Quantifying the Topological Decoupling of the Frontier Electronic Manifold via Persistent Homology

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The organization of molecular properties is frequently reduced to pairwise correlations, leaving the apparent independence of frontier electronic states from macroscopic geometric and thermodynamic structures poorly understood. To quantify this structural orthogonality, we employ persistent homology and optimal transport to measure the topological divergence between a base geometric-thermodynamic manifold and the frontier electronic space. Using the QM9 dataset, we isolate pure structural variance from stoichiometric determinism by extracting gradient-boosted atom-count residuals. We then compute the second-order Wasserstein distance between the persistence diagrams of the augmented property manifolds and localized shuffled controls across distinct heteroatom families. Our results evaluate the magnitude of topological decoupling across various chemical environments, analyzing the extent of manifold rewiring in specific heteroatom groups and assessing the influence of extreme linear geometries on the joint property space. Ultimately, this topological framework provides a mathematical basis for investigating how the integration of the frontier electronic manifold relates to molecular composition and structural complexity.

1 Introduction

The organization of molecular properties is a central theme in cheminformatics and materials discovery. Typically, the relationships between diverse molecular features are analyzed through pairwise correlations or latent space embeddings. However, the frontier electronic manifold—comprising properties such as the highest occupied molecular orbital (HOMO) energy and the electronic spatial extent (r^2)—frequently exhibits a stubborn independence from the macroscopic geometric and thermodynamic structure of the molecule. Understanding the extent and nature of this structural orthogonality remains a significant challenge, as traditional linear metrics and assumptions of multivariate normality fail to capture the complex, high-dimensional geometry of the joint property space.

To rigorously investigate why frontier electronic properties remain structurally orthogonal to global molecular organization, it is necessary to isolate pure architectural effects from stoichiometric determinism. Raw molecular properties are heavily confounded by simple atom counts, which dictate a large variance in baseline thermodynamic and geometric features such as zero-point vibrational energy (ZPVE) or heat capacity (C_v). By extracting non-linear gradient-boosted residuals based on exact atom counts, we can strip away this stoichiometric baseline and observe the pure structural variance—how the specific spatial arrangement of atoms dictates property divergence.

In this work, we employ persistent homology and optimal transport to mathematically quantify the topological decoupling of the frontier electronic manifold. Persistent homology provides a coordinate-free framework to track the intrinsic shape, connected components, and multidimensional holes of point cloud data across varying spatial scales. Using the QM9 dataset, we construct a base geometric-thermodynamic manifold (M_1) using the residuals of ZPVE, C_v, and rotational constants (B, C). We then augment this base manifold with electronic residuals (HOMO, r^2) and compare the resulting topological structures against localized shuffled controls. This localized permutation preserves one-dimensional marginal distributions and local density while destroying joint structural correlations, providing a rigorous null baseline to test whether electronic properties simply "sink" into existing structural noise or introduce fundamentally new topological features.

Family	HOMO Decoupling (W_2)	r^2 Decoupling (W_2)
Pure Hydrocarbons	0.0352 [0.0321, 0.0384]	0.0451 [0.0409, 0.0494]
Nitrogen-containing	0.0356 [0.0332, 0.0380]	0.0412 [0.0368, 0.0456]
Oxygen-containing	0.0321 [0.0296, 0.0347]	0.0369 [0.0320, 0.0419]
Mixed Heteroatoms	0.0296 [0.0272, 0.0321]	0.0354 [0.0311, 0.0396]
Fluorine-containing	0.1167 [0.1089, 0.1245]	0.1555 [0.1483, 0.1627]

By computing the second-order Wasserstein distance (W_2) between the persistence diagrams of the augmented manifolds and their controls, we evaluate the magnitude of topological decoupling across distinct chemical environments. Our analysis reveals that the independence of the frontier electronic manifold is not a universal constant of organic chemistry, but is instead highly conditioned on the specific heteroatom environment. As summarized in Table 1, the magnitude of decoupling varies significantly across structural families.

Notably, fluorine-containing molecules exhibit a fundamental rewiring of the joint property space, with decoupling distances approximately three to four times larger than those of pure hydrocarbons or oxygen-containing compounds. In these highly electronegative environments, the electronic spatial extent (r^2) becomes even more decoupled than the HOMO energy, breaking down conventional assumptions that group r^2 alongside geometric properties. Conversely, molecules with mixed heteroatoms display the lowest decoupling, suggesting that increased structural complexity and diverse substitution force the electronic and geometric manifolds into a more constrained joint organization.

Furthermore, we assess the influence of extreme linear geometries on the manifold boundaries. Our sensitivity analysis demonstrates that these linear molecules act as topological anchors; their removal paradoxically increases the intrinsic decoupling of the remaining non-linear population. Ultimately, this topological framework provides a robust mathematical basis for investigating the complex integration of electronic states with macroscopic molecular architecture, moving beyond pairwise statistics to capture the true shape of chemical property space.

2 Methods

2.1 Data curation and stoichiometric baseline

To isolate pure structural variance from stoichiometric determinism, we first established a baseline using the QM9 dataset. The dataset was deduplicated based on SMILES strings, yielding 133,802 unique molecules. For each molecule, exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted. The dataset was partitioned into five distinct chemical families based on heteroatom presence: pure hydrocarbons, nitrogen-containing, oxygen-containing, fluorine-containing, and mixed heteroatoms. Additionally, nine linear molecules (identified by a rotational constant $A = 0$) were isolated to serve as topological anchors for downstream sensitivity analyses.

We defined two target property manifolds: the geometric-thermodynamic base manifold (M_1), comprising zero-point vibrational energy (ZPVE), heat capacity (C_v), and rotational constants (B, C); and the frontier electronic manifold (M_2), comprising the HOMO energy, HOMO-LUMO gap (replacing LUMO to prevent rank-deficiency), and electronic spatial extent (r^2). To remove the confounding effects of simple atom counts, a gradient boosting regressor was trained for each of the seven properties using only the five atom counts as input features. The residuals from these regressions were extracted and globally scaled using a robust scaler, preserving the relative volume and spread of the manifolds across the different chemical families.

2.2 Localized control generation

To rigorously test whether the electronic properties introduce fundamentally new topological features or merely sink into existing structural noise, we generated localized shuffled controls. For the M_1 residuals, we computed the first principal component (PC1) and binned the dataset into quantiles based on this score. The residuals of the electronic properties were then shuffled strictly within these local PC1 bins. This localized permutation preserves both the one-dimensional marginal distributions and the local density structure of the

base manifold while destroying any specific joint structural correlation, providing a robust null hypothesis for manifold decoupling.

2.3 Point cloud construction and persistent homology

High-dimensional point clouds were constructed for each heteroatom partition. The base space was defined by the 4D M_1 residuals, which was then augmented to create 5D point clouds by appending either the true electronic residuals (M_1 + electronic) or their localized controls (M_1 + control). To ensure computational tractability and statistical robustness, we performed bootstrap sampling, drawing 10 independent random subsamples of exactly 2,000 molecules for each partition. A parallel set of bootstrap samples explicitly excluding the nine linear molecules was generated to assess their anchor effect on the manifold boundaries.

Persistent homology was computed using the `gudhi` library. We constructed a sparse Vietoris-Rips complex directly on the point clouds. To maintain robustness against outliers, the maximum edge length (filtration value) was dynamically set based on the median pairwise distance of the M_1 point cloud for each partition. We computed the persistent homology up to dimension 1, capturing connected components (H_0) and one-dimensional loops (H_1).

2.4 Topological metrics and Wasserstein distance

To quantify the topological shifts, we extracted the Betti numbers (β_0, β_1) from the persistence diagrams. Topological noise was filtered by retaining only H_1 features with a lifetime exceeding 10% of the maximum persistence within their respective diagrams. We calculated the normalized shifts in Betti counts ($\Delta\beta_0, \Delta\beta_1$) between the base M_1 manifold and the augmented spaces, alongside the persistence entropy, to summarize changes in manifold texture.

The magnitude of topological decoupling was formally quantified using the sliced Wasserstein distance (W_2) between the normalized persistence diagrams of the base and augmented spaces. Normalization of the diagrams was applied prior to distance computation to prevent dimensionality penalties from being conflated with topological contributions. Finally, the topological contribution of each electronic property was defined as the difference in the normalized Wasserstein distance shift relative to the shift caused by its matched localized control. These distances were aggregated across the bootstrap iterations to compute means and 95% confidence intervals.

3 Results

3.1 Quantitative measures of manifold decoupling

To evaluate the structural orthogonality of the frontier electronic space, we quantified the decoupling of electronic properties using the second-order Wasserstein distance (W_2). This distance was computed between the persistence diagrams of the real augmented manifolds and their corresponding localized shuffled controls. A non-zero W_2 distance indicates that the electronic property exhibits a joint structural organization with the base geometric-thermodynamic manifold (M_1) that cannot be explained by its marginal distribution or local density alone.

The magnitude of this decoupling varies significantly across heteroatom classes, particularly in the H_1 dimension, which captures the circular connectivity and one-dimensional loops of the point clouds. The mean W_2 decoupling distances for the HOMO energy and the electronic spatial extent (r^2) are summarized in Table 1.

As demonstrated in Figure 1, the most striking feature is the anomalous behavior of fluorine-containing molecules. The decoupling distance for both HOMO and r^2 in the fluorine class is approximately three to four times larger than in any other chemical family. This suggests that the presence of highly electronegative fluorine atoms creates a fundamental rewiring of the joint property space, where the electronic properties deviate most sharply from the geometric-thermodynamic backbone.

Beyond the absolute decoupling distance, we evaluated the topological contribution, which measures the extent to which the addition of an electronic property shifts the manifold's shape relative to the control. For most families, the mean topological contributions are near zero (e.g., -0.0034 for the nitrogen-containing

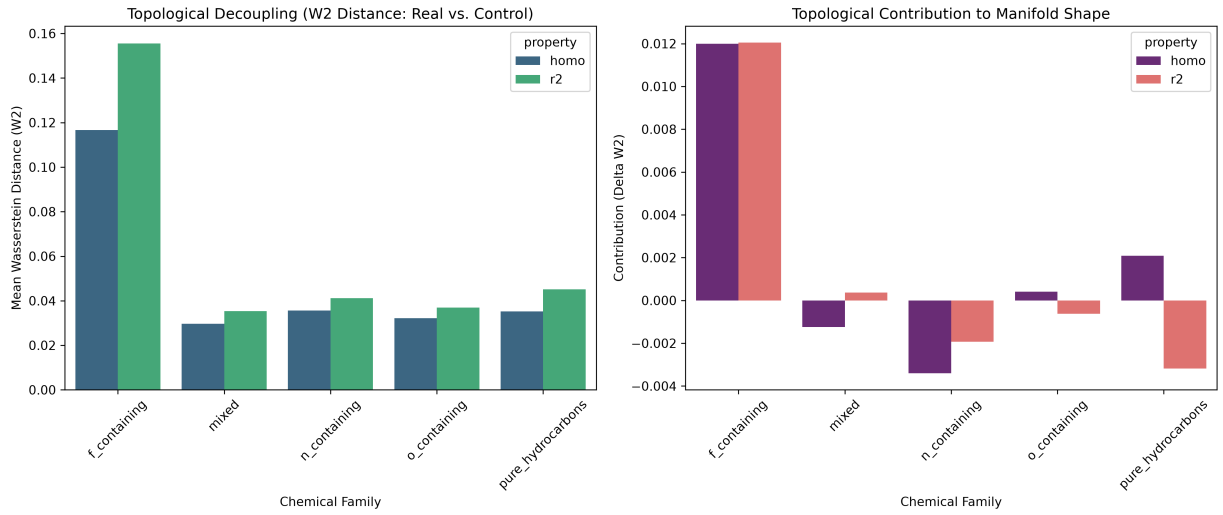

Figure 1: Topological decoupling and contribution of electronic properties (HOMO and r^2) across distinct chemical families. The left panel displays the mean 2nd-order Wasserstein distance (W_2) between real augmented manifolds and localized shuffled controls, quantifying the independence of electronic properties from the base geometric-thermodynamic manifold (M_1). The right panel shows the topological contribution, measuring the shift in manifold shape relative to the control. The plots reveal that fluorine-containing molecules exhibit anomalously high decoupling, with W_2 distances approximately 3 to 4 times larger than other families, alongside unique positive topological contributions. This demonstrates that highly electronegative fluorine substituents cause a fundamental rewiring of the joint property space, introducing topological features absent in the base structure. Conversely, the mixed heteroatom family displays the lowest decoupling, indicating that diverse heteroatom substitution forces the property manifolds into a more constrained joint organization.

HOMO manifold), indicating that the electronic manifold is largely sunk into the existing structural noise. However, fluorine-containing molecules again stand out, exhibiting positive topological contributions for both HOMO (0.0120) and r^2 (0.0120). This implies that fluorine introduces unique topological features—such as new loops or voids—that do not exist in the base geometric space.

3.2 Sensitivity to extreme molecular geometries

A critical question in our topological analysis is whether the observed manifold features are disproportionately driven by extreme geometries, such as the nine linear molecules present in the QM9 dataset. To address this, we performed a sensitivity analysis by comparing the full family populations against subsets that explicitly excluded these linear structures.

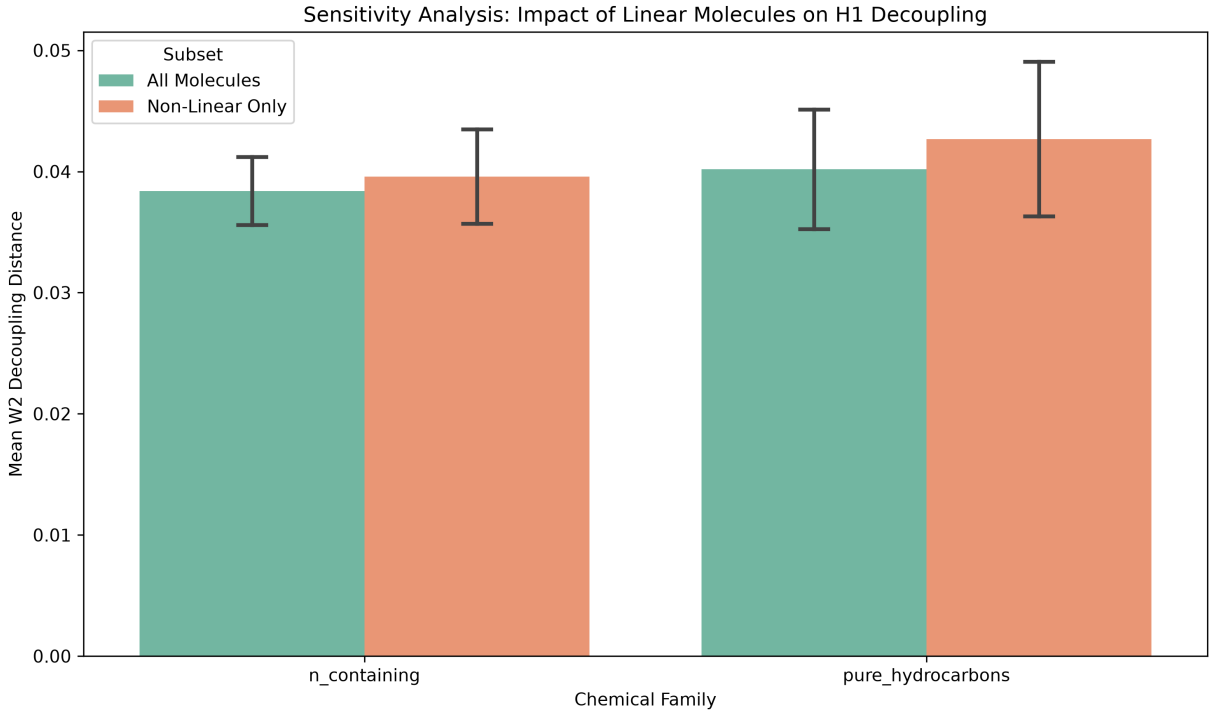

Figure 2: Sensitivity analysis of the mean W_2 decoupling distance in the H_1 dimension for nitrogen-containing (`n_containing`) and pure hydrocarbon (`pure_hydrocarbons`) families, comparing the full molecular population to a subset excluding linear molecules. The plot demonstrates that removing linear geometries increases the decoupling distance. This reveals that extreme linear geometries act as topological anchors that pull the electronic and geometric manifolds closer together, and their removal makes the intrinsic decoupling of the remaining non-linear population more pronounced.

As shown in Figure 2, the removal of linear molecules generally increases the measured decoupling distance. For the H_1 features of the r^2 manifold, the mean decoupling distance shifts from 0.0412 to 0.0435 (a relative increase of 5.56%) in nitrogen-containing molecules, and from 0.0451 to 0.0491 (an increase of 8.73%) in pure hydrocarbons. The HOMO decoupling distances exhibit remarkable stability, with minor shifts of +0.24% and +3.01% for the nitrogen-containing and pure hydrocarbon families, respectively. The increased decoupling observed upon the removal of linear molecules indicates that these extreme geometries act as topological anchors. By pulling the electronic and geometric manifolds closer together, they mask the true extent of structural orthogonality. When these anchors are removed, the intrinsic decoupling of the remaining non-linear population becomes more pronounced.

3.3 Structural reorganization and the property block hypothesis

The persistent homology results reveal distinct patterns of structural reorganization across the chemical families.

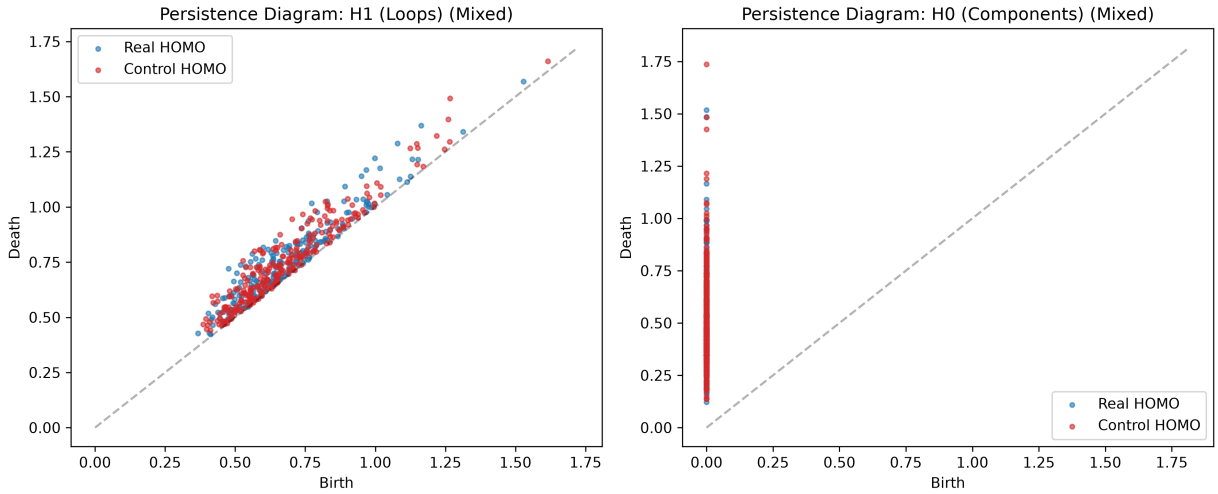

Figure 3: Persistence diagrams comparing the real augmented manifold (M_1 + HOMO) and localized shuffled controls for the mixed heteroatom family. The right panel demonstrates that H_0 connected components are highly similar between the real and control sets, whereas the left panel reveals that H_1 loops in the real electronic manifold exhibit distinct birth-death patterns. Although these topological differences exist, the mixed heteroatom family demonstrates the lowest overall decoupling distance ($W_2 = 0.0296$), revealing that increased molecular complexity through diverse heteroatom substitution forces the electronic and geometric-thermodynamic property manifolds into a more constrained joint organization.

As illustrated by the persistence diagrams in Figure 3, the H_0 connected components remain highly similar between the real augmented manifolds and the shuffled control sets. However, the H_1 features in the real electronic manifold frequently exhibit shorter lifetimes or altered birth-death patterns compared to their controls.

The significant decoupling observed in fluorine-containing molecules highlights a breakdown in the conventional "property block" hypothesis. While the electronic spatial extent (r^2) is often grouped with geometric properties in standard cheminformatics analyses, its W_2 distance of 0.1555 in the fluorine family is the highest measured value across all partitions. This indicates that r^2 is even more decoupled from the base geometric-thermodynamic manifold than the HOMO energy ($W_2 = 0.1167$) in that specific environment.

In contrast, the mixed heteroatom family demonstrates the lowest overall decoupling (HOMO W_2 = 0.0296). This suggests that as molecular complexity increases through diverse heteroatom substitution, the various property manifolds are forced into a more constrained, albeit complex, joint organization. Ultimately, these results demonstrate that the independence of the frontier electronic manifold is not a universal constant of organic chemistry, but rather a variable state heavily conditioned on the specific heteroatom environment, reaching its maximum divergence in the presence of fluorine substituents.

4 Conclusions

In this study, we introduced a topological framework based on persistent homology and optimal transport to quantify the structural orthogonality of the frontier electronic manifold. By isolating pure structural variance from stoichiometric determinism and comparing augmented property manifolds against localized shuffled controls, we moved beyond traditional pairwise correlations to capture the intrinsic, high-dimensional geometry of chemical property space.

Our findings demonstrate that the independence of frontier electronic states from macroscopic geometric and thermodynamic structures is not a universal constant of organic chemistry. Instead, it is heavily conditioned on the specific heteroatom environment. The most profound divergence occurs in fluorine-containing molecules, which exhibit a fundamental rewiring of the joint property space. In these highly electronegative environments, the electronic properties introduce unique topological features that do not exist in the base geometric-thermodynamic manifold, resulting in decoupling distances substantially larger than those observed in pure hydrocarbons or oxygen-containing compounds. Conversely, molecules with mixed heteroatoms display the lowest topological decoupling, indicating that increased structural complexity and diverse substitution force the various property manifolds into a more constrained joint organization.

This environment-dependent decoupling challenges conventional cheminformatics assumptions, particularly the property block hypothesis. The electronic spatial extent (r^2), frequently grouped with geometric properties, was shown to be even more decoupled than the HOMO energy in fluorine-rich environments. This highlights the necessity of evaluating molecular properties within their specific chemical contexts rather than assuming uniform structural integration across all regions of chemical space.

Furthermore, our sensitivity analysis revealed the critical role of extreme molecular geometries in shaping the boundaries of property manifolds. Linear molecules act as topological anchors, artificially pulling the electronic and geometric manifolds closer together at the extremes of rotational space. Their removal unmasks a more pronounced intrinsic decoupling within the non-linear population, emphasizing the importance of boundary conditions in the topological analysis of molecular datasets.

Ultimately, quantifying the topological decoupling of molecular properties provides a rigorous mathematical foundation for understanding how electronic states integrate with macroscopic molecular architecture. By revealing the exact topological thresholds where the stubbornly independent frontier electronic space is forced to coalesce with spatial structure, this approach offers new perspectives for molecular representation learning and the targeted design of functional materials.