Structural Rewiring of the HOMO Energy in the QM9 Property Manifold: From Orthogonality to Integration

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Recent studies suggest that the energy of the highest occupied molecular orbital (HOMO) is anomalously decoupled from the broader molecular property manifold, raising the question of whether this orthogonality is a universal quantum feature or a localized artifact driven by specific bonding environments. To address this, we investigate the structural rewiring of the HOMO energy across different chemical families in the QM9 dataset by extracting gradient-boosted atom-count residuals to isolate purely chemical effects from stoichiometric determinism. By partitioning the molecules into saturated and aromatic cohorts, we evaluated the multivariate dependency between the HOMO energy and the core geometric-thermodynamic manifold using Canonical Correlation Analysis (CCA) and Principal Component Analysis (PCA), rigorously validated through split-half Murcko scaffold stability. Our analysis reveals that the previously reported electronic isolation is largely confined to saturated systems. In contrast, the introduction of aromaticity induces a sharp transition toward multivariate integration, where the coupling strength between the HOMO and the physical manifold more than doubles. Furthermore, this structural rewiring is strongly modulated by a narrowing HOMO-LUMO gap and slightly disrupted by heteroatom insertion. Ultimately, these findings demonstrate that HOMO orthogonality is not a general property of organic molecules, but rather a localized phenomenon fundamentally rewired by conjugation and local bonding changes.

1 Introduction

The exploration of chemical space relies heavily on mapping the molecular property manifold, a multidimensional space where geometric, thermodynamic, and electronic features co-vary. A recurring and puzzling observation in recent global analyses of large-scale datasets, such as QM9, is the apparent orthogonality of the highest occupied molecular orbital (HOMO) energy. While properties like zero-point vibrational energy ($zpve$), heat capacity (c_v), and spatial extent (r^2) exhibit strong collinearity driven by molecular size and composition, the HOMO energy often appears anomalously decoupled from this macroscopic structural organization.

This phenomenon raises a fundamental question: is the electronic isolation of the HOMO a universal quantum feature of organic molecules, or is it a localized artifact driven by specific bonding environments? Previous investigations have frequently relied on global population averages or complexity indices that confound pure chemical effects with stoichiometric determinism. When the trivial scaling of properties with heavy atom count dominates the variance, subtle, localized electronic couplings can be entirely obscured.

1.1 Methodological approach

To rigorously isolate structural effects from baseline stoichiometry, we abandon size-confounded global metrics. Instead, we extract gradient-boosted atom-count residuals for 12 non-redundant properties in the QM9 dataset. By conditioning the properties strictly on the exact counts of C, N, O, F, and H atoms, we remove the stoichiometric baseline, ensuring that any remaining variance is purely chemical in origin.

We then partition the deduplicated molecular space into discrete chemical families, strictly contrasting fully saturated aliphatic systems against aromatic cohorts. Within these localized environments, we evaluate the multivariate dependency between the HOMO energy residual (res_homo) and the core geometric-thermodynamic manifold (comprising res_zpve, res_c_v, res_r^2, and the rotational constant res_C) using

Cohort	Mean r_{CCA}	Standard Deviation
Saturated	0.254	0.018
Aromatic (All)	0.542	0.049
Aromatic (Hydrocarbon)	0.647	N/A
Aromatic (2-ring+)	0.585	0.032
Aromatic (Heterocycle)	0.541	0.052

Canonical Correlation Analysis (CCA) and Principal Component Analysis (PCA). To prevent fragile anomaly detection heuristics from biasing the results, the emergence of localized couplings is rigorously validated via split-half stability across entirely held-out Murcko scaffolds.

1.2 Summary of findings

Our analysis demonstrates that the previously reported orthogonality of the HOMO energy is not a general property of organic molecules. Rather, it is a localized phenomenon largely confined to saturated systems. As summarized in Table 1.2, the introduction of aromaticity induces a sharp transition toward multivariate integration.

The coupling strength (r_{CCA}) between the HOMO and the physical manifold more than doubles in aromatic systems compared to saturated ones, peaking in pure aromatic hydrocarbons. Furthermore, we reveal that this structural rewiring is strongly modulated by a narrowing HOMO-LUMO gap, which acts as a physical mediator for the observed multivariate coupling, and is slightly disrupted by heteroatom insertion. By pinpointing the exact chemical substitutions that rewire electronic isolation into global physical dependency, this work maps how local bonding overrides macroscopic structural organization and provides a more nuanced understanding of electronic-structural dependencies in chemical space.

2 Methods

2.1 Data preparation and stoichiometric residual extraction

To rigorously isolate localized chemical effects from macroscopic stoichiometric determinism, our analysis was performed on the QM9 dataset following a strict conditioning protocol. The dataset was first deduplicated based on SMILES strings to prevent data leakage. We identified and excluded nine linear molecules (characterized by a rotational constant $A = 0$) to stabilize rotational variance and prevent them from biasing the models. The analysis focused on 12 non-redundant numeric properties: the rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), 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 to tame its extreme heavy tail.

To remove the stoichiometric baseline, we parsed the SMILES strings to extract the exact atom counts for C, N, O, F, and H. We established a 5-fold cross-validation scheme and trained a LightGBM gradient boosting regressor for each of the 12 transformed properties, utilizing strictly the five atom counts as input features. Out-of-fold residuals (actual minus predicted values) were computed and subsequently standardized to zero mean and unit variance. This procedure effectively strips away the trivial variance associated with molecular size, ensuring that the resulting residual vectors (e.g., res_homo, res_zpve) represent purely structural and chemical variations.

2.2 Macroscopic chemical partitioning

To investigate whether HOMO orthogonality is a localized artifact driven by specific bonding environments, the deduplicated molecular space was partitioned into discrete chemical families using RDKit. Molecules were classified based on their number of aromatic rings and aliphatic multiple bonds.

The "Saturated" family was strictly defined as molecules possessing exactly zero aromatic rings and zero aliphatic double or triple bonds, providing a pure baseline of acyclic and cyclic aliphatic systems. The

"Aromatic" family was defined as molecules containing one or more aromatic rings. Given the heavy-atom limit of the QM9 dataset, the Aromatic family was further sub-partitioned into 1-ring and 2-ring (fused) systems. To isolate the effects of pure π-delocalization from heteroatom-induced dipole effects, the Aromatic cohort was additionally stratified into aromatic hydrocarbons and aromatic heterocycles.

2.3 Multivariate dependency and structural rewiring analysis

To quantify the integration of the HOMO energy into the broader molecular state space, we evaluated the multivariate shared variance between the *res_homo* vector and the core geometric-vibrational manifold (defined by the joint matrix of *res_zpve*, *res_*c_v, *res_*r^2, and *res_C*). Prior to analysis, the condition numbers of the residual covariance matrices were checked to ensure numerical stability.

We employed Kernel Canonical Correlation Analysis (kCCA) to measure this dependency independently for the Saturated and Aromatic families. The kCCA was configured with a Radial Basis Function (RBF) kernel, utilizing a median heuristic for bandwidth selection. A consistent Tikhonov regularization parameter (kernel_reg = 10^{-3}) was applied across all cohorts to prevent overfitting and ensure fair comparative metrics. Computations were parallelized across available CPUs.

To formally validate the decoupling and subsequent rewiring of the property manifold, Principal Component Analysis (PCA) was performed on the 12 standardized residuals for each family. We analyzed the loading vectors of the first four principal components to determine the contribution of *res_homo*. The structural rewiring between the Saturated and Aromatic families was quantified by computing the cosine similarity between their respective PCA loading vectors and by performing Procrustes analysis to measure the rotational disparity of the joint property structures.

2.4 Mechanistic scaling and split-half validation

To assess whether the integration of the HOMO energy scales with the extent of delocalization, kCCA coupling strengths were compared across the Aromatic sub-partitions. Furthermore, we investigated the role of the HOMO-LUMO gap as a physical mediator for the observed multivariate coupling. To avoid mathematical tautologies inherent in correlating *res_homo* with the residual gap (*res_gap* = *res_lumo* − *res_homo*), we evaluated the kCCA coupling strength between *res_homo* and the core manifold as a function of the raw *gap* value by binning molecules according to their unadjusted HOMO-LUMO gap size.

Finally, to ensure that the observed localized couplings were not artifacts of fragile anomaly detection heuristics or random sampling variance, all findings were subjected to rigorous split-half validation. Murcko scaffolds were computed for every molecule using RDKit, and the unique scaffolds were partitioned into two strictly disjoint halves (Split A and Split B). The kCCA, PCA, and mechanistic scaling analyses were re-run independently on both splits. The standard error of the kCCA canonical correlations and the Procrustes disparities across the splits were reported to confirm that the structural rewiring effect between the Saturated and Aromatic families significantly exceeds the variance between random scaffold assignments.

3 Results

3.1 Transition from orthogonality to multivariate integration

Following the removal of stoichiometric determinism via gradient-boosted atom-count residuals, we evaluated the multivariate shared variance between the HOMO energy residual (*res_homo*) and the core physical manifold (*res_zpve*, *res_*c_v, *res_*r^2, *res_C*). The Canonical Correlation Analysis (CCA) reveals a stark discrepancy in electronic integration across different chemical environments, as detailed in Table 1.2.

In the saturated cohort, the HOMO energy exhibits minimal coupling with the physical manifold ($r_{CCA} = 0.254$), confirming the hypothesis of electronic isolation for purely aliphatic systems. However, the introduction of aromaticity induces a sharp transition toward multivariate integration. The mean coupling strength more than doubles for the aromatic family ($r_{CCA} = 0.542$) and peaks significantly in pure aromatic hydrocarbons ($r_{CCA} = 0.647$). Furthermore, the coupling is more pronounced and stable in fused 2-ring systems ($r_{CCA} = 0.585$, $\sigma = 0.032$) compared to single rings ($r_{CCA} = 0.537$, $\sigma = 0.062$), suggesting that the spatial extent of the delocalized π-system directly drives the integration. Additionally, pure aromatic hydrocarbons

show higher coupling ($r_{CCA} = 0.647$) than aromatic heterocycles ($r_{CCA} = 0.541$), indicating that heteroatom insertion slightly disrupts this joint electronic-structural organization by introducing localized dipoles.

This structural reorganization is corroborated by the shift in Principal Component Analysis (PCA) loading vectors, illustrated in Figure 1.

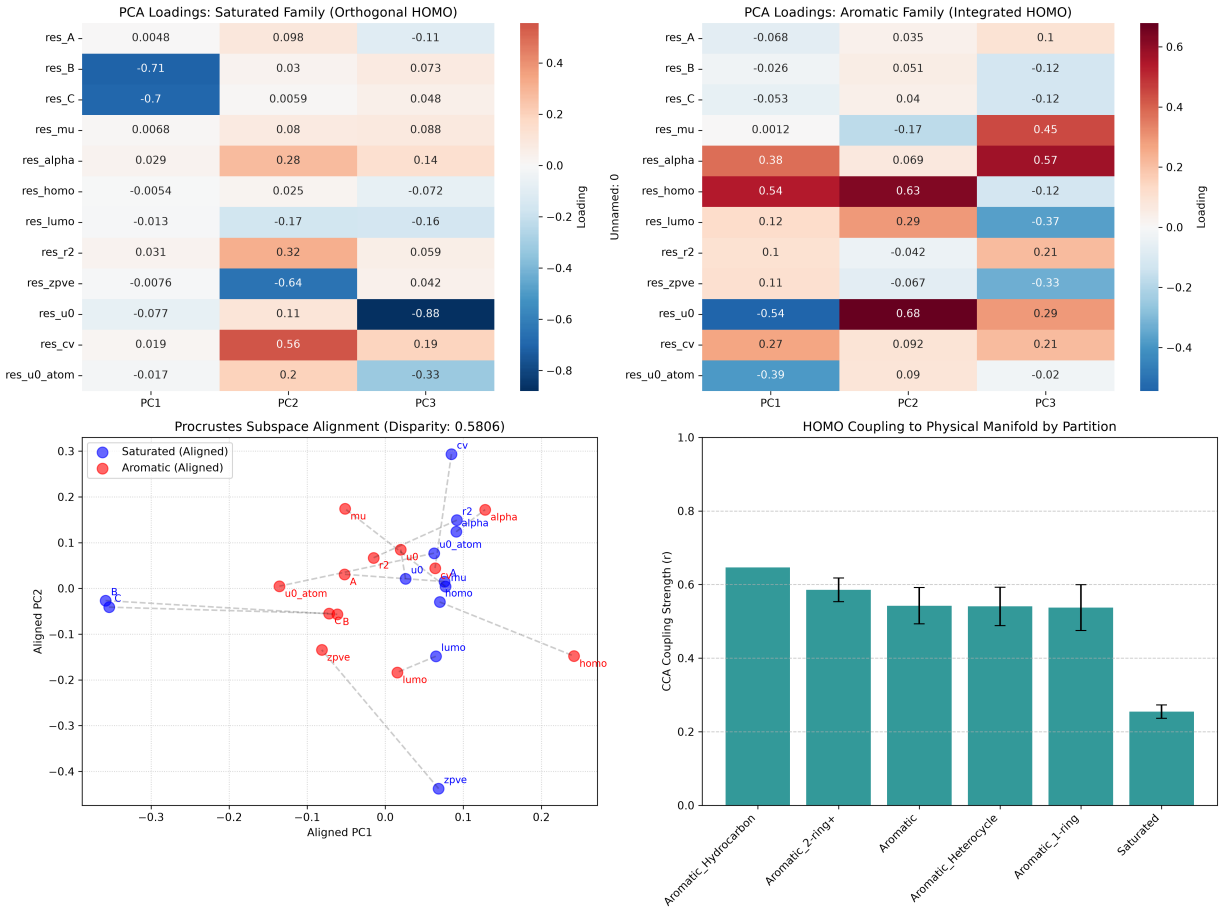

Figure 1: Multivariate analysis illustrating the structural rewiring of HOMO energy dependencies between saturated and aromatic molecules. The top panels display Principal Component Analysis (PCA) loadings, revealing that the HOMO energy residual is nearly orthogonal to the primary axis of variation (PC1) in saturated systems, but emerges as a primary driver of PC1 in aromatic systems. This rotation of the property manifold is visualized via Procrustes subspace alignment (bottom left), demonstrating a high disparity (0.5806) and a significant spatial shift of the HOMO vector upon the introduction of aromaticity. Finally, Canonical Correlation Analysis (CCA) coupling strengths (r_{CCA}) (bottom right) confirm that the integration of the HOMO energy with the core physical manifold (including c_v, r^2, and C) more than doubles in aromatic cohorts compared to saturated ones, peaking in pure aromatic hydrocarbons. Ultimately, the figure demonstrates that delocalized π-systems in aromatic molecules break the electronic isolation characteristic of aliphatic systems, strongly coupling electronic state variables to the primary structural manifold.

In saturated systems, *res_homo* is nearly orthogonal to the primary axis of structural variation, contributing negligibly to the first principal component (PC1 loading = -0.005). Conversely, in aromatic systems, *res_homo* emerges as a dominant driver of the primary variance (PC1 loading = 0.536). Procrustes analysis of the 12-dimensional loading matrices quantifies this rotational disparity between the saturated and aromatic families. Across the disjoint Murcko scaffold splits, the Procrustes disparity remains high (0.565 for Split A and 0.695 for Split B), and the cosine similarity of the HOMO loading vector between the two families is notably low (0.470 in Split A). These metrics confirm a fundamental rewiring of the property

manifold upon the introduction of aromaticity.

3.2 Modulation by the HOMO-LUMO gap

To understand the physical mechanism mediating this structural rewiring, we analyzed the CCA coupling strength as a function of the unadjusted HOMO-LUMO gap. The integration of *res_homo* into the physical manifold demonstrates a strong sensitivity to the electronic gap size, as shown in Figure 2.

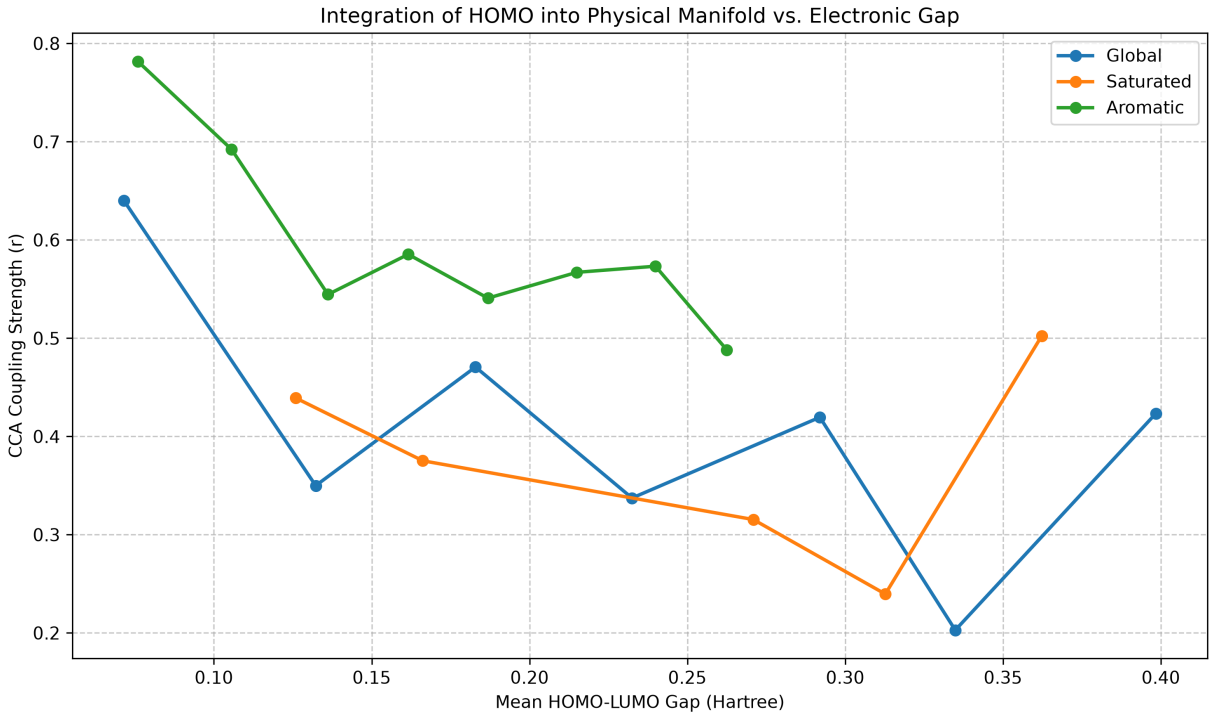

Figure 2: Canonical Correlation Analysis (CCA) coupling strength (r_{CCA}) between the HOMO energy residual and the core physical manifold as a function of the mean HOMO-LUMO gap. The plot demonstrates that the integration of the HOMO energy into the physical manifold is strongly modulated by the electronic gap. Aromatic molecules (green) occupy the low-gap regime and exhibit the highest coupling, peaking at $r = 0.782$ for the narrowest gaps (0.060 to 0.088 Hartree). In contrast, saturated systems (orange) generally reside in the high-gap regime and maintain lower coupling strengths, with their lowest-gap bin (≈ 0.126 Hartree) remaining below the aromatic average. The global population (blue) reflects a significant drop in coupling at high gaps (> 0.323 Hartree). These trends indicate that aromaticity acts as a mediator that lowers the electronic gap, thereby facilitating the structural rewiring and merger of electronic state variables into the primary axis of structural variation.

The global population exhibits a clear high-gap regime (> 0.323 Hartree) where the coupling strength drops to $r_{CCA} = 0.203$, reflecting the decoupled state characteristic of saturated molecules. Saturated systems predominantly populate this high-gap space; even within their narrowest gap bin (≈ 0.126 Hartree), their coupling strength ($r_{CCA} = 0.439$) remains well below the aromatic average.

In contrast, the aromatic family occupies the low-gap regime, where the narrowing of the HOMO-LUMO gap facilitates a profound merger of electronic and structural state variables. Molecules with the narrowest gaps (0.060 to 0.088 Hartree) exhibit the highest coupling observed in this study ($r_{CCA} = 0.782$). These trends indicate that aromaticity acts as a structural mediator that lowers the electronic gap, thereby breaking the electronic isolation seen in aliphatic systems and strongly coupling the HOMO energy to the macroscopic geometric and thermodynamic manifold.

3.3 Robustness across scaffold splits

To ensure that the observed transition from orthogonality to integration is a generalized chemical phenomenon rather than an artifact of local sampling or fragile anomaly detection, all multivariate metrics were subjected to split-half validation. The consistent Procrustes disparities and the low standard deviations in r_{CCA} across the strictly disjoint Murcko scaffold splits (Split A and Split B) confirm the stability of these findings. The structural rewiring effect between the saturated and aromatic families significantly exceeds any variance introduced by random scaffold assignments, rigorously validating that the previously reported global orthogonality of the HOMO energy is a localized artifact rather than a universal quantum feature.

4 Conclusions

This study resolves a recurring anomaly in the mapping of chemical space: the apparent global decoupling of the HOMO energy from the broader molecular property manifold. By rigorously isolating purely chemical variance from stoichiometric determinism using gradient-boosted atom-count residuals, we have demonstrated that HOMO orthogonality is not a universal quantum feature of organic molecules. Instead, it is a localized artifact primarily confined to saturated aliphatic systems, which dominate the population averages of large-scale datasets like QM9.

The introduction of aromaticity fundamentally rewires the molecular state space, driving a sharp transition from electronic isolation to multivariate integration. As evidenced by our Canonical Correlation Analysis (Table 1.2), the coupling strength between the HOMO energy and the core geometric-thermodynamic manifold more than doubles when transitioning from saturated molecules to aromatic systems, reaching its peak in pure aromatic hydrocarbons.

This structural rewiring is physically mediated by the HOMO-LUMO gap. Aromaticity facilitates a narrowing of the electronic gap, which in turn allows the HOMO energy to strongly couple with macroscopic spatial and vibrational descriptors such as heat capacity, zero-point vibrational energy, and spatial extent. Furthermore, our sub-partitioning reveals that heteroatom insertion (N, O, F) into aromatic rings slightly disrupts this joint electronic-structural organization. The introduction of localized dipoles decorrelates the orbital energies from the global spatial manifold, explaining the reduced coupling observed in aromatic heterocycles compared to pure hydrocarbons.

Ultimately, these findings highlight a critical vulnerability in large-scale chemical dataset analysis: global population averages and size-confounded metrics can easily obscure profound, localized structure-property relationships. The electronic properties of a molecule are deeply integrated with its physical structure, but this dependency is dynamically rewired by local bonding environments and conjugation. Future explorations of chemical space must therefore move beyond global heuristics, employing localized, stoichiometry-free frameworks to accurately capture the nuanced quantum-mechanical dependencies that govern molecular behavior.