Variance Decomposition of the HOMO-LUMO Gap to Isolate Heteroatom-Driven Electronic Decoupling

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In chemical datasets, the highest occupied molecular orbital (HOMO) frequently appears decoupled from global thermodynamic and spatial properties, raising the question of whether this orthogonality is a universal phenomenon or a localized consequence of specific chemical motifs. To investigate this, we perform an exact variance decomposition of the HOMO-LUMO gap across 133,802 molecules from the QM9 dataset. We isolate chemical signals from macroscopic size effects by computing property residuals with LightGBM regressors conditioned on an atom-count baseline. We then evaluate the polarizability hub hypothesis using distance correlations, map the local covariance of the residual manifold via UMAP and k-nearest neighbors, and conduct a matched molecular pair perturbation analysis to quantify directional orbital shifts upon heteroatom substitution. Our analysis reveals that the observed HOMO decoupling is a motif-dependent effect driven by heteroatom introduction rather than a global constant. While pure hydrocarbons exhibit strong anti-directional HOMO-LUMO coupling, the addition of electronegative atoms such as nitrogen, oxygen, or fluorine shifts this relationship to a predominantly co-directional regime. Furthermore, mesomeric substitutions exert significantly larger shifts on the HOMO compared to purely inductive substitutions. These findings demonstrate that global orbital decoupling is an averaging artifact of structurally distinct regimes, providing a rigorous, stoichiometry-independent mechanism to pinpoint the chemical drivers of electronic behavior.

1 Introduction

The energy levels of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) are fundamental parameters for predicting chemical reactivity, stability, and optical properties. In large-scale chemical datasets such as QM9, a persistent and curious phenomenon is observed: the HOMO frequently appears decoupled from the global thermodynamic and spatial manifold. While properties like zero-point vibrational energy, enthalpy, and spatial extent scale predictably with molecular size and topology, the HOMO energy exhibits a stubborn orthogonality to these macroscopic features.

This observation raises a critical question: is this electronic decoupling a universal feature of chemical space, or is it a localized consequence of specific chemical motifs? Traditional global correlation analyses often obscure local structural regimes by averaging over highly diverse molecular families. If the decoupling is motif-dependent, treating it as a global constant limits the accuracy of predictive models and obscures the underlying physical organic chemistry governing orbital energies.

To investigate this, we perform an exact variance decomposition of the HOMO-LUMO gap across 133,802 deduplicated molecules from the QM9 dataset. Because the gap is defined strictly as the difference between the LUMO and HOMO energies, the variance of their residuals follows an exact mathematical relationship:

$$\mathrm{Var}(\mathrm{res_gap}) = \mathrm{Var}(\mathrm{res_lumo}) + \mathrm{Var}(\mathrm{res_homo}) - 2\mathrm{Cov}(\mathrm{res_homo}, \mathrm{res_lumo}) \tag{1}$$

By isolating these components, we avoid generic correlation searches and establish a rigorous, stoichiometry-independent framework to pinpoint the precise chemical drivers of electronic behavior.

To separate true chemical signals from macroscopic size effects, we first compute property residuals using gradient-boosted decision trees conditioned on a strict atom-count baseline. This isolates the variance attributable to local bonding environments rather than stoichiometric determinism. Using this residual

manifold, we evaluate the hypothesis that isotropic polarizability acts as a central hub mediating orbital decoupling. Furthermore, we map the local covariance of the residual space using dimensionality reduction and k-nearest neighbors, and we conduct a matched molecular pair (MMP) perturbation analysis to quantify directional orbital shifts upon heteroatom substitution.

Our analysis reveals that the observed HOMO decoupling is not a global constant, but rather a motif-dependent effect driven by the introduction of heteroatoms. We demonstrate that pure hydrocarbons exhibit strong anti-directional HOMO-LUMO coupling, whereas the addition of electronegative atoms such as nitrogen, oxygen, or fluorine shifts this relationship to a predominantly co-directional regime. By categorizing substitutions into inductive and mesomeric contexts, we show that global orbital decoupling is an averaging artifact of these structurally distinct regimes. Ultimately, this work provides a robust mechanism to isolate and understand heteroatom-driven electronic rewiring across chemical space.

2 Methods

2.1 Data curation and structural featurization

The analysis was conducted using the QM9 dataset. To prevent I/O bottlenecks during parallel processing, the RDKit logger was explicitly silenced across all worker functions. Following the parsing of SMILES strings, the dataset was deduplicated to yield 133,802 unique molecular structures. Redundant thermodynamic columns (including zero-point uncorrected and atom-averaged variations of internal energy, enthalpy, and free energy) were removed. The HOMO-LUMO gap was also excluded from the primary feature set, as it is strictly defined as the difference between the lowest unoccupied molecular orbital (LUMO) and the highest occupied molecular orbital (HOMO).

For the remaining 12 non-redundant properties, we extracted a set of baseline structural features designed to capture macroscopic molecular size and basic topology. These included the exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen, alongside ring counts and the number of rotatable bonds. To capture local bonding environments, RDKit was utilized to extract granular structural descriptors via SMARTS matching, atom degree calculations, and conjugation checks. These local descriptors included lone-pair counts, hybridization states, specific functional motifs, and positional contexts (e.g., terminal versus internal, conjugated versus non-conjugated).

2.2 Establishing the stoichiometric baseline via gradient boosting

To isolate true chemical signals from macroscopic size and basic topological effects, we established a stoichiometric baseline using gradient-boosted decision trees. LightGBM regressors, accelerated via GPU, were trained to predict the 12 non-redundant QM9 properties using only the baseline structural features (atom counts, ring counts, and rotatable bonds).

For each property X, we computed the residual res_X by subtracting the LightGBM prediction from the true QM9 value. These residuals represent the property variations stripped of stoichiometric determinism, effectively isolating the variance attributable to local bonding environments and heteroatom perturbations. This 12-dimensional residual space served as the fixed, definitive representation of the chemical space for all subsequent analyses.

2.3 Statistical isolation of the polarizability hub

To evaluate the hypothesis that isotropic polarizability (α) acts as a central hub mediating the decoupling of the HOMO from the global manifold, we employed distance correlation metrics. The distance correlation between res_homo and the other 11 property residuals was computed to quantify non-linear dependencies.

Using the `pingouin` statistical package, we subsequently computed the partial distance correlation between res_homo and the remaining residuals, explicitly controlling for the polarizability residual (res_alpha). We specifically examined whether res_homo maintained a significant correlation with the gap residual (res_lumo − res_homo) after partialling out res_alpha. To further isolate purely local electronic effects, res_homo was regressed on res_alpha, and the resulting "local-only" HOMO residual was tested for coupling against the previously extracted local descriptors, such as lone-pair counts.

Holdout Set	N	Partial R (H/L	α)	Distance Correlation (H/L)	Mean Local Covariance

2.4 Manifold embedding and local covariance mapping

To visualize and quantify the local electronic coupling across different chemical regimes, the 12-dimensional residual space was projected into a two-dimensional embedding using Uniform Manifold Approximation and Projection (UMAP).

Simultaneously, a k-nearest neighbors (k-NN) graph was constructed in the original 12-dimensional residual space using $k = 40$ and a Euclidean distance metric. This neighborhood construction was fully parallelized. For each molecule's local neighborhood, we calculated the Pearson covariance between res_homo and res_lumo. The UMAP projection was then colored by this local covariance using a diverging colormap centered at zero, allowing regions of strong positive (co-directional) or negative (anti-directional) orbital coupling to be visually identified and mapped back to their defining structural motifs.

2.5 Matched molecular pair perturbation analysis

To pinpoint the precise chemical drivers of electronic rewiring, we conducted a perturbation analysis using matched molecular pairs (MMPs). MMPs were defined as pairs of molecules differing by exactly one heavy atom substitution while strictly maintaining the same total number of heavy atoms. To avoid the $O(N^2)$ computational cost of brute-force comparisons, we implemented a hash-based approach that canonicalized SMILES strings after the removal of the substituted atom, allowing pairs to be identified in $O(N)$ time.

The resulting 235,691 unique MMPs were filtered to distinguish between isosteric substitutions (e.g., replacing a carbon atom with a nitrogen atom in a ring) and functional substitutions. The structural context of each substitution was granularly categorized into inductive-only environments (e.g., fluorine on an aliphatic chain), mesomeric-donating environments (e.g., oxygen or nitrogen in a conjugated system), and mesomeric-withdrawing environments.

For each MMP, we calculated the signed shifts in the orbital residuals:

$$\Delta\text{res_homo} = \text{res_homo}_{\text{perturbed}} - \text{res_homo}_{\text{reference}} \tag{2}$$

$$\Delta\text{res_lumo} = \text{res_lumo}_{\text{perturbed}} - \text{res_lumo}_{\text{reference}} \tag{3}$$

Pairs were categorized based on the relative signs of these shifts to determine whether the HOMO and LUMO moved co-directionally or anti-directionally upon heteroatom substitution. Mann-Whitney U tests were performed to statistically compare the magnitudes of orbital shifts caused by purely inductive substitutions versus those with competing mesomeric effects.

2.6 Motif-based holdout validation

To ensure that the predictive rules derived from the MMP and local covariance analyses were not artifacts of the dataset's specific composition, we performed a motif-based holdout validation. The dataset was partitioned into discrete structural classes based on the presence or absence of specific heteroatoms (e.g., pure hydrocarbons, nitrogen-containing molecules, and fluorine-containing molecules).

The local HOMO-LUMO covariance distributions and the partial correlations (testing the polarizability hub hypothesis) were re-evaluated independently on these holdout sets. This validation step, summarized in Table 2.6, confirmed whether the mechanistic rules governing the HOMO's decoupling generalized across entirely different chemical contexts or were specific to localized structural regimes.

3 Results

3.1 Isolation of chemical signals via stoichiometric baseline

To rigorously investigate the electronic decoupling of the highest occupied molecular orbital (HOMO), it is first necessary to separate true chemical signals from macroscopic size effects. As detailed in the methods, all properties were conditioned on an atom-count baseline (C, N, O, F, H) using LightGBM regressors.

The baseline analysis confirms that thermodynamic properties are almost entirely dictated by stoichiometric determinism; for instance, the zero-point vibrational energy (zpve) and internal energy at 0 K (u_0) exhibit R^2 values of 0.997 and approximately 1.00, respectively. In stark contrast, electronic properties remain largely independent of basic atom counts, with the HOMO energy yielding an R^2 of only 0.0855. This pronounced independence necessitates the use of the resulting property residuals (res_X) to identify true joint structures in the electronic manifold, free from the confounding influence of molecular size.

3.2 Evaluation of the polarizability hub hypothesis

We first evaluated the hypothesis that isotropic polarizability (α) acts as a central hub mediating the apparent decoupling of the HOMO from the global manifold. In the 12-dimensional residual space, res_homo exhibits remarkably low coupling with nearly all other properties, yielding distance correlations ($dCor$) consistently below 0.33.

The distance correlation between res_homo and res_alpha is 0.322. However, controlling for polarizability fails to collapse the relationship between orbital energies and other properties. Specifically, the distance correlation between res_homo and res_lumo is 0.141, and this relationship remains significant even after accounting for res_alpha. Furthermore, the gap residual (res_gap = res_lumo − res_homo) shows the strongest coupling to the overall manifold ($dCor = 0.496$). When res_alpha is partialled out, the partial distance correlation between res_homo and res_gap remains robust at 0.237. These results indicate that the decoupling of the HOMO is not merely a byproduct of molecular size or polarizability, but is instead rooted in localized electronic environments.

3.3 Local covariance mapping of the residual manifold

To visualize and quantify these localized electronic environments, the 12-dimensional residual manifold was projected into two dimensions using UMAP, and the local Pearson covariance between res_homo and res_lumo was calculated within a 40-nearest neighbor graph.

While the global mean local covariance is near zero (2.14×10^{-6} Hartree^2), the distribution is highly non-normal, possessing a pronounced heavy tail ($P_{99} = 4.37 \times 10^{-5}$ Hartree^2). As illustrated in Figure 1, this heavy tail demonstrates that high-coupling, co-directional orbital shifts are present within the dataset, contrasting sharply with regions of low or negative covariance.

The UMAP projection (Figure 2) reveals distinct "islands" of electronic behavior, confirming manifold heterogeneity. High-covariance regions (red) are strongly associated with specific heteroatom motifs, whereas pure hydrocarbons occupy the regions of low or negative local covariance (blue). This spatial segregation suggests that the global orthogonality of the HOMO is an averaging artifact of structurally distinct chemical regimes.

3.4 Directional orbital shifts in matched molecular pairs

To pinpoint the precise chemical drivers of this electronic re-wiring, we performed a perturbation analysis using 235,691 unique matched molecular pairs (MMPs) differing by exactly one heavy atom. We measured the signed shifts in residuals (Δres_homo and Δres_lumo) upon substituting a carbon atom with a heteroatom (N, O, or F).

Mann-Whitney U tests confirmed that mesomeric-donating substitutions (e.g., in aromatic or conjugated contexts) induce significantly larger HOMO shifts than purely inductive substitutions (e.g., in saturated aliphatic chains). The mean absolute shift for the HOMO residual (|Δres_homo|) is 0.0094 Hartree for inductive substitutions, compared to 0.0206 Hartree for mesomeric substitutions ($p < 10^{-300}$). Interestingly, the LUMO residual shifts remain relatively constant across both contexts (mean |Δres_lumo| of 0.0126 Hartree for inductive and 0.0132 Hartree for mesomeric).

Analysis of coupling directionality further highlights the role of structural context (Figure 3). In inductive contexts, 55.97% of pairs exhibit co-directional shifts (both orbitals moving in the same direction), while a substantial 44.03% exhibit anti-directional shifts (orbitals moving closer together or further apart). In mesomeric contexts, however, the relationship becomes overwhelmingly co-directional (82.40%), with anti-

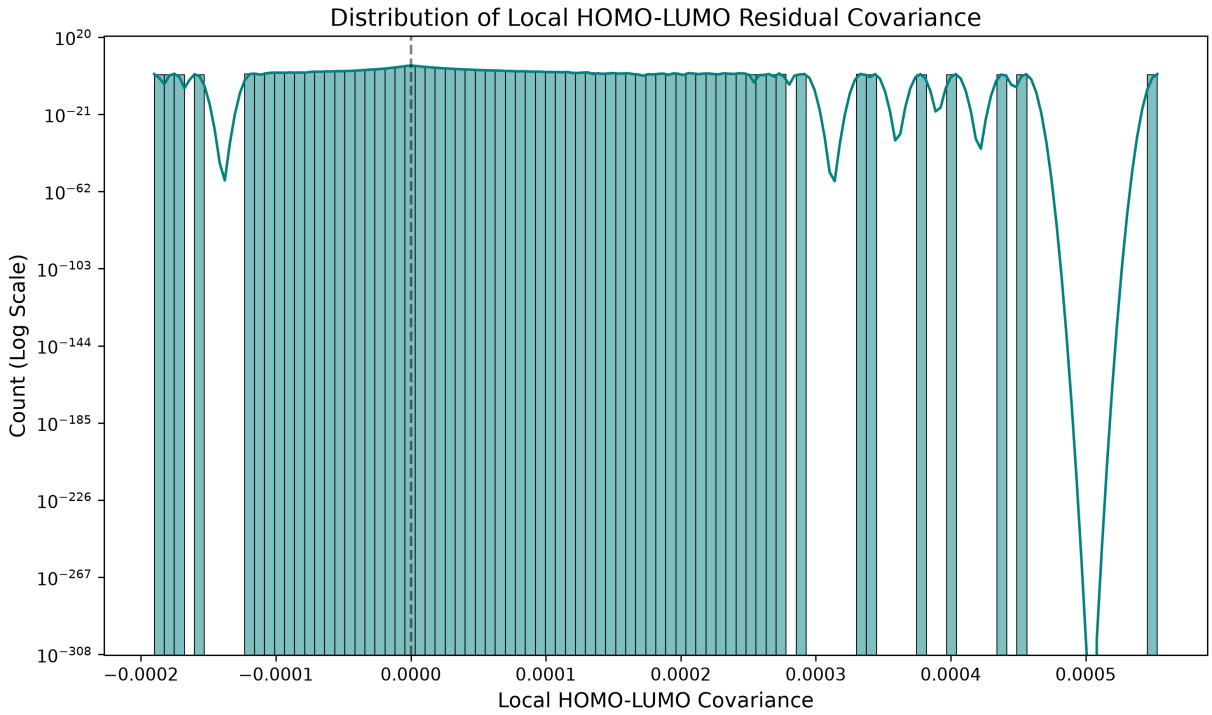

Figure 1: Distribution of local HOMO-LUMO residual covariance across the molecular manifold. The histogram, plotted with a logarithmic y-axis, displays the covariance between res_homo and res_lumo calculated within a 40-nearest neighbor neighborhood for each molecule. Although the global mean local covariance is near zero (2.14×10^{-6} Hartree^2), the distribution features a pronounced heavy tail extending into positive values. This heavy tail demonstrates that high-coupling, co-directional orbital shifts are strongly associated with the introduction of heteroatoms (N, O, F) that re-wire the electronic manifold. Conversely, the regions of low or negative local covariance, visible to the left of the dashed zero line, correspond to pure hydrocarbons that exhibit an anti-correlated, gap-widening regime. Consequently, the plot reveals that the apparent global decoupling of the HOMO is not a uniform property, but rather an averaging effect across these structurally distinct, motif-dependent electronic regimes.

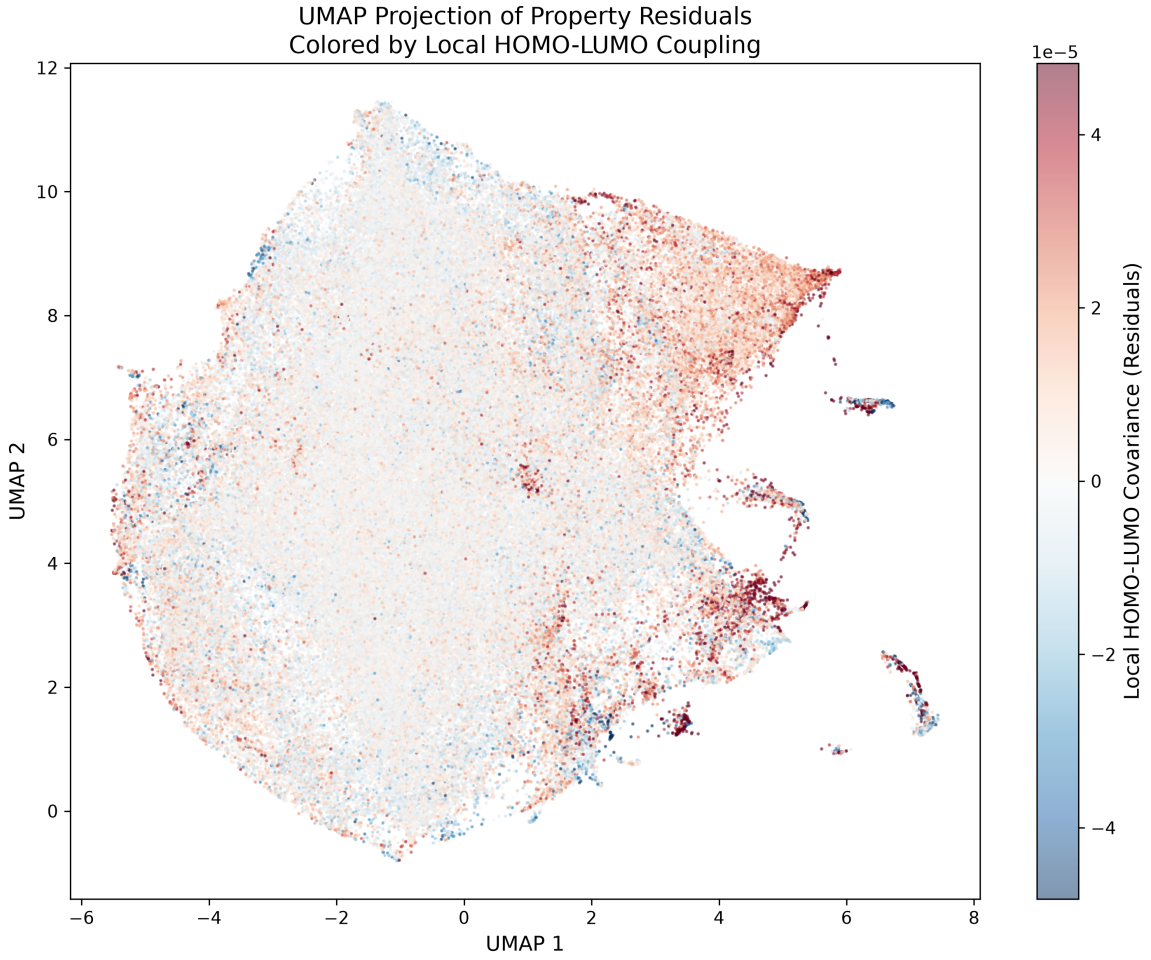

Figure 2: UMAP projection of the 12-dimensional property residual manifold, colored by the local Pearson covariance between HOMO and LUMO residuals. The visualization reveals distinct islands of electronic behavior, highlighting manifold heterogeneity. Regions exhibiting high positive local covariance (red) are strongly associated with heteroatom motifs that induce co-directional orbital shifts. Conversely, regions with low or negative local covariance (blue) correspond to pure hydrocarbons, which follow an anti-directional coupling regime. This demonstrates that the apparent global decoupling of the HOMO is an averaging effect across structurally distinct chemical regimes rather than a uniform molecular property.

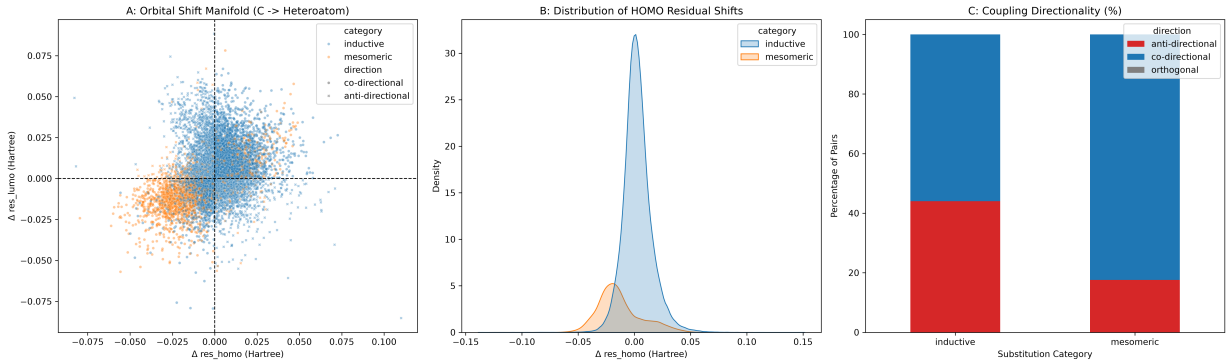

Figure 3: Matched molecular pair perturbation analysis of carbon-to-heteroatom substitutions illustrating the motif-dependent re-wiring of the electronic manifold. Panel A displays the orbital shift manifold of Δres_lumo versus Δres_homo, categorized by inductive and mesomeric effects. Panel B illustrates the density distribution of HOMO residual shifts, demonstrating that mesomeric substitutions induce significantly larger shifts (mean magnitude of 0.0206 Hartree) than purely inductive substitutions (0.0094 Hartree). Panel C presents the coupling directionality, revealing that mesomeric contexts result in predominantly co-directional shifts (82.40%), whereas inductive contexts exhibit a higher proportion of anti-directional shifts (44.03%). Together, these panels demonstrate that heteroatom introduction, particularly in mesomeric contexts, re-wires the electronic manifold by shifting the HOMO-LUMO coupling to a co-directional regime.

directional shifts dropping to 17.60%. This demonstrates that heteroatom introduction, particularly in conjugated systems, forces the HOMO and LUMO to respond in tandem.

3.5 Motif-based validation and the chemical rulebook

The robustness of these patterns was validated across discrete structural holdouts. We partitioned the dataset into pure hydrocarbons ($N = 4,907$), nitrogen-containing molecules ($N = 82,776$), and fluorine-containing molecules ($N = 2,140$).

As shown in Figure 4 and summarized in Table 2.6, the pure hydrocarbon set emerges as a statistical outlier. It exhibits a strong negative partial correlation ($r = -0.4317$) between the HOMO and LUMO residuals when controlling for polarizability, alongside a low mean local covariance (1.00×10^{-6} Hartree^2). In contrast, the introduction of heteroatoms flips this relationship to a positive correlation. Nitrogen-containing molecules show a partial correlation of 0.3001, and fluorine-containing molecules exhibit the strongest co-directional coupling ($r = 0.3873$) and highest mean local covariance (5.85×10^{-6} Hartree^2).

Based on these findings, we define a set of predictive rules governing orbital re-wiring in the QM9 manifold:

- **The Hydrocarbon Rule:** In the absence of heteroatoms, HOMO and LUMO residuals are anti-correlated. Structural changes that stabilize the HOMO tend to destabilize the LUMO, leading to a rigid gap-widening mechanism.
- **The Heteroatom Pinning Rule:** The introduction of nitrogen or oxygen "pins" the electronic manifold, causing the HOMO and LUMO to shift co-directionally (occurring in 82% of mesomeric contexts). This effectively decouples the orbital energies from the spatial extent of the molecule, as both orbitals respond in tandem to the local potential of the heteroatom.
- **The Mesomeric Magnitude Rule:** Mesomeric substitutions exert approximately twice the influence on the HOMO residual compared to purely inductive substitutions, while their effect on the LUMO remains relatively constant.
- **The Fluorine Exception:** Fluorine introduction acts as a powerful local electronic perturbostat, producing the highest mean local covariance and the strongest co-directional coupling, thereby dominating the global residual structure in its vicinity.

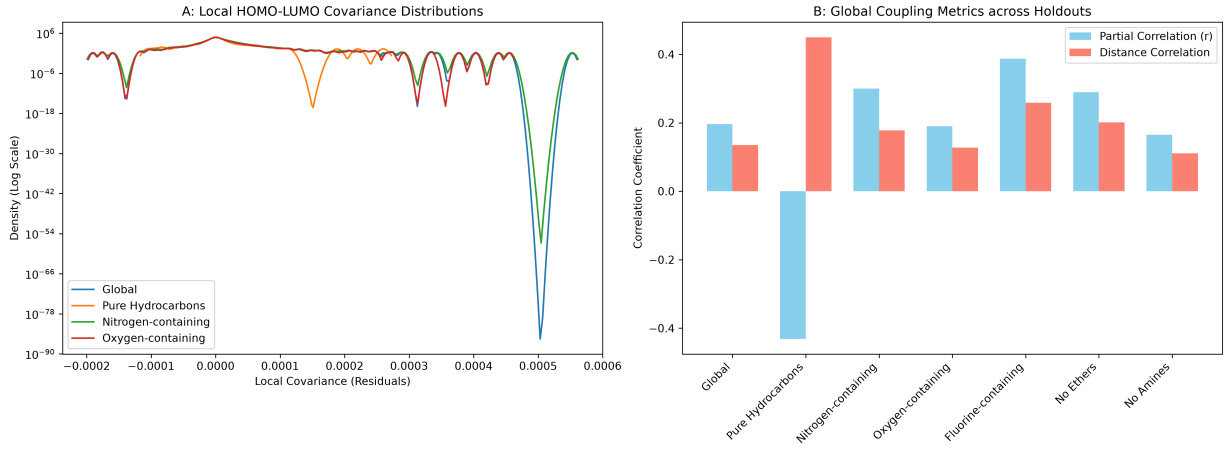

Figure 4: Analysis of HOMO-LUMO coupling regimes across distinct chemical motifs. (A) Density distributions of local HOMO-LUMO covariance for the global dataset and specific structural subsets. The global distribution exhibits a heavy tail driven by heteroatom-containing molecules, while pure hydrocarbons are confined to regions of low or negative local covariance. (B) Global coupling metrics, including partial correlation and distance correlation, across various structural holdouts. Pure hydrocarbons emerge as a statistical outlier, displaying a strong negative partial correlation between HOMO and LUMO residuals. Conversely, the introduction of heteroatoms (nitrogen, oxygen, and fluorine) shifts the relationship to a positive partial correlation. These results demonstrate that the apparent global decoupling of the HOMO is an averaging effect of two distinct behaviors: an anti-correlated regime in pure hydrocarbons and a co-directional, heteroatom-driven regime.

These rules confirm that the global "decoupling" of the HOMO is not a universal constant, but rather an averaging artifact of an anti-correlated regime in hydrocarbons and a co-correlated regime in heteroatom-doped molecules.

4 Conclusions

The persistent decoupling of the highest occupied molecular orbital (HOMO) from global thermodynamic and spatial properties in large chemical datasets has long obscured the underlying electronic behavior of molecules. By performing an exact variance decomposition of the HOMO-LUMO gap on property residuals stripped of stoichiometric determinism, we have demonstrated that this apparent global orthogonality is not a universal chemical constant. Rather, it is a statistical artifact—an averaging of fundamentally distinct, motif-dependent electronic regimes.

Our analysis reveals a stark dichotomy in orbital coupling driven by the presence of heteroatoms. In pure hydrocarbons, the HOMO and LUMO residuals exhibit strong anti-directional coupling, where structural perturbations typically widen the gap. However, the introduction of electronegative atoms—specifically nitrogen, oxygen, and fluorine—fundamentally re-wires this electronic manifold. These heteroatoms pin the local electronic environment, forcing a predominantly co-directional shift where the HOMO and LUMO respond in tandem to the local potential. This tandem response effectively decouples the orbital energies from the macroscopic spatial extent of the molecule. Furthermore, we confirmed that this decoupling is not merely mediated by isotropic polarizability, as significant local HOMO-LUMO covariance persists even when polarizability is partialled out.

Through a matched molecular pair perturbation analysis, we quantified the precise structural contexts of these orbital shifts. We established that mesomeric substitutions in conjugated systems exert significantly larger directional shifts on the HOMO compared to purely inductive substitutions in saturated aliphatic chains, driving the manifold overwhelmingly into a co-directional regime. Fluorine, in particular, emerges as a unique local perturbostat, dominating the residual structure and producing the strongest co-directional

coupling observed in the dataset.

Ultimately, this work provides a rigorous, stoichiometry-independent framework for understanding electronic behavior across vast chemical spaces. By moving beyond global correlation analyses and isolating the local covariance of the residual manifold, we have formulated a concrete set of chemical rules governing orbital re-wiring. Recognizing that global orbital decoupling is merely an averaging effect of localized structural regimes will enable the development of more accurate, motif-aware predictive models and deepen our understanding of the physical organic chemistry that dictates molecular reactivity and stability.