Topological and Statistical Rewiring of the Frontier Electronic Manifold Across Structural Motifs

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how the highest occupied molecular orbital (HOMO) couples with or decouples from global spatial and thermodynamic properties across diverse chemical environments remains a fundamental challenge in physical chemistry. To determine whether the orthogonality of the HOMO is a universally rigid feature or a structurally sensitive property, we investigate the joint structure of molecular properties by isolating chemical effects from stoichiometric scaling. Utilizing the QM9 dataset, we extract gradient-boosted atom-count residuals for frontier orbital energies and spatial response properties, stratifying the molecules into aliphatic, non-aromatic heterocyclic, and aromatic classes. We then apply distance correlation matrices, mutual information networks, and Vietoris-Rips persistent homology to rigorously quantify the multiscale geometric shape of the electronic manifold. Our analysis reveals that the HOMO energy is conditionally isolated, acting as a decoupled local property in over eighty percent of the dataset, specifically within aliphatic and non-aromatic heterocyclic systems. Conversely, the onset of aromaticity acts as a phase transition that integrates the HOMO into the joint manifold, coupling it strongly with the lowest unoccupied molecular orbital and overall molecular geometry. Finally, we show that the introduction of nitrogen or oxygen heteroatoms into aromatic rings acts as a decoherence agent, weakening orbital-orbital coupling and topologically deforming the electronic state space compared to purely carbocyclic systems.

1 Introduction

1.1 The frontier electronic manifold

Frontier molecular orbitals, specifically the highest occupied molecular orbital ($HOMO$) and the lowest unoccupied molecular orbital ($LUMO$), dictate the chemical reactivity, optical characteristics, and thermodynamic stability of organic molecules. Understanding how these localized electronic states couple to, or decouple from, global spatial and thermodynamic properties is a cornerstone of physical chemistry. However, it remains a fundamental challenge to determine whether the energetic orthogonality of the $HOMO$ is a universally rigid feature across all chemical environments, or if it is a structurally sensitive property that dynamically rewires in response to local topological perturbations.

1.2 Isolating chemical effects from stoichiometric scaling

A major obstacle in mapping the joint structure of molecular properties is the overwhelming influence of stoichiometric scaling. Extensive properties, and even intensive properties subject to quantum confinement and screening effects, scale non-linearly with molecular size and heavy-atom counts. Consequently, apparent correlations between electronic and spatial properties often reflect trivial size effects rather than intrinsic chemical coupling. To rigorously evaluate the geometric shape of the local electronic manifold, it is necessary to isolate true chemical effects—such as conjugation, hyperconjugation, and inductive effects—from baseline stoichiometric scaling.

In this work, we investigate the topological and statistical rewiring of the frontier electronic manifold across diverse structural motifs. Utilizing the QM9 dataset, we extract gradient-boosted atom-count residuals for frontier orbital energies and spatial response properties. This procedure effectively eliminates the

Stratum	Population Count	$HOMO$ Isolation Score ($dCor$)
Aliphatic	39,167	0.542
Heterocyclic (Non-Aromatic)	70,865	0.555
Aromatic	23,766	0.837

stoichiometric baseline, yielding a purely chemical state space. We then deploy a combination of distance correlation ($dCor$) matrices, mutual information networks, and Vietoris-Rips persistent homology to quantify the multiscale geometric shape of this manifold.

1.3 Structural stratification and manifold rewiring

By stratifying the deduplicated dataset into aliphatic, non-aromatic heterocyclic, and aromatic classes, we test the hypothesis that the isolated topological signature of the $HOMO$ residual is invariant. Our analysis reveals that the $HOMO$ energy is conditionally isolated, acting as a decoupled local property in over eighty percent of the dataset. As shown in Table 1.3, the isolation score of the $HOMO$ remains low in aliphatic and non-aromatic heterocyclic systems, indicating that it is largely orthogonal to the global spatial and thermodynamic responses of the molecule.

Conversely, the onset of aromaticity acts as a phase transition that integrates the $HOMO$ into the joint manifold. In aromatic systems, the $HOMO$ couples strongly with the $LUMO$ and the overall molecular geometry, reflecting a shift from local orbital character to a regime of delocalized electronic states. Furthermore, we demonstrate that the introduction of nitrogen or oxygen heteroatoms into these aromatic rings acts as a decoherence agent. Differing electronegative lone-pair environments weaken specific orbital-orbital couplings and topologically deform the electronic state space compared to purely carbocyclic systems. Through this information-geometric framework, we establish that the orthogonality of the $HOMO$ is not a universal constant, but rather a highly structured, dynamically rewiring state space.

2 Methods

2.1 Data ingestion and preprocessing

To investigate the joint structure of molecular properties, we utilized the QM9 dataset, which provides computed geometric, energetic, and thermodynamic properties for small organic molecules. The dataset was first parsed using RDKit to extract canonical SMILES strings. To ensure that downstream structural analyses and data splits were not biased by identical molecular representations, the dataset was deduplicated based on these canonical SMILES, yielding a final global population of 133,802 unique molecules.

Special attention was given to the rotational constants. The rotational constant A exhibits an extreme heavy-tailed distribution, which can disproportionately skew residual errors during regression. To mitigate this, we applied a logarithmic transformation, $\log(A+1)$. Furthermore, perfectly linear molecules ($A = 0$) represent a distinct physical regime rather than a continuous structural variation. We therefore created a binary indicator variable, `is_linear`, to explicitly capture this geometric property and prevent it from being treated as a residual error in subsequent modeling.

2.2 Stoichiometric baseline extraction

Because extensive properties and many intensive properties scale non-linearly with molecular size, apparent correlations between electronic and spatial properties often reflect trivial stoichiometric scaling rather than intrinsic chemical coupling. To isolate true chemical effects, we computed exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms for each molecule.

We then fitted gradient-boosted regressors to predict the frontier orbital energies ($HOMO$, $LUMO$), the dipole moment (μ), the static polarizability (α), and the rotational constants ($\log(A+1)$, B, C) using the five atom counts and the `is_linear` indicator as input features. By extracting the residuals from these models (e.g., res_HOMO, res_LUMO), we effectively removed the stoichiometric baseline, leaving behind variations driven purely by chemical environment, such as conjugation and inductive effects. Finally, a

robust scaling transformation was applied to all residuals to ensure that downstream distance correlation and mutual information calculations were not biased by extreme variance or outliers.

2.3 Structural stratification

To test the hypothesis that the topological signature of the $HOMO$ residual is structurally sensitive, we classified the deduplicated molecules into three mutually exclusive structural strata using RDKit's ring information and aromaticity flags:

1. **Aromatic:** Molecules containing at least one aromatic ring.
2. **Heterocyclic (Non-aromatic):** Molecules containing at least one non-aromatic ring with a heteroatom (nitrogen or oxygen).
3. **Aliphatic:** Carbocyclic or acyclic molecules with no aromatic rings and no heterocycles.

The population counts and the corresponding $HOMO$ isolation scores for these strata are detailed in Table 1.3.

To pinpoint the precise chemical drivers of manifold rewiring within the aromatic population, we performed a high-resolution sub-stratification. Using SMARTS matching, the aromatic stratum was further divided based on heteroatom identity, isolating strictly nitrogen-containing aromatic rings, strictly oxygen-containing aromatic rings, and purely carbocyclic aromatic systems.

2.4 Statistical coupling and network analysis

For the global baseline and each structural stratum, we computed pairwise distance correlation ($dCor$) matrices on the seven scaled residuals. To quantify the integration or decoupling of the highest occupied molecular orbital within the broader geometric-electronic manifold, we defined an "isolation score" for res_HOMO. This score was calculated as the ratio of its average coupling strength with all other properties to the average coupling strength among those other properties themselves. A sensitivity check was performed by recomputing the global distance correlation matrix excluding linear molecules, confirming that couplings involving the rotational constant A were not solely driven by this subset.

In parallel, we computed pairwise mutual information (MI) between all seven residuals to construct undirected weighted graphs for each stratum. In these networks, nodes represent the molecular properties and edge weights correspond to the computed MI values. The network topology was quantified by calculating the weighted degree (node strength) of res_HOMO.

To validate the statistical significance of the observed couplings, we generated a null model for each stratum by randomly shuffling the property residual values while keeping the SMILES strings and stratum assignments fixed. This procedure preserves the marginal distributions of the residuals while destroying their joint structure. Empirical coupling metrics were then compared against this null baseline using effect sizes rather than p-values, given the large sample size of the dataset.

2.5 Topological data analysis

To rigorously evaluate the multiscale geometric shape of the local electronic manifold, we applied Vietoris-Rips persistent homology to the three-dimensional state space defined by $\{res_HOMO, res_LUMO, res_\alpha\}$. Point clouds were constructed for chemically distinct subsets, specifically comparing purely carbocyclic aromatics against nitrogen-rich heteroaromatics using balanced samples ($N = 288$) to prevent density-driven topological artifacts.

Vietoris-Rips complexes were built across a range of filtration radii, and persistence diagrams were extracted to capture the birth and death of homological features. By computing the Bottleneck distances between the persistence diagrams of the carbocyclic and heteroaromatic populations for both connected components (H_0) and one-dimensional loops (H_1), we quantified exactly how differing electronegative lone-pair environments topologically deform the frontier-electronic state space.

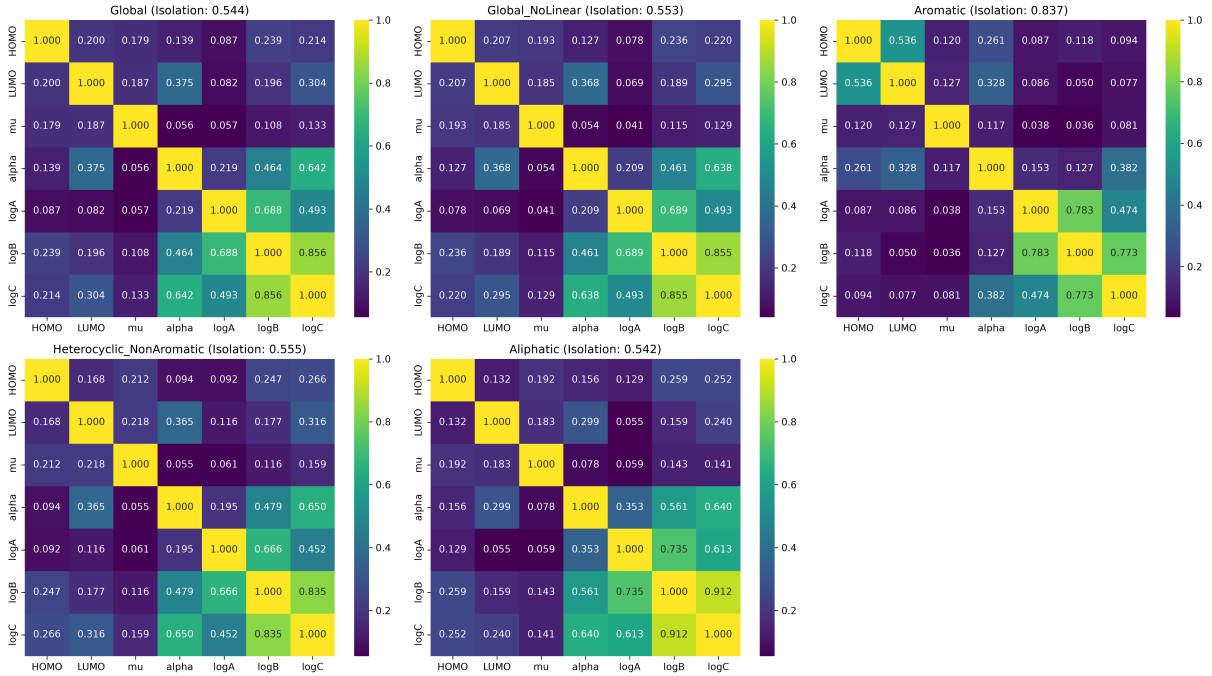


Figure 1: Distance correlation ($dCor$) matrices of molecular properties across distinct structural strata, illustrating the conditional isolation of the $HOMO$ energy. In over 80% of the dataset, including aliphatic and non-aromatic heterocyclic systems, $HOMO$ acts as a decoupled lone property with low isolation scores (≈ 0.54–0.56). The onset of aromaticity induces a phase transition that pulls $HOMO$ into the joint manifold, evidenced by its isolation score rising to 0.837 and the emergence of strong coupling with $LUMO$ ($dCor = 0.536$). Conversely, spatial response properties, such as the rotational constants $\log B$ and $\log C$, remain tightly coupled ($dCor > 0.83$) across all structural regimes.

3 Results

3.1 The stoichiometric baseline and global property isolation

The primary axis of variation in molecular properties is molecular size, which is almost entirely determined by atom counts. By conditioning the properties on the counts of C, N, O, F, and H atoms using gradient-boosted regression, we isolated the intrinsic chemical joint structure from trivial stoichiometric scaling. The resulting residuals represent the portion of each property that is not explained by stoichiometry.

Across the global dataset of 133,802 unique molecules, the energy of the highest occupied molecular orbital ($HOMO$) behaves as a distinct global outlier. While spatial response properties, such as the static polarisability α and the rotational constants B and C, are tightly coupled—exhibiting global distance correlations ($dCor$) as high as 0.856 between $\log B$ and $\log C$—the $HOMO$ residual remains largely decoupled from the rest of the molecular manifold. In this global population, the isolation score for $HOMO$, defined as the ratio of its average coupling to other properties versus the average coupling among those other properties, is 0.544. This global decoupling is visually evident in the distance correlation matrices (Figure 1) and mutual information networks (Figure 2), which highlight $HOMO$ as an isolated node in the majority of the chemical space.

3.2 Stratification of the frontier electronic manifold

The isolation of the $HOMO$ is not a universal feature of organic molecules, but rather a property characteristic of specific structural classes. Stratifying the dataset reveals three distinct regimes of electronic organization, as summarized in Table 1.3.

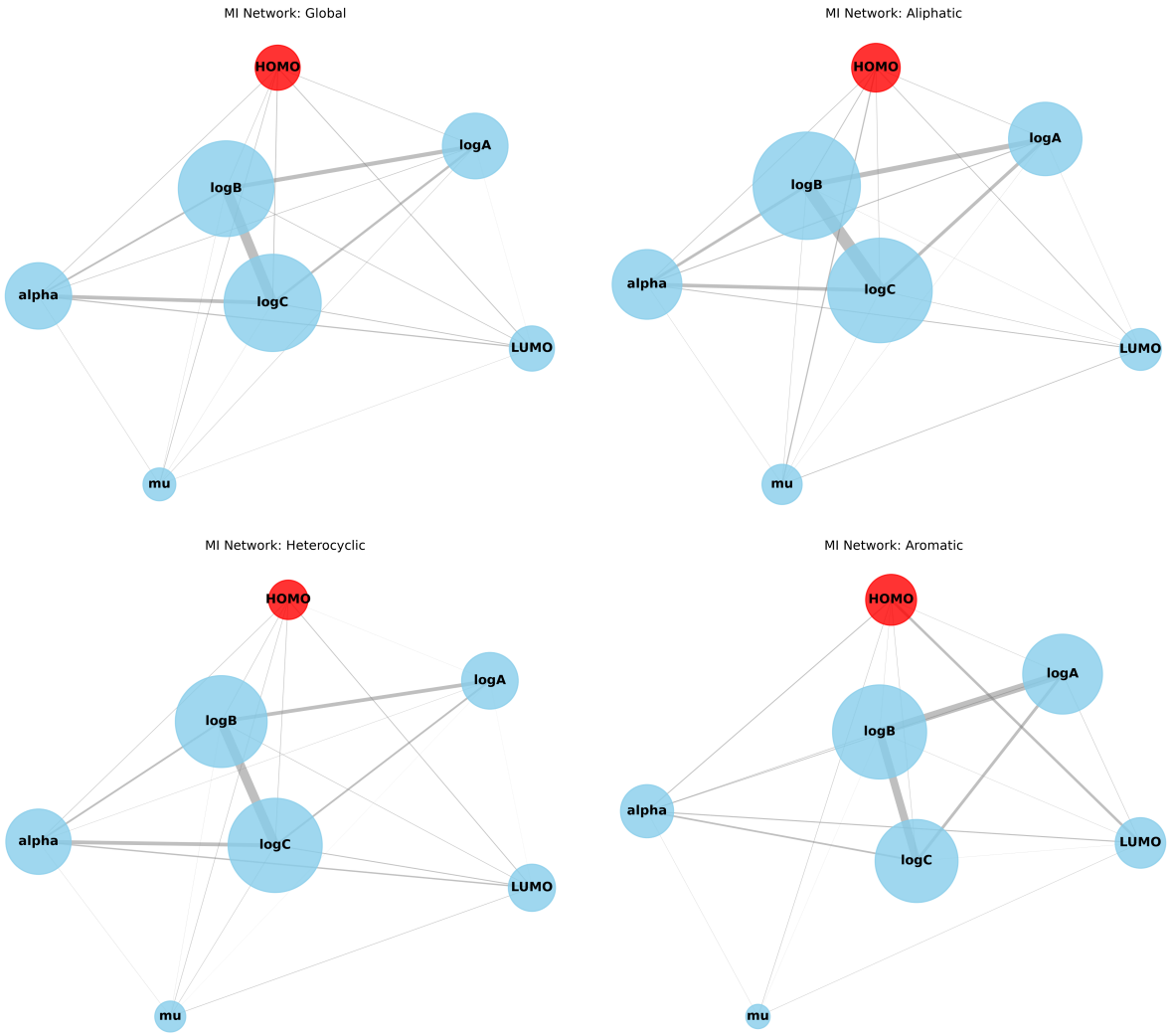


Figure 2: Mutual Information networks illustrating the joint structure of computed molecular properties across distinct structural strata. The graphs highlight the conditional isolation of the $HOMO$ energy residual (red node). In the global, aliphatic, and non-aromatic heterocyclic populations, $HOMO$ remains largely decoupled from spatial and energetic properties. However, the onset of aromaticity induces a structural rewiring, pulling $HOMO$ into the joint manifold where it couples strongly with $LUMO$ and rotational constants. This visualizes a phase transition from local orbital character to delocalized electronic states intrinsically tied to the molecular framework.

In aliphatic and non-aromatic heterocyclic systems, the $HOMO$ energy is governed by local bonding environments that are largely orthogonal to the global spatial and thermodynamic responses of the molecule. However, in aromatic systems, this isolation collapses. The $HOMO$ residual integrates into the joint structure, with its isolation score rising significantly to 0.837. This structural rewiring signifies a phase transition from local orbital character to a regime where electronic states are delocalized and strongly coupled to the molecular framework, as depicted in Figure 3.

3.3 Rewiring in aromatic sub-strata

A finer-grained analysis of the aromatic population distinguishes between purely carbocyclic systems and those containing heteroatoms (nitrogen or oxygen) within the aromatic rings. The coupling between the $HOMO$ and $LUMO$ residuals serves as a sensitive probe of this electronic rewiring (Figure 4).

In purely carbocyclic aromatics, the $HOMO$-$LUMO$ coupling is exceptionally strong ($dCor = 0.6514$), reflecting the high symmetry and uniform conjugation of pure carbon frameworks. In contrast, the introduction of nitrogen or oxygen heteroatoms significantly weakens this specific orbital-orbital coupling. For nitrogen-containing aromatics, the distance correlation drops to 0.5353, and for oxygen-containing or other heteroaromatics, it is 0.5386.

Beyond orbital-orbital interactions, the integration of the $HOMO$ with physical response properties is also heavily modulated by heteroatom inclusion. In carbocyclic aromatics, the $HOMO$ exhibits notable coupling with the rotational constants (e.g., $dCor = 0.5348$ with $\log C$). This suggests that for these planar, symmetric molecules, the electronic energy levels are intrinsically tied to the molecular geometry. This relationship is largely lost in heteroaromatics, where the coupling between the $HOMO$ and rotational constants drops significantly to approximately 0.10 for both nitrogen- and oxygen-containing aromatic systems.

3.4 Topological persistence of electronic states

To rigorously quantify the multiscale geometric shape of these electronic manifolds, we computed Vietoris-Rips persistent homology on the three-dimensional state space defined by the residuals of $HOMO$, $LUMO$, and polarisability α. A comparison between carbocyclic aromatics and N-heteroaromatics, using balanced samples ($N = 288$) to prevent density-driven artifacts, reveals significant topological differences (Figure 5).

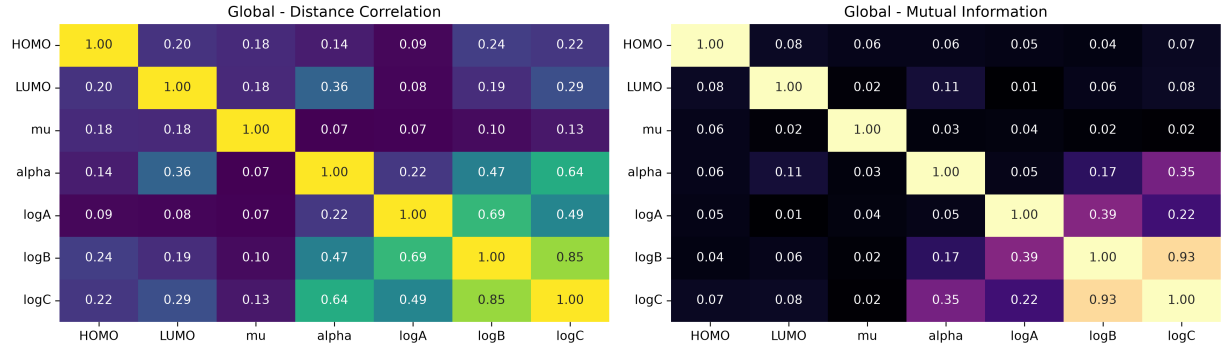
The computed bottleneck distance for connected components (H_0) is 0.372512, indicating that the point cloud of carbocyclic aromatics has a fundamentally different clustering density in the electronic state space compared to nitrogen-containing heteroaromatics. Furthermore, the bottleneck distance for one-dimensional loops (H_1) is 0.108244, reflecting differences in the fundamental voids within the distribution. These topological metrics confirm that carbocyclic systems occupy a more constrained and highly structured manifold, whereas the inclusion of nitrogen acts as a decoherence agent that populates a more diffuse and geometrically deformed electronic space.

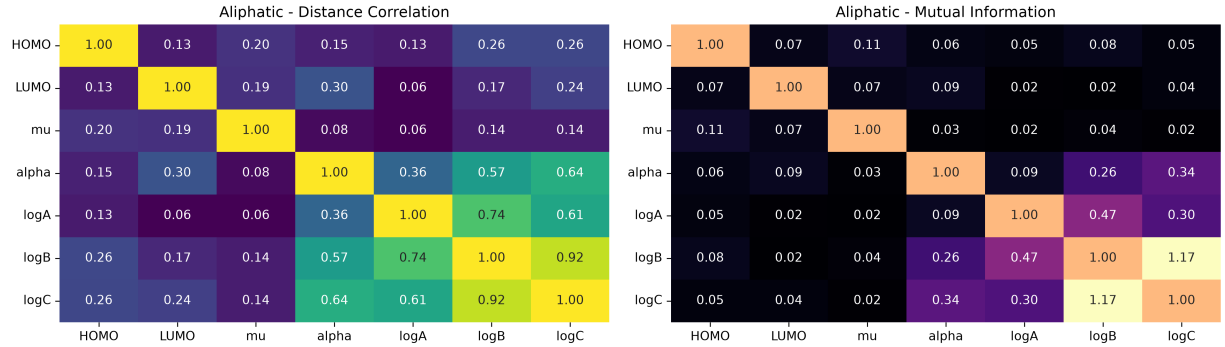
4 Conclusions

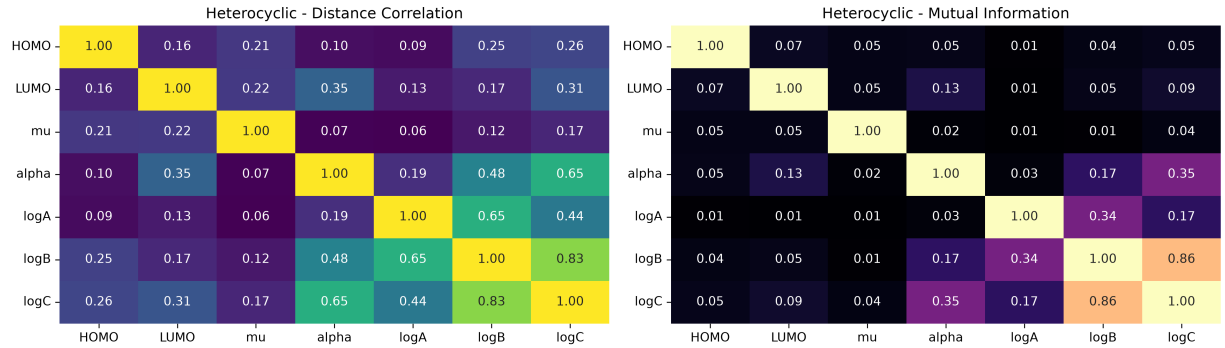
In this work, we investigated the topological and statistical rewiring of the frontier electronic manifold across diverse structural motifs. By extracting gradient-boosted atom-count residuals from the QM9 dataset, we successfully isolated intrinsic chemical effects from baseline stoichiometric scaling. This approach allowed us to rigorously evaluate whether the energetic orthogonality of the highest occupied molecular orbital ($HOMO$) is a universally rigid feature or a structurally sensitive state space.

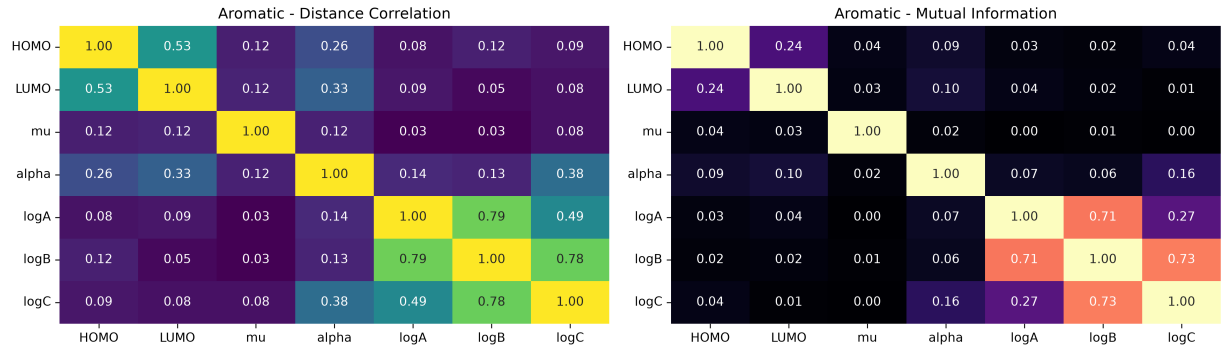
Our analysis reveals that the joint structure of molecular properties is a dynamic network that rewires based on structural motifs. The defining feature of this dataset is the conditional isolation of the $HOMO$. As demonstrated in Table 1.3, in over 80% of the dataset—comprising aliphatic and non-aromatic heterocyclic systems—the $HOMO$ energy acts as a decoupled lone property, largely independent of global spatial and energetic organization.

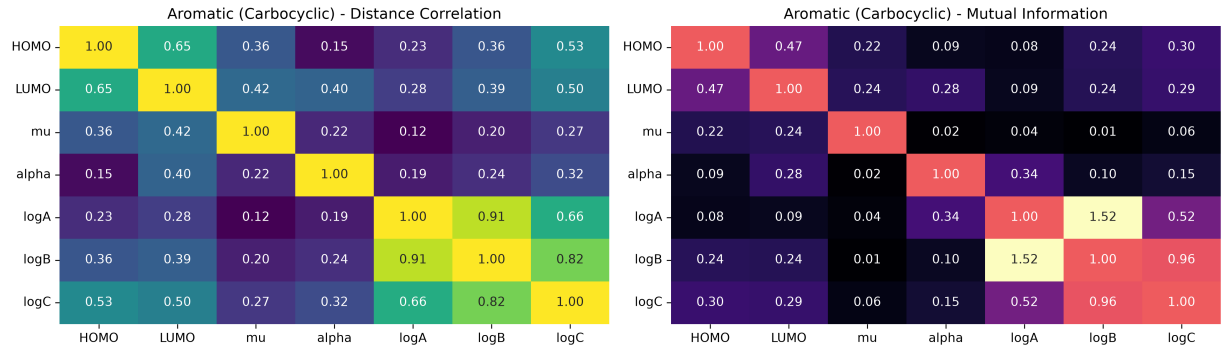
The onset of aromaticity acts as a fundamental phase transition within the electronic manifold. In aromatic systems, the isolation of the $HOMO$ collapses, and it is pulled into the joint manifold where it couples strongly with the lowest unoccupied molecular orbital ($LUMO$) and overall molecular geometry.

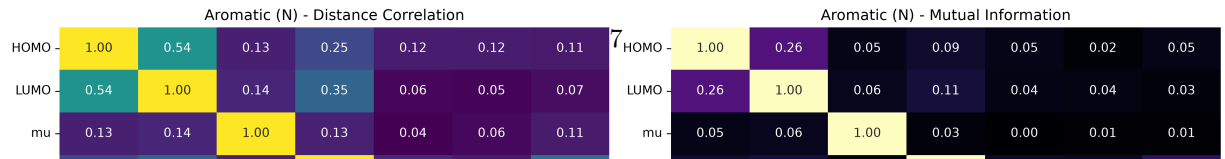

Global - Distance Correlation
HOMO 1.00 0.20 0.18 0.14 0.09 0.24 0.22
LUMO 0.20 1.00 0.18 0.36 0.08 0.19 0.29
mu 0.18 0.18 1.00 0.07 0.07 0.10 0.13
alpha 0.14 0.36 0.07 1.00 0.22 0.47 0.64
logA 0.09 0.08 0.07 0.22 1.00 0.69 0.49
logB 0.24 0.19 0.10 0.47 0.69 1.00 0.85
logC 0.22 0.29 0.13 0.64 0.49 0.85 1.00
HOMO LUMO mu alpha logA logB logC
Global - Mutual Information
HOMO 1.00 0.08 0.06 0.06 0.05 0.04 0.07
LUMO 0.08 1.00 0.02 0.11 0.01 0.06 0.08
mu 0.06 0.02 1.00 0.03 0.04 0.02 0.02
alpha 0.06 0.11 0.03 1.00 0.05 0.17 0.35
logA 0.05 0.01 0.04 0.05 1.00 0.39 0.22
logB 0.04 0.06 0.02 0.17 0.39 1.00 0.93
logC 0.07 0.08 0.02 0.35 0.22 0.93 1.00
HOMO LUMO mu alpha logA logB logC


Aliphatic - Distance Correlation
HOMO 1.00 0.13 0.20 0.15 0.13 0.26 0.26
LUMO 0.13 1.00 0.19 0.30 0.06 0.17 0.24
mu 0.20 0.19 1.00 0.08 0.06 0.14 0.14
alpha 0.15 0.30 0.08 1.00 0.36 0.57 0.64
logA 0.13 0.06 0.06 0.36 1.00 0.74 0.61
logB 0.26 0.17 0.14 0.57 0.74 1.00 0.92
logC 0.26 0.24 0.14 0.64 0.61 0.92 1.00
HOMO LUMO mu alpha logA logB logC
Aliphatic - Mutual Information
HOMO 1.00 0.07 0.11 0.06 0.05 0.08 0.05
LUMO 0.07 1.00 0.07 0.09 0.02 0.02 0.04
mu 0.11 0.07 1.00 0.03 0.02 0.04 0.02
alpha 0.06 0.09 0.03 1.00 0.09 0.26 0.34
logA 0.05 0.02 0.02 0.09 1.00 0.47 0.30
logB 0.08 0.02 0.04 0.26 0.47 1.00 1.17
logC 0.05 0.04 0.02 0.34 0.30 1.17 1.00
HOMO LUMO mu alpha logA logB logC


Heterocyclic - Distance Correlation
HOMO 1.00 0.16 0.21 0.10 0.09 0.25 0.26
LUMO 0.16 1.00 0.22 0.35 0.13 0.17 0.31
mu 0.21 0.22 1.00 0.07 0.06 0.12 0.17
alpha 0.10 0.35 0.07 1.00 0.19 0.48 0.65
logA 0.09 0.13 0.06 0.19 1.00 0.65 0.44
logB 0.25 0.17 0.12 0.48 0.65 1.00 0.83
logC 0.26 0.31 0.17 0.65 0.44 0.83 1.00
HOMO LUMO mu alpha logA logB logC
Heterocyclic - Mutual Information
HOMO 1.00 0.07 0.05 0.05 0.01 0.04 0.05
LUMO 0.07 1.00 0.05 0.13 0.01 0.05 0.09
mu 0.05 0.05 1.00 0.02 0.01 0.01 0.04
alpha 0.05 0.13 0.02 1.00 0.03 0.17 0.35
logA 0.01 0.01 0.01 0.03 1.00 0.34 0.17
logB 0.04 0.05 0.01 0.17 0.34 1.00 0.86
logC 0.05 0.09 0.04 0.35 0.17 0.86 1.00
HOMO LUMO mu alpha logA logB logC


Aromatic - Distance Correlation
HOMO 1.00 0.53 0.12 0.26 0.08 0.12 0.09
LUMO 0.53 1.00 0.12 0.33 0.09 0.05 0.08
mu 0.12 0.12 1.00 0.12 0.03 0.03 0.08
alpha 0.26 0.33 0.12 1.00 0.14 0.13 0.38
logA 0.08 0.09 0.03 0.14 1.00 0.79 0.49
logB 0.12 0.05 0.03 0.13 0.79 1.00 0.78
logC 0.09 0.08 0.08 0.38 0.49 0.78 1.00
HOMO LUMO mu alpha logA logB logC
Aromatic - Mutual Information
HOMO 1.00 0.24 0.04 0.09 0.03 0.02 0.04
LUMO 0.24 1.00 0.03 0.10 0.04 0.02 0.01
mu 0.04 0.03 1.00 0.02 0.00 0.01 0.00
alpha 0.09 0.10 0.02 1.00 0.07 0.06 0.16
logA 0.03 0.04 0.00 0.07 1.00 0.71 0.27
logB 0.02 0.02 0.01 0.06 0.71 1.00 0.73
logC 0.04 0.01 0.00 0.16 0.27 0.73 1.00
HOMO LUMO mu alpha logA logB logC


Aromatic (Carbocyclic) - Distance Correlation
HOMO 1.00 0.65 0.36 0.15 0.23 0.36 0.53
LUMO 0.65 1.00 0.42 0.40 0.28 0.39 0.50
mu 0.36 0.42 1.00 0.22 0.12 0.20 0.27
alpha 0.15 0.40 0.22 1.00 0.19 0.24 0.32
logA 0.23 0.28 0.12 0.19 1.00 0.91 0.66
logB 0.36 0.39 0.20 0.24 0.91 1.00 0.82
logC 0.53 0.50 0.27 0.32 0.66 0.82 1.00
HOMO LUMO mu alpha logA logB logC
Aromatic (Carbocyclic) - Mutual Information
HOMO 1.00 0.47 0.22 0.09 0.08 0.24 0.30
LUMO 0.47 1.00 0.24 0.28 0.09 0.24 0.29
mu 0.22 0.24 1.00 0.02 0.04 0.01 0.06
alpha 0.09 0.28 0.02 1.00 0.34 0.10 0.15
logA 0.08 0.09 0.04 0.34 1.00 1.52 0.52
logB 0.24 0.24 0.01 0.10 1.52 1.00 0.96
logC 0.30 0.29 0.06 0.15 0.52 0.96 1.00
HOMO LUMO mu alpha logA logB logC


Aromatic (N) - Distance Correlation
HOMO 1.00 0.54 0.13 0.25 0.12 0.12 0.11
LUMO 0.54 1.00 0.14 0.35 0.06 0.05 0.07
mu 0.13 0.14 1.00 0.13 0.04 0.06 0.11
Aromatic (N) - Mutual Information
HOMO 1.00 0.26 0.05 0.09 0.05 0.02 0.05
LUMO 0.26 1.00 0.06 0.11 0.04 0.04 0.03
mu 0.05 0.06 1.00 0.03 0.00 0.01 0.01

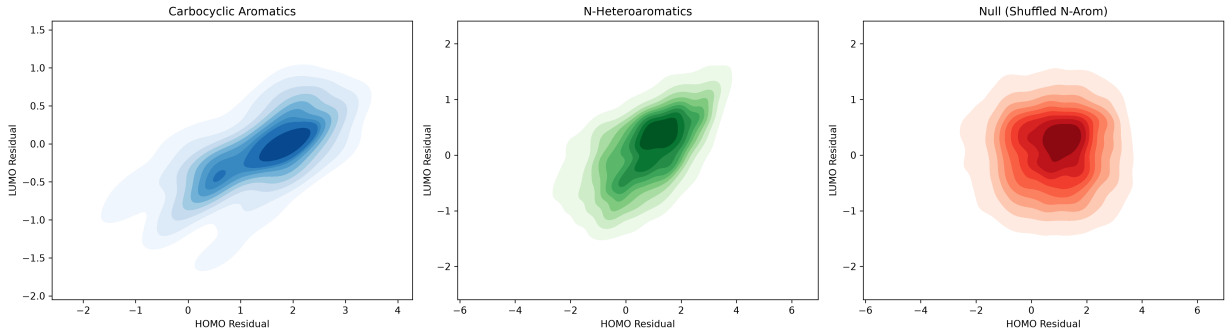


Figure 4: Bivariate density distributions of the $LUMO$ residual versus the $HOMO$ residual for carbocyclic aromatics (left), N-heteroaromatics (center), and a shuffled null baseline (right). By isolating chemical effects from stoichiometric scaling, these distributions visually demonstrate the rewiring of the frontier electronic manifold across different structural motifs. Carbocyclic aromatics exhibit exceptionally strong orbital-orbital coupling ($dCor = 0.6514$), reflecting the high symmetry and uniform conjugation of pure carbon frameworks. In contrast, the inclusion of nitrogen heteroatoms acts as a decoherence agent that significantly weakens this coupling ($dCor = 0.5353$) and populates a more diffuse electronic space, although the heteroaromatic manifold remains distinctly more structured than the completely decoupled null model.

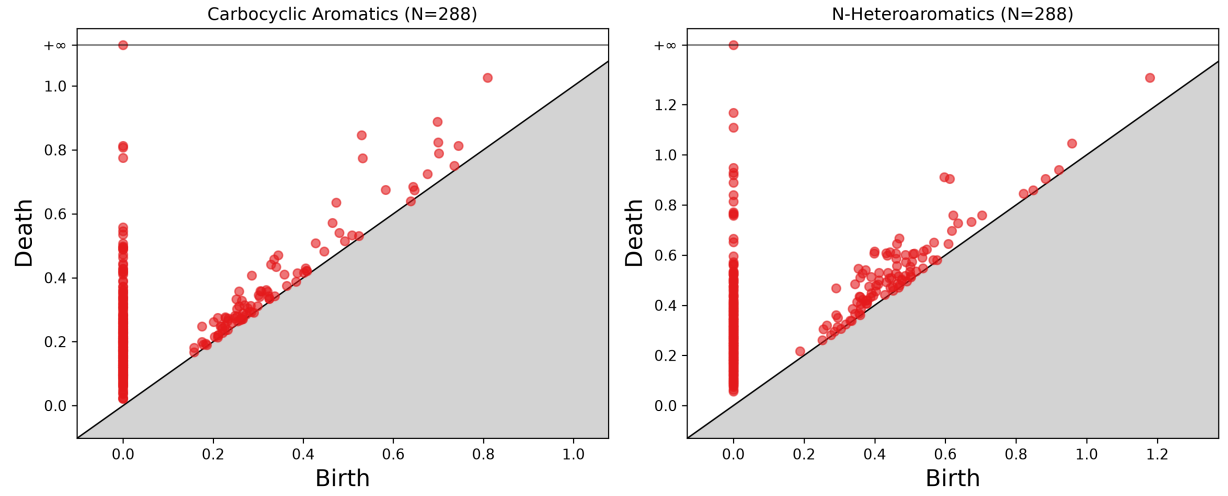


Figure 5: Persistence diagrams comparing the topological features of the electronic manifold for carbocyclic aromatics (left) and N-heteroaromatics (right) using balanced samples ($N = 288$). Computed via Vietoris-Rips persistent homology on the 3D space defined by the residuals of $HOMO$, $LUMO$, and polarisability α, the plots illustrate the geometric deformation caused by nitrogen heteroatoms. The differences in the birth-death distributions yield significant bottleneck distances for connected components ($H_0 = 0.372512$) and loops ($H_1 = 0.108244$). These topological variations demonstrate that while carbocyclic systems occupy a constrained, highly structured electronic manifold, the introduction of nitrogen acts as a decoherence agent, resulting in a more diffuse electronic space with fundamentally altered clustering densities and voids.

This integration signifies a shift from localized orbital character to a regime of delocalized electronic states that are intrinsically tied to the planar molecular framework.

Furthermore, we demonstrated that the introduction of nitrogen or oxygen heteroatoms into aromatic rings acts as a decoherence agent. While purely carbocyclic aromatics exhibit exceptionally strong orbital-orbital coupling, differing electronegative lone-pair environments in heteroaromatics significantly weaken these interactions. This decoherence is accompanied by a geometric deformation of the electronic state space, confirmed by significant topological bottleneck distances for both connected components (H_0) and one-dimensional loops (H_1).

Ultimately, this information-geometric framework establishes that the orthogonality of the $HOMO$ is not a universal constant, but a highly structured property that dynamically rewires in response to local topological perturbations. These findings provide a more nuanced understanding of how localized electronic states couple to global molecular properties, offering valuable insights into the fundamental physical chemistry that governs molecular design and reactivity.