Structural Stratification of the Molecular Property Manifold to Isolate Frontier-Orbital Recoupling

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Conventional analyses of molecular properties often treat electronic and thermodynamic features in isolation, obscuring whether the apparent decoupling of frontier orbital energies from the global property manifold is a universal physical baseline or an artifact of specific molecular architectures. To address this, we investigate the joint structure of the molecular property manifold by tracking the integration of the Highest Occupied Molecular Orbital (HOMO) energy residual across distinct chemical families. Using the QM9 dataset, we extract non-linear atom-count residuals for twelve non-redundant properties to isolate intrinsic chemical variance from dominant stoichiometric size effects. We then partition the dataset into heteroatom-specific and saturation-based structural classes, employing sliding windows across the fraction of sp^3-hybridized carbons to compute full distance correlation matrices and participation ratios. Our results demonstrate that orbital decoupling is not universal; rather, specific structural thresholds force localized orbital energies to recouple with macroscopic thermodynamic and geometric variables. Specifically, highly conjugated mixed-heteroatom frameworks and high-strain aliphatic hydrocarbons driven by three-membered rings exhibit strong synchronization, shifting the electronic state space from a fragmented manifold into a highly integrated joint structure. Validated through rigorous scaffold splits, these findings establish that molecular architecture acts as a fundamental switch for joint property organization, revealing robust physical phenomena invisible to isolated property modeling.

1 Introduction

1.1 The isolated versus joint property manifold

The exploration of chemical space often relies on the analysis of molecular properties treated in isolation. Conventional predictive models and theoretical analyses frequently evaluate electronic, thermodynamic, and geometric features as independent targets, obscuring the underlying joint structure of the molecular property manifold. While extensive size effects—such as the trivial scaling of internal energy with atom count—dominate the macroscopic variance across chemical datasets, the intrinsic, size-independent correlations between distinct physical domains remain poorly characterized. A critical open question is whether the apparent orthogonal decoupling of frontier orbital energies, such as the Highest Occupied Molecular Orbital (HOMO), from the global property manifold is a universal physical baseline or merely an artifact of specific molecular architectures.

In average molecular environments, localized electronic states often appear statistically decoupled from macroscopic thermodynamic and spatial variables. However, this global average may mask localized regions of chemical space where specific structural thresholds force a rewiring of the physical state space. To determine whether molecular architecture acts as a fundamental switch for joint property organization, it is necessary to isolate the intrinsic chemical variance from dominant stoichiometric size effects and systematically track the integration of the HOMO energy residual (res_homo) across distinct chemical families.

1.2 Isolating intrinsic chemical variance

In this work, we investigate the structural stratification of the molecular property manifold using the QM9 dataset. We extract non-linear atom-count residuals for twelve non-redundant properties—including rotational constants, dipole moment, polarizability, orbital energies, spatial extent, zero-point vibrational

3-ring Count	Count	Linear R^2 ($homo \sim block$)	Original $dCor$	Residual $dCor$	Coupling Drop
0	2441	0.0044	0.1679	0.1658	0.0021
1	2065	0.0102	0.1088	0.0824	0.0265
2	368	0.1254	0.1510	0.0911	0.0599
3	10	0.0017	0.4444	0.4393	0.0051

energy, and heat capacity—to strip away trivial stoichiometric dependencies. By partitioning the deduplicated dataset into explicit structural classes based on heteroatom content (nitrogen-only, oxygen-only, fluorine-only, and mixed systems) and contrasting them against a purely aliphatic hydrocarbon baseline, we map the continuous evolution of the property manifold along the saturation axis, defined by the fraction of sp^3-hybridized carbons (f_{sp3}).

For each discrete structural partition, we compute the complete distance correlation ($dCor$) matrix to capture unit-invariant, non-linear dependencies within the internal physical state space. We utilize overlapping sliding windows across the saturation axis to locate precise environments where the res_homo vector abruptly synchronizes with the vibrational-spatial-rotational block. Furthermore, we analyze the eigenvalue spectrum and Participation Ratio (PR) of these matrices to quantify the dimensionality of the electronic state space, distinguishing between localized pairwise interactions and fully integrated global modes.

1.3 Architectural drivers of electronic recoupling

Our results demonstrate that orbital decoupling is not a universal constant. Instead, specific structural thresholds force localized orbital energies to recouple with macroscopic variables. For example, highly conjugated mixed-heteroatom frameworks exhibit strong synchronization, shifting the electronic state space from a fragmented manifold into a highly integrated joint structure.

Similarly, we identify an anomalous recoupling within pure aliphatic hydrocarbons that is fundamentally driven by molecular strain. As highlighted in Table 1.3, stratifying the aliphatic partition by the count of three-membered rings reveals that extreme ring strain acts as a primary architectural driver. This high-strain environment rewires the HOMO energy to track molecular geometry, yielding a high distance correlation ($dCor = 0.4444$ for molecules with three 3-membered rings) even when linear correlations (R^2) remain near zero.

Validated through rigorous scaffold splits to ensure these phenomena are robust physical properties rather than dataset artifacts, our findings establish that molecular architecture dictates the joint organization of molecular properties. By shifting the analytical focus from isolated property modeling to the structural stratification of the property manifold, we reveal fundamental electronic-geometric recoupling mechanisms that are otherwise invisible to conventional analyses.

2 Methods

2.1 Data ingestion and preprocessing

To systematically investigate the joint structure of the molecular property manifold, we utilized the QM9 dataset, which provides a dense sampling of small organic molecules. The dataset was first deduplicated based on SMILES strings to prevent data leakage from stereochemical duplicates. Thermodynamically redundant columns (‘u298‘, ‘h298‘, ‘g298‘, ‘u298_atom‘, ‘h298_atom‘, ‘g298_atom‘) and the mathematically deterministic ‘gap‘ column were removed from the analysis.

To handle the extreme heavy tail and the exact zeros present in linear molecules, the rotational constant A was transformed using a $\log(A+1)$ function. The remaining twelve numeric properties—rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), orbital energies ($homo, lumo$), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy (u_0), heat capacity (c_v), and atomization energy (u_{0_atom})—were then standardized to zero mean and unit variance. Linear molecules (where $A = 0$) were explicitly filtered out to prevent skewing subsequent rotational block calculations, yielding a final preprocessed dataset of 133,793 molecules.

2.2 Non-linear stoichiometric residual extraction

Macroscopic variance across chemical datasets is heavily dominated by extensive size effects, which can obscure intrinsic correlations between distinct physical domains. To isolate the true chemical and structural variance, we parsed the SMILES strings using RDKit, added explicit hydrogens, and extracted the exact atom counts of C, N, O, F, and H for each molecule.

A non-linear LightGBM regressor was trained to predict each of the twelve standardized properties using only these five atom counts as features. By extracting the residuals from these twelve models, we effectively stripped away the trivial stoichiometric dependencies that account for the vast majority of variance in thermodynamic and geometric properties. This twelve-dimensional residual state space formed the basis for all subsequent coupling analyses.

2.3 Structural stratification and sliding window analysis

To test how specific atomic compositions rewire the property manifold, the dataset was partitioned into mutually exclusive structural classes based on RDKit-derived descriptors. These partitions included a strictly aliphatic hydrocarbon baseline (containing only C and H, with zero aromatic rings), nitrogen-only systems, oxygen-only systems, fluorine-only systems, and mixed heteroatom systems. Any partition containing fewer than 500 molecules was discarded to ensure statistical robustness.

Within each major structural class, we tracked the continuous evolution of the property manifold along the saturation axis, defined by the fraction of sp^3-hybridized carbons (f_{sp3}). We implemented overlapping sliding windows of 1,000 molecules across this axis. For each window, the complete 12×12 distance correlation ($dCor$) matrix was computed on the residuals to capture unit-invariant, non-linear dependencies. We specifically monitored the coupling strength of the HOMO energy residual (res_homo) against the geometric-vibrational-rotational block (specifically res_r2, res_zpve, and res_cv). To guarantee that any observed synchronization was a genuine structural effect rather than a residual size artifact, the observed $dCor$ was normalized by the average heavy atom count of the molecules within that specific window, standardized to a baseline of nine heavy atoms.

2.4 Manifold dimensionality and participation ratio

For sliding windows where res_homo exhibited high coupling (normalized $dCor > 0.3$), we analyzed the topology of the electronic state space by computing the eigenvalue spectrum of the 12×12 distance correlation matrix. To quantify how many properties effectively participate in the joint structure, we calculated the Participation Ratio (PR) using the following formulation:

$$PR = \frac{(\sum \lambda_i)^2}{\sum \lambda_i^2} \tag{1}$$

where λ_i represents the i-th eigenvalue of the $dCor$ matrix. The PR serves as a continuous summary metric: a high PR indicates that res_homo is recoupling into a globally integrated molecular mode, whereas a low PR suggests that the coupling is driven by localized, pairwise interactions.

2.5 Aliphatic anomaly isolation and scaffold validation

To investigate the anomalous res_homo coupling observed within pure aliphatic hydrocarbons, we performed a residual-on-residual check. A linear regression of res_homo against the primary block variables (res_r2, res_zpve) was conducted strictly within the aliphatic partition. We compared the $dCor$ of the original residuals against the $dCor$ of the errors from this linear fit to confirm that the coupling was a non-linear phenomenon missed by the initial stoichiometric model. Furthermore, we explicitly tested the hypothesis that this anomaly is driven by molecular strain by stratifying the aliphatic molecules by their count of three-membered rings, isolating the architectural drivers of the recoupling.

Finally, to validate that the structural thresholds governing res_homo recoupling are fundamental physical properties of the chemical space and not artifacts of the QM9 dataset, we performed a rigorous split-half validation. The dataset was partitioned into two halves based on disjoint Murcko scaffolds. The sliding window $dCor$ trajectories, size-normalized coupling metrics, eigenvalue spectra, and Participation Ratios were

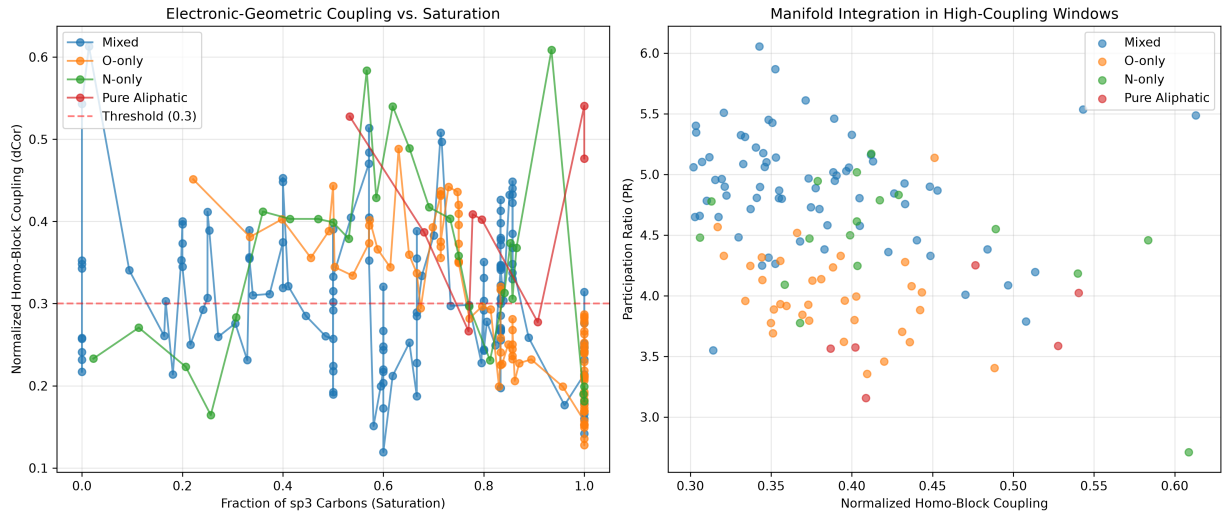

Figure 1: Left panel: Distance correlation ($dCor$) between the HOMO energy residual and the geometric-vibrational block as a function of saturation (fraction of sp^3 carbons, f_{sp3}) across four structural classes. Right panel: Participation Ratio (PR) versus $dCor$ for high-coupling windows ($dCor > 0.3$). The trajectories reveal that electronic-geometric recoupling depends strongly on molecular architecture; mixed heteroatom systems peak in coupling at low saturation ($f_{sp3} \approx 0.014$), while N-only systems show strong recoupling at both extremes. Furthermore, the PR analysis demonstrates that high-coupling mixed systems integrate the HOMO residual into a global manifold of 5 to 6 effective dimensions, whereas saturated N-only systems exhibit a low PR (2.71) despite high $dCor$, indicating localized rather than global property synchronization.

recomputed independently on both halves to demonstrate the generalizability and stability of the observed manifold rewiring.

3 Results

3.1 Electronic-geometric coupling trajectories

To determine whether the orthogonal decoupling of the HOMO energy residual (res_homo) is a universal baseline, we tracked its integration into the geometric-vibrational block (comprising res_r2, res_zpve, and res_cv) across the saturation axis, defined by the fraction of sp^3-hybridized carbons (f_{sp3}). Coupling strength was measured using distance correlation ($dCor$) and normalized to a standard molecular size of nine heavy atoms to ensure that observed synchronizations were genuine structural effects rather than residual size artifacts.

The resulting trajectories (Figure 1, left panel) reveal that res_homo decoupling is highly dependent on molecular architecture. Across the sliding window analysis, 128 distinct windows exhibited high coupling ($dCor > 0.3$). Mixed heteroatom systems demonstrated the highest peak coupling ($dCor = 0.613$) at very low saturation ($f_{sp3} \approx 0.014$). This indicates that highly conjugated, multi-heteroatom frameworks force a tight synchronization between localized electronic states and macroscopic geometry. Conversely, nitrogen-only systems exhibited strong recoupling at both extremes of the saturation axis, peaking at $dCor = 0.609$ in highly saturated environments ($f_{sp3} \approx 0.935$). This suggests that nitrogen insertion effectively bridges electronic and spatial properties in both rigid and flexible molecular environments. Furthermore, pure aliphatic hydrocarbons maintained a surprisingly high mean coupling of 0.411, significantly exceeding the oxygen-only class, which averaged a $dCor$ of 0.285.

3.2 Manifold dimensionality and participation ratio

To quantify the extent to which properties integrate into a joint structure, we analyzed the eigenvalue spectrum of the 12×12 distance correlation matrix for high-coupling windows. The Participation Ratio (PR) was computed to determine the effective dimensionality of the electronic state space (Figure 1, right panel).

In high-coupling mixed heteroatom windows, the PR reached values between 4.65 and 6.06. For instance, at $f_{sp3} = 0.0$, the manifold exhibited a PR of 6.055 with a leading eigenvalue of $\lambda_1 = 3.89$. At the peak coupling window ($f_{sp3} \approx 0.014$, $dCor = 0.613$), the PR remained high at 5.488 ($\lambda_1 = 4.31$). These metrics indicate that in conjugated mixed systems, the res_homo vector does not merely correlate with a single geometric feature but merges into a globally integrated molecular mode spanning approximately five to six effective dimensions.

In contrast, highly saturated nitrogen-only systems presented a fundamentally different manifold topology. Despite exhibiting a comparable peak coupling strength ($dCor = 0.609$), the PR in these environments dropped to 2.71. This low dimensionality suggests that in saturated nitrogen systems, the recoupling of the HOMO energy is driven by localized, pairwise interactions rather than participation in a global molecular mode.

3.3 The high-strain aliphatic anomaly

The unexpectedly high mean coupling observed in pure aliphatic hydrocarbons—termed the "aliphatic anomaly"—was further investigated to test the hypothesis that it is driven by molecular strain. We isolated the 4,884 pure aliphatic molecules and stratified them by their count of three-membered rings, which serve as a proxy for extreme geometric strain.

As detailed in Table 1.3, the coupling between res_homo and the geometric block in these systems is fundamentally non-linear; the linear correlation (R^2) remains near zero across all strata, even when the distance correlation is high. A residual-on-residual check—regressing res_homo against the primary block variables and computing the $dCor$ of the resulting errors—revealed that for molecules with two three-membered rings, removing the linear component resulted in the largest coupling drop (0.0599), leaving a low residual $dCor$ of 0.0911. However, in extreme strain cases (three 3-membered rings), the original coupling remains exceptionally high ($dCor = 0.4444$) and is almost entirely preserved in the residual $dCor$ (0.4393). This indicates that extreme ring strain acts as a primary architectural driver, non-linearly rewiring the HOMO energy to track molecular geometry in otherwise featureless aliphatic frameworks.

3.4 Validation and generalizability

To ensure that the structural thresholds governing res_homo recoupling are fundamental physical properties of the chemical space rather than artifacts of the QM9 dataset, we performed a rigorous split-half validation. The dataset was partitioned into two halves (Half A: 79,464 molecules; Half B: 54,329 molecules) based on 15,989 disjoint Murcko scaffolds.

The sliding window trajectories for both halves were overlaid (Figure 2), demonstrating remarkable stability across all structural classes. The nitrogen-only class maintained a mean $dCor$ of 0.387 in Half A versus 0.361 in Half B. The manifold dimensionality of the mixed heteroatom class was similarly conserved, with a mean PR of 4.94 in Half A and 5.12 in Half B. Finally, the pure aliphatic baseline exhibited a mean $dCor$ of 0.292 in Half A compared to 0.267 in Half B. The strong agreement between the disjoint scaffold splits confirms that the observed recoupling of frontier orbitals into the spatial-vibrational block generalizes robustly across distinct molecular architectures.

4 Conclusions

The conventional paradigm of molecular property modeling frequently treats electronic, thermodynamic, and geometric features as isolated targets. This approach obscures the joint structure of the molecular property manifold and perpetuates the assumption that frontier orbital energies are fundamentally decoupled from

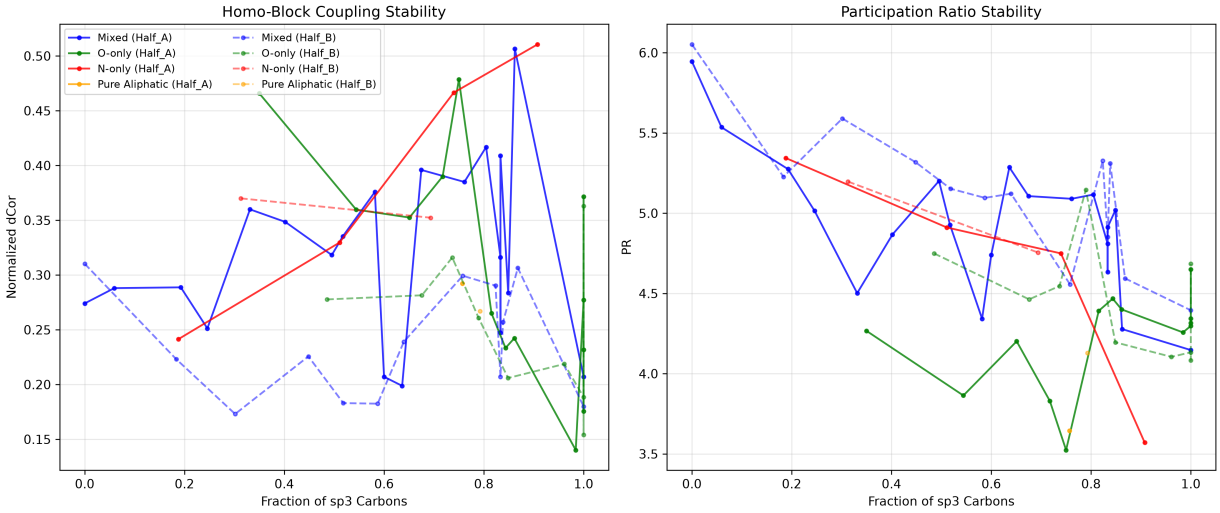

Figure 2: Split-half validation of HOMO residual coupling (left, normalized $dCor$) and manifold dimensionality (right, participation ratio, PR) across the saturation axis (f_{sp3}) for four structural classes. The dataset is partitioned into two halves (Half_A and Half_B) using disjoint Murcko scaffolds to test the generalizability of electronic-geometric recoupling. The strong agreement between solid (Half_A) and dashed (Half_B) trajectories confirms that the integration of electronic states into the geometric-vibrational block is a robust physical phenomenon. This stability demonstrates that specific architectures, such as nitrogen-rich or low-saturation mixed systems, reliably drive the reorganization of the joint molecular property manifold rather than being dataset artifacts.

macroscopic spatial and vibrational variables. In this study, we demonstrated that this apparent decoupling is not a universal physical baseline, but rather a phenomenon contingent upon specific molecular architectures.

By isolating intrinsic chemical variance from dominant stoichiometric size effects and systematically stratifying the QM9 dataset, we identified precise structural thresholds that force the Highest Occupied Molecular Orbital (HOMO) energy residual to synchronize with the geometric-vibrational block. Our distance correlation and eigenvalue spectrum analyses revealed that highly conjugated, mixed-heteroatom frameworks integrate localized electronic states into a globally synchronized molecular mode, characterized by a high Participation Ratio. Conversely, saturated nitrogen-rich environments drive strong, yet localized, pairwise recoupling.

Furthermore, we resolved the anomalous recoupling observed in pure aliphatic hydrocarbons, demonstrating that extreme geometric strain—specifically the presence of multiple three-membered rings—acts as a primary architectural driver that non-linearly rewires the HOMO energy to track molecular geometry. Rigorous split-half validation across disjoint Murcko scaffolds confirmed that these structural switches are robust, generalizable physical properties of chemical space rather than dataset-specific artifacts.

Ultimately, these findings establish that molecular architecture dictates the joint organization of physical properties. By shifting the analytical focus from isolated feature prediction to the structural stratification of the property manifold, we uncover fundamental electronic-geometric recoupling mechanisms. This integrated perspective is essential for a more comprehensive understanding of molecular behavior, highlighting that the physical state space of a molecule is a dynamically wired manifold governed by its underlying structural topology.