Persistent Homology Reveals the Structural Dependence of HOMO Energy Decoupling in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy often appears decoupled from the global thermodynamic and spatial property manifold of molecules. It remains unclear whether this orthogonality is a universal characteristic or a consequence of specific structural topologies. To address this, we analyze the joint structure of molecular properties in the QM9 dataset by tracking the topological shape of their non-linear dependency networks across discrete chemical families. We extract stoichiometric-adjusted residuals for eleven non-redundant properties using gradient-boosted regressors and partition the molecules based on aromaticity and heteroatom identity. By applying persistent homology to Vietoris-Rips complexes constructed from these multidimensional residual point clouds, we generate Persistence Landscapes that capture the robust topological signatures of each structural class. Comparing the landscape distances of the true property set against a statistical control with shuffled HOMO values allows us to isolate the topological rewiring directly attributable to the HOMO energy. Our analysis reveals that the property manifold is highly stable within structural families but undergoes significant topological reorganization across chemical boundaries. Crucially, HOMO decoupling is not a universal constant. While HOMO behaves as a nearly independent variable in aliphatic hydrocarbons and oxygen-containing systems, it becomes topologically reintegrated in aromatic nitrogen heterocycles. This reintegration is evidenced by a strong divergence in the first homology group, reflecting the physical coupling of the HOMO energy to global electronic extent and polarizability in conjugated nitrogen systems.

1 Introduction

The exploration of chemical space increasingly relies on large-scale computational datasets, such as QM9, which provide a comprehensive mapping of molecular structures to their quantum mechanical and thermodynamic properties. A recurring observation in the analysis of these multidimensional property spaces is the apparent decoupling of the highest occupied molecular orbital (HOMO) energy from the global thermodynamic and spatial manifold. While properties such as polarizability (α), electronic extent (r^2), and heat capacity often exhibit strong collinearity driven by molecular size and stoichiometry, the HOMO energy frequently behaves as an orthogonal, independent variable.

It remains an open question whether this orthogonality is a universal characteristic of the quantum chemical space or a localized phenomenon dependent on specific structural topologies. Traditional statistical methods, such as linear covariance estimators, are often insufficient to resolve this anomaly. They are highly sensitive to stoichiometric size-leakage—where trivial atom-count variance dominates the underlying physical signal—and they struggle to capture complex, non-linear dependencies. Consequently, determining if orbital decoupling is a fundamental state of the molecular graph requires a framework capable of isolating genuine electronic rewiring from baseline structural scaling.

To address this challenge, we employ persistent homology, a technique from topological data analysis, to characterize the joint structure of molecular properties across discrete chemical families. We first extract stoichiometric-adjusted residuals for eleven non-redundant properties using gradient-boosted regressors, effectively removing the dominant variance associated with simple atom counts. The dataset is then partitioned into well-defined structural bins based on aromaticity and heteroatom identity. By constructing Vietoris-Rips complexes on these multidimensional residual point clouds, we generate Persistence Landscapes. These landscapes provide a robust, coordinate-free signature of the property manifold's shape, capturing multi-property dependencies as topological features such as connected components (H0) and high-dimensional voids (H1).

Crucially, we isolate the specific topological impact of the HOMO energy by comparing the landscape distances of the true property set against a statistical control in which HOMO values are randomly shuffled. This differential approach allows us to pinpoint exactly which structural transitions force the HOMO to reintegrate into the core spatial-thermodynamic organization.

Our analysis demonstrates that HOMO decoupling is not a universal constant of the QM9 dataset. While the HOMO energy behaves as a nearly independent variable in aliphatic hydrocarbons and oxygen-containing systems, it undergoes significant topological reintegration in aromatic nitrogen heterocycles. This reintegration, evidenced by a strong divergence in the first homology group, reflects the physical coupling of the HOMO energy to global electronic extent and polarizability in conjugated nitrogen systems. By tracking the topological shape of non-linear dependency networks, this work reveals the profound structural dependence of quantum property manifolds.

2 Methods

2.1 Data preparation and stoichiometric baseline

To rigorously evaluate the topological embedding of the highest occupied molecular orbital (*homo*) energy, we first isolated genuine electronic variance from trivial stoichiometric scaling. The QM9 dataset was loaded and deduplicated based on the first occurrence of each SMILES string to prevent data leakage from stereoisomers. We selected eleven non-redundant properties to define our multidimensional space: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), *homo* energy, electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity at 298.15 K (c_v), and atomization energy at 0 K (u_{0_atom}). The lowest unoccupied molecular orbital ($lumo$) energy was explicitly excluded to prevent mathematically constrained bridging signals that could artificially couple the *homo* to the broader manifold.

To address the extreme variance of the rotational constant A, we applied a $\log(A + 1)$ transformation. Furthermore, a binary indicator variable ('is_linear') was created to flag the nine linear molecules where $A = 0$, explicitly capturing this physical symmetry and preventing it from biasing downstream models.

To remove the dominant variance associated with molecular size and atom counts, we established a stoichiometric baseline. Exact atom counts for C, N, O, F, and H were computed from the SMILES strings. We then trained shallow gradient-boosted regressors (maximum depth of 3 to 4) for each of the eleven properties, using only the five atom counts and the 'is_linear' indicator as input features. The residuals from these models—representing the property variance independent of basic stoichiometry—were extracted and globally standardized to zero mean and unit variance.

2.2 Structural partitioning and manifold sampling

To track how the property manifold deforms across chemical boundaries, the dataset was partitioned into discrete structural topologies. Using RDKit, we computed two primary axes for each molecule: an aromaticity boolean flag (indicating the presence of at least one aromatic ring) and a heteroatom identity category (pure hydrocarbons, N-only, O-only, F-only, or mixed heteroatoms). Intersecting these axes yielded distinct structural bins (e.g., "Aliphatic pure hydrocarbon", "Aromatic N-only"). To ensure sufficient density for robust topological analysis, any bin containing fewer than 1,000 molecules was discarded.

Within each valid structural bin, we employed a Farthest Point Sampling (FPS) algorithm to select a representative subset of exactly 500 molecules. FPS was chosen over random sampling because it more effectively preserves the boundary and geometric shape of the underlying manifold. To ensure that subsequent topological comparisons focused strictly on the shape of the manifold rather than its absolute extent or density, the sampled residuals within each bin were MinMax scaled to a unit hypercube.

For each structural bin, we constructed three distinct point clouds: 1. An 11-dimensional cloud containing all property residuals. 2. A 10-dimensional cloud excluding the *homo* residual. 3. A Null 11-dimensional cloud, consisting of the 10-dimensional cloud augmented with randomly shuffled *homo* residuals drawn from the dataset, serving as a statistical control.

2.3 Persistent homology and topological representations

We utilized persistent homology to replace linear dependency metrics with a robust, coordinate-free signature of how the *homo* embeds within the global property manifold. Computations were parallelized across available CPUs. For each of the three point clouds in every structural bin, we constructed Vietoris-Rips complexes.

We computed the persistent homology up to dimension 1, capturing connected components (H0) and one-dimensional cycles or voids (H1). Higher dimensions (H2 and above) were excluded due to the sparsity of 500 points in an 11-dimensional space. The resulting persistence diagrams, which record the birth and death radii of topological features across the filtration, were transformed into Persistence Landscapes. This transformation provided stable, vector-based topological signatures suitable for quantitative distance measurements.

2.4 Distance metrics and signal isolation

To quantify the topological rewiring across structural boundaries, we computed pairwise Wasserstein distances between the persistence diagrams of different structural bins. Wasserstein distance was selected for its stability to noise and varying point densities. Separate pairwise distance matrices were constructed for the 11-dimensional, 10-dimensional, and Null 11-dimensional point clouds.

The specific topological impact of the *homo* energy—the "HOMO Signal"—was isolated by comparing these distance matrices. A genuine topological rewiring of the *homo* dependency network is indicated by a significant divergence between the 11-dimensional and 10-dimensional matrices for a specific pair of bins, provided this divergence is absent when comparing against the Null 11-dimensional matrix. This differential approach rigorously isolates the structural drivers of *homo* reintegration from statistical artifacts.

Finally, to validate the physical significance of the observed topological features, the entire FPS subsampling and persistent homology pipeline was repeated across five independent iterations per bin. We confirmed that intra-bin Wasserstein distances (noise across iterations) were significantly smaller than inter-bin distances (structural signal), ensuring the stability and reliability of the extracted topological manifolds.

3 Results

3.1 Stability of the residual property manifold

By isolating the residuals of eleven non-redundant properties—including rotational constants (A, B, C), electronic extent (r^2), dipole moment (μ), and orbital energies—we constructed a high-dimensional manifold to study the joint structure of molecular properties in the QM9 dataset, which is often obscured by stoichiometric effects. To ensure that the topological features extracted from the Vietoris-Rips complexes represent physical characteristics of the joint property organization rather than stochastic artifacts of the sampling process, we first validated the stability of the residual property manifold. This was achieved by comparing the intra-bin noise (variance across independent sampling iterations) to the inter-bin signal (distance between different structural families) across six distinct chemical classes.

As shown in Figure 1, the H_0 dimension, which captures the connectivity of the property cloud, demonstrates exceptional stability. The global mean stability ratio (inter-bin distance divided by intra-bin noise) consistently exceeds 15.0, with the Aliphatic O-only family achieving the highest H_0 stability ratio of 36.75. The H_1 dimension, representing one-dimensional cycles or "holes" in the property manifold, is inherently more sensitive to local density variations but remains highly stable with a global mean ratio of 2.20. Notably, the Aromatic N-only family exhibits the highest H_1 stability (ratio of 2.90), suggesting that the presence of aromatic nitrogen enforces a particularly rigid joint dependency structure among the molecular properties.

3.2 Topological impact of HOMO energy across structural boundaries

The HOMO energy is frequently observed to be decoupled from the global thermodynamic and spatial manifold in the QM9 dataset. To rigorously quantify the topological "rewiring" induced by the HOMO energy, we computed the L_2 landscape distance between the true 11-dimensional property set and a null

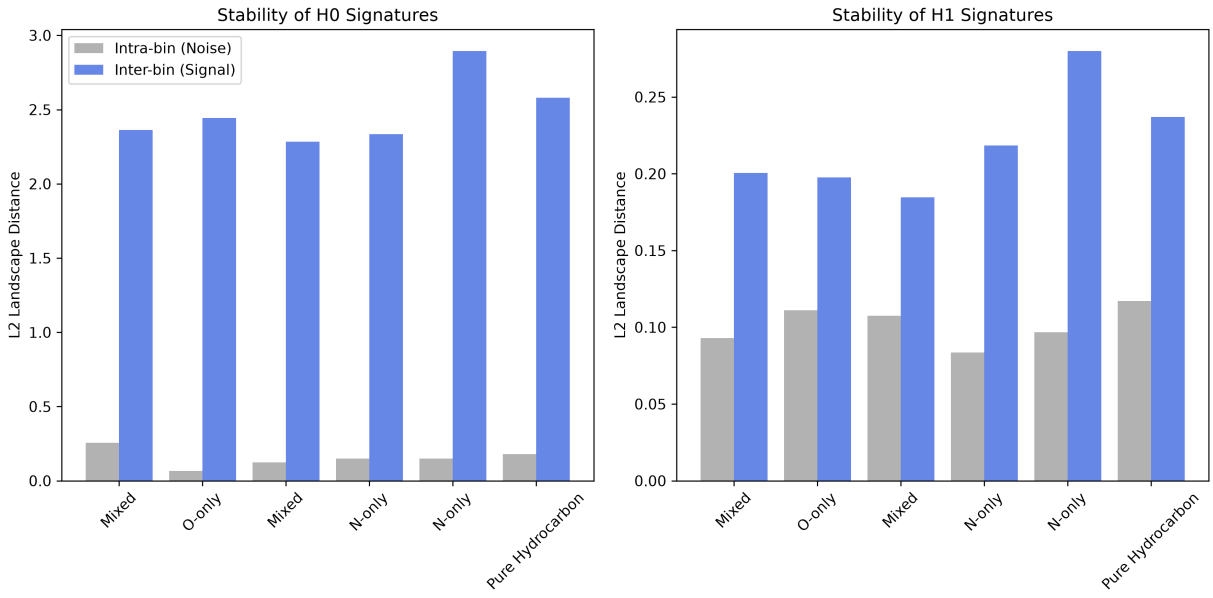

Figure 1: Comparison of intra-bin noise (grey) and inter-bin signal (blue), measured by L_2 landscape distance, for H_0 (left) and H_1 (right) topological signatures across six structural families. The consistently higher inter-bin signal validates the stability of the residual property manifold in the QM9 dataset, demonstrating that these topological features are robust, physical representations of joint property organization rather than stochastic artifacts. While the H_0 dimension (connectivity) exhibits exceptional stability overall, the H_1 dimension (1-cycles) remains stable despite higher sensitivity, with N-only families showing particularly high stability that suggests a rigid joint dependency structure.

11-dimensional set where HOMO values were randomly shuffled. This differential metric isolates the genuine topological signal attributable to the specific embedding of the HOMO energy.

The isolated HOMO signal reveals that the integration of the HOMO energy is not uniform but varies significantly across chemical families (Figure 2). The maximum signal divergence in the H_0 dimension reaches 0.50 between Aliphatic O-only and Aromatic N-only molecules, indicating a profound reorganization of the manifold's connectivity.

In the H_1 dimension, which captures complex multi-property dependencies, the isolation of the HOMO signal highlights distinct structural behaviors. For aromatic N-heterocycles, the signal divergence between Aromatic N-only and Aliphatic Pure Hydrocarbons is 0.17, the highest observed in the H_1 manifold. This pronounced shift suggests that the introduction of aromatic nitrogen fundamentally alters how the HOMO energy relates to the remaining ten properties. This is potentially due to the conjugation of nitrogen lone pairs into the π-system, which couples the electronic state energy to geometric responses (such as electronic extent r^2 and polarizability α) and rotational symmetry in a way not seen in aliphatic systems. Conversely, the Aliphatic O-only family exhibits negative signal shifts relative to several other bins (e.g., -0.09 against Aliphatic N-only). These negative shifts indicate that in aliphatic oxygen systems, the contribution of the HOMO energy to the joint topology closely resembles a random distribution, further confirming its decoupled nature in these specific structural contexts.

3.3 Landscape interpretation of chemical families

Persistence landscapes provide a direct visualization of the magnitude and scale of these topological features, allowing us to interpret the geometric structure of the property manifolds.

Comparing the H_0 landscapes (Figure 3), which represent the connectedness of the property cloud across filtration parameters, reveals stark differences between structural families. Aliphatic pure hydrocarbons exhibit broader, lower-amplitude features, indicative of a more diffuse property manifold. In contrast, aromatic

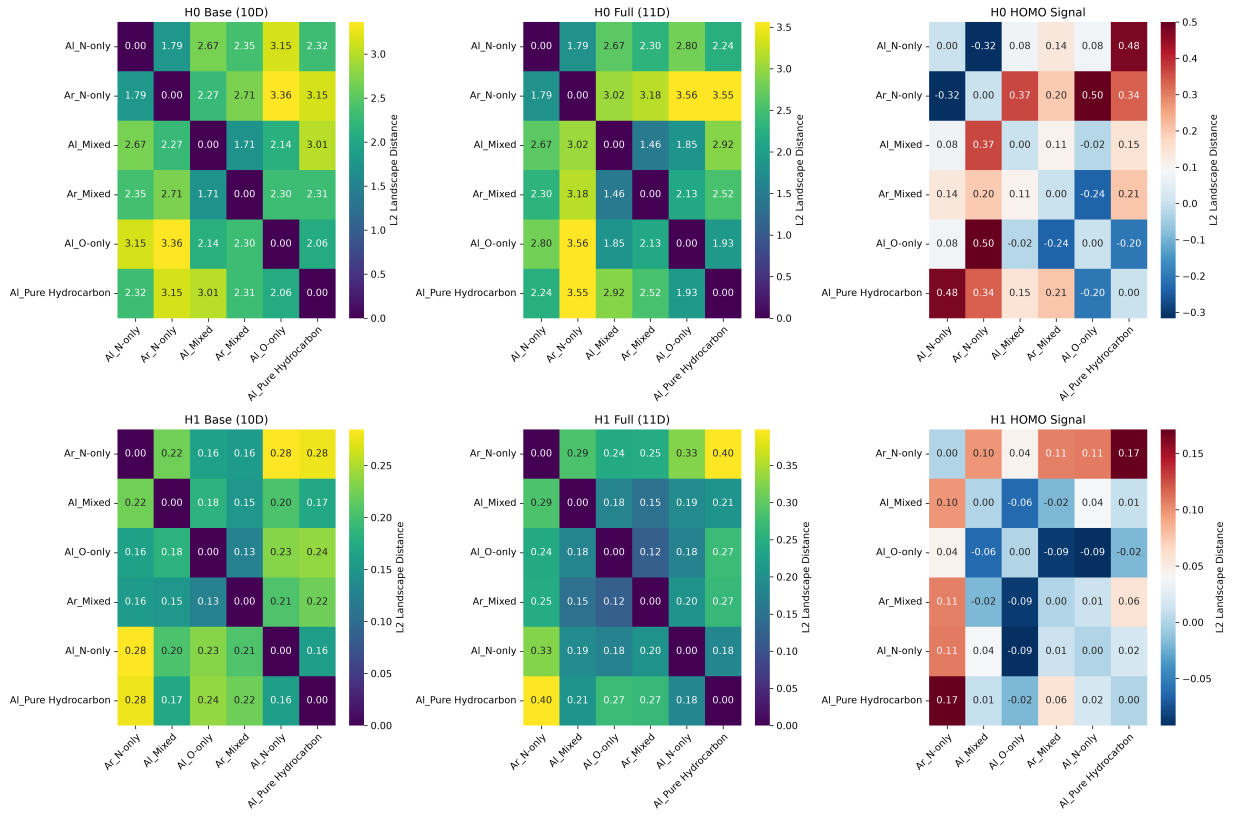

Figure 2: Heatmaps displaying the L_2 landscape distances between six chemical families for the H_0 (top row) and H_1 (bottom row) topological dimensions. The columns compare the 10-dimensional base property manifold (left), the 11-dimensional full manifold including HOMO energy (middle), and the isolated HOMO signal (right). The HOMO signal quantifies the topological rewiring attributable to the specific embedding of the HOMO energy. The right column reveals that this rewiring varies significantly across structural boundaries; the maximum H_0 signal divergence reaches 0.50 between Aliphatic O-only and Aromatic N-only molecules. In the H_1 dimension, which captures complex multi-property dependencies, the highest divergence (0.17) occurs between Aromatic N-only and Aliphatic Pure Hydrocarbons, demonstrating that aromatic nitrogen strongly reintegrates the HOMO energy into the joint property manifold. Conversely, negative shifts in Aliphatic O-only systems (e.g., -0.09 against Aliphatic N-only) indicate that their HOMO contribution is more decoupled, resembling a random distribution rather than a structurally rigid dependency.

N-only systems display a tightly constrained, clumped organization of property residuals, reflecting a highly structured dependency network.

The H_1 landscapes (Figure 4) further elucidate these structural differences. Aromatic N-only molecules possess distinct 1-cycles that persist at significantly larger distance scales than those observed in aliphatic systems. In the context of the residual property space, these persistent cycles represent high-dimensional voids—specific combinations of property values that are physically unreachable for that structural class. The prominence of these voids in aromatic N-only systems underscores a rigid topological structure where properties are tightly interdependent.

3.4 Physical organic interpretation of HOMO reintegration

Collectively, these topological signatures demonstrate that the decoupling of the HOMO energy is not a universal constant of the QM9 dataset, but rather a variable property dictated by structural topology. In aliphatic hydrocarbons and oxygen-containing molecules, the HOMO energy acts as a nearly independent variable, contributing minimally to the persistent homology of the joint manifold.

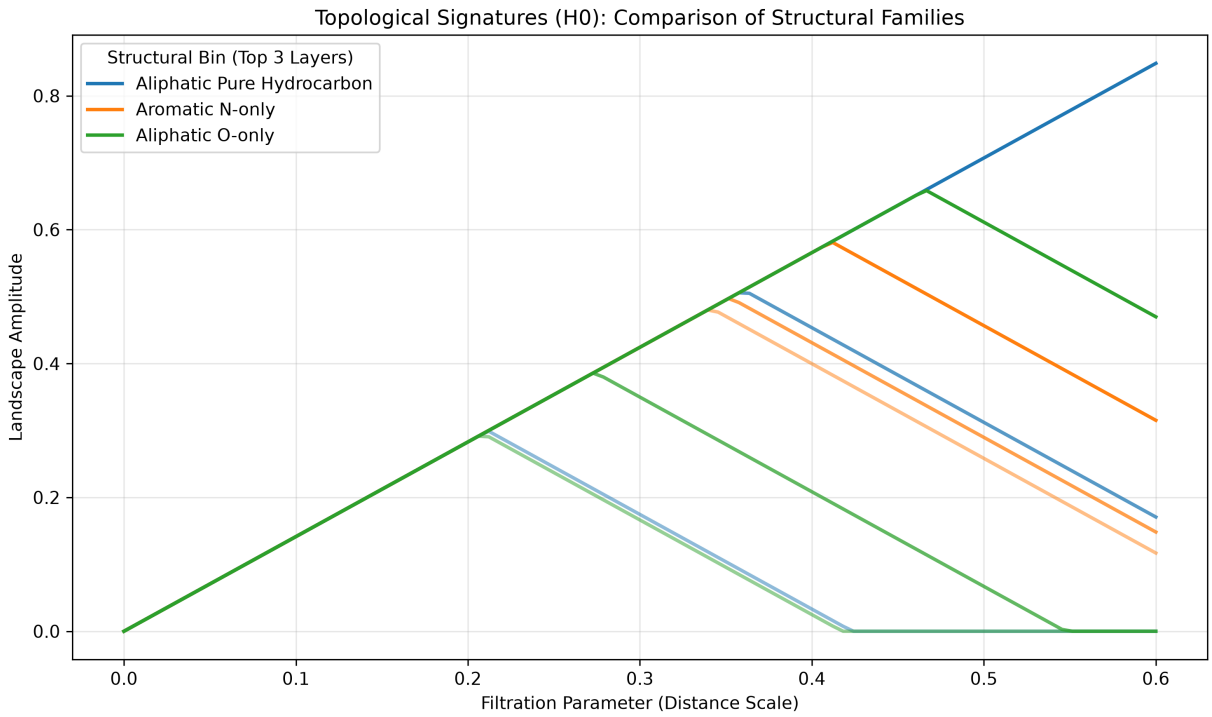

Figure 3: Comparison of H_0 persistence landscapes for three structural families, illustrating the connectedness of the molecular property cloud across filtration parameters. Aliphatic pure hydrocarbons exhibit broader, lower-amplitude features, indicating a diffuse property manifold, whereas aromatic N-only systems display a tightly constrained, clumped organization of property residuals. These topological signatures demonstrate how structural family transitions fundamentally rewires the joint organization of the non-stoichiometric property manifold.

However, in aromatic N-heterocycles, the HOMO energy is topologically reintegrated. This reintegration reflects the physical reality of conjugated nitrogen systems, where the HOMO energy is no longer purely a function of local bonding environments. Instead, it becomes intrinsically linked to the global electronic extent (r^2) and the polarizability (α) of the aromatic ring. The high stability ratio and the large H_1 signal divergence observed for Aromatic N-only molecules confirm that this reintegration is a robust structural feature. The transition across aromatic and heteroatomic boundaries does not merely shift absolute property values; it fundamentally rewires the joint organization of the molecular property manifold.

4 Conclusions

The investigation into the multidimensional property space of the QM9 dataset has resolved a persistent anomaly regarding the highest occupied molecular orbital (*homo*) energy. By employing persistent homology to analyze stoichiometric-adjusted property residuals, we have demonstrated that the apparent orthogonality of the *homo* energy to the global thermodynamic and spatial manifold is not a universal quantum chemical constant. Instead, it is a highly variable characteristic dictated by discrete structural topologies.

Our topological framework, which isolates genuine electronic rewiring from trivial atom-count scaling, reveals that the joint property manifold is exceptionally stable within specific chemical families but undergoes profound reorganization across structural boundaries. Through the differential comparison of persistence landscapes, we quantified the isolated topological impact of the *homo* energy. The results confirm that in aliphatic hydrocarbons and oxygen-containing systems, the *homo* energy behaves as a nearly independent variable, contributing minimally to the persistent homology of the joint manifold and closely resembling a

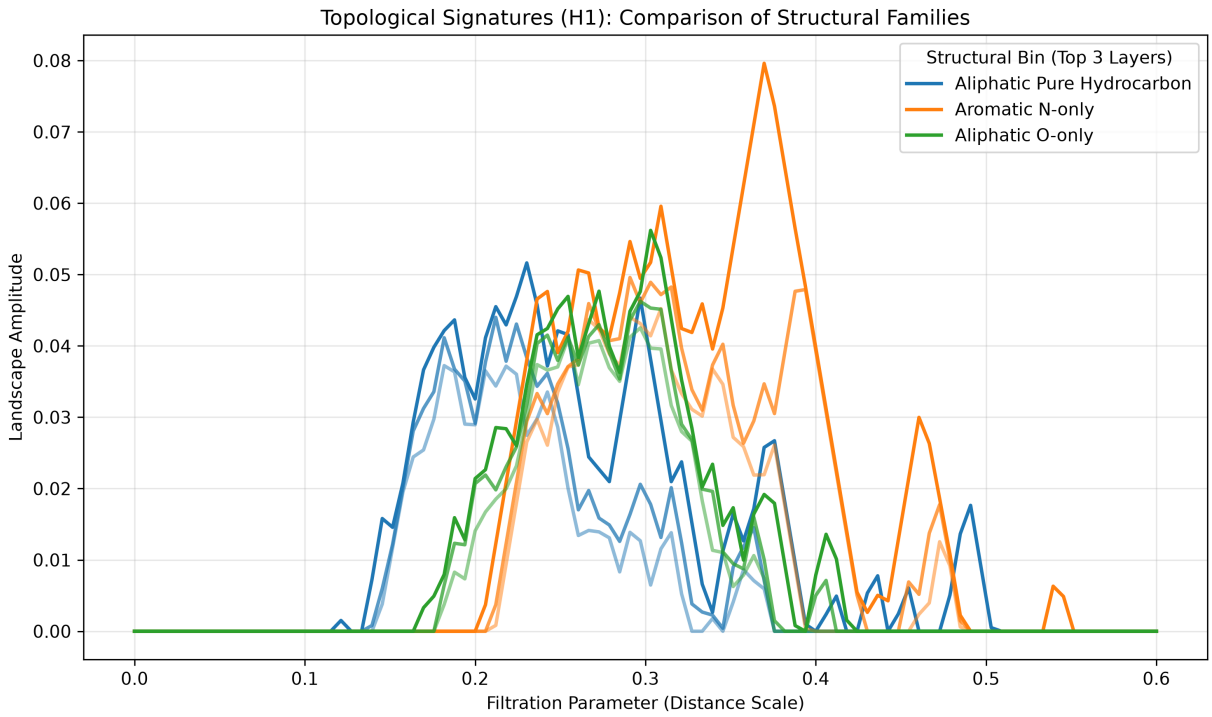

Figure 4: H_1 persistence landscapes comparing the topological signatures of Aliphatic Pure Hydrocarbons, Aromatic N-only, and Aliphatic O-only structural families. The landscape amplitude is plotted against the filtration parameter, representing the distance scale. Aromatic N-only molecules exhibit distinct 1-cycles that persist at significantly larger distance scales than those in aliphatic systems. These persistent cycles indicate high-dimensional voids in the property space, corresponding to specific combinations of property values that are physically unreachable for that structural class. This pronounced topological structure reflects the reintegration of the HOMO energy in conjugated N-systems, where it becomes intrinsically linked to global properties such as electronic extent (r^2) and polarizability (α), contrasting with aliphatic molecules where the HOMO energy acts as a nearly independent variable.

random statistical distribution.

Crucially, this decoupling is broken in aromatic nitrogen heterocycles. The introduction of aromatic nitrogen induces a significant divergence in the first homology group (H_1), indicating the formation of a rigid, highly constrained dependency network. In these conjugated systems, the *homo* energy is topologically reintegrated, reflecting its physical coupling to global molecular properties such as electronic spatial extent (r^2) and isotropic polarizability (α). The persistence of high-dimensional voids in the property space of these molecules underscores that certain combinations of quantum and thermodynamic properties become physically unreachable once this structural boundary is crossed.

Ultimately, this work establishes that fundamental structural transitions—such as the introduction of aromaticity or specific heteroatoms—do not merely shift absolute property values, but fundamentally rewire the underlying dependency graph of the molecular manifold. By moving beyond linear covariance and adopting a coordinate-free topological approach, we provide a more rigorous understanding of how electronic states embed within chemical space. These insights highlight the necessity of accounting for structural topology in the development of predictive machine learning models and the broader exploration of molecular property manifolds.