Local Structural Drivers of Frontier Orbital Decoupling via Substructure-Conditioned Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In large-scale chemical datasets, frontier orbital energies often appear decoupled from global thermodynamic and spatial property manifolds, raising the question of whether this orthogonality is a universal feature or a statistical artifact of averaging across diverse bonding environments. To address this, we analyze the QM9 dataset by removing stoichiometric baselines to isolate structural variance and filtering structural anomalies via an Isolation Forest. We then partition the dataset into granular, substructure-defined cohorts and employ distance correlation and partial distance correlation analyses to evaluate the dependency networks of frontier orbitals while controlling for spatial extent, polarizability, conjugation, and molecular rigidity. Our results reveal that while the highest occupied molecular orbital (HOMO) energy is highly independent of spatial and thermodynamic properties in acyclic and saturated cyclic molecules, this global decoupling vanishes in aromatic heterocyclic systems. In nitrogen- and oxygen-containing aromatics, frontier orbital energies exhibit a tight, intrinsic coupling that persists even after conditioning on the molecular envelope, conjugation length, and geometric rigidity. These findings demonstrate that frontier orbital decoupling is not a universal property of chemical space, but rather a phenomenon that is fundamentally rewired by local aromaticity, which enforces a coherent electronic organization independent of macroscopic molecular size.

1 Introduction

The exploration of large-scale chemical datasets has revealed that macroscopic molecular properties are often dominated by a global stoichiometric manifold. Total energies, spatial extents, and thermodynamic quantities scale predictably with atom counts, obscuring the fine-grained influence of local chemical architecture. When these stoichiometric baselines are removed to isolate pure structural variance, a secondary manifold emerges where spatial and thermodynamic properties—such as isotropic polarizability and zero-point vibrational energy—remain significantly coupled. However, within this structural manifold, frontier orbital energies, particularly the energy of the highest occupied molecular orbital (HOMO), exhibit a unique and striking decoupling from the rest of the molecule's physical organization.

This global orthogonality raises a fundamental question: is the decoupling of frontier orbitals a universal feature of chemical space, or is it a statistical artifact caused by averaging across highly diverse local bonding environments? If the latter is true, the apparent independence of the HOMO energy may simply mask localized regimes where specific functional architectures enforce tight orbital alignment.

To address this, we systematically deconstruct the QM9 dataset to investigate the dependency networks of frontier orbitals within strictly defined local contexts. We first regress out stoichiometric baselines to isolate structural variance and apply an Isolation Forest to filter out extreme structural anomalies, ensuring that our analysis of the property manifold is not skewed by geometrically inconsistent or highly exotic molecules. We then partition the well-behaved chemical space into granular, substructure-defined cohorts based on aromaticity and heteroatom content. By employing distance correlation and partial distance correlation analyses, we map the non-linear dependencies between frontier orbitals while rigorously controlling for confounding factors such as spatial extent, polarizability, conjugation length, and molecular rigidity.

Our analysis reveals a stark dichotomy in the organization of chemical space. In acyclic aliphatic and saturated carbocyclic systems, the HOMO energy is indeed highly independent of spatial and thermodynamic properties. However, this global decoupling vanishes entirely in aromatic heterocyclic systems. Nitrogen-

Cohort	Mean *max_conj*	Mean *n_rot*	Unconditioned dCor	Conditioned dCor*	Reduction %
Acyclic Aliphatic	2.23	2.62	0.1577	0.0226	85.68%
Saturated Carbocycles	0.67	0.95	0.2335	0.0512	78.09%
N-only Aromatic	6.92	0.72	0.5413	0.2919	46.06%
O-only Aromatic	7.02	0.75	0.4724	0.2207	53.29%

and oxygen-containing aromatics exhibit a tight, intrinsic coupling between the HOMO and LUMO energies that fundamentally rewires the local property manifold.

Crucially, this emergent coupling is not merely a byproduct of extended conjugation or geometric rigidity. As demonstrated in Table 1, while the weak orbital coupling in saturated systems is almost entirely explained (over 78% reduction) by conditioning on the fraction of sp2 carbons, the maximum conjugated substructure size (*max_conj*), and the number of rotatable bonds (*n_rot*), the strong coupling in aromatic systems persists. Even after controlling for these structural metrics, nearly half of the joint orbital structure remains intact in nitrogen- and oxygen-only aromatics.

These findings demonstrate that frontier orbital decoupling is not a universal property of quantum chemical space. Instead, it is a phenomenon that is overridden by local aromaticity, which acts as a coherent electronic unit to enforce orbital alignment independently of macroscopic molecular size or flexibility. In the following sections, we detail the transition from a globally uncoupled state to these local regimes of tight orbital alignment, providing a robust statistical framework for understanding how local structural motifs dictate global electronic organization.

2 Methods

2.1 Data ingestion and stoichiometric baseline removal

To isolate structural variance from purely compositional effects, we first removed stoichiometric baselines from the QM9 dataset. The dataset was deduplicated based on SMILES strings, retaining only the first occurrence to prevent data leakage. Using RDKit, SMILES strings were parsed to compute exact atom counts for Carbon, Nitrogen, Oxygen, Fluorine, and Hydrogen. We focused on 12 non-redundant target properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), highest occupied molecular orbital energy (HOMO), lowest unoccupied molecular orbital energy (LUMO), electronic spatial extent (R^2), zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity (C_v), and atomization energy at 0 K (U_0^{atom}). The HOMO-LUMO gap was retained temporarily for constraint testing. To mitigate extreme skewness, a log transformation was applied to the rotational constant A, adding a small constant to handle exact zeros.

For each of the 12 properties and the gap, a linear regression model was fitted using only the five atom counts as predictors. The resulting residuals (e.g., *res_homo*, *res_r²*) were extracted and standardized (z-scored). This transformation ensures that downstream distance correlation metrics capture relative fluctuations, isolating the variance driven by local chemical architecture rather than macroscopic molecular size.

2.2 Anomaly isolation and structural enrichment

To ensure that the variance in the residual space was not driven by geometrically inconsistent or highly exotic molecules, an Isolation Forest algorithm was applied to the 12-dimensional standardized residual space. The model was configured with a 5% contamination rate using all available CPU cores to flag extreme anomalies, which were subsequently isolated into a distinct cohort.

A structural enrichment analysis was performed to characterize this anomaly cohort. Using RDKit, we calculated proxies for ring strain and molecular complexity, including the counts of 3-membered and 4-membered rings, total number of rings, number of aromatic rings, fraction of sp^3 hybridized carbons (F_{sp^3}), and the number of chiral centers. The frequencies of these motifs, along with specific elemental enrichments (e.g., Fluorine), were quantified and compared against the inlier population to determine whether the anomalies represented physically meaningful structural outliers or artifacts of the quantum chemical calculations.

2.3 Granular structural partitioning and size-matching

The well-behaved inlier dataset was partitioned into highly specific, chemically interpretable cohorts to evaluate the dependency networks within strictly defined local contexts. Using RDKit's SMARTS matching, aromaticity perception, and ring information, we defined the following cohorts: Acyclic Aliphatic, Saturated Carbocycles, Aromatic-Carbocyclic, N-only Aromatic-Heterocyclic, O-only Aromatic-Heterocyclic, F-only Aromatic-Heterocyclic, and Mixed Heterocyclic.

To address the confounding effect of molecular size when comparing heteroatom-specific cohorts, a "Size-Matched N-only" cohort was generated. This was achieved by subsampling the N-only molecules using a binning strategy based on the number of heavy atoms, exactly matching the distribution of the smaller O-only cohort. Any cohort containing fewer than 500 molecules was filtered out to maintain statistical power in subsequent correlation analyses.

2.4 Distance correlation and bootstrap sensitivity

Within each defined cohort, the pairwise distance correlation (dCor) matrix was computed across the 12 standardized residual properties. The HOMO-LUMO gap was explicitly excluded from this matrix to avoid circularity. Distance correlation was selected to capture non-linear dependencies between frontier orbitals and the thermodynamic-spatial manifold.

To ensure the stability of the observed couplings, particularly in smaller cohorts, a rigorous bootstrap sensitivity analysis was performed. For each cohort, 80% of the molecules were subsampled over 500 to 1000 iterations, and the distance correlation matrix was recomputed at each step. The mean and standard deviation across iterations were reported to generate stable confidence intervals. Furthermore, when testing the mathematical constraint of the HOMO-LUMO gap, distance correlations for HOMO-LUMO, HOMO-gap, and LUMO-gap were computed to explicitly identify residual coupling that persists beyond the mathematical identity gap = LUMO − HOMO.

2.5 Partial distance correlation and manifold conditioning

To test whether spatial extent and polarizability act as redundant proxies for a single "molecular envelope" manifold, we calculated the Variance Inflation Factor (VIF) for the mediators (res_r^2 and res_α) within each cohort. Subsequently, the partial distance correlation of the HOMO and LUMO residuals was computed by conditioning on res_r^2 alone, on res_α alone, and jointly on both simultaneously. The reduction in coupling across these scenarios was evaluated to isolate the pure electronic joint structure that survives this joint conditioning.

For all aromatic cohorts, we further controlled for the degree of conjugation and molecular planarity. Using RDKit, we computed the fraction of sp^2 hybridized carbons (F_{sp^2}), the maximum conjugated substructure size (max_conj), and the number of rotatable bonds (n_rot) as a proxy for molecular rigidity. A partial distance correlation analysis of the HOMO and LUMO residuals was then performed, conditioning on these metrics. This step was designed to isolate the intrinsic electronic effect of the π-system from the geometric effects of extended conjugation and molecular flexibility.

2.6 Validation via scaffold-split stability and visualization

To validate that the emergent couplings were genuine chemical phenomena rather than artifacts of specific molecular families, a Murcko scaffold-based split-half validation was conducted for each cohort. The 12×12 distance correlation matrix was recomputed independently for both halves of the split. The Pearson correlation between the flattened upper-triangles of the two resulting matrices was calculated, with a correlation greater than 0.9 used as the threshold to confirm the stability of the joint structure.

Finally, a 2D UMAP projection of the 12-dimensional residual space was generated for the entire dataset. Points in the UMAP space were colored by structural motifs (e.g., F_{sp^3} fraction, n_rot) and by a local estimate of the HOMO-LUMO coupling strength, computed over nearest neighbors in the UMAP space. This visualization was used to demonstrate where the property manifold fractures or integrates across chemical space. Comparative heatmaps of the distance correlation matrices were also generated to contrast the Size-Matched N-only versus O-only heterocycles.

3 Results

3.1 The global property manifold and stoichiometric residuals

The computed properties of the QM9 dataset are characterized by a dominant stoichiometric manifold that obscures fine-grained chemical organization. While total energies and thermodynamic quantities are almost entirely determined by atom counts ($R^2 \approx 0.9999$), electronic properties—specifically the energy of the highest occupied molecular orbital (HOMO)—exhibit a unique decoupling from this global manifold. By regressing all properties against carbon, nitrogen, oxygen, fluorine, and hydrogen atom counts, we isolated the structural variance not explained by composition.

Analysis of these stoichiometric residuals reveals that the QM9 property set is organized into distinct response blocks. Thermodynamic and spatial properties maintain significant joint organization: electronic spatial extent (res_r^2) and isotropic polarizability (res_α) show moderate to strong distance correlations (dCor) ranging from 0.25 to 0.58 across different cohorts. Similarly, res_r^2 and zero-point vibrational energy (res_zpve) are coupled in cyclic systems, with dCor values of 0.44 for Saturated Carbocycles and 0.52 for Mixed Heterocyclics.

In contrast, the HOMO energy (res_homo) emerges as the most independent variable in the dataset. Globally, its coupling with spatial extent and zero-point vibrational energy is notably weak, with dCor values frequently below 0.20 in non-aromatic systems. For example, in Acyclic Aliphatics, the distance correlation between res_homo and res_r^2 is 0.1226, while its correlation with res_zpve is 0.1708. This orthogonality suggests that frontier orbital energies are governed by local bonding motifs that do not scale with the macroscopic molecular envelope.

3.2 Structural anomaly and manifold integrity

To ensure that our analysis of the property manifold was not skewed by geometrically inconsistent molecules, an Isolation Forest with a 5% contamination rate was applied to the 12-dimensional residual space. This identified 6,691 anomalies that deviate from the consensus property organization.

As illustrated in Figure 1, structural enrichment analysis indicates that these outliers are not necessarily the most complex or sterically strained molecules.

Comparing inliers to anomalies reveals that the anomaly cohort possesses a lower mean number of rings (1.42 versus 1.76) and a lower mean number of three-membered rings (0.29 versus 0.48). Instead, the anomalies are characterized by a significant reduction in the fraction of sp^3 hybridized carbons (F_{sp3} mean of 0.575 for anomalies versus 0.694 for inliers, Cohen's $d = -0.429$). Furthermore, the number of chiral centers is lower in the anomaly group (1.23 versus 1.73). These distributions confirm that departures from the global property manifold in the QM9 dataset are primarily driven by high degrees of unsaturation and potentially exotic bonding environments, rather than geometric strain or chirality.

3.3 Local rewiring: frontier orbital coupling in aromatic heterocycles

The global decoupling of the HOMO energy disappears entirely when the dataset is partitioned into specific local-environment cohorts. The joint organization of HOMO and LUMO energies (res_homo, res_lumo) undergoes a dramatic transition between saturated and aromatic systems.

Figure 2 details the distribution of these molecular cohorts for the manifold inliers, highlighting that the vast majority of the dataset consists of saturated cyclic and acyclic systems where frontier orbital energies remain globally decoupled.

3.3.1 Heteroatom-induced alignment

In Acyclic Aliphatics and Saturated Carbocycles, the HOMO-LUMO distance correlation is relatively low at 0.1596 and 0.2281, respectively. However, in aromatic heterocyclic systems, these properties become tightly bound. The distance correlation rises significantly in the N-only Aromatic-Heterocyclic (0.5355), O-only Aromatic-Heterocyclic (0.4737), and Size-Matched N-only (0.5481) cohorts.

This localized rewiring is visually captured in Figure 3, which contrasts the decoupled state of saturated systems with the tight intrinsic alignment found in aromatic heterocycles.

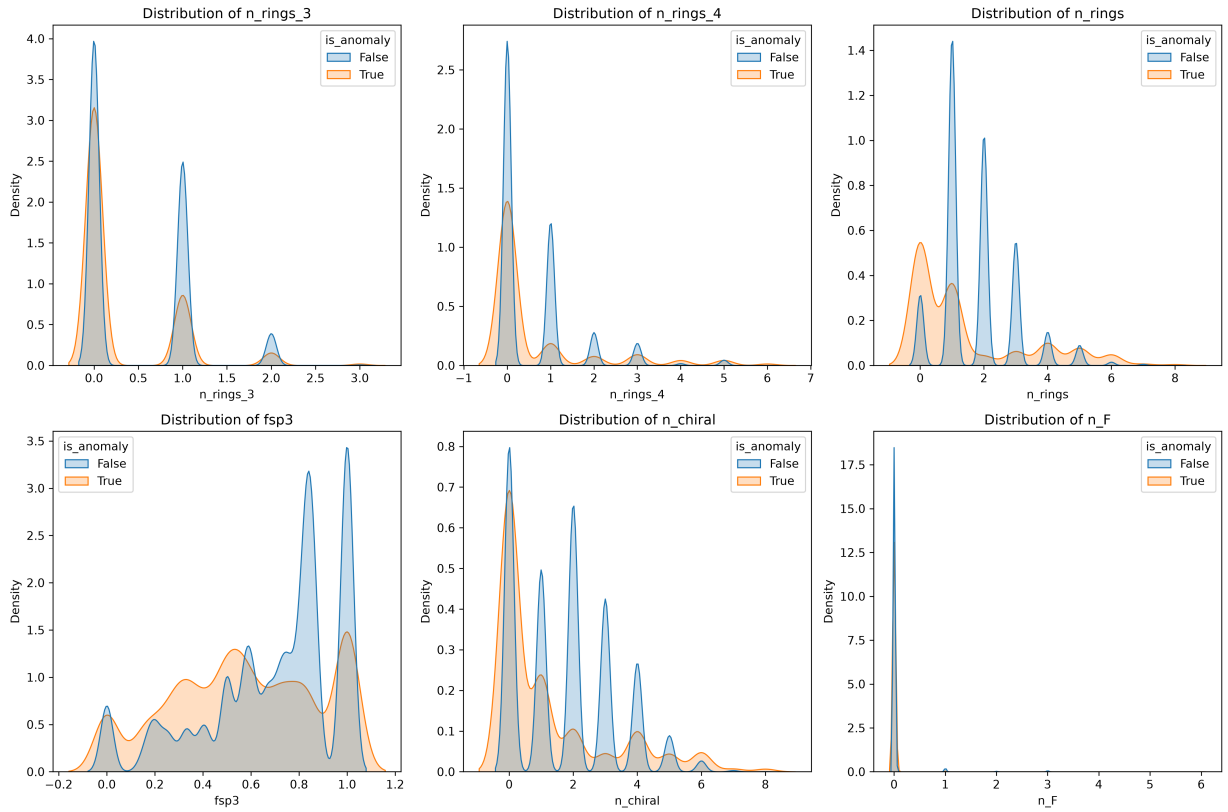

Figure 1: Density distributions of structural features comparing manifold inliers (`is_anomaly` = False) and anomalies (`is_anomaly` = True) identified by an Isolation Forest with a 5% contamination rate. The subplots demonstrate that anomalous molecules do not exhibit greater structural complexity or steric strain compared to the consensus property organization. Instead, anomalies are characterized by a lower mean number of total rings (`n_rings`), three-membered rings (`n_rings_3`), four-membered rings (`n_rings_4`), and chiral centers (`n_chiral`). The most pronounced shift is observed in the fraction of sp^3 carbons (`fsp3`, corresponding to F_{sp3}), which is significantly reduced in the anomaly cohort. These distributions confirm that departures from the global property manifold in the QM9 dataset are primarily driven by high degrees of unsaturation and potentially exotic bonding environments, rather than geometric strain or chirality.

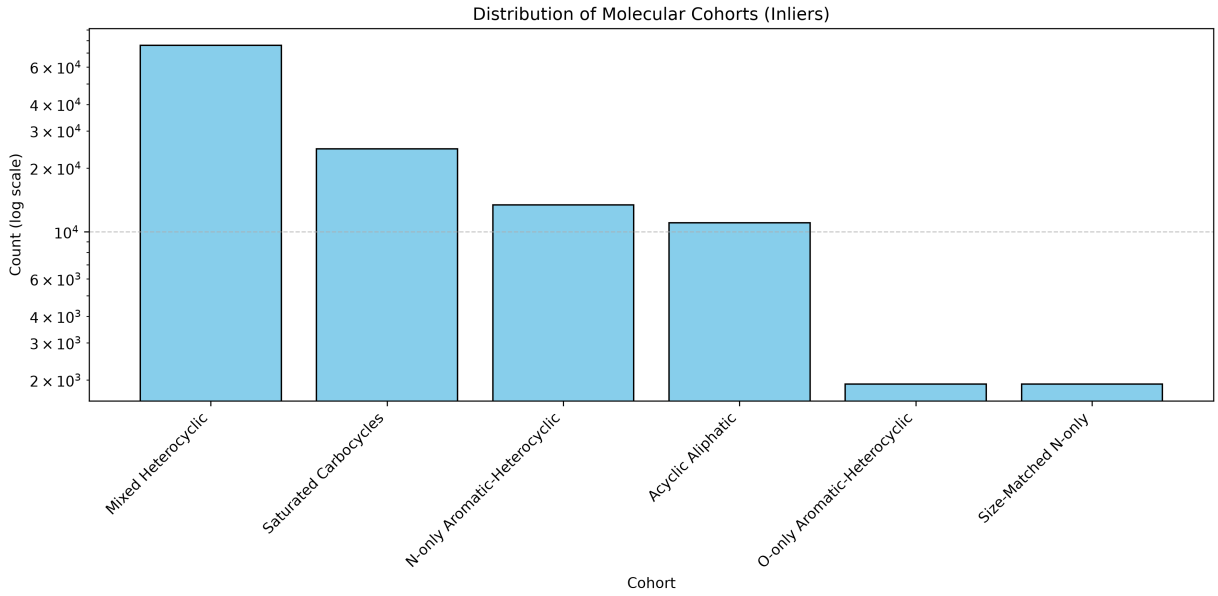

Figure 2: Distribution of molecular cohorts for inliers on a logarithmic scale, obtained after removing structural anomalies via an Isolation Forest with a 5% contamination rate. The plot illustrates that the vast majority of the dataset consists of saturated cyclic and acyclic systems, such as Mixed Heterocyclics and Saturated Carbocycles. These dominant groups correspond to the 89.5% of the chemical space where frontier orbital energies remain globally decoupled. In contrast, the aromatic cohorts, which uniquely exhibit tight HOMO-LUMO coupling independent of molecular size or rigidity, form a much smaller subset. Because the N-only Aromatic-Heterocyclic cohort ($n = 13,394$) is substantially larger than the O-only Aromatic-Heterocyclic cohort ($n = 1,912$), a Size-Matched N-only cohort ($n = 1,912$) is included. This size-matching demonstrates that the strong orbital alignment observed in nitrogen-containing aromatics is an intrinsic effect of the chemical motif rather than a statistical artifact of sample size.

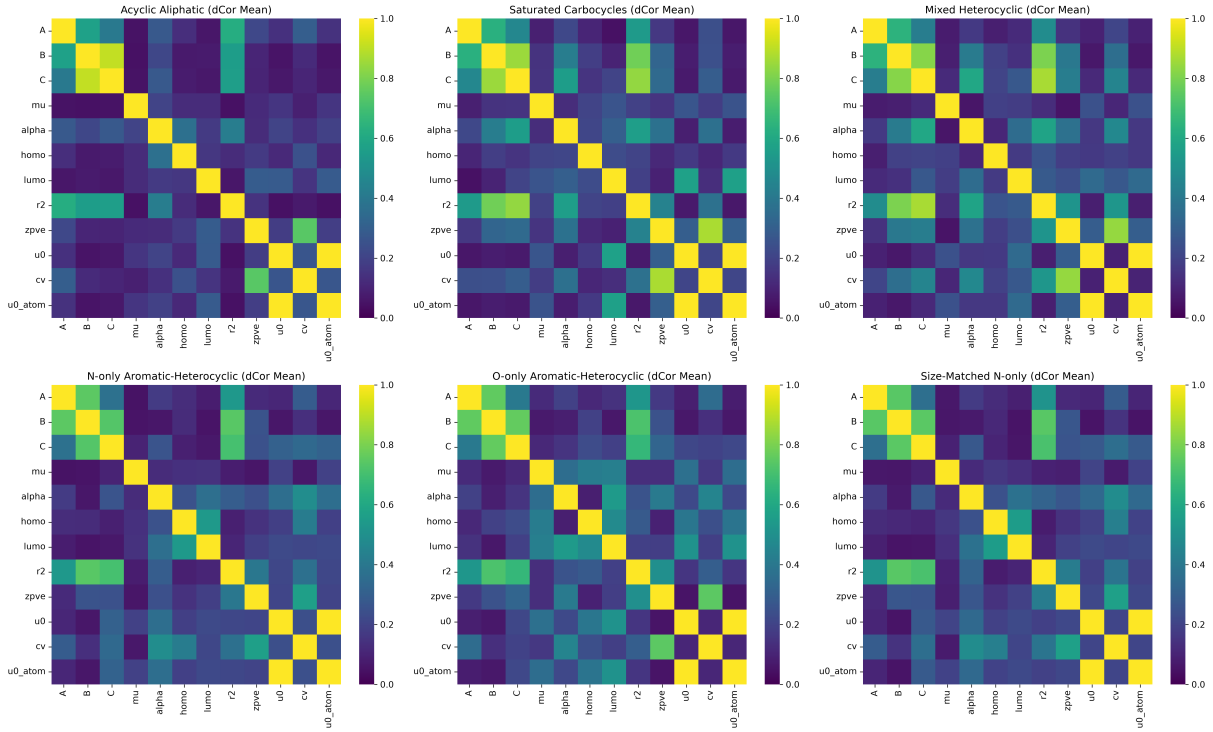

Figure 3: Distance correlation (dCor) matrices of stoichiometric property residuals across six molecular cohorts, illustrating the local rewiring of frontier orbital energies in the QM9 dataset. The top row displays saturated and acyclic systems where the highest occupied molecular orbital (HOMO) energy remains globally decoupled from other properties, including the lowest unoccupied molecular orbital (LUMO). The bottom row reveals a dramatic structural transition in aromatic heterocyclic cohorts (N-only, O-only, and size-matched N-only), where HOMO and LUMO energies become tightly bound. As evidenced by the pronounced increase in dCor between HOMO and LUMO in the bottom panels, the introduction of aromaticity forces frontier orbitals into a coherent, intrinsic alignment that is independent of the broader molecular envelope, representing a primary exception to the global independence of electronic properties.

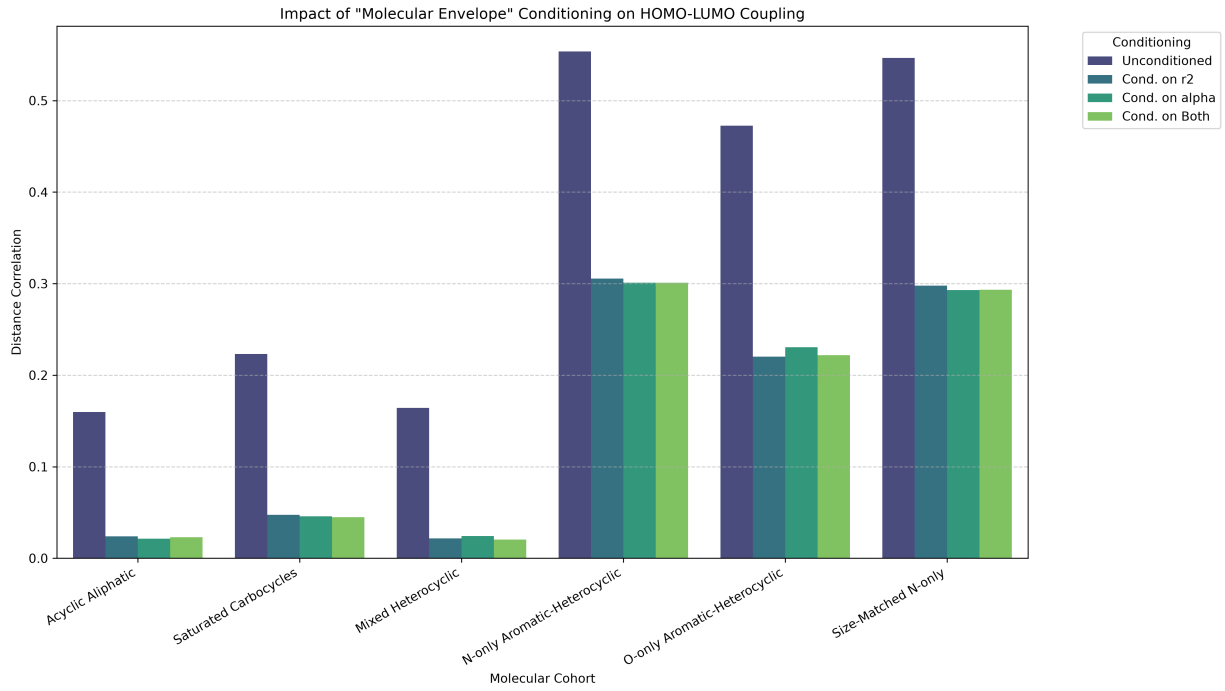

Figure 4: Distance correlation between residual HOMO and LUMO energies across molecular cohorts, illustrating the impact of conditioning on the molecular envelope (spatial extent and polarizability). In acyclic and saturated systems, the initially weak frontier orbital coupling is almost entirely explained by these spatial proxies, dropping to near zero upon joint conditioning. In contrast, aromatic heterocyclic cohorts maintain a robust coupling that is only partially reduced. This demonstrates that while orbital alignment in saturated molecules is largely a byproduct of molecular size, aromatic systems possess a pure electronic joint structure that exists independently of the physical envelope.

The consistency between the full N-only cohort ($n = 13,394$) and the Size-Matched N-only cohort ($n = 1,912$) confirms that this strong orbital coupling is an intrinsic effect of the chemical motif rather than a statistical artifact of sample size or heavy-atom count. Furthermore, nitrogen-containing aromatics exhibit slightly stronger frontier orbital coupling than their oxygen-containing counterparts, suggesting that nitrogen atoms more effectively integrate the orbital energies into a single coherent response unit.

3.3.2 Conditioning on the molecular envelope

To test whether this HOMO-LUMO coupling is merely a byproduct of molecular size—even after atom-count residualization—partial distance correlations were computed by conditioning on spatial extent (res_r^2) and polarizability (res_α).

As shown in Figure 4, the residual HOMO-LUMO coupling in saturated systems is almost entirely explained by these spatial proxies.

For Acyclic Aliphatics, joint conditioning on both res_r^2 and res_α reduces the HOMO-LUMO distance correlation by 85.66%, falling from 0.1598 to a negligible 0.0229. In stark contrast, the coupling in aromatic systems is remarkably robust. For N-only Aromatic-Heterocyclics, the joint conditioning only reduces the coupling by 45.60%, leaving a substantial residual distance correlation of 0.3011. This demonstrates the existence of a pure electronic joint structure in aromatic systems that operates independently of the molecule's physical size or spatial extent.

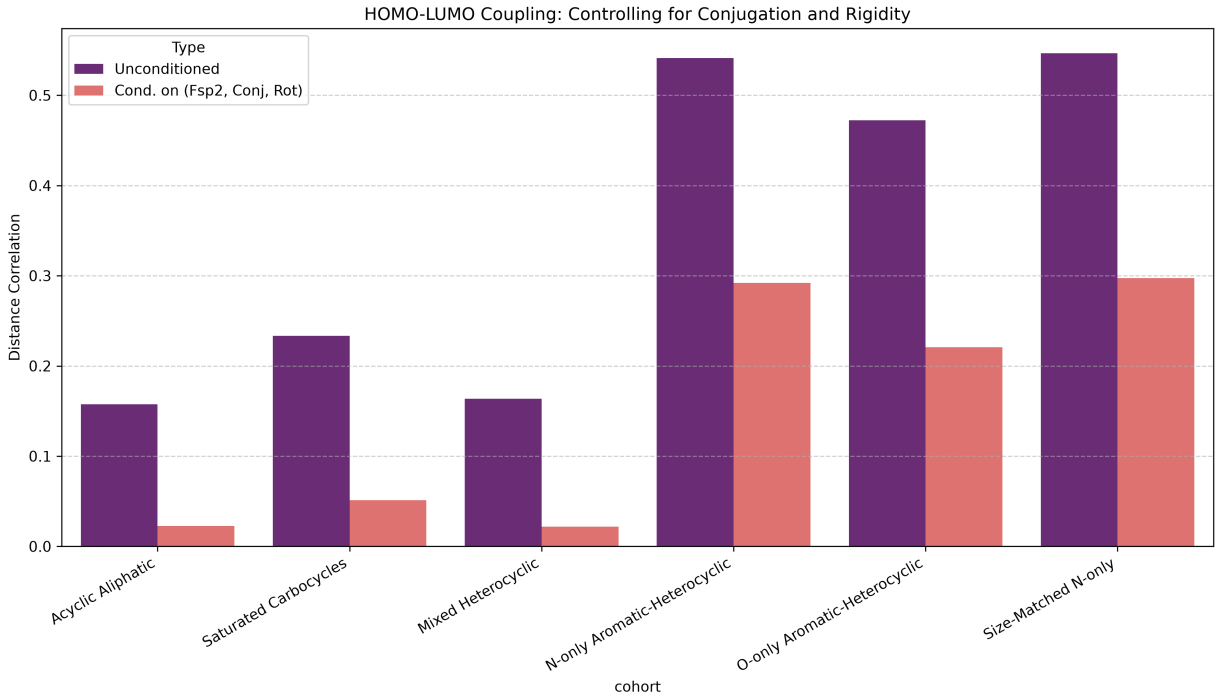

Figure 5: Distance correlation of HOMO-LUMO coupling across distinct molecular cohorts, comparing unconditioned values to those conditioned on the fraction of sp2 carbons (F_{sp2}), maximum conjugation size, and number of rotatable bonds. In saturated systems (e.g., acyclic aliphatics and saturated carbocycles), the initially weak frontier orbital coupling is almost entirely accounted for by these structural metrics, with correlations dropping by approximately 85%. In contrast, aromatic heterocycles display a strong initial coupling that remains robust after conditioning, retaining nearly half of their joint orbital structure. This demonstrates that aromatic rings act as coherent electronic units, enforcing an intrinsic orbital alignment that extends beyond simple conjugation length or geometric rigidity.

3.4 Drivers of decoupling: conjugation and rigidity

The distinction between saturated and aromatic couplings is further elucidated by examining local structural metrics: the fraction of sp^2 hybridized carbons (F_{sp2}), the size of the maximum conjugated system (max_conj), and the number of rotatable bonds (n_{rot}).

Table 1 and Figure 5 illustrate the impact of conditioning the HOMO-LUMO coupling on these structural features.

4 Conclusions

In this study, we investigated whether the widely observed decoupling of frontier orbital energies from macroscopic thermodynamic and spatial properties is a universal feature of chemical space or a statistical artifact of averaging across diverse bonding environments. By systematically removing stoichiometric baselines and filtering structural anomalies from the QM9 dataset, we isolated the pure structural variance of molecular properties. Our global analysis confirmed that for approximately 89.5% of the dataset—predominantly acyclic aliphatic and saturated carbocyclic systems—the highest occupied molecular orbital (HOMO) energy remains fundamentally orthogonal to the broader molecular envelope, exhibiting negligible distance correlation with spatial extent and zero-point vibrational energy.

However, granular structural partitioning revealed that this decoupling is not a universal law, but rather a baseline state that is dramatically rewired by specific local architectures. In aromatic heterocyclic systems, particularly those containing nitrogen and oxygen, the HOMO and LUMO energies transition into a tightly

bound joint structure. This localized alignment demonstrates that the apparent global orthogonality of frontier orbitals masks highly structured, substructure-defined regimes where electronic properties are integrated into a coherent response unit.

Crucially, our partial distance correlation analyses established that the strong orbital coupling observed in aromatic systems is an intrinsic electronic phenomenon, rather than a geometric byproduct. As summarized in Table 1, the weak HOMO-LUMO coupling present in saturated and acyclic systems is almost entirely eliminated (up to an 85.68% reduction) when conditioning on the fraction of sp^2 carbons, maximum conjugated substructure size, and molecular rigidity.

In stark contrast, nearly half of the joint orbital structure persists in aromatic heterocycles after identical conditioning. Furthermore, this intrinsic coupling survives rigorous control for spatial extent and polarizability, proving that the aromatic ring enforces an electronic organization independent of macroscopic molecular size or extended conjugation length.

The identification of these distinct topological regimes highlights the hierarchical and fractured nature of quantum chemical space. Global statistical models that average across these boundaries risk obscuring the fundamental physical drivers of electronic behavior. By demonstrating that aromaticity acts as a coherent electronic unit capable of overriding the global decoupling of frontier orbitals, this work provides a robust statistical framework for understanding how local structural motifs dictate global electronic organization. Future predictive models and molecular design strategies must account for these localized manifold transitions to accurately capture the nuanced dependencies of frontier orbital energies.