Localized Structural Drivers and Topological Organization of HOMO Decoupling in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular properties often scale predictably with molecular size, obscuring the underlying structural drivers of specific quantum features such as the highest occupied molecular orbital (HOMO) energy. To isolate genuine structural effects from stoichiometric determinism, we establish a baseline using shallow gradient-boosted regressors to extract property residuals for molecules in the QM9 dataset. We then train XGBoost models on Morgan fingerprints to predict these residuals and apply feature attribution to pinpoint the local bonding environments governing HOMO variance. Additionally, we use topological data analysis via persistent homology to characterize the global structure of the residual space. Our results confirm that HOMO energy residuals are highly decoupled from global molecular geometry and electronegativity, maintaining near-zero partial correlations across most structural families. Feature attribution reveals that this decoupling is driven by highly localized structural motifs—predominantly nitrile groups and specific nitrogen configurations—rather than macroscopic topology. Furthermore, topological data analysis identifies significant voids in the residual space, which represent physically inaccessible regions of chemical property space. Ultimately, these findings demonstrate that while the joint organization of molecular properties is heavily constrained, specific quantum features like HOMO energy remain independently governed by localized structural environments.

1 Introduction

The exploration of chemical space relies on understanding how molecular structure dictates physical and quantum properties. In comprehensive databases such as the QM9 dataset, the joint organization of molecular properties is heavily dominated by stoichiometric determinism. Extensive macroscopic features, such as molecular size and basic atom counts, scale predictably with thermodynamic and spatial properties. While this global covariance is useful for broad predictive modeling, it obscures the underlying structural drivers of specific quantum features, such as the highest occupied molecular orbital (HOMO) energy.

1.1 Stoichiometric determinism and property scaling

The HOMO energy plays a fundamental role in determining chemical reactivity, electron donation capacity, and molecular stability. However, its variance often appears uniquely decoupled from the global spatial-thermodynamic manifold. To rigorously investigate this phenomenon, it is necessary to isolate genuine structural effects from baseline stoichiometric scaling. By extracting property residuals—representing the variance unexplained by simple atom counts and total electronegativity—we can expose a "clean" manifold of chemical properties. This transformation removes the overwhelming influence of molecular size, allowing for a direct examination of whether the HOMO energy is governed by macroscopic molecular topology or by highly localized bonding environments.

1.2 Isolating localized structural drivers

To pinpoint the exact structural motifs responsible for this decoupling, we employ gradient-boosted regressors trained on Morgan fingerprints to predict the HOMO energy residuals (*res_homo*). By extracting native tree-based feature attributions, we bypass global covariance washing and directly map the predictive

dominance of specific local environments. Stratifying these attributions across distinct macroscopic structural families, such as Murcko scaffolds and heteroatom compositions, reveals how the structural drivers of the HOMO residual shift depending on the global molecular context. This approach allows us to test the hypothesis that specific quantum features remain independently governed by localized structural environments, even when the broader molecule undergoes significant topological changes.

1.3 Topological organization of chemical space

Beyond localized feature attribution, understanding the global organization of this decoupled property space requires advanced geometric techniques. Topological data analysis, specifically persistent homology, offers a mathematical framework to characterize the shape of high-dimensional data. By computing the persistence of topological features such as voids and loops in the residual space, we can identify physically inaccessible regions of chemical property space. Comparing these topological signatures against synthetic null models allows us to distinguish genuine physical constraints from mere sampling artifacts inherent to the dataset generation process.

In this work, we systematically dismantle the stoichiometric baseline of the QM9 dataset to investigate the localized structural drivers and topological organization of HOMO decoupling. We demonstrate that HOMO energy residuals maintain near-zero partial correlations with global geometry and electronegativity across most structural families. Feature attribution reveals that this independence is driven by highly localized structural motifs—predominantly nitrile groups and specific nitrogen configurations. Furthermore, topological data analysis identifies significant voids in the residual space, confirming that while the joint organization of molecular properties is heavily constrained, specific quantum features remain independently governed by localized structural environments.

2 Methods

2.1 Data preprocessing and stoichiometric baseline

To isolate the structural drivers of molecular properties from stoichiometric determinism, we first established a robust baseline using the QM9 dataset. The dataset was loaded and deduplicated based on canonical SMILES strings, removing 83 duplicate structures to prevent bias in downstream topological and graph-based analyses. Molecular graphs were parsed using RDKit, with explicit hydrogen atoms added to accurately reflect the full molecular geometry.

To prevent rank-deficient covariance matrices in subsequent multivariate analyses, we removed highly correlated thermodynamic restatements (specifically `u298`, `h298`, `g298`, `u298_atom`, `h298_atom`, and `g298_atom`) as well as the HOMO-LUMO `gap`, which is a linear combination of two other features. The properties `u0`, `u0_atom`, and `zpve` were retained. Additionally, a $\log(1 + x)$ transformation was applied to the rotational constant A to mitigate its extreme heavy tail and to handle the exact zeros present in linear molecules.

We then extracted the exact atom counts for C, N, O, F, and H, alongside the sum of Pauling electronegativities for each molecule. A shallow LightGBM regressor, constrained to a maximum depth of 3 to prevent the absorption of complex structural variance, was trained to predict the 12 remaining physical and quantum properties strictly from these basic stoichiometric and electronic counts. The residuals from these predictions represent the "clean" property manifold. To identify geometrically inconsistent molecules—where the computational methodology (B3LYP/6-31G(2df,p)) may have struggled—we computed a normalized reconstruction error for each molecule:

$$E = \frac{1}{N_{\text{heavy}}} \sum_{i=1}^{12} \left(\frac{r_i - \mu_i}{\sigma_i} \right)^2 \tag{1}$$

where N_{heavy} is the number of heavy atoms, r_i is the residual for property i, and μ_i and σ_i are the mean and standard deviation of that residual across the dataset. Molecules exhibiting the highest normalized reconstruction errors were flagged and separated from the clean subset.

2.2 Structural enrichment and partial correlation networks

To determine if the geometrically inconsistent subset represented a systematic failure for specific structural motifs, we performed a structural enrichment analysis. We defined a set of SMARTS patterns for highly strained geometries (e.g., cyclopropanes, aziridines, oxiranes), cumulenes, and fused polycyclics, and explicitly identified linear molecules where the rotational constant A was zero. Hypergeometric tests were used to quantify the overrepresentation or underrepresentation of these motifs in the inconsistent subset relative to the clean data.

For the clean subset, molecules were grouped into macroscopic structural families based on heteroatom composition (e.g., pure hydrocarbons, strictly N-containing) and Murcko scaffold topology (acyclic, monocyclic, fused polycyclic). Families with fewer than 500 samples were excluded to ensure statistical stability. Within each family, we computed the partial correlation matrix of the 12 property residuals using Spearman rank correlations. A Ledoit-Wolf shrinkage estimator was applied to the covariance matrix to ensure well-conditioned partial correlations. The decoupling of the HOMO energy residual (res_homo) was evaluated by examining its partial correlations with the other properties and with the electronegativity-weighted atom counts. Shifts in the global joint organization of properties across different structural families were quantified by calculating the Frobenius norm of the differences between their respective partial correlation matrices.

2.3 Gradient-boosted feature attribution

To pinpoint the specific local bonding environments governing the variance of the decoupled HOMO energy, we trained XGBoost regressors to predict res_homo directly from RDKit Morgan fingerprints. Morgan fingerprints were chosen to capture the exact connectivity, hybridization, and heteroatom positioning that simple stoichiometric counts miss. As a control, an identical XGBoost model was trained to predict the residual of isotropic polarizability (res_alpha), a property known to be highly coupled to molecular geometry.

Following model training, we extracted native tree-based feature attributions (Gini importance) to map the predictive dominance of specific structural motifs. By stratifying these attributions across the previously defined Murcko scaffold and heteroatom families, we tracked how the structural drivers of the HOMO residual shift across macroscopic boundaries. This approach allowed us to bypass global covariance washing and directly identify whether localized motifs (such as specific nitrogen configurations or nitriles) dictate HOMO variance independently of the global molecular context.

2.4 Topological data analysis of the residual space

To characterize the global structure and boundaries of the stable chemical manifold, we applied persistent homology to the 12-dimensional residual space. For computational tractability, a representative subsample of 5,000 molecules from the clean dataset was used. We constructed a Vietoris-Rips complex and computed the persistence barcodes and Betti numbers (H_1, H_2) to identify topological features such as loops and voids.

To distinguish genuine physical constraints from sampling artifacts, the topological signatures of the real data were compared against two synthetic null models: a multivariate Gaussian dataset matching the mean and covariance of the residuals, and a shuffled null model where property values were permuted across existing structures to break structural correlations. Topological voids present in the real data but absent in the null models were subjected to a local density validation. We calculated the average distance to the k-nearest neighbors for the molecules bounding these voids. If the local density of a void region fell below the 5th percentile of the global density distribution, it was flagged as a potential sampling artifact of the GDB-9 enumeration rules; otherwise, it was confirmed as a genuine forbidden region in chemical property space.

3 Results

3.1 Stoichiometric baseline and geometrically inconsistent structures

The joint organization of molecular properties in the QM9 dataset is heavily dominated by molecular size and basic atomic composition. As outlined in our methodology, we first established a stoichiometric baseline using

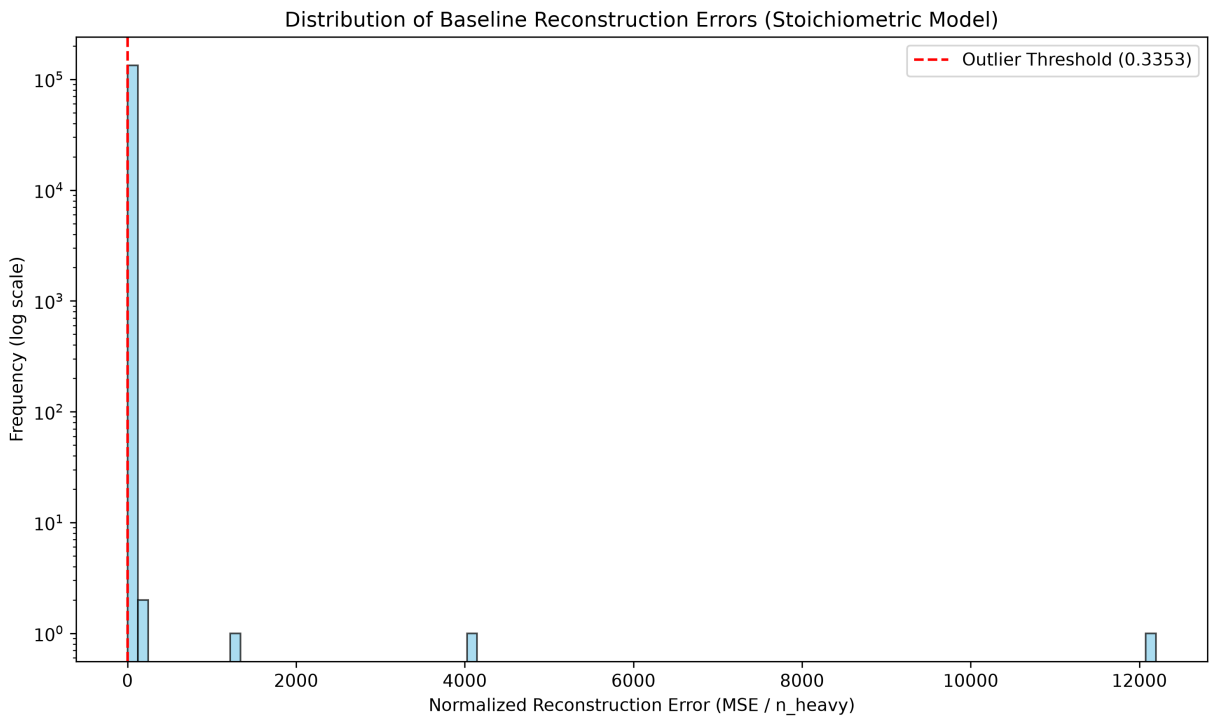

Figure 1: Distribution of normalized reconstruction errors (MSE / n_{heavy}) from the stoichiometric baseline model, displayed on a logarithmic scale. The dense concentration of errors near zero confirms that thermodynamic and vibrational properties are predominantly determined by simple atom counts. A red dashed line marks the outlier threshold at 0.3353, separating the bulk of the data from 3,054 geometrically inconsistent molecules (2.28% of the dataset). The extreme outliers extending far along the x-axis are significantly enriched for linear geometries, such as acetylene derivatives, demonstrating that while the underlying B3LYP/6-31G(2df,p) methodology is robust for most structures including small strained rings, it struggles with the coordinate systems or electronic descriptions of near-linear molecules.

a shallow LightGBM regressor to predict 12 non-redundant physical and quantum properties strictly from atom counts and total electronegativity. The residuals from these predictions represent a "clean" manifold of chemical properties, stripped of trivial size-scaling effects.

By analyzing the normalized reconstruction errors of these residuals, we identified a subset of 3,054 molecules (2.28% of the dataset) that exhibited significant geometric inconsistency, suggesting that the underlying B3LYP/6-31G(2df,p) computational methodology struggled with their specific coordinate systems or electronic descriptions. The distribution of these reconstruction errors is shown in Figure 1.

To understand the structural basis of these computational failures, we performed a structural enrichment analysis comparing the inconsistent subset to the clean data. The results demonstrate that extreme outliers are not driven by typical ring strain. In fact, common strained motifs such as cyclopropane (enrichment factor 0.38×) and aziridine (0.46×) are significantly underrepresented in the outlier set. Instead, the inconsistent subset is overwhelmingly enriched for linear geometries, such as acetylene derivatives, which appear 43.81 times more frequently than in the clean manifold ($p = 1.66 \times 10^{-15}$). This stark contrast is illustrated in Figure 2.

3.2 Decoupling of the HOMO energy from the global manifold

Having isolated the clean subset of molecules, we investigated the joint organization of the property residuals. While properties such as isotropic polarizability (res_α) and electronic spatial extent (res_{r^2}) remain tightly bound to the molecular moments of inertia (res_B, res_C), the HOMO energy residual (res_{homo}) exhibits a

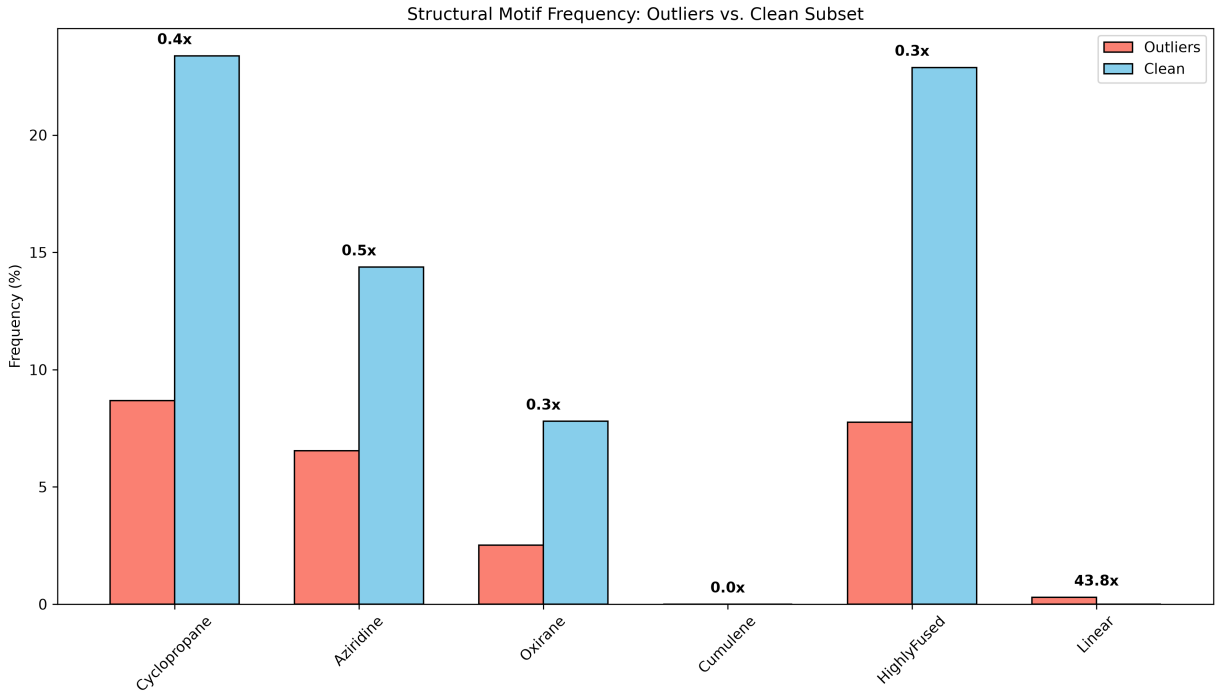

Figure 2: Comparison of structural motif frequencies (%) between the geometrically inconsistent outlier set and the clean subset of the QM9 dataset. The enrichment factors above the bars highlight that linear geometries are significantly overrepresented in the outlier set (43.8×), whereas common strained motifs like cyclopropane (0.4×) and aziridine (0.5×) are underrepresented. This distribution suggests that the underlying computational methodology is robust for small strained rings but struggles with the coordinate systems or electronic descriptions of near-linear structures.

remarkable independence from the global spatial-thermodynamic manifold.

Crucially, this decoupling is not merely a failure to account for basic electronic properties; the correlation between res_{homo} and the total Pauling electronegativity of the molecule is only 0.0339. As shown in Figure 3, partial correlation analysis across macroscopic structural families confirms that res_{homo} maintains near-zero partial correlations with other properties. This isolation is most pronounced in cyclic families, with monocyclic and polycyclic structures yielding partial correlations of 0.016 and 0.018, respectively. Acyclic molecules exhibit a slightly higher, yet still modest, coupling of 0.335.

Furthermore, the broader partial correlation network of the remaining properties shifts significantly depending on heteroatom identity. The Frobenius norm of the difference between partial correlation matrices is substantial when comparing pure hydrocarbons to fluorine-containing molecules (2.704) and nitrogen-containing molecules (2.253). This indicates that while heteroatom identity fundamentally reconfigures the joint organization of most physical properties, the HOMO energy remains uniquely isolated across these macroscopic topological boundaries.

3.3 Localized structural drivers of HOMO variance

To pinpoint the exact structural motifs responsible for the variance of the decoupled HOMO energy, we trained XGBoost regressors on Morgan fingerprints to predict res_{homo} and, as a control, res_{α}. The models successfully captured the variance for both properties, achieving comparable predictive performance with $R^2 = 0.793$ for res_{homo} and $R^2 = 0.784$ for res_{α} (Figure 4).

Despite the similar predictive accuracy, feature attribution analysis reveals that the underlying structural drivers are distinct. Localization metrics indicate that the attributions for res_{α} (Gini = 0.9245) are slightly more concentrated than those for res_{homo} (Gini = 0.9211). More importantly, the specific motifs driving the predictions differ fundamentally. As detailed in Figure 5, the top structural driver for res_{homo} across the entire dataset is the nitrile group (C≡N) environment (Bit ID 790).

Stratifying these attributions by structural family confirms that the predictive dominance of this motif is heavily weighted in nitrogen-containing (importance: 0.0059) and monocyclic families (importance: 0.0034). In contrast, hydrocarbon and oxygen-containing families rely more heavily on general carbon hybridization features (Bit ID 1873). These results confirm our hypothesis that the HOMO energy is governed by highly localized bonding motifs—principally nitriles and specific nitrogen configurations—rather than global molecular geometry or size.

3.4 Topological organization and forbidden regions

Beyond localized feature attribution, we applied Topological Data Analysis (TDA) to characterize the global structure of the 12-dimensional residual space. Using persistent homology on a representative subsample of 5,000 molecules, we found that the chemical property space is not a simple, dense Gaussian cloud, but rather contains genuine structural voids.

As shown in Figure 6, the real data exhibits a higher maximum H_1 persistence (0.6057) compared to both a multivariate Gaussian null model (0.4628) and a shuffled null model (0.4741). The presence of these H_1 topological loops, which have lifetimes significantly exceeding those in the null models, indicates the existence of "forbidden regions" in the chemical property space. These regions represent combinations of property residuals that are physically or energetically inaccessible under the GDB-9 enumeration rules.

To ensure these voids were not merely sampling artifacts, we performed a local density validation. The 5th percentile of the k-nearest neighbors density for the real data (0.4119) is comparable to, and slightly higher than, that of the Gaussian null model (0.4056). This confirms that the boundaries of these voids are well-sampled. Furthermore, the computed Betti numbers ($H_0 = 218$, $H_1 = 295$, $H_2 = 150$) reveal a high H_1 count relative to H_2, suggesting a filamentary or "holey" structure characteristic of a constrained physical system. Ultimately, these topological signatures demonstrate that while specific quantum features like the HOMO energy can be independently governed by local motifs, the joint organization of the broader property manifold remains subject to strict global physical constraints.

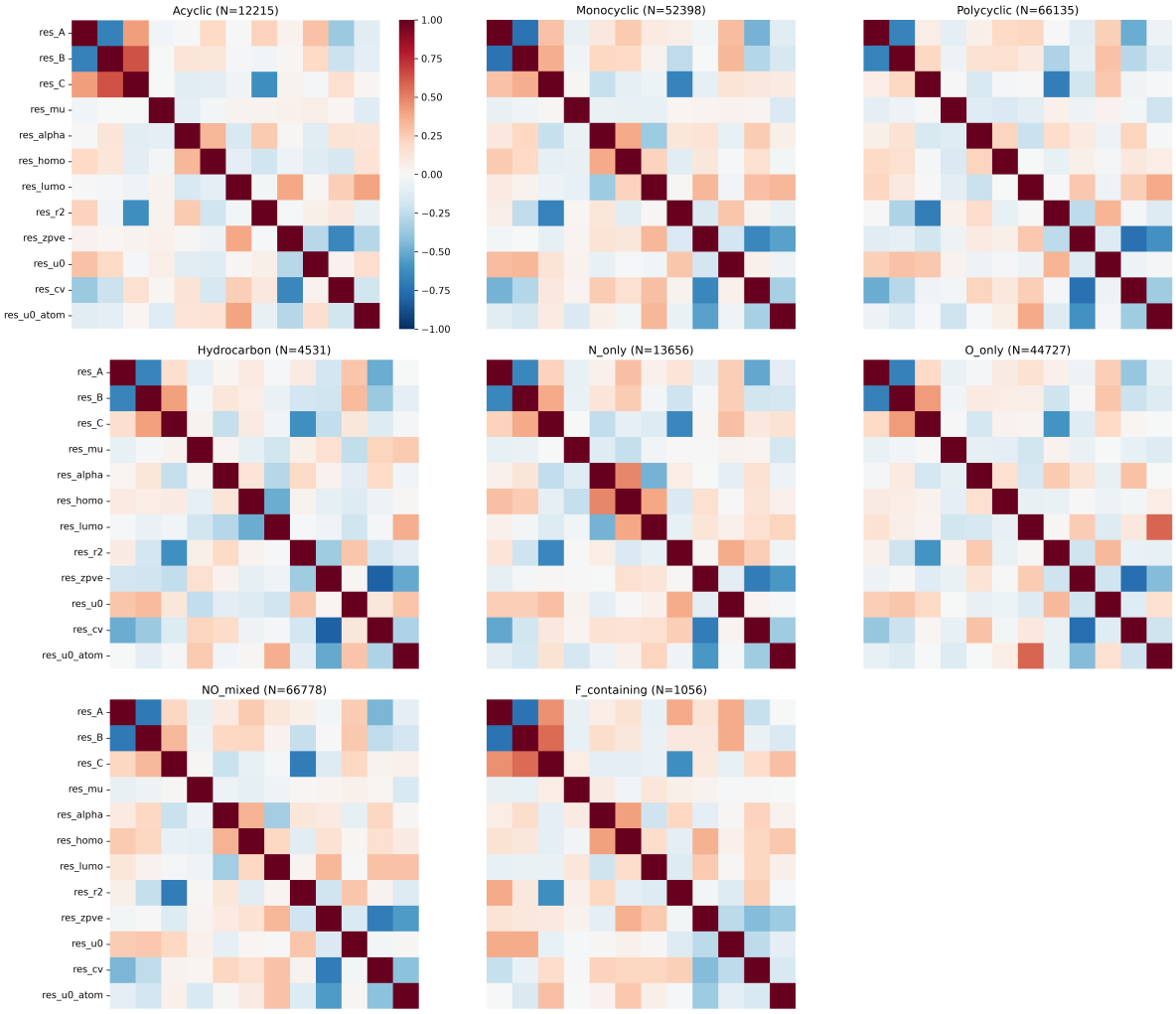

Figure 3: Partial correlation matrices of 12 property residuals stratified by molecular family and elemental composition. The heatmaps demonstrate the remarkable independence of the HOMO energy residuals (res_{homo}) from the global spatial-thermodynamic manifold, maintaining near-zero partial correlations across most families, particularly in monocyclic and polycyclic structures. Acyclic molecules exhibit a slightly higher, yet modest, coupling. Conversely, properties such as isotropic polarizability (res_{α}) and electronic spatial extent (res_{r^2}) remain tightly bound to molecular moments of inertia (res_B, res_C). Furthermore, comparing families like hydrocarbons to F-containing molecules reveals that heteroatom identity significantly reconfigures the broader correlation network of the remaining properties, even while the HOMO energy remains uniquely isolated.

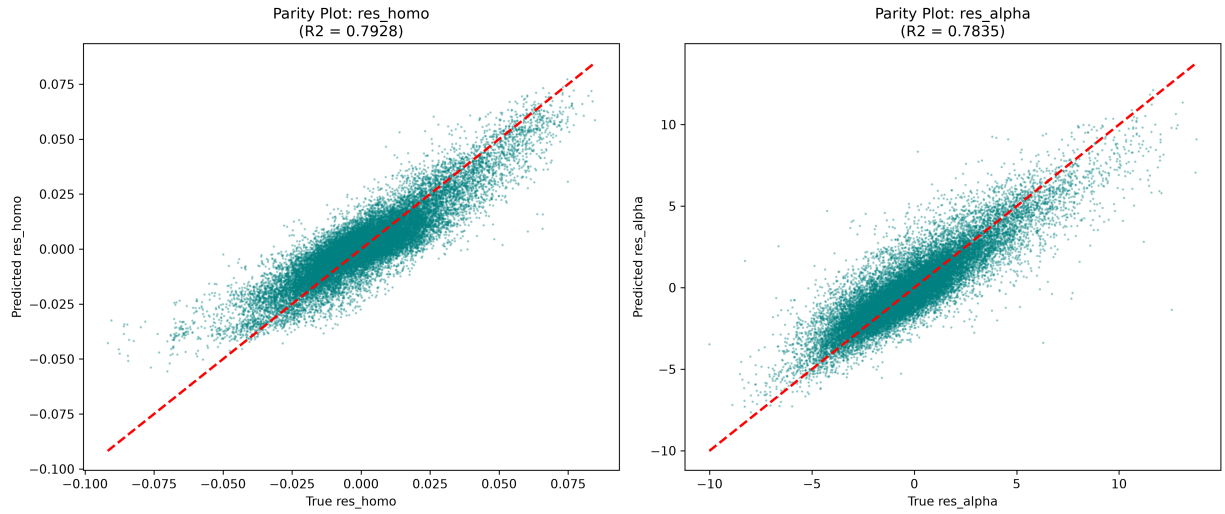

Figure 4: Parity plots comparing true and predicted residuals for the highest occupied molecular orbital energy (res_{homo}, left) and isotropic polarizability (res_{α}, right) using XGBoost models trained on Morgan fingerprints. The scatter distributions around the red dashed identity lines illustrate comparable predictive performance ($R^2 \approx 0.79$), demonstrating that both property residuals are successfully captured by local structural features. Despite this similar accuracy, attribution analysis indicates that their underlying structural drivers are distinct: res_{homo} is uniquely governed by specific local bonding motifs like nitriles rather than global molecular geometry, while the structural attributions for res_{α} are slightly more concentrated.

4 Conclusions

The exploration of the QM9 dataset reveals that the joint organization of molecular properties is heavily dominated by stoichiometric determinism. By establishing a shallow gradient-boosted baseline, we successfully stripped away these trivial size-scaling effects to expose a high-dimensional manifold of chemical residuals. This approach allowed us to isolate genuine structural drivers from macroscopic covariance and rigorously investigate the uniquely decoupled nature of the highest occupied molecular orbital (HOMO) energy.

Our analysis of the stoichiometric baseline also highlighted specific limitations in the underlying computational methodology. We found that geometrically inconsistent structures—those with the highest reconstruction errors—are overwhelmingly enriched for near-linear geometries, such as acetylene derivatives, rather than highly strained rings. This indicates that the B3LYP/6-31G(2df,p) level of theory struggles primarily with the coordinate systems or electronic descriptions of linear molecules, while remaining robust for typical strained motifs.

Within the clean property manifold, the HOMO energy exhibits a remarkable decoupling from both global molecular geometry and total electronegativity, maintaining near-zero partial correlations across most structural families. By employing gradient-boosted regressors trained on Morgan fingerprints, we demonstrated that this independence is driven by highly localized structural environments. Feature attribution pinpointed specific nitrogen-containing motifs—most notably the nitrile group—as the primary drivers of HOMO variance. These localized motifs dictate the quantum behavior of the HOMO independently of the broader topological family or macroscopic scaffold of the molecule.

Furthermore, topological data analysis of the residual space revealed that the underlying chemical manifold is subject to strict global physical constraints. The identification of significant topological voids, validated against synthetic null models and local density checks, confirms the existence of physically or energetically inaccessible regions in property space. The filamentary structure of this space demonstrates that while individual quantum features like the HOMO energy can be governed by localized motifs, the co-variance of the broader property manifold remains fundamentally constrained by the rules of chemical stability.

In summary, the property space of small organic molecules is a composite of a dominant stoichiometric baseline and a topologically complex residual manifold. Our findings emphasize that while global geometric

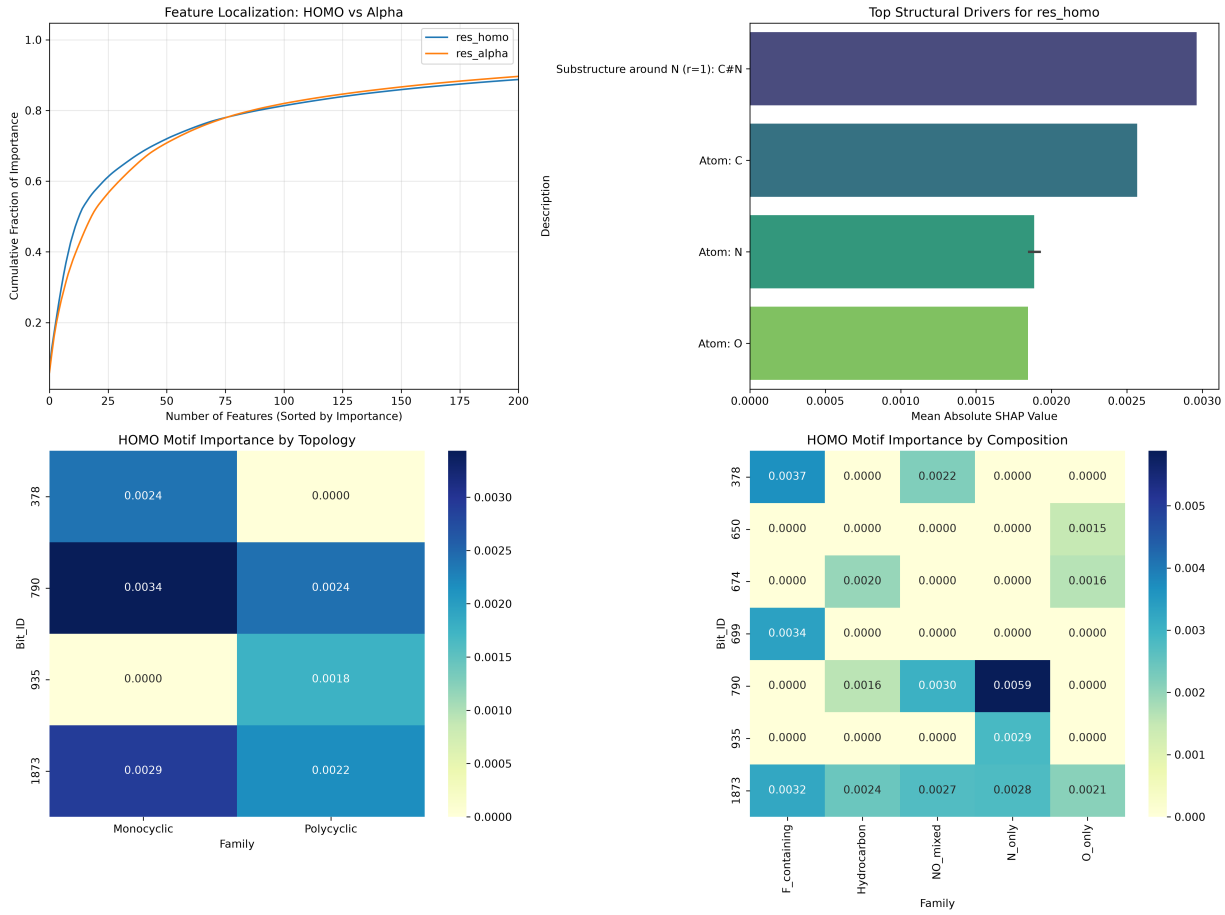

Figure 5: Structural attribution analysis of the HOMO energy and isotropic polarizability residuals (res_{homo} and res_{α}) using XGBoost models trained on Morgan fingerprints. The cumulative fraction of feature importance demonstrates that both properties are highly localized, with res_{α} exhibiting slightly higher concentration than res_{homo}. Mean absolute SHAP values reveal that the primary structural driver for res_{homo} across the dataset is the nitrile group (C≡N) environment. Stratification of motif importance by topological family and elemental composition shows that this specific motif (Bit ID 790) is predominantly critical for monocyclic and nitrogen-only molecules, whereas general carbon hybridization (Bit ID 1873) drives predictions for hydrocarbons and oxygen-containing families. These localized attributions confirm that the HOMO energy is uniquely decoupled from global molecular geometry and is instead governed by specific nitrogen-containing bonding motifs.

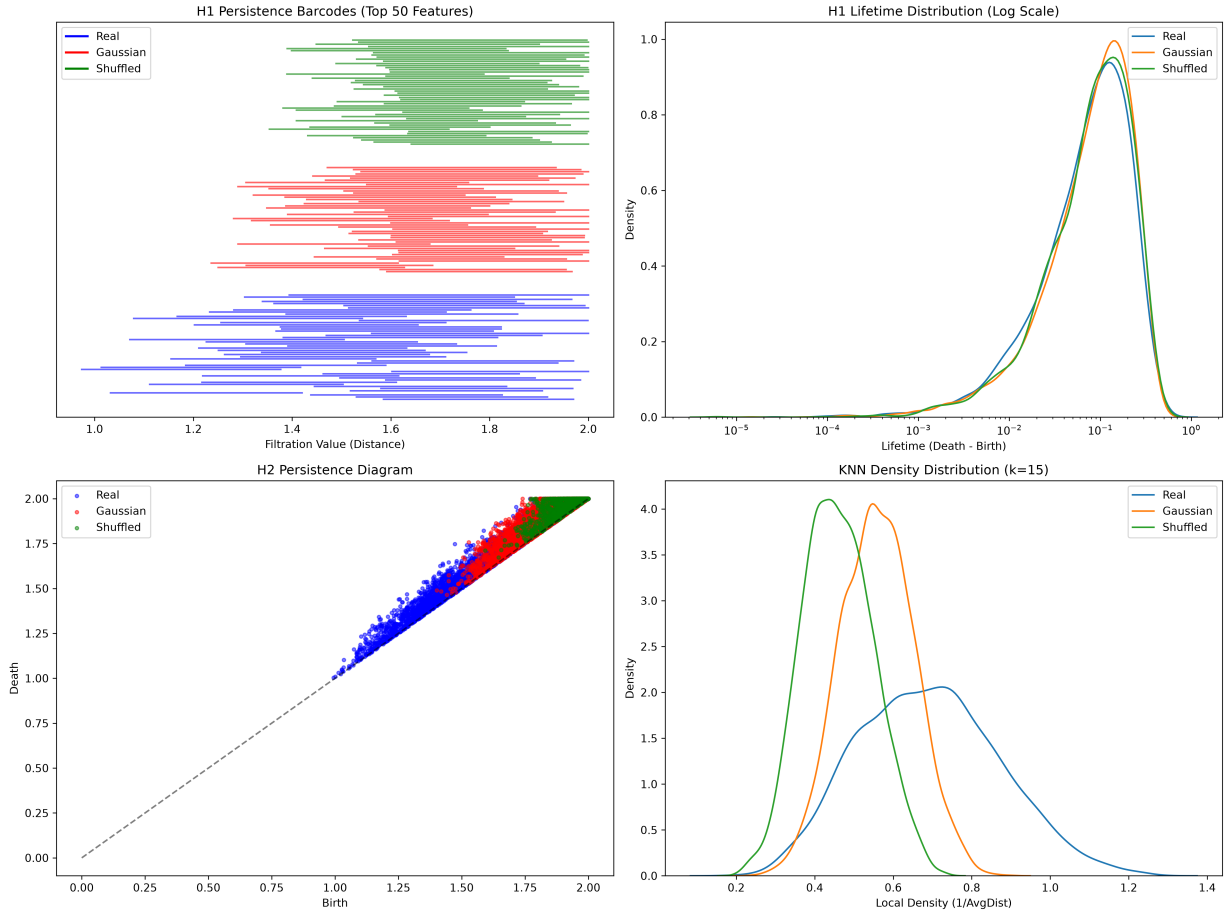

Figure 6: Topological Data Analysis (TDA) of the standardized residual space for a subsample of 5,000 molecules, comparing real data against Gaussian and shuffled null models. The H_1 persistence barcodes and lifetime distributions show that the real chemical space exhibits a higher maximum H_1 persistence (0.6057) compared to the Gaussian (0.4628) and shuffled (0.4741) baselines. Together with the H_2 persistence diagram, this high H_1 persistence reveals a filamentary structure containing significant topological loops, which correspond to forbidden regions of physically or energetically inaccessible property combinations. Furthermore, the k-nearest neighbors (KNN) density distribution confirms these structural voids are genuine physical constraints rather than sampling artifacts, as the real data demonstrates a robust local density that is comparable to or exceeds that of the null models.

and thermodynamic properties are tightly coupled, specific quantum features can remain independently governed by precise local bonding environments. This framework provides a more nuanced understanding of molecular property organization, highlighting the interplay between global physical constraints and localized quantum behavior.