Conditional Mutual Information Mapping of HOMO Integration and Frontier Orbital Decoupling

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energy of the Highest Occupied Molecular Orbital (HOMO) often appears decoupled from macroscopic spatial and thermodynamic properties in small organic molecules. To determine if this electronic isolation is structurally contingent, we investigate the architectural conditions that mediate frontier orbital coupling using the QM9 dataset of 133,802 molecules. We isolate genuine chemical variance from compositional size effects via polynomial residualization and construct a consolidated spatial-thermodynamic manifold using Principal Component Analysis. By applying partial correlation analysis, Conditional Mutual Information, and k-nearest neighbor distance correlation mapping, we quantify the dependence of frontier orbitals on specific structural features. Our results indicate that while π-system extension drives frontier orbital variation, the lowest unoccupied molecular orbital (LUMO) coordinates more strongly with the global manifold than the HOMO. Crucially, we identify a distinct mode of integration in saturated, non-aromatic systems, where strained architectures—such as three- and four-membered rings—exhibit significantly higher coupling than unstrained acyclic counterparts. Validated through a leave-one-scaffold-out approach, these findings demonstrate that frontier electronic isolation is not absolute but is governed by specific architectural triggers, emphasizing the role of structural strain in coupling electronic states to global molecular physics.

1 Introduction

The frontier molecular orbitals, specifically the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO), govern the chemical reactivity, stability, and optical properties of organic molecules. In large-scale computational datasets such as QM9, a persistent observation is that the HOMO energy often appears statistically decoupled from macroscopic spatial and thermodynamic property manifolds. This apparent electronic isolation complicates the development of unified predictive models and raises fundamental questions about the relationship between global molecular physics and localized electronic states.

It remains unclear whether this decoupling is an intrinsic feature of frontier orbitals or a structurally contingent phenomenon driven by specific architectural features. Traditional analyses often conflate genuine chemical variance with extensive compositional size effects, obscuring the subtle topological triggers that might mediate orbital coupling. Understanding the precise structural conditions under which frontier electronic states integrate with or decouple from the global spatial-thermodynamic manifold is essential for rational molecular design and the interpretation of large chemical spaces.

To determine if this electronic isolation is structurally contingent, we investigate the architectural conditions that mediate frontier orbital coupling across 133,802 unique molecules from the QM9 dataset. We employ a rigorous statistical framework to isolate genuine chemical variance from stoichiometric size effects via polynomial residualization. By constructing a consolidated spatial-thermodynamic manifold using Principal Component Analysis (PCA), we establish a baseline of global molecular properties. We then deploy partial correlation analysis, Conditional Mutual Information (CMI), and k-nearest neighbor distance correlation mapping to quantify the dependence of frontier orbitals on specific structural features, such as π-system extension, heteroatom presence, and ring strain.

Our findings demonstrate that frontier electronic isolation is not absolute. While π-system extension is a primary driver of frontier orbital variation, we find that the LUMO coordinates more strongly with the global

manifold than the HOMO. Crucially, by inverting the search to identify highly coupled subpopulations, we uncover a distinct mode of integration in saturated, non-aromatic systems. Strained architectures, particularly three- and four-membered rings, exhibit significantly higher coupling than their unstrained acyclic counterparts. Validated through a leave-one-scaffold-out approach, these results emphasize the critical role of structural strain in forcing electronic states to couple with global molecular physics, providing a nuanced understanding of how molecular architecture dictates electronic behavior.

2 Methods

2.1 Data curation and structural feature engineering

To investigate the structural conditions mediating frontier orbital coupling, we utilized the QM9 dataset. To prevent data leakage during structural validation, we first deduplicated the dataset by removing redundant SMILES strings, yielding 133,802 unique molecules. The 19 original numeric properties were filtered to 12 non-redundant features by excluding thermodynamic and atomization restatements, as well as the HOMO-LUMO gap, which is a direct linear combination of the frontier orbital energies.

To handle the extreme heavy-tailed distribution of the rotational constant A, we applied a $\log(A+1)$ transformation. Furthermore, to prevent near-collinearity in downstream manifold construction, the highly correlated rotational constants B and C were averaged into a single feature, as they are identical for the nine linear molecules in the dataset. Molecular structures were parsed to extract exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen. Targeted structural features were engineered to quantify architectural triggers: conjugation length, aromatic ring count, the fraction of sp^2-hybridized carbons, specific heteroatom counts, and Murcko scaffolds.

2.2 Compositional residualization and manifold consolidation

Extensive compositional size effects often obscure genuine topological coupling in large chemical spaces. To isolate structural variance from simple stoichiometric scaling, we fitted a second-degree polynomial regression—incorporating both interaction and squared terms—using strictly the five atom counts to predict each of the transformed, non-redundant properties. The residuals from these models (e.g., *res_homo*, *res_alpha*) represent the size-independent chemical variance. Cook's Distance was computed for these fits to flag and exclude molecules where the baseline model fit poorly, ensuring that downstream outliers were not merely regression artifacts.

The extracted residuals were z-score standardized to prevent variables with larger absolute variances from artificially dominating distance metrics. We then performed Principal Component Analysis (PCA) on the standardized residuals to collapse redundant dimensions into a consolidated, orthogonal spatial-thermodynamic manifold, retaining the principal components necessary to explain 95% of the variance. A sensitivity check was conducted by computing the PCA components both with and without the linear molecules to ensure the resulting manifold was not biased by extreme geometries.

2.3 Partial correlation and conditional mutual information

To quantitatively isolate whether frontier electronic isolation is driven by π-system extension or general molecular complexity, we performed a targeted sensitivity analysis. We computed the partial correlation between *res_homo* and the engineered structural descriptors, strictly controlling for the consolidated principal components of the other property residuals.

We then deployed a Conditional Mutual Information (CMI) framework to test whether the dependence of *res_homo* on polarizability (*res_alpha*) changes as a function of specific architectural triggers. Structural features, such as the fraction of sp^2 carbons, were discretized into bins with balanced sample sizes. The CMI between *res_homo* and *res_alpha* was computed conditioned on these structural bins. To ensure physical specificity, the LUMO energy residual (*res_lumo*) was included as a continuous control and evaluated under identical conditions, testing whether topological integration is unique to the HOMO or coordinated across the frontier orbital gap. A rigorous null baseline was established via a permutation test, shuffling the residual

vectors 1,000 times strictly within each structural bin to preserve the marginal distributions and generate empirical p-values.

2.4 Local distance correlation and scaffold-aware validation

Because the HOMO energy globally resists integration into the joint property structure, we inverted the search to identify specific molecular subpopulations where *res_homo* actively correlates with the other properties. We defined local chemical neighborhoods using a k-nearest neighbors (k-NN) approach ($k = 400$) in the structural descriptor space. Within these local neighborhoods, we computed the Distance Correlation (dCor) between *res_homo* and the consolidated PCA manifold.

Populations exhibiting statistically elevated dCor values compared to the global baseline were isolated for mechanistic analysis. To systematically evaluate the role of molecular strain in these coupled subpopulations, the dataset was partitioned into three distinct structural regimes: saturated strained (three- and four-membered rings), saturated unstrained (acyclic and large rings), and conjugated/aromatic systems. Finally, the discovered structural rules were validated using a strict leave-one-scaffold-out (LOGO) approach based on Murcko scaffolds. This ensured that the observed coupling behaviors, particularly in strained architectures, were general physical phenomena rather than artifacts driven by isolated structural families.

3 Results

3.1 Dimensionality of the chemical residual manifold

To isolate genuine chemical variance from extensive compositional size effects, molecular properties were residualized against a second-order polynomial of atom counts (C, N, O, F, H). Principal Component Analysis (PCA) on the resulting 12-dimensional residual manifold demonstrates that the global spatial-thermodynamic structure is highly consolidated. We find that seven principal components are sufficient to explain over 97% of the cumulative variance in the dataset. The first two components, PC1 and PC2, account for 31.9% and 17.2% of the variance, respectively. Sensitivity analysis confirms that this manifold structure is robust to the exclusion of the nine linear molecules in the dataset, indicating that the baseline joint property structure is driven by fundamental chemical variance rather than extreme geometric outliers.

3.2 Targeted sensitivity of HOMO coupling to structural features

To determine which architectural features mediate the relationship between the HOMO energy residual (*res_homo*) and the global molecular manifold, we performed a partial correlation analysis controlling for the seven consolidated principal components. The results reveal that π-system extension is a primary driver of frontier orbital variation (). Specifically, the number of aromatic rings ($r = 0.309$), the fraction of sp^2-hybridized carbons ($r = 0.289$), and the number of conjugated bonds ($r = 0.200$) exhibit the strongest partial correlations with *res_homo*.

In contrast, the partial correlations for heteroatom counts (nitrogen, oxygen, and fluorine) are near zero ($|r| < 10^{-13}$). This suggests that the mere presence of electronegative atoms does not inherently rewire the HOMO's relationship to the global manifold; instead, the architectural arrangement of carbon hybridization dictates this coupling.

3.3 Conditional mutual information and frontier coordination

To rigorously quantify whether the dependence of frontier orbitals on the spatial-thermodynamic manifold is statistically significant when conditioned on these structural features, we computed the Conditional Mutual Information (CMI). We evaluated the coupling of both *res_homo* and the LUMO energy residual (*res_lumo*) to polarizability (*res_alpha*), conditioned on structural bins such as the fraction of sp^2 carbons.

Permutation tests ($N = 1000$) establish that the integration of both frontier orbitals is statistically significant ($p < 0.001$), as shown in . The observed CMI for the HOMO is 0.0319 nats, whereas the LUMO exhibits a substantially higher CMI of 0.1008 nats. This disparity indicates that the lowest unoccupied orbital

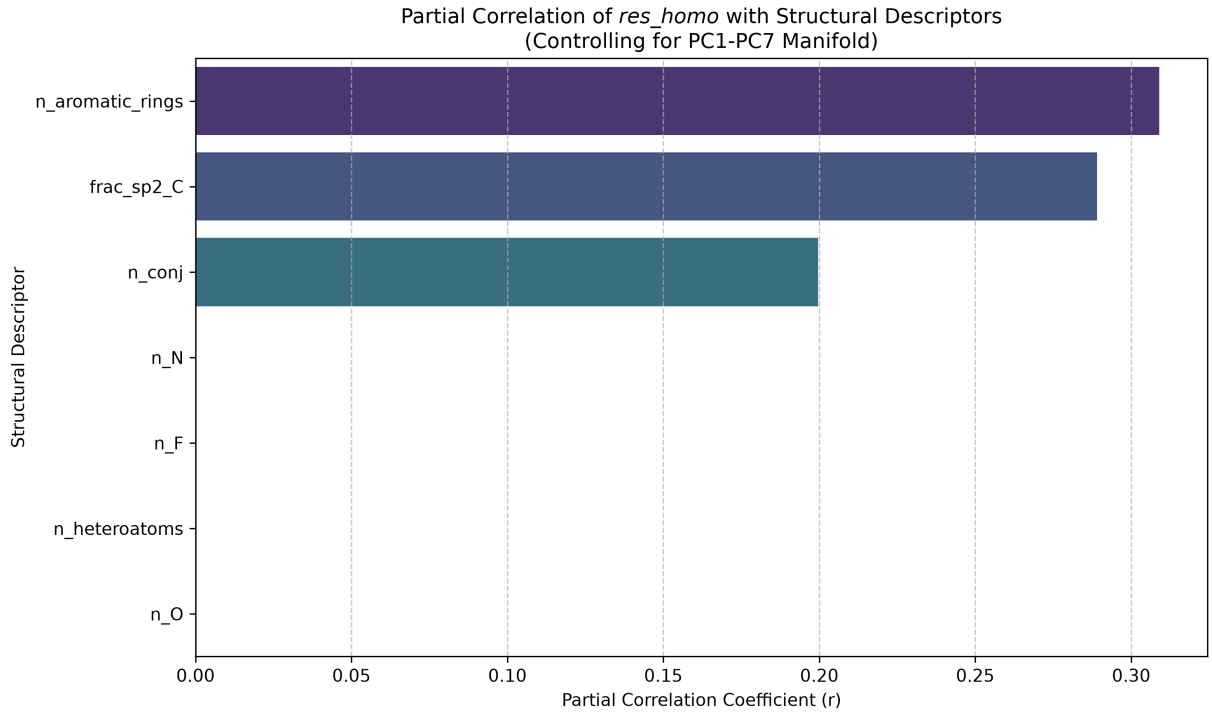

Figure 1: Partial correlation coefficients of HOMO energy residuals (`res_homo`) with specific structural descriptors, controlling for the 7-component spatial-thermodynamic manifold (PC1-PC7). The plot reveals that π-system extension features—specifically aromaticity (`n_aromatic_rings`, $r = 0.309$), carbon hybridization (`frac_sp2_C`, $r = 0.289$), and conjugation (`n_conj`, $r = 0.200$)—are the primary drivers of frontier orbital variation. Conversely, heteroatom counts (`n_N`, `n_F`, `n_heteroatoms`, and `n_O`) exhibit near-zero partial correlations. This demonstrates that the architectural arrangement of carbon hybridization, rather than the mere presence of nitrogen or oxygen, is the controlling factor in rewiring the relationship between the HOMO and the global molecular manifold.

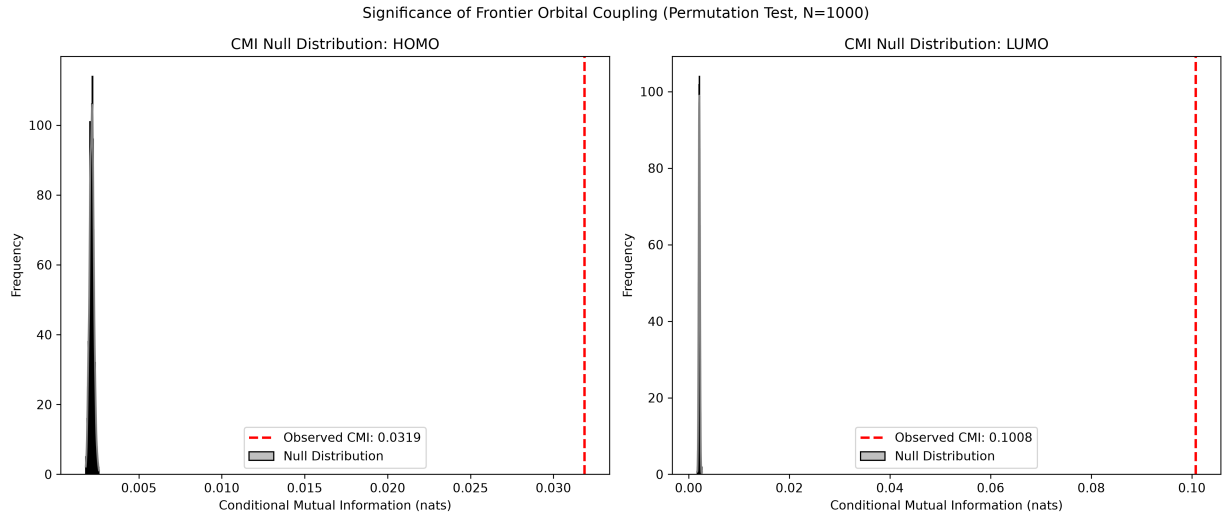

Figure 2: Permutation test results ($N = 1000$) evaluating the statistical significance of Conditional Mutual Information (CMI) for HOMO (left) and LUMO (right) coupling to the structural-spatial manifold. The histograms represent the null distributions, while the red dashed lines indicate the observed CMI values of 0.0319 nats and 0.1008 nats, respectively. Both observed values fall far outside their null distributions, demonstrating statistically significant integration ($p < 0.001$). The substantially higher observed CMI for the LUMO reveals that the lowest unoccupied orbital is more strongly coordinated with the spatial-thermodynamic manifold than the HOMO, even after accounting for stoichiometric effects.

is more strongly coordinated with the structural-spatial manifold than the HOMO, even after rigorously accounting for stoichiometric size effects.

3.4 Characterization of highly coupled subpopulations

Given that the HOMO energy residual is often decoupled from the global joint property structure, we sought to identify specific local chemical neighborhoods where this coupling is elevated. Using a k-nearest neighbor (k-NN) approach ($k = 400$) in the structural descriptor space, we computed local distance correlations ($dCor$) between res_homo and the PCA manifold. While the global baseline $dCor$ for a random sample of this size is 0.1277, the mean local $dCor$ across the descriptor space is 0.4020, reaching a maximum of 0.8372 in the most integrated neighborhoods.

Comparing the structural profiles of the 70 most strongly coupled anchor neighborhoods against the global population reveals a stark contrast (). The highly coupled subpopulations are almost entirely devoid of aromaticity (mean aromatic rings = 0.000 vs. 0.136 globally) and exhibit negligible conjugation (mean conjugated bonds = 0.043 vs. 1.707) and sp^2 carbon fractions (0.004 vs. 0.216). Furthermore, these coupled islands show a slight enrichment in nitrogen (1.000 vs. 0.938) and a depletion in oxygen (0.986 vs. 1.451) relative to the global means. This structural profile demonstrates that res_homo integrates most strongly with global physical properties in saturated, non-aromatic systems, challenging the assumption that delocalized π-systems are the sole mechanism for frontier orbital coupling.

3.5 The role of structural strain in frontier orbital integration

To systematically evaluate the hypothesis that molecular strain modulates this unexpected coupling in saturated systems, we partitioned the dataset into three distinct structural regimes: saturated strained (containing three- and four-membered rings), saturated unstrained (acyclic and larger rings), and conjugated/aromatic systems.

As illustrated in , while conjugated and aromatic systems exhibit the highest overall coupling ($dCor = 0.2727$, $N = 103,635$), saturated strained systems display significantly elevated coupling ($dCor = 0.2540$,

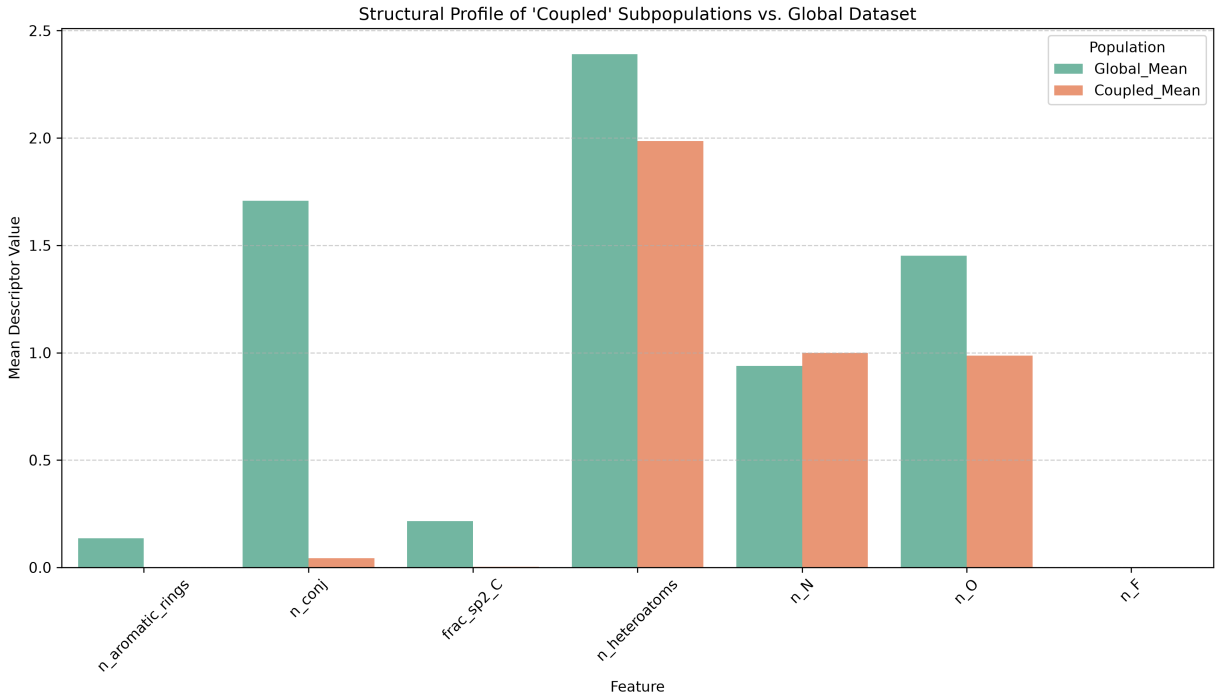

Figure 3: Comparison of mean structural descriptor values (`Global_Mean` versus `Coupled_Mean`) for the QM9 dataset. The bar chart illustrates that the highly coupled subpopulations possess near-zero values for `n_aromatic_rings`, `n_conj`, and `frac_sp2_C`, demonstrating a stark lack of aromaticity and conjugation compared to the global population. Additionally, while overall `n_heteroatoms` and oxygen (`n_O`) are depleted in the coupled groups, nitrogen (`n_N`) is slightly enriched. From this structural profile, we learn that the HOMO energy residual (`res_homo`) integrates most strongly with the global spatial-thermodynamic manifold in saturated, non-aromatic systems, contrary to the expectation that delocalized π-systems provide the primary coupling mechanism.

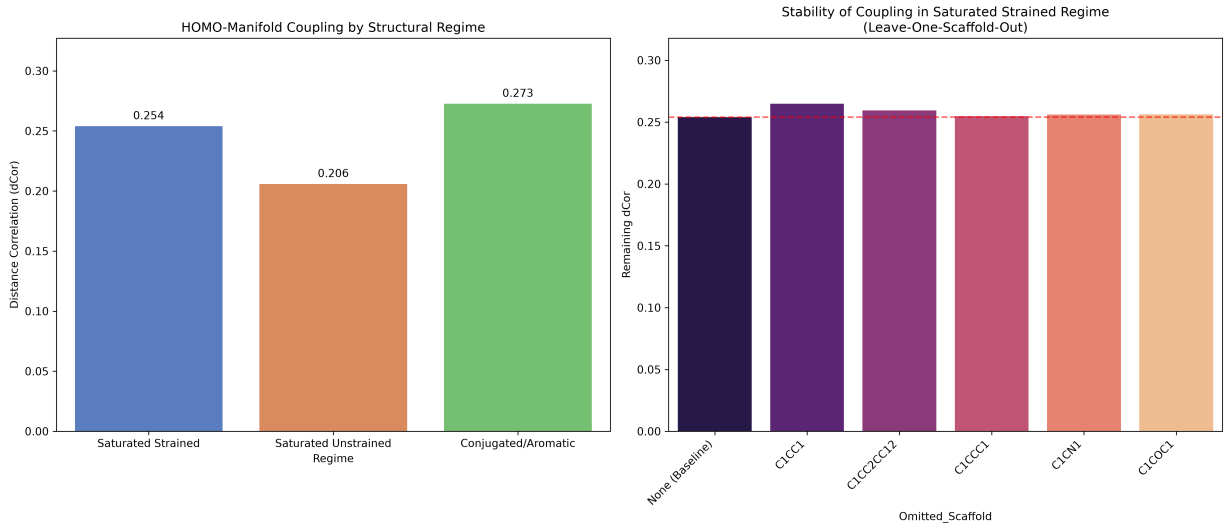

Figure 4: Distance correlation ($dCor$) between HOMO energy residuals and the spatial-thermodynamic manifold across distinct molecular architectures. The left panel demonstrates that while conjugated and aromatic systems exhibit the highest coupling ($dCor = 0.273$), saturated strained systems display significantly elevated coupling ($dCor = 0.254$) compared to unstrained saturated systems ($dCor = 0.206$). The right panel presents a Leave-One-Scaffold-Out validation within the saturated strained regime, showing that omitting common scaffolds (such as `C1CC1` or `C1CC2CC12` on the `Omitted_Scaffold` axis) maintains a stable $dCor$ near the baseline. These results reveal that frontier orbital integration is structurally contingent, demonstrating that the robust coupling observed in strained architectures is a general property of molecular strain rather than an artifact of specific scaffold families.

$N = 27,261$) compared to their unstrained saturated counterparts ($dCor = 0.2059$, $N = 2,906$). The coupling in strained architectures is approximately 1.23 times higher than in flexible, acyclic saturated molecules. Strikingly, 70% of the highly coupled anchor molecules identified in the k-NN analysis belong to the saturated strained regime, despite this class comprising only about 20% of the total dataset.

To ensure these findings are not artifacts of specific, overrepresented structural families, we performed a Leave-One-Scaffold-Out (LOGO) validation within the strained regime. Omitting highly frequent scaffolds, such as cyclopropane (`C1CC1`) or bicyclo[1.1.1]pentane (`C1CC2CC12`), does not collapse the observed correlation; the $dCor$ remains highly stable between 0.254 and 0.265 (, right panel). This confirms that the robust integration of the HOMO into the spatial-thermodynamic manifold is a general physical consequence of structural strain, forcing electronic states to couple with global molecular properties in a manner absent in unstrained saturated systems.

4 Conclusions

In this study, we demonstrated that the apparent electronic isolation of the Highest Occupied Molecular Orbital (HOMO) in large chemical spaces is not an absolute intrinsic property, but rather a structurally contingent phenomenon. By rigorously isolating genuine chemical variance from compositional size effects through polynomial residualization and manifold consolidation, we established a robust framework for quantifying the dependence of frontier orbitals on specific architectural triggers across the QM9 dataset.

Our targeted sensitivity and Conditional Mutual Information analyses revealed that while π-system extension—driven by aromaticity and carbon hybridization—is a primary mediator of frontier orbital variation, the frontier orbitals do not behave uniformly. The Lowest Unoccupied Molecular Orbital (LUMO) coordinates significantly more strongly with the global spatial-thermodynamic manifold than the HOMO, even after accounting for stoichiometric scaling. Furthermore, the mere presence of electronegative heteroatoms does not inherently rewire this electronic-structural relationship, pointing to the topological arrangement of

the carbon skeleton as the fundamental driver.

Crucially, by inverting our search to map highly coupled local chemical neighborhoods, we identified a distinct and unexpected mode of integration in saturated, non-aromatic systems. We found that structural strain, particularly within three- and four-membered rings, forces the HOMO to couple with global molecular physics ($dCor = 0.254$) to a degree that is significantly elevated compared to flexible, unstrained acyclic molecules ($dCor = 0.206$). The stability of this coupling, confirmed via rigorous leave-one-scaffold-out validation, establishes that this behavior is a general physical consequence of molecular strain rather than a statistical artifact of specific, overrepresented scaffold families.

Ultimately, these findings provide a nuanced understanding of how molecular architecture dictates electronic behavior. Recognizing that frontier orbital decoupling is governed by specific topological triggers—most notably structural strain and π-delocalization—offers critical insights for rational molecular design. It highlights the necessity of structurally aware approaches in the development of predictive models, ensuring that the subtle physical rules governing localized electronic states and global molecular properties are accurately captured in large chemical spaces.