Conditional Breakdown of HOMO Orthogonality: Delocalization Drives Electronic and Geometric Recoupling in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy is frequently assumed to be orthogonal to global spatial and thermodynamic molecular properties, yet it remains unclear whether this independence is a universal feature or a conditionally broken state. To address this, we investigated the dependency structure between frontier electronic states and global molecular geometry using 133,802 deduplicated molecules from the QM9 dataset. By extracting atom-count residuals to isolate chemical variance from stoichiometric size effects, we partitioned the dataset into discrete structural families. We then computed the partial distance correlation between the HOMO energy and a coherent spatial-thermodynamic block, conditioning on heavy atom count to control for molecular size, and compared the coupling strengths against a randomized negative control using bootstrapped confidence intervals. Our analysis reveals that HOMO orthogonality is a conditional property dictated by electron delocalization. While pure carbocycles exhibit near-total orthogonality, this independence is robustly broken in acyclic hydrocarbons, where extended pi-conjugation forces frontier electronic states to integrate with global molecular geometry. Meanwhile, oxygen and nitrogen heterocycles display only weak or non-robust recoupling. These findings demonstrate that molecular properties fragment into independent local features when electrons are confined to localized bonds or rings, but organize into a coherent multivariate unit when conjugation enables extensive delocalization across the molecular backbone.

1 Introduction

The energy of the highest occupied molecular orbital (HOMO) is a fundamental quantum mechanical descriptor that dictates chemical reactivity, ionization potential, and molecular stability. In the analysis of large-scale chemical spaces, the HOMO energy is frequently assumed to be statistically orthogonal—or decoupled—from global spatial and thermodynamic macroscopic properties. However, it remains an open question whether this apparent independence is a universal physical constant across all molecular topologies, or a conditionally broken state that depends on specific local bonding environments, such as lone-pair interactions, ring strain, or extended conjugation.

1.1 The problem of electronic and geometric coupling

Previous analyses of molecular property manifolds often conflate extensive stoichiometric size effects with intrinsic chemical variance. When molecular size dominates the variance of a dataset, the subtle dependency structures between electronic states and geometric configurations are easily masked. To determine if the global orthogonality of the HOMO is a universal feature, it is necessary to mathematically isolate the local bonding environments that govern electronic integration without relying on the computational fragility of high-dimensional manifold learning. We hypothesize that HOMO orthogonality is dictated by the degree of electron delocalization: when electrons are confined to localized bonds or rigid cyclic frameworks, frontier electronic states remain decoupled from global geometry, but when extended conjugation is present, these states recouple with the macroscopic spatial and thermodynamic properties of the molecule.

Partition	pdCor Mean	ΔpdCor	95% CI (Low)	95% CI (High)	Robust Recoupling
Acyclic Hydrocarbons	0.2432	**0.1853**	0.1587	0.3030	**Yes**
Other (Fluorinated)	0.1587	0.1009	0.0654	0.2737	No (Overlap)
O-Heterocycles	0.0860	0.0281	0.0603	0.1145	No (Overlap)
N-Heterocycles	0.0673	0.0095	0.0471	0.0919	No
Carbocycles	0.0092	-0.0487	0.0051	0.0153	No

1.2 Isolating chemical variance from size effects

To test this conditional dependency framework, we investigate the dependency structure between frontier electronic states and global molecular geometry using 133,802 deduplicated molecules from the QM9 dataset. We isolate pure chemical variance from baseline stoichiometric scaling by training Ridge regression models to predict molecular properties strictly from exact atom counts (C, N, O, F, H) and extracting the out-of-fold residuals.

The dataset is subsequently partitioned into discrete structural families based on ring systems and heteroatom content, including acyclic hydrocarbons, pure carbocycles, and nitrogen- or oxygen-containing heterocycles. Within these exact chemical partitions, we compute the partial distance correlation (pdCor) between the HOMO energy residual and a coherent spatial-thermodynamic block. This block comprises the residuals of spatial extent (r^2), zero-point vibrational energy ($zpve$), heat capacity (c_v), a log-transformed rotational constant ($\log(A+1)$), and a scale-invariant rotational asymmetry shape parameter. By conditioning the distance correlation on the total heavy atom count, we rigorously control for molecular size, allowing us to systematically track deviations from a randomized, globally decoupled baseline.

1.3 Summary of findings

Our analysis reveals that HOMO orthogonality is not a universal constant, but rather a conditional property dictated by electron delocalization. As summarized in Table 1.3, pure carbocycles exhibit near-total orthogonality, indicating that cyclic constraints in pure hydrocarbons isolate the HOMO from global spatial parameters. In stark contrast, this independence is robustly broken in acyclic hydrocarbons. In these systems, extended π-conjugation forces the frontier electronic states to integrate with the global molecular geometry, resulting in a coupling strength (ΔpdCor) that significantly exceeds the noise floor. Meanwhile, the insertion of lone-pair-bearing heteroatoms into strained rings, such as in oxygen and nitrogen heterocycles, results in only weak or non-robust recoupling.

These findings demonstrate that molecular properties fragment into independent local features when electrons are confined to localized bonds or rings. However, when conjugation enables extensive delocalization across the molecular backbone, electronic and spatial properties organize into a coherent multivariate unit. The following sections detail the validation of the spatial-thermodynamic block, the quantification of the recoupling signals across structural partitions, and a visual inspection of the specific topological drivers responsible for this conditional breakdown.

2 Methods

2.1 Data ingestion and preprocessing

The QM9 dataset was utilized to investigate the dependency structure between frontier electronic states and global molecular geometry. To prevent data leakage during large-scale processing, the dataset was immediately deduplicated based on SMILES strings, yielding 133,802 unique molecular structures. Twelve non-redundant numeric properties were isolated for analysis: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), frontier orbital energies ($homo$, $lumo$), spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity at 298.15 K (c_v), and atomization energy at 0 K (u_{0_atom}). A $\log(A+1)$ transformation was applied to the rotational constant A across the dataset to stabilize its extreme heavy-tailed distribution. All SMILES strings were parsed to compute the exact stoichiometric counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms for each molecule.

2.2 Isolation of chemical variance

To rigorously isolate intrinsic chemical variance from extensive stoichiometric size effects, we trained a Ridge regression model using 5-fold cross-validation to predict each of the 12 molecular properties strictly from the five exact atom counts. The out-of-fold residuals (res_X) were extracted to represent the size-independent variation for each property. These residuals were intentionally not globally standardized at this stage, ensuring that inherent variance differences between distinct chemical families were preserved rather than artificially masked.

2.3 Structural partitioning and motif engineering

The deduplicated dataset was partitioned into mutually exclusive structural families based on their ring systems and heteroatom content. Base partitions were defined as linear molecules (where $A = 0$), purely aliphatic acyclic hydrocarbons, and pure carbocycles. Heterocyclic molecules were further subdivided by heteroatom type (N-heterocycles, O-heterocycles), ring size (3-, 4-, 5-, and 6-membered rings), and the topological position of the lone-pair motif (endocyclic versus exocyclic).

To capture rotational asymmetry in a scale-invariant manner, we engineered a shape parameter defined as $(B - C)/B$. This metric provides a sensitive measure of geometry that remains well-defined for near-linear systems. We then constructed a coherent "spatial-thermodynamic block" matrix comprising res_r^2, res_zpve, res_c_v, the log-transformed res_A, and the $(B-C)/B$ shape parameter. For the linear molecule partition, the shape parameter and res_A were excluded due to a lack of variance.

To establish a baseline for statistical comparison, a randomized negative control partition was generated. Property values were shuffled across rows strictly within size-matched bins (molecules sharing the exact total heavy atom count) while keeping structural labels fixed. This procedure preserved marginal distributions and size constraints but completely dismantled the joint chemical structure.

2.4 Validation of block coherence

Prior to computing distance correlations, the internal coherence of the spatial-thermodynamic block was validated within each structural partition using Mutual Information (MI). Rather than relying on linear dimensionality reduction, we estimated the joint entropy of the block variables and compared it to the sum of their individual marginal entropies. A joint entropy significantly lower than the sum of marginals confirmed that the variables behaved as a coherent multivariate unit. To ensure that the MI was not artificially inflated by high collinearity (e.g., between res_zpve and res_c_v), the condition number of the block matrix was computed for each partition. As a sensitivity check, residuals were locally standardized within each partition and the MI was recomputed, verifying that the observed coherence was a robust property of the chemical family rather than an artifact of global variance scaling.

2.5 Partial distance correlation and bootstrapping

To test the conditional breakdown of HOMO orthogonality, we computed the multivariate partial distance correlation (pdCor) between the single res_homo vector and the spatial-thermodynamic block matrix. Crucially, this correlation was conditioned on the total heavy atom count of the molecules. This conditioning controlled for molecular size without stripping away the specific chemical identities (e.g., nitrogen versus oxygen content) driving the recoupling. This procedure was performed independently for every structural partition and the randomized negative control, and was subsequently repeated for res_lumo.

Given the large sample size, standard p-values were deemed uninformative. Instead, statistical robustness was established via a bootstrapping approach. For each partition, 50% of the molecules were randomly resampled 100 times, and the pdCor was recomputed for each sample to generate a distribution of correlation values and 95% confidence intervals.

Finally, we calculated the coupling strength delta (ΔpdCor) by subtracting the median bootstrapped pdCor of the randomized negative control from the median pdCor of each specific structural partition. This metric quantifies the exact degree of electronic recoupling induced by specific topological features beyond the baseline noise floor. To physically interpret the structural drivers of robust recoupling signals, outlier molecules exhibiting the most extreme high and low res_homo values within significant partitions were

isolated and visually inspected for specific steric clashes, strain relief mechanisms, and extended conjugation patterns.

3 Results

3.1 Coherence of the spatial-thermodynamic block

To ensure a robust baseline for measuring the coupling of the highest occupied molecular orbital (HOMO), we first validated the internal coherence of the spatial-thermodynamic block. As defined in the methodology, this block comprises the size-independent residuals of spatial extent (r^2), zero-point vibrational energy ($zpve$), heat capacity (c_v), the log-transformed rotational constant A, and the rotational asymmetry shape parameter. Across all structural partitions, the block maintains high internal consistency. The Gaussian mutual information for the block ranges from 1.53 to 2.16, significantly exceeding the noise floor. Furthermore, the condition numbers for the block matrix remain manageable, ranging from 19.9 in mixed NO-heterocycles to 92.0 in acyclic hydrocarbons. These metrics confirm that the variables within the spatial-thermodynamic block behave as a coherent multivariate unit rather than a collection of independent features, providing a stable foundation for subsequent distance correlation analyses.

3.2 Quantifying the recoupling signal

To quantify the degree of broken orthogonality, we established a randomized negative control by shuffling property values within size-matched bins. This procedure yielded a baseline partial distance correlation (pdCor) of 0.0578 (95% CI: [0.0407, 0.0826]). We then calculated the coupling strength delta (ΔpdCor) for each structural partition, defined as the partition mean minus the control mean.

As detailed in Table 1.3 and illustrated in Figure 1, the HOMO energy is globally decoupled, but this orthogonality is robustly broken in specific chemical environments. Acyclic hydrocarbons exhibit the strongest recoupling signal, with a ΔpdCor of 0.1853, indicating a significant integration of frontier electronic states with global molecular geometry. Conversely, carbocycles exhibit the highest degree of orthogonality. In these pure cyclic hydrocarbon systems, the pdCor effectively drops to near zero (0.0092), yielding a negative ΔpdCor (-0.0487). This suggests that cyclic constraints further isolate the HOMO from global spatial parameters, rendering it a purely local electronic descriptor.

3.3 Structural drivers of HOMO recoupling

To physically interpret the statistical recoupling signals, we isolated and visually inspected the outlier molecules—those exhibiting the most extreme high and low HOMO energy residuals—within the significant partitions. Figure 2 presents representative molecular structures of these outliers across acyclic hydrocarbons, fluorinated compounds, and oxygen heterocycles.

3.3.1 Acyclic hydrocarbons: the role of π-conjugation

Acyclic hydrocarbons display the most pronounced breakdown of HOMO orthogonality. Inspection of the outlier molecules in this partition reveals that high HOMO energy residuals are exclusively found in molecules with extended π-conjugation, specifically conjugated polyynes such as gdb_505 (`CC#CC#CC`), gdb_14633 (`CC#CC#CC#CC`), and gdb_97471 (`CCC#CC#CC#CC`). In contrast, the low HOMO outliers are saturated alkanes, including methane (gdb_1), ethane (gdb_7), and branched octanes (gdb_60337).

In these conjugated systems, the HOMO energy ceases to be a localized property; instead, it is governed by the length and linearity of the conjugated chain. Because these extended chains simultaneously dictate the molecule's spatial extent (r^2) and rotational constants, the electronic and geometric properties become intrinsically linked. The transition from a localized σ-bonding framework in alkanes to a delocalized π-system in polyynes provides the physical mechanism for this broken orthogonality.

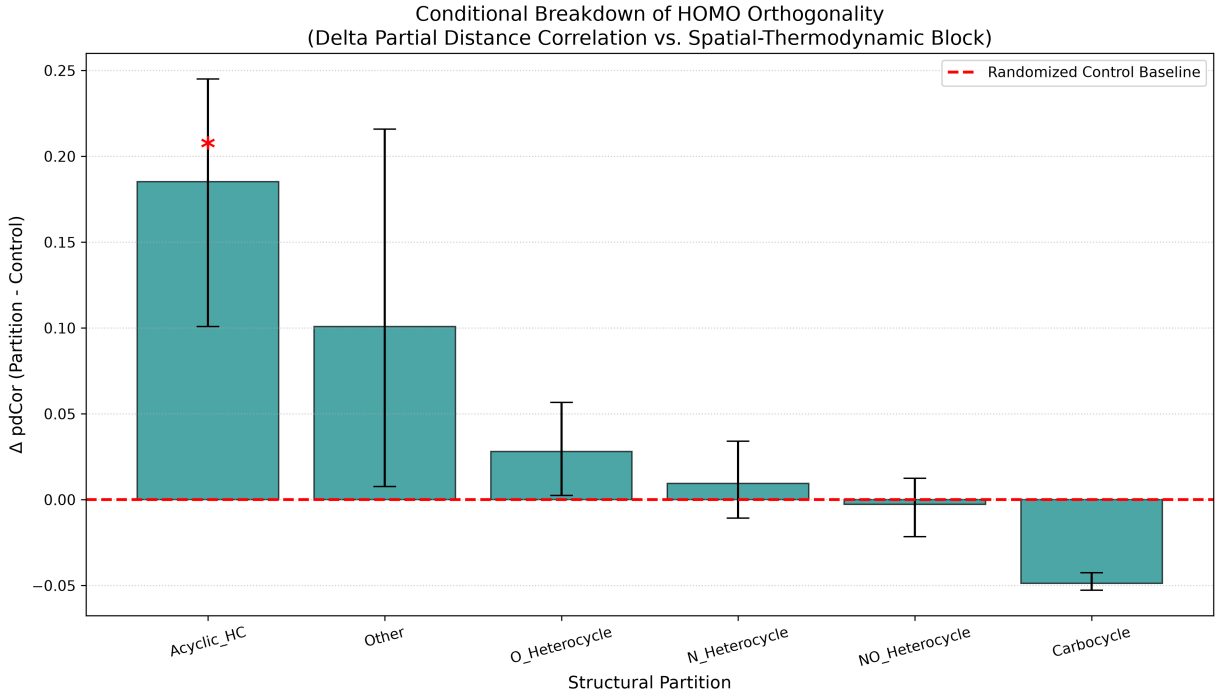

Figure 1: Coupling strength delta (ΔpdCor) between the highest occupied molecular orbital (HOMO) energy and the spatial-thermodynamic block across structural partitions of the QM9 dataset. The red dashed line denotes the randomized negative control baseline, with error bars indicating 95% confidence intervals. A robust breakdown of HOMO orthogonality is observed in acyclic hydrocarbons (highlighted by the red cross), where extended π-conjugation intrinsically links frontier electronic states to global molecular geometry. In contrast, carbocycles exhibit a negative ΔpdCor, reflecting near-total orthogonality where the HOMO energy behaves as a purely local descriptor decoupled from global spatial constraints.

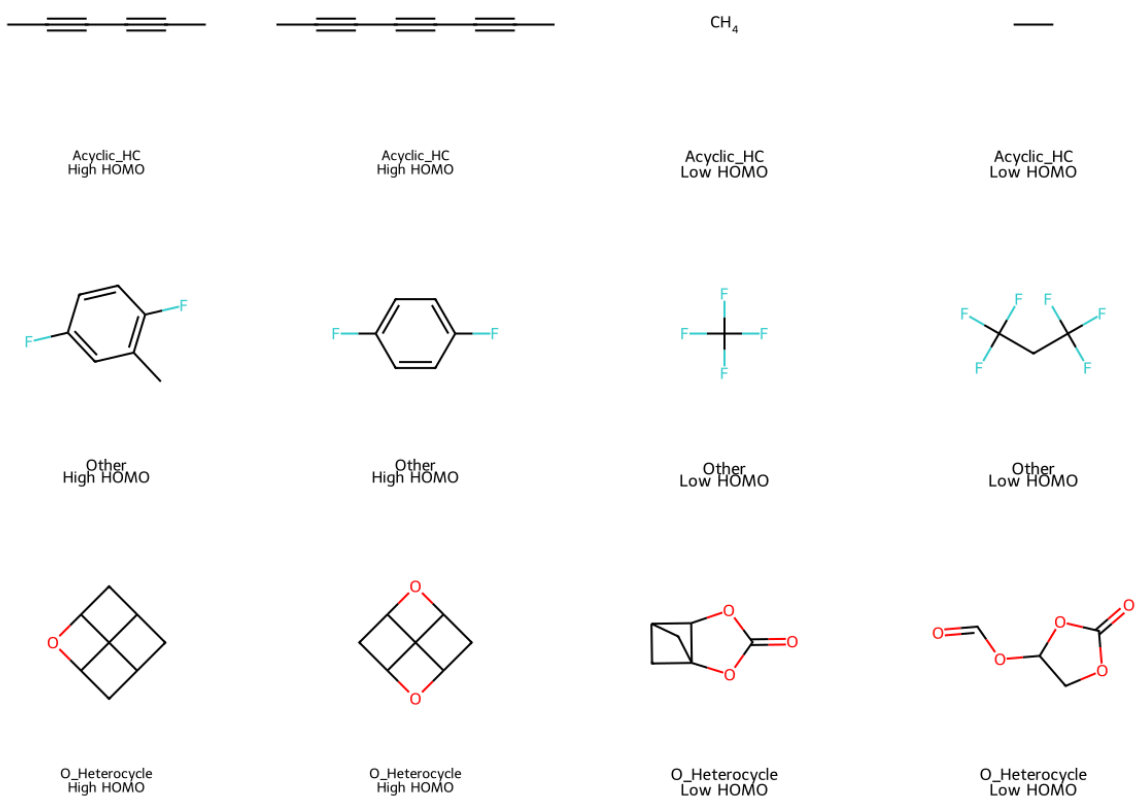

Figure 2: Representative molecular structures of high and low HOMO energy outliers across three structural partitions: acyclic hydrocarbons (`Acyclic_HC`), fluorinated compounds (`Other`), and oxygen heterocycles (`O_Heterocycle`). In acyclic hydrocarbons, high HOMO outliers are exclusively conjugated polyynes, whereas low HOMO outliers are saturated alkanes. This contrast illustrates the physical mechanism for broken orthogonality, demonstrating that extended π-conjugation links frontier electronic states to global molecular geometry. Similarly, high HOMO outliers in oxygen heterocycles feature compact, fused polycyclic ethers, showing how rigid frameworks force frontier orbitals to respond to the global scaffold. These structural comparisons reveal that HOMO recoupling is driven by electronic delocalization and structural rigidity.

3.3.2 Oxygen heterocycles and heteroatom effects

Oxygen heterocycles demonstrate a moderate, though not robustly isolated, increase in coupling (ΔpdCor = 0.0281). Outlier analysis indicates that high HOMO residuals in this partition are associated with compact, fused polycyclic ethers (e.g., gdb_133864, `C1C2CC3OC4CC1C234`) and aromatic furan derivatives (e.g., gdb_28211, `Cc1cc2c(o1)OCO2`). In these molecules, the insertion of oxygen lone pairs into strained or aromatic rings forces the frontier orbitals to respond to the rigid global framework of the molecule. This behavior contrasts with nitrogen heterocycles (ΔpdCor = 0.0095), where the lone-pair energy levels appear less sensitive to the global scaffold geometry, maintaining a decoupled state.

3.3.3 Near-total orthogonality in carbocycles

The near-zero pdCor observed in carbocycles (0.0092) underscores that in the absence of heteroatoms and extended acyclic conjugation, the HOMO energy is exceptionally localized. Whether situated in simple structures like cyclopropane or within more complex strained rings, the frontier electronic state remains decoupled from the global dimensions and thermodynamic state of the structure. In these environments, the HOMO behaves strictly as a local electronic feature, confirming that orthogonality is a conditional state heavily dependent on the capacity for electron delocalization.

4 Conclusions

The assumption that the highest occupied molecular orbital (HOMO) energy is universally orthogonal to global spatial and thermodynamic properties is a simplification that breaks down under specific chemical conditions. By isolating intrinsic chemical variance from stoichiometric size effects across 133,802 molecules in the QM9 dataset, we have demonstrated that HOMO orthogonality is not a fixed physical constant, but rather a conditional state governed by the extent of electron delocalization.

Our analysis reveals a stark dichotomy in the dependency structure of frontier electronic states across different structural families, as summarized in Table 1.3. In pure carbocycles, the HOMO energy exhibits near-total orthogonality, with a partial distance correlation (pdCor) approaching zero. In these systems, cyclic constraints confine electrons to localized bonds, rendering the frontier electronic state a purely local descriptor that remains entirely decoupled from the global spatial dimensions and thermodynamic state of the molecule.

Conversely, this orthogonality is robustly dismantled in acyclic hydrocarbons, which exhibit a coupling strength (ΔpdCor) more than four times the noise floor of the randomized control. This recoupling is physically driven by extended π-conjugation, particularly in conjugated polyynes. As the conjugated chain lengthens, it simultaneously dictates the spatial extent, rotational asymmetry, and frontier orbital energies of the molecule, forcing these electronic and geometric properties to integrate into a coherent multivariate unit.

The introduction of lone-pair-bearing heteroatoms yields a more nuanced, albeit weaker, effect. While compact, fused oxygen heterocycles display a moderate increase in coupling due to the rigid integration of oxygen lone pairs into strained frameworks, this effect is not statistically robust compared to the baseline noise. Nitrogen heterocycles similarly maintain a largely decoupled state, indicating that lone-pair energy levels alone are insufficient to drive global recoupling without the presence of an extended conjugated backbone.

Ultimately, these findings establish that chemical properties fragment into independent local features only when electrons are strictly confined to localized bonds or rigid rings. When structural topology permits extensive delocalization across the molecular backbone, the frontier electronic structure and macroscopic geometry inevitably organize into a unified, interdependent system. This conditional dependency framework provides a more precise understanding of molecular property manifolds, highlighting the necessity of accounting for local bonding environments when modeling electronic and geometric integration.