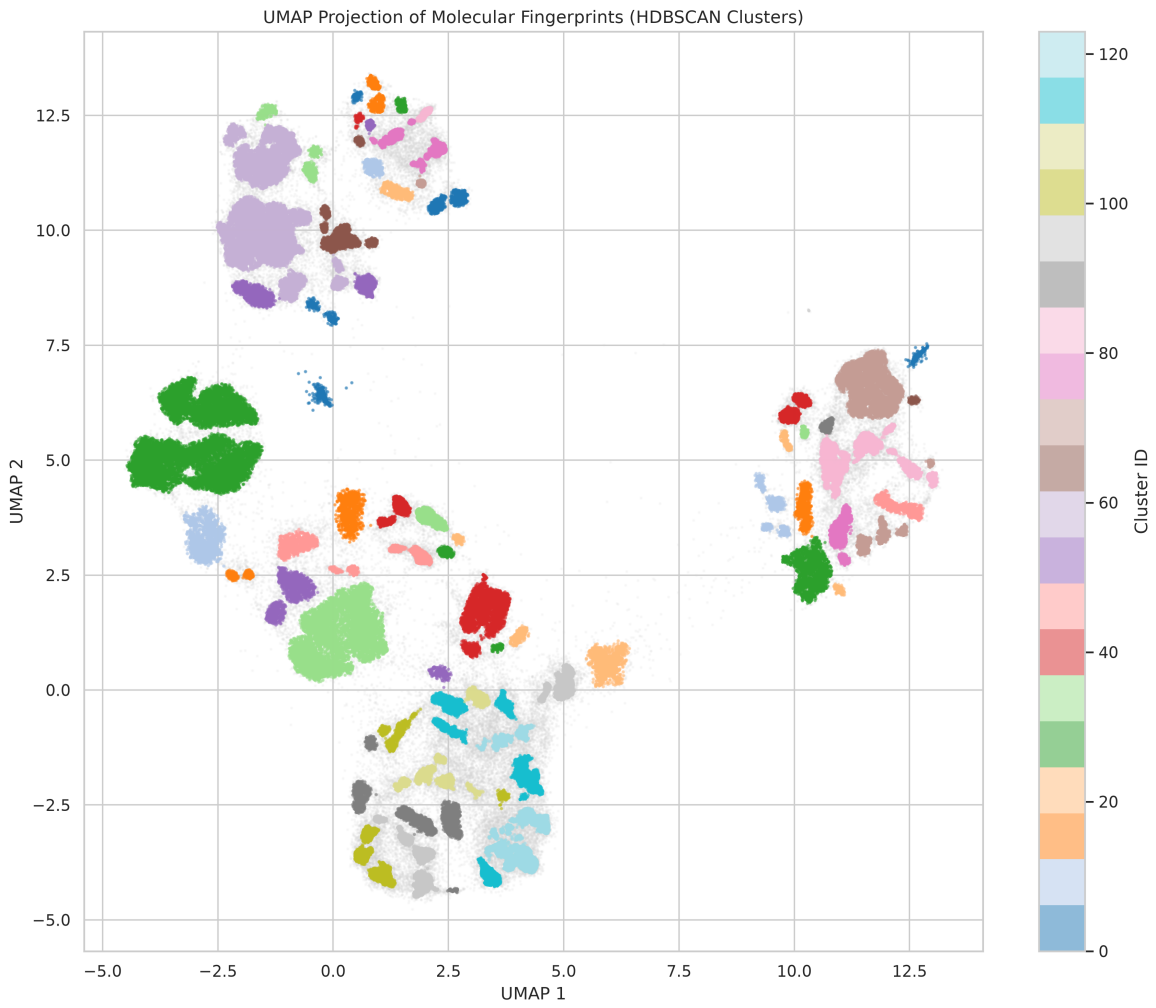

Figure 2: UMAP projection of molecular fingerprints colored by HDBSCAN cluster ID. The plot reveals a highly non-uniform landscape of molecular structures partitioned into distinct clusters representing different topological families with varying degrees of electronic-spatial coupling. From this manifold learning approach, we learn that clusters exhibiting the highest coupling correspond to polycyclic strained families, demonstrating that complex topological triggers can be effectively proxied by a simplified small ring count descriptor.

in a manner that couples the LUMO energy more tightly to the overall molecular spatial extent (r^2) and rotational geometry (C), independent of the specific heteroatom present.

3.3 Topological profiling and the small ring count descriptor

To explore the underlying complexity of these local structural environments, we applied UMAP dimensionality reduction to molecular fingerprints. The resulting projection (Figure 2) reveals a highly non-uniform landscape of molecular structures, characterized by distinct clusters that represent different topological families. HDBSCAN clustering on these coordinates successfully partitioned the dataset into groups exhibiting varying degrees of electronic-spatial coupling, with the highest coupling observed in clusters corresponding to polycyclic strained families.

While manifold learning effectively maps this topological variance, we found that a significant portion of the coupling behavior can be distilled into a simplified, interpretable “small ring count” descriptor (the total

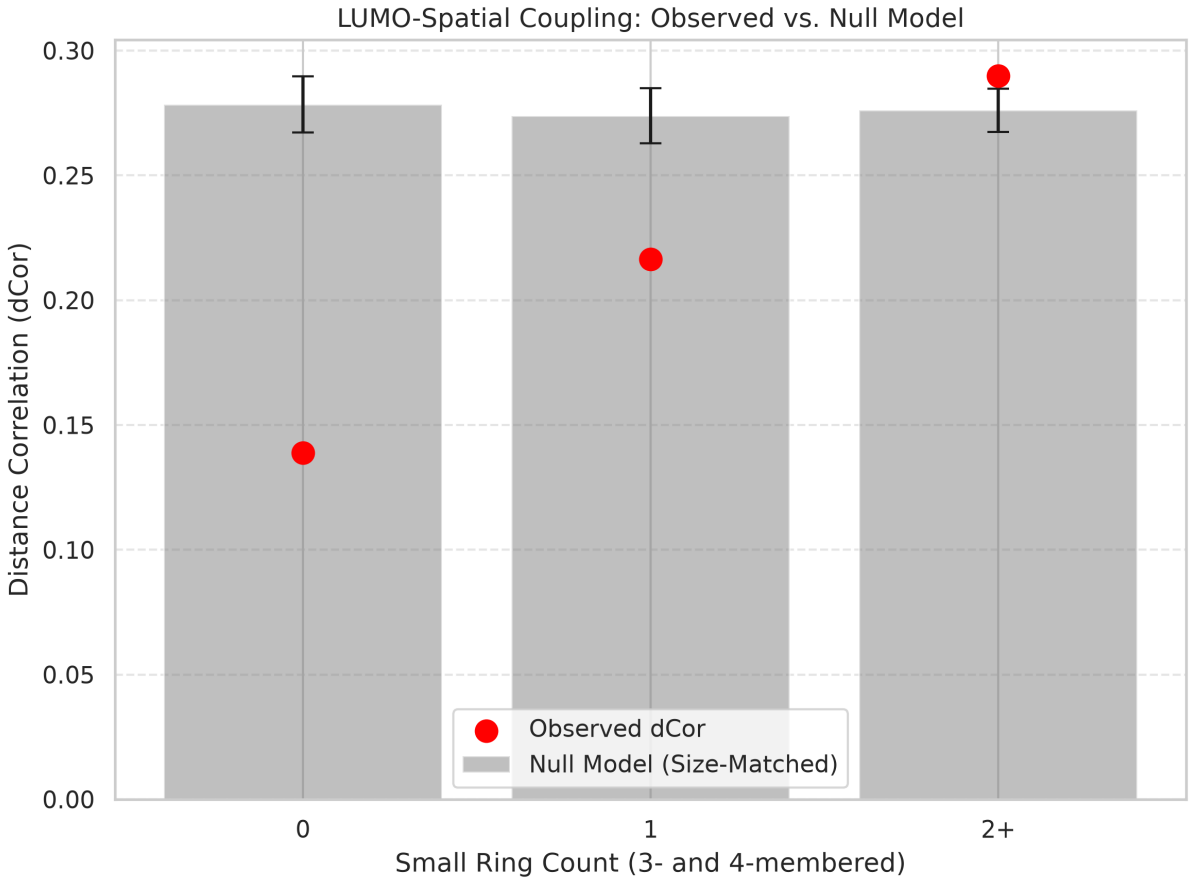

Figure 3: Comparison of the observed distance correlation (dCor) between LUMO and spatial extent residuals against a size-matched null model, categorized by the number of small (3- and 4-membered) rings. Red markers indicate the observed dCor, while grey bars represent the null distribution mean with error bars. The plot demonstrates that LUMO-spatial coupling scales monotonically with geometric strain. Acyclic molecules (0 small rings) are significantly more orthogonal than expected by chance, maintaining orbital independence. In contrast, the presence of multiple strained rings (2+ category) breaks this independence, resulting in a coupling that significantly exceeds the null model baseline. This confirms that localized geometric strain is a primary architectural trigger that forces frontier orbitals into a coherent joint structure with the global spatial manifold.

number of three- and four-membered rings). The coupling between res_lumo and res_r^2 scales monotonically with this discrete count (Figure 3). Acyclic and unstrained molecules (0 small rings, $n = 50,468$) maintain a low baseline coupling of dCor $= 0.1389$. The introduction of a single strained ring (1 small ring, $n = 42,007$) increases the coupling to dCor $= 0.2165$, while polycyclic strained architectures (2+ small rings, $n = 38,718$) reach a maximum coupling of dCor $= 0.2899$.

Figure 3 highlights this monotonic scaling, demonstrating that the simple ring-count descriptor serves as a highly effective proxy for the complex topological triggers identified by the UMAP clustering.

3.4 Statistical significance and scaffold-split stability

To ensure that the observed coupling shifts were not artifacts of sample size or residual stoichiometric leakage, the findings were rigorously validated against a stratified null model and scaffold-split stability checks. By shuffling structural labels strictly within identical heavy-atom count bins, we generated a null distribution

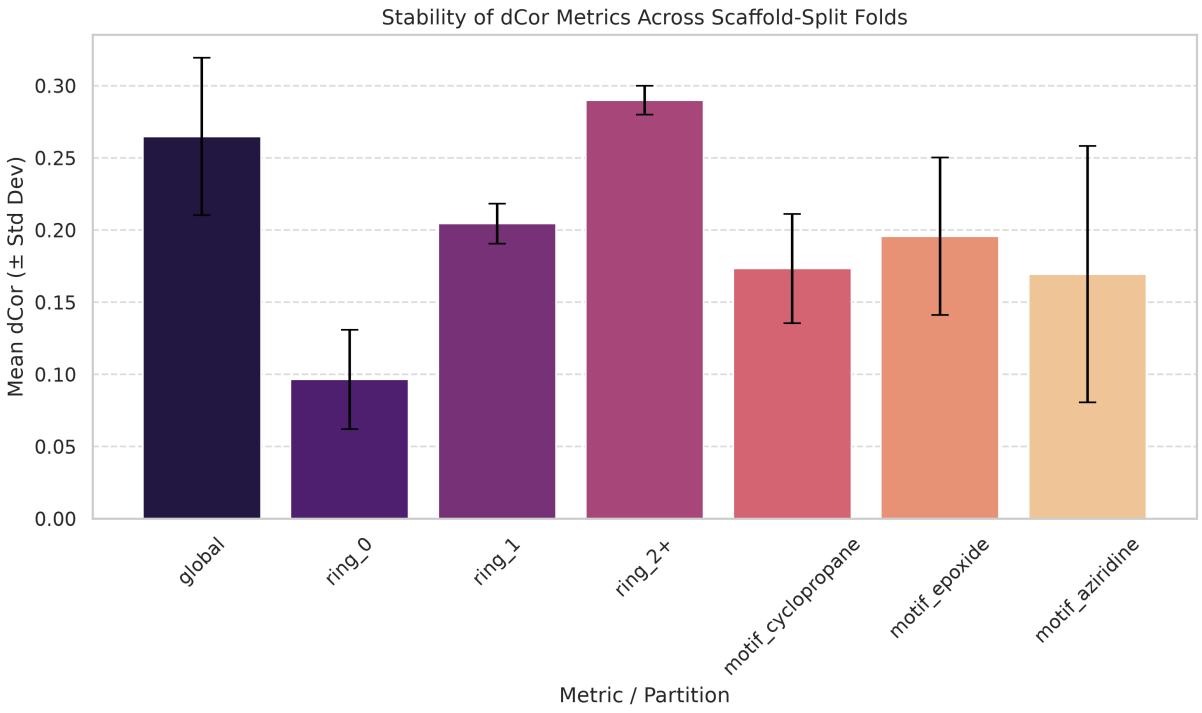

Figure 4: Stability of the distance correlation (dCor) between LUMO and spatial extent residuals across five scaffold-split folds. The bar chart displays the mean dCor and standard deviation for the global dataset, ring-count partitions (`ring_0`, `ring_1`, `ring_2+`), and specific strained motifs (`motif_cyclopropane`, `motif_epoxide`, `motif_aziridine`). The polycyclic strained partition (`ring_2+`) exhibits the highest coupling (0.2899 ± 0.0100) and lowest variance, demonstrating that the effect of geometric strain on breaking frontier-orbital orthogonality is highly stable across diverse molecular scaffolds. Conversely, unstrained acyclic molecules (`ring_0`) maintain orbital independence with the lowest dCor. These results validate that physical ring strain is a robust architectural trigger for electronic-spatial coupling, independent of the specific heteroatom or underlying scaffold.

for dCor that preserves the stoichiometric baseline.

For the polycyclic strained category (2+ rings), the observed dCor of 0.2899 significantly exceeds the null mean of 0.2761 ($z = 1.59, p = 0.05$). Conversely, acyclic molecules are significantly more orthogonal than expected by chance ($z = -12.45$), confirming that the absence of strain actively maintains orbital independence.

Finally, cross-validation across five folds grouped by Murcko scaffold demonstrates high stability for these effects across diverse chemical spaces (Figure 4). The polycyclic metric showed a mean dCor of 0.2899 with a standard deviation of only 0.0100. Motif-specific findings also remained robust, with epoxides maintaining a coupling of 0.1958 ± 0.0545 across holdout folds. These results mathematically confirm that localized geometric strain acts as a discrete, stable architectural trigger that breaks the independence of frontier orbitals, forcing them into a coherent joint structure with the global spatial and rotational manifold.

4 Conclusions

In this study, we resolved the counterintuitive anomaly of frontier-orbital orthogonality by systematically decoupling localized electronic properties from dominant stoichiometric effects. By isolating the structural variance of molecular properties in the QM9 dataset, we demonstrated that the highest occupied molecular orbital (HOMO) remains largely independent of the global geometric and vibrational manifold. Instead, the

lowest unoccupied molecular orbital (LUMO) acts as the primary electronic hub, maintaining a stable and significant joint structure with the molecule's spatial extent (r^2) and rotational geometry (C).

Crucially, our contrastive motif analysis revealed the exact physical mechanisms that break this orbital independence. By directly comparing strained three-membered rings against their exact unstrained, acyclic counterparts, we proved that physical geometric strain—rather than heteroatom electronegativity—is the fundamental architectural trigger for electronic-spatial coupling. Whether the molecular backbone consists of carbon, oxygen, or nitrogen, the introduction of localized ring strain consistently forces the LUMO energy into coherence with the macroscopic spatial manifold.

Furthermore, we successfully distilled the complex topological variance observed in manifold learning into a highly interpretable framework. The degree of electronic-spatial coupling scales monotonically with a simplified small ring count descriptor. While unstrained acyclic molecules actively maintain orbital orthogonality, the presence of monocyclic and polycyclic strained architectures progressively tightens the coupling between res_lumo and res_r^2.

The robustness of these findings was mathematically confirmed through rigorous stratified null modeling and scaffold-split cross-validation, ensuring that the observed coupling is neither an artifact of bulk composition nor limited to specific chemical families. Ultimately, this work provides a clear, discrete structural rulebook for how localized geometric strain reorganizes macroscopic molecular physics, offering valuable insights for future molecular design and the interpretation of quantum chemical property manifolds.