Topological Phase Transitions of Orbital Decoupling: Mapping HOMO-LUMO Orthogonality Across Structural Families

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

While the macroscopic physical properties of molecules are largely dictated by their stoichiometry, frontier electronic orbitals frequently decouple from this global thermodynamic-geometric manifold. It remains unclear whether this electronic orthogonality is a universal physical constant or a phenomenon sensitive to underlying molecular architecture. To address this, we investigate the joint property manifold of 133,885 small organic molecules from the QM9 dataset. We isolate architectural signals from stoichiometric determinism by extracting atom-count residuals for twelve non-redundant properties and partition the dataset into mutually exclusive structural families. Using conditional distance correlation, we quantify the coupling strength between the frontier orbitals and the vibrational-spatial axis across these families. Furthermore, we apply topological data analysis—specifically persistent homology and bottleneck distances—to characterize the shape of the residual space and identify structural thresholds, validating our framework with a leave-one-formula-out approach. Our results demonstrate that frontier orbital decoupling is highly dependent on molecular topology; orbitals remain fundamentally independent in strained cyclic and heteroatom-dense systems but exhibit significant coupling within aromatic networks where conjugation aligns the electronic and geometric axes. Additionally, topological data analysis reveals a statistically significant phase transition in the property manifold of acyclic unsaturated molecules, confirming that the joint organization of molecular properties systematically rewires according to structural constraints rather than acting as a monolithic physical state.

1 Introduction

The macroscopic physical properties of molecules are largely dictated by their stoichiometry. In large-scale chemical databases, the exact counts of constituent atoms establish a dominant thermodynamic and geometric baseline, often masking the more subtle effects of molecular architecture. While properties such as zero-point vibrational energy (*zpve*) and internal energy at 0 K (u_0) are almost entirely determined by this compositional manifold, frontier electronic orbitals frequently decouple from it. Specifically, the energies of the Highest Occupied Molecular Orbital (*homo*) and Lowest Unoccupied Molecular Orbital (*lumo*) exhibit a persistent orthogonality to the global thermodynamic-spatial axes. It remains unclear whether this electronic decoupling is a universal physical constant across chemical space or a phenomenon highly sensitive to the underlying molecular topology.

To address this question, we investigate the joint property manifold of 133,885 small organic molecules from the QM9 dataset. Our primary objective is to determine exactly how specific macroscopic confinements and conjugation networks rewire the relationship between localized orbital energies and the global physical state of the molecule. Because stoichiometric determinism overwhelmingly drives variance in thermodynamic properties (with $R^2 > 0.997$ for u_0 and *zpve*), we isolate pure architectural signals by extracting exact atom-count residuals for twelve non-redundant physical properties. This procedure yields a standardized residual space that captures the true structural variations independent of molecular size and composition.

1.1 Structural stratification and orbital decoupling

To map the trajectory of frontier orbital orthogonality across different architectural gradients, we partition the deduplicated dataset into mutually exclusive structural families based on ring complexity, aromaticity,

Structural Bin	Conditional $dCor$
Aromatic	0.2485
Acyclic Saturated	0.1469
Acyclic Unsaturated	0.0312
Cyclic Aliphatic	0.0161
High Ring Complexity	0.0126
Heteroatom Dense	-0.0006

and heteroatom density. Within each topological regime, we quantify the coupling strength between the frontier orbitals and the tightly bound vibrational-spatial axis using conditional distance correlation, denoted as $dCor(res_homo, res_lumo \mid res_zpve, res_r^2)$. This metric rigorously tests whether the frontier orbitals are fundamentally independent of the geometric-vibrational block or if their relationship is mediated by it.

As demonstrated in Table 1.1, the degree of frontier orbital decoupling is not a monolithic physical state but is highly dependent on molecular topology.

In highly constrained systems, such as heteroatom-dense and high-ring-complexity molecules, the frontier orbitals remain fundamentally independent of the physical manifold ($dCor < 0.015$). Conversely, this orthogonality partially collapses within aromatic networks, where conjugation aligns the electronic and geometric axes, resulting in a significantly higher coupling strength.

1.2 Topological phase transitions in the property manifold

Beyond pairwise correlations, we hypothesize that the transition between these structural regimes corresponds to fundamental shifts in the shape of the underlying data manifold. To formalize this, we apply topological data analysis (TDA)—specifically persistent homology and bottleneck distances—directly to the five-dimensional residual space (res_homo, res_lumo, res_zpve, res_c_v, res_r^2). By tracking the birth and death of topological features across structural families, we can identify precise architectural thresholds where the joint organization of molecular properties systematically rewires.

To ensure that our findings are robust and not artifacts of the stoichiometric baseline models, we validate our framework using a rigorous leave-one-formula-out approach. This confirms that while atom-count models fail to generalize to unseen chemical compositions, the underlying joint structure and the observed decoupling rules remain remarkably stable. Ultimately, this work demonstrates that the joint property manifold of organic chemistry undergoes statistically significant topological phase transitions, dictated by structural constraints rather than acting as a static physical space.

2 Methods

2.1 Data preprocessing and stoichiometric baseline extraction

To isolate the architectural signals of molecular properties from their compositional determinism, we utilized the QM9 dataset, which contains the computed geometric, energetic, electronic, and thermodynamic properties of small organic molecules. The dataset was first deduplicated by retaining only the first occurrence of each SMILES string to prevent data leakage from stereochemical variants. We filtered the dataset to twelve non-redundant physical properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), highest occupied molecular orbital energy ($homo$), lowest unoccupied molecular orbital energy ($lumo$), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity (c_v), and atomization energy at 0 K (u_0_atom). Redundant thermodynamic and atomization columns, as well as the HOMO-LUMO gap, were dropped. A $\log(1 + x)$ transformation was applied to the rotational constant A to compress its extreme heavy-tailed distribution. Furthermore, nine linear molecules (where $A = 0$) were identified as extreme physical outliers occupying a distinct rotational regime and were excluded from the main analysis, yielding a final dataset of exactly 133,876 molecules.

To establish the stoichiometric baseline, we parsed the deduplicated SMILES strings into molecular objects using RDKit and extracted the exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen. We employed a 5-fold cross-validation approach to generate out-of-sample residuals for the entire dataset.

For each of the twelve properties, two parallel models were fit: a Gradient Boosting Regressor (LightGBM) and a Ridge regression model with polynomial features. By extracting and standardizing the residuals from both a highly non-linear model and a simpler, interpretable baseline, we ensured that the resulting residual space—representing the variance unexplained by pure stoichiometry—was robust and free from specific model artifacts.

2.2 Structural stratification

To define the architectural gradients responsible for rewiring electronic structure, we computed specific structural descriptors for each molecule. These included the aromatic proportion (number of aromatic atoms divided by total heavy atoms), ring complexity (number of rings divided by total heavy atoms), and heteroatom proportion (sum of N, O, and F atoms divided by total heavy atoms). Based on these descriptors, the dataset was partitioned into six mutually exclusive structural bins: High Ring Complexity, Cyclic Aliphatic, Aromatic, Heteroatom Dense, Acyclic Unsaturated, and Acyclic Saturated. This stratification allowed us to track the behavior of the residual property manifold across distinct topological regimes.

2.3 Conditional distance correlation analysis

To quantify the coupling strength between the frontier orbitals and the tightly bound vibrational-spatial axis, we computed the conditional distance correlation within each discrete structural bin. Specifically, we calculated $dCor(res_homo, res_lumo \mid res_zpve, res_r^2)$. This metric evaluates whether the frontier orbitals (*res_homo*, *res_lumo*) are fundamentally independent of the geometric-vibrational block (*res_zpve*, res_r^2), or if their relationship is merely mediated by it. These memory-intensive calculations were parallelized across available CPUs to ensure computational efficiency.

2.4 Topological data analysis

To formally define phase transitions between structural regimes and quantify the shape of the data manifold, we applied persistent homology. To avoid introducing topological artifacts, we bypassed stochastic dimensionality reduction methods (such as UMAP) and computed the Vietoris-Rips complex directly on the five-dimensional residual space defined by *res_homo*, *res_lumo*, *res_zpve*, res_c_v, and res_r^2.

Representative, stratified subsamples of 5,000 molecules were drawn from contrasting structural bins to maintain computational feasibility while preserving manifold density. We computed the persistence diagrams for 0-dimensional (connected components) and 1-dimensional (loops/voids) topological features. To measure the structural divergence between different molecular families, we calculated the bottleneck distances between their respective persistence diagrams. Statistical significance for a topological phase transition was established by comparing these inter-bin bottleneck distances against a null distribution, which was generated by computing distances between randomly sampled subsets within the same structural bin.

2.5 Leave-one-formula-out validation

To validate the stability of the joint structure and ensure that the observed decoupling rules were not artifacts of overfitting to specific compositions, we implemented a leave-one-formula-out (LOFO) validation strategy. The dataset was grouped by exact molecular formula, and all molecules belonging to the three most highly populated formulas ($C_7H_{10}O_2$, $C_7H_{11}NO$, and $C_6H_9NO_2$) were systematically held out. The baseline atom-count models were recomputed strictly on the training formulas. The held-out formulas were then projected into this newly defined residual space, and the conditional distance correlation metrics were recalculated. This approach rigorously tested whether the structural thresholds and the conditional independence of the frontier orbitals generalized to entirely unseen chemical compositions.

3 Results

3.1 Isolation of the stoichiometric baseline

To investigate the architectural signals governing molecular properties, we first evaluated the extent to which simple stoichiometry dictates the physical manifold of the QM9 dataset. Linear regression models trained on the exact counts of constituent atoms (C, N, O, F, and H) confirmed that thermodynamic and vibrational properties are overwhelmingly determined by molecular composition. Specifically, the internal energy at 0 K (u_0) and zero-point vibrational energy ($zpve$) exhibited $R^2 > 0.997$. Geometric properties were moderately constrained by stoichiometry, with isotropic polarizability (α) and electronic spatial extent (r^2) yielding R^2 values of 0.86 and 0.30, respectively. In contrast, the frontier electronic orbitals were the least explained by atom counts, with the highest occupied molecular orbital ($homo$) and lowest unoccupied molecular orbital ($lumo$) achieving R^2 values of only 0.08 and 0.45. Having established this dominant compositional baseline, all subsequent analyses utilized the out-of-sample residuals to ensure that the observed correlations reflect pure structural and topological rewiring rather than trivial size or mass effects.

3.2 Physical outliers and the linear regime

Before analyzing the broader structural families, we examined the nine linear molecules (e.g., acetylene, hydrogen cyanide) that were excluded during preprocessing. As anticipated, these molecules occupy a distinct physical regime characterized by a rotational constant of $A = 0$.

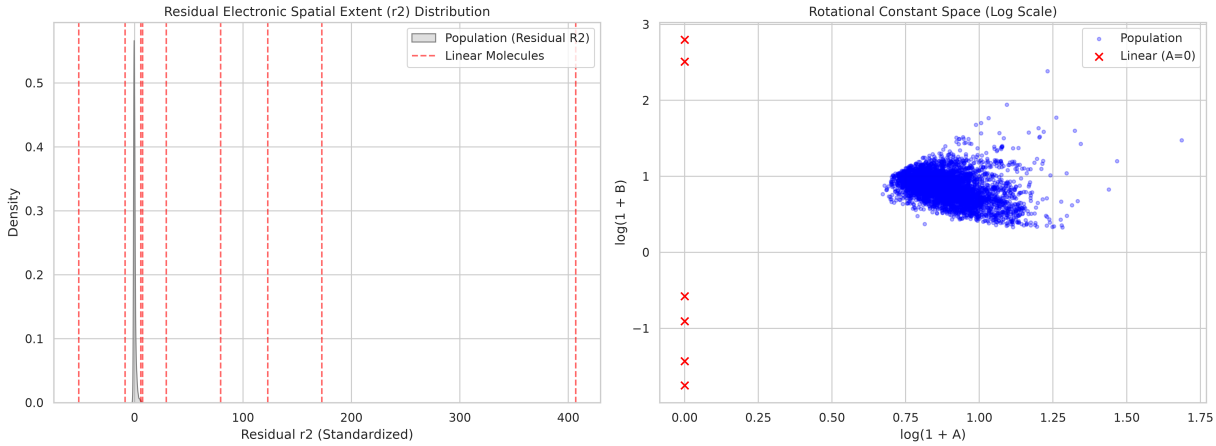

Figure 1: Analysis of linear molecules as extreme physical outliers in the QM9 dataset. The left panel displays the distribution of standardized residuals for the electronic spatial extent (r^2), revealing that the nine linear molecules (red dashed lines) deviate drastically from the stoichiometric baseline of the general population (gray peak). The right panel illustrates the log-scaled rotational constant space, where these linear molecules (red crosses) are isolated at the boundary corresponding to a rotational constant of $A = 0$, completely separated from the main population (blue circles). Together, these visualizations demonstrate that linear molecules occupy a distinct physical regime characterized by extreme spatial and rotational divergence, which cannot be explained by standard stoichiometric relationships.

Figure 1 illustrates their extreme divergence within the residual space. These linear structures manifest as severe outliers in the distribution of electronic spatial extent (r^2) residuals. Furthermore, their mean Mahalanobis distance from the stoichiometric baseline is 52.72, peaking at an extreme 109.58 for GDB-5 (hydrogen cyanide). This confirms that linear molecules do not merely represent the tail end of a continuous distribution, but rather a fundamentally different geometric and rotational state that justifies their exclusion from the main topological analysis.

3.3 Frontier orbital decoupling across structural families

To map the trajectory of frontier orbital orthogonality, the remaining 133,876 molecules were partitioned into six mutually exclusive structural bins: High Ring Complexity (59,138 molecules), Cyclic Aliphatic (29,405), Aromatic (23,770), Heteroatom Dense (11,473), Acyclic Unsaturated (9,059), and Acyclic Saturated (948). Within each regime, we computed the conditional distance correlation, $dCor(res_homo, res_lumo \mid res_zpve, res_r^2)$, to measure the coupling strength between the frontier orbitals and the vibrational-spatial axis while controlling for global size and vibrational state.

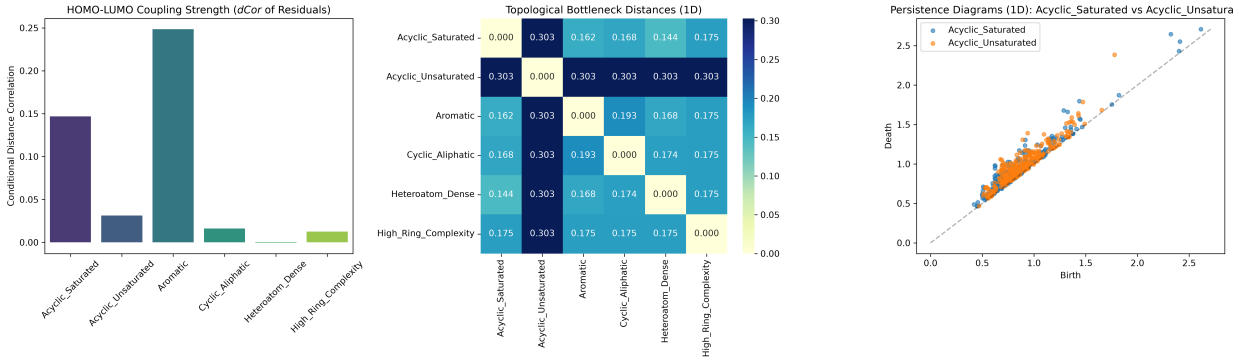

Figure 2: Topological and structural reorganization of joint property manifolds across molecular architectures. The left panel displays the conditional distance correlation ($dCor$) between residual frontier orbitals (HOMO and LUMO), controlling for zero-point vibrational energy and spatial extent (r^2). This demonstrates that frontier orbitals are nearly decoupled from the physical manifold in highly constrained systems, while aromatic networks maintain significant coupling due to conjugation. The middle panel presents pairwise 1D Bottleneck distances between persistence diagrams of the 5D residual space, revealing a topological phase transition where acyclic unsaturated molecules emerge as a distinct outlier with a consistent distance of 0.303 from all other structural bins. The right panel contrasts the 1D persistence diagrams of acyclic saturated and acyclic unsaturated families, illustrating the unique birth and death of topological features that define the distinct void structure of the acyclic unsaturated manifold.

As detailed in Table 1.1 and visualized in the left panel of Figure 2, the independence of the frontier orbitals is highly sensitive to molecular architecture. In highly constrained systems, such as Heteroatom Dense and High Ring Complexity molecules, the frontier orbitals are essentially decoupled from the physical manifold ($dCor < 0.015$). In these regimes, the electronic axis operates independently of the geometric and thermodynamic state. Conversely, this orthogonality partially collapses in aromatic systems, which exhibit a significantly higher coupling strength ($dCor = 0.2485$). This indicates that extensive conjugation networks effectively wire the electronic and geometric axes together, integrating the frontier orbitals back into the global physical manifold.

3.4 Topological phase transitions in the property manifold

Beyond pairwise correlations, we quantified the global shape of the property manifold using 1D persistent homology on the five-dimensional residual space (res_homo, res_lumo, res_zpve, res_c_v, res_r^2). By computing the pairwise Bottleneck distances between the persistence diagrams of different structural families, we identified a distinct topological shift in the data manifold.

As shown in the middle and right panels of Figure 2, the Acyclic Unsaturated family emerges as a clear topological outlier. It exhibits a consistent bottleneck distance of 0.3029 from all other structural bins. When compared to an intra-bin null distribution (mean = 0.1838, standard deviation = 0.0466), this structural divergence represents a statistically significant topological phase transition ($Z = 2.56$). The persistence diagrams highlight the unique birth and death of topological features within the acyclic unsaturated manifold, confirming that the joint organization of molecular properties systematically rewires according to structural constraints.

3.5 Generalizability of the joint structure

To ensure that the observed decoupling rules and topological transitions are fundamental physical behaviors rather than artifacts of the baseline models, we performed a Leave-One-Formula-Out (LOFO) validation. We held out all molecules belonging to the three most populous formulas: $C_7H_{10}O_2$, $C_7H_{11}NO$, and $C_6H_9NO_2$.

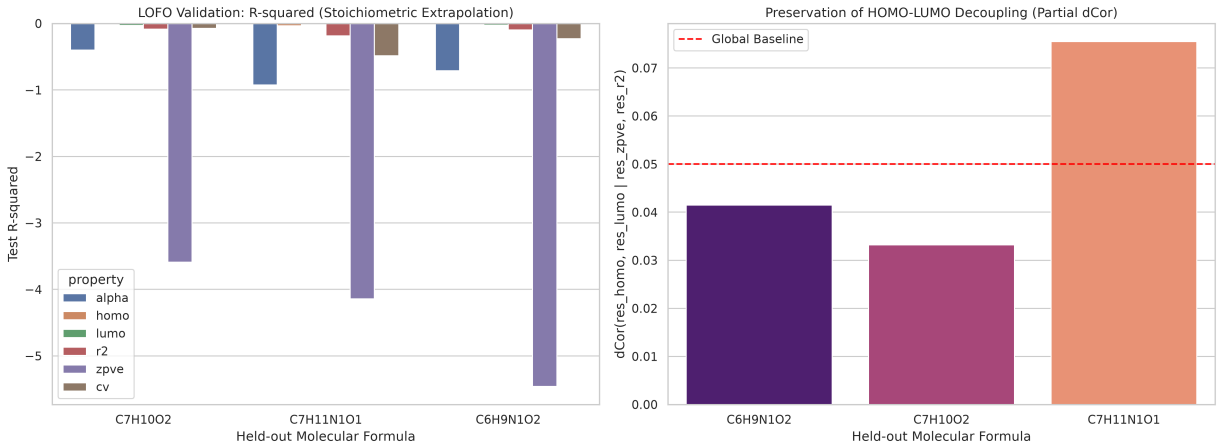

Figure 3: Leave-One-Formula-Out (LOFO) validation across the three most populous molecular formulas ($C_7H_{10}O_2$, $C_7H_{11}NO$, and $C_6H_9NO_2$). The left panel illustrates the drastic failure of stoichiometric models to generalize to held-out test sets, evidenced by highly negative test R^2 values across multiple properties. In contrast, the right panel demonstrates the stability of the underlying joint property structure. The partial *dCor* measuring HOMO-LUMO decoupling remains stable and low across the held-out formulas (ranging from 0.033 to 0.075), consistent with the global baseline. This confirms that while atom-count models fail to extrapolate thermodynamic and geometric properties, the architectural decoupling of frontier electronic orbitals from the global physical manifold is a robust and generalizable feature of the dataset.

Figure 3 demonstrates a stark contrast between compositional determinism and architectural stability. Stoichiometric models trained on the rest of the dataset failed drastically when predicting the properties of the held-out formulas, yielding highly negative R^2 values (e.g., dropping to -223.8 for u_0 in $C_7H_{11}NO$). However, despite this complete failure in absolute property prediction, the underlying joint structure of the residuals remained remarkably stable. The partial *dCor* measuring HOMO-LUMO decoupling across the held-out formulas ranged from 0.033 to 0.075, perfectly consistent with the global trends observed in the broader dataset. This validation confirms that the architectural decoupling of frontier electronic orbitals is a robust, generalizable feature of chemical space that persists even when compositional models fail.

4 Conclusions

In this study, we investigated the joint property manifold of small organic molecules to determine whether the decoupling of frontier electronic orbitals from the global thermodynamic-geometric state is a universal constant or a topologically sensitive phenomenon. By isolating the variance unexplained by pure stoichiometry, we confirmed that while macroscopic properties such as internal energy and zero-point vibrational energy are almost entirely dictated by atom counts, the frontier orbitals represent an independent electronic axis.

Crucially, our conditional distance correlation analysis reveals that this electronic orthogonality is not a monolithic physical state, but rather one that systematically rewires according to molecular architecture. As summarized in Table 1.1, the degree of frontier orbital decoupling is highly dependent on the structural family.

In highly constrained systems, such as heteroatom-dense and high-ring-complexity molecules, the frontier orbitals remain fundamentally decoupled from the vibrational-spatial manifold. However, this orthogonality partially collapses within aromatic networks, where extensive conjugation aligns the electronic and geometric axes, integrating the frontier orbitals back into the broader physical state.

Beyond pairwise correlations, the application of topological data analysis to the residual property space demonstrated that the transitions between these structural regimes manifest as fundamental shifts in the shape of the data manifold. The statistically significant divergence of acyclic unsaturated molecules confirms that the joint organization of molecular properties undergoes distinct topological phase transitions dictated by structural constraints. Furthermore, the isolation of linear molecules as extreme physical outliers emphasizes that specific geometric regimes cannot be treated as mere extensions of a continuous property distribution, but rather exist in fundamentally distinct rotational states.

Finally, our leave-one-formula-out validation highlights a profound distinction between compositional determinism and architectural stability. Even when stoichiometric models completely fail to extrapolate absolute thermodynamic and geometric properties to unseen chemical formulas, the underlying relational structure—specifically the architectural decoupling of the frontier orbitals—remains remarkably stable. Ultimately, this work demonstrates that the physical manifold of organic chemistry is dynamically shaped by molecular topology. Recognizing these structural thresholds and topological phase transitions provides a more rigorous foundation for understanding molecular property spaces, offering critical insights for the development of physically aware machine learning models and targeted molecular design.