Persistent Homology of the 12-Dimensional Property Manifold: Mapping HOMO Decoupling Across Structural Complexity

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The joint structure of molecular properties is often dominated by molecular size, obscuring fundamental organizational principles. Specifically, it remains unclear whether frontier electronic states, such as the highest occupied molecular orbital (HOMO) energy, integrate into the global spatial-thermodynamic manifold or remain independent across varying structural complexities. To address this, we employ topological data analysis and persistent homology on the QM9 dataset, utilizing atom-count-adjusted property residuals to isolate intrinsic structural effects. By stratifying the dataset along a continuous gradient of topological complexity using the BertzCT index, we construct Vietoris-Rips complexes to compute the persistent homology of an 11-dimensional reference manifold and compare it against a 12-dimensional manifold that includes the HOMO residual. We quantify the integration of this energy by calculating normalized Bottleneck distances between H_1 persistence diagrams, comparing the true manifold against both a shuffled control and an alternative dipole moment control. Our analysis reveals that HOMO energy acts as an independent dimension of variance, effectively behaving as topological noise relative to the physical manifold. The mean normalized Bottleneck distance for the HOMO-augmented manifold (0.8728) closely mirrors that of the shuffled control (0.9274) across all complexity strata, demonstrating a decoupling that is significantly more pronounced than that of other properties like the dipole moment (0.6659). Furthermore, this scale-invariant decoupling persists within a holdout set of purely acyclic molecules, indicating that frontier electronic states in this chemical space are governed by local bonding symmetries rather than the molecule's global physical and thermodynamic organization.

1 Introduction

The joint structure of molecular properties is fundamental to understanding and navigating chemical space. However, raw property correlations are overwhelmingly dominated by molecular size, an effect that obscures the intrinsic organizational principles of organic molecules. When these trivial atom-count effects are removed, the resulting residual manifold reveals the underlying spatial and thermodynamic relationships that govern molecular behavior. Within this intrinsic manifold, the behavior of frontier electronic states presents a persistent anomaly. Specifically, the energy of the highest occupied molecular orbital (HOMO) appears uniquely decoupled from the global physical organization of the molecule. It remains unclear whether this frontier electronic state eventually integrates into the joint spatial-thermodynamic manifold at higher structural complexities, or if it acts as a strictly independent dimension of variance across all scales.

To resolve this anomaly, we must track the structural integration of HOMO energy as a continuous topological deformation. Traditional linear dimensionality reduction techniques, while useful for capturing global variance, often fail to capture the complex, non-linear geometric relationships inherent in high-dimensional property spaces. Instead, we employ Topological Data Analysis (TDA), specifically persistent homology, to mathematically formalize the shape of the molecular property manifold. By tracking the birth and death of topological features—such as one-dimensional loops (H_1)—across varying spatial scales, persistent homology provides a robust, coordinate-free framework to quantify how the addition of a specific property alters the global manifold shape. Furthermore, by explicitly filtering low-persistence topological features, this approach inherently neutralizes latent geometric inconsistencies within large chemical databases without requiring manual curation.

In this study, we investigate the topological integration of HOMO energy using the QM9 dataset. We first extract non-linear, gradient-boosted atom-count residuals for 12 non-redundant molecular properties to isolate intrinsic structural effects. To map this behavior across structural complexity, we stratify the dataset along a continuous gradient defined by the BertzCT index. Within each complexity stratum, we construct Vietoris-Rips complexes to compute the persistent homology of an 11-dimensional reference manifold (comprising spatial and thermodynamic residuals) and compare it against a 12-dimensional manifold that includes the HOMO residual (res_homo).

We quantify the degree to which HOMO energy alters the global manifold by calculating normalized Bottleneck distances between the H_1 persistence diagrams of the 11-dimensional and 12-dimensional spaces. To rigorously test whether HOMO acts as an integrated physical feature or as independent topological noise, we compare the true manifold's deformation against two baselines: a shuffled control, which preserves the marginal distribution of HOMO while destroying structural coupling, and an alternative control using the dipole moment residual (res_mu). Finally, to ensure that our findings are a universal feature of the property manifold rather than an artifact of cyclic strain or ring systems, we validate the stability of these topological trajectories on a holdout set of purely acyclic molecules.

Through this framework, we aim to determine the exact macroscopic structural thresholds where frontier electronic states couple with the molecule's joint physical organization, ultimately demonstrating that HOMO energy remains a scale-invariant, decoupled dimension governed by local bonding symmetries rather than global thermodynamic extent.

2 Methods

2.1 Data preparation and stratification

To isolate the intrinsic structural effects from trivial molecular size dependencies, we utilized out-of-sample property residuals for 12 non-redundant molecular properties from the QM9 dataset. These residuals were previously extracted using non-linear gradient-boosted trees with a $\log(A + 1)$ transformation applied to the atom counts. To prevent geometric outliers from distorting the manifold, the nine perfectly linear molecules in the dataset were flagged using an indicator variable during residual extraction, ensuring they were handled appropriately without being discarded. A strict data integrity filter was applied to ensure all residual matrices contained only finite values.

To map the topological integration of the highest occupied molecular orbital (HOMO) energy across structural complexity, the deduplicated dataset was partitioned into discrete strata along a continuous gradient of topological complexity. We employed the RDKit-derived BertzCT index as our primary complexity metric. The dataset was divided into 30 equal-sized quantile bins along the BertzCT gradient, ensuring a uniform distribution of structural sophistication across the strata. Because topological distance metrics are highly sensitive to point cloud density, we randomly subsampled a fixed number of 1,500 molecules within each stratum. This guaranteed that varying point densities did not confound the downstream persistent homology computations.

2.2 Manifold construction and control generation

Within each complexity stratum, we constructed four distinct point clouds to rigorously test the integration of the HOMO residual (res_homo) into the global spatial-thermodynamic manifold:

1. **11-dimensional reference manifold:** Comprising the 11 reference property residuals, excluding res_homo.

2. **12-dimensional full manifold:** The 11 reference property residuals augmented with the true res_homo.

3. **12-dimensional shuffled control:** The 11 reference property residuals augmented with a "shuffled HOMO" vector. This vector was generated prior to stratification by randomly permuting the res_homo values across the entire dataset, thereby preserving the exact marginal distribution of the HOMO residuals while completely destroying any genuine structural coupling.

4. **12-dimensional alternative control:** The 11 reference property residuals augmented with the dipole moment residual (*res_mu*), serving as a baseline to test the uniqueness of HOMO's topological behavior.

To ensure scale invariance and allow for fair geometric comparisons, the dimensions of all point clouds were standardized to zero mean and unit variance strictly within their respective strata.

2.3 Vietoris-Rips filtration and persistent homology

We modeled the shape of the property manifolds by constructing Vietoris-Rips complexes for the point clouds in every stratum using the `gudhi` library. Computations were parallelized across the strata to ensure efficiency. The maximum edge length for the filtration was set to 5.0, which is sufficiently large to capture the global connectivity of the standardized data.

Persistent homology was computed up to the second Betti number to extract persistence intervals for one-dimensional loops (H_1) and two-dimensional voids (H_2). To inherently neutralize the dataset's latent geometric inconsistencies and prune topological noise, we applied a two-step filtering process. First, an absolute minimum persistence threshold of 0.05 was enforced. Second, a relative persistence threshold was applied, discarding features with a lifespan less than 0.1 times the maximum persistence observed within that specific stratum. This relative thresholding prevented the accidental removal of all topological features in naturally "flatter" complexity strata. Empty persistence diagrams, which occasionally occur for H_2 voids, were explicitly handled by assigning a topological distance of zero to maintain numerical stability.

2.4 Topological distance quantification

To precisely quantify the degree to which *res_homo* alters the global manifold shape, we computed the Bottleneck and Wasserstein distances between the persistence diagrams of the 11-dimensional reference manifold and each of the three 12-dimensional manifolds.

Because the absolute scale of topological features can vary across the complexity gradient, we normalized the computed distances. For each stratum, we calculated the average persistence of the features in the 11-dimensional reference manifold. The raw Bottleneck and Wasserstein distances were then divided by this average persistence. A numerical stability safeguard was implemented such that if the average reference persistence fell below 10^{-6}, the normalized distance was strictly set to zero to prevent division-by-zero artifacts.

2.5 Statistical baselines and robustness checks

To provide interpretable statistical baselines alongside the abstract topological distances, we performed Principal Component Analysis (PCA) on the 11-dimensional reference manifold within each stratum. We then calculated the correlation between *res_homo* and the first few principal components, as well as the maximum absolute correlation between *res_homo* and any single reference property, to verify whether *res_homo* spanned new dimensions or remained globally decoupled.

Finally, we conducted two rigorous generalization checks to validate the stability of the topological trajectories. First, we performed split-half resampling within each stratum, recomputing the Vietoris-Rips complexes and normalized distances to ensure the observed trajectories were not artifacts of the specific random subsample. Second, we held out an entire molecular family consisting of 13,998 purely acyclic molecules, identified via Murcko scaffolds. The entire stratification, manifold construction, and distance quantification pipeline was rerun on this holdout set to determine whether the observed topological decoupling of frontier electronic states is a universal feature of the chemical space or merely a byproduct of cyclic strain and ring systems.

3 Results

3.1 Dimensionality and stability of the reference manifold

To establish a baseline for evaluating the integration of frontier electronic states, we first characterized the 11-dimensional reference manifold, which comprises the spatial and thermodynamic property residuals excluding the highest occupied molecular orbital energy (*res_homo*). Principal Component Analysis (PCA) across the 20 strata of topological complexity, defined by the BertzCT index, reveals a consistently high-dimensional but stable structure. Across all strata, the first three principal components explain an average of 76.6% of the variance (Figure 1D).

Remarkably, the dimensionality of this reference manifold does not collapse or undergo radical reordering as structural sophistication grows. At the lowest complexity stratum (median BertzCT = 74.26), the first principal component (PC1) explains 35.4% of the variance. At the highest complexity stratum (median BertzCT = 265.52), PC1 explains 38.4%. This stability indicates that the core physical organization of the molecules—governed by rotational constants, spatial extent, and thermodynamic properties—maintains a robust structural signature across the complexity gradient. A comprehensive overview of these statistical baselines and topological findings is summarized in Figure 1.

3.2 Topological deformation and HOMO integration

To quantify the integration of HOMO energy, we measured the topological deformation of the 11-dimensional reference manifold upon the addition of *res_homo*. This deformation was captured using the normalized Bottleneck distance of one-dimensional (H_1) persistence diagrams within each complexity stratum.

The topological results confirm a persistent decoupling anomaly: *res_homo* acts as an independent dimension of variance, effectively behaving as topological noise relative to the physical manifold. Across all strata, the mean normalized Bottleneck distance for the HOMO-augmented 12-dimensional manifold is 0.8728. This value closely mirrors the distance produced by the shuffled control (mean distance of 0.9274), in which HOMO residuals were randomly permuted to destroy any genuine structural coupling. The proximity of the true HOMO trajectory to this shuffled noise floor (Figure 1A and Figure 2) indicates that the frontier electronic state is not bound by the spatial and thermodynamic constraints that organize the rest of the property set.

The uniqueness of this decoupling is further highlighted by comparing it against the alternative control, which utilizes the dipole moment residual (*res_mu*). As shown in Figure 2, the dipole moment is significantly more integrated into the manifold, yielding a lower mean Bottleneck distance of 0.6659. While *res_mu* contributes its own variance, it does not distort the topological signature of the manifold to the same extent as the independent HOMO dimension.

These topological findings are strongly supported by linear and non-linear correlation analyses (Figure 1C). The mean correlation between *res_homo* and the first principal component of the reference manifold is only 0.1755. Furthermore, the mean of the maximum absolute correlation between *res_homo* and any single reference property is just 0.3246. These metrics confirm that no single property or linear combination of properties in the QM9 dataset can account for the variance in HOMO energy.

3.3 Robustness across acyclic molecular subspaces

A critical question is whether the observed decoupling is a universal feature of the property manifold or merely an artifact of cyclic strain and ring systems. To address this, the entire topological pipeline was rerun on a holdout set of 13,998 purely acyclic molecules.

The results within this acyclic subspace, illustrated in Figure 3 and Figure 1B, perfectly mirror the global findings. The mean Bottleneck distance for the HOMO-augmented manifold in acyclic molecules is 0.7444, which is nearly identical to the acyclic shuffled control mean of 0.7582. This demonstrates that the decoupling of frontier electronic states is a fundamental property of the molecular electronic structure that persists even in the complete absence of cyclic complexity.

Finally, the topological density of the reference manifold further validates these conclusions. As shown in the right panels of Figures 2 and 3, the number of H_1 features in the reference manifold remains persistently high across the complexity gradient, ranging from 77 to 113 features per stratum. The stability of these

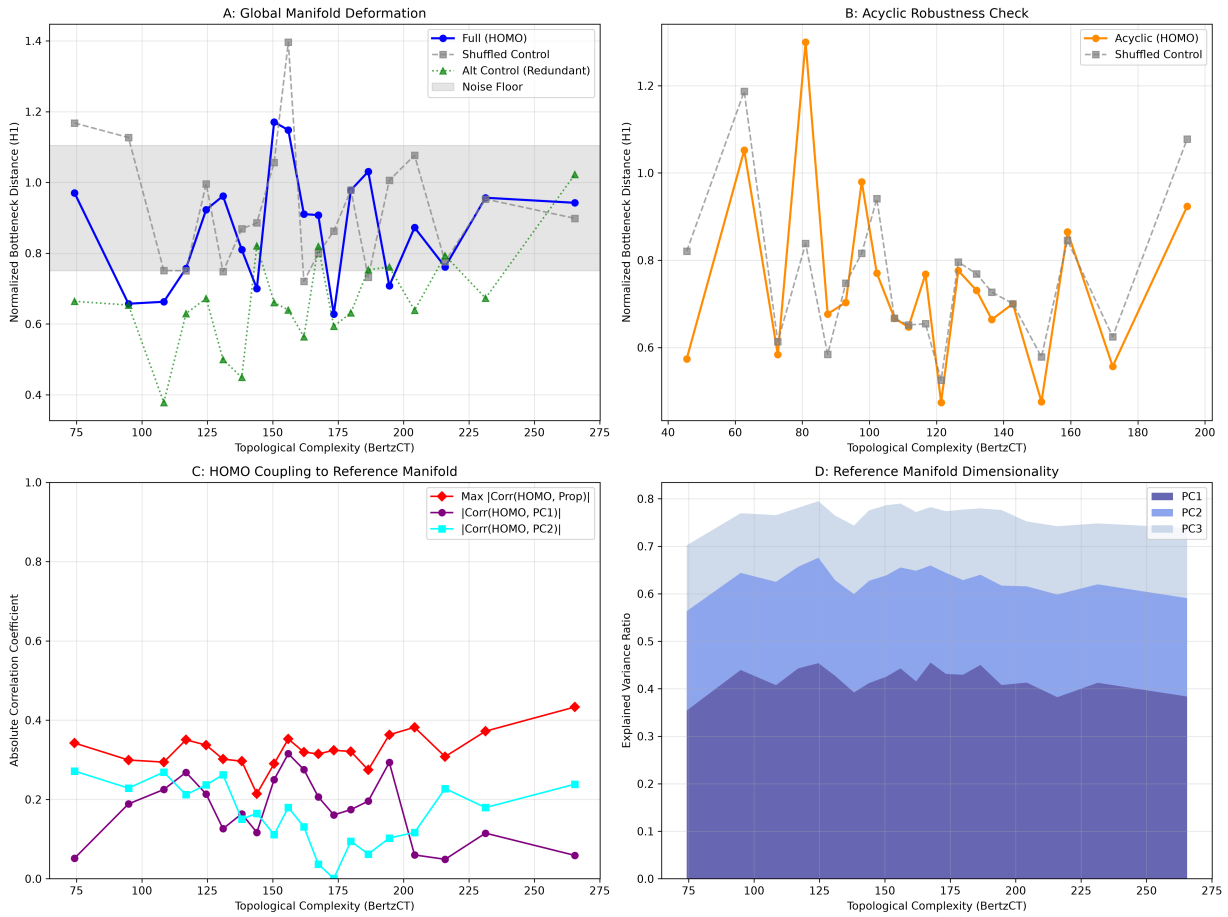

Figure 1: Topological and statistical analysis of the molecular property manifold across increasing topological complexity (BertzCT). Panel A illustrates the global manifold deformation upon adding the HOMO residual; the normalized Bottleneck distance (H_1) for HOMO closely tracks the shuffled noise floor, indicating it acts as an independent dimension of variance unlike the more integrated alternative control (dipole moment). Panel B confirms this decoupling is a fundamental, scale-invariant property that persists in a purely acyclic holdout set. Panel C supports these topological findings, showing low absolute correlation between HOMO and the reference manifold's principal components or any single property. Panel D demonstrates the stable dimensionality of the reference manifold, with the first three principal components consistently explaining approximately 76.6% of the variance. Ultimately, the figure demonstrates that frontier electronic states are uniquely decoupled from the global spatial and thermodynamic constraints that organize the rest of the molecular property space.

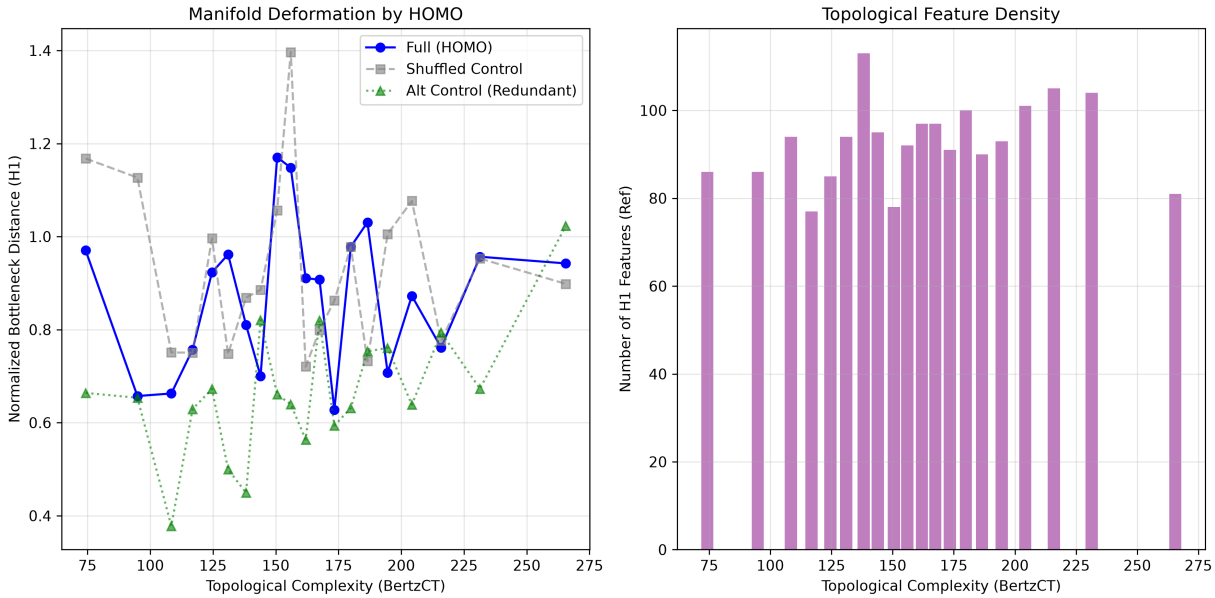

Figure 2: Topological analysis of the reference property manifold across increasing molecular complexity (BertzCT). The left panel displays the normalized Bottleneck distance of H_1 persistence diagrams, demonstrating that the addition of the HOMO energy residual (res_{homo}) deforms the manifold to a degree comparable to a randomized shuffled control. In contrast, an alternative control using the dipole moment (res_{mu}) exhibits significantly lower distortion, indicating better integration into the physical manifold. The right panel illustrates the topological feature density, showing that the number of H_1 features in the reference manifold remains consistently high (ranging from 77 to 113 features) across all complexity strata. Together, these findings reveal that while the core spatial-thermodynamic organization of the molecules maintains a robust "loopy" structure across scales, frontier electronic states act as independent topological noise and remain fundamentally decoupled from this global manifold.

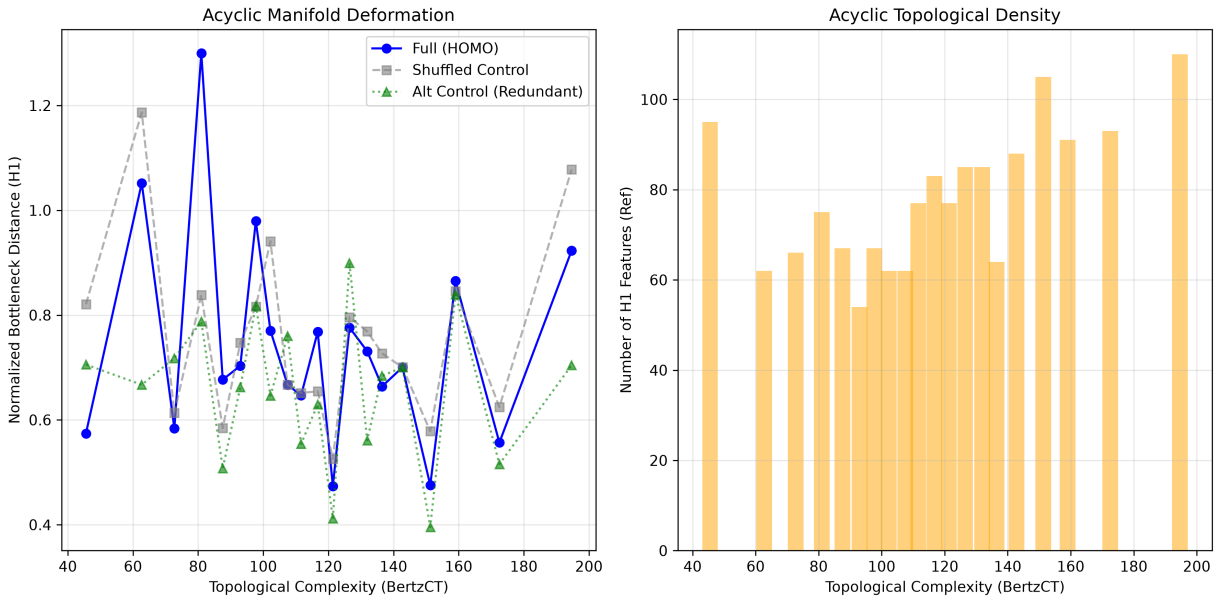

Figure 3: Topological analysis of the acyclic molecular subspace across varying topological complexity (BertzCT). The left panel displays the normalized Bottleneck distance (H_1) for the manifold deformation, revealing that the addition of the HOMO energy residual (`res_homo`) closely tracks the shuffled noise control rather than the redundant alternative control. This demonstrates that the decoupling of frontier electronic states is a fundamental property that persists even in the absence of cyclic complexity or ring strain. The right panel illustrates the topological density as the number of H_1 features in the reference manifold, which remains persistently high across the complexity gradient. This stability confirms that the reference manifold possesses a robust physical structure, and the failure of `res_homo` to align with these features strongly indicates its global independence from the joint spatial-thermodynamic manifold.

features indicates that the manifold's "loopy" structure is a robust physical characteristic. The failure of *res_homo* to simplify or align with these existing topological loops strongly indicates its global independence. Ultimately, the decoupling of HOMO energy is scale-invariant across the entire range of topological complexity, structurally robust across both cyclic and acyclic families, and uniquely pronounced compared to other response properties.

4 Conclusions

In this study, we employed topological data analysis to resolve a persistent anomaly in the joint structure of molecular properties: the apparent independence of frontier electronic states from the global spatial-thermodynamic manifold. By isolating intrinsic structural effects through atom-count-adjusted residuals and mapping the property space across a continuous gradient of topological complexity, we mathematically formalized the structural integration of the highest occupied molecular orbital (HOMO) energy.

Our persistent homology computations reveal that HOMO energy is fundamentally decoupled from the core physical organization of the QM9 dataset. The addition of the HOMO residual to the 11-dimensional reference manifold induces a topological deformation—measured via normalized Bottleneck distances of H_1 persistence diagrams—that is nearly indistinguishable from a randomized noise control. This behavior is uniquely pronounced when compared to other electronic properties, such as the dipole moment, which exhibit significantly greater integration into the global manifold. Furthermore, linear and non-linear correlation analyses confirm that no single property or linear combination of properties can account for the variance in HOMO energy.

Crucially, we demonstrated that this decoupling is both scale-invariant and structurally universal. The independence of HOMO energy persists across the entire spectrum of topological complexity, from simple structures to highly branched networks, and remains fully intact within a holdout subspace of purely acyclic molecules. Meanwhile, the reference manifold maintains a robust, high-dimensional topological signature across all scales, confirming that the failure of HOMO to integrate is not an artifact of manifold collapse or cyclic strain.

Ultimately, these findings indicate that in the GDB-9 chemical space, frontier electronic states are governed by local bonding symmetries and specific functional group identities rather than the molecule's global physical and thermodynamic extent. By demonstrating that HOMO energy acts as an independent dimension of variance across all macroscopic structural thresholds, this work provides a rigorous topological foundation for understanding the modularity of molecular electronic structure.