Identifying Structural Triggers of Electronic Coupling in the QM9 Dataset via Residual Space Analysis

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy typically exhibits extreme orthogonality to other physical properties, remaining uniquely isolated across chemical space. It remains unclear whether this electronic decoupling is a universal baseline or if it breaks down under specific architectural conditions. To investigate this, we utilize the QM9 dataset and train non-linear gradient-boosted regressors to extract stoichiometric residuals for 11 non-redundant physical properties, effectively stripping out extensive size effects. By grouping molecules into Murcko scaffolds, we compute the multivariate distance correlation between the HOMO residual and the remaining 10-dimensional property space to identify localized coupling neighborhoods. Our analysis reveals that while the HOMO energy is globally decoupled, this isolation vanishes within specific structural families. We find that molecular rigidity, indicated by fewer rotatable bonds, along with higher fractions of sp^3 carbons and increased ring counts, act as primary architectural triggers that force the electronic state to synchronize with the molecular framework. Furthermore, a high-strain outlier subset comprising one percent of the population exhibits extreme coupling and is significantly enriched in both fluorine substitutions and rotatable bonds. Granular analysis within highly coupled scaffolds demonstrates that the HOMO energy synchronizes strongly with isotropic polarizability and dipole moment, while block-wise partial distance correlations reveal that its unique coupling is primarily driven by the molecule's thermodynamic and geometric-vibrational states rather than rotational geometry. These findings provide a robust framework for understanding exactly where and why electronic isolation breaks down in molecular property spaces.

1 Introduction

The exploration of chemical space relies heavily on understanding the joint distribution of molecular properties. In large-scale computational databases such as the QM9 dataset, which contains extensive quantum mechanical properties for small organic molecules, most physical and thermodynamic features exhibit strong interdependencies. However, the highest occupied molecular orbital (HOMO) energy presents a striking anomaly. It typically exhibits extreme orthogonality to other physical properties, remaining uniquely isolated across the broader chemical space.

This profound electronic decoupling raises a fundamental question: is the isolation of the HOMO energy a universal baseline across all organic molecules, or does it break down under specific architectural conditions? Identifying the boundaries of this orthogonality is crucial for understanding structure-property relationships and for the targeted design of molecules where electronic states must be precisely tuned in tandem with thermodynamic or geometric constraints.

To investigate the limits of this electronic isolation, we introduce a local sensitivity analysis within a property residual space. Because extensive properties are heavily dominated by trivial molecular size and stoichiometric effects, we first train non-linear gradient-boosted regressors to predict 11 non-redundant physical properties using only basic atom counts. By extracting the residuals from these models, we effectively strip out stoichiometric variance, isolating the fluctuations driven purely by molecular architecture. We then group the dataset into structural families based on Murcko scaffolds and compute the multivariate distance correlation between the HOMO residual and the remaining 10-dimensional property space. This approach allows us to systematically identify localized coupling neighborhoods where the global orthogonality vanishes.

Our analysis reveals that while the HOMO energy is globally decoupled, this isolation is highly context-dependent. We identify specific architectural triggers that force the electronic state to synchronize with the molecular framework. Notably, molecular rigidity—indicated by fewer rotatable bonds—along with higher fractions of sp^3 carbons and increased ring counts, drive the molecule into a coupled regime. Furthermore, we isolate a high-strain outlier subset comprising one percent of the population that exhibits extreme coupling, characterized by significant enrichment in fluorine substitutions and rotatable bonds.

Granular analysis within these highly coupled scaffolds demonstrates that the HOMO energy synchronizes most strongly with isotropic polarizability and dipole moment. Block-wise partial distance correlations further reveal that this unique coupling is primarily driven by the molecule's thermodynamic and geometric-vibrational states, rather than its rotational geometry. Ultimately, these findings provide a robust, computationally efficient framework for understanding exactly where and why electronic isolation breaks down in molecular property spaces, naturally segregating latent geometric artifacts into identifiable outlier neighborhoods.

2 Methods

2.1 Data preparation and stoichiometric baseline

To isolate the variance in molecular properties driven purely by architectural differences, we first established a stoichiometric baseline to strip out extensive size effects. The QM9 dataset was loaded and deduplicated based on canonical SMILES strings, retaining the first occurrence of each molecule. Nine perfectly linear molecules, identified by a rotational constant $A = 0$, were explicitly flagged to prevent the model from attributing linear geometric artifacts to the residuals of other properties. To compress the extreme heavy tail of the rotational constant A, a logarithmic transformation was applied:

$$A_{\text{trans}} = \log(A + 1) \tag{1}$$

The dataset was then subset to 11 non-redundant physical properties: the transformed rotational constant A_{trans}, rotational constants B and C, dipole moment (μ), isotropic polarizability (α), highest occupied molecular orbital energy (*homo*), 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 lowest unoccupied molecular orbital energy (*lumo*) and the HOMO-LUMO gap (*gap*) were explicitly excluded from the analysis to prevent mathematical circularity when evaluating *homo* isolation. Multicollinearity among the remaining 10 properties (excluding *homo*) was assessed to ensure stability for downstream multivariate distance correlation calculations.

For each of the 11 properties, a non-linear gradient boosting regressor (LightGBM) was trained using only exact atom counts (C, N, O, F, H) and the binary linearity indicator as features. The stoichiometric residual for each property y was extracted as:

$$res_y = y - \hat{y}_{\text{LightGBM}}(\text{C, N, O, F, H, linearity}) \tag{2}$$

Molecules exhibiting extreme reconstruction errors—defined as the top 1% of mean squared error across all 11 properties—were flagged as a "high-strain" subset and retained for targeted outlier analysis.

2.2 Structural feature engineering and scaffold generation

To link the extracted property residuals to molecular architecture, structural descriptors were generated from the deduplicated SMILES strings. Topological and compositional features were computed, including the number of aromatic rings, total ring count, fraction of sp^3 hybridized carbons, number of rotatable bonds, counts of 3- and 4-membered rings (as indicators of local ring strain), and specific heteroatom counts. Additionally, molecules were grouped into structural families by generating their corresponding Murcko scaffolds, which represent the core ring and linker framework of the molecule stripped of terminal side chains.

2.3 Scaffold-level coupling and distance correlation

To identify localized neighborhoods where the global electronic decoupling of the HOMO energy breaks down, the dataset was grouped by Murcko scaffold. Linear molecules were excluded from this step to prevent outlier skew. Only scaffolds containing a statistically sufficient number of molecules ($N \geq 30$) were retained for analysis.

For each valid scaffold, the multivariate distance correlation ($dCor$) was computed between the HOMO residual (res_homo) and the 10-dimensional residual space comprising the remaining physical properties. This computation was parallelized across 14 CPUs. To ensure robustness against small-sample noise, bootstrapping was utilized to estimate confidence intervals for the $dCor$ value of each scaffold, yielding a stable "coupling score".

Representative structural descriptors (e.g., average fraction of sp^3 carbons, average number of rotatable bonds) were computed for each valid scaffold. Pearson and Spearman correlation coefficients were then calculated between the scaffold-level coupling scores and these structural descriptors. To ensure statistical significance and confirm that observed correlations were not due to chance, permutation testing was performed by shuffling the coupling scores 1,000 times to establish a null distribution and compute rigorous p-values.

2.4 High-strain outlier analysis

The "high-strain" subset identified during the residual extraction phase was isolated to investigate the behavior of the HOMO energy under extreme geometric or compositional variance. The global distance correlation between res_homo and the 10-dimensional residual space was computed specifically for this subset. A permutation test, which involved shuffling the "high-strain" labels across the entire dataset, was performed to ensure that any observed difference in $dCor$ compared to the normal population was statistically significant and not an artifact of the smaller sample size. Finally, Mann-Whitney U tests were conducted to determine if this outlier subset was significantly enriched in specific structural motifs, such as fluorine substitutions or rotatable bonds, compared to the rest of the dataset.

2.5 Granular coupling and block-wise partial distance correlation

To determine exactly which physical properties synchronize with the HOMO energy in highly coupled regimes, a granular analysis was performed on the top decile of scaffolds ranked by coupling score. Within this subset, pairwise distance correlations were calculated between res_homo and each individual property residual.

Furthermore, to isolate the unique contributions of different physical domains, the 10-dimensional residual space was partitioned into three distinct blocks:

- **Thermodynamic:** res_u_0, $res_u_0_atom$
- **Geometric-Vibrational:** res_r^2, res_zpve, res_c_v
- **Rotational:** res_A_{trans}, res_B, res_C

The partial distance correlation of res_homo with each block was computed while controlling for the variance explained by the other two blocks. This allowed us to distinguish primary coupling drivers from secondary, mediated effects.

2.6 Validation via split-half stability testing

To validate the robustness of the identified architectural triggers, a split-half stability test was conducted. The set of valid Murcko scaffolds ($N \geq 30$) was randomly partitioned into two independent halves. The Pearson and Spearman correlations between the scaffold-level coupling scores and the structural descriptors were re-evaluated separately on both halves. The stability of these correlation coefficients across the independent splits was reported to confirm that the identified structural motifs consistently predict HOMO coupling across diverse molecular families.

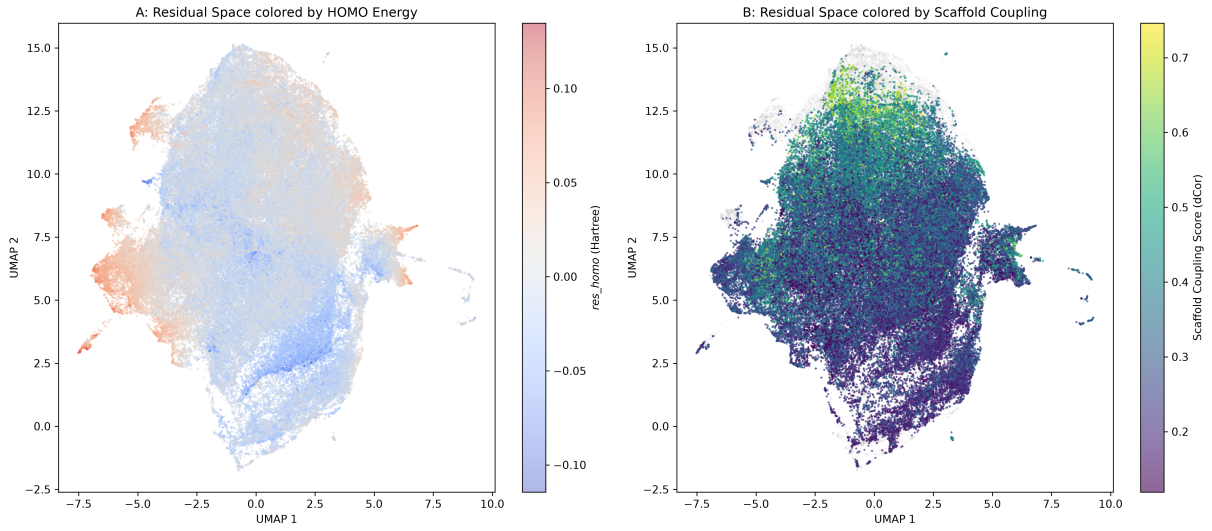

Figure 1: UMAP projection of the physical residual space isolating variance independent of molecular composition. Panel A displays the embedding colored by the HOMO energy residual (`res_homo`), illustrating its global distribution and general orthogonality to other physical properties. Panel B shows the same space colored by the Murcko scaffold coupling score ($dCor$), highlighting localized structural neighborhoods that exhibit high electronic-geometric synchronization. These visualizations demonstrate that while the HOMO energy is globally decoupled from thermodynamic and geometric states, this isolation breaks down within specific architectural regimes, such as rigid or high-strain molecular frameworks, where the electronic state is forced to strongly synchronize with the physical geometry.

3 Results

3.1 Isolation of residual variance via gradient boosting

To characterize the joint structure of the QM9 dataset independently of trivial molecular size effects, a stoichiometric baseline was established using non-linear gradient-boosted regressors (LightGBM). These models predicted 11 non-redundant properties using only exact atom counts (C, N, O, F, H) and a linearity indicator. The resulting residuals successfully isolated the variance not attributable to molecular composition.

The baseline models revealed a stark contrast in property dependencies. While thermodynamic properties such as internal energy at 0 K (u_0) and zero-point vibrational energy ($zpve$) are almost entirely determined by atom counts ($R^2 > 0.99$), the HOMO energy ($homo$) remains largely independent ($R^2 = 0.0855$). This confirms that fluctuations in the HOMO energy are driven by architectural rather than stoichiometric factors. A global analysis of these residuals confirms the extreme orthogonality of $homo$; across the entire 133,802-molecule population, the distance correlation ($dCor$) between the $homo$ residual (`res_homo`) and the 10-dimensional space of other physical residuals is remarkably low, typically remaining below 0.3 in global samples. However, projecting this residual space into a lower-dimensional embedding reveals that this global decoupling is not uniform, but rather contains localized structural neighborhoods where the electronic state and physical geometry synchronize (Figure 1).

3.2 Structural triggers of HOMO coupling

By grouping molecules into Murcko scaffolds and computing a multivariate coupling score ($dCor$) for each family, we identified the specific architectural conditions under which the global electronic decoupling of the HOMO energy breaks down. Statistical analysis of 551 valid scaffolds ($N \geq 30$) reveals that the degree of $homo$ coupling is strongly correlated with specific structural features, as visualized in Figure 2.

The primary architectural determinants of coupling include:

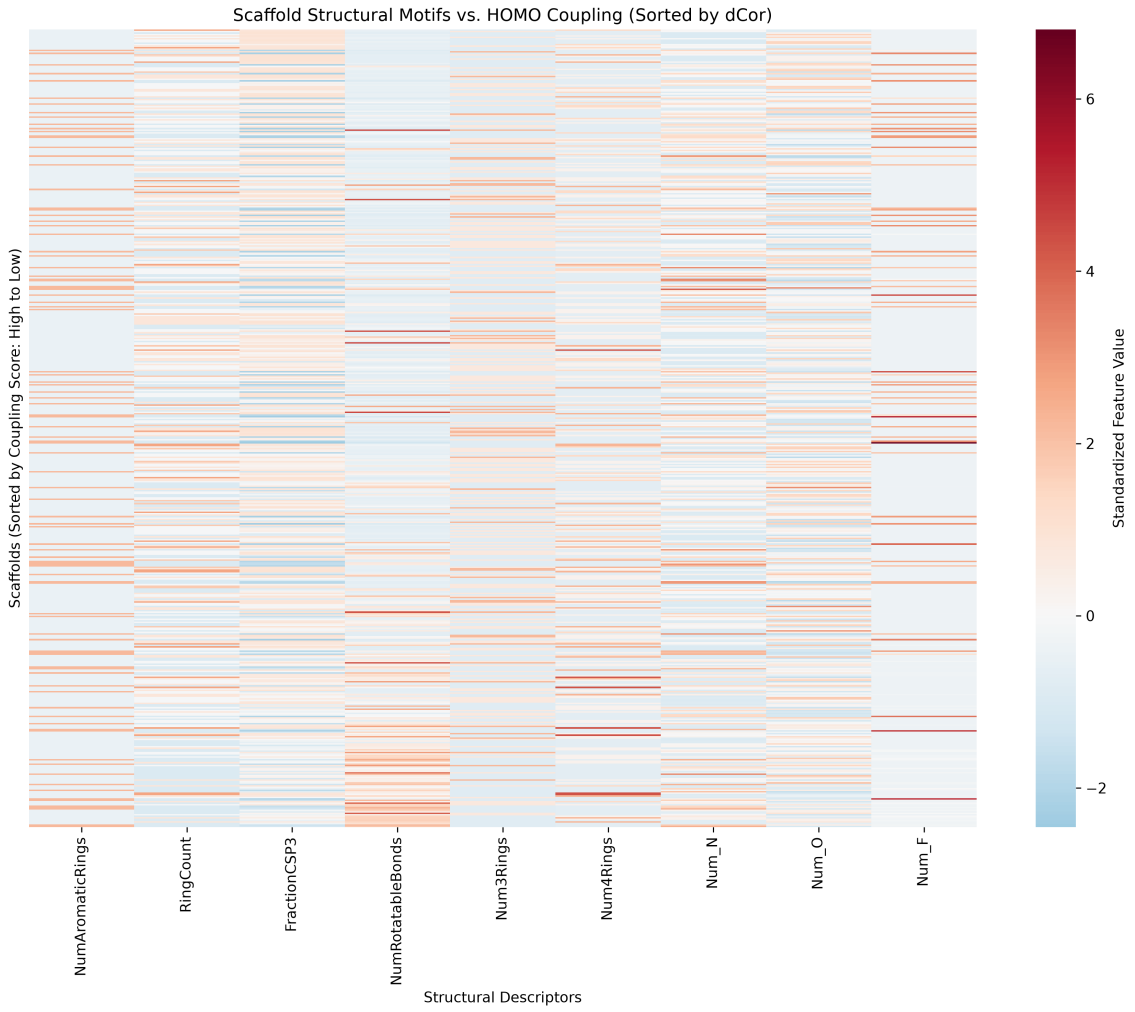

Figure 2: Heatmap of standardized structural descriptors for molecular scaffolds, sorted vertically by their HOMO coupling score ($dCor$) from highest (top) to lowest (bottom). The visualization confirms the architectural determinants of electronic coupling by demonstrating a strong inverse relationship between molecular flexibility and coupling; specifically, the number of rotatable bonds (`NumRotatableBonds`) visibly increases in the lower, weakly coupled regions. Conversely, features associated with rigidity and saturation, such as the fraction of sp^3 carbons (`FractionCSP3`) and total ring count (`RingCount`), show higher standardized values among the highly coupled scaffolds at the top. The distribution of fluorine atoms (`Num_F`) further illustrates its role in reducing the synchronization between the HOMO energy and the broader physical geometry of the molecule.

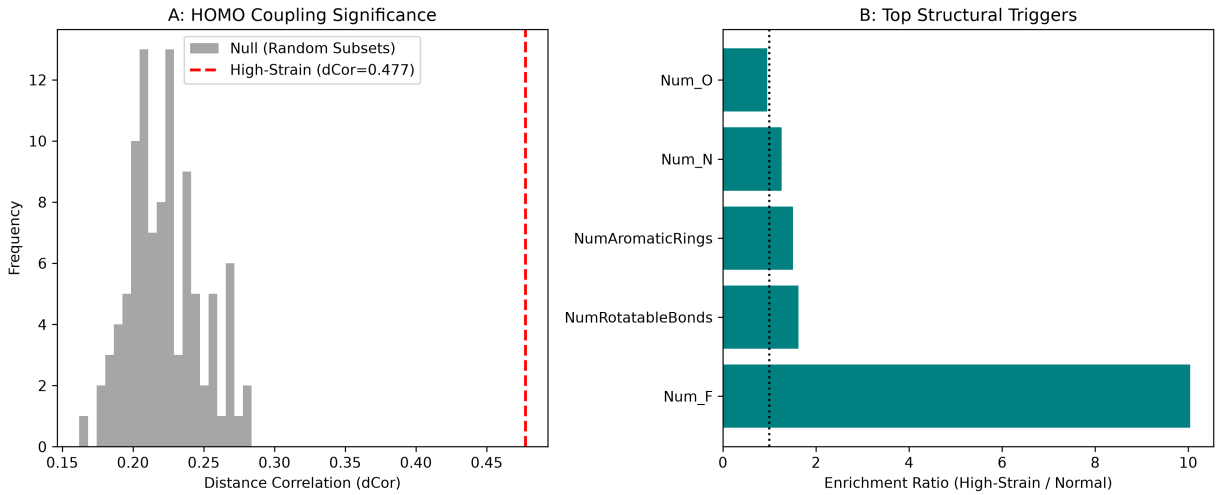

Figure 3: High-strain outlier analysis. Panel A displays the distance correlation ($dCor$) between the HOMO energy residual and the 10-dimensional physical property space, highlighting that the high-strain outlier subset exhibits a significantly elevated coupling ($dCor = 0.477$) compared to the null distribution of random subsets. Panel B illustrates the structural triggers driving this synchronization by plotting the enrichment ratio of molecular features in the high-strain subset relative to the normal population. The extreme enrichment of fluorine atoms (`Num_F`) and the increased presence of rotatable bonds (`NumRotatableBonds`) demonstrate that specific chemical substitutions and geometric strain break the typical electronic isolation of the HOMO energy, forcing a strong joint structure between the electronic and thermodynamic states.

- **Molecular rigidity:** The number of rotatable bonds shows the strongest inverse relationship with coupling (Spearman $r = -0.4226$, permutation $p < 0.001$). Rigid, highly fused scaffolds tend to exhibit significantly higher coupling scores, forcing the electronic state to respond more directly to the molecular framework.
- **Saturation and topology:** The fraction of sp^3 carbons is positively associated with coupling (Spearman $r = 0.1981$, $p < 0.001$). Similarly, total ring count correlates with increased coupling ($r = 0.1284$, $p = 0.004$).
- **Heteroatom influence:** The presence of fluorine atoms acts as a significant rewriter of joint structure (Spearman $r = -0.1369$ between fluorine count and $dCor$, $p = 0.002$), generally reducing the coupling between the HOMO and other physical properties.

To ensure these relationships are not artifacts of the particular sample, a split-half stability test was performed. The Spearman correlations for the primary triggers remained stable across independent scaffold subsets. For the number of rotatable bonds, the correlations were $r_A = -0.4629$ and $r_B = -0.3910$ (a difference of 0.0719). For the fraction of sp^3 carbons, the correlations were $r_A = 0.1855$ and $r_B = 0.2309$ (a difference of 0.0454). The highest coupling score was observed for the scaffold `C1C2CN3CC2C13` ($dCor = 0.7461$), a compact, rigid polycyclic structure that perfectly exemplifies the highly coupled regime.

3.3 High-strain outlier neighborhoods

The residual extraction phase identified a "high-strain" subset comprising 1,339 molecules (1.0% of the population) that exhibited the highest reconstruction errors across the stoichiometric baseline. In this subset, the electronic isolation of *homo* vanishes entirely (Figure 3).

The multivariate distance correlation between `res_homo` and the 10-dimensional property space in this high-strain subset is 0.4774, compared to a null mean of 0.2228 established via permutation testing (Z-score = 10.02, $p = 0.0099$). This confirms that extreme geometric strain or unusual bonding motifs force a strong

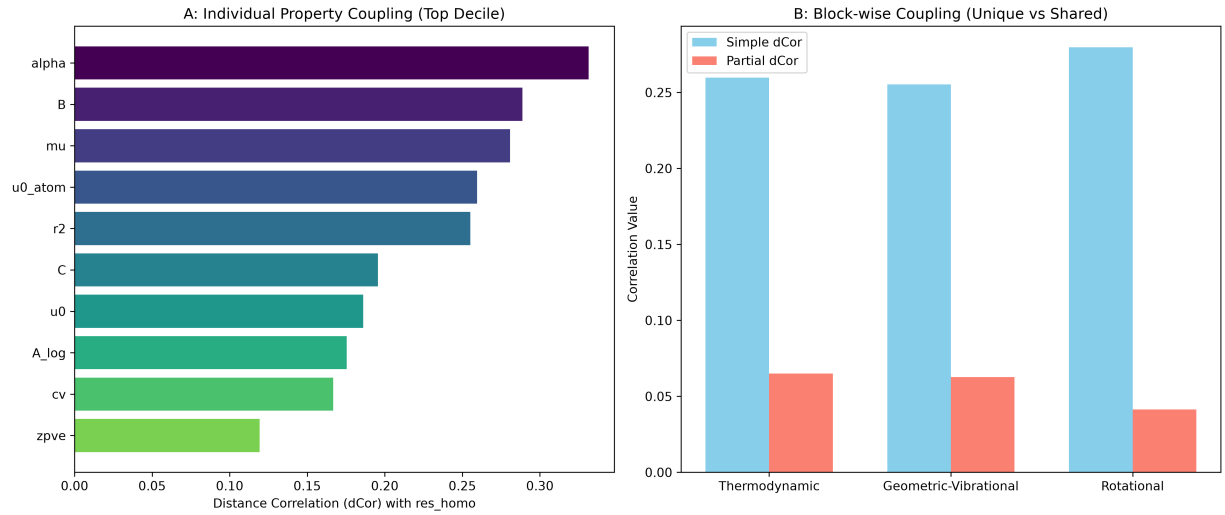

Figure 4: Granular and block-wise distance correlation ($dCor$) analysis of the HOMO energy residual (`res_homo`) within the top decile of highly coupled molecular scaffolds. Panel A displays pairwise couplings, revealing that `res_homo` synchronizes most strongly with isotropic polarizability (α), the rotational constant B, and the dipole moment (μ), while exhibiting the weakest coupling with zero-point vibrational energy ($zpve$). Panel B compares simple and partial $dCor$ across Thermodynamic, Geometric-Vibrational, and Rotational property blocks to isolate unique variance contributions. Although the Rotational block exhibits high simple coupling, its unique signal (partial $dCor$) is the lowest. This hierarchy demonstrates that when the HOMO energy breaks its electronic isolation, it does so primarily by synchronizing with the molecule's thermodynamic and vibrational states, whereas its relationship to geometric rotation is a secondary, mediated effect.

joint structure between electronic and thermodynamic states. Mann-Whitney U tests show that this high-strain subset is significantly enriched in fluorine (mean count = 0.226 vs 0.023, $p < 10^{-194}$) and rotatable bonds (mean count = 1.488 vs 0.917, $p < 10^{-38}$). This suggests that specific chemical substitutions drive molecules into these coupled, high-variance regimes, overriding the general decoupling trends observed in the broader dataset.

3.4 Granular and block-wise coupling in extreme scaffolds

To determine exactly which physical properties the HOMO energy synchronizes with when its isolation breaks down, we performed a granular analysis within the top decile of highly coupled scaffolds (Figure 4).

Pairwise distance correlations reveal that in these regimes, `res_homo` is most strongly tethered to isotropic polarizability (α, $dCor = 0.3314$), followed by the rotational constant B ($dCor = 0.2886$) and the dipole moment (μ, $dCor = 0.2808$). The weakest coupling remains with zero-point vibrational energy ($zpve$, $dCor = 0.1192$).

To isolate unique contributions and distinguish primary drivers from mediated effects, we grouped the 10 properties into three blocks: Thermodynamic (u_0, u_0_atom), Geometric-Vibrational (r^2, $zpve$, c_v), and Rotational (A, B, C). The block-wise partial distance correlations yielded the following hierarchy:

- **Thermodynamic block:** Exhibits the highest simple coupling ($dCor = 0.2598$) and retains the highest unique signal after controlling for the other blocks (Partial $dCor = 0.0650$).
- **Geometric-Vibrational block:** Shows a simple $dCor$ of 0.2553 and a partial $dCor$ of 0.0627.
- **Rotational block:** While exhibiting high simple coupling ($dCor = 0.2797$), its unique signal is the lowest (Partial $dCor = 0.0414$).

This indicates that the relationship between the HOMO energy and rotational geometry is largely mediated by the molecule's broader thermodynamic and spatial organization. When the HOMO energy breaks its isolation, it does so primarily by synchronizing with the molecule's thermodynamic and vibrational states, with rotational coupling emerging as a secondary, shared effect.

4 Conclusions

In this study, we investigated the boundaries of electronic isolation within the QM9 dataset by mapping the joint distribution of molecular properties in a stoichiometric residual space. By employing non-linear gradient-boosted regressors to strip out extensive size effects, we isolated the variance driven purely by molecular architecture. This approach enabled a localized sensitivity analysis, revealing that the extreme global orthogonality of the highest occupied molecular orbital (HOMO) energy is not a universal baseline, but rather a context-dependent phenomenon that breaks down under specific structural conditions.

Our scaffold-level analysis identified distinct architectural triggers that force the electronic state to synchronize with the broader physical geometry of the molecule. We demonstrated that molecular rigidity—characterized by a reduction in rotatable bonds—along with increased ring counts and higher fractions of sp^3 hybridized carbons, act as primary drivers of this coupling. The robustness of these structural determinants was confirmed through split-half stability testing, underscoring that rigid, highly fused frameworks inherently restrict the independent fluctuation of electronic states.

Furthermore, we isolated a high-strain outlier subset, comprising one percent of the dataset, where the electronic isolation of the HOMO energy vanishes entirely. This highly coupled regime is significantly enriched in fluorine substitutions and rotatable bonds, suggesting that extreme geometric strain and specific heteroatom motifs can override general decoupling trends, forcing a strong joint structure between electronic and thermodynamic states.

Granular analysis within the most highly coupled structural families provided a precise hierarchy of this synchronization. When the HOMO energy breaks its isolation, it tethers most strongly to isotropic polarizability and the dipole moment. Crucially, block-wise partial distance correlations revealed that this coupling is primarily driven by the molecule's thermodynamic and geometric-vibrational states. The apparent synchronization with rotational geometry was shown to be a secondary effect, largely mediated by the molecule's broader spatial and thermodynamic organization.

Ultimately, these findings provide a computationally efficient and statistically robust framework for understanding exactly where and why electronic isolation breaks down in chemical space. By pinpointing the structural triggers of HOMO coupling, this work offers valuable insights for the targeted design of organic molecules, particularly in applications where electronic properties must be precisely tuned in tandem with thermodynamic and geometric constraints.