Topological Re-coupling of Frontier Orbitals via Conjugation and Aromaticity Gradients

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The anomalous decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from broader spatial and geometric properties presents a significant challenge in understanding molecular property manifolds. To determine whether this isolation is an inherent orbital feature or a state-dependent structural constraint, we investigate the topological conditions under which this decoupling physically breaks down. Using 133,802 deduplicated molecules from the QM9 dataset, we extract size-conditioned residuals for frontier orbitals and geometric response variables via exact atom count regression to strictly isolate chemical architecture from stoichiometric determinism. We then stratify the dataset along semi-continuous gradients of conjugation length and aromaticity, evaluating joint structural coupling through Canonical Correlation Analysis (CCA) and Distance Correlation metrics validated against exact-count null distributions. Our analysis reveals that increased electronic delocalization forcefully re-integrates both the HOMO and LUMO into the global geometric state space. Specifically, the primary canonical correlation between frontier orbitals and geometric responses increases from 0.536 in short-conjugated molecules to 0.627 in long-conjugated systems, and reaches 0.681 in aromatic molecules where the coupling manifold collapses into a single dominant dimension. Furthermore, heteroatom identity—particularly fluorine and aromatic nitrogen—strongly modulates this rewiring. Ultimately, these findings demonstrate that frontier orbital decoupling is a feature of localized, aliphatic systems, whereas extended π-systems enforce a globalized, low-dimensional integration of electronic and spatial properties.

1 Introduction

1.1 The frontier orbital decoupling anomaly

Understanding the joint distribution of electronic and geometric properties is a central challenge in mapping molecular property manifolds. In large-scale chemical spaces, macroscopic physical properties often exhibit strong collinearity, largely driven by extensive scaling with molecular size and heavy atom count. However, within datasets such as QM9, the Highest Occupied Molecular Orbital (HOMO) energy exhibits an anomalous global decoupling from the broader spatial and geometric property manifold. While properties such as polarizability (α), spatial extent (R^2), and dipole moment (μ) share significant mutual information, the HOMO energy appears isolated, varying independently of these global geometric constraints.

This isolation presents a significant conceptual and practical hurdle. It remains unclear whether this decoupling is an inherent quantum mechanical feature of the HOMO itself, or if it is a state-dependent structural constraint reflecting the specific compositional bias of the chemical space being studied. Because large enumerative datasets are overwhelmingly dominated by small, localized, and aliphatic molecules, the observed global decoupling may simply be a macroscopic artifact of localized bonding environments rather than a universal physical rule.

1.2 Disentangling size from topology

To rigorously determine under what structural conditions this decoupling physically breaks down, it is necessary to isolate the effects of chemical architecture from stoichiometric determinism. Molecular size and exact atom counts dictate the vast majority of variance in extensive properties; for instance, stoichiometry alone determines up to 99.99% of total energy variance and 86% of polarizability variance.

To strictly isolate topological effects, we extract size-conditioned residuals for the frontier orbitals (*res_homo*, *res_lumo*) and the geometric response variables (*res_*μ, *res_*R^2, *res_*α). By regressing these target properties against exact atom counts (C, N, O, F, H), we remove the deterministic baseline of molecular size. The resulting residual space represents the variance uniquely attributable to molecular topology, bonding architecture, and electron delocalization, providing a rigorous foundation for multivariate analysis.

1.3 Topological gradients and joint structure

We hypothesize that extended electron delocalization forcefully re-integrates the frontier orbitals into the molecule's global geometric state space. Rather than relying on heterogeneous functional group classifications, we stratify 133,802 deduplicated molecules from the QM9 dataset along semi-continuous topological gradients: maximum conjugation length and aromatic ring fraction.

Across these stratified bins, we evaluate the joint structural coupling between the frontier orbital pair and the geometric-response group. We employ Canonical Correlation Analysis (CCA) to capture linear manifold alignments and Distance Correlation metrics to track non-linear dependencies. To rigorously distinguish genuine electronic rewiring from the inherently high variance of frontier orbital residuals, all observed coupling shifts are validated against exact-count null distributions generated by permuting residuals strictly within highly granular strata of identical atom counts.

1.4 Scope and implications

In this work, we map the divergence and convergence of HOMO and LUMO coupling trajectories as conjugation and aromaticity increase. We demonstrate that the anomalous decoupling of the HOMO is not a universal feature, but rather a characteristic of localized, aliphatic systems. As the π-system extends, the frontier orbitals undergo a fundamental reorganization, re-coupling with the geometric manifold. Furthermore, we show that aromaticity induces a phase transition in this coupling, collapsing the joint property manifold into a single dominant dimension, an effect that is strongly modulated by heteroatom identity and position. Ultimately, these findings reveal exactly how macroscopic structural delocalization controls the integration of frontier orbitals into global physical properties, suggesting that future predictive models for quantum properties must account for these distinct topological regimes.

2 Methods

2.1 Data curation and preprocessing

To investigate the structural conditions under which frontier orbitals integrate into the global geometric state space, we utilized the QM9 dataset. Initial preprocessing involved deduplicating the dataset based on SMILES strings to prevent data leakage from stereochemical duplicates, yielding a final working set of 133,802 molecules. Molecular structures were parsed using RDKit, from which exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted.

To prevent manifold distortion in subsequent multivariate analyses, we identified and flagged perfectly linear molecules—defined geometrically as having a rotational constant A equal to zero (or approaching infinity) and B equal to C. These 9 linear molecules were explicitly excluded from the regression training and coupling analyses. Furthermore, redundant thermodynamic columns and the deterministic HOMO-LUMO gap column were removed to focus strictly on the primary electronic and spatial variables.

2.2 Size-conditioning and residual extraction

As established, molecular size and exact atom counts dictate the vast majority of variance in extensive properties. To strictly isolate chemical architecture from this stoichiometric determinism, we regressed the target properties—HOMO energy (*homo*), LUMO energy (*lumo*), dipole moment (μ), spatial extent (R^2), and isotropic polarizability (α)—against the exact atom counts (C, N, O, F, H).

Linear regression models were trained for each property, explicitly excluding the flagged linear molecules to avoid skewing the global fit. We then extracted the residuals from these models, yielding a size-conditioned

feature space: res_homo, res_lumo, res_μ, res_R^2, and res_α. These residuals represent the variance uniquely attributable to molecular topology and bonding architecture, forming the basis for all subsequent multivariate analyses.

2.3 Topological stratification and sparsity control

To map the divergence and convergence of coupling trajectories, the dataset was stratified along two semi-continuous topological gradients: conjugation length and aromaticity. Using RDKit, conjugation length was quantified by counting the total number of conjugated bonds within each molecule. The dataset was binned into three conjugation tiers: "Short" (0–3 conjugated bonds), "Medium" (4–6 conjugated bonds), and "Long" ($\geq$ 7 conjugated bonds). Aromaticity was treated as a binary phase transition, categorizing molecules as either aliphatic (zero aromatic rings) or aromatic (at least one aromatic ring).

To rigorously control for structural sparsity bias and ensure that observed shifts in coupling were not artifacts of differing molecular size distributions within these bins, we performed stratified subsampling. Specifically, the "Short" and "Medium" conjugation bins were subsampled such that their heavy-atom count distributions exactly matched the heavy-atom count distribution of the "Long" conjugation bin.

2.4 Multivariate coupling analysis

We evaluated the joint structural coupling between the frontier orbital pair (res_homo, res_lumo) and the geometric-response group ($res_\mu, res_R^2, res_\alpha$) across the matched topological bins. Canonical Correlation Analysis (CCA) was employed to capture the linear alignment of the frontier-geometric manifold, yielding canonical correlations (CC_1, CC_2) that quantify the dimensionality and strength of the joint structure.

To ensure that complex, non-linear dependencies were not obscured by the linear constraints of CCA, we concurrently computed Distance Correlation ($dCor$) metrics between the individual frontier residuals and the geometric residuals. Computations were parallelized across the stratified bins to efficiently handle the dataset scale.

2.5 Null distribution testing

To rigorously distinguish genuine electronic rewiring from the inherently high variance of frontier orbital residuals, all observed $dCor$ and CCA trajectories were validated against a statistical null distribution. This null distribution was generated by repeatedly permuting the residuals strictly within highly granular strata of identical exact atom counts. By restricting permutations to molecules with the exact same number of C, N, O, F, and H atoms, we preserved the exact heteroatom distribution of the dataset while destroying any true topological coupling. The observed coupling metrics were then compared against the mean and variance of this null distribution to confirm the physical significance of the structural rewiring.

2.6 Manifold visualization and heteroatom analysis

To visualize the phase transition induced by aromaticity, we isolated the highest aromaticity bin and the zero-aromaticity (aliphatic) bin. A single Uniform Manifold Approximation and Projection (UMAP) embedding was fit on the combined subset to create a consistent coordinate system. Both bins were projected into this shared 5-dimensional residual space and colored by secondary variables, such as heteroatom count, to verify that manifold clustering was driven by aromaticity rather than secondary compositional effects.

Furthermore, we investigated the mechanism by which specific heteroatoms modulate frontier coupling. Using RDKit, oxygen, nitrogen, and fluorine atoms were classified based on their structural context, distinguishing between endocyclic heteroatoms (those participating in an aromatic ring) and exocyclic/aliphatic heteroatoms. CCA and $dCor$ metrics were recomputed for these specific sub-populations within the highly conjugated bins to determine whether spatial re-integration is a global effect of π-system participation or a localized electronegativity effect.

2.7 Cross-scaffold validation

Finally, to audit the robustness of the observed coupling trends and ensure they generalize across diverse molecular families, we executed a cross-scaffold validation using Murcko scaffolds. The dataset was partitioned using a 5-fold cross-validation scheme based on scaffold identity. Additionally, a leave-one-scaffold-out (LOSO) validation was performed on the top 10 most frequent scaffolds (including common structures like acyclic chains, cyclopropane, and oxirane). This rigorous validation ensures that the observed re-coupling of frontier orbitals is an intrinsic feature of delocalized electronic systems rather than a statistical artifact driven by a few overrepresented molecular geometries.

3 Results

3.1 Baseline isolation via size-conditioning

To ensure that our analysis captures genuine topological effects rather than extensive size scaling, we first evaluated the baseline variance explained by exact atom counts. As established in the methodology, molecular size and stoichiometry dictate the vast majority of variance in extensive properties. As demonstrated in , stoichiometry alone dictates approximately 86% of the variance in polarizability, yet only 8.5% of the HOMO energy variance. By extracting the size-conditioned residuals ($res_$homo, $res_$lumo, res_μ, res_R^2, res_α), we successfully isolated the variance uniquely attributable to molecular architecture, providing a rigorous foundation for evaluating structural coupling.

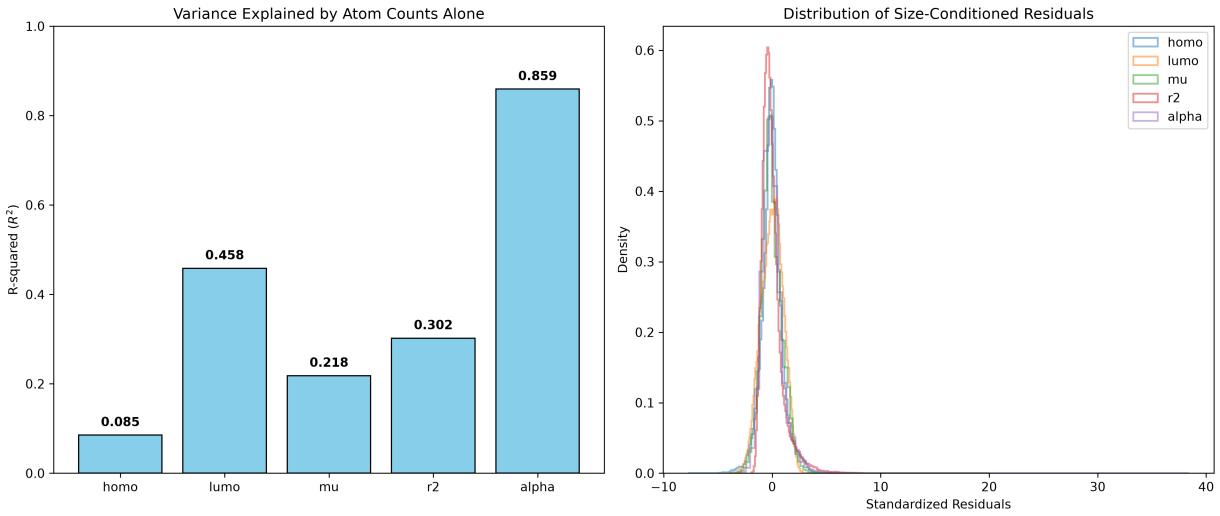

Figure 1: The left panel illustrates the variance (R^2) of molecular properties explained solely by exact atom counts, demonstrating that stoichiometry dictates approximately 86% of polarizability variance but only 8.5% of HOMO energy variance. The right panel displays the density distributions of the resulting size-conditioned standardized residuals (`res_homo`, `res_lumo`, `res_mu`, `res_r2`, and `res_alpha`). By regressing these properties against atom counts, the dominant effects of molecular size are removed to isolate the variance uniquely attributable to molecular topology. We learn that this baseline size-conditioning is essential to ensure that the observed topological re-coupling and integration of frontier orbitals into the global geometric state space are genuine physical effects driven by conjugation and aromaticity, rather than statistical artifacts of molecular size.

3.2 Topological stratification and sparsity control

Following residual extraction, the dataset was stratified along semi-continuous gradients of conjugation length and aromaticity. To prevent structural sparsity bias from confounding our coupling metrics, we

applied stratified subsampling to perfectly align the heavy-atom count distributions across the "Short", "Medium", and "Long" conjugation bins. confirms that this matching process eliminates underlying size variations across strata, ensuring that subsequent shifts in joint structure are strictly driven by electronic delocalization rather than statistical artifacts of the underlying size distribution.

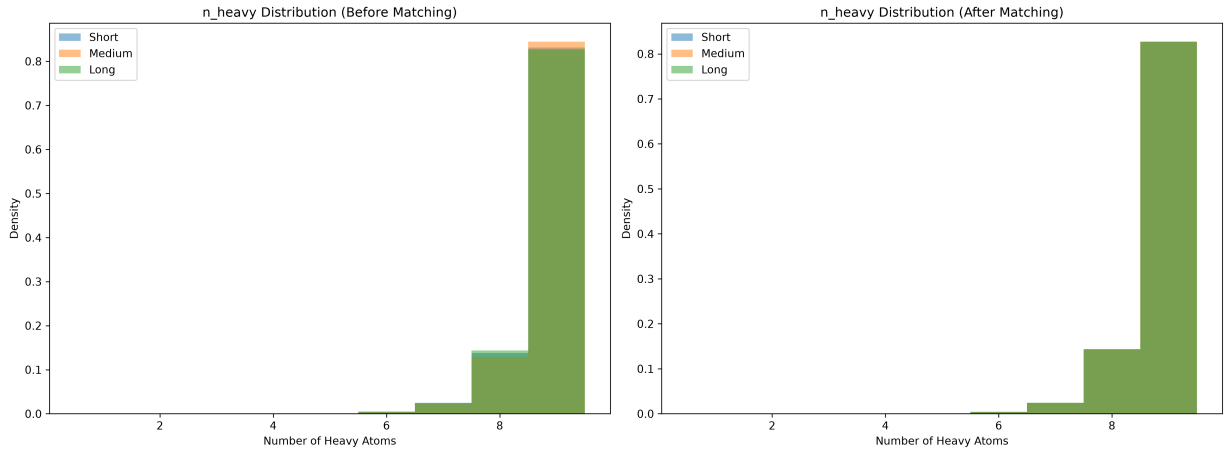

Figure 2: Density distributions of the number of heavy atoms (`n_heavy`) for Short, Medium, and Long conjugation length bins before (left) and after (right) stratified subsampling. While the initial dataset exhibits slight variations in molecular size across conjugation strata, the matching process perfectly aligns the heavy-atom count distributions across all levels. This ensures that the observed structural rewiring and increased integration of frontier orbitals in highly conjugated systems are genuine topological effects rather than statistical artifacts driven by underlying differences in molecular size.

3.3 Conjugation-driven integration of frontier orbitals

With size and sparsity biases controlled, we evaluated the joint structure between the frontier orbitals and the geometric response group. As conjugation length increases, the frontier-geometric manifold undergoes a fundamental reorganization. In the "Short" conjugation bin, the primary canonical correlation (CC_1) is 0.536, with distance correlation ($dCor$) values suggesting a localized, non-linear relationship ($dCor_{\text{homo}} = 0.372$, $dCor_{\text{lumo}} = 0.287$).

However, in the "Long" conjugation bin, CC_1 rises to 0.627, indicating a stronger linear alignment of the manifold. Concurrently, the absolute $dCor$ values decrease ($dCor_{\text{homo}} = 0.133$, $dCor_{\text{lumo}} = 0.126$) but remain significantly above their respective exact-count null means (0.055 and 0.052). This trajectory, visualized in , demonstrates that extended delocalization forcefully re-integrates frontier orbitals from localized, bond-specific variance into a constrained, global electronic state.

3.4 Aromaticity as a phase transition in coupling

The integration of frontier orbitals reaches a critical phase transition in the presence of aromaticity. For purely aliphatic molecules, the coupling is relatively weak and distributed across multiple dimensions, with $CC_1 = 0.449$ and a secondary correlation $CC_2 = 0.405$. In contrast, aromatic systems exhibit a dramatic manifold collapse: CC_1 jumps to 0.681 while CC_2 collapses to 0.080 (, bottom row).

This shift indicates that in aromatic molecules, the frontier orbitals and geometric properties are coupled along a single, dominant axis of variance. The anomalous isolation of the HOMO observed in global dataset averages is therefore a feature of the aliphatic majority; once aromaticity is introduced, the HOMO is forcefully shunted into the global property manifold, losing its independent degrees of freedom.

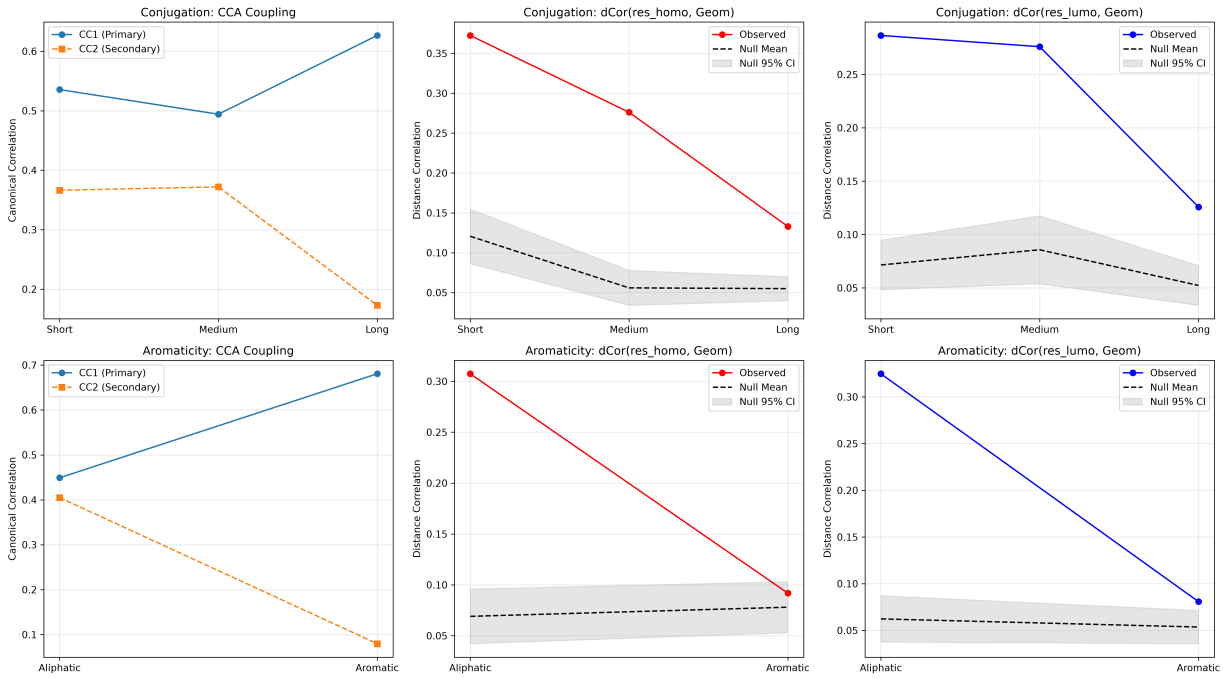

Figure 3: Joint structure analysis illustrating the topological re-coupling of frontier orbitals across conjugation length (top row) and aromaticity (bottom row). The left panels display Canonical Correlation Analysis (CCA) metrics, revealing that the primary canonical correlation (CC_1) strengthens while the secondary correlation (CC_2) collapses as delocalization increases. This is particularly evident in aromatic systems, where the manifold dimensionality collapses into a single dominant axis of variance. The middle and right panels show the distance correlation ($dCor$) for size-conditioned HOMO (`res_homo`) and LUMO (`res_lumo`) residuals against geometric responses. As conjugation and aromaticity increase, the absolute $dCor$ values decrease but remain significantly above the permuted null means (dashed lines with 95% confidence intervals). Together, these trends demonstrate that increased electronic delocalization forcefully re-integrates frontier orbitals from localized, non-linear relationships into a constrained, global geometric state space.

3.5 Heteroatom positional effects

Beyond carbon-backbone topology, heteroatom identity and structural position strongly modulate this re-coupling. Pure hydrocarbons exhibit a moderate primary coupling ($CC_1 = 0.498$). The introduction of fluorine drives the strongest integration ($CC_1 = 0.737$), suggesting that its high electronegativity and localized orbitals impose rigid dipole-driven constraints on the frontier manifold.

Furthermore, heteroatoms participating directly in aromatic rings enforce the one-dimensional coupling regime ($CC_1 = 0.678$, $CC_2 = 0.078$). Interestingly, exocyclic nitrogen yields higher integration ($CC_1 = 0.648$) than endocyclic nitrogen in non-aromatic rings ($CC_1 = 0.558$), implying that nitrogen lone pairs in flexible, conjugated environments more effectively bridge electronic and geometric states than those constrained within saturated cycles.

3.6 Global manifold visualization

To contextualize these high-dimensional shifts, we projected the 5-dimensional residual space into a Uniform Manifold Approximation and Projection (UMAP) embedding. As shown in , the embedding reveals distinct clustering governed by molecular delocalization. Aliphatic molecules form a broad, diffuse manifold characterized by high variance in the dipole moment residual (res_μ). The transition to longer conjugation lengths and aromaticity manifests as a continuous drift toward a constrained, high-density region. This visualizes the structural collapse where frontier orbitals in delocalized systems are integrated into the global geometric state space, confirming that positional heteroatom types dictate the local structure of this re-coupling.

3.7 Robustness and cross-scaffold generalization

Finally, to ensure these findings are intrinsic to delocalized electronic systems rather than artifacts of specific overrepresented molecular geometries, we performed rigorous cross-scaffold validation. In a 5-fold cross-validation split by Murcko scaffolds, the "Long" conjugation bin consistently maintained a higher CC_1 than the "Short" bin, with a mean gap of 0.145 and a standard deviation of 0.079.

Furthermore, a leave-one-scaffold-out (LOSO) analysis on the top ten most frequent scaffolds—including acyclic chains, cyclopropane, and oxirane—demonstrated that CC_1 for long-conjugated molecules remains highly stable, fluctuating only between 0.626 and 0.669 regardless of the excluded scaffold. This confirms that the topological re-coupling of frontier orbitals is a robust, generalized physical phenomenon across diverse chemical families, underscoring the necessity of stratifying predictive models by topological class.

4 Conclusions

4.1 Resolution of the frontier orbital anomaly

This study resolves the anomalous global decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from the broader geometric and spatial property manifold. By rigorously isolating chemical architecture from stoichiometric determinism through size-conditioned residuals, we have demonstrated that the apparent isolation of the HOMO is not an intrinsic quantum mechanical feature. Rather, it is a state-dependent structural constraint driven by the overwhelming prevalence of localized, aliphatic molecules in large-scale enumerative datasets like QM9.

Our topological stratification reveals that extended electron delocalization forcefully re-integrates both the HOMO and LUMO into the global geometric state space. In short-conjugated and aliphatic systems, frontier orbital energies are primarily governed by local bonding environments, exhibiting weak and multi-dimensional coupling with macroscopic spatial properties. However, as the π-system extends, these orbitals undergo a fundamental structural rewiring. This integration culminates in a phase transition induced by aromaticity, where the joint property manifold collapses into a single dominant dimension (with the primary canonical correlation rising to $CC_1 = 0.681$ while the secondary correlation drops to $CC_2 = 0.080$). Furthermore, we found that this low-dimensional coupling regime is strongly enforced by specific heteroatom environments, particularly the highly electronegative localized orbitals of fluorine and the delocalized lone pairs of aromatic nitrogen.

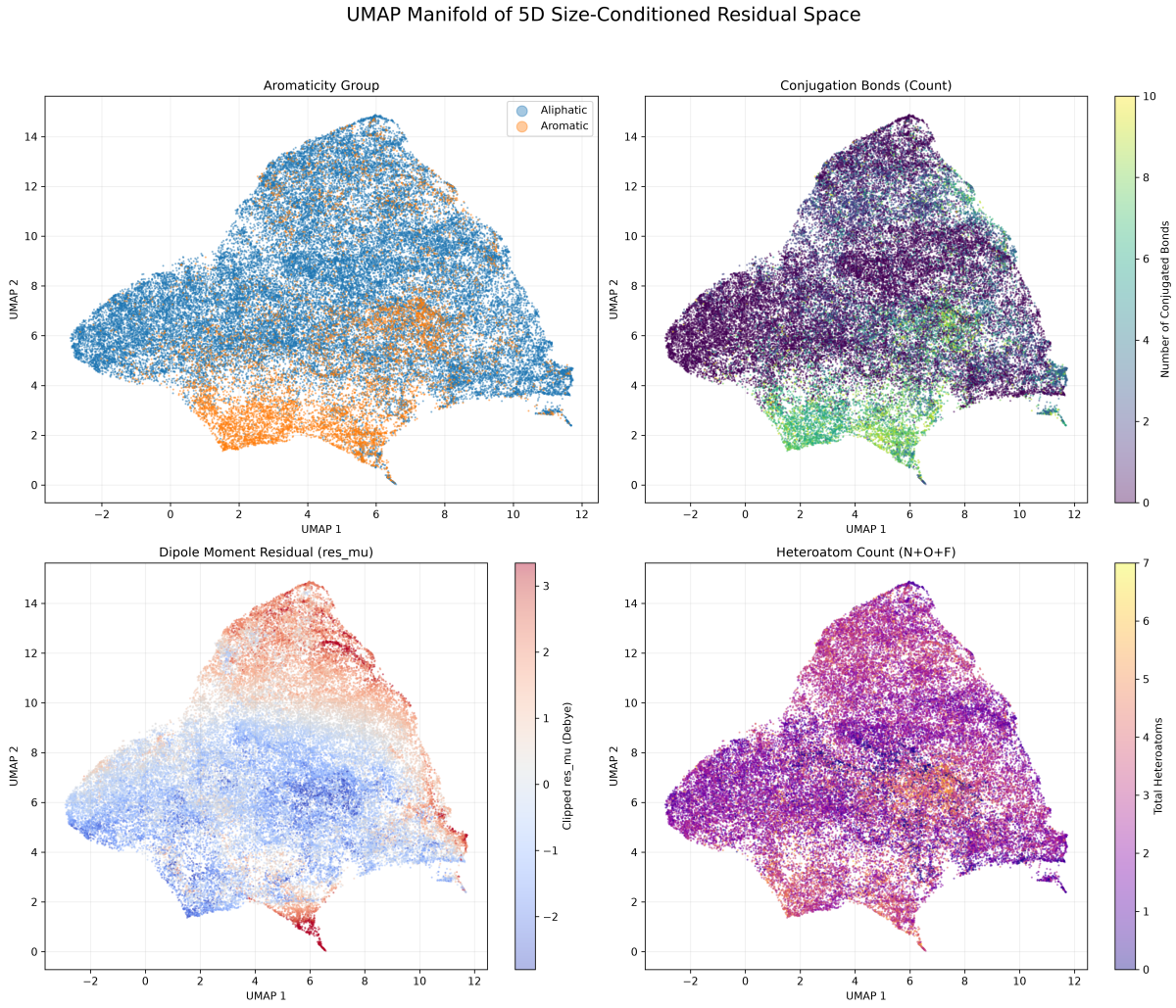

Figure 4: UMAP visualization of the 5D size-conditioned residual space demonstrating the topological recoupling of frontier orbitals. The embedding reveals distinct clustering governed by molecular delocalization, where aliphatic molecules form a broad, diffuse manifold exhibiting high variance in the dipole moment residual (res_μ). In contrast, the transition to longer conjugation lengths and the introduction of aromaticity drive a continuous drift toward a more constrained, high-density region. This structural collapse visualizes a fundamental phase transition where frontier orbitals in delocalized systems are forcefully integrated into the global geometric state space. Furthermore, the heteroatom count panel indicates that while total composition shifts the global manifold position, the local structural rewiring is primarily dictated by the topological placement of these heteroatoms rather than their absolute number.

4.2 Implications for chemical space modeling

The identification of this topological phase transition has significant implications for the development of predictive models in computational chemistry. Global machine learning models trained across highly heterogeneous chemical spaces often struggle to accurately predict HOMO energies because they attempt to fit a single continuous relationship to a population that spans fundamentally distinct coupling regimes. Our findings indicate that averaging over these diverse topological classes obscures the underlying physics of the property manifold.

To capture the true variance of frontier orbitals, future predictive frameworks and representation learning models must account for these distinct topological states, potentially by stratifying training regimes or explicitly encoding conjugation and aromaticity gradients. Ultimately, this work establishes that while frontier orbitals in saturated systems may act as independent electronic degrees of freedom, extended π-systems enforce a globalized, low-dimensional integration of electronic and spatial properties.