Mapping the Orthogonal Frontier-Electronic Manifold via Local Structural Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

While the thermodynamic and geometric properties of molecules are largely determined by global stoichiometry, frontier molecular orbitals often exhibit behavior that is orthogonal to macroscopic molecular architecture. To understand the local structural drivers of this decoupled electronic manifold, we analyzed 133,802 molecules from the QM9 dataset by extracting stoichiometric residuals for the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) using gradient boosting to remove macroscopic size effects. We then computed non-linear distance correlations between these residuals and local structural descriptors, alongside applying topological data analysis via Vietoris-Rips complexes to quantify manifold fragmentation. Our results indicate that LUMO residuals are primarily governed by carbonyl density and ring topology, whereas HOMO residuals are driven by nitrile and aromatic nitrogen environments. Furthermore, topological persistence statistics reveal that the introduction of nitrogen significantly increases the complexity and fragmentation of the frontier manifold compared to strictly matched hydrocarbon null models. Finally, matched-pair substitution walks demonstrate that single heteroatom replacements induce distinct, context-dependent trajectory shifts in the electronic space, confirming that frontier orbital physics is dictated by specific local bonding motifs rather than global molecular structure.

1 Introduction

1.1 Motivation and background

The mapping of molecular structure to quantum mechanical properties is a foundational challenge in computational chemistry and materials discovery. While macroscopic thermodynamic and geometric properties typically scale predictably with global molecular stoichiometry, the frontier molecular orbitals—specifically the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO)—often exhibit behavior that is orthogonal to this macroscopic architecture. This decoupling suggests that frontier-electronic manifolds are governed by localized physical rules rather than global structural determinism. Consequently, global molecular descriptors frequently fail to capture the nuanced behavior of the HOMO and LUMO, as they force these electronic properties into macroscopic geometric frameworks that obscure the underlying physics. To truly understand the isolated physical state space of frontier orbitals, it is necessary to strip away the confounding variance introduced by molecular size and global stoichiometry.

1.2 Methodological approach

To isolate this unique frontier-electronic behavior from stoichiometric determinism, we investigate the orthogonal frontier-electronic manifold across 133,802 deduplicated molecules from the QM9 dataset. We extract stoichiometric baseline residuals (res_homo and res_lumo) via gradient boosting on exact atom counts, effectively removing the macroscopic size effect without discarding highly strained structural outliers. We then generate a targeted suite of local structural descriptors, including hybridization ratios (F_{sp^3}), specific heteroatom fragment counts, and topological polar surface area.

By computing non-linear distance correlation matrices between this local structural feature space and the frontier orbital residuals, we identify the precise functional motifs governing this decoupled electronic

Metric	Hydrocarbon Null	Nitrogen-rich
$^{**}H_1$ Count (Cycles)**	259	308
$^{**}H_1$ Max Lifetime**	0.2121	0.2940
$^{**}H_1$ Stable Features**	103	180
$^{**}H_1$ Entropy**	4.9588	5.2583

manifold. To understand the higher-order structure of this space, we apply Topological Data Analysis (TDA) via Vietoris-Rips complexes, quantifying how the frontier manifold topologically fragments into discrete regimes upon the introduction of specific heteroatoms. Finally, we execute matched-pair substitution walks to track the trajectory of frontier residuals across single-atom replacements (e.g., carbon-to-nitrogen or carbon-to-oxygen substitutions) within identical Murcko scaffolds, revealing whether manifold fragmentation manifests as sharp phase transitions or continuous drifts.

1.3 Summary of contributions

Our analysis reveals that the HOMO and LUMO manifolds are governed by distinct, largely non-overlapping structural drivers. We find that *res_lumo* is primarily sensitive to carbonyl density and ring topology, whereas *res_homo* is driven by nitrile and aromatic nitrogen environments. Scaffold-disjoint split-half validation confirms that these structural drivers are robust, general features of the QM9 manifold rather than artifacts of specific molecular families.

Furthermore, topological persistence statistics demonstrate that the introduction of nitrogen significantly increases the complexity and fragmentation of the frontier manifold. When contrasting a nitrogen-rich cohort against a strictly matched hydrocarbon null model (matched by heavy atom and ring counts), the nitrogen-rich manifold exhibits a higher number of stable 1-cycles, greater maximum lifetimes, and elevated persistence entropy. These topological differences are summarized in Table 1.3.

Matched-pair substitution walks further demonstrate that single heteroatom replacements induce distinct, context-dependent trajectory shifts in the electronic space. Together, these findings confirm that frontier orbital physics is dictated by specific local bonding motifs and hybridization states, providing a mechanistic explanation for why the HOMO and LUMO refuse to align with global property networks.

2 Methods

2.1 Data ingestion and preprocessing

The QM9 dataset, comprising 133,802 organic molecules, was utilized for all analyses. Initial preprocessing involved the suppression of RDKit loggers and the deduplication of canonical SMILES strings to ensure a unique structural set. To isolate the frontier-electronic behavior from macroscopic stoichiometric determinism, we extracted baseline residuals for the highest occupied and lowest unoccupied molecular orbitals (*res_homo* and *res_lumo*). This was achieved by training gradient boosting regressors on exact atom counts (C, N, O, F, H) to predict the raw HOMO and LUMO energies. The residuals from these models represent the variance in frontier orbital energies that is strictly orthogonal to global molecular size and atomic composition.

2.2 Local structural descriptors and distance correlation

To quantify local bonding environments, a targeted suite of structural descriptors was generated using RDKit. These included the fraction of sp^3 hybridized carbons (F_{sp^3}), specific heteroatom fragment counts (such as carbonyls, nitriles, and aromatic nitrogens), topological polar surface area, and total ring counts. The dataset was subsequently stratified by fixed heavy-atom counts, with a specific focus on the most populated strata (7, 8, and 9 heavy atoms) to maximize statistical power.

We computed non-linear distance correlation ($dCor$) matrices between the local structural feature space and the frontier orbital residuals. To manage computational complexity, calculations were parallelized across 14 CPUs using the `joblib` library, and representative scaffold-stratified subsamples of 10,000 to 20,000

molecules were utilized where necessary. Features were ranked by their association magnitude, with $dCor < 0.2$ defined as a low-magnitude association threshold.

To validate the structural robustness of these associations, we performed a scaffold-disjoint split-half validation. The dataset was partitioned into two independent halves based on Murcko scaffolds. Distance correlations were computed separately for each half, and their agreement was quantified using Pearson and Spearman correlation coefficients to ensure the identified drivers were not artifacts of specific molecular clusters.

2.3 Topological data analysis

To investigate the higher-order structure and potential fragmentation of the frontier-electronic manifold, Topological Data Analysis (TDA) was applied. Vietoris-Rips complexes were constructed on the two-dimensional (res_homo, res_lumo) space using the `gudhi` library. We generated full persistence diagrams across all spatial scales to identify stable topological features, specifically focusing on connected components (H_0) and one-dimensional cycles (H_1).

The topological complexity was quantified using several summary statistics: the total count of H_1 cycles, the maximum lifetime of these cycles, the number of stable features (defined by lifetimes exceeding the median), and the persistence entropy. To isolate the topological influence of specific heteroatoms, we contrasted a nitrogen-rich cohort (molecules containing ≥ 2 nitrogen atoms) against a strictly matched hydrocarbon null model. The null model was constructed to exactly match the distribution of both heavy-atom counts and the number of rings present in the nitrogen-rich cohort.

2.4 Matched-pair substitution walks

Finally, to mechanistically trace how local structural changes rewire the electronic manifold, we executed a targeted substitution walk via a matched-pair analysis. Using RDKit's substructure matching and canonical SMILES comparisons, we identified existing pairs of molecules within the dataset that shared an identical Murcko scaffold but differed by exactly one atom (e.g., a single carbon-to-nitrogen or carbon-to-oxygen substitution). By tracking the trajectory of the res_homo and res_lumo residuals across these matched pairs, we quantified the magnitude and direction of the electronic shifts (Δres_homo and Δres_lumo). This approach allowed us to determine whether specific heteroatom substitutions induce sharp phase transitions or continuous drifts within the isolated frontier manifold.

3 Results

3.1 Governance of the orthogonal frontier manifold

To isolate the chemical variance of the frontier orbitals from macroscopic size effects, we analyzed the stoichiometric residuals (res_homo and res_lumo) extracted via gradient boosting. These residuals represent the electronic shifts driven purely by local bonding topology and functional group environments, independent of global atomic composition.

Distance correlation ($dCor$) analysis reveals that the res_homo and res_lumo manifolds are governed by distinct, largely non-overlapping sets of structural features. As illustrated in Figure 1, the res_lumo manifold is most strongly associated with the presence of carbonyl groups ($dCor = 0.4916$) and molecular ring count ($dCor = 0.3819$). This indicates that LUMO energy levels are primarily sensitive to π-system extension and the introduction of electron-withdrawing oxygen centers. In contrast, res_homo is most strongly coupled to nitrile groups ($dCor = 0.4340$) and aromatic nitrogen content ($dCor = 0.2451$), showing significantly lower sensitivity to carbonyls ($dCor = 0.1536$) and total ring count ($dCor = 0.1995$). Both residuals exhibit moderate coupling to the fraction of sp^3 hybridized carbons (f_{sp3}), though res_lumo ($dCor = 0.3599$) is more sensitive to this geometric feature than res_homo ($dCor = 0.2709$).

To ensure these structural associations are not artifacts of specific molecular families, we performed a scaffold-disjoint split-half validation (Figure 2). The distance correlations computed for two independent sets of Murcko scaffolds demonstrate remarkable stability, yielding a Pearson correlation of $r = 0.9044$ and

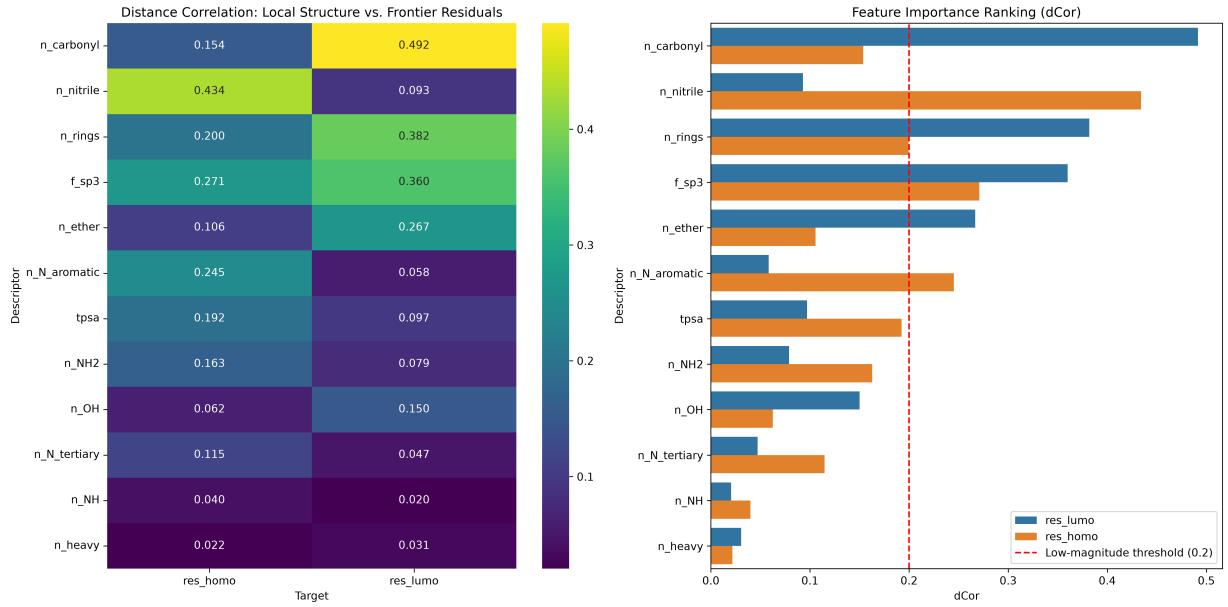

Figure 1: Distance correlation (dCor) analysis between local structural descriptors and stoichiometric frontier orbital residuals (*res_homo* and *res_lumo*). The left panel presents a heatmap of dCor values, while the right panel ranks feature importance against a low-magnitude threshold of 0.2. The results demonstrate that the two residuals are governed by distinct, largely non-overlapping structural drivers. The *res_lumo* manifold is primarily sensitive to carbonyl groups (`n_carbonyl`) and molecular ring count (`n_rings`), reflecting a dependence on π-system extension and electron-withdrawing oxygen centers. In contrast, *res_homo* is most strongly coupled to nitrile groups (`n_nitrile`) and aromatic nitrogen content (`n_N_aromatic`). Both residuals exhibit moderate coupling to the fraction of sp^3 hybridized carbons (f_{sp3}). Ultimately, this reveals that the decoupled electronic manifold is rewired by specific local bonding motifs and heteroatom identities rather than global molecular architecture.

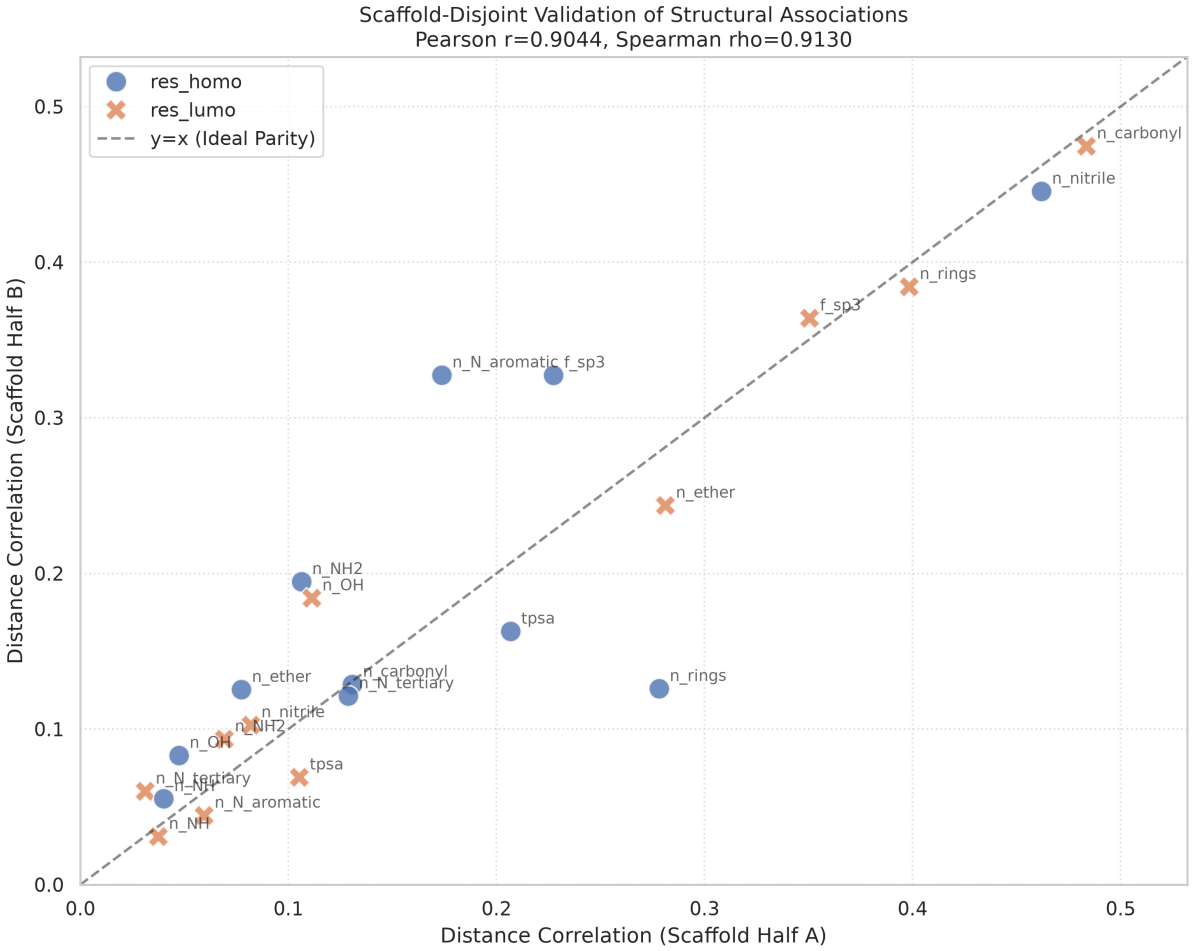

Figure 2: Scaffold-disjoint split-half validation of structural associations for stoichiometric residuals (*res_homo* and *res_lumo*). The scatter plot compares the distance correlation of various structural features with the frontier orbital residuals across two independent sets of Murcko scaffolds. Blue circles and orange crosses represent associations for *res_homo* and *res_lumo*, respectively. The tight alignment of points along the ideal parity line ($y = x$), quantified by high Pearson ($r = 0.9044$) and Spearman ($\rho = 0.9130$) correlations, demonstrates remarkable stability. This confirms that the identified local structural drivers—such as carbonyl groups for LUMO and nitrile groups for HOMO—are general features governing the orthogonal frontier-electronic manifold of the QM9 dataset, rather than artifacts of specific molecular clusters.

a Spearman correlation of $\rho = 0.9130$. This parity confirms that the identified structural drivers are general, robust features of the QM9 manifold rather than localized correlations within specific molecular clusters.

3.2 Topological fragmentation and heteroatom influence

To understand the higher-order structure of the decoupled electronic space, we constructed Vietoris-Rips complexes in the two-dimensional (*res_homo*, *res_lumo*) space. We compared a nitrogen-rich cohort (molecules containing ≥ 2 nitrogen atoms) against a strictly matched hydrocarbon null model, which was constrained to have identical heavy-atom and ring count distributions.

Topological Data Analysis (TDA) summary statistics reveal that the introduction of nitrogen significantly increases the topological complexity of the frontier manifold (Table 1.3 and Figure 3). The nitrogen-rich manifold exhibits a higher number of stable 1-cycles (H_1) compared to the hydrocarbon null (180 vs. 103) and

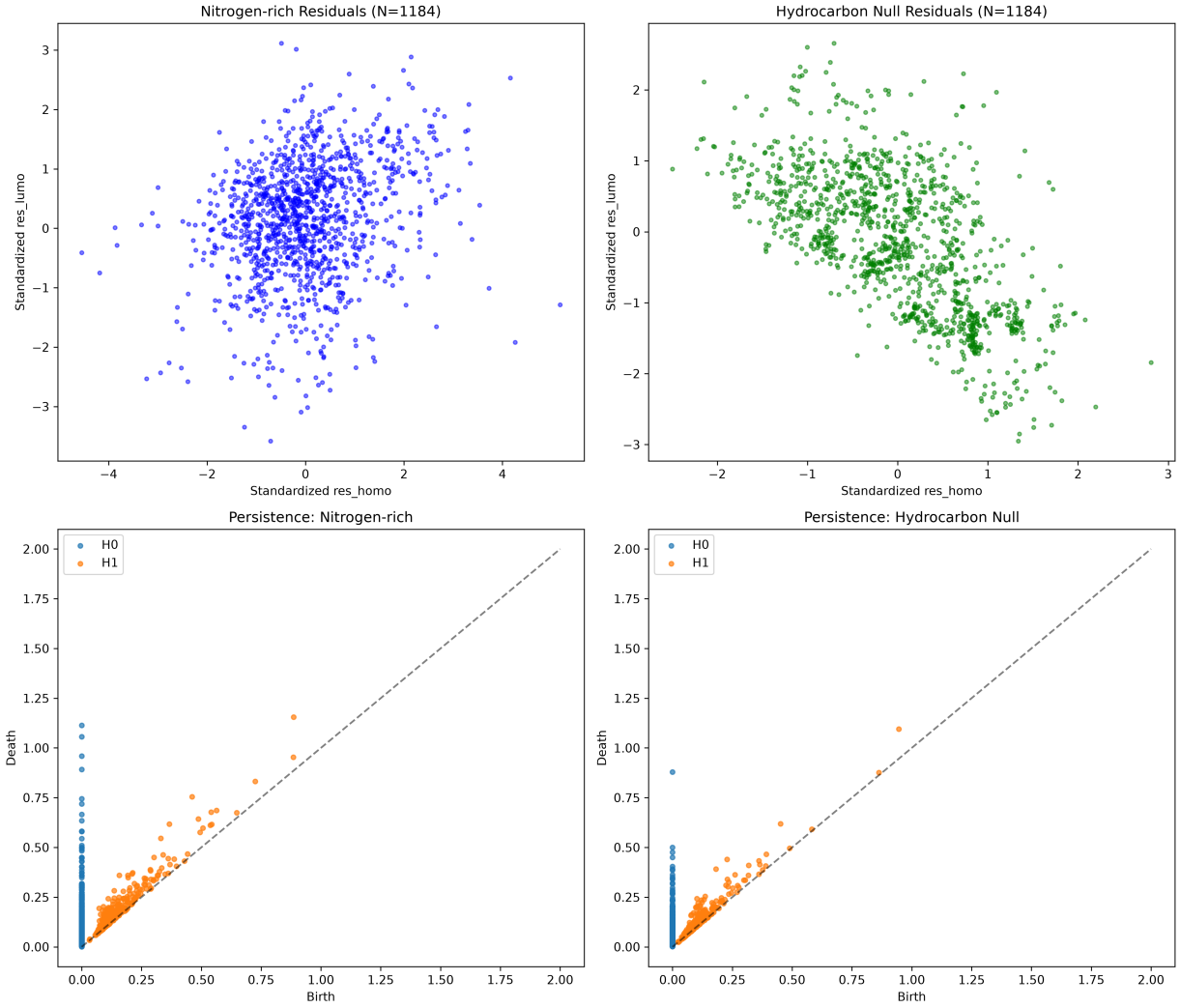

Figure 3: Top panels display the distribution of standardized stoichiometric residuals (*res_homo* and *res_lumo*) for a nitrogen-rich cohort and a strictly matched hydrocarbon null model. Bottom panels show the corresponding topological persistence diagrams for connected components (H_0) and 1-cycles (H_1). The scatter plots illustrate that the nitrogen-rich cohort occupies the frontier orbital space more sparsely and forms distinct clusters, whereas the hydrocarbon null merges into a single component more rapidly. The persistence diagrams reveal that the introduction of nitrogen increases the topological complexity of the manifold, evidenced by a higher number of stable H_1 features with greater maximum lifetimes. Additionally, the extended H_0 lifetimes in the nitrogen-rich cohort confirm that heteroatom substitution drives topological fragmentation, creating distinct electronic regimes that rewire the frontier manifold independently of macroscopic molecular architecture.

a higher persistence entropy (5.2583 vs. 4.9588), indicating a more fragmented and complex loop structure. Furthermore, the maximum lifetime of these H_1 features is substantially higher in the nitrogen-rich cohort (0.2940 vs. 0.2121), suggesting that nitrogen creates distinct, persistent electronic regimes across spatial scales.

Analysis of manifold connectivity further supports this fragmentation. The median H_0 lifetime, which represents the distance at which isolated components merge, is notably higher for the nitrogen-rich cohort (0.0701) than for the hydrocarbon null (0.0460). This implies that nitrogen-containing molecules occupy the frontier orbital space more sparsely and form more distinct clusters, whereas hydrocarbons merge into a single connected component more rapidly.

3.3 Mechanistic insights from substitution walks

To mechanistically trace how local structural changes rewire the electronic manifold, we tracked matched molecular pairs that differ by a single heteroatom substitution within an identical Murcko scaffold (Figure 4).

A single carbon-to-nitrogen (C $\rightarrow$ N) substitution induces a systematic increase in *res_lumo* (mean $\Delta = +0.0153$ Hartree) while leaving the mean *res_homo* nearly unchanged ($\Delta = -0.0001$ Hartree). However, the high standard deviation in the *res_homo* shift (± 0.0211 Hartree) indicates that the effect of nitrogen on the HOMO is highly context-dependent. This variance is likely contingent on the specific local environment, such as whether the nitrogen is incorporated into an aromatic ring or forms a nitrile group, aligning with our distance correlation findings.

Conversely, a carbon-to-oxygen (C $\rightarrow$ O) substitution results in a simultaneous increase in both residuals (mean $\Delta res_homo = +0.0033$ Hartree, mean $\Delta res_lumo = +0.0172$ Hartree). Oxygen substitutions exhibit a tighter distribution of shifts (standard deviation for *res_homo* = 0.0129 Hartree), suggesting a more uniform electronic response across different scaffolds compared to nitrogen. These trajectory shifts confirm that frontier orbital physics is dictated by specific local bonding motifs and heteroatom identities, driving sharp phase transitions in the electronic space that operate independently of the molecule's macroscopic architecture.

4 Conclusions

In this study, we addressed the fundamental decoupling of frontier molecular orbitals from macroscopic molecular architecture. By isolating stoichiometric residuals (*res_homo* and *res_lumo*) from the QM9 dataset and applying non-linear distance correlation alongside topological data analysis, we successfully mapped the local structural drivers that govern this orthogonal electronic manifold.

Our findings demonstrate that the HOMO and LUMO manifolds are not monolithic structures, but rather collections of distinct regimes governed by non-overlapping local bonding motifs. Specifically, the LUMO manifold is primarily dictated by carbonyl density and ring topology, reflecting a strong sensitivity to π-system extension and electron-withdrawing centers. Conversely, the HOMO manifold is driven by nitrile groups and aromatic nitrogen environments. The robustness of these structural drivers, confirmed via scaffold-disjoint split-half validation, underscores that these associations are general physical rules of the frontier manifold rather than localized artifacts of specific molecular families.

Furthermore, topological data analysis revealed that the introduction of specific heteroatoms fundamentally rewires the complexity of the frontier space. As summarized in Table 1.3, a nitrogen-rich cohort exhibits a significantly more fragmented and complex loop structure compared to a strictly matched hydrocarbon null model. The nitrogen-rich manifold is characterized by a higher number of stable 1-cycles, greater maximum lifetimes, and elevated persistence entropy, indicating that nitrogen creates distinct, persistent electronic regimes across spatial scales.

Matched-pair substitution walks provided mechanistic confirmation of these topological shifts. Single heteroatom replacements induce distinct trajectory shifts in the electronic space; for instance, carbon-to-nitrogen substitutions yield highly context-dependent HOMO shifts, whereas carbon-to-oxygen substitutions produce more uniform electronic responses across different scaffolds.

Ultimately, these findings explain the frequent failure of global molecular descriptors to capture frontier orbital behavior. The HOMO and LUMO manifolds are rewired by local heteroatom identity and hybridiza-

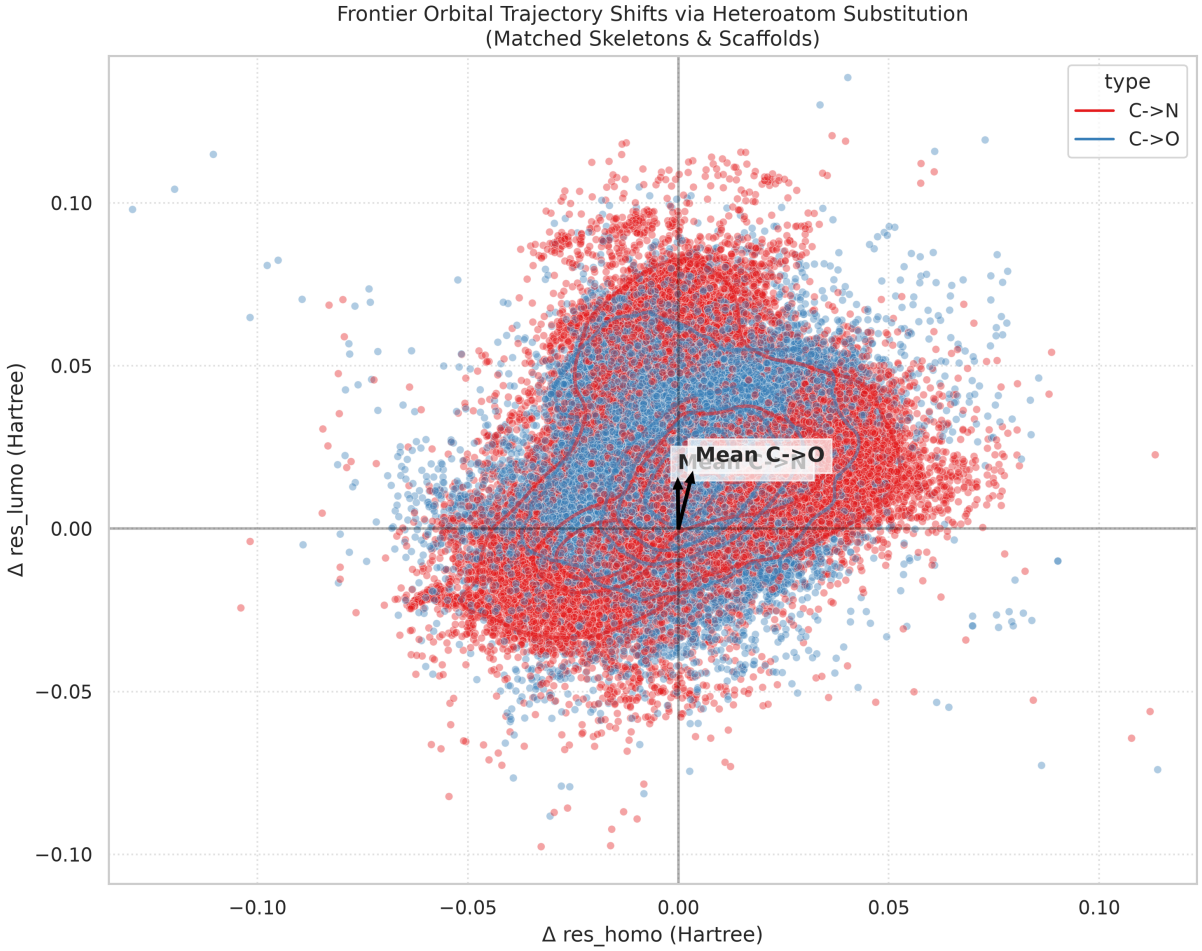

Figure 4: Frontier orbital trajectory shifts (Δres_homo vs. Δres_lumo) for matched molecular pairs differing by a single heteroatom substitution within the same skeleton and Murcko scaffold. Carbon to nitrogen (C $\rightarrow$ N, red) substitutions induce a systematic increase in the LUMO residual while leaving the mean HOMO residual nearly unchanged. However, the wide horizontal spread for C $\rightarrow$ N indicates that the effect on the HOMO is highly context-dependent, likely relying on specific local environments such as aromatic rings or nitrile groups. In contrast, carbon to oxygen (C $\rightarrow$ O, blue) substitutions cause a simultaneous increase in both residuals with a tighter distribution, reflecting a more uniform electronic response across different scaffolds. Arrows indicate the mean trajectory shifts for each substitution type, demonstrating how local heteroatom identity rewires the decoupled electronic manifold independently of macroscopic molecular architecture.

tion states that operate independently of global molecular structure. By shifting the analytical focus from macroscopic geometry to local structural distance correlation, this work provides a robust framework for understanding and predicting the isolated physical state space of frontier-electronic manifolds.