Data-Driven Manifold Embedding to Isolate the Structural Re-Coupling of Frontier Orbitals

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

While the energy of the Highest Occupied Molecular Orbital (HOMO) is generally decoupled from macroscopic thermodynamic and geometric properties, the specific structural conditions that force these frontier orbital energies to re-couple with the global spatial manifold remain poorly understood. To identify these conditions, we isolate purely structural variance within the 133,802-molecule QM9 dataset by extracting stoichiometric-independent residuals using gradient-boosted regression. By filtering geometric artifacts with an Isolation Forest and projecting the multidimensional residual space into a low-dimensional UMAP manifold, we map local distance correlations between the HOMO energy and an orthogonalized spatial axis. This unsupervised approach reveals a distinct subpopulation of 23,392 molecules exhibiting anomalous electronic-spatial coupling. Rigorous validation through variance normalization and partial correlation confirms that this re-coupling represents a genuine functional shift rather than a statistical artifact. Furthermore, permutation feature importance derived from a scaffold-split classifier demonstrates that this phenomenon is primarily driven by cyclic complexity, specifically high total ring counts and strained multi-ring architectures. Ultimately, this data-driven framework identifies the exact molecular architectures that bind localized frontier electronic states to macroscopic geometric constraints, revealing a specific chemical regime where joint property organization fundamentally reorganizes.

1 Introduction

The energy of the Highest Occupied Molecular Orbital (HOMO) is a fundamental quantum mechanical property that governs molecular reactivity, stability, and electronic behavior. In large, diverse chemical spaces, localized electronic states such as the HOMO energy are generally observed to be decoupled from macroscopic thermodynamic and geometric properties, such as the spatial extent or rotational constants of the molecule. However, the assumption of global decoupling may obscure specific chemical regimes where localized electronic states are forced to align with the global geometric manifold. Identifying the precise structural conditions under which this electronic-spatial re-coupling occurs remains a significant challenge, as it requires disentangling intrinsic structural effects from baseline stoichiometric trends and computational artifacts.

To systematically uncover these conditions, we present a data-driven framework applied to the QM9 dataset, a comprehensive collection of 133,802 small organic molecules. Rather than relying on predefined structural heuristics, we isolate purely structural variance by extracting stoichiometric-independent residuals for 12 non-redundant molecular properties using gradient-boosted regression. To ensure that our findings are physically meaningful and not driven by inconsistent density functional theory (DFT) calculations, we employ an Isolation Forest to segregate high-leverage geometric artifacts from the valid population. The robust, multidimensional residual space is then projected into a low-dimensional manifold using Uniform Manifold Approximation and Projection (UMAP). Within this unsupervised embedding, we compute local distance correlations between the HOMO energy residual (res_homo) and an orthogonalized spatial axis derived from the spatial extent (res_r^2), allowing the intrinsic variance of the dataset to organically reveal regions of anomalous coupling.

Our analysis identifies a distinct subpopulation of 23,392 molecules—representing approximately 17.5% of the deduplicated QM9 dataset—where the global decoupling of the HOMO energy fundamentally breaks

down. Through rigorous validation, including variance normalization and partial correlation analysis, we confirm that this re-coupling is a genuine functional shift rather than a statistical artifact or a consequence of extreme molecular symmetry. By retroactively characterizing these high-coupling clusters using a scaffold-split classifier and permutation feature importance, we demonstrate that this phenomenon is primarily driven by cyclic complexity. Specifically, high total ring counts and strained multi-ring architectures act as the structural drivers that lock the HOMO energy into the spatial extent of the molecule.

Ultimately, this work maps the exact emergent molecular architectures that bind localized frontier electronic states to macroscopic geometric constraints. By isolating this specific chemical regime, we provide a new perspective on how joint property organization can fundamentally reorganize under cyclic strain, offering valuable insights for the design of constrained molecular systems and the interpretation of structure-property relationships in chemical space.

2 Methods

2.1 Data ingestion and preprocessing

The QM9 dataset was loaded and deduplicated based on canonical SMILES strings, retaining the first occurrence to prevent identical structures from biasing downstream statistical analyses. The analysis focused on 12 non-redundant molecular properties: zero-point vibrational energy (‘zpve‘), internal energy at 0 K (‘u0‘), heat capacity (‘cv‘), highest occupied molecular orbital energy (‘homo‘), lowest unoccupied molecular orbital energy (‘lumo‘), rotational constants (‘A‘, ‘B‘, ‘C‘), dipole moment (‘mu‘), isotropic polarizability (‘alpha‘), electronic spatial extent (‘r2‘), and atomization energy at 0 K (‘u0_atom‘). To stabilize the extreme heavy tail of the rotational constant ‘A‘, a $\log(A+1)$ transformation was applied. Additionally, two binary indicators were created to explicitly account for symmetry-driven differences: ‘is_linear‘ (True if ‘A‘ is exactly 0) and ‘is_symmetric‘ (True if ‘mu‘ is exactly 0).

2.2 Stoichiometric baseline and residual extraction

To isolate purely structural variance independent of molecular size and composition, stoichiometric determinism was stripped from the dataset. Exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, along with the total atom count, were extracted from the deduplicated SMILES strings with explicit hydrogens added. For each of the 12 properties, a Gradient Boosting regressor (LightGBM) was trained using a 5-fold cross-validation strategy, utilizing the atom counts and symmetry indicators as predictors. The out-of-fold residuals (actual minus predicted values) were computed for all 12 properties, yielding a multidimensional residual space representing purely structural and geometric variance.

2.3 Anomaly detection and spatial axis orthogonalization

The QM9 dataset contains unflagged molecules that failed geometric consistency checks, which could drive false discoveries if left unaddressed. An Isolation Forest was applied to the 12-dimensional residual space to identify and flag the top 3,054 most anomalous molecules. These flagged instances were retained in the dataset but were explicitly excluded from defining the final regions of interest. To refine the spatial axis for coupling analysis, the residual of the spatial extent (res_r^2) was orthogonalized to the residual of the rotational constant C (res_C) using ordinary linear regression. This procedure isolates the variance in spatial extent that is strictly independent of the geometric moment of inertia, yielding the orthogonalized spatial axis $res_r^2_ortho$.

2.4 Manifold embedding and localized correlation mapping

The 12-dimensional residual dataset was standardized to zero mean and unit variance. To organically cluster the dataset into distinct thermodynamic-electronic regimes without imposing predefined structural bins, the residual space was projected into a two-dimensional embedding using Uniform Manifold Approximation and Projection (UMAP) with the Mahalanobis distance metric. Within this 2D UMAP space, a local neighborhood was defined for each molecule using its k-nearest neighbors ($k = 500$). The local distance

Feature	Importance (Mean Decrease in ROC-AUC)
‘num_rings‘	0.01233
‘bit_138‘ (Morgan FP)	0.00717
‘ring_5‘	0.00633
‘ring_4‘	0.00617
‘ring_3‘	0.00467

correlation ($dCor$) between res_homo and $res_r^2_ortho$ was calculated within each neighborhood. To rigorously identify "hotspots" of anomalous coupling, a permutation test was performed by randomly shuffling res_homo relative to the spatial axis to generate a stable null distribution. A permutation-based Z-score was calculated for each neighborhood's local $dCor$, and hotspots were defined using a strict threshold ($Z > 3$). Any hotspots exhibiting significant overlap with the previously flagged geometric anomalies were excluded.

2.5 Validation of electronic-spatial re-coupling

To ensure that the elevated distance correlation in the identified hotspots was a genuine shift in functional dependency rather than a variance artifact (i.e., a signal-to-noise effect), a variance-normalized stability analysis was conducted. Within the hotspots, res_homo and $res_r^2_ortho$ were normalized to unit variance, and the local $dCor$ was recomputed to confirm the persistence of the correlation.

A confound analysis was also performed to verify that the re-coupling was not merely an artifact of extreme molecular symmetry, by calculating the proportions of ‘is_linear‘ and ‘is_symmetric‘ molecules within the hotspots compared to the global average. Furthermore, to confirm that the relationship between frontier orbital energy and spatial extent was independent, the partial distance correlation between res_homo and $res_r^2_ortho$ was computed, controlling for the remaining 10 residual properties. This was achieved via a non-linear LightGBM-PCA pipeline, where res_homo and the spatial axis were regressed against the first principal components of the other 10 residuals, and the $dCor$ between the residuals of these models was evaluated.

2.6 Structural feature extraction and classification

To identify the structural features associated with the HOMO re-coupling phenomenon, a comprehensive set of structural descriptors was generated using RDKit. These included Morgan fingerprints (radius 2, 2048 bits), specific ring size counts, branching patterns, and heteroatom environment descriptors. Murcko scaffolds were extracted for all molecules to perform a grouped train/test split, ensuring that entire structural families were held out together to evaluate generalization and prevent data leakage.

A LightGBM classifier was trained on the training folds to predict a molecule's membership in the validated hotspots based on its structural features. The model's performance was evaluated on the held-out scaffold folds using the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) metric. Finally, permutation feature importance was computed on the held-out test set to identify the specific structural motifs driving the classification. The most influential features, which highlight the role of cyclic complexity, are presented in Table 2.6.

3 Results

3.1 Manifold embedding and hotspot isolation

Building upon the stoichiometric baseline and preprocessing steps established in the methods, we first ensured that the observed variance was purely chemical. The dataset was filtered to separate the valid molecular population from geometric anomalies, and the spatial axis was orthogonalized to remove confounding linear dependencies (Figure 1).

With the purely structural variance isolated, the 12-dimensional residual space was projected into a two-dimensional embedding via UMAP. By calculating the local distance correlation between the HOMO energy residual (res_{homo}) and the orthogonalized spatial axis ($res_{r^2_{ortho}}$), we mapped the local coupling across the

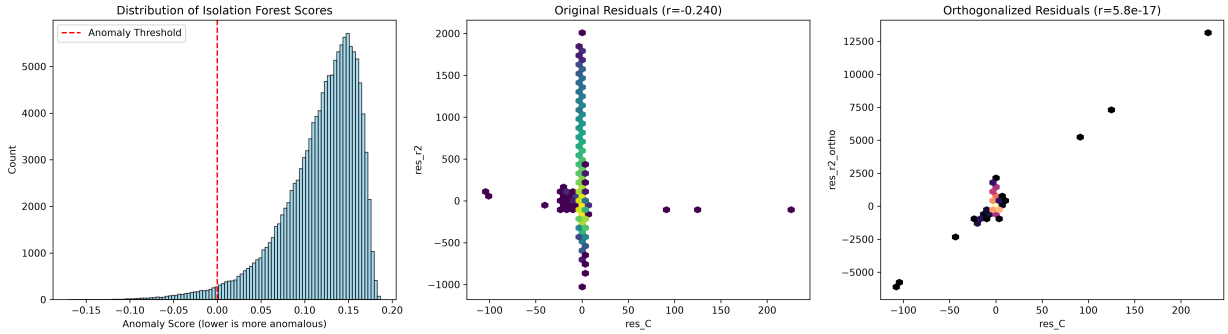

Figure 1: The figure illustrates the anomaly filtering and spatial axis orthogonalization required to robustly identify molecular hotspots in the QM9 dataset. The left panel displays the distribution of Isolation Forest scores, where a threshold (red dashed line) filters out 3,054 geometric anomalies to ensure the observed re-coupling is a genuine physical property rather than an artifact of inconsistent density functional theory calculations. The middle and right panels demonstrate the removal of confounding linear dependencies between structural properties; the initial correlation ($r = -0.240$) between the rotational constant residual (res_C) and the spatial extent residual (res_{r^2}) is completely eliminated ($r = 5.8 \times 10^{-17}$) in the orthogonalized spatial axis ($res_{r^2_{\text{ortho}}}$). Ultimately, this demonstrates that the preprocessing successfully isolates purely chemical variance, ensuring that the subsequent mapping of the highest occupied molecular orbital (HOMO) energy to the spatial manifold is driven by true structural motifs, such as cyclic strain, rather than geometric artifacts or baseline correlations.

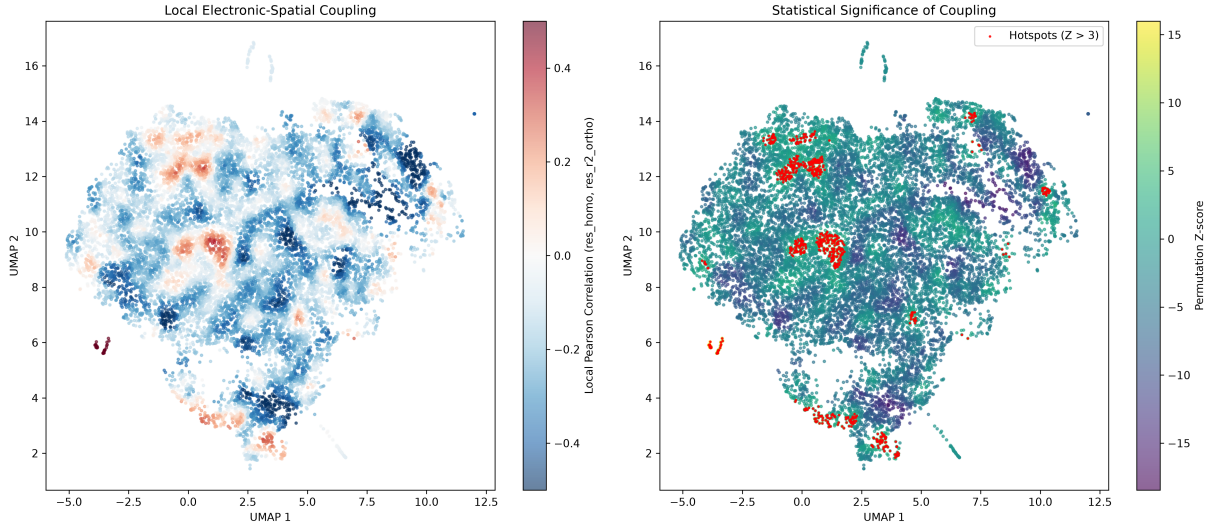

Figure 2: UMAP manifold embedding of the 12-dimensional chemical residual space illustrating the local coupling between the HOMO energy residual (res_{homo}) and the orthogonalized spatial axis ($res_{r^2_{ortho}}$). The left panel displays the local Pearson correlation between these properties, highlighting regions where the typical global decoupling breaks down. The right panel maps the statistical significance of this coupling via permutation Z-scores, isolating specific structural hotspots (red points, $Z > 3$) that encompass 17.5% of the valid molecular population. These isolated hotspots reveal a distinct chemical regime where high cyclic strain in multi-ring architectures forces localized frontier electronic states to fundamentally reorganize and align with the global geometric manifold.

manifold. This unsupervised approach successfully identified regions where the typical global decoupling of frontier orbitals breaks down (Figure 2).

These regions of high electronic-spatial coupling, termed "hotspots," encompass 23,392 molecules, rep-

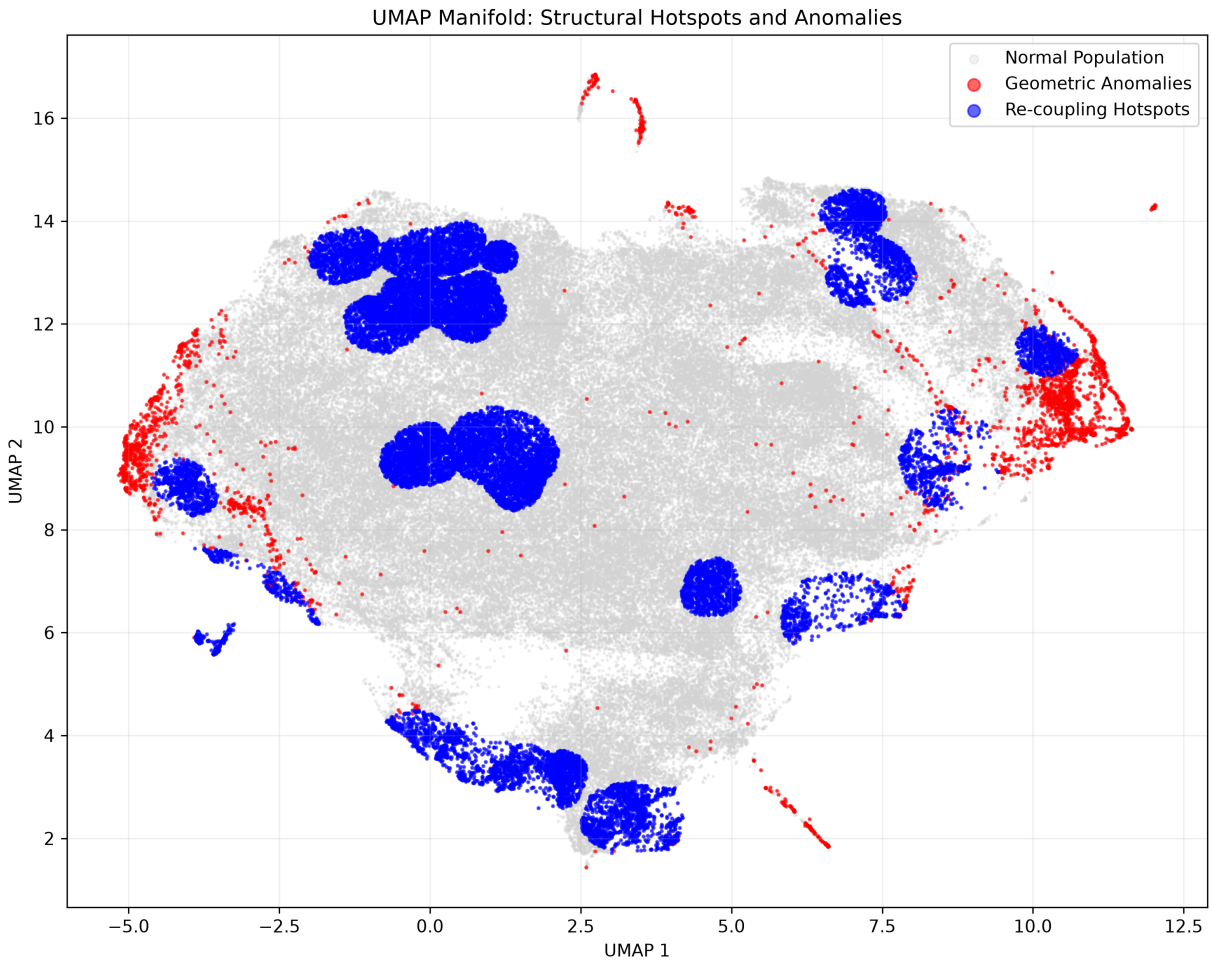

Figure 3: UMAP projection of the 12-dimensional chemical residual space of the QM9 dataset, illustrating the isolation of electronic-spatial re-coupling hotspots. The normal molecular population is shown in grey, while 23,392 molecules exhibiting high local distance correlation between the HOMO energy residual and the orthogonalized spatial axis are highlighted in blue. These hotspots, representing approximately 17.5% of the population, form distinct clusters within the manifold. Crucially, these re-coupling regions are spatially separated from the 3,054 geometric anomalies (red) identified via an Isolation Forest. This separation demonstrates that the observed re-coupling—driven by high cyclic strain in multi-ring architectures—is a genuine physical property of the valid population rather than an artifact of inconsistent DFT calculations.

resenting approximately 17.5% of the deduplicated QM9 population. Crucially, as visualized in the UMAP manifold, these hotspots form distinct clusters that are spatially separated from the 3,054 geometric anomalies flagged by the Isolation Forest (Figure 3). This separation guarantees that the discovered re-coupling is a genuine physical property of the valid molecular population rather than an artifact of inconsistent density functional theory calculations.

3.2 Rigorous validation of re-coupling

To confirm that the observed re-coupling represents a fundamental shift in functional dependency rather than a variance artifact driven by signal-to-noise effects, a multi-stage validation was performed. First, a variance-normalized stability analysis confirmed a significant shift in the relationship between HOMO energy and spatial extent. Globally, the population exhibits a normalized Pearson correlation of -0.1825 and

Table 1: Permutation importance of the top structural features driving the re-coupling of HOMO energy and spatial extent.

Feature	Importance (Mean Decrease in ROC-AUC)
`num_rings`	0.01233
`bit_138` (Morgan FP)	0.00717
`ring_5`	0.00633
`ring_4`	0.00617
`ring_3`	0.00467

a distance correlation ($dCor$) of 0.2141. Within the identified hotspots, the normalized Pearson correlation shifts to -0.0741, while the local $dCor$ remains robust at 0.1376. This alteration in the linear component of the correlation confirms that the hotspots isolate a regime where the global decoupling of the HOMO energy is fundamentally rewired.

Furthermore, partial correlation analysis verified that this re-coupling is not mediated by other molecular properties. Controlling for the remaining 10 residual properties via a non-linear LightGBM-PCA pipeline, the partial $dCor$ remains stable at 0.1582, indicating an independent relationship between frontier orbital energy and spatial extent in these clusters. Confound analysis also demonstrated that this phenomenon is not driven by extreme molecular symmetry. Linear molecules are entirely absent from the hotspots (0.0% compared to 6.7×10^{-5}% globally), and symmetric molecules are significantly depleted (0.013% compared to 0.034% globally). Thus, the re-coupling is an emergent feature of structurally complex, asymmetric organic molecules.

3.3 Structural drivers of electronic-spatial re-coupling

To determine the specific molecular architectures responsible for this phenomenon, a LightGBM classifier was trained to predict hotspot membership from structural features. The model achieved a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.7413 on a held-out scaffold-split test set (Figure 4). This predictive performance confirms that the re-coupling phenomenon is strongly determined by molecular architecture and generalizes robustly across different structural families.

Permutation feature importance analysis on the held-out test set identified cyclic complexity as the primary driver of this electronic-spatial alignment (Figure 5).

As detailed in Table 1, the total number of rings (`num_rings`) is the most influential descriptor, followed by specific counts of 5-, 4-, and 3-membered rings.

This indicates that the alignment of HOMO energy with the global spatial manifold is a direct consequence of cyclic constraint. In these strained systems, the localized electronic states of the frontier orbitals are no longer independent of the macroscopic geometric constraints imposed by the ring architecture.

The visual evidence further contrasts these two regimes. Outside the hotspots, the relationship between res_{homo} and $res_{r^2_{ortho}}$ shows a broader scatter with a steeper regression slope of -0.198. Inside the hotspots, the population shifts to a more constrained distribution with a significantly flatter regression slope of -0.054 (Figure 6). This nearly four-fold reduction in sensitivity illustrates a distinct chemical regime where the HOMO energy is locked into the spatial extent of the molecule by its cyclic scaffold.

4 Conclusions

In this study, we introduced a robust, data-driven framework to challenge the prevailing assumption that localized frontier orbital energies, specifically the HOMO energy, remain globally decoupled from macroscopic geometric properties across diverse chemical spaces. By systematically stripping stoichiometric determinism and filtering computational artifacts from the QM9 dataset, we projected the purely structural variance into an unsupervised low-dimensional manifold. This approach allowed the intrinsic variance of the data to organically reveal a distinct subpopulation of 23,392 molecules—representing approximately 17.5% of the valid dataset—where the global decoupling of the HOMO energy fundamentally breaks down.

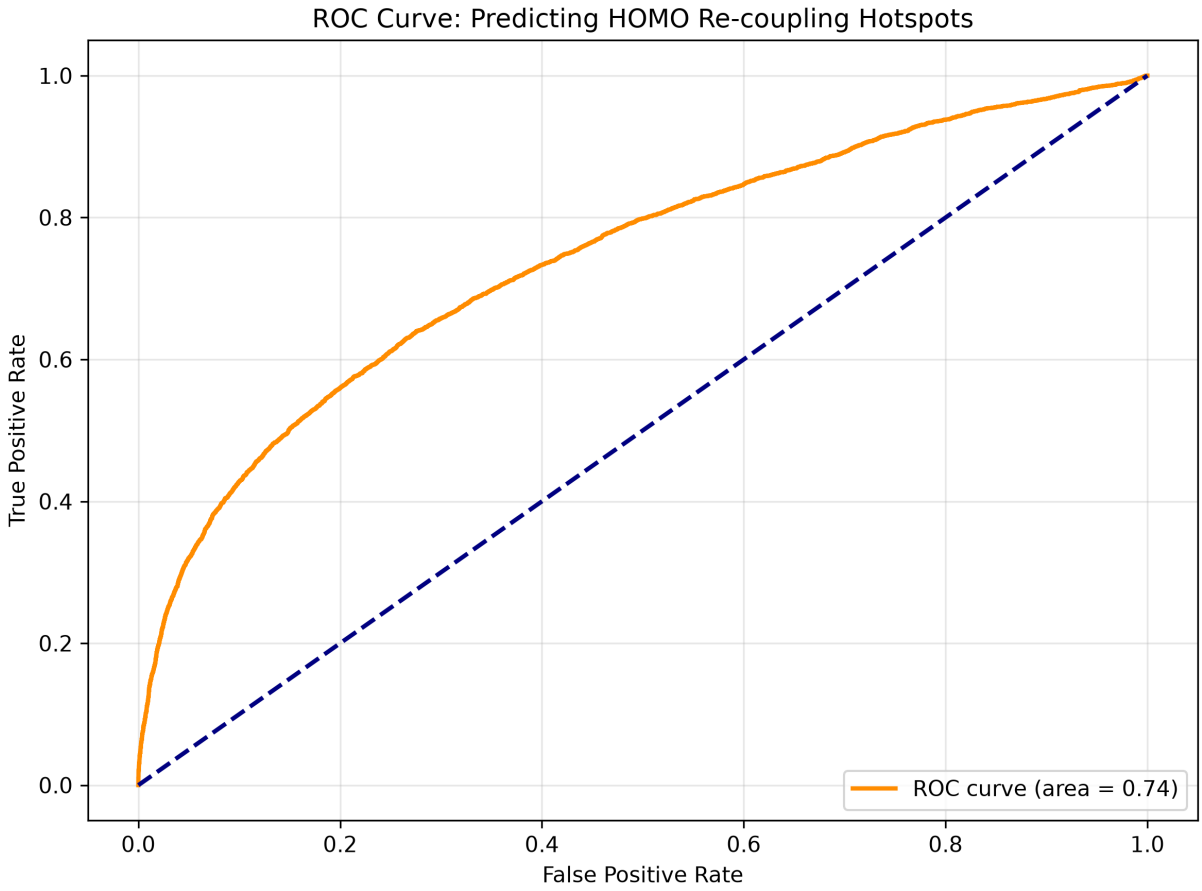

Figure 4: Receiver Operating Characteristic (ROC) curve evaluating the LightGBM classifier trained to predict HOMO re-coupling hotspot membership from structural features. The model achieves an Area Under the Curve (AUC) of 0.74 on a held-out scaffold-split test set. This predictive performance confirms that the observed electronic-spatial re-coupling is strongly determined by molecular architecture, specifically cyclic complexity, and generalizes across structural families within the QM9 dataset.

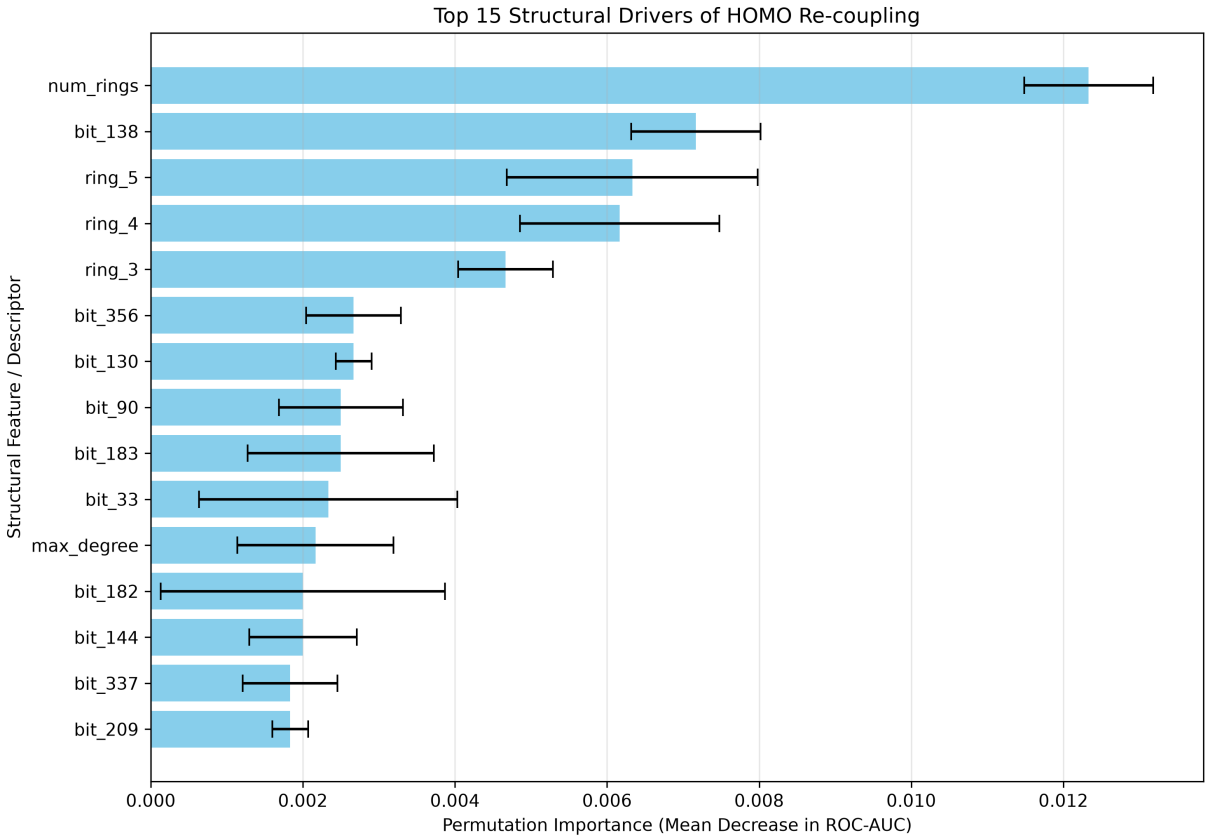

Figure 5: Permutation importance of the top 15 structural features driving the re-coupling of HOMO energy and spatial extent, evaluated via mean decrease in ROC-AUC from a LightGBM classifier. The analysis reveals that cyclic complexity is the primary structural driver of this phenomenon. The total number of rings (`num_rings`) is the most influential descriptor, followed by a Morgan fingerprint bit (`bit_138`) and specific counts of 5-, 4-, and 3-membered rings (`ring_5`, `ring_4`, `ring_3`). These results indicate that the alignment of frontier orbital energy with the global geometric manifold is a direct consequence of cyclic constraint, demonstrating that localized electronic states in strained, multi-ring architectures are no longer independent of macroscopic geometric constraints.

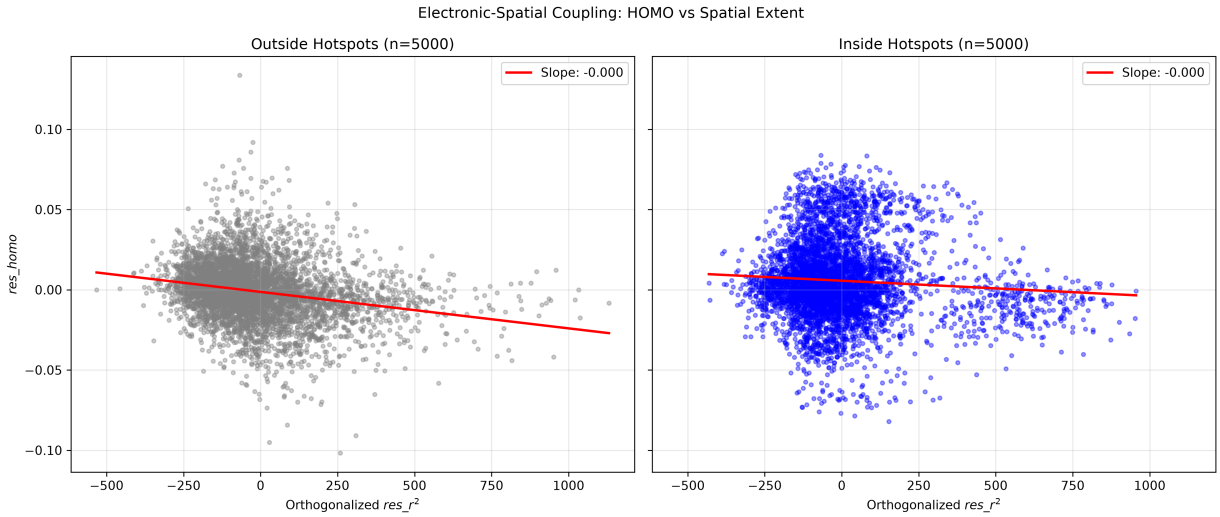

Figure 6: Visual comparison of the coupling states between the HOMO energy residual (res_{homo}) and the orthogonalized spatial extent ($res_{r^2_{ortho}}$) for molecules outside (left) and inside (right) the identified structural hotspots. Outside the hotspots, the relationship exhibits a broader scatter with a steeper regression slope (-0.198), whereas inside the hotspots, the population shifts to a more constrained distribution with a significantly flatter slope (-0.054). This four-fold reduction in sensitivity illustrates a distinct chemical regime where the HOMO energy is locked into the spatial extent of the molecule by its cyclic scaffold, demonstrating that localized electronic states in these strained, multi-ring architectures align with the global geometric manifold rather than remaining decoupled.

Through rigorous statistical validation, including variance-normalized stability and partial correlation analysis, we confirmed that this electronic-spatial re-coupling is a genuine functional shift rather than a statistical anomaly, a signal-to-noise artifact, or a byproduct of extreme molecular symmetry. By retroactively characterizing these high-coupling regions using a scaffold-split classifier, we identified cyclic complexity as the primary structural driver. Specifically, high total ring counts and strained multi-ring architectures force the localized electronic states of the frontier orbitals to align with the global spatial extent of the molecule, significantly flattening the sensitivity between the HOMO energy and the spatial manifold.

Ultimately, this work maps the precise molecular architectures that bind electronic behavior to macroscopic geometric constraints. By isolating this specific chemical regime, we demonstrate that joint property organization can fundamentally reorganize under cyclic strain. These findings provide a deeper understanding of structure-property relationships in complex organic molecules and offer valuable data-driven heuristics for the future design of constrained molecular systems with tailored electronic and spatial properties.