Heteroatom-Driven Rewiring of the Joint Molecular Property Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular properties are typically dominated by size and stoichiometry, masking how specific chemical architectures and heteroatoms dictate the relationships between electronic and thermodynamic states. A key anomaly in this space is the decoupling of the highest occupied molecular orbital (HOMO) energy from the broader molecular property network. To separate structural effects from simple atomic counting, we analyzed the QM9 dataset by calculating non-linear, gradient-boosted atom-count residuals for twelve key properties. We stratified the molecules into pure hydrocarbons and distinct nitrogen-, oxygen-, and fluorine-containing families, applying distance correlation networks, hub-scoring, and differential rewiring analysis to track changes in property relationships. We found that while thermodynamic and geometric properties form a highly correlated core in pure hydrocarbons, HOMO energy remains isolated, driven instead by local bonding environments. Introducing heteroatoms fundamentally rewires this network: nitrogen breaks the standard geometric scaling of polarizability, while oxygen decouples the lowest unoccupied molecular orbital (LUMO) from the molecule's vibrational state. Robustness checks reveal that although some spatial correlation shifts are artifacts of linear molecular symmetries, the reorganization of frontier orbitals is a genuine effect of heteroatom substitution. Finally, network hub analysis shows that the central organizing property shifts from polarizability in hydrocarbons to spatial extent and mass distribution in heteroatom systems, with terminal polar groups like aldehydes driving the most significant deviations from the standard property manifold.

1 Introduction

The exploration of chemical space frequently reveals that macroscopic molecular properties are heavily dominated by size and stoichiometry. While these fundamental scaling laws provide a robust baseline for predicting thermodynamic and electronic behaviors, they simultaneously mask the more subtle ways in which specific chemical architectures and heteroatom substitutions dictate the relationships between different molecular states. Within the joint manifold of molecular properties, a key anomaly is the apparent decoupling of the highest occupied molecular orbital (HOMO) energy from the broader spatial and thermodynamic network. In pure hydrocarbons, thermodynamic properties (such as internal energy and heat capacity) and geometric response properties (such as polarizability and spatial extent) form a tightly bound, highly correlated core. However, the HOMO energy remains largely isolated. It remains an open question whether this decoupling is a universal physical constant of the molecular property manifold or a conditional state that undergoes rewiring upon the introduction of highly electronegative atoms.

To isolate the influence of chemical architecture from stoichiometric determinism, we investigate how specific lone-pair-bearing heteroatoms (nitrogen, oxygen, and fluorine) modulate the integration of frontier orbitals into the joint molecular state. Using the QM9 dataset, we extract non-linear, gradient-boosted atom-count residuals for twelve non-redundant properties. This residualization strategy effectively removes the variance explained by simple atomic tallies, allowing us to focus purely on structural and topological drivers. We stratify the dataset into pure hydrocarbons and distinct heteroatom-containing families, applying distance correlation networks, systematic hub-scoring, and differential rewiring analysis to quantify shifts in multivariate coupling strengths.

Our analysis reveals that the introduction of heteroatoms fundamentally rewires the molecular property network. We demonstrate that nitrogen breaks the standard geometric scaling of polarizability observed in

hydrocarbons, while oxygen decouples the lowest unoccupied molecular orbital (LUMO) from the molecule's vibrational state. Through rigorous linear sensitivity checks, we confirm that while some spatial correlation shifts are artifacts of linear molecular symmetries, the internal reorganization of frontier orbitals is a genuine, robust consequence of heteroatom substitution.

Furthermore, network hub analysis indicates a shift in the central organizing property from polarizability in pure hydrocarbons to spatial extent and mass distribution in heteroatom systems. By grounding these statistical shifts in chemical structure, we identify that terminal polar groups drive the most significant deviations from the standard property manifold, forcibly entangling localized frontier orbitals with the macroscopic geometric landscape of the molecule.

2 Methods

2.1 Data initialization and epsilon-transformation

To investigate the joint molecular property manifold, we utilized the QM9 dataset, which was first deduplicated based on canonical SMILES strings. We isolated twelve non-redundant molecular properties for analysis: zero-point vibrational energy ('zpve'), internal energy at 0 K ('u0'), heat capacity ('cv'), rotational constants ('A', 'B', 'C'), dipole moment ('mu'), isotropic polarizability ('alpha'), highest occupied molecular orbital energy ('homo'), lowest unoccupied molecular orbital energy ('lumo'), electronic spatial extent ('r2'), and atomization energy at 0 K ('u0_atom').

Linear molecules possess undefined rotational constants along the molecular axis, which can distort variance calculations. To retain these molecules without introducing numerical instability, we applied a logarithmic transformation to the first rotational constant across the entire dataset: $\log(A + \epsilon)$, where $\epsilon = 10^{-3}$. This epsilon-transformation maintains the distinction of linear geometries while stabilizing the global variance.

2.2 Compositional residualization

To strictly isolate the effects of chemical architecture from stoichiometric determinism, we transformed the raw property values into a non-linear residualized space. For each of the twelve properties, we trained a histogram-based gradient boosting regressor ('HistGradientBoostingRegressor') on the full dataset. The feature space for these models consisted exclusively of exact atomic counts (C, N, O, F, and H), the number of heavy atoms ('n_heavy'), and the number of rings ('n_rings'). By extracting the residuals from these models, we ensured that the remaining variance in the property space was driven entirely by structural arrangement and topological features rather than simple atomic tallies.

2.3 Stratification and strict subsampling control

Following residualization, we parsed the SMILES strings to partition the dataset into mutually exclusive compositional strata: pure acyclic hydrocarbons, pure cyclic hydrocarbons, nitrogen-only (N-only), oxygen-only (O-only), and fluorine-only (F-only) heteroatom systems.

Because distance correlation metrics are highly sensitive to sample size variations, we implemented a strict subsampling control. We determined the population size of the smallest stratum (F-only) and randomly subsampled all other strata to exactly match this integer. This procedure eliminated statistical artifacts that could arise from comparing networks of vastly different sizes.

2.4 Distance correlation and differential rewiring extraction

For each size-controlled stratum, we computed the pairwise distance correlation ($dCor$) matrix for the twelve residualized properties. Distance correlation was chosen over Pearson correlation to capture both linear and non-linear dependencies between properties. These matrices were subsequently converted into distance matrices using the transformation $1 - dCor$.

To quantify the heteroatom-induced reorganization of the property manifold, we performed a differential network analysis. We defined the pure cyclic hydrocarbon matrix as the baseline and subtracted it from

Functional Group	High Anomaly (Top 50)	Low Anomaly (Bottom 50)
Ether ($R-O-R$)	31	42
Alcohol ($-OH$)	23	20
Ketone ($C=O$)	1	11
Aldehyde ($-CHO$)	9	2
Epoxide	7	10

the N-only and O-only matrices to generate differential Rewiring Matrices. We then isolated the top 5% of absolute shifts in $dCor$. This thresholding isolated the specific property-pair relationships that were most significantly created or destroyed by the introduction of a given heteroatom.

2.5 Systematic hub search

To identify which properties act as structural mediators within the manifold, we executed a full partial-correlation scan. For each stratum, we computed the partial distance correlation of the HOMO residual (‘res_homo‘) with all other properties, iteratively conditioning on every other property in the network.

We defined a "Hubness Score" for each property to quantify its mediating influence. The score was calculated as the median magnitude of the reduction in correlation across the network when conditioning on that variable:

$$\Delta dCor = |dCor_{\text{unconditioned}} - dCor_{\text{conditioned}}|$$

Using the median magnitude prevents isolated outlier relationships from skewing the assessment of a property's global centrality. The property with the highest Hubness Score in each stratum was identified as the primary structural hub, allowing us to track how the central organizing property shifts from hydrocarbons to heteroatom systems.

2.6 Linear sensitivity check

To ensure that the observed rewiring of the HOMO energy was a fundamental chemical phenomenon rather than a geometric artifact, we performed a binary sensitivity check. We recomputed the Rewiring Matrices twice: once including the epsilon-transformed linear molecules, and once excluding them entirely. We then compared the top 5% of rewired edges between the two outputs. This robustness check allowed us to distinguish genuine electronic reorganizations of the frontier orbitals from spatial correlation shifts driven by high-symmetry linear geometries.

2.7 Chemical grounding of outliers

To physically ground the statistical rewiring, we analyzed the specific molecules driving the deviations in the O-only stratum. We calculated an anomaly score for each molecule, defined as the Euclidean distance from the manifold centroid in the residual space of the tightly bound spatial-thermodynamic core (res_r^2, res_C, res_zpve, res_cv).

We extracted the SMILES strings for the molecules exhibiting the highest (top 50) and lowest (bottom 50) anomaly scores. These outliers were then categorized by their dominant functional groups to link the abstract joint structure directly to localized bonding environments. The distribution of these functional groups among the most and least anomalous molecules is detailed in Table 2.7.

3 Results

3.1 Global property manifold and stoichiometric decoupling

The joint structure of the QM9 dataset exhibits a clear hierarchical organization when the variance explained by simple atomic tallies is removed. In pure hydrocarbons, thermodynamic properties (such as internal energy u_0, zero-point vibrational energy (ZPVE), and heat capacity c_v) and geometric response properties (such as isotropic polarizability α and electronic spatial extent r^2) form a tightly bound, highly correlated

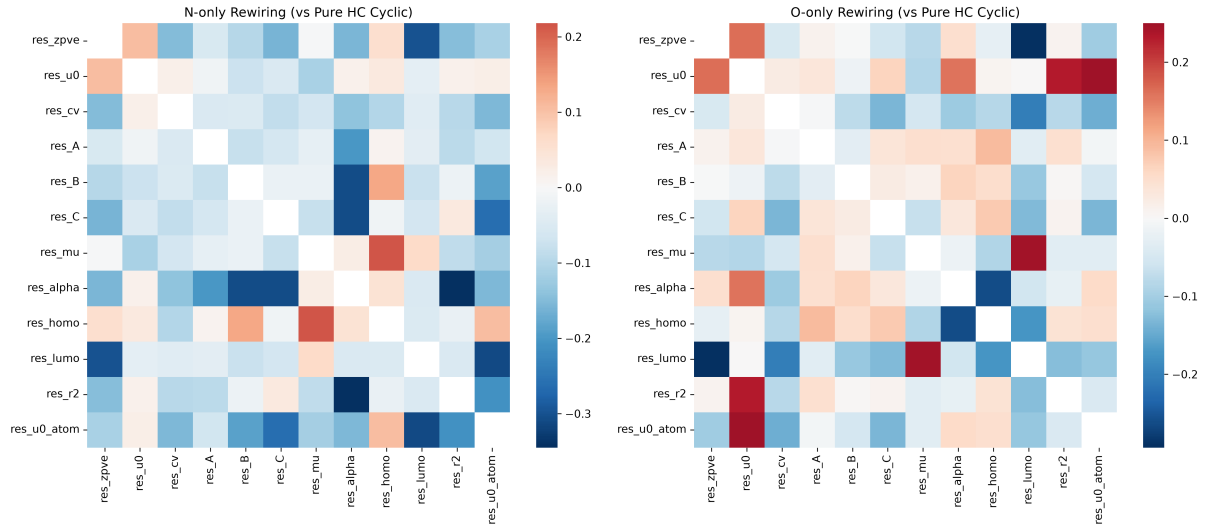

Figure 1: Heatmaps illustrating the systematic rewiring of the joint molecular property manifold upon heteroatom substitution, showing the shift in distance correlation ($dCor$) for property residuals in nitrogen-only (left) and oxygen-only (right) strata compared to pure cyclic hydrocarbons. Negative values (blue) highlight the dissolution of standard hydrocarbon couplings. In the nitrogen-only stratum, the strong negative shifts between polarizability (res_α) and both spatial extent (res_{r^2}) and rotational constants (e.g., res_C) demonstrate that nitrogen breaks the simple geometric scaling of polarizability found in pure hydrocarbons. In the oxygen-only stratum, a significant negative shift between LUMO energy (res_{LUMO}) and zero-point vibrational energy (res_{ZPVE}) indicates a decoupling of the unoccupied orbital landscape from the vibrational manifold. Overall, the visualizations reveal that heteroatom introduction fundamentally reorganizes the relationships between electronic, thermodynamic, and geometric properties into new, composition-specific dependencies.

core. Within this baseline manifold, spatial extent acts as a primary structural hub. For instance, in cyclic hydrocarbons, the residual spatial extent (res_{r^2}) mediates the relationships between electronic extent and rotational constants with a high Hubness Score of 0.252.

In contrast, the highest occupied molecular orbital (HOMO) energy residuals (res_{HOMO}) exhibit a unique decoupling from this core. Unlike the thermodynamic and geometric properties, res_{HOMO} shows significantly lower integration into the global network across all compositional strata. Its Hubness Scores remain consistently among the lowest (e.g., 0.204 in the nitrogen-only stratum and 0.262 in acyclic hydrocarbons). This isolation suggests that the HOMO energy is governed by localized bonding environments that are largely orthogonal to the molecule's macroscopic spatial and vibrational manifold.

3.2 Heteroatom-induced rewiring

The introduction of lone-pair-bearing heteroatoms, specifically nitrogen and oxygen, induces a systematic rewiring of these property relationships. This reorganization is characterized by the dissolution of standard hydrocarbon couplings and the emergence of new, composition-specific dependencies.

As illustrated in Figure 1, the transition from cyclic hydrocarbons to nitrogen-containing molecules is marked by a sharp decrease in the coupling between isotropic polarizability and spatial extent. The shift in distance correlation ($dCor$) for the pair (res_α, res_{r^2}) is -0.345, dropping from a baseline $dCor$ of 0.645 in cyclic hydrocarbons to 0.299 in the nitrogen-only stratum. Similarly, the relationship between rotational constants and polarizability weakens significantly (e.g., a shift of -0.308 for res_C versus res_α). This indicates that nitrogen atoms break the simple geometric scaling of polarizability found in pure hydrocarbons, likely through the introduction of localized lone pairs and more complex electronic distributions.

In the oxygen-only stratum, heteroatom substitution significantly decouples the lowest unoccupied molec-

ular orbital (LUMO) energy from the vibrational manifold. The relationship between res_{ZPVE} and res_{LUMO} exhibits a negative shift of -0.294 (from a baseline $dCor$ of 0.435 to 0.141). This demonstrates that in oxygenated systems, the unoccupied orbital landscape becomes substantially less sensitive to the zero-point vibrational state of the molecule than it is in pure hydrocarbons.

3.3 Robustness and geometric artifacts

To determine whether these rewiring effects were genuine electronic reorganizations or merely artifacts of high-symmetry linear molecules (which possess undefined rotational constants along the molecular axis), we performed a strict linear sensitivity check. Excluding 129 linear and extreme molecules from the dataset revealed that many of the initial spatial correlation shifts were indeed driven by geometric outliers; the robustness score for the top 5% of rewired edges was only 12.5%.

However, specific electronic relationships emerged as highly robust within the non-linear subset. Most notably, the coupling between the HOMO and LUMO residuals (res_{HOMO} versus res_{LUMO}) showed a strong negative shift upon heteroatom introduction: -0.277 for nitrogen and -0.389 for oxygen. This indicates that while the global spatial manifold is highly sensitive to linear geometry artifacts, the internal reorganization and decoupling of the frontier orbital levels is a fundamental and robust consequence of heteroatom substitution.

3.4 Structural mediators and hubness shifts

Beyond pairwise rewiring, the global role of specific properties as structural mediators (hubs) shifts dramatically with chemical composition.

As shown in Figure 2, in acyclic hydrocarbons, polarizability (res_{α}) acts as the dominant hub, boasting a Hubness Score of 0.319. However, upon the introduction of nitrogen and oxygen, the network reorganizes, and electronic spatial extent (res_{r^2}) takes over as the primary hub (Hubness Scores of 0.217 and 0.246, respectively). In the fluorine-only stratum, the network undergoes an even more extreme reorganization: the third rotational constant (res_C) becomes the primary hub (Hubness Score of 0.263). This suggests that for highly electronegative substituents, the mass distribution and overall geometry become the central organizing factors for the residual property manifold.

3.5 Chemical grounding of manifold deviations

To physically ground these statistical shifts, we analyzed the specific functional groups driving manifold deviations within the oxygen-only stratum (comprising 45,771 molecules). We calculated an anomaly score for each molecule, defined as the Euclidean distance from the manifold centroid in the residual space of the tightly bound core (res_{r^2}, res_C, res_{ZPVE}, res_{c_v}).

Figure 3 and Table 2.7 detail the distribution of functional groups among the most and least anomalous molecules. Molecules with high anomaly scores—those that most strongly violate the expected joint structure—exhibit a 4.5-fold increase in the frequency of aldehydes (9 in the top 50) compared to low-anomaly molecules (2 in the bottom 50). Conversely, ketones (1 vs. 11) and esters (1 vs. 6) are significantly more prevalent in molecules that adhere strictly to the global manifold. This structural dichotomy reveals that terminal, highly polar oxygen groups (like aldehydes) are the primary drivers of property rewiring. In contrast, internal carbonyl groups (like ketones) are more easily integrated into the standard spatial-thermodynamic manifold without causing severe structural deviations.

4 Conclusions

This study demonstrates that the joint manifold of molecular properties is not a static, universally scaled network, but rather a dynamic structure that is fundamentally rewired by chemical architecture and heteroatom substitution. By isolating non-linear, atom-count residuals, we successfully separated topological and structural effects from stoichiometric determinism. Our analysis confirms that in pure hydrocarbons, thermodynamic and geometric properties form a tightly integrated core, whereas the highest occupied molecular

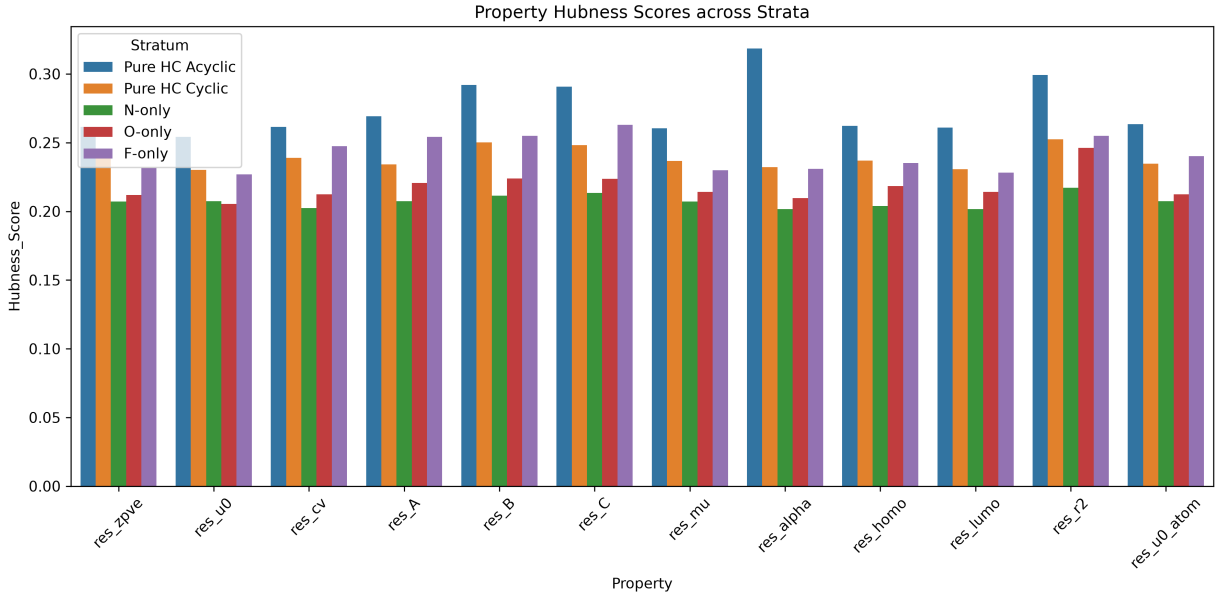

Figure 2: Property Hubness Scores for molecular property residuals across different compositional strata, illustrating the heteroatom-induced rewiring of the joint property manifold. The bar chart demonstrates how the role of specific properties as structural mediators shifts with chemical composition. In pure acyclic hydrocarbons, residual polarizability (res_α) acts as the dominant hub. Upon heteroatom substitution, the network reorganizes: residual spatial extent (res_{r^2}) becomes the primary hub for nitrogen- and oxygen-only strata, whereas the third rotational constant (res_C) dominates in fluorine-containing molecules, indicating that mass distribution and geometry become the central organizing factors for highly electronegative substituents. Additionally, HOMO energy residuals (res_{HOMO}) exhibit consistently low hubness scores across all strata, highlighting the unique decoupling of the highest occupied orbital energy from the global spatial and vibrational manifold.

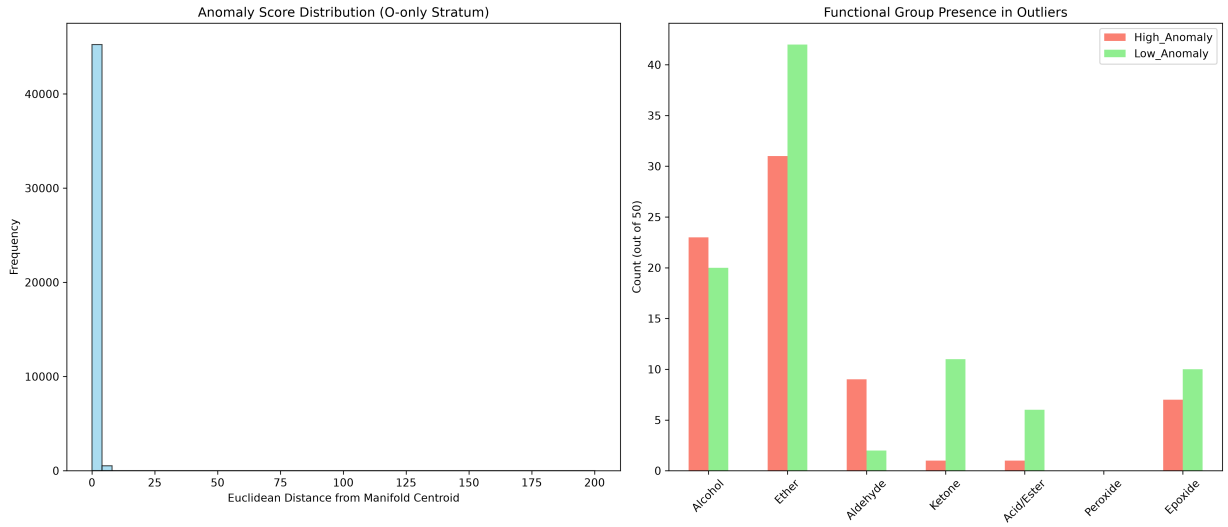

Figure 3: Distribution of anomaly scores and functional group drivers for the oxygen-only stratum. The left panel shows the frequency of anomaly scores, calculated as the Euclidean distance from the manifold centroid in the residual space (res_{r^2}, res_C, res_{ZPVE}, res_{c_v}). The right panel compares functional group counts between the top 50 high-anomaly and bottom 50 low-anomaly molecules. We observe a 4.5-fold higher prevalence of aldehydes in the high-anomaly subset, whereas ketones and esters are more common in the low-anomaly subset. This reveals that terminal, highly polar oxygen groups primarily drive property rewiring by violating the expected joint structure, while internal carbonyl groups integrate more readily into the standard spatial-thermodynamic manifold.

orbital (HOMO) energy remains uniquely decoupled, governed instead by localized bonding environments that are largely orthogonal to the macroscopic spatial and vibrational manifold.

The introduction of highly electronegative heteroatoms disrupts this baseline structure, dissolving standard hydrocarbon couplings and establishing new, composition-specific dependencies. Specifically, nitrogen substitution breaks the geometric scaling of polarizability, while oxygen decouples the lowest unoccupied molecular orbital (LUMO) from the molecule's zero-point vibrational state. Crucially, our linear sensitivity checks reveal that while some spatial correlation shifts are artifacts of high-symmetry linear geometries, the internal reorganization of frontier orbitals—particularly the decoupling between HOMO and LUMO energies—is a robust and genuine consequence of heteroatom introduction.

Furthermore, our network hub analysis highlights a profound shift in the central organizing properties of these molecules. The dominant structural mediator transitions from polarizability in acyclic hydrocarbons to electronic spatial extent in nitrogen and oxygen systems, and ultimately to mass distribution (the third rotational constant) in fluorine-containing molecules. This indicates that as substituent electronegativity increases, macroscopic geometry and mass distribution become the primary organizing factors of the residual property space.

Finally, by grounding these statistical deviations in specific chemical structures, we identified the localized drivers of manifold rewiring. As detailed in Table 2.7, terminal polar groups such as aldehydes are disproportionately represented among the most anomalous molecules in the oxygen-only stratum, whereas internal carbonyls like ketones are more easily integrated into the standard spatial-thermodynamic core.

Ultimately, these findings reveal that the decoupling of frontier orbitals is not a universal constant, but a conditional state modulated by specific functional groups. Understanding these heteroatom-driven topological shifts provides a more nuanced view of chemical space, highlighting how localized electronic perturbations forcibly entangle with or decouple from a molecule's macroscopic geometric landscape.