Stratified Distance Correlation to Map the Breakdown of HOMO Decoupling Across Local Electronic Environments

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energy of the highest occupied molecular orbital (HOMO) typically exhibits an unexplained orthogonality to global thermodynamic and spatial property manifolds in large chemical datasets. To determine whether this physical isolation is a universal constant or conditional on local bonding environments, we systematically map the breakdown of HOMO decoupling across discrete structural regimes. Using the QM9 dataset of 133,802 molecules, we isolate chemical architecture from stoichiometric determinism by extracting non-linear gradient-boosted atom-count residuals for twelve core molecular properties. We then partition the chemical space into distinct strata via density-based clustering on local structural descriptors. Within each stratum, we compute the partial distance correlation between the HOMO residual and a principal component representation of the non-frontier manifold, explicitly controlling for LUMO dependencies. Our analysis reveals that global HOMO orthogonality is largely an artifact of dataset-wide averaging. We observe significant rewirings of the electronic-geometric relationship in specific local environments, with the strongest coupling occurring in saturated, heteroatom-rich, and cyclic systems. Conversely, highly aromatic and conjugated systems maintain strict decoupling, acting as electronic buffers. These findings demonstrate that frontier orbital energies are tightly constrained by molecular geometry in specific aliphatic and polar systems, providing a targeted map of where joint property models must account for frontier-electronic coupling.

1 Introduction

The energy of the highest occupied molecular orbital (HOMO) is a fundamental descriptor of chemical reactivity, ionization potential, and overall molecular stability. However, in large-scale chemical datasets such as the QM9 database of small organic molecules, a persistent anomaly has been observed: the HOMO energy exhibits a striking orthogonality to global thermodynamic and spatial property manifolds. While geometric, rotational, and thermodynamic properties often display tight joint organization when conditioned on basic stoichiometric variables, the HOMO energy appears physically isolated. This raises a critical question: is this electronic decoupling a universal physical constant of small organic molecules, or is it merely a statistical artifact of dataset-wide averaging?

Traditional approaches to molecular property analysis often treat vast chemical spaces as monolithic distributions. By computing global correlations or training global predictive models, these methods inadvertently average over highly diverse local bonding environments. Consequently, if the relationship between frontier orbital energies and the spatial-vibrational manifold is highly conditional on specific chemical architectures, global metrics will wash out these localized dependencies, falsely presenting the HOMO as universally decoupled.

To resolve this unexplained orthogonal behavior, we systematically map the breakdown of HOMO decoupling across discrete structural regimes. We first isolate intrinsic chemical architecture from stoichiometric determinism by extracting non-linear gradient-boosted atom-count residuals for twelve core molecular properties. This ensures that our analysis captures purely structural and electronic variance, avoiding the confounding effects of molecular size and atom-count scaling. We then partition the deduplicated QM9 dataset into distinct structural strata using density-based clustering on localized structural descriptors, such as hybridization fractions, heteroatom density, and aromaticity.

Cluster ID	Size	Representative Chemical Motif (Interpretation)
0	23,527	Aromatic+, Conjugated+ (Planar systems)
24	2,623	Saturated+, Hetero-rich (Polar aliphatic)
14	1,756	Conjugated+, Aromatic (Extended π systems)
3	2,522	Saturated+, Cyclic (Fused/strained rings)
32	1,605	Cyclic+, Saturated+ (Polycyclic cages)
-1	32,591	Cyclic+, Saturated (Noise/Miscellaneous)

This stratification yields a diverse set of coherent molecular families, ranging from highly aromatic systems to dense, heteroatom-rich saturated clusters. Table 1 highlights several representative strata identified in this study, illustrating the distinct chemical motifs that define these local environments.

Within each specific local-environment stratum, we compute the partial distance correlation between the HOMO residual and a principal component representation of the non-frontier manifold. Crucially, this approach explicitly controls for the mathematical dependency on the lowest unoccupied molecular orbital (LUMO) energy, ensuring that any observed coupling is a genuine reflection of HOMO behavior rather than a shared dependency on the band gap.

By identifying the exact structural thresholds where the HOMO suddenly couples to the geometric-vibrational axes, we provide a targeted map of where global frontier-electronic independence breaks down. Our findings demonstrate that global HOMO orthogonality is indeed an artifact of dataset-wide averaging. We reveal that significant rewirings of the electronic-geometric relationship occur in specific local environments—most notably in saturated, heteroatom-rich, and cyclic systems—where frontier orbital energies are tightly constrained by molecular geometry. Conversely, highly aromatic and conjugated systems maintain strict decoupling, acting as electronic buffers. These results provide a targeted framework for understanding where joint property models must explicitly account for frontier-electronic coupling.

2 Methods

2.1 Data preprocessing and redundancy elimination

The QM9 dataset was deduplicated using SMILES strings, yielding 133,802 unique molecules. To prevent mathematical identities from confounding multivariate analysis, the HOMO-LUMO gap was removed, as it is a linear combination of the HOMO and LUMO energies. Redundant thermodynamic properties (e.g., zero-point energies and atom-averaged thermodynamic quantities) were also excluded, leaving 12 core molecular properties. For linear molecules exhibiting an exact zero for the rotational constant A, a binary `is_linear` flag was introduced. The zero values were subsequently replaced with half the minimum observed non-zero A value in the dataset to maintain class separation without introducing discontinuities. Finally, a $\log(1+x)$ transformation was applied to A to mitigate the distortion of variance calculations by its heavy-tailed distribution.

2.2 Non-linear stoichiometric baseline extraction

To isolate intrinsic chemical architecture from stoichiometric determinism, we extracted size-orthogonal property residuals. Relying on linear models for this task often traps residuals into acting as mere proxies for molecular volume. Instead, we employed a non-linear baseline by training a Gradient Boosting Regressor (LightGBM) for each of the 12 core properties. The features consisted of the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, alongside the `is_linear` flag. Out-of-fold predictions were generated via 5-fold cross-validation, and the residuals were computed as the difference between the true and predicted values ($res_prop = true - predicted$). These residuals represent purely structural and electronic variance, independent of basic atom-count scaling.

2.3 Local structural descriptors and density-based stratification

To partition the chemical space into discrete structural strata, we computed five intensive, size-independent structural descriptors using RDKit: the fraction of sp^3 hybridized carbons, the heteroatom-to-carbon ratio, the aromatic heavy atom fraction, the maximum conjugated path length normalized by heavy atom count, and the number of rings per heavy atom. Molecules lacking specific atom types were handled to avoid division-by-zero errors. The descriptors were then standardized to zero mean and unit variance to ensure equal weighting during distance-based clustering.

We applied Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to the standardized descriptors. The minimum cluster size was set to 1,000 molecules to guarantee statistical robustness within each stratum. This density-based approach partitioned the dataset into 37 distinct strata defined by intrinsic chemical motifs, such as high-aromaticity or high-heteroatom-density clusters. The centroids of these clusters in the original descriptor space were extracted to assign chemical interpretations to each group (Table 1).

2.4 Stratum-specific manifold extraction

To reduce the multiple-testing burden and evaluate the joint response of the spatial and thermodynamic properties, we represented the non-frontier properties as a single local manifold within each stratum. The 10 non-frontier property residuals (excluding HOMO and LUMO residuals) were isolated, and Principal Component Analysis (PCA) was performed. The first principal component (PC1) was extracted to represent the manifold. If PC1 explained less than 30% of the variance within a given stratum, both PC1 and PC2 were retained to ensure sufficient representation of the local property variance.

2.5 Conditional distance correlation mapping

To measure the localized shifts in joint dependency between the HOMO residual and the non-frontier manifold, we utilized Distance Correlation. Because the HOMO and LUMO residuals are mathematically linked to the physical gap, it was imperative to ensure that any observed coupling was not merely a shared dependence on the LUMO. We computed a partial distance correlation ($pdCor$) using a regression-based residualization approach. Within each stratum, a Random Forest regressor was used to regress the LUMO residual out of both the HOMO residual and the manifold representation (PC1, or the joint vector of PC1 and PC2). The standard Distance Correlation was then computed between the residuals of these two regressions. This metric explicitly quantifies the coupling of the HOMO to the rest of the manifold after controlling for the LUMO. A global baseline $pdCor$ was also computed across the unstratified dataset for comparison.

2.6 Statistical validation and scaffold holdout

Distance correlation is sensitive to sample size and local density variations. To verify that the observed shifts in $pdCor$ were genuine chemical rewirings rather than statistical artifacts of a stratum's internal distribution, we conducted permutation tests. For each stratum, the HOMO residuals were permuted 500 times while keeping the manifold representation and LUMO residuals fixed. The residualization and distance correlation computations were repeated for each permutation to generate a null distribution. A genuine breakdown of HOMO isolation was defined as an observed $pdCor$ that strictly exceeded the 95th percentile of this null distribution and exhibited a substantial absolute shift relative to the global baseline.

Finally, to ensure that the discovered couplings were not driven by a single, highly populated rigid molecular family within a stratum, we performed a scaffold-split holdout validation. The top five most frequent Murcko scaffolds within each significant stratum were identified and removed. The $pdCor$ was then recomputed on the remaining subset to confirm the generalizability of the electronic-geometric coupling across diverse structures within the motif.

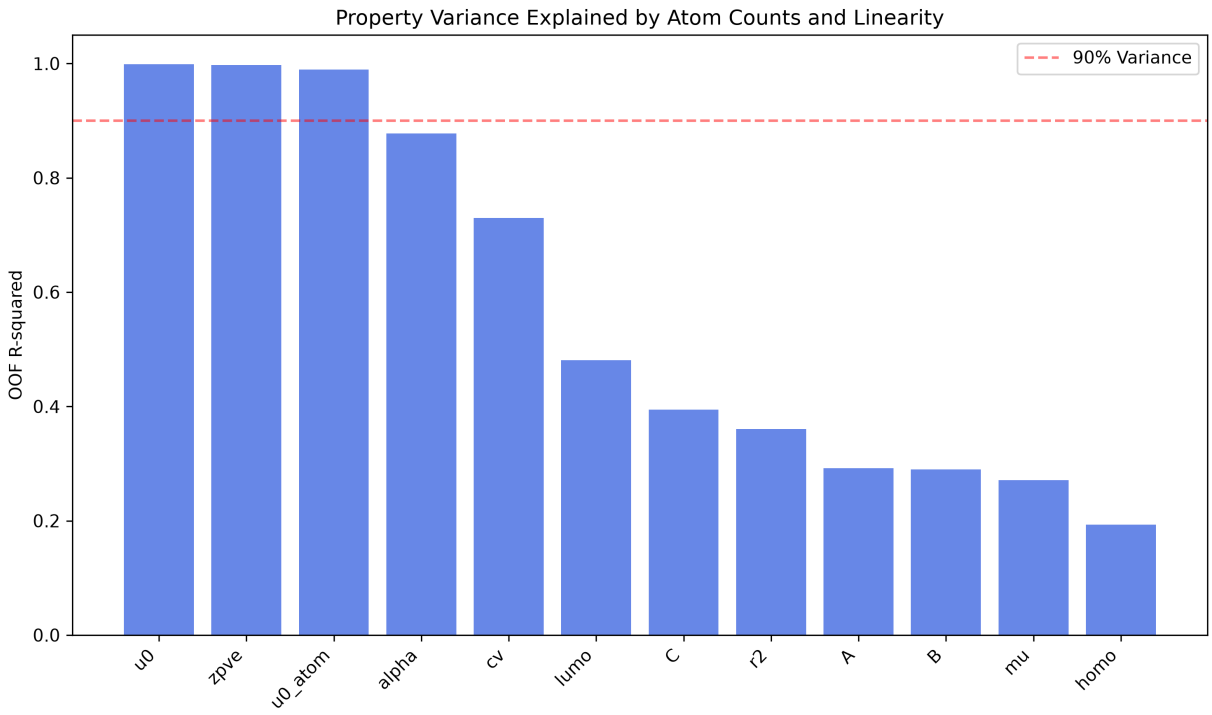

Figure 1: Out-of-fold (OOF) R-squared values representing the property variance explained by atom counts and linearity across the QM9 dataset. A red dashed line marks the 90% variance threshold. The plot demonstrates that while thermodynamic properties (such as u_0 and zpve) exhibit tight joint organization and are highly predictable based on atom counts, the HOMO energy shows the lowest explained variance. This visualizes the global "HOMO isolation" from the rest of the molecular property manifold, confirming that HOMO energy maintains near-zero correlation globally and motivating the search for specific local structural environments where this decoupling breaks down.

3 Results

3.1 Global isolation of HOMO energy

The energy of the highest occupied molecular orbital (HOMO) in the QM9 dataset is known for its remarkable orthogonality to the rest of the molecular property manifold. As established during our non-linear stoichiometric baseline extraction, while thermodynamic and geometric properties show tight joint organization once conditioned on atom counts, the residuals of HOMO energy (res_{homo}) exhibit the lowest explained variance (Figure 1). This visualizes the global "HOMO isolation" from the rest of the molecular property manifold. However, our subsequent analysis reveals that this isolation is not a universal physical constant, but rather a conditional property that breaks down within specific local electronic and structural environments.

3.2 Stratification of the chemical space

To move beyond global averages, the 133,802 unique molecules were partitioned into 37 discrete structural strata using HDBSCAN clustering on five standardized descriptors: sp^3 fraction, heteroatom-to-carbon ratio, aromatic fraction, normalized maximum conjugated path length, and rings per heavy atom. The highly skewed and multimodal profiles of these features (Figure 2) illustrate the extreme structural heterogeneity of the chemical space, necessitating a density-based approach to isolate coherent molecular families.

This stratification successfully isolated distinct chemical motifs, ranging from highly aromatic systems to dense, heteroatom-rich saturated clusters, as summarized in Table 1.

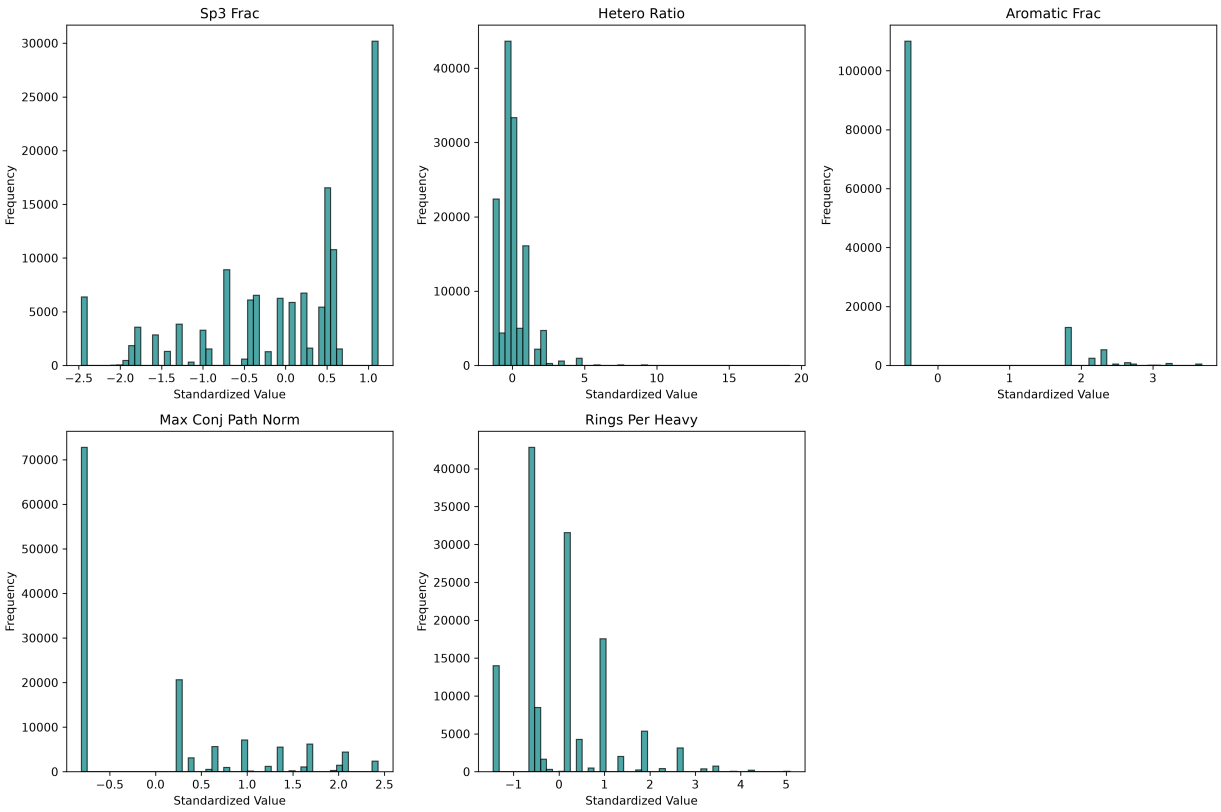

Figure 2: Frequency distributions of the five standardized structural descriptors—sp^3 fraction, heteroatom-to-carbon ratio, aromatic fraction, normalized maximum conjugated path length, and rings per heavy atom—used to partition the 133,802 molecules of the QM9 dataset. The highly skewed and multimodal profiles of these features illustrate the extreme structural heterogeneity of the chemical space, demonstrating why density-based HDBSCAN clustering is required to successfully isolate coherent molecular families. Stratifying the dataset along these specific dimensions is what enables the discovery that the apparent global isolation of HOMO energy residuals (res_{homo}) is an artifact; within specific local environments defined by these features, such as polar aliphatic or extended conjugated systems, the frontier electronic state strongly couples to the molecule's spatial and vibrational manifold.

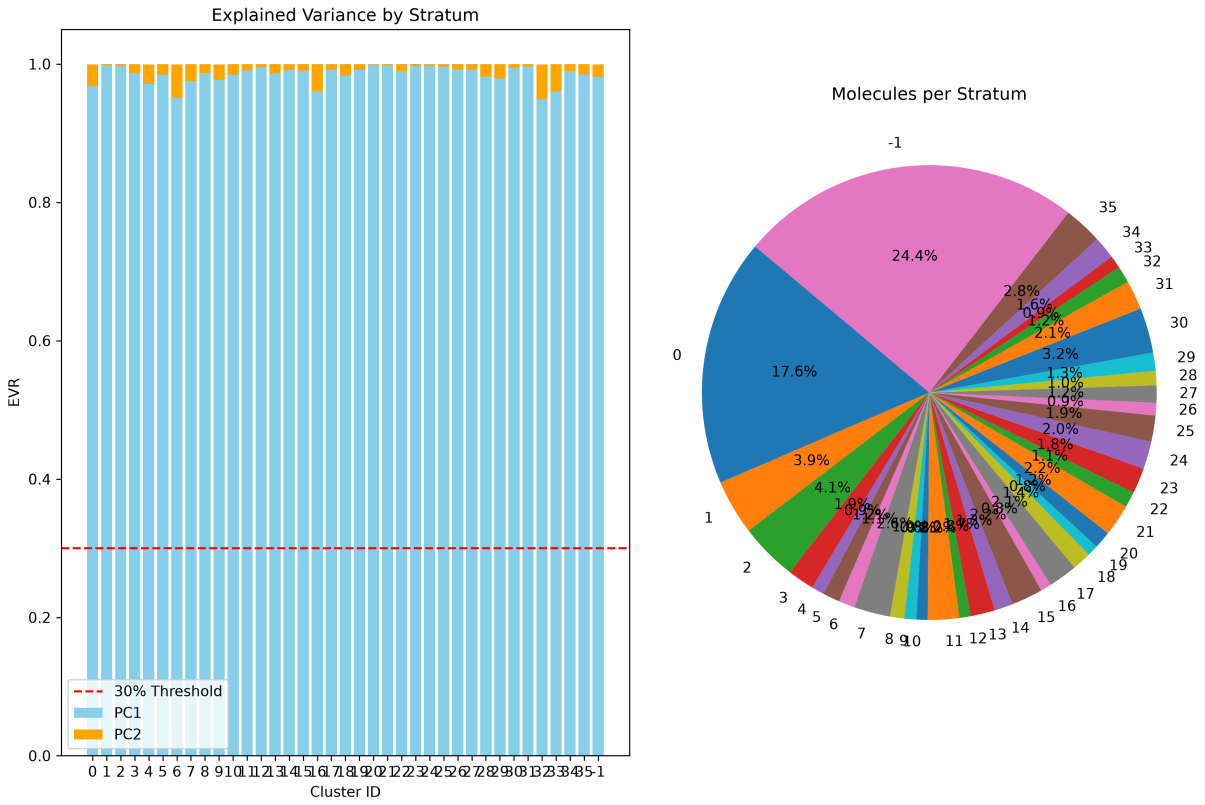

Figure 3: Explained variance ratio (EVR) of the first two principal components (left) and the distribution of molecules (right) across the 37 structural strata identified via HDBSCAN clustering. The left panel demonstrates that the first principal component (PC1) captures the vast majority of the variance in non-frontier properties for all clusters, far exceeding the 30% threshold. This confirms that PC1 serves as a robust representation of the local spatial and vibrational manifold within each specific chemical environment. The right panel illustrates the stratification of the 133,802 molecules in the dataset, revealing an uneven distribution where the largest coherent structural family consists of planar aromatic systems (Cluster 0, 17.6%), while unclassified noise or miscellaneous structures (Cluster -1) account for 24.4% of the chemical space.

Within each stratum, the non-frontier properties (rotational, thermodynamic, and spatial residuals) were condensed into a local manifold using Principal Component Analysis (PCA). As shown in Figure 3, the first principal component (PC1) captures the vast majority of the variance in non-frontier properties for all clusters, serving as a robust representation of the joint response of the molecule's spatial and vibrational state.

3.3 Breakdown of HOMO decoupling across local environments

The coupling between res_{homo} and the local manifold was measured using Partial Distance Correlation ($pdCor$), which accounts for the mathematical dependency between HOMO and LUMO energy residuals by residualizing both against res_{lumo}. A global baseline measurement across the entire dataset yielded a $pdCor$ of 0.161. However, evaluating this metric across the 37 structural clusters reveals significant variation (Figure 4), demonstrating that the global decoupling of HOMO energy is not a universal constant.

Significant rewirings of the electronic-geometric relationship were identified in 21 of the 37 strata ($p < 0.05$ via 500-permutation null checks). The strongest observed coupling occurred in Cluster 24 (Saturated+, Hetero-rich), where the $pdCor$ rose to 0.243, a 51% increase over the global baseline (Figure 5). This suggests that in polar aliphatic systems, where frontier orbitals are dominated by lone pairs and polar bonds,

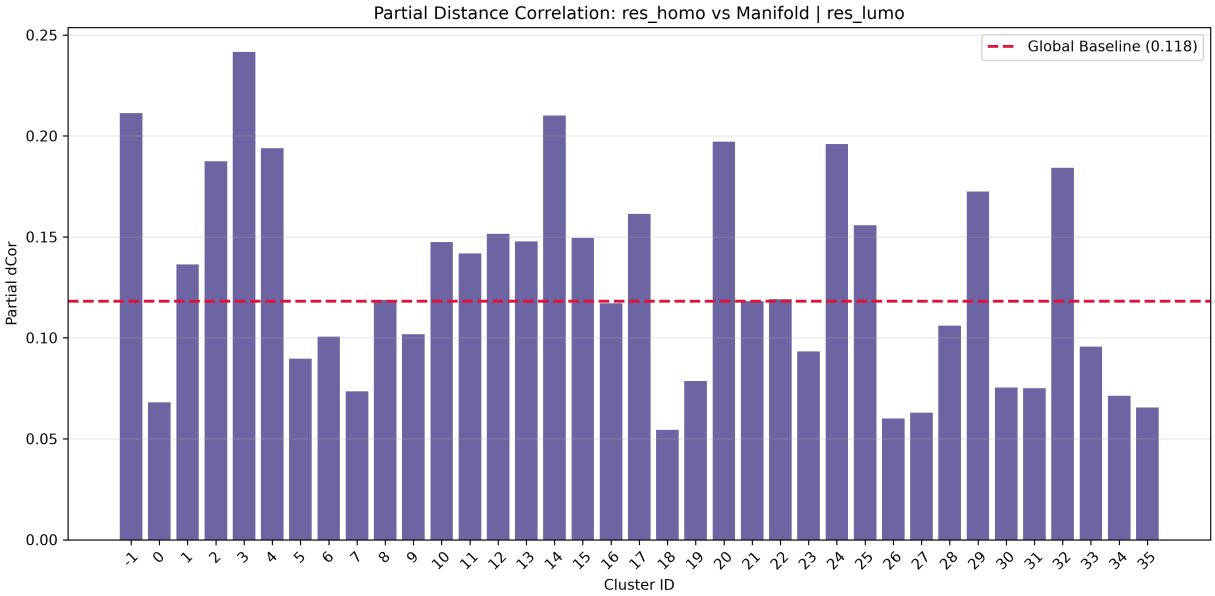

Figure 4: Partial distance correlation ($pdCor$) between HOMO energy residuals (res_{homo}) and the local molecular property manifold, conditioned on LUMO energy residuals (res_{lumo}), across 37 structural clusters. The red dashed line represents the global baseline. The significant variation in $pdCor$ demonstrates that the global decoupling of HOMO energy is not a universal constant but breaks down within specific local environments. Clusters corresponding to saturated, hetero-rich, and cyclic systems exhibit strong coupling that substantially exceeds the global average, indicating that their frontier electronic levels are tightly constrained by molecular geometry. Conversely, aromatic environments (e.g., Cluster 0) remain below the baseline, confirming that delocalized π systems act as an electronic buffer that maintains the independence of frontier orbitals from the vibrational-spatial manifold.

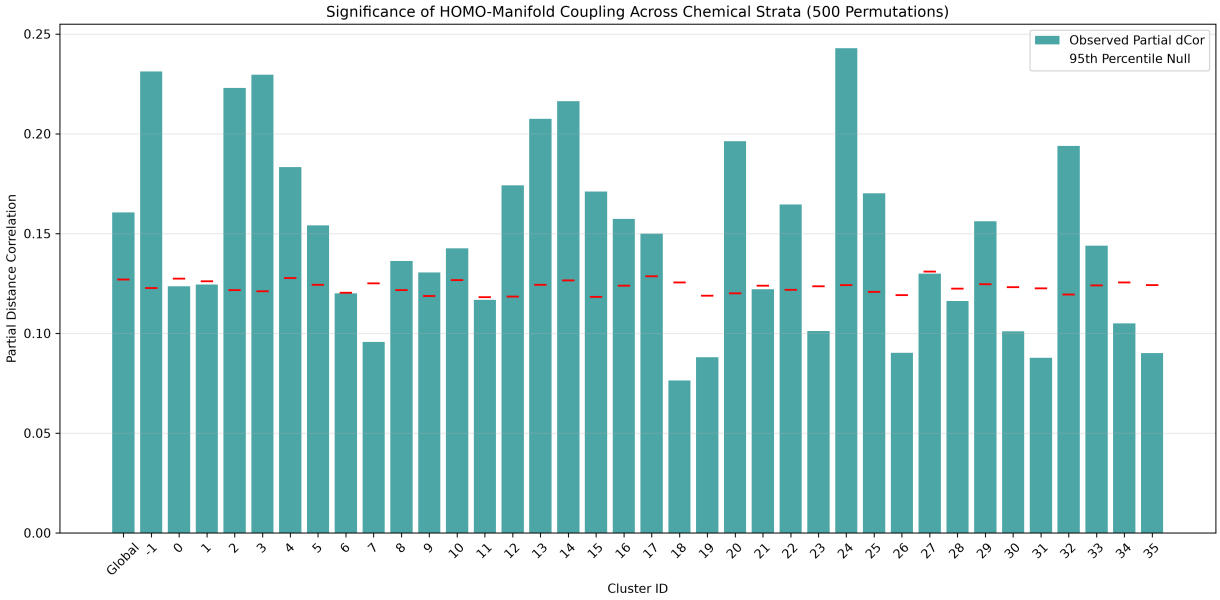

Figure 5: Observed Partial Distance Correlation ($pdCor$) between HOMO energy residuals and the local molecular manifold across 37 chemical strata, compared to the global baseline. Teal bars represent the observed $pdCor$, and red horizontal lines denote the 95th percentile of the null distribution derived from 500 permutations. The plot demonstrates that the assumed global decoupling of HOMO energy is not universal; 21 strata exhibit significant coupling where the observed $pdCor$ strictly exceeds the null threshold. The strongest breakdown of decoupling occurs in Cluster 24 (saturated, heteroatom-rich systems), which shows a 51% increase over the global baseline, indicating that frontier electronic levels in polar aliphatic systems are tightly constrained by molecular geometry. Conversely, large aromatic families like Cluster 0 remain decoupled (observed value below the red threshold), revealing that frontier electronic-spatial coupling is highly dependent on the local chemical environment.

electronic levels are tightly constrained by the overall molecular geometry. Other notable environments showing significant breakdown include Cluster 14 (Conjugated+, Aromatic) with a $pdCor$ of 0.216, and Cluster 3 (Saturated+, Cyclic) with a $pdCor$ of 0.230, pointing to the role of extended conjugation paths and ring strain in modulating frontier orbital levels.

Conversely, the decoupling persists in the most aromatic environments. Cluster 0 (Aromatic+, Conjugated+), despite being the largest coherent family (23,527 molecules), exhibited a $pdCor$ of only 0.124, failing to reach significance ($p = 0.058$) against its null distribution. This indicates that the delocalized π system in planar aromatic molecules acts as an electronic buffer, maintaining independence from the vibrational-spatial manifold.

3.4 Robustness and scaffold holdout validation

The validity of these findings was rigorously tested through permutation-based null distributions and Murcko scaffold holdout validation. For each significant stratum, the observed $pdCor$ strictly exceeded the 95th percentile of the null distribution. For instance, in Cluster 24, the observed value of 0.243 was nearly double the null threshold of 0.124, and Cluster 3 (0.230) similarly exceeded its 0.121 threshold. To ensure the observed coupling was a general feature of the structural motif rather than an artifact of a few highly populated scaffolds, the top five Murcko scaffolds were removed from each significant stratum (Figure 6).

Under this scaffold holdout validation, Cluster 3 proved exceptionally robust, with its $pdCor$ increasing to 0.312 after removing dominant scaffolds (such as `C1OC2CC12` and `C1NC2CC12`). This confirms that the coupling is an intrinsic feature of the "Saturated, Cyclic" motif across diverse structures. Cluster 15 (Saturated+, Cyclic) also showed increased coupling under holdout, rising from 0.171 to 0.217. Across the 17

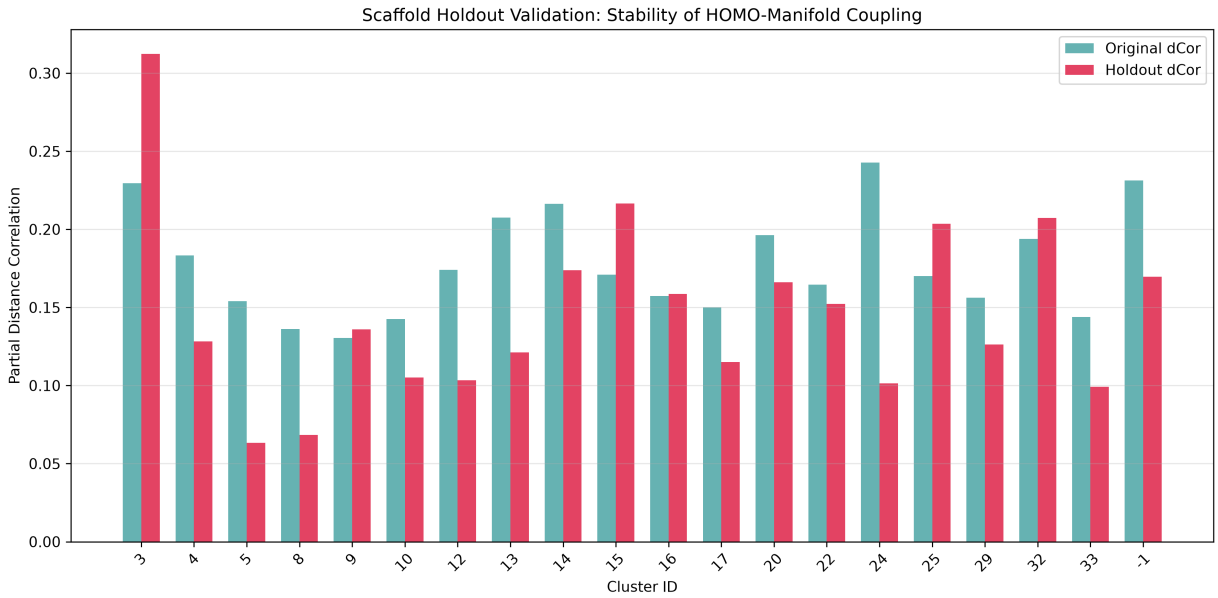

Figure 6: Comparison of the original partial distance correlation (Original dCor) and the scaffold holdout correlation (Holdout dCor) across significant molecular clusters, evaluating the stability of coupling between HOMO energy residuals and the local spatial-vibrational manifold. By removing the top five Murcko scaffolds from each stratum, this validation tests whether the observed coupling is an intrinsic feature of the structural motif or an artifact of highly populated scaffolds. The results demonstrate that the breakdown of HOMO decoupling persists across diverse structures, with environments such as Cluster 3 (Saturated+, Cyclic) and Cluster 15 exhibiting increased correlation under holdout. The overall high retention of correlation across robust strata confirms that the discovered rewiring of electronic-geometric relationships generalizes across molecular families rather than being driven by isolated structural outliers.

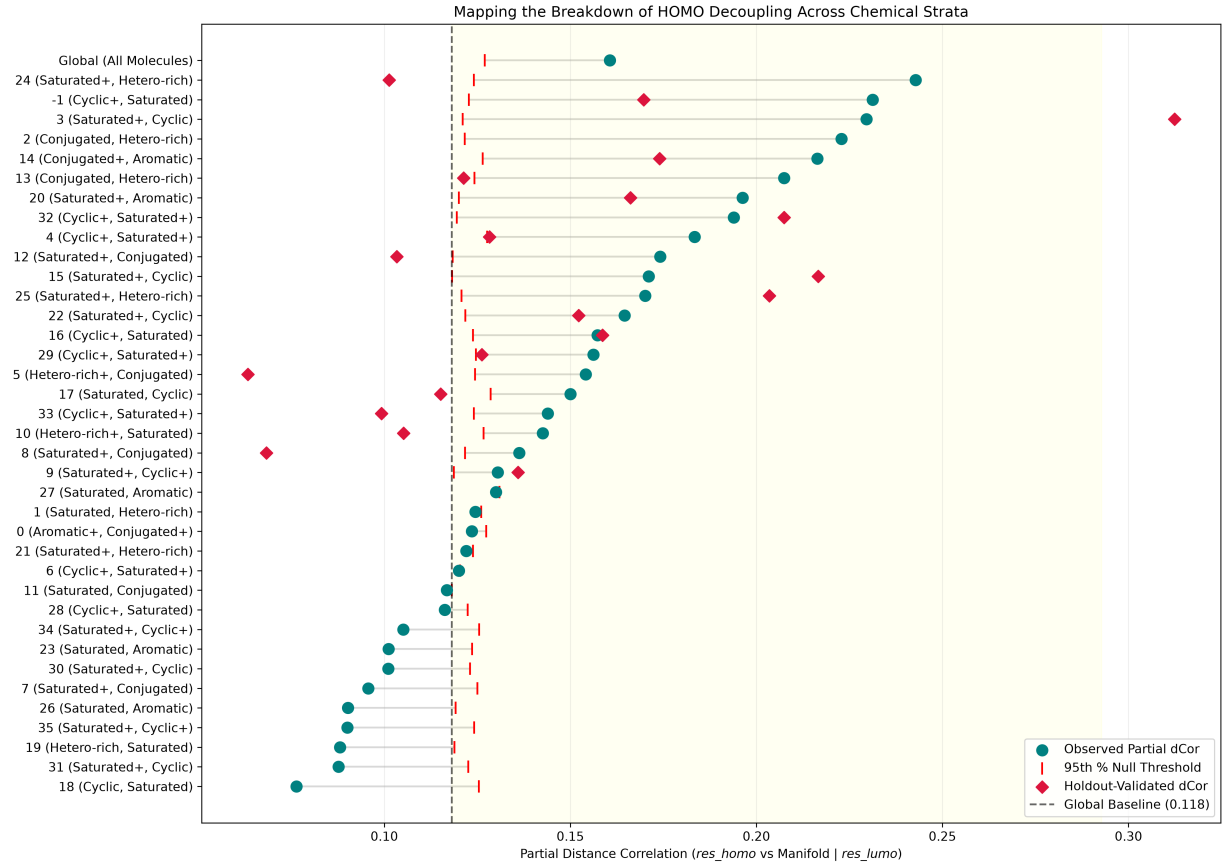

Figure 7: Partial distance correlation ($pdCor$) of HOMO energy residuals against the local spatial-vibrational manifold, conditioned on LUMO residuals, across distinct chemical strata. Teal circles denote the observed $pdCor$ for each structural cluster, evaluated against the 95th percentile null threshold (red vertical lines) and the global baseline (dashed line). The plot demonstrates that HOMO decoupling is not a universal constant but breaks down in specific local environments. Saturated, heteroatom-rich systems (Cluster 24) and saturated, cyclic structures (Cluster 3) exhibit strong electronic-geometric coupling that significantly exceeds null thresholds. Red diamonds represent the $pdCor$ after Murcko scaffold holdout validation, confirming that this rewiring is an intrinsic property of the chemical motifs rather than an artifact of dominant scaffolds, with Cluster 3 showing a particularly marked increase in coupling. In contrast, aromatic environments (e.g., Cluster 0) fall below the significance threshold, indicating that delocalized π systems act as buffers that preserve the independence of the frontier electronic state from the geometric manifold.

strata identified as robust (significant $p < 0.05$ and holdout $pdCor > 0.1$), the mean stability ratio (Holdout/Original) was 0.823. The final forest plot (Figure 7) summarizes these validated couplings, confirming that the discovered rewiring of electronic-geometric relationships generalizes across molecular families and providing a targeted map for where joint property models must explicitly account for frontier-electronic coupling.

4 Conclusions

The assumption that the energy of the highest occupied molecular orbital (HOMO) is universally decoupled from the spatial and thermodynamic property manifolds is a byproduct of treating vast chemical spaces as monolithic distributions. By extracting non-linear stoichiometric baselines and stratifying the QM9 dataset into discrete structural environments, this study demonstrates that the physical isolation of the HOMO is

highly conditional.

Our conditional distance correlation mapping reveals that the electronic-geometric relationship rewires significantly in specific local environments. In saturated, heteroatom-rich systems (such as polar aliphatics) and strained cyclic structures, the frontier orbitals are tightly constrained by the overall molecular geometry. This breakdown of decoupling is evidenced by partial distance correlations that substantially exceed the global baseline and remain robust under rigorous scaffold holdout validation.

Conversely, we found that the apparent global decoupling is primarily driven by highly aromatic and conjugated systems. In these planar environments, delocalized π networks act as electronic buffers, maintaining the independence of the frontier electronic state from the vibrational-spatial manifold. Table 1 summarizes these contrasting structural regimes, highlighting the distinct chemical motifs that either enforce or break frontier-electronic isolation.

Ultimately, the identification of 17 robust environments where HOMO decoupling breaks down provides a targeted map for future computational chemistry efforts. Joint property models and machine learning architectures must explicitly account for frontier-electronic coupling in aliphatic and polar systems, moving beyond the traditional, one-size-fits-all assumption of HOMO orthogonality.