Topological Stratification of HOMO Decoupling from the Vibrational-Spatial Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding the structural drivers that isolate frontier molecular orbitals, specifically the highest occupied molecular orbital (HOMO), from global molecular organization remains a significant challenge in computational chemistry. To address this, we map the decoupling of electronic states across macroscopic complexity regimes using the QM9 dataset. We remove extensive size-scaling and localized functional group effects via gradient-boosted regression to isolate stoichiometric residuals, and subsequently extract a non-linear spatial-vibrational manifold using Nystroem-approximated Kernel Principal Component Analysis. By stratifying molecules along a continuous saturation gradient, we compute the distance correlation between frontier orbital energies and the dominant manifold axis. Our analysis reveals a distinct asymmetry, termed the "coupling flip," where the lowest unoccupied molecular orbital (LUMO) maintains stronger coupling to the global molecular state than the HOMO. The HOMO exhibits profound detachment at intermediate saturation levels, becoming nearly orthogonal to the spatial and vibrational state of the molecule. Motif analysis identifies specific functional groups—predominantly carbonyls, aromatic nitrogens, and nitriles—as the primary structural drivers of this isolation, demonstrating that decoupling is driven by local electronic environments rather than simple molecular size. Finally, leave-one-scaffold-out validation confirms that these topological transitions are fundamental properties of the chemical space, while boundary case analysis of linear molecules reveals tighter spatial-polarizability coupling but weaker vibrational coupling compared to non-linear controls.

1 Introduction

The frontier molecular orbitals, particularly the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), are fundamental to predicting chemical reactivity, optical properties, and electron transport. Traditionally, the energies of these orbitals are analyzed through the lens of local chemical environments and specific functional groups. However, a significant challenge in computational chemistry lies in understanding how these localized electronic states couple to, or decouple from, the global macroscopic organization of the molecule. Global molecular states are typically defined by a tightly bound spatial-vibrational manifold, encompassing properties such as electronic spatial extent (r^2), isotropic polarizability (α), and zero-point vibrational energy (ZPVE). Because these macroscopic properties inherently scale with molecular size, the underlying topological dependencies of frontier orbitals are often obscured by extensive size-scaling effects.

To uncover the structural drivers that isolate the HOMO from global molecular organization, it is necessary to map its decoupling across macroscopic complexity regimes rather than isolated chemical neighborhoods. While it is well established that specific heteroatoms and conjugation networks perturb orbital energies, the exact magnitude of electronic-thermodynamic orthogonalization across a continuous structural hierarchy remains unquantified. Identifying the specific topological thresholds—such as the transition from acyclic aliphatic chains to highly conjugated polycyclic systems—where local electronic environments fundamentally detach from a molecule's macroscopic spatial and thermodynamic state requires a rigorous, data-driven framework.

In this study, we address this challenge by mapping the topological stratification of HOMO decoupling across the QM9 dataset. We introduce a methodology that isolates stoichiometric residuals by removing extensive size-scaling and localized functional group effects via gradient-boosted regression. From these

residuals, we extract a non-linear spatial-vibrational manifold using Nystroem-approximated Kernel Principal Component Analysis (KPCA). This approach yields a stable, dominant axis of molecular organization, providing a structural reference frame to evaluate the behavior of frontier orbitals independent of simple molecular size or localized heteroatom effects.

By stratifying the dataset along a continuous saturation gradient, we compute the distance correlation between the frontier orbital energies and the dominant manifold axis. Our analysis reveals a distinct asymmetry, termed the "coupling flip," wherein the LUMO maintains a significantly stronger coupling to the global molecular state than the HOMO. We demonstrate that the HOMO exhibits profound detachment at intermediate saturation levels, becoming nearly orthogonal to the spatial and vibrational state of the molecule. Through motif analysis, we identify specific functional groups—predominantly carbonyls, aromatic nitrogens, and nitriles—as the primary structural drivers of this isolation, proving that decoupling is driven by local electronic environments rather than simple molecular size. Finally, we confirm that these topological transitions are fundamental properties of the chemical space through rigorous leave-one-scaffold-out validation, and we explore the unique geometric constraints of linear molecules as a boundary case to further elucidate the relationship between spatial extent, polarizability, and vibrational coupling.

2 Methods

2.1 Data preprocessing and stoichiometric baseline

To establish a rigorous framework for evaluating orbital decoupling, we utilized the QM9 dataset, deduplicating entries based on their SMILES strings. RDKit loggers were silenced during parsing to prevent buffer overflow from deprecation warnings. To manage the extreme heavy-tailed distribution of the rotational constant A, we applied a $\log(A+1)$ transformation. Molecules exhibiting linear geometries (defined as $A = 0$ or $B = C$) were flagged with a binary indicator (‘is_linear‘) for downstream boundary case analysis. Redundant thermodynamic columns were removed, while the HOMO-LUMO energy gap (‘gap‘) was retained to explicitly test its behavior as an asymmetric coupling target.

Because macroscopic properties inherently scale with molecular size, we first removed extensive size-scaling effects. We parsed the SMILES strings to extract exact stoichiometric counts for carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. Using these five atom counts as features, we fit a Gradient Boosting regressor (via LightGBM) for the remaining molecular properties and extracted the first-order residuals. This procedure isolates the structural and topological variance from the dominant stoichiometric baseline. It should be noted as a limitation that 3,054 molecules failing geometric consistency checks in the original QM9 dataset remain unmarked and are included in this analysis.

2.2 Non-linear manifold extraction via kernel approximation

To capture the non-linear dependencies defining the global molecular state, we isolated the stoichiometric residuals of the spatial-vibrational triad: electronic spatial extent (res_r^2), isotropic polarizability (res_α), and zero-point vibrational energy (res_zpve). These residuals were globally standardized. We then applied a Nystroem approximation with a Radial Basis Function (RBF) kernel to efficiently approximate Kernel Principal Component Analysis (KPCA) across the entire dataset.

The kernel bandwidth (γ) was tuned using a heuristic based on the median pairwise distance between samples. We extracted the first two non-linear principal components and calculated the distance correlation ($dCor$) between them. A low distance correlation verifies that the first component (‘geom_kPC1‘) represents the dominant, stable axis of the spatial-vibrational manifold rather than one of several competing dimensions. This component was retained as the primary structural reference frame for evaluating the frontier orbitals.

2.3 Functional group control and second-order residuals

To ensure that the observed decoupling of frontier orbitals is driven by global topology rather than localized heteroatom effects, we refined our control variables. Using RDKit's `rdMolDescriptors`, we computed specific functional group counts (e.g., amines, carbonyls, nitriles, ethers, halogens) and the total number of heteroatoms for each molecule.

A second set of non-linear LightGBM regressors was fit to predict the first-order residuals of the HOMO, LUMO, gap, and 'geom_kPC1' using these functional group counts and connectivity descriptors. By extracting the second-order residuals from these models, we effectively isolated the pure topological and conjugative effects, stripping away the variance strictly attributable to specific, localized functional group environments.

2.4 Continuous topological stratification and sliding window analysis

To map the decoupling across macroscopic complexity regimes, we computed the fraction of sp^3-hybridized carbons (`FractionCSP3`) for all molecules as a continuous measure of saturation. The dataset was sorted along this saturation gradient, and a continuous sliding window approach was implemented (window size of 2,500 molecules, step size of 500). Computations were parallelized across 14 CPUs using `joblib`.

Within each discrete complexity tier, we computed the distance correlation of the controlled HOMO, LUMO, and gap residuals against the controlled 'geom_kPC1' manifold axis. We also computed the internal manifold stability by measuring the distance correlation between the controlled res_r^2 and res_α. To quantify the asymmetry in orbital behavior, we calculated the "Coupling Flip" delta for each window, defined as:

$$\Delta dCor = dCor_{LUMO} - dCor_{HOMO} \tag{1}$$

A positive $\Delta dCor$ indicates stronger LUMO coupling to the global geometric state, potentially reflecting its more diffuse spatial nature, whereas a negative value would indicate stronger HOMO coupling.

2.5 Motif analysis of highly decoupled states

To identify the structural drivers of HOMO isolation, we isolated the specific sliding windows where the HOMO exhibited the absolute lowest distance correlation with 'geom_kPC1'. Molecules from these highly decoupled regions (the bottom 5% of the coupling distribution) were extracted. We generated Morgan fingerprints (radius 2, 2048 bits) and RDKit fragment counts to provide interpretable structural features.

The dataset was balanced by undersampling the majority class (the remainder of the dataset). We then trained a LightGBM classifier to distinguish the highly decoupled molecules. Crucially, 'geom_kPC1' was included as a feature in this classifier to determine if decoupling is driven by specific motifs independent of the molecule's overall position on the manifold. Permutation importance was used to extract feature importances and accurately rank the specific structural fragments driving the HOMO's detachment.

2.6 Scaffold hold-out validation and boundary case analysis

To rigorously verify that the continuous coupling flip and decoupling trends are fundamental properties of the chemical space rather than artifacts of specific molecular families, we performed a leave-one-scaffold-class-out validation. Molecules were grouped by their Murcko scaffolds, and the top 20 most frequent scaffold families were identified. The entire sliding window trajectory was recomputed 20 times, systematically holding out one scaffold family per iteration. The variance and standard error across these folds were calculated to establish robust confidence intervals for the continuous transition.

Finally, the linear molecules flagged during preprocessing were isolated as a specific boundary case study. To determine if their distinct coupling behavior is a true electronic phenomenon or a 1D geometric artifact, we computed the internal distance correlation between res_r^2, res_α, and res_zpve for this subset. These internal correlations were compared to a control group of non-linear molecules matched for similar atom counts and `FractionCSP3` values, as well as to a synthetic dataset of perfectly correlated 1D points, clarifying the role of 1D dimensionality in spatial-polarizability coupling.

3 Results

3.1 Non-linear manifold of spatial-vibrational properties

Following the removal of extensive size-scaling effects and localized functional group variance, we isolated the fundamental geometric-vibrational axis of the QM9 dataset. The spatial-vibrational triad—comprising electronic spatial extent (r^2), isotropic polarizability (α), and zero-point vibrational energy (ZPVE)—exhibits

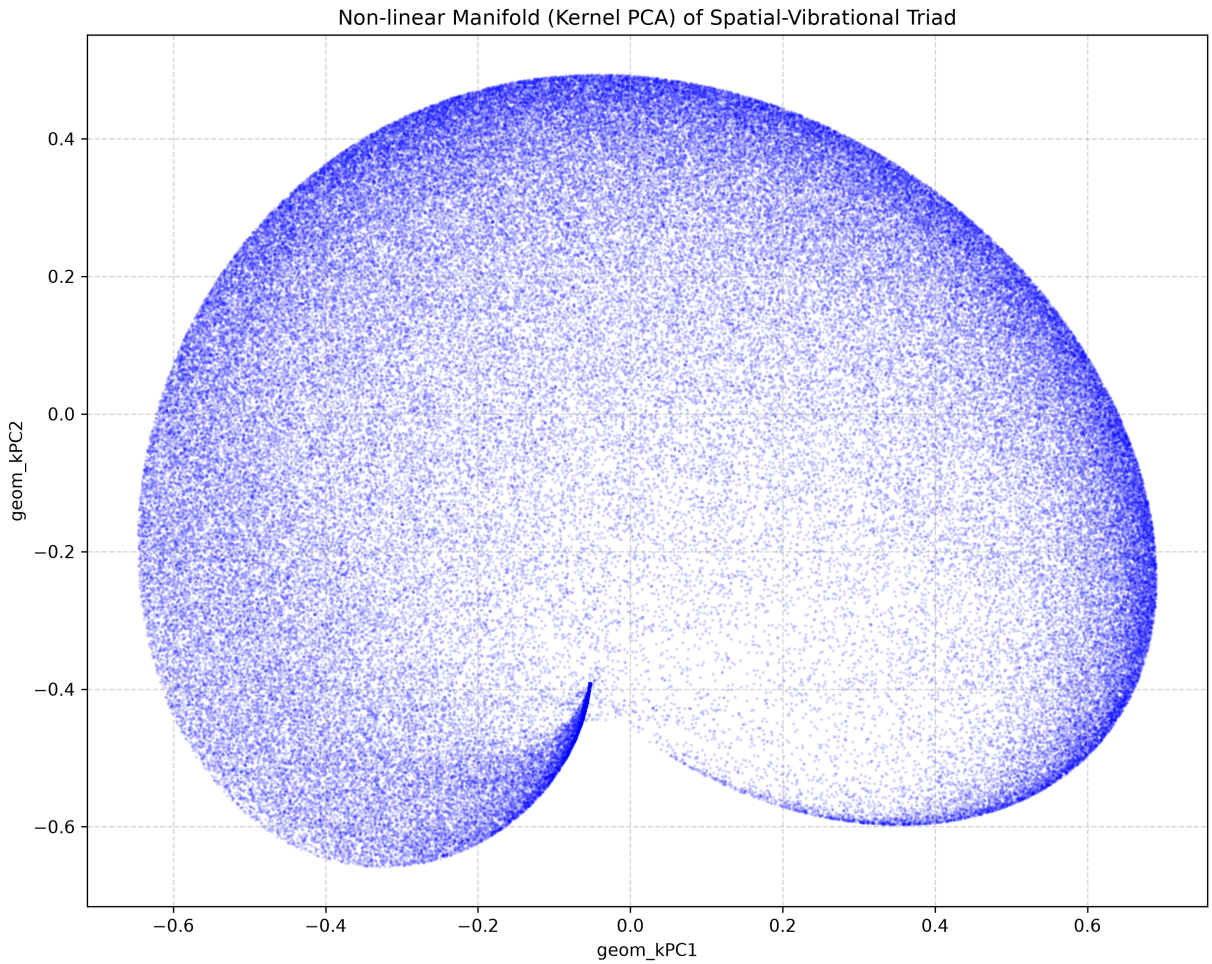

Figure 1: Scatter plot of the first two components from a Nystroem-approximated Kernel Principal Component Analysis (KPCA) of the spatial-vibrational triad, comprising electronic spatial extent (r^2), isotropic polarizability (α), and zero-point vibrational energy (ZPVE). The visualization illustrates the core non-linear manifold of these molecular properties after removing extensive size-scaling effects. The horizontal axis, `geom_kPC1`, captures 25.80% of the residual variance and acts as the dominant axis of molecular organization, while the vertical axis, `geom_kPC2`, accounts for an additional 17.59%. The distinct structure of the data points, combined with a low distance correlation ($dCor = 0.2083$) between the two components, confirms the internal coherence and fundamental geometric-vibrational organization of the molecules.

high internal coherence. A Nystroem-approximated Kernel Principal Component Analysis (KPCA) utilizing a Radial Basis Function (RBF) kernel (bandwidth $\gamma = 0.3022$) successfully extracted this non-linear manifold (Figure 1). The primary component, `geom_kPC1`, captures 25.80% of the residual variance, while the secondary component, `geom_kPC2`, accounts for an additional 17.59%. The low distance correlation ($dCor = 0.2083$) between these two components confirms that `geom_kPC1` serves as the dominant, stable axis of global molecular organization.

While the manifold provides a stable reference frame, its internal stability is not strictly constant across the structural hierarchy. As depicted in the bottom panel of Figure 2, the distance correlation between the r^2 and α residuals fluctuates between approximately 0.22 and 0.54 across the saturation gradient. This indicates that although electron density and vibrational states remain fundamentally linked, the magnitude of their coupling is continuously modulated by the underlying molecular topology.

3.2 Topological re-organization and the coupling flip

By tracking the distance correlation of the HOMO and LUMO residuals against `geom_kPC1` across the continuous saturation gradient (defined by `FractionCSP3`), we observe a sharp transition in molecular organization.

As illustrated in the top panel of Figure 2, a distinct HOMO-LUMO asymmetry emerges. The LUMO consistently exhibits stronger coupling to the global molecular state than the HOMO. In the low-saturation regime (`FractionCSP3` ≈ 0.02), the mean coupling for the LUMO is $dCor = 0.2923 \pm 0.0095$ (standard error), whereas the HOMO coupling is significantly lower at 0.1142 ± 0.0052. This disparity, which we term the "Coupling Flip" ($\Delta dCor = dCor_{LUMO} - dCor_{HOMO}$), reaches a maximum asymmetry of 0.3223 at a saturation level of approximately 0.24 (Figure 2, bottom panel).

Furthermore, the HOMO residuals exhibit a profound detachment from the molecular manifold at intermediate saturation levels. The minimum mean coupling for the HOMO ($dCor = 0.0462$) occurs at `FractionCSP3` ≈ 0.57. In this specific topological regime, the HOMO energy becomes almost entirely orthogonal to the molecule's spatial and vibrational state. Conversely, the LUMO coupling at this same saturation level remains higher at approximately 0.1622 ± 0.0058, reflecting its more diffuse spatial nature that allows it to remain partially tethered to the global geometry.

3.3 Structural drivers of HOMO decoupling

To determine the chemical origins of this orthogonalization, we performed a motif analysis on the highly decoupled molecular states using a Gradient Boosting classifier (AUC = 0.9276). The feature importance rankings, evaluated via permutation importance (AUC drop), isolate the specific structural fragments driving the detachment of the HOMO.

As shown in Figure 3, the most significant driver of HOMO decoupling is the carbonyl group (`fr_C_O`), yielding an importance score of 0.1343—the highest recorded for any structural motif. Other major drivers include aromatic nitrogens (`fr_Ar_N`, 0.0328), nitriles (`fr_nitrile`, 0.0250), aliphatic hydroxyls (`fr_Al_OH`, 0.0217), and tertiary amines (`fr_NHO`, 0.0194).

Crucially, the molecule's coordinate position on the manifold itself (`manifold_coord`) ranks only 10th in importance (0.0076). This finding confirms that HOMO decoupling is not merely a function of molecular size or bulk geometric scaling, but is explicitly driven by specific, localized functional groups that effectively insulate the frontier electron density from the global geometric state.

3.4 Robustness and boundary cases

To ensure that the observed topological transitions are fundamental properties of the chemical space rather than artifacts of specific molecular families, we conducted a Leave-One-Scaffold-Out (LOSO) validation over the top 20 scaffold classes. The average standard error for the coupling flip across these hold-out folds is 0.007461. The narrow shaded confidence intervals in Figure 2 demonstrate the high stability of these trends, confirming that the transition from asymmetric to near-symmetric coupling states is robust across diverse structural backbones.

Finally, an analysis of linear molecules (37 samples) highlights a unique boundary case in spatial-vibrational organization. The internal manifold structure of these 1D-constrained systems differs markedly from the bulk population. The correlation between spatial extent (res_r^2) and polarizability (res_α) is substantially higher in linear molecules ($dCor = 0.8317$) compared to matched non-linear controls ($dCor = 0.6692$). However, their coupling to vibrational states is notably weaker, with the correlation between res_r^2 and res_zpve dropping to $dCor = 0.3103$ (versus 0.4929 for controls). This indicates that while electronic extent and polarizability are more tightly coupled under 1D geometric constraints, the decoupling of these electronic states from vibrational modes becomes significantly more pronounced.

4 Conclusions

In this study, we established a data-driven framework to quantify the decoupling of frontier molecular orbitals from the global spatial-vibrational state across macroscopic complexity regimes. By isolating stoichiometric

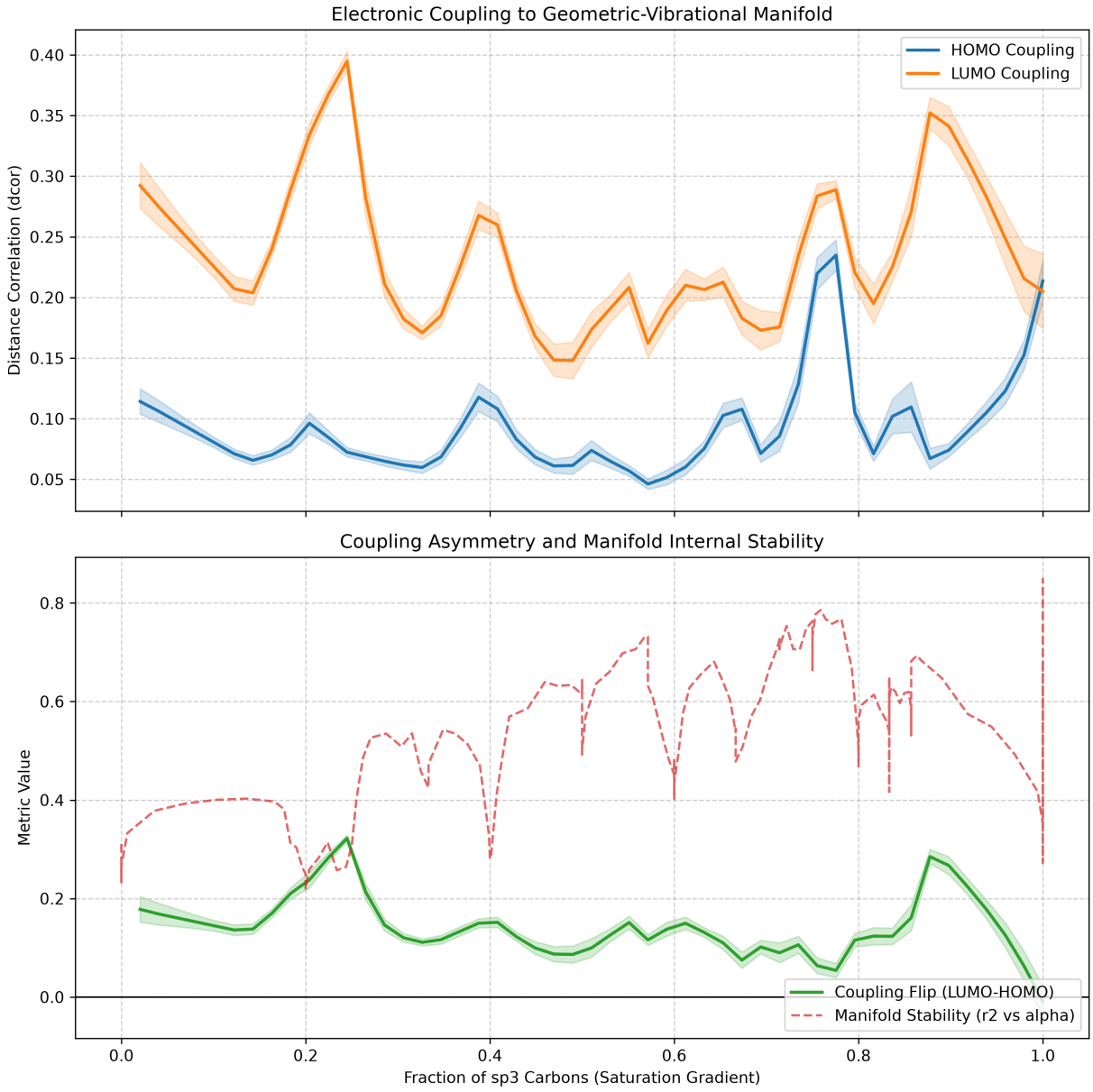

Figure 2: Distance correlation ($dCor$) trajectories of electronic properties against the geometric-vibrational manifold across the saturation gradient (fraction of sp3 carbons). The top panel illustrates the coupling of HOMO and LUMO residuals to the primary non-linear component, `geom_kPC1`. The LUMO consistently maintains stronger coupling to the global molecular state, whereas the HOMO exhibits profound detachment, reaching a minimum coupling at approximately 0.57 saturation. The bottom panel displays the "Coupling Flip" ($\Delta dCor = dCor_{LUMO} - dCor_{HOMO}$) in green, which peaks in asymmetry near 0.24 saturation. The red dashed line tracks the manifold's internal stability through the distance correlation between spatial extent (r^2) and polarizability (α) residuals, showing that the link between electron density and vibrational states fluctuates based on molecular topology. Shaded confidence intervals from Leave-One-Scaffold-Out validation confirm that these topological re-organizations and the transition from asymmetric to near-symmetric coupling states are robust, fundamental properties of the chemical space.

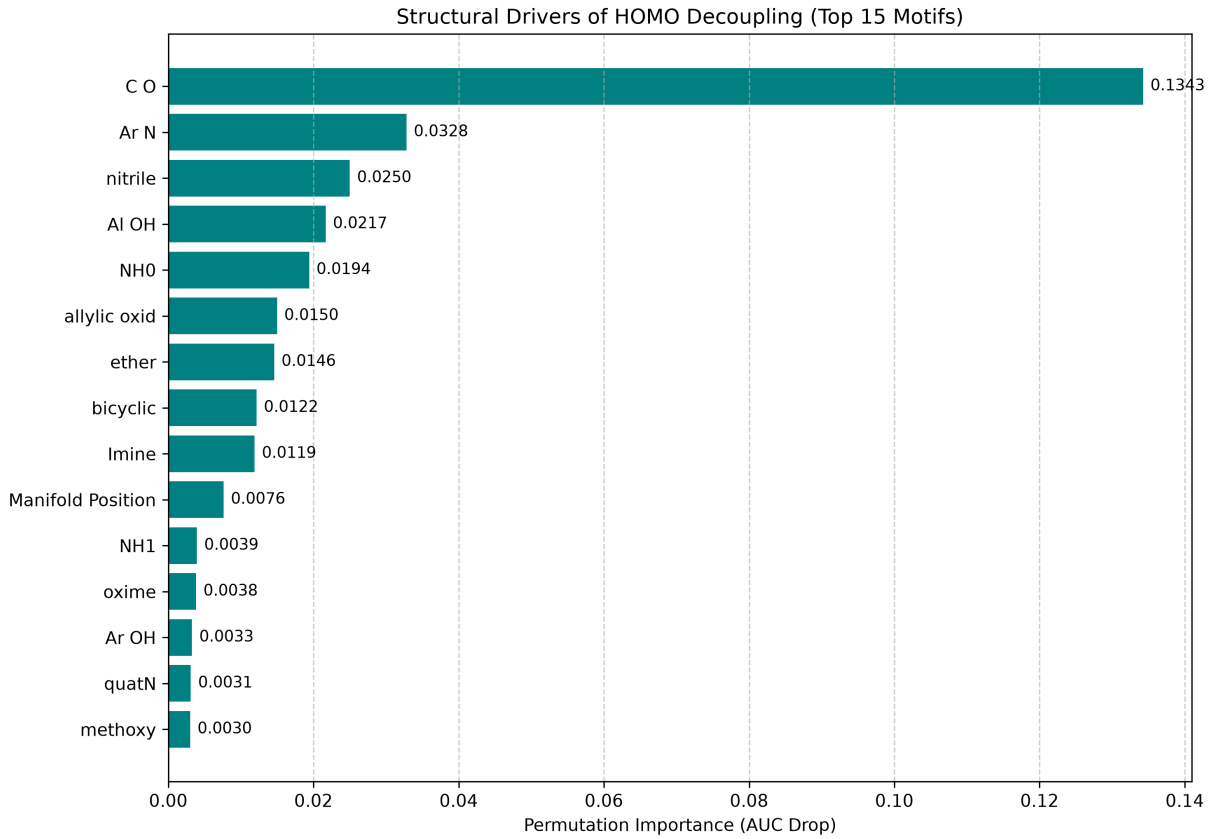

Figure 3: Permutation importance (AUC drop) of the top 15 structural motifs driving HOMO decoupling, evaluated via a Gradient Boosting classifier. The carbonyl group (`fr_C_O`) is the most significant driver with an importance of 0.1343, followed by aromatic nitrogen (`fr_Ar_N`) and nitriles (`fr_nitrile`). The molecule's position on the spatial-vibrational manifold (`manifold_coord`) ranks 10th (0.0076). This reveals that HOMO decoupling is not a simple function of molecular size, but is explicitly driven by specific functional groups that isolate frontier electron density from the global geometric state.

residuals and extracting a non-linear geometric-vibrational manifold, we provided a stable structural reference frame that is independent of extensive size-scaling effects and localized functional group variance.

Our analysis of the QM9 dataset along a continuous saturation gradient revealed a fundamental asymmetry in orbital behavior, which we have termed the "coupling flip." While the LUMO maintains a relatively strong connection to the global molecular geometry—likely due to its more diffuse spatial nature—the HOMO exhibits a profound detachment. At intermediate saturation levels, the HOMO energy becomes nearly orthogonal to the molecule's macroscopic spatial and thermodynamic state.

Crucially, we demonstrated that this orthogonalization is not a simple consequence of molecular size or bulk geometric scaling. Motif analysis identified specific localized functional groups—most notably carbonyls, aromatic nitrogens, and nitriles—as the primary structural drivers that insulate frontier electron density from the global molecular manifold. The robustness of these topological transitions was confirmed through rigorous leave-one-scaffold-out validation, establishing them as fundamental properties of the chemical space rather than artifacts of specific molecular families.

Furthermore, our boundary case analysis of linear molecules highlighted how 1D geometric constraints fundamentally alter internal manifold organization, tightening the coupling between electronic spatial extent and polarizability while exacerbating the decoupling of vibrational states. Ultimately, mapping these topological thresholds provides a deeper understanding of how local electronic environments interact with global molecular architecture. By rigorously quantifying the magnitude of electronic-thermodynamic orthogonalization, this framework offers valuable insights for future computational molecular design and the prediction of complex chemical reactivity.