Identifying Topological Thresholds of HOMO Integration via Conditional Dependency Analysis

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy in small organic molecules frequently appears orthogonal to the broader spatial and thermodynamic manifold. To determine whether this electronic isolation is a universal property or breaks down under specific architectural constraints, we systematically investigate the QM9 dataset by tracking the conditional dependencies of HOMO energies across gradients of topological complexity. We employ stoichiometric residualization via gradient-boosted atom-count residuals to strictly eliminate molecular size effects, followed by partial distance correlation analysis along a continuous BertzCT topological gradient. By mapping these trajectories, we identify distinct topological phase transitions where frontier orbitals integrate with the molecule's vibrational-rotational manifold. Our results reveal that for pure hydrocarbons, this integration occurs at a complexity threshold of approximately 259 BertzCT, driven by acyclic branching and simple ring fusions. Conversely, the introduction of heteroatoms fundamentally rewires these property dependencies; heteroatoms act as electronic insulators that delay integration, shifting the threshold to approximately 334 BertzCT. This pronounced shift demonstrates that heteroatom-rich molecules require significantly more complex, rigid polycyclic architectures to force the coupling of frontier electronic states with the global molecular geometry.

1 Introduction

The comprehensive mapping of chemical space through large-scale datasets, such as QM9, has revealed deep underlying correlations among the spatial, thermodynamic, and electronic properties of small organic molecules. Because many of these macroscopic properties scale predictably with molecular size and atomic composition, they form a tightly bound spatial-thermodynamic manifold. However, within this highly correlated space, the highest occupied molecular orbital (*homo*) energy frequently emerges as an anomaly. Across diverse molecular architectures, the *homo* appears remarkably orthogonal to the broader physical manifold, exhibiting minimal correlation with global geometric or thermodynamic descriptors.

This electronic isolation raises a fundamental physical question: is the decoupling of frontier orbitals a universal constant of small organic molecules, or does it break down under specific architectural constraints? Because frontier orbitals dictate chemical reactivity, understanding precisely when and how these localized electronic states are forced to integrate with the global molecular geometry is critical for rational molecular design.

1.1 Isolating structural dependencies from stoichiometric determinism

To rigorously investigate the orthogonality of the *homo*, it is necessary to separate genuine structural and topological dependencies from the dominant, confounding effects of molecular size. In standard datasets, thermodynamic properties such as zero-point vibrational energy (*zpve*) and internal energy at 0 K (u_0) are almost entirely determined by atomic composition.

To eliminate this stoichiometric determinism, we employ a strict residualization framework. By regressing the core molecular properties against a baseline of exact atom counts (C, N, O, F, and H) using cross-validated gradient boosting, we extract atom-count residuals. This transformation isolates the variance in properties that is purely attributable to molecular architecture—such as branching, ring strain, and spatial

Property Residual	Mean PDC (Hydrocarbons)	Mean PDC (Heteroatoms)
res_lumo	0.2982	0.0897
res_cv	0.0768	0.0472
res_mu	0.0701	0.0471
res_zpve	0.0664	0.0439
res_u0_atom	0.1065	0.0343

confinement. Within this decoupled residual manifold, the *homo* remains highly isolated, confirming that its lack of correlation is a fundamental characteristic of frontier electronic states rather than a statistical artifact of size scaling.

1.2 Topological thresholds and heteroatom rewiring

To determine the conditions under which this isolation breaks down, we systematically track the conditional dependencies of the *homo* energy across gradients of topological complexity. We utilize Partial Distance Correlation (PDC) to quantify the non-linear coupling between the *homo* residual and the remaining spatial-vibrational-rotational manifold, evaluated across a continuous sliding window of the BertzCT topological index.

By mapping these multidimensional PDC trajectories, we identify distinct topological phase transitions where spatial confinement and structural rigidification force the frontier orbitals to strongly couple with the molecule's joint physical structure. Crucially, our analysis reveals that this integration is not uniform across all chemical classes; rather, it bifurcates sharply based on the presence of electronegative atoms.

To illustrate the magnitude of this bifurcation, we can observe the baseline mean partial distance correlations of the *homo* residual against other key property residuals across the two chemical populations (Table 1.2).

The stark collapse in coupling—most notably between the *homo* and *lumo* residuals, which drops from approximately 0.30 in pure hydrocarbons to 0.09 in heteroatom-rich systems—demonstrates that heteroatoms fundamentally rewire the electronic dependencies of the molecule. Heteroatoms act as electronic insulators, localizing the frontier orbitals and delaying their integration into the global manifold. Consequently, heteroatom-rich molecules require significantly more complex, rigid polycyclic architectures to force the coupling of frontier electronic states with the global molecular geometry. In the following sections, we detail the stoichiometric residualization process, map the exact topological thresholds of these phase transitions, and extract the specific chemical motifs responsible for driving HOMO integration.

2 Methods

2.1 Data initialization and robust scaling

To prepare the QM9 dataset for topological and conditional dependency analysis, we first deduplicated the dataset by removing 83 redundant SMILES strings. Redundant thermodynamic columns (u_{298}, h_{298}, g_{298}, u_{298_atom}, h_{298_atom}, g_{298_atom}) and the energy gap (*gap*) were excluded to focus the analysis on the core spatial, vibrational, and rotational properties. We identified and isolated the 9 linear molecules (characterized by a rotational constant $A = 0$) to prevent extreme-value distortion during scaling.

For the rotational constants B and C, we applied a quantile transformation to normalize their distributions. To ensure consistency and prevent the linear molecules from skewing the transformation, the transformer was fitted exclusively on the non-linear population first, then applied to the entire dataset. The transformed B and C values were subsequently averaged into a single feature, BC_{avg}. Finally, we applied a $\log(A + 1)$ transformation to the rotational constant A for the non-linear molecules to neutralize extreme outliers and stabilize the variance.

2.2 Stoichiometric residualization

As outlined in the introduction, thermodynamic and spatial properties in small organic molecules are heavily confounded by molecular size. To strictly eliminate this stoichiometric determinism, we extracted exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen for each molecule. We then regressed the 10 core properties against these atom counts using cross-validated gradient boosting.

The resulting predictions were subtracted from the original values to isolate the atom-count residuals (e.g., *res_homo*, *res_zpve*, *res_cv*). We enforced the exclusive use of these gradient-boosted residuals for all subsequent multivariate analyses. This transformation ensures that any observed correlations or phase transitions are purely attributable to molecular architecture—such as branching, ring strain, and spatial confinement—rather than simple size scaling.

2.3 Heteroatom track bifurcation and topological gradients

To investigate the hypothesis that electronegative atoms rewire the electronic dependencies of the molecule, we bifurcated the dataset into two parallel analytical tracks based on a strict quantitative definition. The "hydrocarbon-like" population was defined as molecules containing only carbon and hydrogen (sum of N, O, and F atom counts equals zero), while the "heteroatom-rich" population included all molecules with at least one heteroatom (sum of N, O, and F atom counts greater than zero).

For both tracks, we computed the BertzCT topological index using RDKit. The BertzCT index quantifies the complexity of a molecular graph based on its symmetry and branching, serving as our primary continuous axis of topological complexity across which the integration of the *homo* could be tracked.

2.4 Partial distance correlation and phase transition identification

To pinpoint the exact structural thresholds where spatial confinement forces the *homo* to couple with the tightly bound vibrational-spatial manifold, we computed the Partial Distance Correlation (PDC) of *res_homo* against the remaining 9 property residuals. This calculation was explicitly conditioned on the atom-count residuals of the surrounding manifold to maintain the size-decoupled framework.

We executed these PDC calculations across a continuous sliding window of the BertzCT space using density-normalized Gaussian kernel weighting. The weights were strictly normalized by the local density of the descriptor space to prevent sample-size artifacts from masquerading as structural thresholds. Given the computational intensity of multidimensional distance correlations, we parallelized the grid-point iterations across all available CPUs, employing a representative subsample of up to 10,000 molecules per grid point where necessary to accurately reflect the local density.

By tracking the evolution of these PDC trajectories independently for both the hydrocarbon and heteroatom tracks, we identified the exact topological coordinates (BertzCT values) where the trajectories exhibited phase transitions—marking the integration of the frontier orbitals into the structural manifold. We performed a sensitivity analysis by varying the Gaussian kernel bandwidth to confirm that the identified thresholds were stable physical transitions ($\sigma_{stability}$) and not smoothing artifacts. Finally, we extracted Murcko scaffolds for molecules at these identified phase transitions to map the mathematical thresholds directly to specific chemical motifs, such as acyclic branching or polycyclic fusions.

2.5 Validation and manifold projection

To validate the robustness of the identified phase transitions, we isolated a strict hold-out set of 3,054 geometrically inconsistent molecules (as flagged in the original QM9 dataset) and verified that the integration thresholds persisted when this subset was removed.

To visually confirm the spatial localization of the *homo* integration, we projected the residualized molecular data onto a Uniform Manifold Approximation and Projection (UMAP) space. The projected points were colored by the local PDC value of *res_homo*, allowing us to observe the segregation between the hydrocarbon and heteroatom tracks and confirm that the integration of frontier electronic states is localized to specific, highly complex regions of the topological-chemical space.

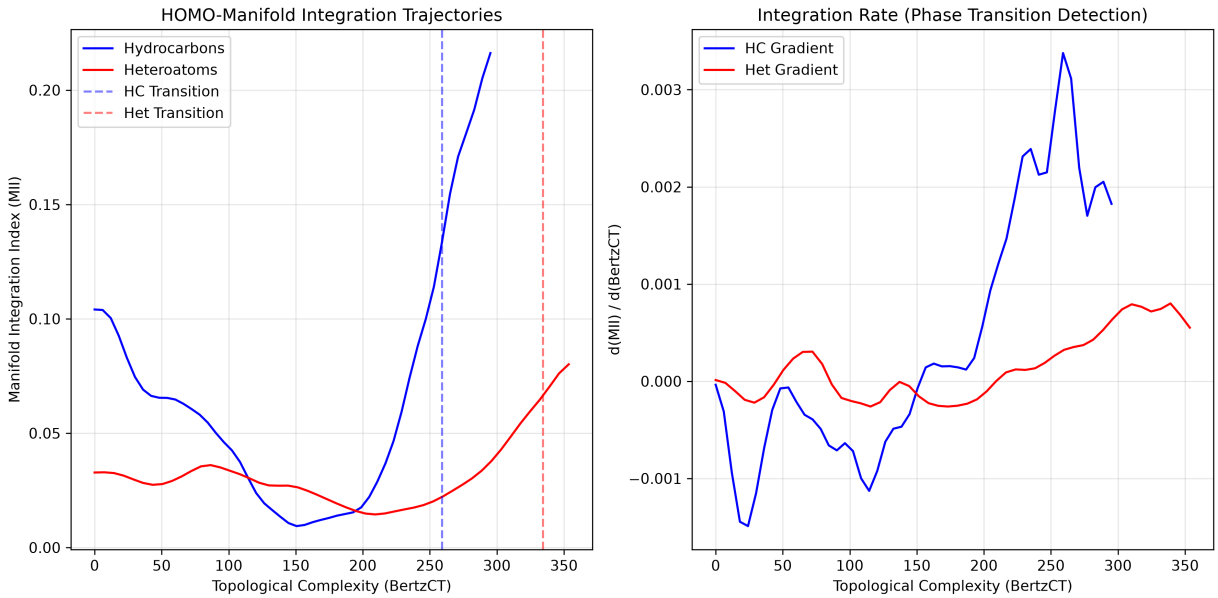

Figure 1: The plot displays the Manifold Integration Index (MII) trajectories (left) and their corresponding integration rates (right) as a function of topological complexity (BertzCT) for pure hydrocarbons and heteroatom-rich molecules. The left panel illustrates that `res_homo` remains largely decoupled from the spatial-thermodynamic manifold until reaching distinct topological thresholds, marked by vertical dashed lines. For hydrocarbons, this integration phase transition occurs at a BertzCT of approximately 259, whereas the inclusion of heteroatoms shifts this threshold to roughly 334. The right panel highlights that the maximum integration rate for heteroatoms is significantly shallower than for hydrocarbons, demonstrating that heteroatoms act as electronic insulators that maintain the isolation of the *homo* energy across a broader range of molecular architectures.

3 Results

3.1 Stoichiometric residualization and the decoupled manifold

Building upon the residualization framework established in the methods, we first evaluated the baseline isolation of the highest occupied molecular orbital (*homo*). By regressing the original properties against exact atom counts, we isolated 10 core chemical features (*res_A_log*, *res_BC_avg*, *res_mu*, *res_alpha*, *res_lumo*, *res_r2*, *res_zpve*, *res_u0*, *res_cv*, *res_u0_atom*) that are strictly decoupled from molecular size. Within this residual manifold, the *homo* remains remarkably isolated across the majority of the population, exhibiting a mean partial distance correlation (PDC) with other residuals of less than 0.1. This confirms that the lack of correlation observed for the *homo* is not merely a byproduct of size scaling, but a fundamental characteristic of frontier electronic states in small organic molecules.

3.2 Topological phase transitions in HOMO integration

To determine where this electronic isolation breaks down, we analyzed the Manifold Integration Index (MII)—defined as the mean PDC of *res_homo* with the spatial-vibrational-rotational manifold (*res_A_log*, *res_BC_avg*, *res_zpve*, *res_cv*, *res_r2*)—across a continuous gradient of topological complexity (*BertzCT*), as shown in Figure 1.

3.2.1 The hydrocarbon baseline

For pure hydrocarbons, the coupling of the *homo* to the broader manifold is initially high in very simple structures, such as acyclic alkanes, but decays as complexity increases (Figure 2). However, as the carbon

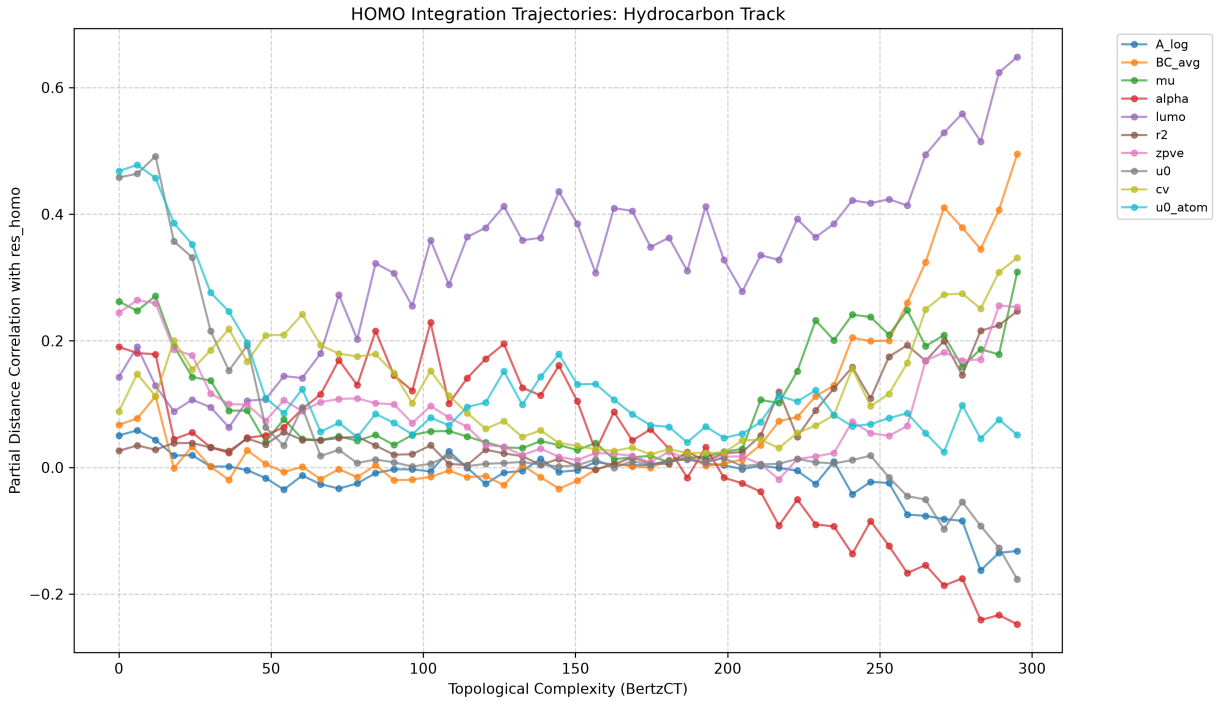

Figure 2: Partial distance correlation (PDC) trajectories of residualized molecular properties with the HOMO energy (`res_homo`) across topological complexity (BertzCT) for pure hydrocarbons. The plot illustrates the breakdown of the HOMO's electronic isolation, revealing a distinct integration threshold at approximately 259 BertzCT. While initial coupling in simple acyclic structures decays with increasing complexity, a sharp rise in correlation with properties such as `res_lumo` and `res_BC_avg` occurs at this threshold. This demonstrates that as hydrocarbon frameworks transition into highly branched or strained cyclic systems, the carbon topology alone becomes sufficient to force the coupling of frontier electronic states with the broader spatial and thermodynamic manifold.

framework transitions into increasingly branched or strained cyclic systems, the trajectory reaches a distinct integration peak. We identify this integration threshold at a $BertzCT$ of 259.02, where the MII reaches 0.1219 with a maximum gradient of 0.0034.

This transition is highly stable across smoothing scales ($\sigma_{stability} = 0.00$). The emergence of this integrated phase at a relatively low complexity level suggests that, in the absence of electronegative atoms, the topology of the carbon framework alone is sufficient to force electronic-spatial coupling.

3.2.2 The heteroatom-rich shift

The inclusion of nitrogen, oxygen, and fluorine fundamentally rewires the integration trajectory, shifting the threshold significantly toward higher topological complexity (Figure 3). For heteroatom-rich molecules, the integration threshold occurs at a $BertzCT$ of 334.32, representing a substantial shift of +75.31 units relative to the hydrocarbon baseline.

At this transition, the MII is 0.0627, and the maximum integration gradient (0.0008) is roughly 4.2 times shallower than that of the hydrocarbon track. Furthermore, this threshold exhibits higher variance ($\sigma_{stability} = 12.26$), indicating a more diverse set of structural pathways to integration. These results demonstrate that heteroatoms act as electronic insulators, maintaining the isolation of the *homo* energy across a much wider range of molecular architectures.

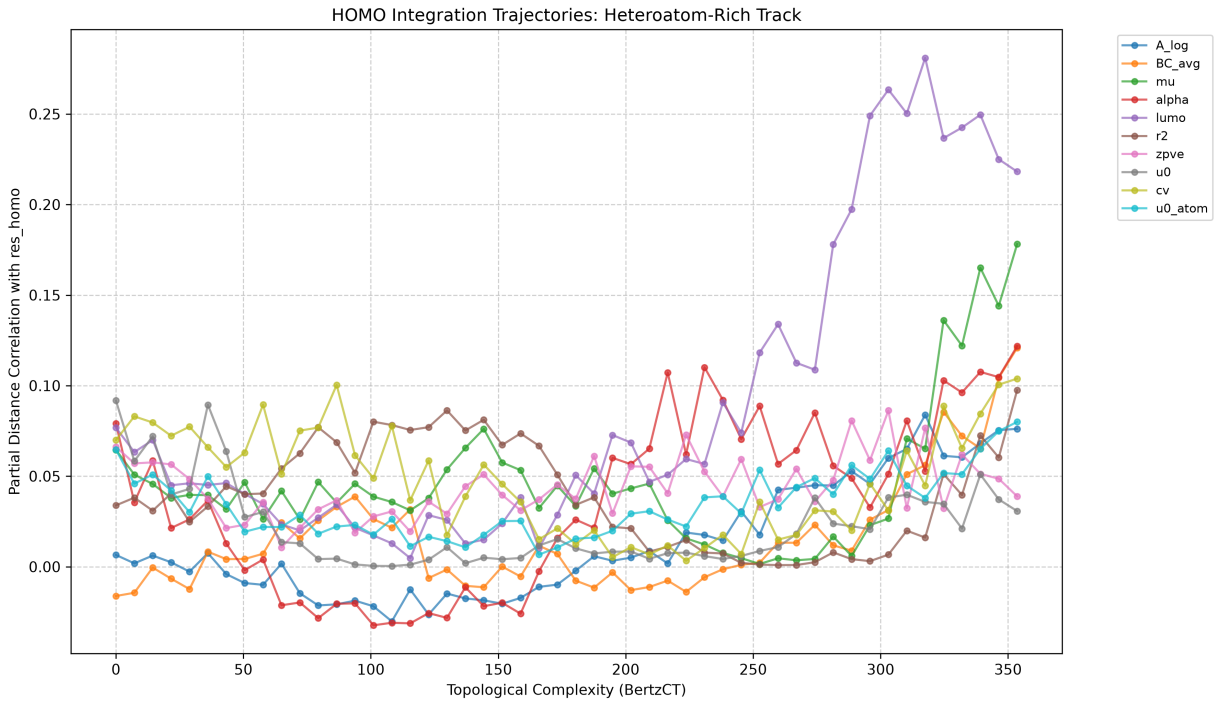

Figure 3: Partial distance correlation (PDC) trajectories of residualized molecular properties with the residualized HOMO energy (`res_homo`) across topological complexity (BertzCT) for heteroatom-rich molecules. The plot illustrates the extended decoupling of the HOMO from the broader spatial and thermodynamic manifold, maintaining a low correlation (PDC < 0.1) across most structural architectures due to the electronic insulating effect of incorporated heteroatoms. Significant coupling, most notably with the residualized LUMO and dipole moment, only emerges at high topological complexities as the system approaches the integration threshold of approximately 334 BertzCT. This delayed integration visually confirms that heteroatom-rich species require highly complex, rigid polycyclic scaffolds to overcome electronic isolation and force frontier electronic states to couple with the global molecular structure.

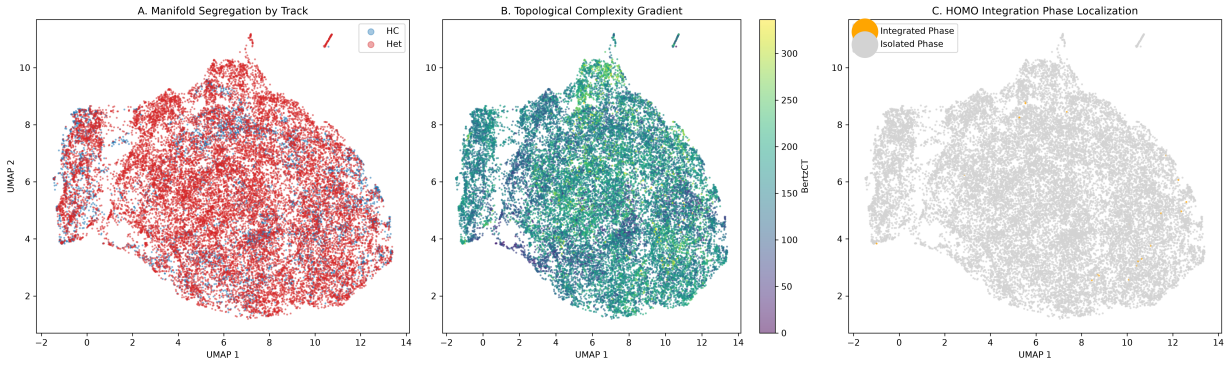

Figure 4: UMAP projection of the residual manifold illustrating the segregation of molecular tracks and the localization of HOMO integration. Panel A demonstrates a clear segregation between pure hydrocarbons (HC) and heteroatom-rich molecules (Het), with a centroid separation distance of 0.2319. Panel B maps the topological complexity gradient (BertzCT) across this space. Panel C highlights the integrated phase molecules, which exceed the topological thresholds of approximately 259 BertzCT for hydrocarbons and 334 BertzCT for heteroatoms. Together, these projections reveal that the coupling of the HOMO to the broader spatial-thermodynamic manifold is not a random occurrence, but is instead localized to specific regions of the topological-chemical space dictated by structural complexity and heteroatom incorporation.

3.3 Rewiring of property dependencies and scaffold interpretations

The specific properties that *res_homo* couples with differ markedly between the two tracks, further highlighting the insulating effect of heteroatoms (Table 1.2).

In pure hydrocarbons, the *homo* and *lumo* residuals are moderately coupled (mean PDC ≈ 0.2982), suggesting a shared dependency on the π-system of the carbon skeleton. However, in heteroatom-rich molecules, this coupling collapses to 0.0897. In these systems, the *homo* is uniquely sensitive to local lone-pair environments or specific heteroatom placements that do not affect the *lumo* or the macroscopic spatial extent to the same degree.

These mathematical thresholds correspond directly to distinct chemical motifs. The hydrocarbon integration threshold ($BertzCT \approx 259$) is dominated by acyclic branched structures and simple small-ring fusions, where the spatial extent and vibrational modes become sufficiently constrained that frontier orbitals must respond to the global geometry. Conversely, the heteroatom integration threshold ($BertzCT \approx 334$) is characterized by complex bicyclic heterocycles, such as `N=c1[nH]c2ccoc2[nH]1` and `N=c1oc2ncoc2o1`. These highly substituted, fused-ring systems represent the critical level of rigid polycyclic architecture required to force the localized electronic state to interact with the broader molecular manifold.

3.4 Manifold segregation and spatial localization

To visually confirm the global structure of the residual manifold and the localization of these phase transitions, we projected the residualized molecular data onto a UMAP space (Figure 4).

The projection reveals a clear segregation between the two analytical tracks, with a Euclidean distance of 0.2319 between the centroids of the hydrocarbon and heteroatom manifolds. Crucially, the integrated phase molecules—those exceeding their respective $BertzCT$ thresholds—occupy a restricted, localized region of the UMAP space, particularly within the heteroatom track. This spatial localization confirms that the integration of the *homo* is not a random occurrence, but is strictly dictated by specific, highly complex regions of the topological-chemical space.

4 Conclusions

The apparent orthogonality of the highest occupied molecular orbital (*homo*) to the broader spatial and thermodynamic manifold is not a universal constant of small organic molecules, but rather a complexity-

dependent phenomenon. By employing stoichiometric residualization to strictly eliminate molecular size effects, we isolated the genuine structural dependencies of frontier electronic states. Tracking partial distance correlation trajectories across gradients of topological complexity revealed that the *homo* remains effectively decoupled in the majority of simple molecular architectures. However, under specific structural constraints, this electronic isolation breaks down, forcing the frontier orbitals to integrate with the molecule's joint physical structure.

Our conditional dependency analysis identified distinct topological phase transitions that govern this integration, which bifurcate sharply based on atomic composition. For pure hydrocarbons, integration occurs at a relatively low complexity threshold of approximately 259 $BertzCT$. In the absence of electronegative atoms, the spatial confinement introduced by acyclic branching and simple ring fusions is sufficient to couple the *homo* to the global molecular geometry. Conversely, the introduction of heteroatoms fundamentally rewires these property dependencies. Heteroatoms act as potent electronic insulators, localizing the frontier orbitals and delaying their integration into the global manifold. This insulating effect shifts the integration threshold to approximately 334 $BertzCT$.

This shift of over 75 $BertzCT$ units represents the "electronic cost" of heteroatom incorporation. It demonstrates that heteroatom-rich molecules require significantly more complex, rigid polycyclic architectures—such as highly substituted, fused bicyclic heterocycles—to overcome electronic isolation. Furthermore, the nature of the integration itself is altered; while the *homo* and *lumo* share a moderate dependency on the carbon skeleton's π-system in pure hydrocarbons, this coupling collapses in heteroatom-rich species, where the *homo* becomes uniquely sensitive to local lone-pair environments rather than the macroscopic spatial extent.

Ultimately, the spatial localization of these integrated phases within the residual manifold confirms that HOMO integration is a highly specific, structurally deterministic event. By mapping the exact topological thresholds where localized electronic states are forced to interact with the macroscopic constraints of the molecule, this framework provides a quantitative basis for understanding frontier orbital behavior. These insights offer critical guidelines for rational molecular design, highlighting that achieving electronic-spatial coupling in heteroatom-rich compounds demands specific, rigid polycyclic scaffolds.