Mapping the Structural Drivers of Frontier Orbital Decoupling in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In molecular property datasets, the energy of the highest occupied molecular orbital (HOMO) uniquely decouples from the global thermodynamic and spatial manifold once basic stoichiometric effects are removed. To investigate the structural conditions governing this phenomenon, we analyzed 133,802 deduplicated molecules from the QM9 dataset to determine whether this orthogonality is a universal property or a structurally dependent state. We applied stoichiometric residualization to isolate purely structural variance and computed multivariate distance correlations between HOMO energy residuals and a global manifold proxy, defined by the first principal component of zero-point vibrational energy, heat capacity, and electronic spatial extent. By partitioning the chemical space using graph-derived structural descriptors, we evaluated coupling gradients across different molecular environments, validating our findings through bootstrap resampling and scaffold hold-outs. Our results indicate that frontier orbital decoupling is primarily driven by electron delocalization, as coupling strength drops significantly in highly conjugated systems. Conversely, the presence of peripheral heteroatoms or extreme geometric constraints, such as molecular linearity, forces the localized electronic state to align more closely with the macroscopic physical manifold. These findings demonstrate that the decoupling between electronic and thermodynamic states is not uniform across chemical space, but is systematically modulated by local topology and geometric boundaries.

1 Introduction

The exploration of chemical space through large-scale computational datasets, such as QM9, has fundamentally advanced our understanding of structure-property relationships. In such datasets, the dominant source of variance for most molecular properties is trivial molecular size and stoichiometry. When these first-order compositional effects are mathematically removed, the underlying structural dependencies of the molecules are revealed. Among these residualized properties, a robust joint structure typically emerges between macroscopic thermodynamic and spatial features, forming a tightly coupled global physical manifold. However, the energy of the highest occupied molecular orbital (HOMO) exhibits a striking anomaly: it uniquely decouples from this global manifold. While macroscopic properties remain highly correlated with one another, the HOMO energy appears largely orthogonal to the molecule's broader spatial and thermodynamic state.

This phenomenon raises a fundamental question regarding the behavior of frontier orbitals across diverse chemical environments. Is the orthogonality of the HOMO energy a universal quantum mechanical feature of local electronic independence, or is it a structurally dependent state modulated by specific chemical motifs? Because HOMO variance is primarily dictated by local bonding topology rather than macroscopic architecture, understanding the boundary conditions of this decoupling is critical. It provides insight into when localized electronic states operate independently of the global molecular structure and when they are forced into alignment with the macroscopic physical state.

To investigate the structural drivers governing this phenomenon, we analyzed 133,802 deduplicated molecules from the QM9 dataset. We applied stoichiometric residualization to regress out the effects of atom counts (C, N, O, F, H), isolating purely structural variance to yield residualized properties such as *res_homo*, *res_zpve*, *res_cv*, and *res_r2*. To quantify the baseline physical state of each molecule, we defined a global manifold proxy using the first principal component of a highly consistent spatial-vibrational

triad: zero-point vibrational energy, heat capacity, and electronic spatial extent. We then employed multivariate distance correlation ($dCor$) to measure the non-linear coupling between the HOMO energy residuals and this global manifold proxy.

Rather than treating the dataset as a single homogeneous entity, we partitioned the chemical space using interpretable, graph-derived structural descriptors. By clustering molecules based on their conjugation extent, heteroatom proximity, and geometric boundary conditions, we evaluated how the $dCor$ coupling gradient fluctuates across different local environments.

In this work, we demonstrate that frontier orbital decoupling is not a uniform property across chemical space, but rather a structurally dependent phenomenon. We show that electron delocalization acts as the primary driver of decoupling, whereas peripheral heteroatoms and extreme geometric constraints, such as molecular linearity, act as structural bridges that force the HOMO to align with the global thermodynamic and spatial manifold. Through rigorous statistical validation, including bootstrap resampling and scaffold hold-outs, we map the specific topological conditions under which the fundamental divide between electronic and physical molecular states either persists or breaks down.

2 Methods

2.1 Data preparation and stoichiometric residualization

To isolate the purely structural variance of molecular properties from trivial compositional effects, we utilized the QM9 dataset. The dataset was first filtered to remove 83 duplicate SMILES strings, preventing data leakage during downstream validation and yielding a final set of 133,802 deduplicated molecules. Prior to regression, a $\log(A + 1)$ transformation was applied to the rotational constant A to tame its extreme heavy-tailed distribution, ensuring it did not disproportionately dominate variance-based metrics.

Stoichiometric residualization was performed by fitting ordinary linear regression models for 12 non-redundant molecular properties against the raw atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen. The residuals from these regressions were extracted to represent the size- and composition-independent structural variance of each property. This procedure yielded residualized vectors for the HOMO energy (res_homo), LUMO energy (res_lumo), zero-point vibrational energy (res_zpve), heat capacity (res_cv), and electronic spatial extent (res_r2), among others.

2.2 Computation of structural descriptors

To partition the chemical space into interpretable local environments, structural descriptors were computed directly from the molecular graphs using RDKit. Global logging for RDKit was silenced to optimize parallel processing. For each molecule, we derived the following features:

- **Conjugation index:** Calculated as the sum of aromatic rings and conjugated double bonds, providing a quantitative measure of electron delocalization.
- **Heteroatom proximity and density:** To robustly proxy how buried or exposed heteroatoms are within the molecular topology, we computed the average topological shortest-path distance from all heteroatoms (N, O, F) to the topological center of the molecule (defined as the atom with the highest degree). The total count of heteroatoms was also recorded to distinguish isolated heteroatoms from dense heteroatomic environments.
- **Symmetry proxy:** Topological symmetry classes were derived directly from the molecular graph. This graph-based approach was chosen over computed physical properties (such as the dipole moment μ or rotational constants A, B, C) to prevent circularity when defining structural partitions.

2.3 Definition of the global manifold and distance correlation

Following residualization, we verified the internal consistency of a spatial-vibrational triad comprising res_zpve, res_cv, and res_r2 by computing their pairwise distance correlation ($dCor$) matrix. To simplify the gradient analysis and establish a single baseline physical state, we extracted the first principal component (PC1) of this triad. This PC1 vector served as the "Global Manifold" proxy.

To quantify the non-linear coupling between the frontier orbitals and the macroscopic physical state, we computed the multivariate distance correlation between the *res_homo* vector and the Global Manifold proxy. The same metric was computed for *res_lumo* to determine whether the observed decoupling was uniquely a HOMO phenomenon or a broader characteristic of frontier orbitals. Distance correlation was selected over Pearson correlation to capture both linear and non-linear dependencies between the highly localized electronic states and the global manifold.

2.4 Structural partitioning and geometric boundary analysis

To evaluate how the HOMO-manifold coupling fluctuates across different chemical environments, the dataset was partitioned into distinct subsets based on the computed structural descriptors. Molecules were binned by their conjugation index and heteroatom proximity. Within each subset, the internal $dCor$ stability of the spatial-vibrational triad was first verified to ensure the baseline manifold remained intact. The $dCor$ between *res_homo* and the Global Manifold proxy was then calculated for each partition, allowing us to analyze the gradient of coupling strength as a function of specific structural motifs.

Furthermore, we isolated geometric boundary conditions to test the limits of HOMO decoupling. Linear molecules were identified by a rotational constant of $A = 0$. For this specific subset, A was removed from the feature set, as it acts as a constant and $B = C$. The remaining non-linear molecules were partitioned based on the graph-derived symmetry proxy and raw dipole moment (μ) to contrast highly symmetric, non-polar molecules against the polar baseline.

2.5 Statistical validation and computational details

Given the large sample size of the dataset, p-values are generally uninformative; therefore, our analysis strictly relied on absolute $dCor$ values and their effect sizes (gradients) across structural partitions. To ensure the statistical robustness of the observed coupling gradients, we employed bootstrap resampling within each structural partition to generate 95% confidence intervals for the $dCor$ estimates.

To confirm that the observed structural dependencies were not artifacts of overrepresented molecular families, we performed a hold-out validation based on Murcko scaffolds. The dataset was grouped by scaffold, and the top 10 most frequent scaffolds were systematically excluded from the bins. The coupling metrics were then re-evaluated to verify that the relative gradients persisted independently of the dominant scaffolds in the QM9 dataset. All computationally intensive tasks, including the distance matrix computations for $dCor$ and the bootstrap resampling procedures, were parallelized across 14 CPUs using the `joblib` library.

3 Results

3.1 Stoichiometric residualization and the global manifold

The organization of molecular properties in the QM9 dataset is fundamentally dominated by molecular size and stoichiometry. Prior to residualization, extensive properties such as zero-point vibrational energy (`zpve`) and heat capacity (`cv`) exhibit near-perfect linear dependence on basic atom counts (C, N, O, F, H). In contrast, the variance of the HOMO energy (`homo`) is primarily dictated by local bonding topology and shows minimal dependence on these first-order compositional effects. This stoichiometric determinism is illustrated in Figure 1, which justifies the necessity of residualizing these properties to isolate purely structural drivers.

Once these trivial size and composition effects are mathematically removed, a robust joint structure emerges among the thermodynamic and spatial properties. The spatial-vibrational triad—comprising the residuals of zero-point vibrational energy (`res_zpve`), heat capacity (`res_cv`), and electronic spatial extent (`res_r2`)—maintains high internal coherence across the dataset. Pairwise distance correlations confirm this tight coupling, notably yielding $dCor(\texttt{res_zpve}, \texttt{res_cv}) = 0.832$ and $dCor(\texttt{res_cv}, \texttt{res_r2}) = 0.552$.

However, when comparing the frontier orbital energies to the first principal component (PC1) of this triad—our proxy for the global physical manifold—a striking decoupling is observed. As shown in Figure 2, the coupling between the HOMO energy residuals (`res_homo`) and the global manifold is remarkably weak ($dCor = 0.259$). The LUMO residuals (`res_lumo`) exhibit a slightly higher, yet still moderate, correlation

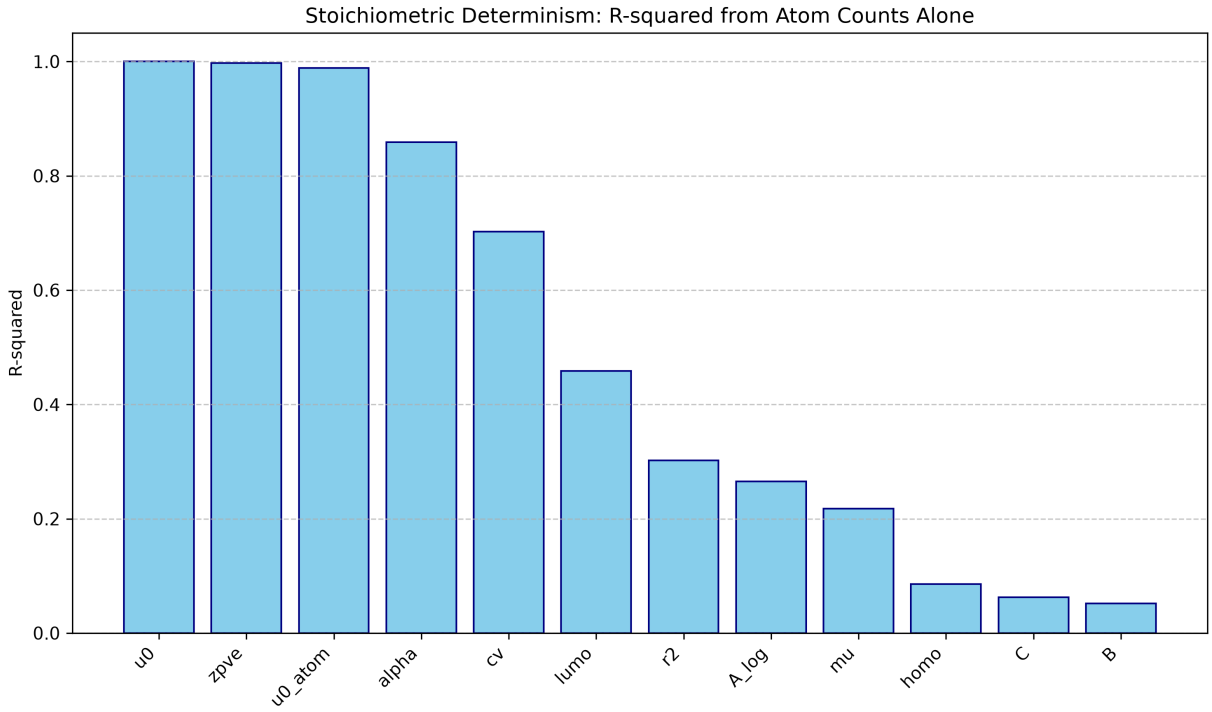

Figure 1: Coefficient of determination (R^2) for twelve molecular properties predicted solely from atom counts (C, N, O, F, H). The plot illustrates the first-order stoichiometric determinism in the QM9 dataset prior to residualization. Extensive properties like zero-point vibrational energy (`zpve`) and heat capacity (`cv`) are highly dominated by stoichiometry, whereas the HOMO energy (`homo`) exhibits minimal dependence on atom counts. This demonstrates that frontier orbital energies are uniquely orthogonal to basic stoichiometric effects, justifying the need to residualize these properties to isolate purely structural drivers within the global manifold.

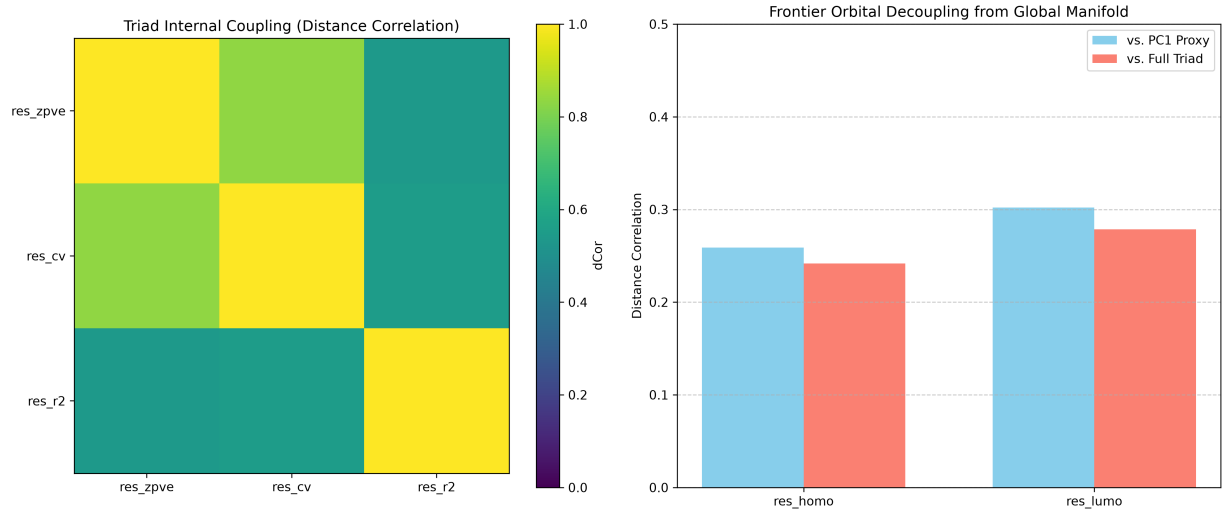

Figure 2: The figure illustrates the distance correlation ($dCor$) metrics defining the global molecular manifold and the decoupling of frontier orbitals. The left panel displays a heatmap of the internal coupling within the spatial-vibrational triad, demonstrating high coherence among the zero-point vibrational energy (`res_zpve`), heat capacity (`res_cv`), and electronic spatial extent (`res_r2`), with strong pairwise correlations such as $dCor = 0.832$ between `res_zpve` and `res_cv`. The right panel compares the coupling of the frontier orbital energy residuals to this global manifold, represented by both the principal component proxy (PC1) and the full triad. The HOMO energy residuals (`res_homo`) exhibit a remarkably weak coupling ($dCor = 0.259$), whereas the LUMO residuals (`res_lumo`) show a slightly higher but still moderate correlation ($dCor = 0.302$). These results demonstrate that the variance of the HOMO energy is uniquely orthogonal to the macroscopic physical state defined by the spatial and vibrational characteristics of the molecules.

($dCor = 0.302$). These results confirm that the variance of the HOMO energy is uniquely orthogonal to the macroscopic physical state defined by the molecule's spatial and vibrational extent.

3.2 Structural drivers of the coupling gradient

To determine whether this decoupling is a universal quantum mechanical feature or a structurally dependent state, we partitioned the chemical space using graph-derived structural descriptors. The distributions of these key structural properties across the QM9 dataset are detailed in Figure 3. The dataset is predominantly composed of saturated systems, alongside molecules with low symmetry and intermediate heteroatom placements.

By evaluating the coupling gradients across these structural bins, we found that the decoupling of the HOMO is not uniform. Figure 4 illustrates how specific local environments modulate the relationship between the localized electronic state and the global manifold.

First, the extent of electron delocalization acts as a primary driver of decoupling. The coupling between `res_homo` and the global manifold follows a clear downward trend as the conjugation index increases. Saturated systems (Conjugation Index 0) exhibit the highest baseline coupling ($dCor = 0.339$). As conjugation increases to moderate levels (Index 1–2), the coupling drops to $dCor \approx 0.18$–0.19, and reaches a minimum of $dCor = 0.100$ in highly conjugated systems (Index 3). This indicates that in delocalized systems, the HOMO energy is dictated almost entirely by the specific topology of the π-system, further isolating the frontier orbitals from the molecule's macroscopic spatial and vibrational characteristics.

Conversely, the topological placement of heteroatoms (N, O, F) can bridge this divide. Pure hydrocarbons exhibit the lowest coupling ($dCor = 0.148$). However, as heteroatoms are introduced and positioned further toward the molecular periphery, the coupling steadily increases, reaching a maximum of $dCor = 0.333$ for molecules in the most distant topological quantile (`Q4_Far`). Peripheral heteroatoms appear to force the localized electronic state of the HOMO to align more closely with the global physical extent of the molecule.

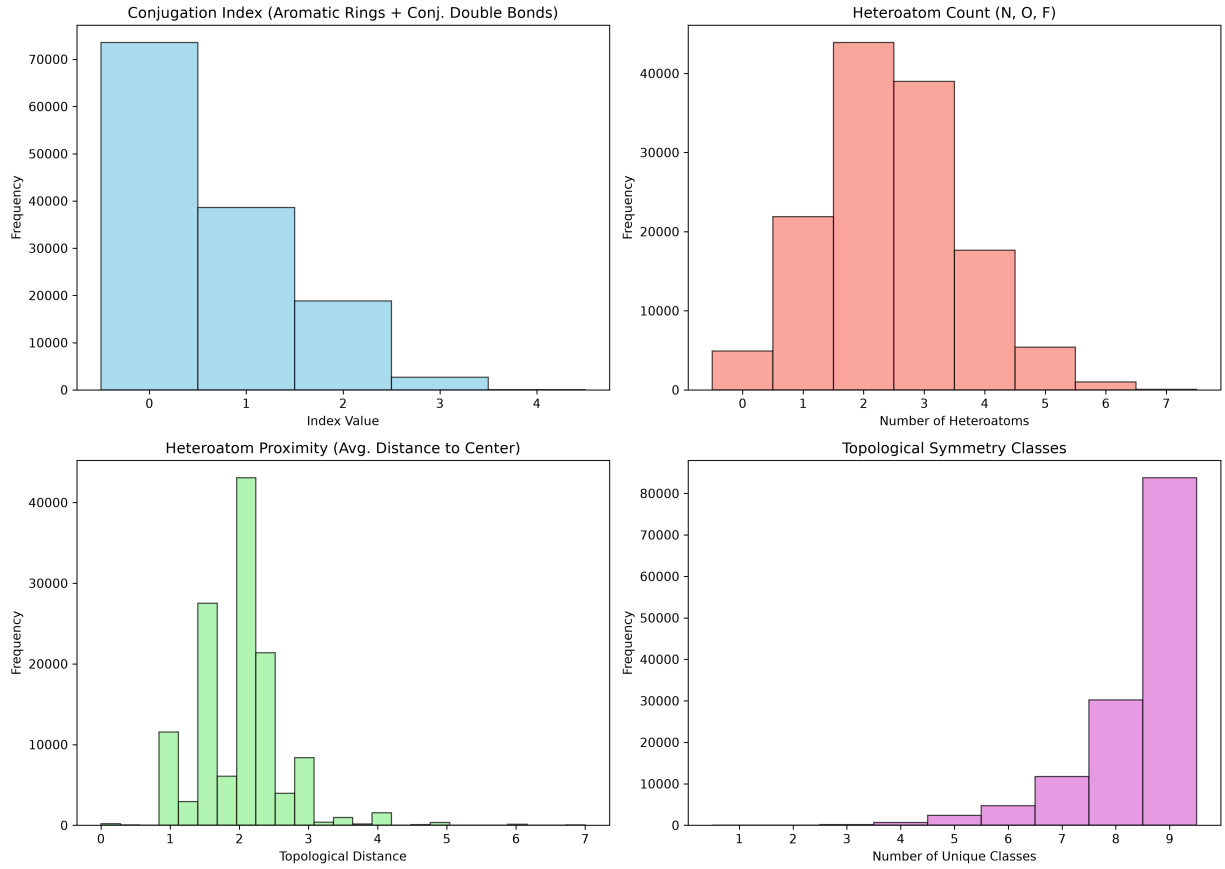

Figure 3: Distributions of key structural properties across the QM9 dataset that modulate the coupling between the HOMO energy (`homo`) and the global thermodynamic-spatial manifold. The panels display the frequencies of the conjugation index, heteroatom count, heteroatom proximity to the topological center, and topological symmetry classes. The dataset is predominantly composed of saturated systems (conjugation index 0), which exhibit the highest coupling, alongside molecules with low symmetry (indicated by a high number of unique topological classes) and intermediate heteroatom placements. Because strong alignment between the localized electronic structure and the macroscopic manifold requires extreme geometric constraints, high symmetry, or highly peripheral heteroatoms, these distributions illustrate why the global baseline coupling remains weak. The relative rarity of these bridging structural conditions ensures that the frontier orbitals remain conspicuously decoupled from the spatial and vibrational extent for the vast majority of organic molecules.

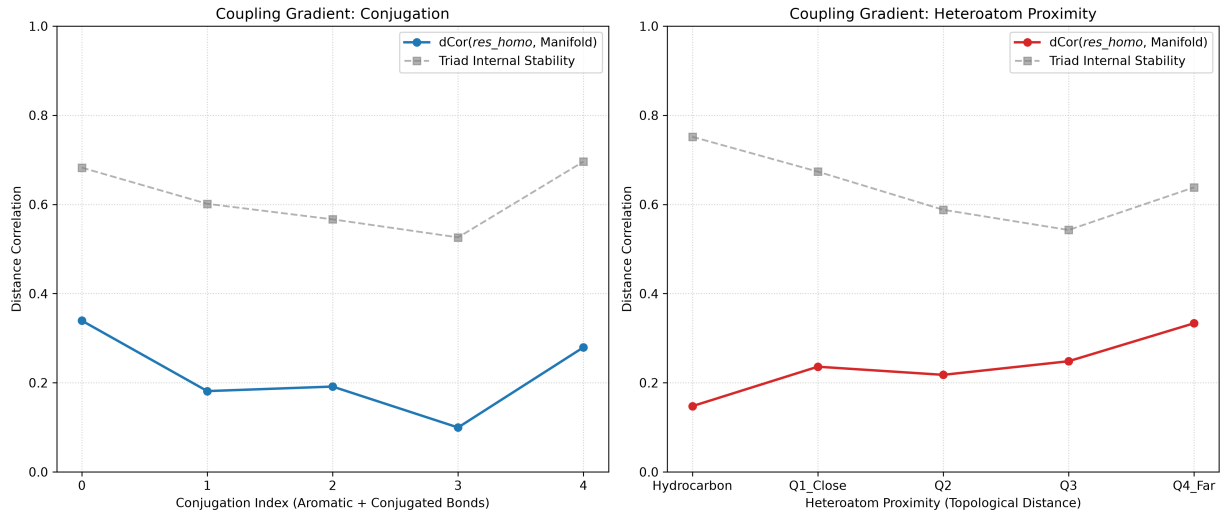

Figure 4: Distance correlation ($dCor$) between the HOMO energy residuals (`res_homo`) and the global thermodynamic-spatial manifold across structural bins. The left panel demonstrates that coupling decreases as the conjugation index increases from saturated systems (index 0) to highly conjugated systems (index 3), indicating that electron delocalization isolates frontier orbitals from macroscopic molecular properties. The right panel illustrates that pure hydrocarbons exhibit the lowest coupling, which steadily increases as heteroatoms are introduced and moved toward the molecular periphery (`Q4_Far`). These gradients reveal that HOMO decoupling is structurally driven: specific π-system topologies dictate the electronic state independently of global volume, whereas peripheral heteroatoms act as bridges that force the localized HOMO to align more closely with the overall physical extent of the molecule.

3.3 Geometric and electrostatic boundary conditions

Beyond local bonding topology, extreme geometric constraints significantly rewire the HOMO-manifold relationship. While the global non-linear baseline coupling of 0.263 is standard for most organic structures in 3D space, specific geometric and electrostatic subsets deviate sharply from this norm, as shown in Figure 5.

The most dramatic shift occurs in linear molecules (identified by a rotational constant of $A = 0$). In this small subset of 9 molecules, the extreme reduction in spatial dimensionality forces the frontier orbitals to correlate strongly with the global vibrational and spatial state, resulting in a highly elevated coupling of $dCor = 0.640$. Furthermore, highly symmetric and non-polar molecules ($\mu = 0$) exhibit a slightly elevated coupling ($dCor = 0.298$) compared to the polar baseline ($dCor = 0.254$). These findings indicate that the enforcement of strict geometric boundaries or high symmetry can effectively "squeeze" the electronic structure into alignment with the macroscopic manifold, overcoming the baseline decoupling observed in less constrained systems.

3.4 Statistical stability and scaffold diversity

To ensure that the observed structural dependencies were robust and not artifacts of sampling or over-represented molecular families, the gradients were subjected to rigorous statistical validation. Bootstrap resampling within each structural partition yielded narrow 95% confidence intervals for the $dCor$ estimates (e.g., $[0.316, 0.402]$ for saturated molecules), confirming the statistical stability of the coupling values.

Additionally, a hold-out validation was performed by excluding the top 10 most frequent Murcko scaffolds from the dataset. When these dominant structural families were removed, the absolute magnitude of the coupling values often shifted downward; for instance, the mean coupling in saturated molecules dropped from 0.355 to a hold-out $dCor$ of 0.165. This reduction indicates that the absolute magnitude of coupling in certain bins is partially sustained by the most common structural families in the QM9 dataset. Crucially, however, the relative gradients—such as the significant drop in coupling with increased conjugation and

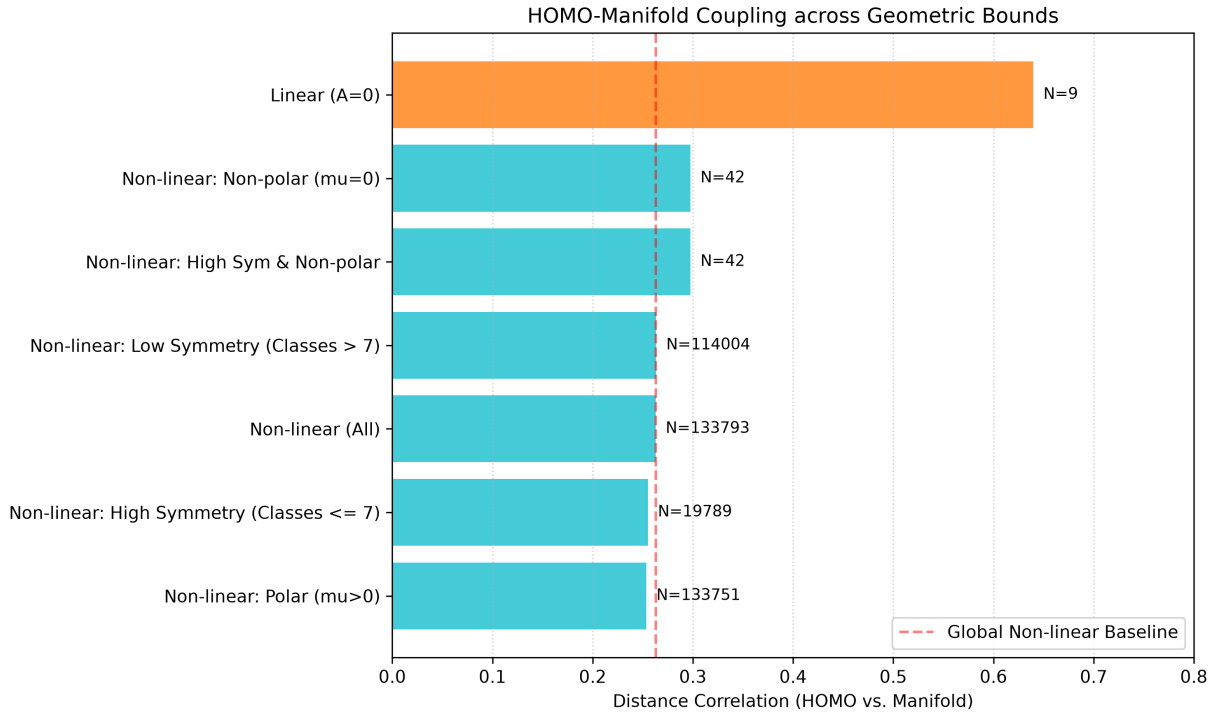

Figure 5: Distance correlation ($dCor$) between the HOMO energy residuals and the global thermodynamic-spatial manifold across various geometric and electrostatic subsets. The vertical dashed line represents the global non-linear baseline coupling ($dCor = 0.263$), which is standard for most organic structures in the dataset. The plot demonstrates that while the HOMO is generally decoupled in 3D non-linear space, specific constraints can force the electronic structure into alignment with the macroscopic manifold. Most notably, the extreme reduction in spatial dimensionality in linear molecules ($A = 0$, $N = 9$) results in a highly elevated coupling ($dCor = 0.640$). Furthermore, highly symmetric and non-polar molecules ($\mu = 0$) show a slight increase in coupling compared to the polar baseline, indicating that strict geometric boundaries can bridge the fundamental divide between localized frontier orbitals and global spatial-vibrational properties.

the rise in coupling with peripheral heteroatoms—persisted in the hold-out sets. This confirms that the structural modulation of HOMO decoupling is a genuine physical phenomenon rather than a scaffold-driven artifact.

4 Conclusions

In this study, we investigated the structural boundary conditions governing the decoupling of the highest occupied molecular orbital (HOMO) from the global thermodynamic and spatial manifold in the QM9 dataset. By applying stoichiometric residualization to remove first-order compositional effects, we confirmed that while macroscopic properties—specifically zero-point vibrational energy, heat capacity, and electronic spatial extent—form a highly coherent physical state, the HOMO energy remains conspicuously orthogonal.

Our structural partitioning analysis reveals that this frontier orbital decoupling is not a universal constant across chemical space, but rather a structurally modulated phenomenon. We identified electron delocalization as the primary driver of this orthogonality; as conjugation increases, the HOMO energy becomes almost entirely dictated by the local topology of the π-system, further isolating it from the molecule's global spatial and vibrational characteristics. Conversely, the introduction of peripheral heteroatoms acts as a structural bridge. As heteroatoms are positioned further toward the molecular periphery, the localized electronic state is forced into closer alignment with the macroscopic physical extent of the molecule.

Furthermore, we demonstrated that extreme geometric constraints can fundamentally rewire the relationship between electronic and macroscopic states. While decoupling is the standard baseline for non-linear organic structures in three-dimensional space, the extreme reduction in spatial dimensionality observed in linear molecules ($A = 0$) drastically elevates the coupling between the HOMO and the global manifold. High symmetry and a lack of permanent dipole moment similarly, though less dramatically, squeeze the electronic structure into alignment with the global physical state.

Rigorous statistical validation through bootstrap resampling and scaffold hold-outs confirms that these coupling gradients are genuine physical phenomena rather than artifacts of overrepresented molecular families in the QM9 dataset. Ultimately, these findings demonstrate that the fundamental divide between localized electronic states and global thermodynamic properties is not absolute. Instead, it is systematically governed by local bonding topology and geometric boundaries, providing a more nuanced understanding of structure-property relationships in computational chemical space.