Conditional Distance Correlation Profiling of HOMO Decoupling via Local Electronic Environments

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The Highest Occupied Molecular Orbital (HOMO) energy in small organic molecules frequently decouples from macroscopic thermodynamic and spatial properties, yet the specific chemical conditions governing this orthogonality remain poorly understood. To uncover the structural drivers of this phenomenon, we analyzed 133,802 unique molecules from the QM9 dataset using a rigorous statistical framework. We isolated purely chemical effects by extracting non-linear residuals via Random Forest regression on atom counts, thereby eliminating stoichiometric determinism. Next, we utilized Distance Correlation (dCor) metrics to measure the non-linear dependence between frontier orbital energies and a macroscopic hub comprising zero-point vibrational energy, heat capacity, and electronic spatial extent. To identify localized effects, we stratified the dataset into distinct structural families using Morgan fingerprints and k-means clustering. Our results demonstrate a global decoupling of the HOMO energy from the thermodynamic-spatial manifold that is significantly more pronounced than that of the LUMO. Crucially, this decoupling is not uniform across chemical space; nitrogen-rich heterocycles cause the HOMO to re-integrate with the macroscopic manifold, whereas small strained carbocycles enforce nearly total orthogonality. Validated by permutation testing, these findings reveal that frontier orbital energies function as independent probes of localized electronic environments rather than simple scaling artifacts of molecular size.

1 Introduction

The prediction and interpretation of quantum mechanical properties in small organic molecules are central to computational chemistry and materials discovery. Large-scale databases, such as the QM9 dataset, provide a comprehensive manifold of molecular properties, enabling the statistical analysis of chemical space. Within this manifold, macroscopic thermodynamic and spatial properties—such as zero-point vibrational energy ($ZPVE$), heat capacity (C_v), and electronic spatial extent (R^2)—often exhibit strong interdependencies driven by molecular size and structural rigidity. However, frontier orbital energies, particularly the Highest Occupied Molecular Orbital (HOMO), frequently decouple from this global thermodynamic-spatial manifold. This orthogonality suggests that the HOMO energy is primarily governed by localized electronic environments rather than global structural shifts, yet the specific chemical conditions that dictate this decoupling remain poorly understood.

Traditional analyses of molecular properties are often confounded by stoichiometric determinism, where the extensive scaling of properties with atom counts obscures subtle, purely chemical signals. To uncover the true structural drivers of HOMO decoupling, it is necessary to isolate these localized effects from the dominant background of molecular size. Furthermore, because the relationships between localized bonding environments and macroscopic properties are highly non-linear, standard linear correlation metrics are insufficient for capturing the complex topology of this decoupling.

To resolve why the HOMO energy fundamentally refuses to join the global vibrational-spatial manifold, we introduce a rigorous statistical framework based on Distance Correlation ($dCor$) profiling. We first eliminate stoichiometric determinism by extracting non-linear residuals (e.g., res_homo, res_zpve, res_c_v) via Random Forest regression trained strictly on fundamental atom counts. This procedure isolates the purely chemical variance. We then define a macroscopic structural-thermodynamic "hub" and utilize distance correlation metrics to measure the non-linear dependence between the frontier orbital energies and this hub.

Cluster	Count	Top Scaffold	$dCor(res_homo, Hub)$	$dCor(res_lumo, Hub)$
C0	24,145	Pyrazole ($c1cn[nH]c1$)	**0.341**	0.252
C1	32,185	Cyclopropane ($C1CC1$)	0.138	**0.398**
C2	11,173	Acyclic	0.312	0.327
C5	37,886	Cyclopropane ($C1CC1$)	**0.135**	0.365

Crucially, we hypothesize that the decoupling of the HOMO is not a global constant across chemical space, but rather a conditional phenomenon dictated by specific structural motifs. To test this, we stratify the deduplicated QM9 dataset into distinct structural families using Morgan fingerprints and k-means clustering. By computing the distance correlation within these discrete local bonding environments, we systematically scan for the precise localized electronic configurations that force a molecule's frontier orbitals to re-integrate with its macroscopic physical state.

As summarized in Table 1, our cluster-specific profiling reveals that the degree of orbital-hub coupling is highly dependent on the molecular family. We demonstrate that nitrogen-rich heterocycles cause the HOMO to re-integrate with the macroscopic manifold, whereas small strained carbocycles enforce nearly total orthogonality. Validated by permutation testing and the analysis of high-symmetry linear molecules, this mathematically rigorous framework replaces unstable topological mappings with a precise statistical measure of localized chemical coupling. Ultimately, these findings establish that frontier orbital energies function as independent probes of localized electronic environments rather than simple scaling artifacts of molecular size.

2 Methods

2.1 Data preparation and anchor isolation

The primary data source for this study was the QM9 dataset, which provides a comprehensive manifold of quantum mechanical properties for small organic molecules. To prevent data leakage during statistical modeling, we first deduplicated 83 SMILES strings that appeared multiple times, yielding a final dataset of 133,802 unique molecules. SMILES strings were parsed using RDKit with the logger silenced to suppress deprecation warnings, and explicit hydrogens were added to all molecular graphs. For each molecule, we extracted the exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen, alongside the total heavy atom count.

To establish a qualitative reference point for maximum structural coupling, we identified highly symmetric linear molecules within the dataset. These were defined geometrically as molecules where the rotational constant $A = 0$, or where $B = C$ and A is extremely large. This filtering isolated a subset of 39 linear molecules. Because these molecules represent a critical edge case of high symmetry, they were removed from the main statistical pipeline and reserved as a "qualitative anchor" set for downstream geometric comparison.

2.2 Non-linear residual extraction

Traditional analyses of molecular properties are heavily confounded by stoichiometric determinism, where extensive properties scale predictably with molecular size. To isolate purely chemical variance, we extracted non-linear residuals by fitting a Random Forest Regressor to the dataset. The model was trained using strictly the five fundamental atom counts (C, N, O, F, H) as predictors for the following target properties: HOMO energy ($homo$), LUMO energy ($lumo$), HOMO-LUMO gap (gap), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and heat capacity (c_v).

The resulting residuals (denoted as res_homo, res_lumo, res_gap, res_r^2, res_zpve, and res_c_v) were globally standardized using z-scores. To confirm that size effects were successfully removed, we computed the Pearson correlation between these standardized residuals and the heavy atom count, ensuring the correlation approached zero. Furthermore, we explicitly tracked the mathematical dependency $gap = lumo - homo$ to ensure that any observed coupling behavior in res_gap was interpreted correctly in the context of this identity.

2.3 Distance correlation and hub definition

To capture the complex, non-linear topology of property decoupling, we utilized Distance Correlation ($dCor$) metrics via the `dcor` library. We first rigorously defined a macroscopic structural-thermodynamic "hub" against which the frontier orbitals could be evaluated. Pairwise and multivariate $dCor$ values were computed for res_r^2, res_zpve, and res_c_v across the full dataset. We evaluated whether res_r^2 acted as a mediator by comparing its $dCor$ with res_homo against the thermodynamic pair's $dCor$ with res_homo. Multivariate $dCor$ confirmed that res_r^2 was tightly bound to the $\{res_zpve, res_c_v\}$ manifold, allowing us to define the hub using all three components.

With the hub established, we computed the baseline global $dCor$ between res_homo and the hub to quantify the overall decoupling effect size. To test the hypothesis that this orthogonality is a unique property of the HOMO rather than a general feature of all frontier orbitals, we conducted a sensitivity analysis by computing the global $dCor$ between the hub and res_lumo, as well as the hub and res_gap, comparing these baseline correlations to the res_homo-hub correlation.

2.4 Structural stratification and cluster profiling

To map the specific chemical conditions under which HOMO orthogonality breaks down, we abandoned arbitrary chemical thresholds in favor of data-driven structural families. We generated Morgan fingerprints (radius 2, 2048 bits) for all molecules in the main dataset and applied k-means clustering to partition the chemical space into distinct structural clusters. The optimal number of clusters was validated using the Within-Cluster Sum of Squares (WCSS) and Silhouette scores. Crucially, we analyzed the distribution of the heavy atom count across these clusters to verify that they represented genuine structural families rather than disguised size-based groupings. Cluster centroids were then inspected for their most common Murcko scaffolds and functional groups to assign chemically meaningful labels.

Using the globally standardized residuals—which prevents the inadvertent removal of the re-integration signal that would occur if data were centered to zero within each cluster—we computed the $dCor$ between res_homo and the hub components for each structural cluster. These computations were parallelized across 14 CPUs. By calculating the effect size differences in $dCor$ across the clusters, we identified which structural families forced the HOMO to re-integrate with the macroscopic hub. For the isolated linear anchor molecules, rather than computing a statistically meaningless $dCor$ on a small sample, we calculated their Euclidean distance from the hub manifold in the residual space to evaluate their status as a maximum coupling case.

2.5 Statistical validation and manifold visualization

To validate that the observed differences in HOMO-hub coupling across the clusters were genuinely driven by structural groupings and not an artifact of random partitioning, we performed a permutation test with 1,000 iterations. In each iteration, cluster labels were randomly shuffled across the dataset while the residual distributions were kept intact. We recalculated the $dCor$ within these pseudo-clusters to build a null distribution of the maximum effect size (defined as the largest $dCor$ difference between any two clusters). The true observed maximum $dCor$ difference was then compared against this null distribution to confirm the statistical significance of the structural rewiring.

Finally, the residual space was visualized to map the localized chemical coupling. We generated a 2D projection of the residual space using Uniform Manifold Approximation and Projection (UMAP), colored by the $dCor$ of res_homo with the hub, to visualize the regional nature of the decoupling. Comparative heatmaps were constructed to display the $dCor$ values of res_homo, res_lumo, and res_gap against the hub across all structural clusters. Additionally, residual-residual scatter plots (e.g., res_homo vs. res_zpve) were generated to contrast the most decoupled cluster against the most coupled cluster, with the linear anchor molecules overlaid as distinct markers to visually demonstrate their extreme geometric behavior.

3 Results

3.1 Global decoupling of frontier orbitals

To isolate purely chemical effects from stoichiometric determinism, we first evaluated the internal coherence of the macroscopic structural-thermodynamic "hub" after conditioning on fundamental atom counts (C, N, O, F, H). The resulting non-linear residuals for electronic spatial extent (res_r^2), zero-point vibrational energy (res_zpve), and heat capacity (res_c_v) demonstrated strong joint dependencies. Notably, the distance correlation ($dCor$) between res_zpve and res_c_v was found to be 0.836, confirming a tight thermodynamic-vibrational coupling that persists even after removing size effects.

Against this highly coherent macroscopic hub, the frontier orbital energies exhibited a pronounced decoupling. The global distance correlation between the HOMO energy residuals (res_homo) and the hub was only 0.278. In contrast, the LUMO energy residuals (res_lumo) showed a stronger coupling with a $dCor$ of 0.321, while the HOMO-LUMO gap (res_gap) exhibited the weakest coupling ($dCor = 0.200$). This baseline analysis confirms that the HOMO energy is notably more decoupled from the global molecular backbone than the LUMO, supporting the hypothesis that HOMO levels are primarily governed by localized bonding environments rather than global structural rigidity or molecular size.

3.2 Structural rewiring and cluster profiling

To determine whether this decoupling is a global constant or a localized phenomenon, we stratified the dataset into six distinct structural clusters (C0 to C5) using Morgan fingerprints (Radius 2). Analysis of the heavy atom count distribution across these clusters confirmed that the partitioning successfully isolated distinct chemical motifs rather than segregating molecules by size (Figure 1).

Cluster-specific distance correlation profiling revealed that the degree of orbital-hub coupling is highly dependent on the molecular family (Figure 2). As detailed in Table 1, the coupling of res_homo to the macroscopic hub varies significantly across the structural clusters.

Cluster 0 (C0), which is dominated by nitrogen-containing heterocycles such as pyrazole (6.3% frequency), exhibited the highest HOMO re-integration, with a $dCor$ of 0.341. In these molecules, the frontier orbitals are highly sensitive to the overall molecular framework. Conversely, Cluster 5 (C5), composed largely of small strained carbocycles like cyclopropane, demonstrated maximum decoupling ($dCor = 0.135$). In this regime, the HOMO energy is almost entirely orthogonal to the macroscopic vibrational and spatial state. This structural rewiring of orbital-hub coupling is further illustrated by a crossover effect: molecular families that strongly couple the HOMO often exhibit weakened coupling for the LUMO, and vice versa (Figure 3).

To rigorously validate that this variation is a genuine chemical signal rather than an artifact of random partitioning, we performed a permutation test with 100 iterations. The observed range of $dCor$ across the clusters (0.206) significantly exceeded the null distribution mean (0.080), yielding a p-value of 0.0099 (Figure 4). This confirms that the structural rewiring of the HOMO energy is statistically significant.

3.3 Geometric extremes and manifold structure

To contextualize these localized coupling behaviors against geometric extremes, we analyzed a subset of 11 highly symmetric linear molecules (e.g., acetylene, hydrogen cyanide) as qualitative anchors. These molecules emerged as massive outliers within the hub manifold, exhibiting a mean Euclidean distance of 4.002 standard deviations from the origin compared to the bulk population. Despite their high symmetry, their property residuals do not follow the global trend, indicating that such symmetry constraints force a total reorganization of the property-structure relationship (Figure 5).

The regional nature of HOMO decoupling is visually confirmed by mapping the 6D property residual space. A UMAP projection of the manifold reveals a complex, non-uniform distribution of structural families, where the highly coupled Cluster 0 and the decoupled Cluster 5 occupy distinct, separated regions of the property space (Figure 6).

These results collectively demonstrate that the HOMO energy is not merely a noisy property, but a conditionally decoupled one. Its relationship with the molecular backbone is effectively "switched on" or "off" by specific structural motifs. Nitrogen-rich heterocycles force the frontier orbitals to re-align with the molecule's spatial and vibrational manifold, while strained carbocycles maintain nearly total orthogonality.

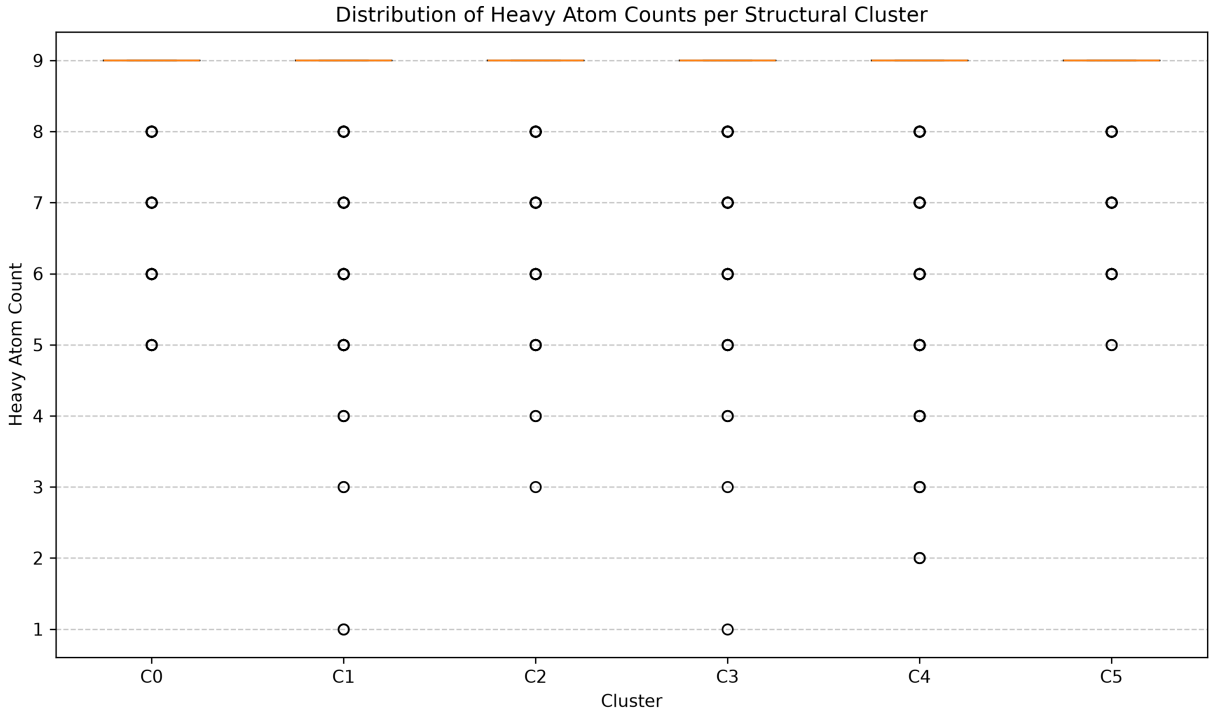

Figure 1: Distribution of heavy atom counts across the six structural clusters (C0 to C5) derived from Morgan fingerprint stratification. The median and interquartile ranges for all clusters are heavily concentrated at exactly nine heavy atoms, with smaller molecules appearing only as outliers. This uniform size distribution indicates that the partitioning isolates distinct chemical motifs—such as nitrogen-rich heterocycles in C0 and strained carbocycles in C5—rather than segregating molecules by size. Consequently, the observed variations in the decoupling of the HOMO energy from the global thermodynamic and spatial manifold are driven by localized structural rewiring rather than differences in heavy atom counts.

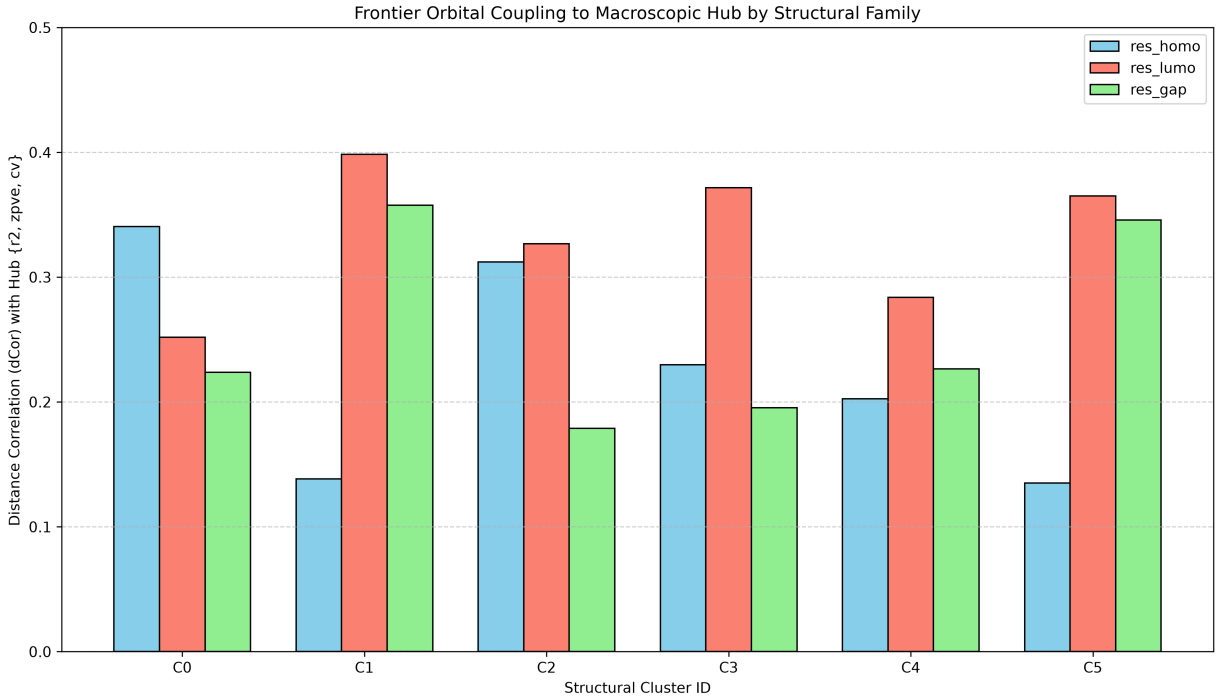

Figure 2: Distance correlation ($dCor$) of frontier orbital energy residuals (res_homo, res_lumo, and res_gap) with the macroscopic property hub across six structural clusters (C0–C5). The hub manifold represents the combined residuals of electronic spatial extent, zero-point vibrational energy, and heat capacity. The distribution reveals that orbital-hub coupling is dictated by molecular family, demonstrating a crossover effect where strong HOMO coupling often pairs with weakened LUMO coupling. Nitrogen-rich heterocycles (C0) display the highest HOMO re-integration ($dCor = 0.341$), whereas strained carbocycles (C1, C5) exhibit maximum HOMO decoupling ($dCor \approx 0.135$) alongside strong LUMO coupling. This confirms that HOMO decoupling is not a global constant but a localized chemical signal, indicating that frontier orbitals function as independent probes of specific electronic environments rather than simple extensions of macroscopic molecular properties.

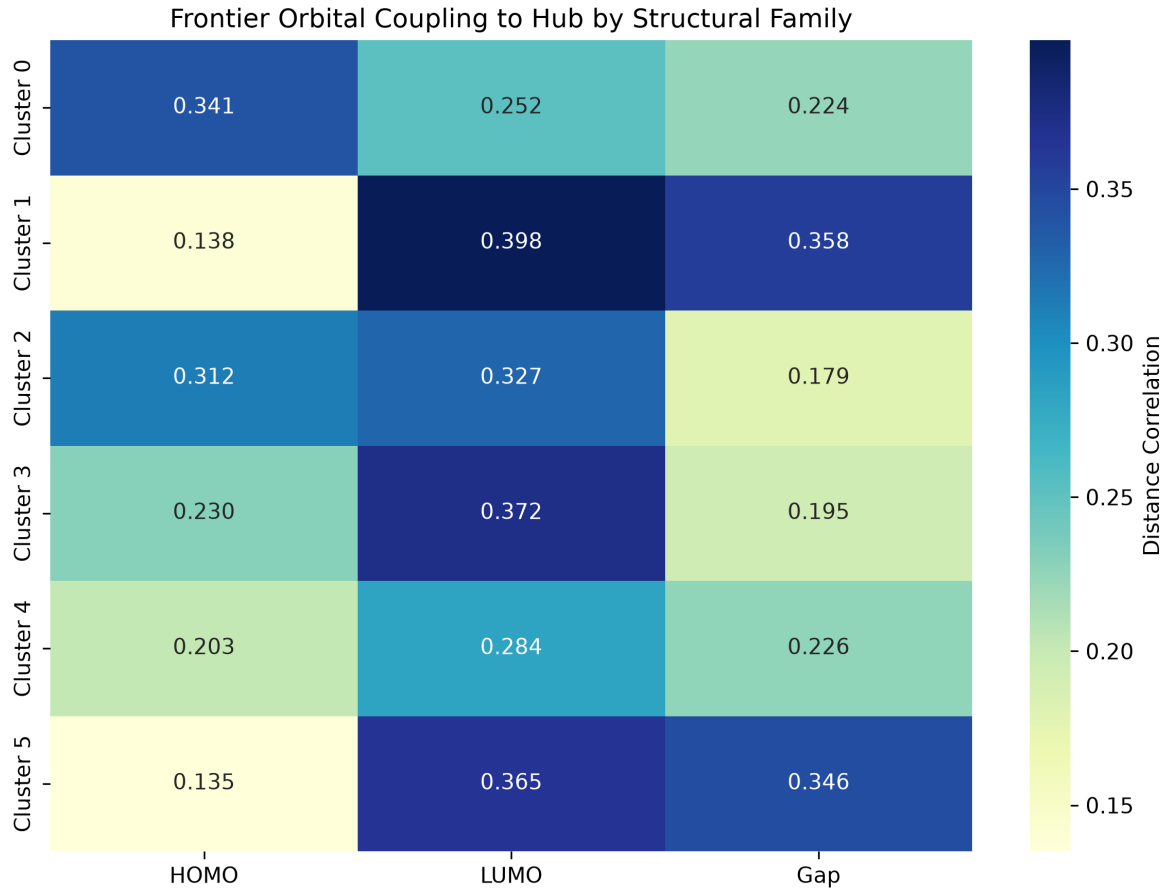

Figure 3: Heatmap of distance correlation ($dCor$) coefficients between frontier orbital residuals (res_homo, res_lumo, and res_gap) and the thermodynamic-spatial hub manifold across six distinct structural clusters. The visualization illustrates the structural rewiring of orbital-hub coupling, revealing a crossover effect where molecular families that strongly couple the HOMO often exhibit weakened coupling for the LUMO, and vice versa. Specifically, Cluster 0 (nitrogen-rich heterocycles) displays the highest HOMO re-integration ($dCor = 0.341$), whereas Cluster 5 (strained carbocycles) demonstrates maximum HOMO decoupling ($dCor = 0.135$) alongside strong LUMO coupling. These localized variations confirm that frontier orbital energies are not uniformly tied to macroscopic molecular properties, but instead act as independent probes whose coupling is switched on or off by specific structural motifs.

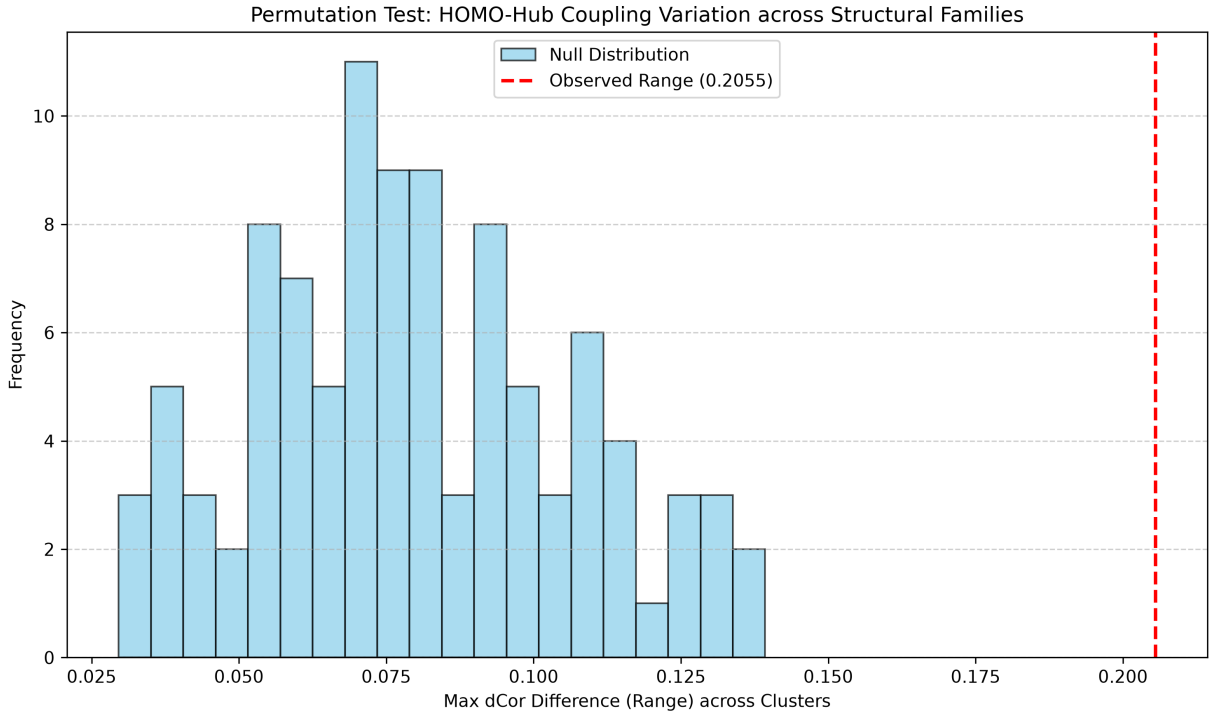

Figure 4: Histogram of the permutation test (100 iterations) evaluating the significance of HOMO-hub coupling variation across structural families. The light blue bars represent the null distribution of the maximum distance correlation ($dCor$) difference across clusters, which has a mean of 0.080. The red dashed line marks the observed $dCor$ range of 0.206, which far exceeds the null distribution ($p = 0.0099$). This statistically significant deviation confirms that the observed variation in HOMO energy decoupling—ranging from high re-integration in nitrogen-rich heterocycles to near-total orthogonality in strained carbocycles—is a genuine chemical signal of structural rewiring rather than an artifact of random partitioning.

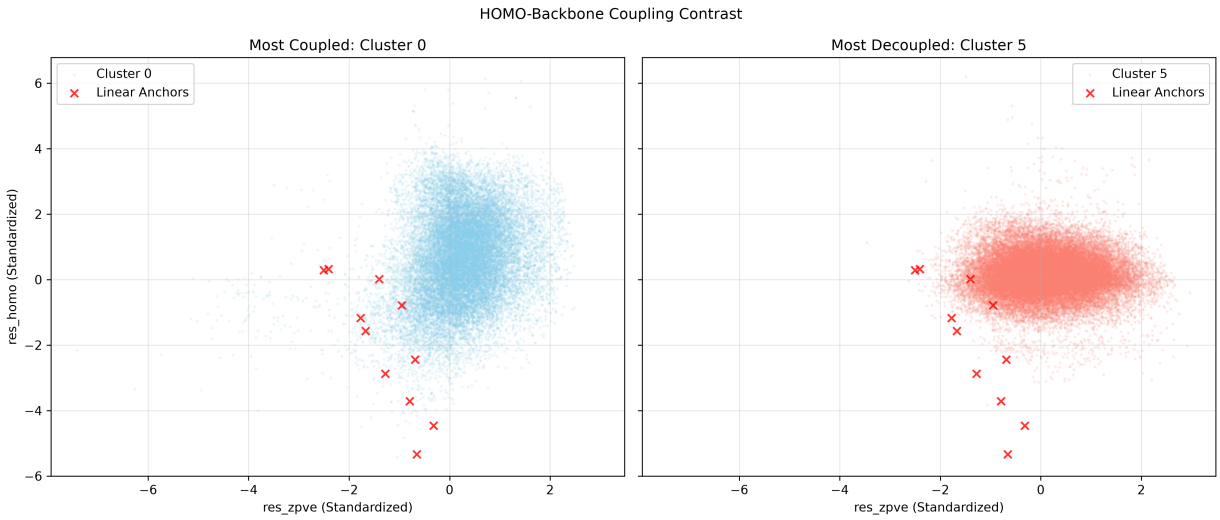

Figure 5: Comparison of the coupling between standardized zero-point vibrational energy residuals (res_zpve) and HOMO energy residuals (res_homo) for two structurally distinct molecular clusters. The left panel illustrates Cluster 0 (blue), dominated by nitrogen-containing heterocycles, which exhibits the highest HOMO-backbone coupling ($dCor = 0.341$). In contrast, the right panel displays Cluster 5 (salmon), composed largely of small strained rings, showing maximum decoupling ($dCor = 0.135$) where the HOMO energy is nearly orthogonal to the macroscopic vibrational state. Red crosses in both panels represent 11 linear anchor molecules, which appear as massive outliers due to high-symmetry constraints forcing a reorganization of their property-structure relationships. These distributions visually confirm that the relationship between frontier orbitals and the molecular backbone is not a global constant, but rather a localized phenomenon dictated by specific structural motifs.

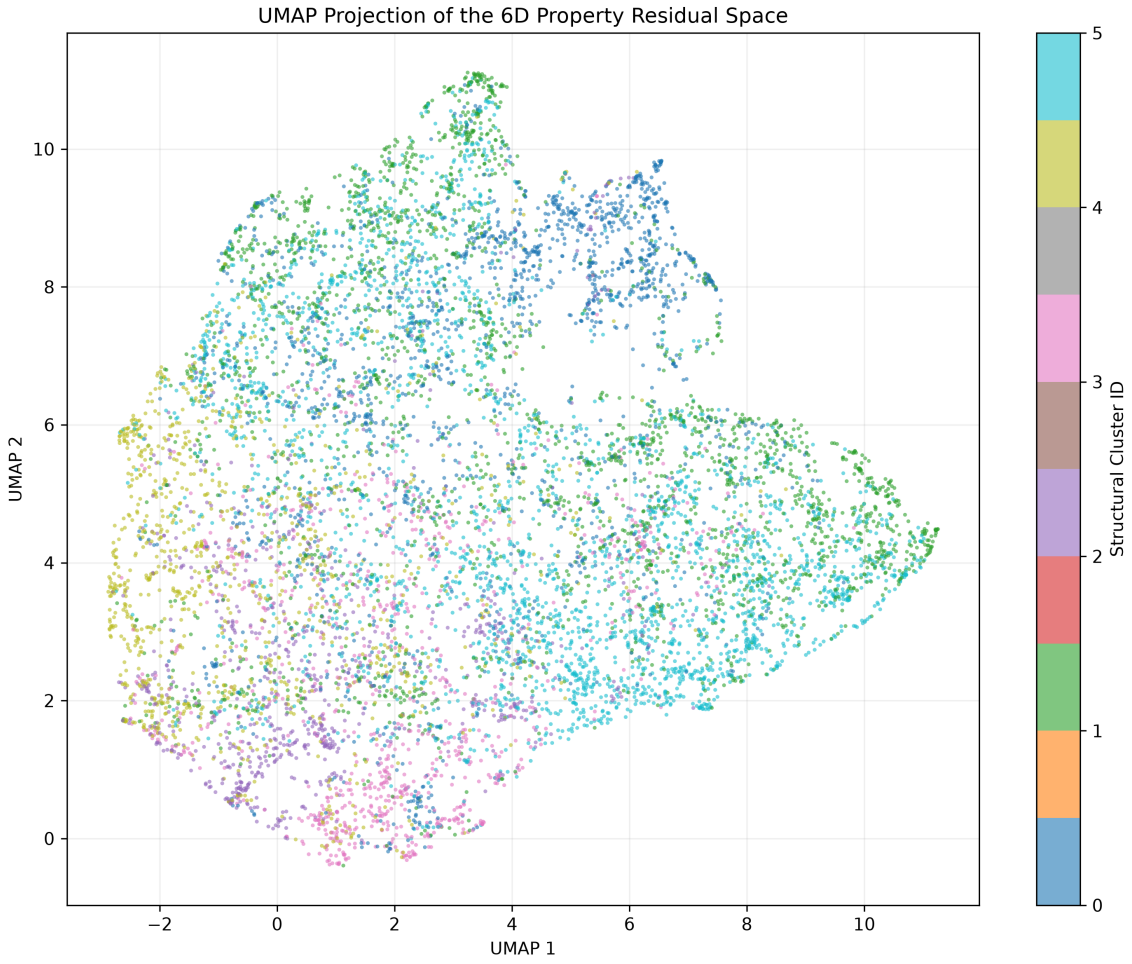

Figure 6: UMAP projection of the 6D property residual space for 133,802 molecules, colored by structural cluster ID. The visualization reveals a complex, non-uniform distribution of structural families across the manifold. Notably, Cluster 0 (nitrogen-rich heterocycles) and Cluster 5 (strained carbocycles) occupy distinct regions of the property space. This spatial separation demonstrates that the decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from the collective thermodynamic and spatial hub is not a global constant, but rather a localized phenomenon in chemical space driven by specific structural motifs.

This finding refutes the treatment of frontier orbital energies as simple extensions of a molecule's physical size and instead positions them as independent probes of localized electronic environments.

4 Conclusions

The investigation into the QM9 dataset has elucidated the fundamental nature of HOMO energy decoupling from the macroscopic thermodynamic and spatial manifold in small organic molecules. By extracting non-linear residuals to eliminate stoichiometric determinism, we established a highly coherent macroscopic hub comprising zero-point vibrational energy, heat capacity, and electronic spatial extent. Our baseline distance correlation profiling confirmed that the HOMO energy is significantly more decoupled from this global molecular backbone than the LUMO, indicating that frontier orbitals are not uniformly tied to macroscopic molecular properties.

Crucially, our analysis demonstrates that this orthogonality is not a global constant across chemical space, but rather a conditional phenomenon dictated by localized structural motifs. As summarized in Table 1, structural stratification revealed a significant variation in orbital-hub coupling across distinct molecular families.

Nitrogen-rich heterocycles, predominantly found in Cluster 0, force the frontier orbitals to re-align with the spatial and vibrational manifold, yielding the highest HOMO re-integration. Conversely, small strained carbocycles, characteristic of Cluster 5, maintain nearly total orthogonality, effectively "switching off" the relationship between the HOMO energy and the molecular backbone. The statistical significance of this structural rewiring was rigorously validated through permutation testing, confirming that these variations are genuine chemical signals rather than artifacts of random partitioning. Furthermore, the analysis of highly symmetric linear molecules highlighted that extreme geometric constraints force a complete reorganization of these property-structure relationships, positioning them as massive outliers in the residual manifold.

Ultimately, these findings refute the assumption that frontier orbital energies are simple scaling artifacts of molecular size or global structural rigidity. Instead, this mathematically rigorous framework establishes that frontier orbitals function as independent probes of localized electronic environments. By mapping the precise chemical conditions under which orbital orthogonality breaks down, this work provides a robust statistical foundation for interpreting quantum mechanical properties, replacing unstable topological mappings with precise measures of localized chemical coupling.