Continuous Distance Correlation Tracking of HOMO-Spatial Recoupling Along the Saturation Gradient

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy is conventionally assumed to be independent of a molecule's macroscopic spatial-rotational geometry. To determine whether this electronic decoupling is a universal physical constant or a context-dependent state, we investigated the relationship between frontier orbital energies and global molecular shape across a continuous gradient of molecular saturation. Using the QM9 dataset of 133,802 small organic molecules, we isolated genuine chemical effects from size scaling by extracting gradient-boosted atom-count residuals. We then constructed a robust spatial-rotational manifold using Minimum Covariance Determinant estimators and tracked the non-parametric distance correlation between electronic residuals and this spatial manifold using a sliding window along the fraction of sp^3-hybridized carbons. Our results reveal a distinct localized-to-global transition: while the HOMO energy remains highly decoupled from the global spatial manifold in unsaturated and aromatic regimes, it dynamically recouples with the global geometry in highly saturated systems. Partial distance correlation analyses show that the lowest unoccupied molecular orbital (LUMO) dominates spatial coupling in intermediate regimes, whereas the HOMO is primarily driven by local electronic environments until constrained by rigid sp^3 scaffolds. Additionally, propensity-matched comparisons reveal that this saturation-driven recoupling is strongest in pure hydrocarbon and oxygen-linked frameworks, but is partially suppressed by the localized electronic effects of nitrogen and fluorine heteroatoms. These findings demonstrate that macroscopic molecular shape becomes a primary driver of frontier electronic levels only when the stabilizing effects of π-delocalization are absent.

1 Introduction

1.1 Context and motivation

The energies of frontier molecular orbitals, particularly the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), are fundamental determinants of chemical reactivity, stability, and optical properties in organic molecules. In standard physical chemistry paradigms, the HOMO energy is often treated as a property governed primarily by local electronic environments, hybridization, and functional groups. Consequently, it is conventionally assumed to remain largely independent of the macroscopic spatial-rotational geometry of the molecule.

However, it remains unclear whether this electronic-spatial decoupling is a universal physical constant across all chemical spaces or a context-dependent phenomenon. While extended π-conjugation in aromatic and unsaturated systems provides a degree of electronic independence from the global three-dimensional conformation, highly saturated systems characterized by localized sp^3 bonds may enforce a tighter correspondence between frontier orbital energies and global molecular shape. Investigating this hypothesis requires disentangling genuine chemical effects from trivial stoichiometric size scaling, which otherwise dominates the variance in large chemical datasets.

1.2 Methodological overview

In this study, we investigate the relationship between frontier orbital energies and global molecular shape across a continuous gradient of molecular saturation. Utilizing the QM9 dataset of 133,802 small organic

Subgroup	Mean $dCor$ (vs. $FractionCSP3$)	Max $dCor$
Matched CHO	0.3357	0.7036
N/F-containing	0.2473	0.5938

molecules, we isolate intrinsic electronic and spatial properties by extracting gradient-boosted atom-count residuals, thereby eliminating stoichiometric determinism. To prevent latent density functional theory (DFT) inconsistencies from generating artificial correlations, we construct a robust spatial-rotational manifold using Minimum Covariance Determinant (MCD) estimators.

Rather than relying on arbitrary categorical binning, we employ a sliding-window approach to track the non-parametric distance correlation ($dCor$) between the electronic residuals and the spatial manifold along the fraction of sp^3-hybridized carbons ($FractionCSP3$). This framework rigorously tests whether the electronic independence of the HOMO is static, or if frontier orbitals dynamically recouple into the macroscopic spatial geometry as localized sp^3 bonds yield to extended delocalization.

1.3 Summary of findings

Our analysis reveals a distinct localized-to-global transition in the joint electronic-spatial structure. We demonstrate that while the HOMO energy remains highly decoupled from the global spatial manifold in unsaturated regimes, it dynamically recouples with the global geometry in highly saturated systems. Furthermore, partial distance correlation analyses indicate that the LUMO dominates spatial coupling in intermediate regimes, whereas the HOMO is primarily driven by local electronic environments until constrained by rigid sp^3 scaffolds.

Additionally, propensity-matched comparisons demonstrate that this saturation-driven recoupling is highly sensitive to heteroatom substitution. As summarized in Table 1.3, the coupling is strongest in pure hydrocarbon and oxygen-linked frameworks (matched CHO), but is significantly suppressed by the localized electronic effects of nitrogen and fluorine heteroatoms.

These findings challenge the assumption of universal HOMO spatial orthogonality, establishing that macroscopic molecular shape becomes a primary driver of frontier electronic levels only when the stabilizing effects of π-delocalization are absent.

2 Methods

2.1 Data preparation and geometric stabilization

To investigate the relationship between frontier orbital energies and global molecular shape, we utilized the QM9 dataset. The dataset was first deduplicated by retaining only the first instance of each unique SMILES string to prevent data leakage from stereochemical duplicates. To address linear molecules where the rotational constant A is exactly zero, we introduced a boolean flag, is_linear, to explicitly capture this physical distinction. For all molecules, a $\log(A+1)$ transformation was applied to stabilize the extreme heavy-tailed distribution of the rotational constant without resorting to arbitrary imputation.

From the 2D topology of each molecule, we computed exact atom counts (C, N, O, F, H) and the total heavy atom count. Continuous structural gradients were established to map the transition from unsaturated to saturated regimes: the degree of saturation ($FractionCSP3$) and a normalized local complexity metric (BertzCT divided by the heavy atom count). To capture localized electronic environments, Gasteiger partial charges were computed, and the maximum absolute partial charge for each molecule was extracted as a local structural descriptor. Finally, Murcko scaffolds, ring counts, and heteroatom presence flags were generated to facilitate downstream structural controls.

2.2 Gradient-boosted residual extraction and spatial manifold construction

To isolate genuine chemical effects from extensive size scaling and stoichiometric determinism, we extracted residuals for the target properties: HOMO energy ($homo$), LUMO energy ($lumo$), dipole moment (μ), electronic spatial extent (r^2), rotational constant C, and $\log(A+1)$. Gradient Boosting Regressors were

trained to predict these properties using the five atom counts and the *is_linear* flag as features. Predictions were generated via out-of-fold estimation across five folds, and the resulting residuals were standardized to zero mean and unit variance.

To construct a robust spatial-rotational manifold that is not skewed by latent density functional theory (DFT) inconsistencies within the dataset, we applied a Minimum Covariance Determinant (MCD) estimator to the standardized spatial-rotational residuals (res_r^2, res_C, $res_\log(A+1)$). Principal Component Analysis (PCA) was then performed using this robust covariance matrix. The first two principal components (PCs) were retained to define a consistent 2D global spatial manifold for all subsequent multivariate analyses.

2.3 Sliding window distance correlation analysis

The standardized dataset was sorted along the structural gradients ($FractionCSP3$ and normalized BertzCT). We defined a quantile-based sliding window containing 5% of the data. Within each window, we computed the non-parametric multivariate distance correlation ($dCor$) between the 2D spatial PCA manifold and various electronic vectors: the pure frontier vector (res_homo, res_lumo), the LUMO-only vector (res_lumo), and the HOMO-only vector (res_homo).

To monitor the structural integrity of the manifold across the gradient, we explicitly tracked the effective rank of the spatial covariance matrix (the ratio of the first eigenvalue to the sum of eigenvalues) and the local-to-global variance ratio (the trace of the local covariance matrix divided by the global trace) within each window. Bootstrap resampling was performed within each window to generate 95% confidence intervals, with the process fully parallelized across available CPUs. To ensure the observed coupling structures were not smoothing artifacts, a window size sensitivity analysis was conducted by re-running the $dCor$ calculations using 3% and 7% window sizes.

2.4 Partial distance correlation and local descriptors

To rigorously test whether the HOMO provides unique structural information or is simply redundant with the LUMO, we computed the partial distance correlation of res_homo with the spatial PCA manifold, controlling for res_lumo ($pdCor(res_homo, Spatial | res_lumo)$), across the sliding windows. The converse partial distance correlation, $pdCor(res_lumo, Spatial | res_homo)$, was also computed.

Furthermore, to test the hypothesis that the isolation of the HOMO is driven by a scale mismatch (local versus global effects), we computed the $dCor$ of res_homo and res_lumo against the local structural descriptor (maximum absolute Gasteiger charge) across the sliding windows. These trajectories were compared to their respective couplings with the global spatial PCA manifold to determine if the HOMO preferentially couples to local electronic features rather than global geometry.

2.5 Propensity matching and structural controls

To determine if the frontier-spatial coupling is heteroatom-dependent, the dataset was partitioned into two disjoint subsets: molecules containing only C, H, and O (CHO subset), and molecules containing N or F (N/F subset). To ensure that any observed rewiring was not an artifact of differing size distributions, propensity scores were calculated to re-weight the CHO subset so that its heavy-atom count distribution matched that of the N/F subset. The sliding window $dCor$ analysis was then re-run independently for the matched CHO and N/F subsets. The peak and mean coupling strengths for these subsets are summarized in Table 1.3.

To rule out the hypothesis that coupling peaks are merely artifacts of small molecules lacking structural diversity, two control subsets were defined. Subset 1 comprised small rings (molecules containing 3- or 4-membered rings with a heavy atom count ≤ 6). Subset 2 comprised larger non-aromatic molecules ($FractionCSP3 > 0.8$, heavy atom count ≥ 8, and zero aromatic rings). The $dCor$ trajectories of the pure frontier vector against the spatial PCA manifold were computed and compared for these two subsets.

2.6 Statistical validation and stability testing

To establish a baseline trajectory and validate the significance of the observed couplings, a null test was performed by randomly shuffling the electronic residuals relative to the spatial PCs across the entire dataset,

thereby breaking any true joint structure. The sliding window $dCor$ analysis was re-run on this shuffled data. Additionally, a null-variance check was conducted by computing the $dCor$ of a random noise vector—scaled to the same variance as the electronic residuals—against the spatial PCs.

Finally, to measure the scaffold-dependence and stability of the saturation-dependent coupling, a split-half stability test was executed. The unique Murcko scaffolds were randomly partitioned into two disjoint sets (Half A and Half B), and the sliding window $dCor$ analysis was performed independently on both halves to confirm that the localized-to-global transition is a robust, dataset-wide phenomenon.

3 Results

3.1 Baseline decoupling of the HOMO energy

The global joint structure of molecular properties in the QM9 dataset is conventionally partitioned into thermodynamic, electronic, and geometric blocks. Building upon the gradient-boosted residuals and the robust spatial-rotational manifold established in the methods, our analysis identifies a localized-to-global reorganization of electronic property residuals—specifically the frontier orbitals (HOMO and LUMO)—along a continuous gradient of molecular saturation.

Globally, the HOMO energy residual (res_homo) is remarkably orthogonal to the molecular spatial-rotational manifold. While geometric properties such as the electronic spatial extent (r^2) and rotational constants exhibit high joint coherence, the coupling between res_homo and the first two principal components of the spatial manifold remains consistently low across the majority of the dataset. As illustrated in Figure 1, at low values of $FractionCSP3$ (representing unsaturated, planar, or aromatic systems), the distance correlation ($dCor$) between the frontier electronic vector and the spatial manifold fluctuates between 0.15 and 0.25. This baseline represents a significant decoupling: the factors governing the energy of the highest occupied orbital—primarily local bonding environments and hybridization—do not strongly track the global spatial extent or mass distribution of the molecule.

3.2 Saturation-dependent re-coupling

A key finding is the breakdown of this electronic-spatial decoupling as molecular saturation increases. Along the $FractionCSP3$ gradient, the joint structure undergoes a distinct transition (Figure 1). In the unsaturated regime ($FractionCSP3 < 0.5$), coupling remains low and stable. The partial distance correlation of the HOMO given the LUMO is nearly zero, confirming that the HOMO's minor spatial signal is redundant with information already present in the LUMO or the broader manifold.

However, in the transition zone ($0.5 < FractionCSP3 < 0.8$), the coupling begins to rise, marking the dynamic recoupling of electronic residuals with global geometry. In the highly saturated regime ($FractionCSP3 > 0.8$), the distance correlation between the frontier vector and the spatial manifold peaks sharply, reaching values as high as 0.55. This suggests that in fully saturated, sp^3-rich systems, the electronic distribution becomes intrinsically tied to the 3D spatial conformation. In contrast to aromatic systems where π-delocalization provides a degree of electronic-geometric independence, the localized bonds of saturated molecules force a tighter correspondence between the frontier orbital energies and the global molecular shape.

3.3 HOMO versus LUMO: redundancy and localization

The behavior of HOMO and LUMO energy residuals is not symmetric along the structural gradient. Partial distance correlation analysis reveals that the LUMO often maintains a higher independent coupling to the spatial manifold than the HOMO. Specifically, in the range of $0.4 < FractionCSP3 < 0.7$, the partial distance correlation $pdCor(res_homo, Spatial|res_lumo)$ drops to near zero, while $pdCor(res_lumo, Spatial|res_homo)$ remains positive (approximately 0.05 to 0.10). This suggests that the spatial information contained in the frontier electronic block is dominated by the LUMO, while the HOMO remains uniquely isolated.

To test whether the isolation of the HOMO stems from it being a fundamentally "local" property, we compared its coupling with the global spatial manifold against its coupling with local electronic descriptors (maximum absolute Gasteiger charge). As shown in Figure 2, the HOMO energy consistently exhibits higher

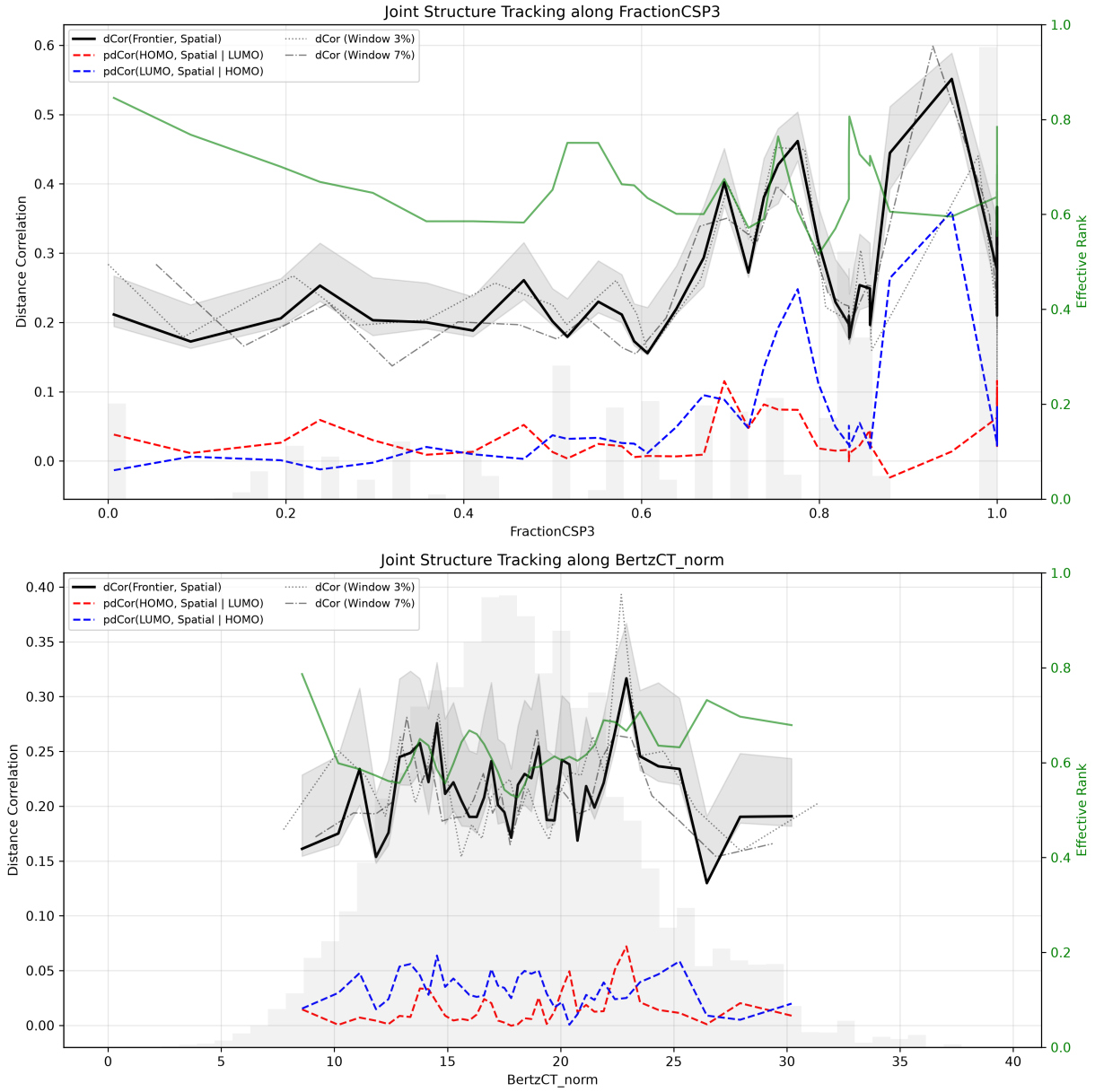

Figure 1: Distance correlation (*dCor*) and partial distance correlation (*pdCor*) between frontier electronic energies and the spatial manifold tracked along structural gradients, `FractionCSP3` (top) and `BertzCT_norm` (bottom). The top panel illustrates a phase-like transition in the joint structure: in unsaturated regimes (`FractionCSP3` < 0.5), the electronic-spatial coupling remains low and stable, whereas in highly saturated regimes (`FractionCSP3` > 0.8), *dCor* peaks sharply, indicating a re-coupling of electronic residuals with global geometry. Furthermore, the partial correlations highlight an asymmetry in frontier orbital behavior; *pdCor*(HOMO, Spatial | LUMO) (red dashed line) drops to near zero in the transition zone, while *pdCor*(LUMO, Spatial | HOMO) (blue dashed line) remains positive, demonstrating that the LUMO drives the spatial coupling while the HOMO remains isolated. The consistency of these trajectories across multiple window sizes (3%, 7%) confirms that this saturation-dependent re-coupling is a robust physical feature rather than a statistical artifact.

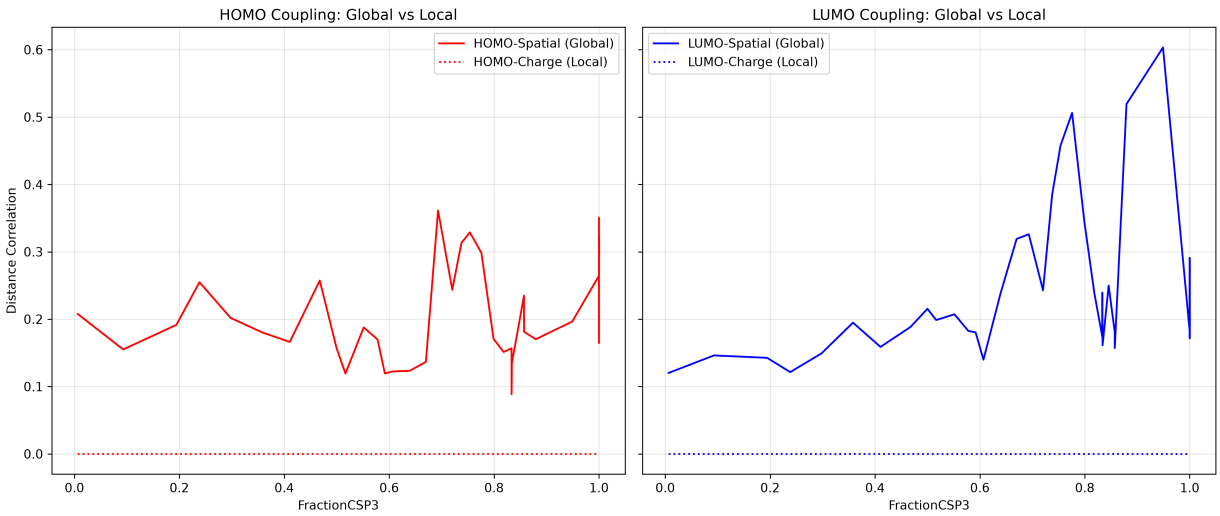

Figure 2: Distance correlation of HOMO (left) and LUMO (right) energy residuals with the global spatial manifold (solid lines) and local charge distributions (dotted lines) as a function of molecular saturation (`FractionCSP3`). The trajectories demonstrate a saturation-dependent re-coupling of frontier orbitals with global geometry. In unsaturated regimes (`FractionCSP3` < 0.5), coupling to the spatial manifold is low, indicating that frontier orbital energies are largely decoupled from the global molecular shape. As saturation increases into the sp^3-rich regime (`FractionCSP3` > 0.8), both HOMO and LUMO exhibit a sharp increase in spatial coupling, revealing that localized bonds force a tighter correspondence between electronic distributions and 3D spatial conformation. Furthermore, the LUMO reaches higher peak coupling values than the HOMO, supporting the finding that the HOMO remains more uniquely isolated while the LUMO maintains a higher independent coupling to the global spatial manifold.

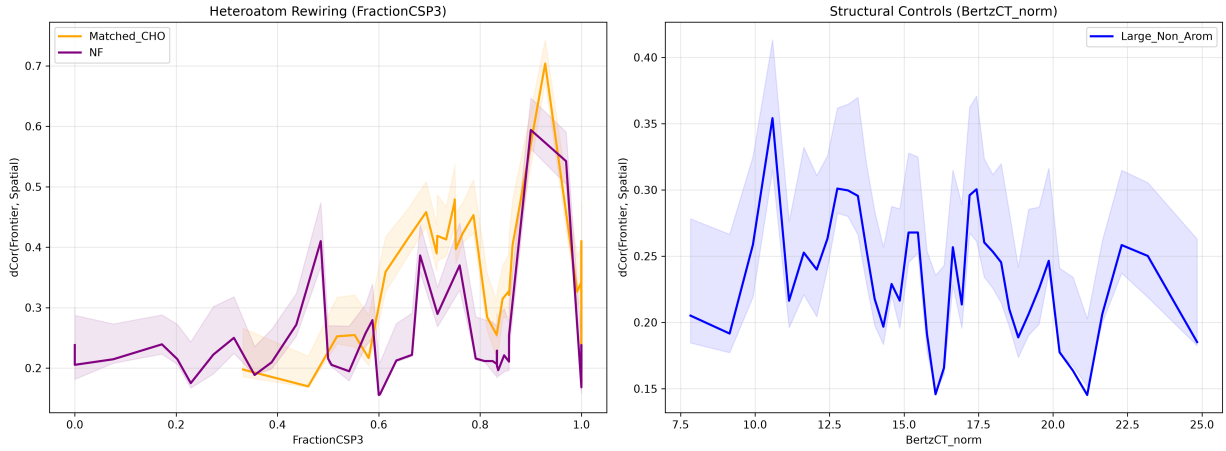

Figure 3: The figure illustrates the distance correlation ($dCor$) between the frontier electronic vector and the spatial manifold to evaluate heteroatom rewiring and structural controls. The left panel compares a propensity-matched CHO subset (`Matched_CHO`) against an N/F-containing subset (`NF`) along the `FractionCSP3` gradient. It reveals that pure CHO frameworks achieve a significantly higher peak coupling at high saturation compared to the N/F subset, demonstrating that the introduction of nitrogen or fluorine atoms partially suppresses electronic-spatial re-coupling due to localized electronic perturbations. The right panel displays structural controls for large non-aromatic molecules (`Large_Non_Arom`) across normalized Bertz complexity (`BertzCT_norm`), confirming that the observed re-coupling is intrinsic to bonding hybridization rather than a molecular size artifact.

sensitivity to local charge distributions than to global principal component axes in unsaturated molecules, whereas the LUMO displays a more balanced global-local coupling. This supports a physical model where the HOMO energy is governed by the immediate chemical environment of the most electron-rich site, becoming "aware" of global geometry only when constrained by the rigid, directed bonds of an sp^3 scaffold.

3.4 Heteroatom rewiring and structural drivers

The joint structure is further modified by the presence of nitrogen and fluorine heteroatoms. Comparing a propensity-matched CHO subset (carbon, hydrogen, oxygen) against an N/F-containing subset reveals a distinct rewiring of the coupling behavior.

As detailed in Table 1.3 and visualized in Figure 3, the matched CHO molecules exhibit a significantly higher peak coupling (0.70) than the N/F subset (0.59). The introduction of nitrogen or fluorine atoms, which often contribute lone pairs or induce strong inductive effects, partially restores the electronic-spatial decoupling even in highly saturated systems. This suggests that the high-saturation recoupling is a prominent feature of pure hydrocarbon or oxygen-linked frameworks, which is subsequently diluted by the more complex, localized electronic perturbations introduced by nitrogen and fluorine.

Structural control tests confirmed that this coupling is not an artifact of molecular size. Large non-aromatic molecules (heavy atom count ≥ 8, $FractionCSP3 > 0.8$) maintained a mean coupling of 0.25, while small rings (heavy atom count ≤ 6) showed high variance but followed the same upward trajectory. This indicates that the recoupling effect is intrinsic to the bonding hybridization rather than a trivial consequence of molecular size constraints.

3.5 Stability and statistical validation

The reported structures are statistically robust and stable across molecular families. Randomly partitioning the dataset by unique SMILES into two split halves (Half A versus Half B) yielded nearly identical mean $dCor$ values (0.2648 vs. 0.2628), with an absolute difference of only 0.002.

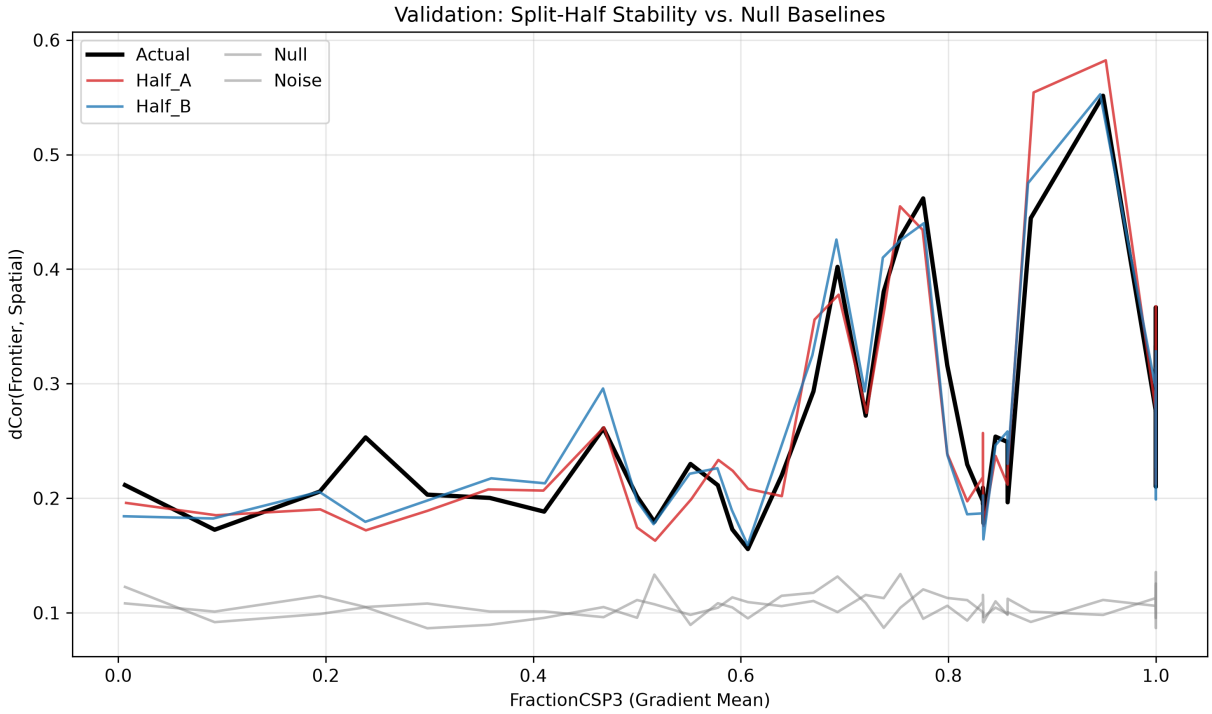

Figure 4: Validation of the joint electronic-spatial structure stability across the `FractionCSP3` gradient. The plot displays the distance correlation ($dCor$) between the frontier electronic vector and the spatial manifold for the full dataset, two random split-half partitions (`Half_A` and `Half_B`), and two baselines (Null and Noise). The nearly identical trajectories of the actual data and the split halves demonstrate that the observed saturation-dependent re-coupling is statistically robust across molecular families. Furthermore, the substantial signal-to-null gap between the actual data and the flat baseline models confirms that the transition from global independence to spatial re-coupling in highly saturated molecules represents genuine chemical information rather than statistical noise.

As shown in Figure 4, the measured signal ($dCor \approx 0.266$) significantly exceeds the null hypothesis baseline generated by shuffled residuals ($dCor \approx 0.106$) and the random noise baseline ($dCor \approx 0.106$). This substantial signal-to-null gap of 0.160 confirms that the observed joint structure represents genuine chemical information rather than statistical noise. Furthermore, the trajectories remained stable across window sizes of 3%, 5%, and 7%, ensuring that the observed localized-to-global transition is not an artifact of the smoothing or sampling frequency.

4 Conclusions

The conventional assumption in physical chemistry dictates that the highest occupied molecular orbital (HOMO) energy is governed primarily by local electronic environments and remains largely independent of a molecule's macroscopic spatial-rotational geometry. In this study, we challenged this paradigm by tracking the joint structure of frontier orbital energies and global molecular shape across a continuous gradient of molecular saturation. By isolating genuine chemical effects from stoichiometric size scaling through gradient-boosted residuals and robust Minimum Covariance Determinant (MCD) estimators, we demonstrated that the spatial orthogonality of the HOMO is not a universal physical constant, but rather a context-dependent state.

Our non-parametric distance correlation analysis reveals a distinct localized-to-global transition within the QM9 dataset. In unsaturated and aromatic regimes, the HOMO energy is indeed highly decoupled from the global spatial manifold. However, as the fraction of sp^3-hybridized carbons ($FractionCSP3$) increases, the frontier electronic residuals dynamically recouple with the global 3D geometry. Partial distance correlation analyses further highlight a fundamental asymmetry in this process: the lowest unoccupied molecular orbital (LUMO) drives spatial coupling in intermediate transition zones, while the HOMO remains uniquely isolated. The HOMO couples to the global manifold only when constrained by the rigid, directed bonds of fully saturated scaffolds.

Furthermore, we found that this saturation-dependent recoupling is highly sensitive to heteroatom substitution. As summarized in Table 1.3, propensity-matched comparisons demonstrate that the electronic-spatial coupling is most pronounced in pure hydrocarbon and oxygen-linked frameworks.

The introduction of nitrogen and fluorine atoms partially restores the electronic-spatial decoupling, even in highly saturated systems. This indicates that the lone pairs and strong inductive effects of these heteroatoms introduce localized electronic perturbations that override global geometric drivers, diluting the structural coherence seen in pure CHO frameworks.

Ultimately, these findings establish that macroscopic molecular shape becomes a primary driver of frontier electronic levels only when the stabilizing effects of π-delocalization are absent. Recognizing the dynamic, saturation-dependent nature of HOMO-spatial recoupling provides a more nuanced understanding of molecular property spaces. This localized-to-global transition offers valuable physical insights that can inform the design of saturated organic materials and guide the refinement of structure-property predictive models.