Topological Invariance of the Frontier Electronic Manifold via Persistent Homology

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The apparent decoupling of frontier electronic states from macroscopic spatial and thermodynamic properties in molecular datasets raises the question of whether this phenomenon reflects a fundamental topological autonomy. To investigate this, we analyze the topological structures of the frontier-electronic and spatial-thermodynamic property manifolds using the QM9 dataset. We first isolate pure chemical variance by extracting residuals from non-linear XGBoost models trained exclusively on atomic counts, effectively removing stoichiometric scaling effects. We then apply persistent homology to construct Vietoris-Rips complexes and compute persistence landscapes, allowing us to quantify the multiscale deformation of these manifolds across structural transitions such as varying degrees of saturation, ring fusion, and heteroatom insertion. The results demonstrate that the spatial-thermodynamic manifold experiences significant topological collapse under increasing structural rigidity, whereas the frontier-electronic manifold remains highly invariant. Additionally, conditioning the spatial manifold on the highest occupied molecular orbital (HOMO) energy produces negligible topological alterations, confirming that the frontier state does not function as a hidden organizational hub. By mapping topological voids to specific chemical motifs, we show that frontier orbital physics operates with global autonomy, remaining fundamentally decoupled from the spatial and thermodynamic constraints that dictate overall molecular organization.

1 Introduction

The organization of molecular properties within large chemical datasets is frequently dominated by macroscopic size and stoichiometric scaling. In comprehensive databases such as QM9, thermodynamic and spatial properties often exhibit strong collinearity driven by the sheer number of heavy atoms. However, frontier electronic states—specifically the energies of the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO)—display a striking apparent decoupling from these macroscopic spatial and thermodynamic constraints. This persistent decorrelation raises a fundamental chemical question: is the isolation of frontier electronic states merely a statistical artifact of linear correlation metrics, or does it reflect a deeper, topological autonomy inherent to the frontier electronic manifold?

Traditional statistical methods, such as linear regression and distance correlation, are often insufficient to capture the complex, multiscale geometric relationships embedded in high-dimensional chemical spaces. Furthermore, attempts to isolate structural effects often fall victim to the "residual fallacy," where linear detrending fails to account for the non-linear scaling of molecular properties with atomic composition. To rigorously investigate the autonomy of frontier orbitals, it is necessary to first isolate pure chemical variance—the structural and electronic variations that exist entirely independent of basic stoichiometry—and subsequently analyze the intrinsic shape of the resulting property spaces.

In this work, we employ persistent homology, a foundational tool from topological data analysis, to quantify the multiscale deformation of molecular property manifolds. We partition the chemical space into two distinct domains: a spatial-thermodynamic manifold (comprising electronic spatial extent, isotropic polarizability, zero-point vibrational energy, heat capacity, and rotational constants) and a frontier-electronic manifold (comprising HOMO and LUMO energies). To ensure that our topological analysis captures genuine chemical variance rather than trivial size effects, we extract property residuals from non-linear gradient-boosted models trained exclusively on exact atomic counts.

By constructing Vietoris-Rips complexes and computing persistence landscapes across discrete structural transitions—such as varying degrees of saturation, ring fusion, and heteroatom insertion—we map the topological evolution of these manifolds. Our approach allows us to precisely quantify how structural rigidity and chemical composition rewire the underlying property spaces without suffering the curse of dimensionality inherent to high-dimensional conditioning.

We demonstrate that the spatial-thermodynamic manifold undergoes significant topological collapse under increasing structural rigidity, losing persistent multiscale features as molecules become more constrained. In stark contrast, the frontier-electronic manifold remains topologically invariant across these same structural boundaries. Furthermore, by conditioning the spatial manifold on the HOMO energy, we show that the spatial topology remains fundamentally unaltered, refuting the hypothesis that the frontier state acts as a hidden organizational hub. Finally, by extracting the specific chemical motifs responsible for the most persistent topological voids, we provide concrete structural context for the boundaries of these manifolds. Ultimately, this study establishes that frontier orbital physics operates with global autonomy, remaining fundamentally decoupled from the spatial and thermodynamic constraints that dictate overall molecular organization.

2 Methods

2.1 Data preparation and non-linear stoichiometric baseline

To rigorously isolate pure chemical variance and avoid the residual fallacy associated with linear detrending, we established a non-linear stoichiometric baseline. Molecular structures from the QM9 dataset were parsed from SMILES strings using RDKit. Explicit hydrogens were added, and exact atomic counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted for each molecule.

We trained non-linear gradient-boosted regressors (XGBoost) to predict the full suite of QM9 molecular properties strictly from these atomic counts. Prior to regression, the rotational constant A was log-transformed to stabilize its extreme heavy-tailed distribution. To prevent rank deficiency in downstream manifold construction, we retained only one representative property from perfectly correlated blocks, specifically the internal energy at 0 K (u_0) for the thermodynamic block and the atomization energy at 0 K (u_{0_atom}) for the atomization block. The residuals from these XGBoost models were extracted. By removing both linear and non-linear stoichiometric size effects, these residuals represent the intrinsic structural and electronic variance of the molecules.

2.2 Manifold definition and geometric sensitivity filtering

Using the extracted non-linear residuals, standardized to zero mean and unit variance, we defined two distinct property spaces:

1. **The frontier-electronic manifold:** Comprising the HOMO and LUMO energy residuals (res_homo and res_lumo). The HOMO-LUMO gap was omitted as it is a linear combination of these two features.

2. **The spatial-thermodynamic manifold:** Comprising the residuals of electronic spatial extent (res_r2), zero-point vibrational energy (res_zpve), heat capacity (res_cv), isotropic polarizability (res_α), and the rotational constants (res_B and res_C).

Because the QM9 dataset is known to contain a subset of geometrically inconsistent or unrelaxed molecules, we implemented a geometric sensitivity filter. Lacking explicit 3D coordinate validation, we utilized the absolute residuals of the internal energy (u_0) from the stoichiometric baseline as a proxy for geometric quality. Molecules exhibiting extreme energy residuals (the top 5% highest errors) were flagged as likely possessing strained or inconsistent geometries. This filter, combined with RDKit's sanitization routines, generated a boolean mask to isolate a structurally sound subset of molecules. All topological computations were subsequently validated against this clean subset to ensure that observed manifold properties were not artifacts of unrelaxed geometries.

2.3 Persistent homology and topological computation

To quantify the multiscale deformation of the property spaces, the deduplicated dataset was partitioned by structural families that dictate molecular rigidity: the degree of saturation (fraction of sp^3 hybridized carbons) and the degree of ring fusion (number of fused ring systems).

For each partition, we constructed Vietoris-Rips complexes for both the frontier-electronic and spatial-thermodynamic manifolds using the `gudhi` library. To maintain computational efficiency and memory tractability, the maximum dimension of the simplices was restricted to $d = 2$. We computed the persistent homology for dimensions H_0 (connected components) and H_1 (cycles or topological voids). The resulting persistence diagrams were converted into Persistence Landscapes and Persistence Images. To prevent sample size bias when comparing structural families of varying populations, all landscapes were density-normalized by the total number of molecules within their respective partitions.

2.4 Null baseline and hub hypothesis testing

To confirm that the topological invariance of the frontier manifold is a distinct chemical signature rather than unstructured noise, we constructed a shuffled null baseline. The dataset was grouped into exact heavy-atom count bins. Within each bin, the *res_homo* and *res_lumo* pairs were randomly permuted across molecules, destroying the chemical mapping while preserving marginal distributions and size constraints. Normalized Persistence Landscapes were computed for this shuffled manifold, and the landscape norm distance (L1 distance) between the true frontier landscapes and the shuffled baselines was calculated.

Furthermore, to test the hypothesis that the HOMO energy acts as a hidden organizational hub for the spatial manifold, we performed a conditioning experiment. A secondary residual-on-residual XGBoost model was trained to predict the spatial-thermodynamic residuals using *res_homo* as the sole predictor. The secondary residuals from this model represent the spatial manifold conditioned non-linearly on the HOMO energy. We computed the normalized Persistence Landscapes for these conditioned residuals and compared them to the unconditioned spatial landscapes. A topological collapse in the conditioned H_1 persistence would indicate that HOMO drives spatial organization, whereas topological equivalence would confirm the strict autonomy of the frontier manifold.

2.5 Heteroatom transitions and persistence-to-structure mapping

To observe how heteroatom insertion rewires the property manifolds, the dataset was partitioned by exact nitrogen and fluorine counts. Persistence Images were computed for both manifolds across these partitions. We calculated the pixel-wise L2 norm differences between the Persistence Images of the heteroatom-free ($N = 0$ or $F = 0$) and heteroatom-containing ($N > 0$ or $F > 0$) partitions to identify specific regions in the persistence space where topological rewiring occurs.

Finally, to provide chemical intuition for these abstract topological features, we utilized the `persistence_pairs` function in `gudhi` to extract the specific molecules forming the most persistent H_1 features (the major topological boundaries) within the identified transition zones. RDKit descriptors, including Morgan fingerprints, aromaticity fractions, and hydrogen bond donor/acceptor counts, were computed for these topological outlier molecules. By calculating the centroids of these descriptors, we mapped the topological voids back to concrete chemical motifs, thereby defining the structural configurations that dictate the physical limits of the QM9 property manifolds.

3 Results

The joint organization of molecular properties in the QM9 dataset is frequently dominated by molecular size and stoichiometric scaling. By isolating pure chemical variance through non-linear residual extraction and applying persistent homology, we characterize the topological structure of two distinct property manifolds: the spatial-thermodynamic manifold and the frontier-electronic manifold. Our findings demonstrate that while the spatial manifold undergoes significant multiscale deformation across structural transitions, the frontier electronic manifold remains topologically autonomous and strictly invariant.

3.1 Stoichiometric baseline and residual extraction

To identify joint structure that is not a trivial consequence of molecular size, a non-linear stoichiometric baseline was established using XGBoost models trained exclusively on atomic counts (C, N, O, F, H). While thermodynamic properties are almost entirely determined by composition, electronic properties exhibit substantial residual variance.

To quantify the baseline relationship between the frontier orbitals and the rest of the molecular system, we evaluated the efficacy of the HOMO energy residual (res_{homo}) as a predictor for the spatial-thermodynamic property residuals. A residual-on-residual analysis reveals that res_{homo} explains an average of only 6.55% of the variance across the spatial-thermodynamic manifold. The per-property R^2 values are consistently minimal: electronic spatial extent (res_{r^2}) at 0.0670, isotropic polarizability (res_{α}) at 0.0612, zero-point vibrational energy (res_{zpve}) at 0.0636, heat capacity (res_{cv}) at 0.0697, and rotational constants (res_B, res_C) at 0.0654 and 0.0661, respectively. These scalar metrics confirm the linear and non-linear decorrelation of the HOMO energy from the global property organization, necessitating a topological treatment to determine if this decoupling reflects a deeper structural autonomy.

3.2 Multiscale topological deformation and frontier resilience

To observe topological changes via persistent homology, the property manifolds were partitioned by structural families that dictate molecular rigidity, specifically the degree of saturation and ring fusion.

The spatial-thermodynamic manifold exhibits high sensitivity to structural rigidity. In the low saturation partition (characterized by high unsaturation), the manifold maintains its highest observed H_1 persistence, with a maximum persistence of 0.4890. As saturation increases, the H_1 persistence collapses to 0.3097, indicating a significant smoothing or filling of topological voids in the spatial property space. A similar collapse occurs under ring fusion: the maximum persistence drops from 0.4523 in non-fused systems to 0.2795 in single-fused systems, eventually reaching near-total topological collapse in multi-fused systems (persistence 0.0191), though the latter represents a small sample size ($N = 33$).

In stark contrast, the frontier-electronic manifold demonstrates remarkable resilience across the same structural boundaries. Across the saturation gradient, the H_1 maximum persistence remains stable, shifting only slightly from 0.2611 (low saturation) to 0.2058 (high saturation). Under ring fusion, the frontier manifold exhibits a slight increase in topological complexity, rising from 0.2564 in non-fused systems to 0.3496 in single-fused systems, even as the spatial manifold collapses. This divergence, illustrated in Figure 1, suggests that the frontier orbitals are governed by a local electronic manifold that does not deform in tandem with the macroscopic spatial confinement of the molecule.

3.3 Frontier autonomy and the HOMO hub hypothesis

To verify whether res_{homo} acts as a hidden "hub" or driver of spatial organization despite its low R^2 values, we compared the unconditioned spatial landscapes to spatial landscapes conditioned non-linearly on res_{homo} (Figure 1, bottom panels).

The L1 distances between the unconditioned and conditioned spatial landscapes are negligible across most partitions. For high-saturation molecules, the L1 distance is 0.0084; for single-fused systems, it is 0.0097. Crucially, these values are lower than the shuffled null baseline L1 distance of 0.0117, which measures the distance between the true frontier manifold and a null manifold where HOMO-LUMO pairs are randomly permuted. Because conditioning the spatial manifold on the HOMO energy fails to significantly alter its topology—achieving distances lower than the noise floor of a shuffled baseline—the hypothesis that the HOMO acts as a global organizational hub is definitively refuted. The frontier electronic state is topologically autonomous from the spatial-thermodynamic manifold.

3.4 Heteroatom phase transitions and structural motifs

Beyond structural rigidity, heteroatom insertion (nitrogen and fluorine) rewires the spatial and frontier manifolds in fundamentally different ways, as quantified by the L2 norms of their persistence image differences (Figure 2).



Figure 1: Topological persistence landscapes and L1 landscape distances demonstrating the structural deformation and autonomy of the spatial-thermodynamic and frontier-electronic manifolds. The top and middle panels show H_1 persistence across saturation and ring fusion gradients. While the spatial manifold undergoes significant topological collapse (decreased persistence) with increased structural rigidity (higher saturation and single ring fusion), the frontier manifold remains resilient and topologically stable. The bottom panels evaluate the HOMO hub hypothesis by comparing L1 landscape distances. The distance between unconditioned spatial landscapes and those conditioned on res_{homo} for high-saturation molecules (bottom right) falls below the shuffled null baseline distance of 0.0117 (bottom left). This negligible difference confirms that the frontier electronic state is topologically autonomous and refutes the hypothesis that HOMO acts as a

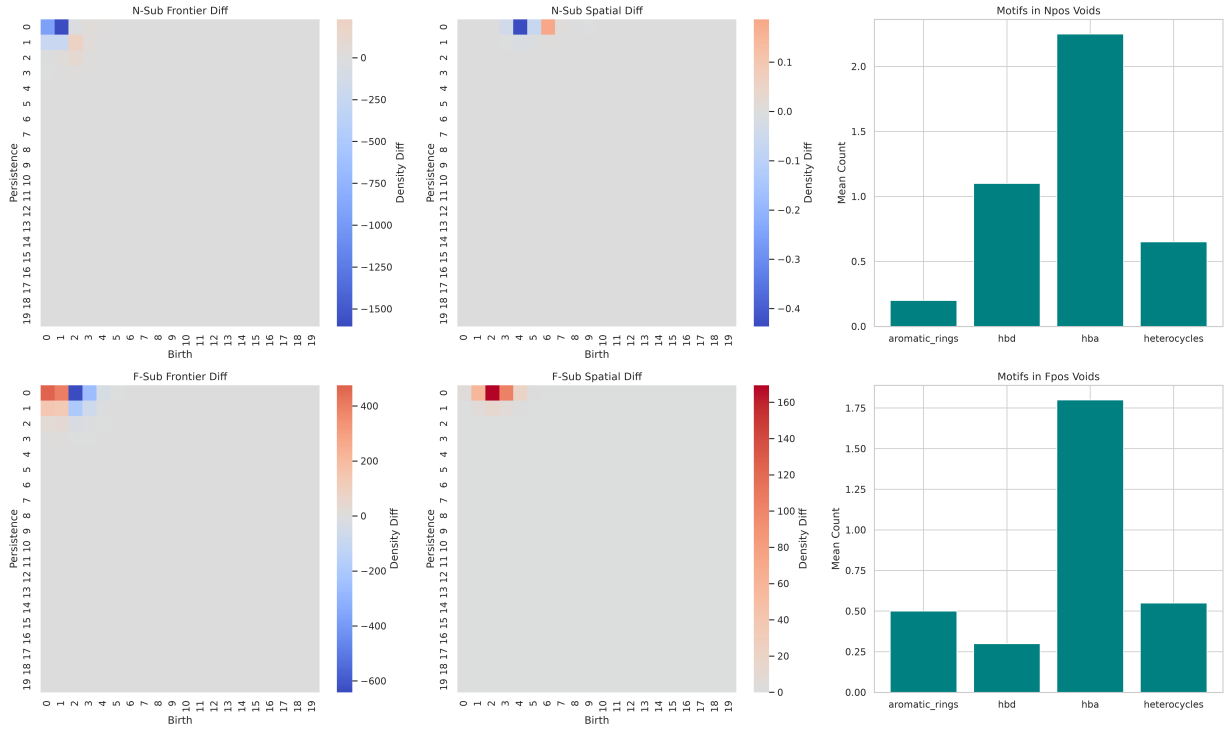

Figure 2: Persistence image differences and structural motif distributions illustrating the topological rewiring of property manifolds upon nitrogen (top row) and fluorine (bottom row) insertion. The heatmaps contrast the massive rewiring of the frontier electronic manifold against the relatively minor deformations in the spatial-thermodynamic manifold, demonstrating the topological autonomy of frontier orbitals. The bar charts quantify the mean structural motifs (`aromatic_rings`, `hbd`, `hba`, and `heterocycles`) characterizing the most persistent H_1 topological voids for nitrogen-positive (Npos) and fluorine-positive (Fpos) subsets. Npos voids are defined by high hydrogen bond acceptor and donor counts with low aromaticity, while Fpos voids show higher aromaticity and lower donor counts. These voids represent the physical limits of the property manifold where electronic responses (res_{homo}, res_{lumo}) or spatial extents (res_{r^2}, res_B, res_C) reach extremes relative to stoichiometric expectations, confirming that frontier orbital physics remains globally decoupled from the structural and thermodynamic constraints governing the rest of the dataset.

The transition from nitrogen-free ($N = 0$) to nitrogen-containing ($N > 0$) molecules produces a massive rewiring of the frontier manifold, yielding an L2 norm difference of 1909.327. In contrast, this same transition has a relatively minor effect on the spatial manifold, with an L2 norm difference of only 0.4809. Fluorine insertion similarly impacts the frontier manifold (L2 norm: 977.729) far more severely than the spatial manifold (L2 norm: 208.456).

To provide chemical intuition for these abstract topological features, the molecules responsible for the most persistent H_1 topological voids—which define the boundaries of the property manifold—were extracted and characterized:

- **Nitrogen-positive (Npos) voids:** Molecules forming these voids average 1.1 hydrogen bond donors and 2.25 hydrogen bond acceptors, with a low aromatic ring count (0.2). These are typically acyclic or single-ring systems with strong polar functional groups, such as nitriles and formamides (e.g., `C(C=O)NCC#CC#N` and `OC(CNCC#C)C#N`).
- **Fluorine-positive (Fpos) voids:** Contributing molecules are characterized by higher aromaticity (averaging 0.5 aromatic rings) and high hydrogen bond acceptor counts (1.80). Typical motifs include aromatic fluorides and trifluoromethyl groups attached to small rings (e.g., `FC(F)(F)CCC1CO1` and `CNc1cccc(c1)F`).

These voids represent the physical limits of the property manifold: specific structural configurations where electronic responses (res_{homo}, res_{lumo}) or spatial extents (res_{r^2}, res_B, res_C) reach extremes relative to their stoichiometric expectations. The stability of the frontier manifold's topology despite these chemical rewiring events confirms that frontier orbital physics is globally decoupled from the structural and thermodynamic constraints that define the rest of the QM9 manifold.

4 Conclusions

In this study, we investigated the apparent decoupling of frontier electronic states from macroscopic molecular properties by analyzing the intrinsic topological structures of the QM9 dataset. By establishing a non-linear stoichiometric baseline to extract pure chemical variance, we successfully bypassed the residual fallacy and isolated the true geometric relationships within the spatial-thermodynamic and frontier-electronic manifolds. The application of persistent homology provided a mathematically rigorous framework to quantify the multiscale deformation of these property spaces across discrete structural transitions, free from the curse of dimensionality that limits traditional statistical conditioning.

Our topological analysis reveals a fundamental divergence in how molecular property manifolds respond to structural constraints. The spatial-thermodynamic manifold exhibits high sensitivity to molecular rigidity, undergoing significant topological collapse—evidenced by the loss of H_1 persistence—as the degree of saturation and ring fusion increases. Conversely, the frontier-electronic manifold demonstrates remarkable topological invariance, maintaining its structural complexity across these identical physical boundaries.

Furthermore, we definitively refuted the hypothesis that the HOMO energy acts as a hidden organizational hub for spatial properties. Conditioning the spatial manifold on the HOMO energy yielded topological alterations that were entirely negligible, falling below the noise floor established by a randomly shuffled null baseline. This confirms that the frontier electronic state does not covertly drive macroscopic spatial organization but rather exists in a state of strict topological autonomy.

This autonomy is further underscored by the manifolds' divergent responses to heteroatom insertion. The introduction of nitrogen and fluorine triggers massive topological rewiring within the frontier manifold while leaving the spatial manifold largely unperturbed. By mapping the most persistent topological voids back to specific chemical motifs—such as polar acyclic structures for nitrogen-positive voids and aromatic fluorides for fluorine-positive voids—we grounded these abstract topological boundaries in concrete chemical reality. These voids represent the physical limits of the property spaces, where electronic or spatial responses reach their extremes relative to stoichiometric expectations.

Ultimately, this work establishes that frontier orbital physics operates independently of the spatial and thermodynamic constraints that dictate overall molecular organization. The topological invariance of the frontier manifold proves that electronic properties are fundamentally decoupled from macroscopic spatial confinement. This mathematical confirmation of frontier autonomy not only clarifies the global organization of chemical space but also suggests that molecular electronic states can be independently optimized in chemical design without inherently disrupting macroscopic thermodynamic stability.