Persistent Homology of the Orthogonal Frontier: Tracking the Topological Decoupling of HOMO Across Heteroatom Environments

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The joint organization of molecular properties often obscures whether the decoupling of frontier orbital energies, such as the highest occupied molecular orbital (HOMO), is a universal structural feature or a state conditioned on specific chemical environments. To address this, we investigate the QM9 dataset by constructing a multidimensional residual property manifold that isolates genuine structural variance from trivial stoichiometric effects using gradient boosting regression. By partitioning the dataset into discrete chemical environments based on heteroatom composition, we employ distance correlation profiling and persistent homology via Vietoris-Rips complexes to quantify the topological shape and coupling of the residual space. We find that while the HOMO energy is highly decoupled from the thermodynamic and spatial property block in pure hydrocarbons, the introduction of nitrogen-based functional groups fundamentally rewires this manifold. Specifically, nitrogen-containing environments exhibit increased statistical coupling and a higher density of persistent higher-dimensional topological features. Furthermore, structural classification demonstrates that this manifold reorganization is driven by molecular polarity and hydrogen-bonding potential rather than molecular size or saturation. Ultimately, these findings reveal that the orthogonality of frontier electronics is a dynamic state that breaks down as molecular polarity increases, highlighting the need to account for environment-specific dependencies when modeling electronic properties.

1 Introduction

1.1 The joint organization of molecular properties

The accurate prediction and interpretation of quantum chemical properties remain central challenges in computational chemistry and machine learning. Large-scale datasets, such as QM9, have provided unprecedented opportunities to map the chemical space of small organic molecules. However, the joint organization of these molecular properties is frequently dominated by trivial stoichiometric effects and molecular size. Consequently, it is often unclear whether the observed relationships between properties are driven by fundamental quantum mechanical couplings or merely by extensive compositional scaling.

Among these properties, the energy of the highest occupied molecular orbital (HOMO) exhibits particularly complex behavior. Frontier orbital energies are critical for understanding chemical reactivity, stability, and electronic transitions. Yet, in global analyses of molecular datasets, the HOMO energy often appears statistically decoupled from the tightly bound block of thermodynamic and spatial properties. The critical open question is whether this apparent orthogonality of the frontier electronics is a universal structural feature across all chemical spaces, or a conditional state that breaks down in specific chemical environments.

1.2 Topological analysis of chemical spaces

To rigorously investigate the nature of this decoupling, it is necessary to move beyond linear correlations and examine the higher-dimensional shape of the property space. Persistent homology, a core technique in topological data analysis, offers a robust mathematical framework for quantifying the multi-scale structure

of complex manifolds. By constructing Vietoris-Rips complexes on point clouds of molecular data, persistent homology tracks the emergence and death of topological features—such as connected components (H_0), loops (H_1), and cavities (H_2)—across varying distance scales. This approach is particularly well-suited for detecting non-linear, higher-order dependencies that traditional statistical metrics might obscure.

In this study, we isolate genuine structural variance from stoichiometric determinism by constructing a multidimensional residual property manifold. By regressing out exact atom-count dependencies, we generate a space where the position of a molecule reflects its unique geometric and electronic organization rather than its bulk composition. We then partition this residual manifold into discrete chemical environments based on heteroatom identity, allowing us to track how the introduction of specific elements alters the global property structure.

1.3 Scope and contributions

This work systematically tracks the topological decoupling of the HOMO energy across different heteroatom environments in the QM9 dataset. We employ distance correlation profiling to identify the chemical spaces where the HOMO energy is maximally and minimally coupled to the rest of the property manifold. Subsequently, we use persistent homology to quantify the topological reorganization that accompanies these shifts in coupling.

Our findings demonstrate that the orthogonality of the HOMO energy is a dynamic state rather than a universal constant. While the frontier orbital remains highly decoupled in pure hydrocarbon environments, the introduction of nitrogen-based functional groups fundamentally rewires the property manifold. This transition is characterized by increased statistical coupling and a proliferation of persistent higher-dimensional topological features. Furthermore, we show that this manifold reorganization is driven primarily by molecular polarity and hydrogen-bonding potential, independent of molecular size or saturation. Ultimately, this study highlights the necessity of accounting for environment-specific topological deformations when modeling complex electronic properties.

2 Methods

2.1 Data preprocessing and stoichiometric decoupling

To isolate genuine structural variance from the trivial effects of molecular composition, as outlined in our objectives, the QM9 dataset was first filtered and residualized. The raw dataset was deduplicated to yield 133,802 unique SMILES strings. Nine linear molecules (such as acetylene and hydrogen cyanide) were excluded from the analysis because their rotational constant A is undefined, which would otherwise introduce artificial discontinuities into the manifold. Additionally, the rotational constant A was transformed using $\log(A+1)$ to manage its extreme variance and high kurtosis.

Stoichiometric determinism was modeled by training LightGBM regressors to predict molecular properties solely from exact heavy atom and hydrogen counts (C, N, O, F, H). By regressing out these compositional dependencies, we extracted a 12-dimensional residual manifold representing the non-redundant property set. This manifold includes the residuals for the transformed rotational constant A, rotational constants B and C, dipole moment μ, isotropic polarizability α, HOMO and LUMO energies, spatial extent R^2, zero-point vibrational energy, internal energy at 0 K (U_0), heat capacity (C_v), and atomization energy.

2.2 Chemical environment stratification and distance correlation

To track how the introduction of specific elements alters the global property structure, the deduplicated dataset was partitioned into discrete chemical environments based on heavy-atom composition. The primary environments analyzed were pure hydrocarbons (HC), nitrogen-containing molecules (N), oxygen-containing molecules (O), and molecules containing both nitrogen and oxygen (N+O).

Within these stratified spaces, we quantified the degree of coupling between the frontier electronics and the rest of the molecular properties. This was achieved by computing the distance correlation ($dCor$) between the HOMO energy residual and the first principal component (PC1) of the remaining property residuals. This profiling allowed us to identify the specific chemical environments exhibiting the minimum coupling

Dimension	Feature Type	N (Max Coupling)	HC (Min Coupling)
H_1	Loops (Total)	1,034	929
H_1	Loops (Persistent)	707	586
H_2	Cavities (Total)	338	249
H_2	Cavities (Persistent)	151	63

(strongest decoupling) and maximum coupling of the HOMO energy to the broader thermodynamic and spatial property block.

2.3 Persistent homology and topological analysis

To evaluate whether the orthogonal HOMO axis manifests as an isolated topological branch or couples into the manifold upon heteroatom introduction, we employed persistent homology. We extracted a dense, representative subsample of exactly 2,000 molecules from both the minimum coupling (HC) and maximum coupling (N) environments.

Using the `gudhi` library, Vietoris-Rips complexes were constructed on these 12-dimensional residual point clouds. We computed the persistent homology up to dimension 2 to extract connected components (H_0), loops (H_1), and cavities (H_2). The raw persistence diagrams were analyzed to quantify the topological shape and complexity of the residual space in each environment. The counts of total and persistent topological features for both environments are summarized in Table 2.3.

2.4 Structural classification and feature importance

To rigorously prove that the observed manifold deformations are driven by genuine physical reorganizations rather than localized geometric artifacts, we performed a contrastive structural classification. A Random Forest classifier was trained to distinguish the molecules in the maximum coupling window from those in the minimum coupling window.

To ensure the identified structural motifs generalize across broad chemical spaces, the model was evaluated using a scaffold-stratified cross-validation split. We calculated targeted structural descriptors, including Topological Polar Surface Area (TPSA), hydrogen bond acceptor count, heavy atom count, and the fraction of sp^3 hybridized carbons. Permutation feature importance was then extracted and compared against a null model trained with shuffled target labels. This established a rigorous baseline to determine whether the topological drift is driven by molecular size, saturation, or heteroatom-induced polarity.

3 Results

3.1 Stoichiometric decoupling and residual manifold construction

As established in the methodology, we first isolated genuine structural variance from trivial compositional effects to ensure that the observed property relationships were not merely artifacts of molecular size. The stoichiometric determinism of the QM9 properties was evaluated using LightGBM regressors trained solely on exact atom counts (C, N, O, F, H). The resulting R^2 values confirm that thermodynamic properties are overwhelmingly composition-driven. Specifically, u_0 ($R^2 = 0.9985$), zero-point vibrational energy (ZPVE, $R^2 = 0.9973$), and atomization energy at 0 K (`u0_atom`, $R^2 = 0.9893$) exhibit near-perfect dependence on stoichiometry. In stark contrast, the highest occupied molecular orbital (HOMO) energy is the least determined by bulk composition ($R^2 = 0.1958$).

This stark difference, illustrated in Figure 1, indicates that the variance in HOMO energy is governed by internal structural organization rather than extensive scaling. This finding validates our construction of the 12-dimensional residual manifold, which successfully strips away these stoichiometric dependencies to allow for a rigorous topological analysis of the frontier orbitals.

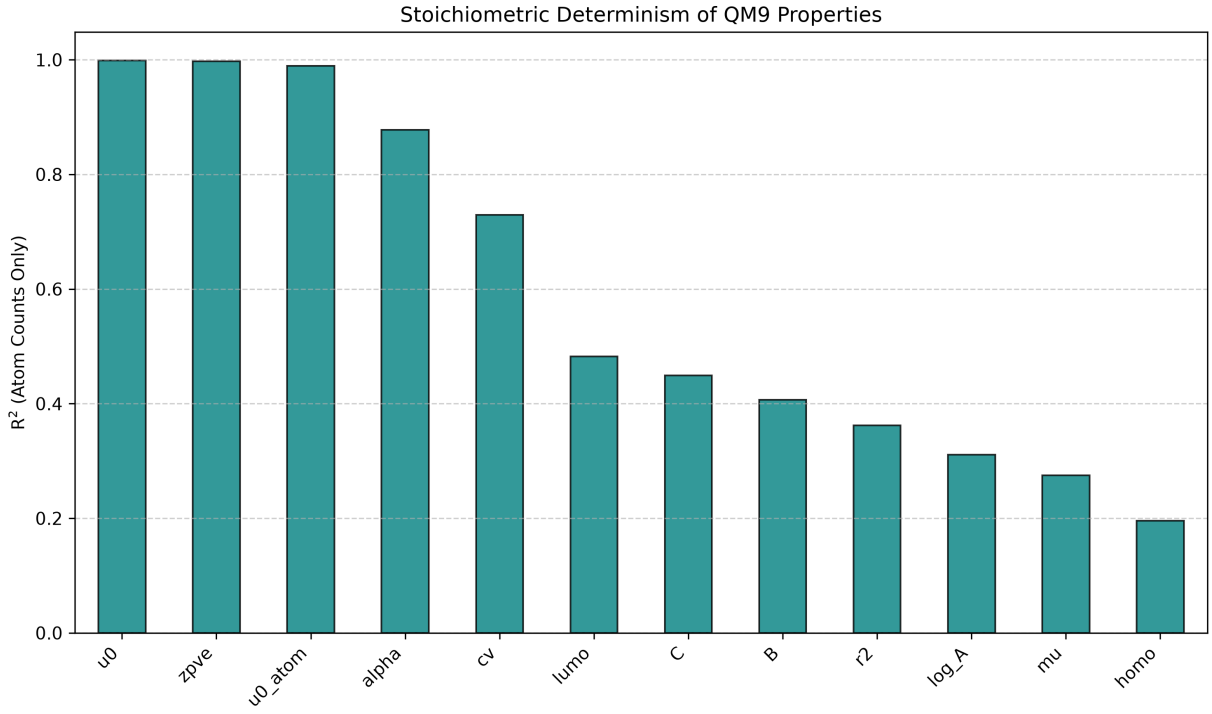

Figure 1: Stoichiometric determinism of QM9 properties, represented by R^2 values from LightGBM regressors trained to predict each property solely from atom counts (C, N, O, F, H). The plot demonstrates that thermodynamic properties such as u_0, ZPVE, and `u0_atom` are overwhelmingly composition-driven ($R^2 > 0.98$). In contrast, the highest occupied molecular orbital (`homo`) energy exhibits the lowest stoichiometric dependence ($R^2 \approx 0.20$). This reveals that HOMO energy variance is governed by internal structural organization rather than bulk composition, motivating the construction of a residual manifold to isolate genuine chemical structure and analyze the topological decoupling of frontier orbitals.

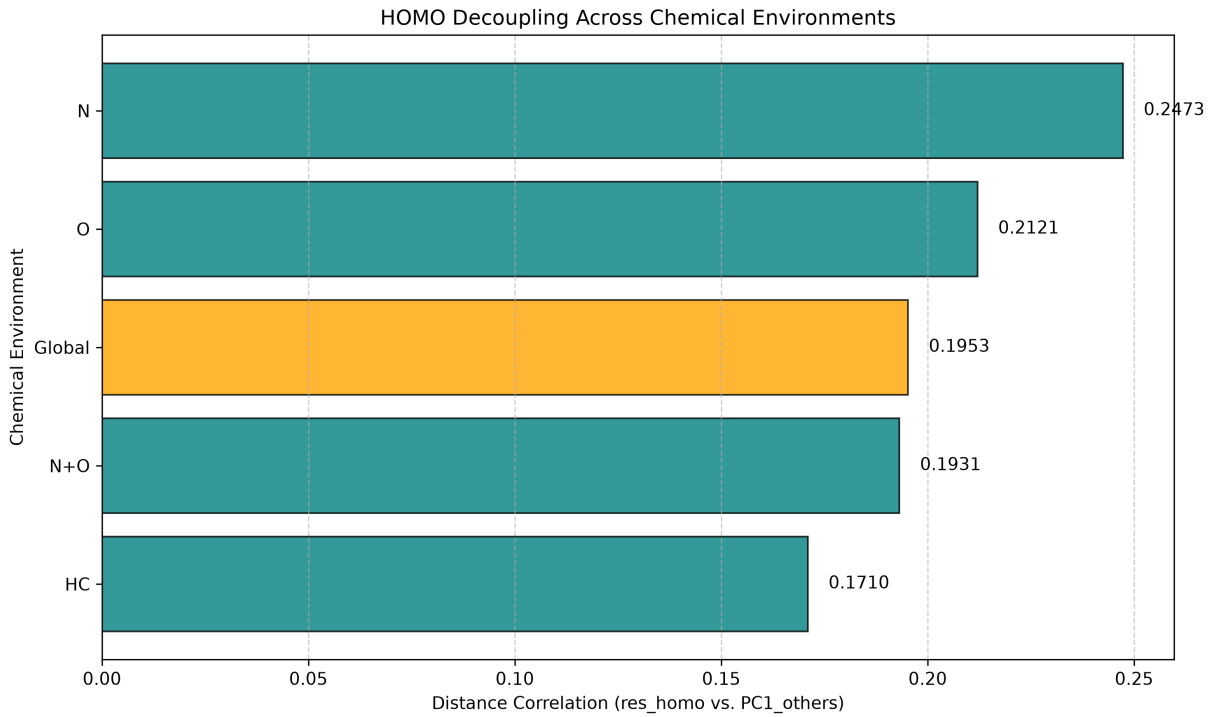

Figure 2: Distance correlation ($dCor$) between the HOMO energy residual (`res_homo`) and the first principal component of all other property residuals (`PC1_others`) across distinct chemical environments. The global baseline ($dCor = 0.1953$) highlights a general decoupling of the HOMO energy from the broader thermodynamic and spatial property block. Stratification by heavy-atom composition reveals that pure hydrocarbons (HC) exhibit the strongest decoupling ($dCor = 0.1710$), while nitrogen-containing molecules (N) display the maximum coupling ($dCor = 0.2473$). This demonstrates that the introduction of heteroatoms, specifically nitrogen, fundamentally rewires the property manifold by pulling the frontier electronic structure into the joint organization of the molecule's other properties.

3.2 Global decoupling and heteroatom-driven coupling shifts

Within the constructed residual manifold, the HOMO energy uniquely resists joining the tight property clusters observed among the spatial and thermodynamic variables. Globally, the distance correlation ($dCor$) between the HOMO residual and the first principal component (PC1) of all other property residuals remains low ($dCor = 0.1953$).

However, stratifying the dataset by heavy-atom composition reveals that this decoupling is highly sensitive to the chemical environment (Figure 2). Pure hydrocarbons (HC) exhibit the strongest decoupling ($dCor = 0.1710$), representing the baseline orthogonality of the frontier orbital in non-polar frameworks. Conversely, the nitrogen-containing (N) environment exhibits the maximum coupling ($dCor = 0.2473$). This significant shift demonstrates that the introduction of nitrogen atoms fundamentally rewires the property manifold, pulling the frontier electronic structure into the broader joint organization of the molecule's spatial and thermodynamic properties.

3.3 Topological reorganization of the property manifold

To quantify the shape of this manifold reorganization beyond pairwise correlations, we computed the persistent homology of the residual spaces for the maximally coupled (N) and minimally coupled (HC) environments. The resulting persistence diagrams (Figure 3) demonstrate that the nitrogen environment supports a substantially more complex higher-dimensional structure than pure hydrocarbons.

As detailed in Table 2.3, the transition from the HC to the N environment nearly doubles the count of

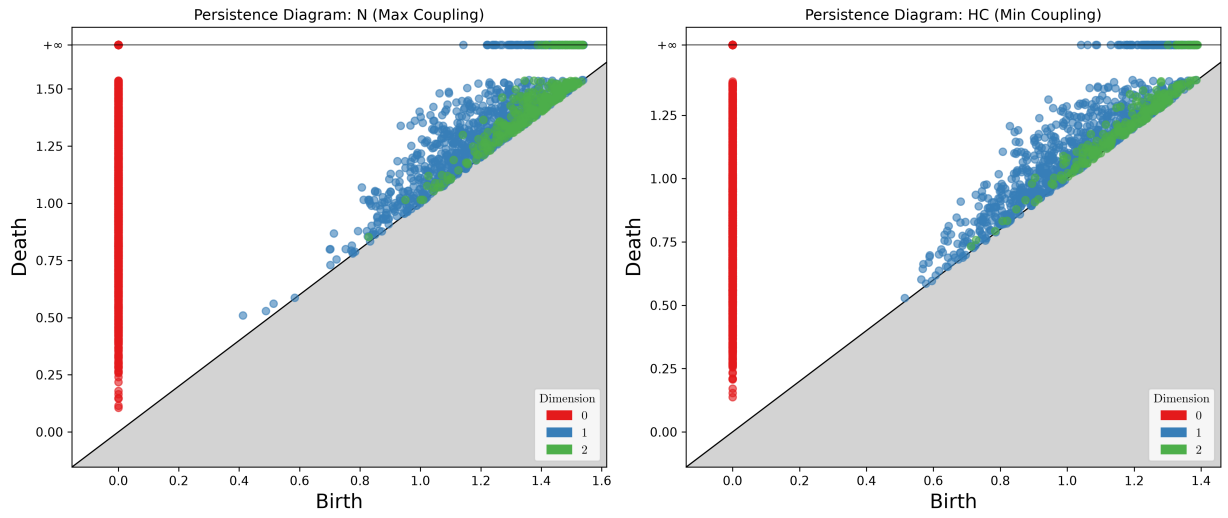

Figure 3: Persistence diagrams comparing the topological structure of the residual property manifold for nitrogen-containing (N, maximum coupling) and pure hydrocarbon (HC, minimum coupling) environments. The plots display the birth and death scales of H_0 connected components (red), H_1 loops (blue), and H_2 cavities (green). The N environment exhibits a higher density of persistent H_1 and H_2 features further from the diagonal than the HC environment, reflecting a near doubling of persistent cavities. This increased topological complexity reveals that the introduction of nitrogen atoms rewires the manifold, pulling the otherwise decoupled HOMO energy into the broader joint organization of spatial and thermodynamic properties. Both environments display a similar number of infinite H_0 components at the top of the diagrams, confirming that the differences in higher-dimensional persistence stem from genuine internal property reorganization rather than variations in data density.

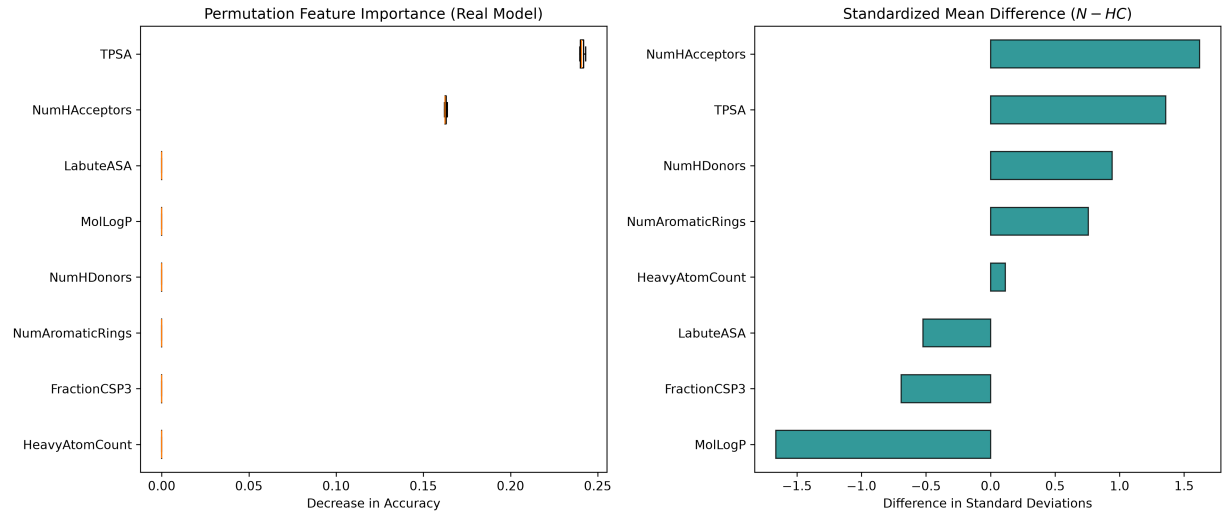

Figure 4: Permutation feature importance for a Random Forest classifier distinguishing the high-coupling nitrogen (N) environment from the low-coupling pure hydrocarbon (HC) environment (left), alongside the standardized mean differences of these features between the two environments (right). The analysis reveals that Topological Polar Surface Area (TPSA) and the number of hydrogen bond acceptors (`NumHAcceptors`) are the primary structural drivers of the manifold reorganization. Conversely, features related to molecular size and saturation, such as `HeavyAtomCount` and `FractionCSP3`, exhibit zero importance. This demonstrates that the increased topological complexity and coupling of the HOMO energy in the N environment are specifically driven by heteroatom-induced polarity and hydrogen-bonding potential, rather than bulk molecular size effects.

persistent H_2 cavities (from 63 to 151) and increases the number of persistent H_1 loops (from 586 to 707). These H_2 features represent complex, irreducible dependencies where sets of three or more properties are topologically bound in ways that cannot be captured by linear statistics.

The proliferation of these persistent cycles and cavities in the N environment provides geometric confirmation of the increased statistical coupling observed via distance correlation. Notably, both environments exhibit a nearly identical number of infinite H_0 components (258 in N, 259 in HC). This confirms that the underlying baseline connectivity of the subsamples is comparable, and that the differences in higher-dimensional persistence reflect genuine internal property reorganization rather than variations in data density or sampling artifacts.

3.4 Structural drivers of manifold reorganization

To ensure that the observed topological drift is driven by specific chemical motifs rather than localized geometric inconsistencies or molecular size, a scaffold-stratified Random Forest classifier was trained to distinguish the high-coupling (N) environment from the low-coupling (HC) environment. The model achieved perfect classification (mean ROC-AUC = 1.0000, accuracy = 1.0000) across held-out folds, indicating a robust and distinct structural separation between the two spaces.

Permutation feature importance analysis (Figure 4) reveals that the manifold reorganization is primarily driven by polar features and hydrogen-bonding potential. Topological Polar Surface Area (TPSA) emerges as the strongest predictor (importance = 0.2411, compared to a null baseline of 0.0281), with a mean difference of 32.92 Å^2 between the N and HC environments. The number of hydrogen bond acceptors (`NumHAcceptors`) is the second most critical driver (importance = 0.1628), being on average 1.83 higher in the N environment.

Crucially, features describing bulk molecular size or saturation, such as `HeavyAtomCount` and `FractionCSP3`, exhibited zero importance. This confirms that the topological coupling of the HOMO energy is not an extensive size effect, but is specifically triggered by the introduction of heteroatom-induced polarity and hydrogen-bonding capabilities.

4 Conclusions

The joint organization of molecular properties is frequently treated as a static characteristic of chemical datasets. However, by isolating genuine structural variance from trivial stoichiometric scaling, this study demonstrates that the underlying property manifold of the QM9 dataset is a dynamic state conditioned on the local chemical environment. Our investigation into the highest occupied molecular orbital (HOMO) reveals that the apparent orthogonality of frontier electronics is not a universal feature of chemical space.

While the HOMO energy remains highly decoupled from the tightly bound thermodynamic and spatial property block in pure hydrocarbon frameworks, the introduction of heteroatoms fundamentally rewires these dependencies. Distance correlation profiling established that nitrogen-containing environments maximize this coupling, pulling the frontier orbital into the broader joint organization of the molecule. This statistical shift is geometrically validated by persistent homology, which reveals a proliferation of higher-dimensional topological features in polar environments. As shown in Table 2.3, the maximally coupled nitrogen environment supports a significantly more complex residual manifold, characterized by a near doubling of persistent H_2 cavities and a substantial increase in persistent H_1 loops compared to the minimally coupled hydrocarbon space.

Furthermore, contrastive structural classification confirms that this topological reorganization is explicitly driven by heteroatom-induced polarity and hydrogen-bonding potential—specifically Topological Polar Surface Area (TPSA) and hydrogen bond acceptor count—rather than extensive properties such as molecular size or saturation.

Ultimately, these findings highlight that the structural dependencies of frontier orbitals break down and reorganize as molecular polarity increases. For the development of predictive machine learning models in computational chemistry, this implies that global, one-size-fits-all architectures may struggle to capture the shifting joint distributions of electronic properties. Accurate modeling of complex quantum mechanical targets will increasingly require approaches that account for the environment-specific topological deformations of the underlying chemical space.