Topological Discovery of HOMO-Coupling Regimes: Structural Rewiring of the QM9 Property Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The Highest Occupied Molecular Orbital (HOMO) energy typically exhibits global orthogonality, remaining decoupled from the tight joint structures formed by thermodynamic and spatial properties. To identify the architectural triggers that break this decoupling, we analyzed 133,802 molecules from the QM9 dataset by conditioning molecular properties on a stoichiometric baseline and clustering the resulting residual space. We employed type-aware matched-pair analysis to isolate the effects of single heteroatom substitutions and applied persistent homology to distance correlation matrices to quantify the topological rewiring of property networks against an acyclic baseline. Our analysis reveals six distinct physical regimes corresponding to specific structural motifs, such as strained fused systems, polycyclic aromatics, and simple chains. Matched-pair analysis demonstrates that specific substitutions, notably nitrogen to fluorine and carbon to nitrogen, act as primary drivers in breaking HOMO orthogonality, significantly increasing its coupling with heat capacity and spatial extent, respectively. Furthermore, topological stability analysis using Wasserstein distances identifies unique architectural states—particularly within specific ring-dense intermediate regimes—where electronic, spatial, and thermodynamic properties reorganize into complex, cyclic dependency networks. Ultimately, these findings demonstrate that the orthogonality of frontier electronic states is not a constant feature but a state actively modulated by specific structural constraints and heteroatom substitutions.

1 Introduction

The exploration of chemical space often relies on large-scale datasets like QM9, which provides comprehensive quantum mechanical, thermodynamic, and spatial properties for small organic molecules. Traditionally, these computed properties are analyzed as independent targets, or their variances are dominated by extensive size and stoichiometric effects. When these baseline size effects are decoupled, a persistent feature of the molecular property manifold emerges: the global orthogonality of the Highest Occupied Molecular Orbital (HOMO) energy. Across the broader chemical space, the HOMO energy typically remains isolated, refusing to participate in the tight joint structures formed by thermodynamic properties, such as heat capacity (c_v), and spatial properties, such as electronic spatial extent (r^2).

While this global decoupling is a well-documented baseline behavior, it is not an immutable physical law. Localized quantum states can be forced to integrate with global thermodynamic and spatial properties under specific structural constraints. However, identifying the exact architectural triggers—such as specific ring fusions, scaffold topologies, or heteroatom substitutions—that break this orthogonality remains a significant challenge. Understanding these triggers is crucial for mapping how fundamental physical laws reorganize within distinct molecular architectures, transitioning from independent property behaviors to integrated dependency networks.

In this study, we systematically identify the structural and chemical drivers that rewire the QM9 property manifold. We analyze 133,802 deduplicated molecules by first conditioning their properties on a stoichiometric baseline (C, N, O, F, and H atom counts) to isolate pure chemical organization from size effects. By clustering the resulting 12-dimensional residual space, we discover discrete physical behavior regimes. To rigorously quantify the reorganization of property relationships within these regimes, we apply persistent homology to distance correlation ($dCor$) matrices. By converting these matrices into metric spaces and constructing Vietoris-Rips complexes, we measure the topological rewiring of property networks against an

Substitution	Target Property	Count	Baseline $dCor$	New $dCor$	Shift ($\Delta dCor$)
N→F	Heat Capacity (c_v)	1,654	0.2285	0.3699	+0.1414
N→F	LUMO Energy	1,654	0.4485	0.5841	+0.1356
C→N	Spatial Extent (r^2)	95,963	0.1730	0.2995	+0.1265
O→F	ZPVE	1,534	0.1194	0.2290	+0.1096
C→N	Atomization (u_{0_atom})	95,963	0.2831	0.1620	-0.1211

Table 1: Significant HOMO Coupling Shifts ($\Delta dCor$) from Matched Pairs

acyclic baseline using Wasserstein distances. Furthermore, we employ a type-aware matched-pair analysis to isolate the precise effects of single heteroatom substitutions on property entanglement.

Our analysis reveals six distinct physical regimes corresponding to specific structural motifs, ranging from highly flexible simple chains to rigid, strained fused systems. Crucially, we demonstrate that the orthogonality of frontier electronic states is actively modulated by specific chemical transitions. As detailed in Table 1, matched-pair analysis identifies substitutions such as Nitrogen to Fluorine (N→F) and Carbon to Nitrogen (C→N) as primary drivers in breaking HOMO orthogonality, significantly increasing its coupling with heat capacity and spatial extent, respectively.

Finally, topological stability analysis uncovers unique architectural states—particularly within specific ring-dense intermediate regimes—where electronic, spatial, and thermodynamic properties reorganize into complex, cyclic dependency networks. Ultimately, these findings demonstrate that the orthogonality of frontier electronic states is a dynamic condition, actively rewired by structural constraints and heteroatom substitutions to form novel, integrated physical regimes.

2 Methods

2.1 Data preprocessing and stoichiometric decoupling

To isolate the fundamental chemical organization of molecules from extensive size effects, we first decoupled the molecular properties from their stoichiometric baseline. The dataset comprised 133,802 deduplicated molecules from the QM9 database. We restricted our analysis to 12 non-redundant properties, explicitly removing the perfectly correlated total energy metrics to prevent degenerate correlation structures.

For each property, we trained gradient-boosted decision trees using a five-feature stoichiometric baseline consisting of the exact atomic counts of Carbon, Nitrogen, Oxygen, Fluorine, and Hydrogen. The predictions of these baseline models were subtracted from the raw property values to yield stoichiometric residuals. To stabilize the variance across the diverse property scales and handle skewed distributions, a logarithmic transformation, specifically $\log(A+1)$, was applied to the residuals. This procedure mapped each molecule into a continuous 12-dimensional residual space, representing pure chemical and structural variance independent of basic stoichiometry.

2.2 Density-based regime discovery and structural attribution

To identify discrete physical behavior regimes, the 133,802 molecules were partitioned by clustering their 12-dimensional residual space. This density-based regime discovery allowed us to group molecules based on their intrinsic property coupling behaviors rather than predefined structural classes.

Following the clustering, we performed structural attribution by mapping the discovered regimes back to molecular motifs derived using RDKit. We quantified structural features such as Murcko scaffolds, ring density, ring fusion rates, aromaticity, and heteroatom patterns to characterize the architectural signatures of each regime. This mapping provided the physical basis for understanding how specific structural constraints dictate the organization of the residual space.

2.3 Distance correlation profiling and matched-pair analysis

Within each clustered regime, we computed unbiased, parallelized 12×12 distance correlation ($dCor$) matrices to quantify the non-linear dependencies between the property residuals. Particular attention was given to

the coupling between the typically isolated frontier orbital energies (HOMO and LUMO) and the spatial-thermodynamic manifold, including zero-point vibrational energy ($zpve$), heat capacity (c_v), and electronic spatial extent (r^2).

To isolate the exact architectural triggers responsible for breaking HOMO orthogonality, we implemented a type-aware hash-based matched-pair analysis. This algorithm utilized a hash lookup that replaced specific heteroatoms with atomic number placeholders, thereby preserving the substitution identity without collapsing distinct chemical structures. By comparing matched pairs of molecules differing by a single heteroatom substitution, we calculated the shift in distance correlation ($\Delta dCor$) for HOMO couplings. Linear molecule edge cases were explicitly handled during the hashing process to ensure robust pair matching across all structural families.

2.4 Topological stability and persistent homology

To rigorously quantify the structural rewiring of the property networks, we applied persistent homology to the correlation manifolds. The regime-specific distance correlation matrices were converted into metric spaces by defining the distance metric as $D = 1 - dCor$. Using these metric spaces, we constructed Vietoris-Rips complexes and utilized the `gudhi` library to extract their persistent homology.

We focused on 0D persistence, which captures the merging of property identities into connected components, and 1D persistence, which identifies the emergence of complex, non-transitive dependency cycles (topological "holes"). To measure the extent of topological rewiring, we computed the Wasserstein distances between the persistence diagrams of each regime and an acyclic baseline comprising 13,998 molecules.

Finally, to ensure the statistical stability of the identified topological features and rule out sampling artifacts, we employed a bootstrap validation strategy. We recomputed the unbiased $dCor$ matrices and persistence diagrams by subsampling 80% of the molecules over 10 independent iterations. The resulting Wasserstein distances were averaged, and standard deviations were calculated to confirm the robustness of the topological shifts against the global baseline.

3 Results

3.1 The breakdown of HOMO orthogonality

The computed properties of the 133,802 deduplicated molecules in the QM9 dataset are typically analyzed as independent targets. By conditioning all properties on a five-feature stoichiometric baseline (C, N, O, F, H counts) and analyzing the residuals, we successfully isolated chemical organization from extensive size effects. Global analysis of these residuals confirms a persistent decoupling of the Highest Occupied Molecular Orbital (HOMO) energy, which typically refuses to participate in the tight joint structures formed by thermodynamic and spatial properties.

However, our analysis reveals that this global orthogonality is not an immutable physical law. By partitioning the residual space, we identified specific architectural triggers that actively break this decoupling. Figure 1 illustrates the distance correlation ($dCor$) matrices across the discovered structural regimes, highlighting how the typically isolated HOMO residual (visible as dark orthogonal bands in the baseline) exhibits anomalous, strong dependencies with the spatial-thermodynamic manifold under specific structural constraints.

3.1.1 Architectural triggers via matched-pair analysis

To isolate the exact chemical transitions responsible for this rewiring, we employed a type-aware hash-based matched-pair analysis. This approach identified single heteroatom substitutions that force frontier electronic states to integrate with global thermodynamic and spatial properties. The global shifts in property couplings driven by these substitutions are visualized in Figure 2, while the specific impacts on HOMO orthogonality are detailed in Figure 3.

As quantified in Table 1, the substitution of nitrogen for fluorine (N→F) and carbon for nitrogen (C→N) act as primary drivers of this structural rewiring. The introduction of fluorine in place of nitrogen significantly strengthens the coupling between the HOMO and the vibrational density of states (encoded in the heat

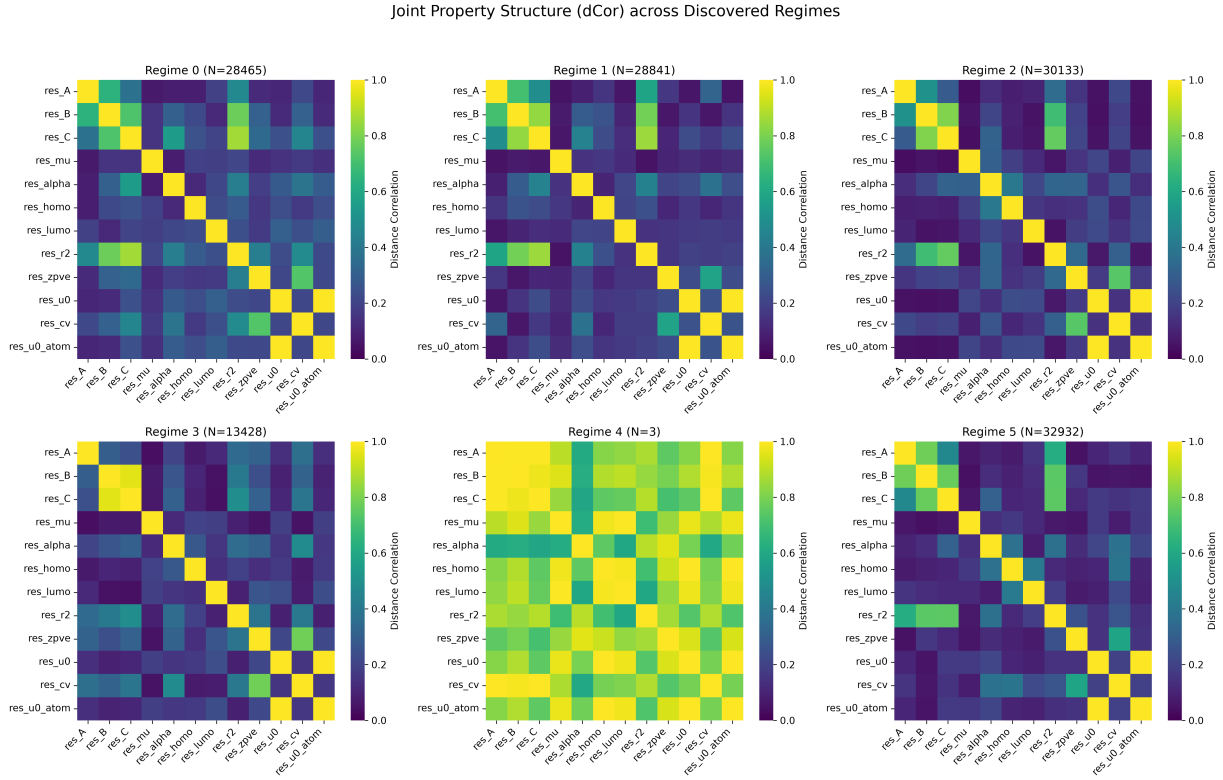


Figure 1: Distance correlation ($dCor$) matrices of the 12-dimensional residual property space across the six discovered structural regimes. The heatmaps visualize the joint property structures and demonstrate how specific molecular architectures rewire these relationships after conditioning on a stoichiometric baseline to remove size effects. The generally low correlation of the HOMO energy (visible as dark bands corresponding to the `res_homo` variable) reflects its typical orthogonality from the spatial-thermodynamic manifold, a state that is actively altered by complex ring architectures. Distinct correlation patterns across the regimes confirm the topological rewiring of property networks, ranging from the highly fragmented acyclic structures in Regime 3 to the complex, cyclic dependency networks in the ring-dense Regime 1, ultimately revealing that joint property structures are highly sensitive to molecular architecture.

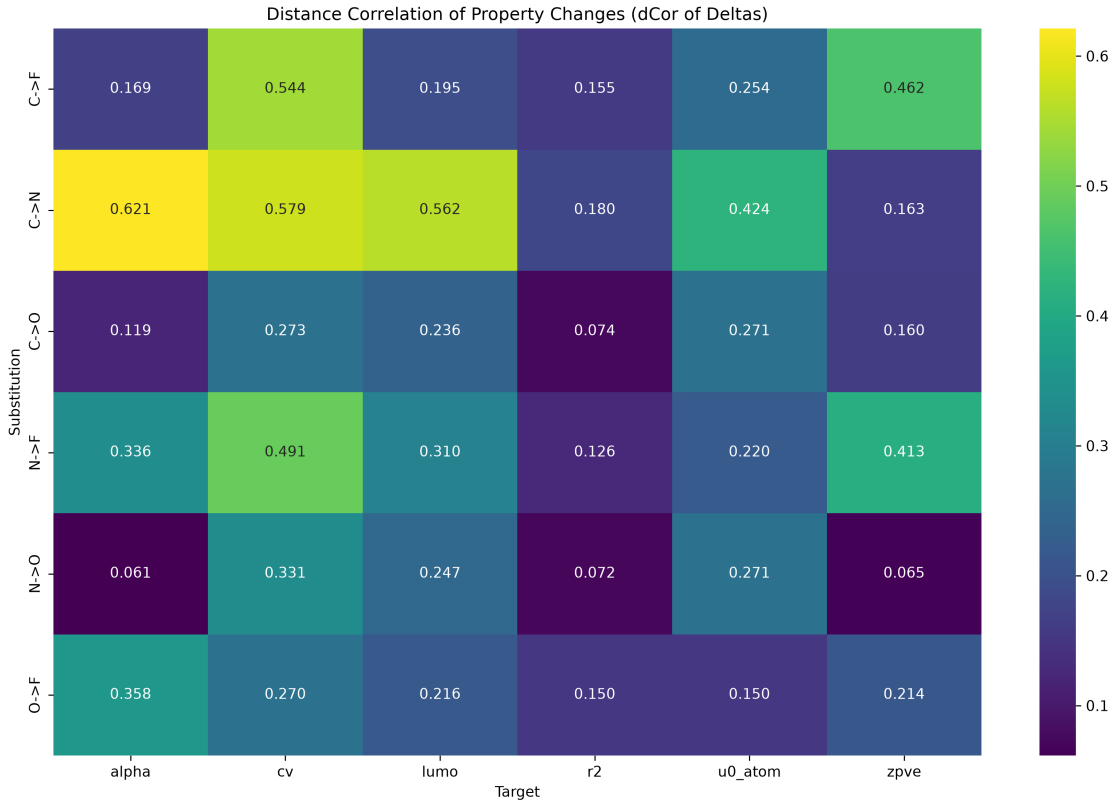


Figure 2: Heatmap illustrating the distance correlation ($dCor$) of property changes for single heteroatom substitutions, derived from matched-pair analysis. The plot identifies specific chemical transitions, such as C→N and N→F, as primary architectural triggers that drive the structural rewiring of the molecular property manifold. These substitutions break the typical orthogonality of frontier electronic states, forcing them to integrate with spatial and thermodynamic properties. Notably, C→N substitutions exhibit strong correlations with changes in polarizability (α), heat capacity (c_v), and LUMO energy, while N→F transitions strongly couple with c_v and zero-point vibrational energy (ZPVE). Ultimately, this demonstrates how specific structural modifications actively reorganize independent molecular properties into integrated dependency networks.

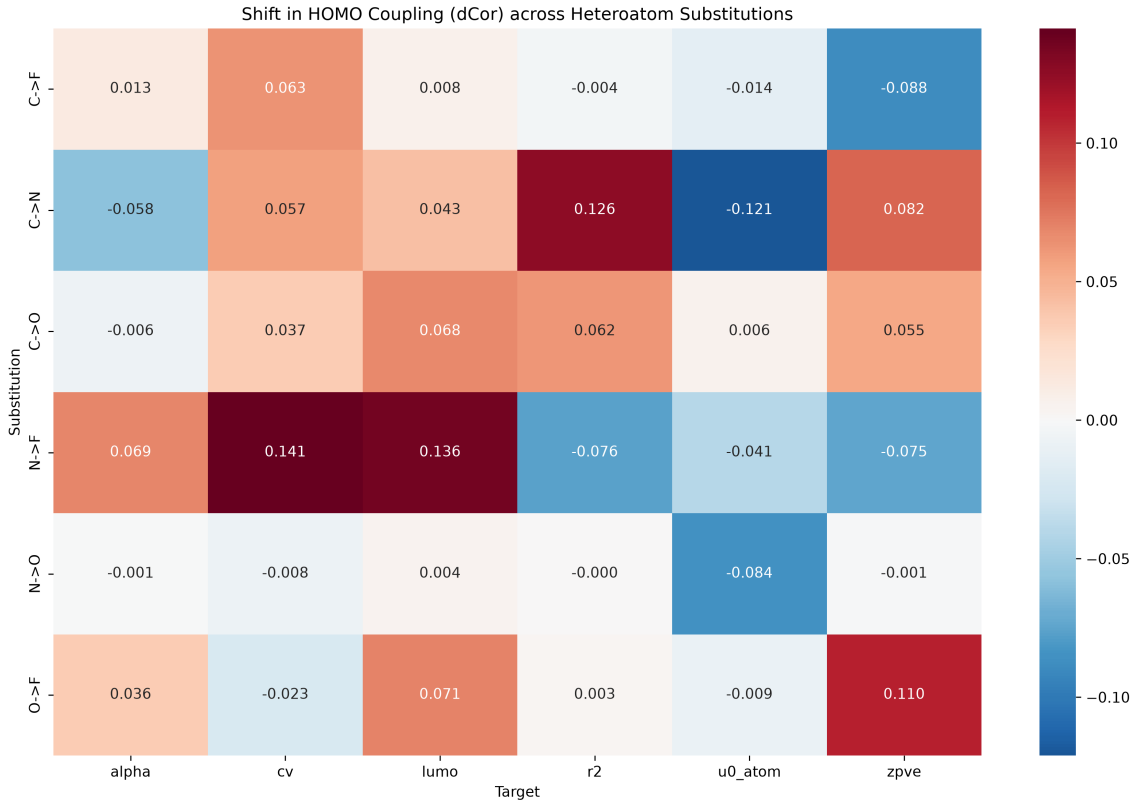


Figure 3: Heatmap illustrating the shift in distance correlation ($\Delta dCor$) between the HOMO energy and target properties across single heteroatom substitutions. Derived from matched-pair analysis, the plot demonstrates how specific architectural triggers break the typical orthogonality of the HOMO, forcing frontier electronic states to integrate with the spatial-thermodynamic manifold. Notably, introducing fluorine in place of nitrogen (N→F) significantly strengthens HOMO coupling with the vibrational density of states (heat capacity, c_v) and LUMO energy. Conversely, carbon-to-nitrogen (C→N) substitutions decouple the HOMO from the total bonding energy (atomization energy) while increasing its alignment with the electronic spatial extent (r^2). Additionally, O→F substitutions increase coupling with the zero-point vibrational energy (ZPVE). These shifts reveal that HOMO orthogonality is not a constant baseline, but a state that is actively rewired by specific chemical transitions to form integrated, cyclic property networks.

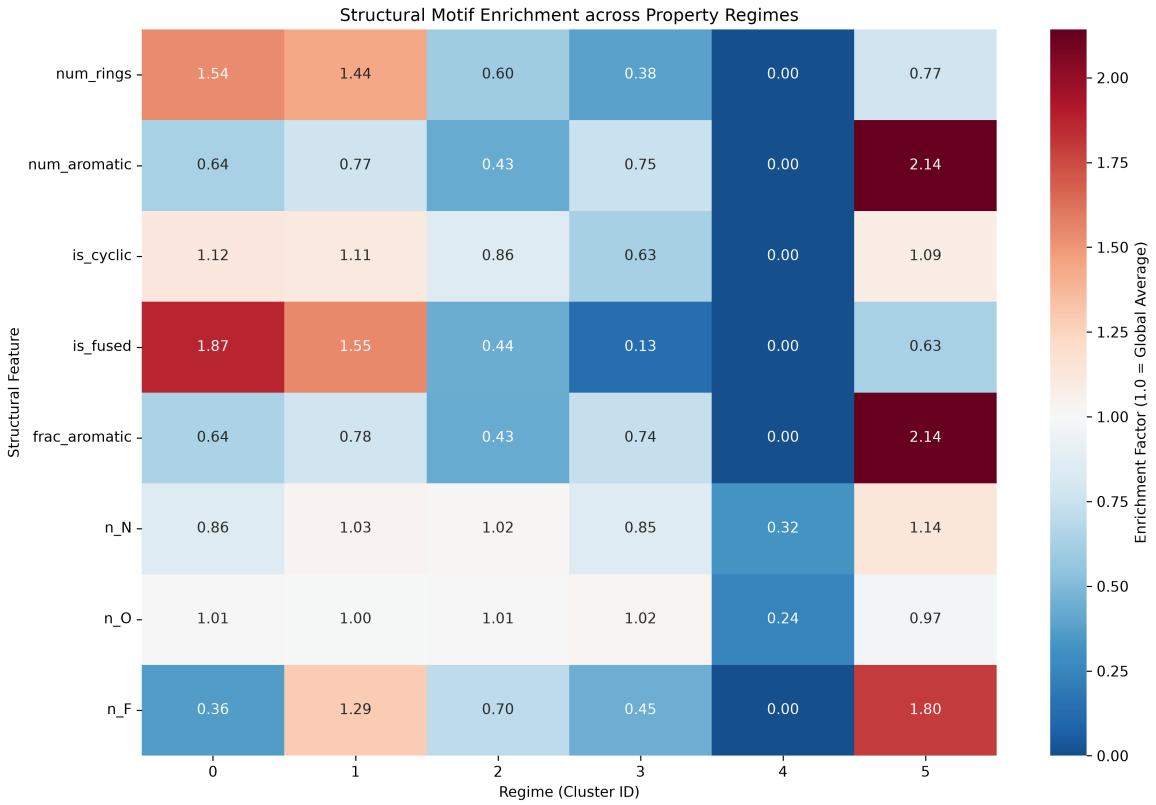


Figure 4: Heatmap detailing the structural motif enrichment across six discrete property regimes relative to the global average (1.0). The visualization confirms that the regimes, discovered by clustering a 12-dimensional residual space, correspond to coherent chemical families rather than mere statistical partitions. Regime 0 is heavily enriched in fused and ring structures (`is_fused`, `num_rings`), representing strained or fused systems. Regime 5 exhibits the highest enrichment in aromatic features (`num_aromatic`, `frac_aromatic`) and fluorine atoms (`n_F`), characterizing polycyclic aromatics. Furthermore, Regime 1 serves as an intermediate state, displaying a high density of rings but a lower fusion rate compared to Regime 0. Ultimately, this structural attribution demonstrates how specific molecular architectures actively drive the topological rewiring of property networks, forcing transitions from independent behaviors to integrated regimes.

capacity, c_v), increasing the distance correlation by 0.1414. Conversely, C→N substitutions decouple the HOMO from the total bonding energy (u_{0_atom}, $\Delta dCor = -0.1211$) while simultaneously increasing its alignment with the electronic spatial extent (r^2, $\Delta dCor = +0.1265$).

3.2 Discovery of discrete physical regimes

Clustering the 12-dimensional residual space revealed six distinct regimes of physical behavior. Rather than representing mere statistical partitions, these regimes correspond to coherent chemical families that organize molecular properties in fundamentally different ways.

3.2.1 Structural enrichment and motif mapping

By mapping the discovered regimes back to RDKit-derived structural motifs, we characterized the architectural signatures of each regime (Figure 4 and Figure 5).

The structural attribution confirms distinct chemical identities:

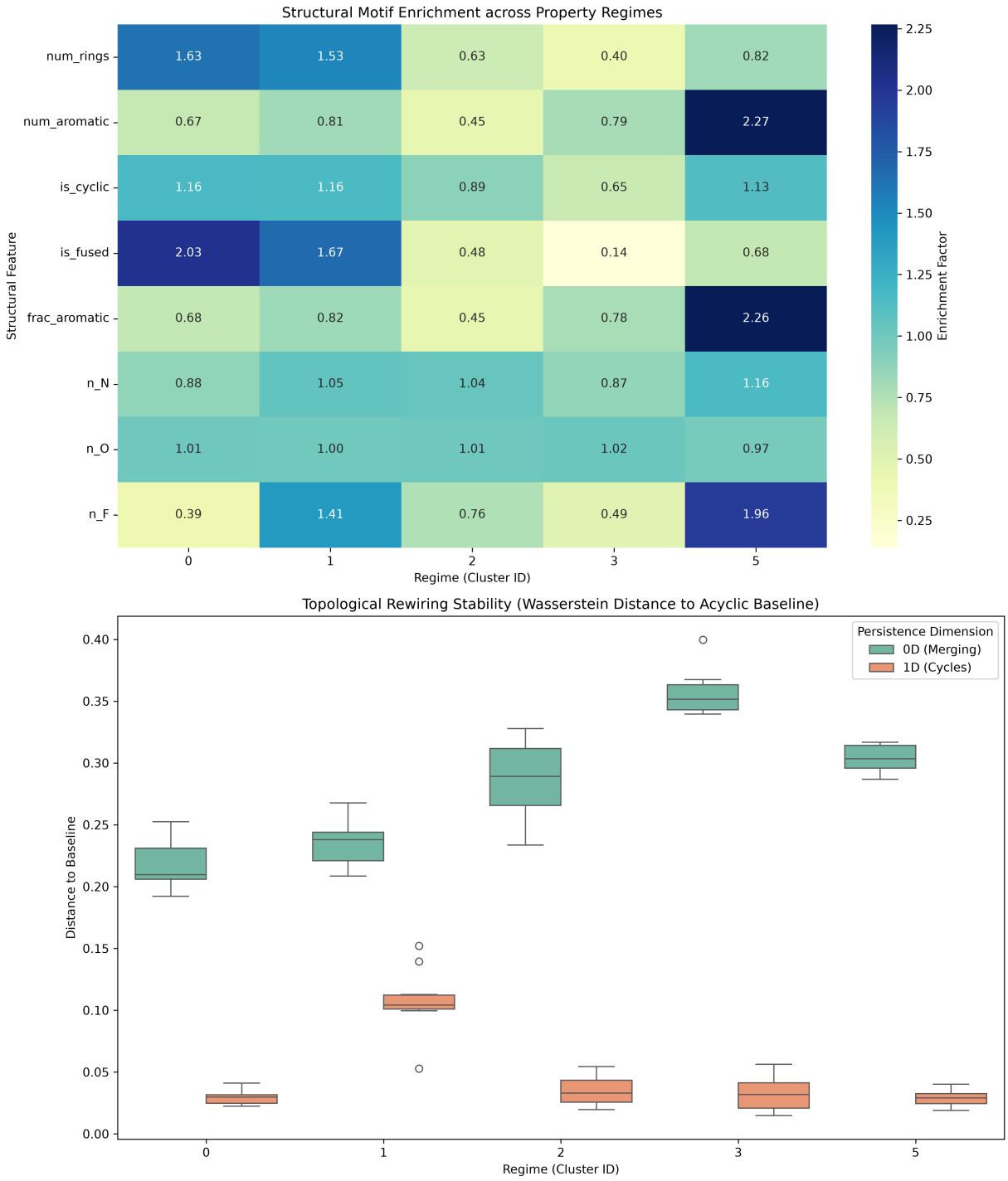


Figure 5: Structural motif enrichment and topological rewiring stability across discrete physical regimes. The top heatmap displays the enrichment factors of structural features, confirming distinct chemical identities: Regime 0 is highly enriched in fused systems (`is_fused`) and rings (`num_rings`), Regime 5 is dominated by aromaticity (`num_aromatic`), and Regime 3 represents flexible molecules with low ring counts. The bottom boxplot quantifies the topological reorganization of property networks via the Wasserstein distance to an acyclic baseline across 0D (merging) and 1D (cycles) persistence dimensions. Low variance across bootstrap iterations demonstrates the robustness of these topological shifts. Notably, Regime 3 exhibits the most significant 0D global rewiring (W_{Dist_Dim0}), indicating a highly fragmented property network, whereas Regime 0 remains topologically similar to the baseline despite its structural complexity. Furthermore, Regime 1 displays an extraordinary 1D topological shift (W_{Dist_Dim1}), revealing the emergence of complex, non-transitive dependency cycles where properties are rewired into closed loops.

- **Regime 0 (Strained/fused systems):** Exhibits the highest ring density (2.677 rings per molecule) and the highest fusion rate (85.8%).
- **Regime 5 (Polycyclic aromatics):** Dominated by aromaticity (mean 2.077 aromatic atoms) and the highest heavy atom count (8.48).
- **Regime 3 (Acyclic/simple chains):** Features the lowest ring count (0.656), representing the most structurally flexible molecules in the dataset.

Furthermore, Regime 1 and Regime 2 serve as intermediate states. Regime 1, in particular, exhibits a high density of rings (2.507) but a significantly lower fusion rate than Regime 0, setting the stage for unique topological property behaviors.

3.3 Topological rewiring of property networks

To rigorously quantify how these structural regimes reorganize property relationships, we applied persistent homology to the distance correlation manifold ($D = 1 - dCor$). We measured the topological distance (Wasserstein distance) of each regime from a baseline of 13,998 acyclic molecules. The resulting persistence barcodes and diagrams (Figure 6) reveal profound differences in how properties entangle across regimes.

3.3.1 0D persistence: the merging of property identities

The 0D Wasserstein distance (W_{Dist_Dim0}) measures the overall shift in how properties merge into a single connected component. Regime 3 exhibits the most significant global rewiring ($W_{Dist_Dim0} = 0.3560 \pm 0.0182$), indicating a property network that is highly fragmented compared to the acyclic baseline. Interestingly, Regime 0, despite its high structural complexity and ring fusion, remains topologically closest to the baseline ($W_{Dist_Dim0} = 0.2173 \pm 0.0195$). This suggests that the extreme rigidity of fused rings enforces a hierarchy of property couplings that is topologically similar to that of simple chains.

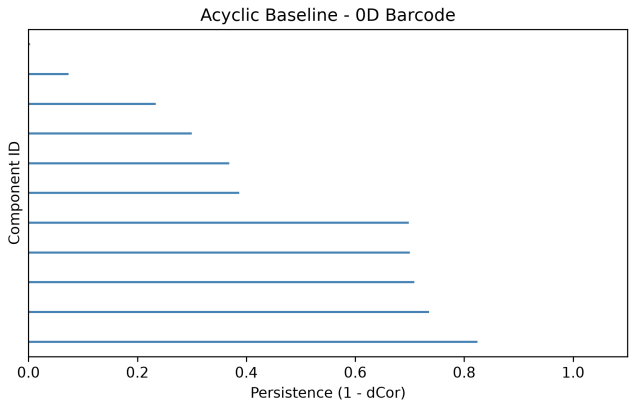
3.3.2 1D persistence: emergence of dependency cycles

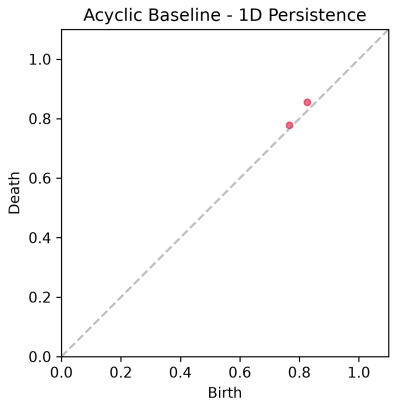
1D persistence captures topological "holes" in the correlation manifold—complex, non-transitive dependency cycles where multiple properties constrain each other in a closed loop that cannot be explained by a single shared latent factor. Here, Regime 1 shows an extraordinary 1D topological shift ($W_{Dist_Dim1} = 0.1079 \pm 0.0262$). This shift is nearly four times larger than that observed in the highly aromatic Regime 5 (0.0295) or the fused Regime 0 (0.0293). This prominent 1D signature identifies Regime 1 as a unique architectural state where electronic, spatial, and thermodynamic properties are fundamentally rewired into a complex, cyclic dependency network.

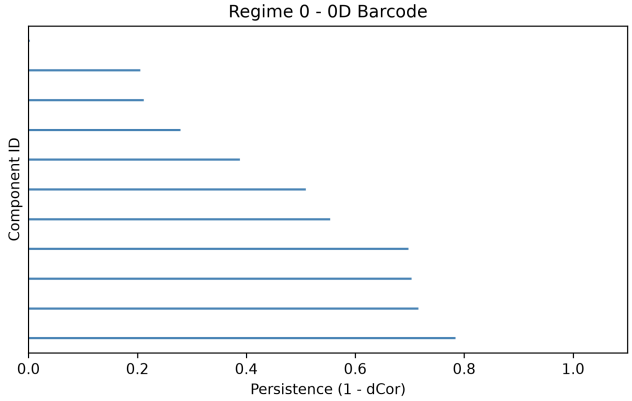
3.4 Stability and robustness

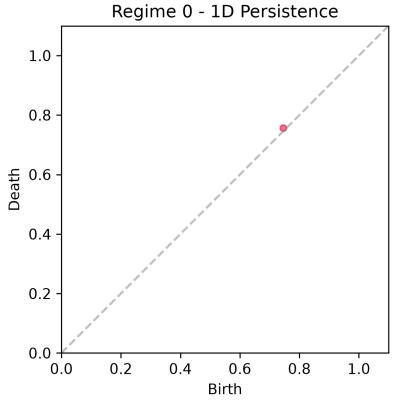
To ensure that the identified regimes and their topological features are robust characteristics of the QM9 population rather than sampling artifacts, we performed bootstrap validation using 10 iterations of 80% subsampling.

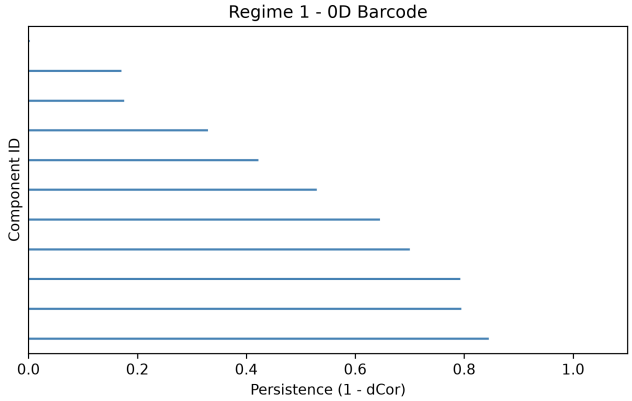
As demonstrated in Figure 7 (and corroborated by the stability analysis in Figure 5), the standard deviations for the topological distances across all regimes are consistently low. The narrow distributions confirm the statistical stability of the topological rewiring. By removing the dominant stoichiometric baseline and redundant energy metrics, we have shown that the remaining joint structure of the QM9 property manifold is highly sensitive to molecular architecture. The orthogonality of the HOMO is not a constant baseline; rather, it is a dynamic state that is actively broken by specific heteroatom substitutions and complex ring architectures, forcing a transition from independent property behaviors to integrated, cyclic physical networks.

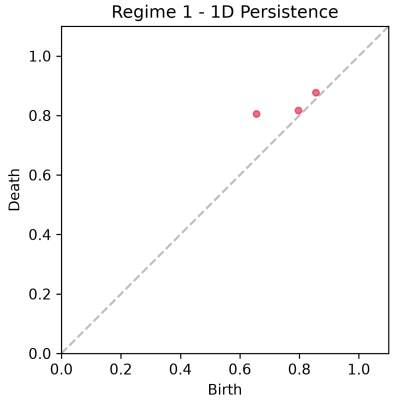

Acyclic Baseline - 0D Barcode
Component ID
Persistence (1 - dCor)
0.0
0.2
0.4
0.6
0.8
1.0

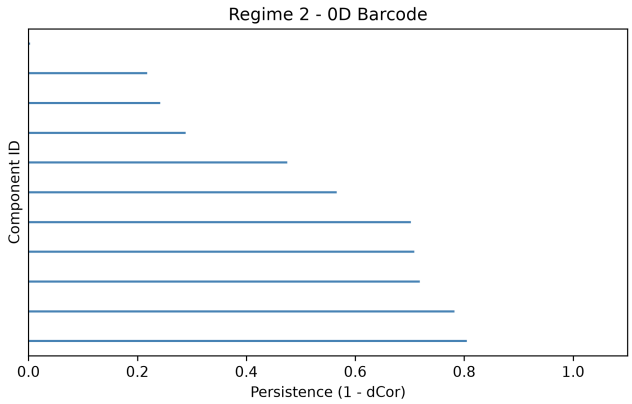

Acyclic Baseline - 1D Persistence
Death
Birth
0.0
0.2
0.4
0.6
0.8
1.0

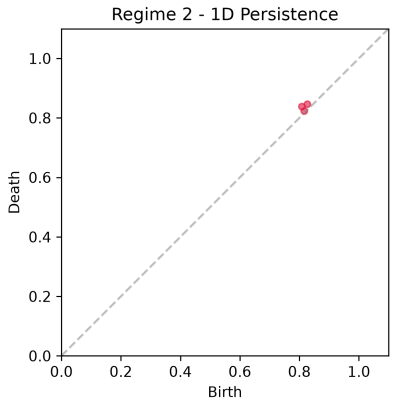

Regime 0 - 0D Barcode
Component ID
Persistence (1 - dCor)
0.0
0.2
0.4
0.6
0.8
1.0

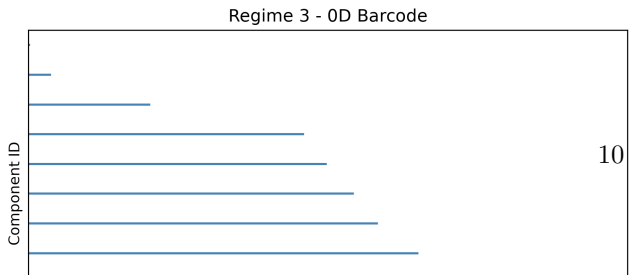

Regime 0 - 1D Persistence
Death
Birth
0.0
0.2
0.4
0.6
0.8
1.0

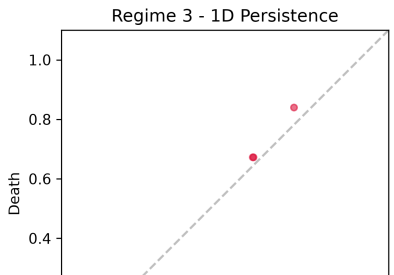

Regime 1 - 0D Barcode
Component ID
Persistence (1 - dCor)
0.0
0.2
0.4
0.6
0.8
1.0


Regime 1 - 1D Persistence
Death
Birth
0.0
0.2
0.4
0.6
0.8
1.0


Regime 2 - 0D Barcode
Component ID
Persistence (1 - dCor)
0.0
0.2
0.4
0.6
0.8
1.0


Regime 2 - 1D Persistence
Death
Birth
0.0
0.2
0.4
0.6
0.8
1.0


Regime 3 - 0D Barcode
Component ID
Persistence (1 - dCor)
0.0
0.2
0.4
0.6
0.8
1.0


Regime 3 - 1D Persistence
Death
Birth
0.0
0.2
0.4
0.6
0.8
1.0

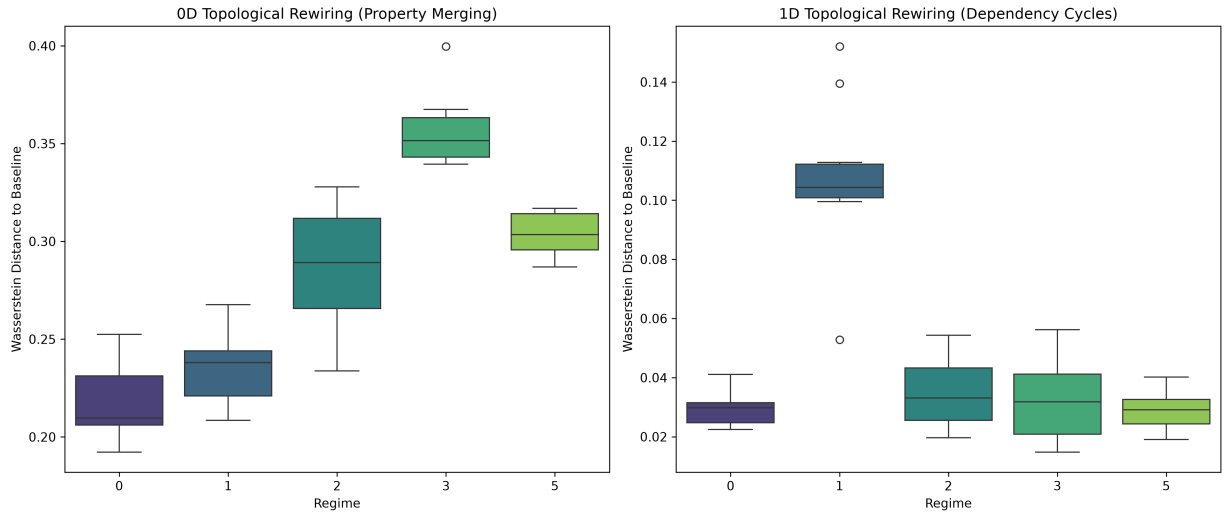


Figure 7: Topological rewiring of property networks across distinct structural regimes, quantified by the Wasserstein distance to an acyclic baseline. The left panel illustrates 0D persistence (property merging), revealing that Regime 3 undergoes the most significant global rewiring, while the highly fused Regime 0 remains topologically closest to the baseline. The right panel displays 1D persistence (dependency cycles), highlighting an extraordinary topological shift in Regime 1. This prominent 1D signature indicates that Regime 1 represents a unique architectural state where electronic, spatial, and thermodynamic properties reorganize into complex, non-transitive cyclic dependency networks. The narrow distributions in the box-plots, generated via bootstrap validation with 10 iterations of 80% subsampling, confirm that these discrete physical regimes and their topological features are highly stable and robust across the molecular population.

4 Conclusions

This study demonstrates that the global orthogonality of the Highest Occupied Molecular Orbital (HOMO) energy—a persistent feature across large chemical datasets like QM9—is not an immutable physical law, but rather a dynamic state actively modulated by molecular architecture. By decoupling extensive stoichiometric effects and analyzing the residual property space, we shifted the analytical focus from independent property variances to the underlying joint structures that govern chemical behavior.

Through type-aware matched-pair analysis, we identified specific heteroatom substitutions that serve as primary architectural triggers for breaking HOMO orthogonality. Notably, the substitution of nitrogen for fluorine (N→F) forces frontier electronic states to integrate with the vibrational density of states, while carbon-to-nitrogen (C→N) transitions couple the HOMO with the electronic spatial extent. These targeted chemical shifts prove that localized quantum states can be systematically entangled with global thermodynamic and spatial properties.

Furthermore, the application of persistent homology to distance correlation manifolds revealed that molecular properties organize into discrete, topologically distinct regimes. We discovered that intermediate, ring-dense architectures—specifically Regime 1—exhibit extraordinary 1D topological shifts. In these structures, the property manifold is fundamentally rewired, giving rise to complex, non-transitive dependency cycles where electronic, spatial, and thermodynamic properties constrain one another in closed loops. Conversely, highly strained and fused systems (Regime 0) enforce a rigid hierarchy of property couplings that topologically mirrors the behavior of simple acyclic chains.

Ultimately, our findings provide a rigorous quantitative framework for understanding how fundamental physical relationships reorganize within distinct molecular architectures. By proving that joint property structures are highly sensitive to structural constraints and heteroatom substitutions, this work highlights the necessity of treating molecular properties not as isolated targets, but as interconnected networks whose topologies are dictated by the underlying chemical scaffold.