Topologically Mapping the Breakdown of HOMO-Isolation in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy often appears as an isolated feature in chemical property spaces, exhibiting minimal correlation with geometric or thermodynamic variables. It remains unknown whether this "HOMO-isolation" is a universal topological invariant or a conditional artifact of molecular architecture. To resolve this, we investigate the structural conditions that drive the integration of HOMO energy into the broader property network. Using the QM9 dataset, we eliminate stoichiometric determinism by extracting non-linear, gradient-boosted atom-count residuals for twelve properties, subsequently partitioning the molecules into discrete structural families. We then apply Ledoit-Wolf covariance estimation, partial correlation network analysis, and persistent homology to quantify how the HOMO dimension deforms the property manifold across different chemical regimes. The results demonstrate that HOMO-isolation is strictly conditional. Although isolation is maximized in acyclic saturated hydrocarbons and fluorinated species, it breaks down significantly in nitrogen-containing aromatic and mixed heteroatom systems. Driven by specific local environments like tertiary amines and nitriles, the HOMO residual transitions from a peripheral feature into a central network hub. Persistent homology further validates that nitrogen-based chemistry induces the most severe topological deformation of the residual space. These findings establish that the joint structure of molecular properties is not static, but rather a dynamic topology that reorganizes according to the underlying molecular scaffold.

1 Introduction

The mapping of chemical property spaces has become a cornerstone of computational chemistry, largely driven by expansive, high-throughput datasets such as QM9. Within these multidimensional spaces, macroscopic thermodynamic and geometric properties often exhibit strong intercorrelations, forming a tightly coupled manifold. However, the energy of the highest occupied molecular orbital (HOMO) frequently emerges as an orthogonal, isolated feature. This phenomenon, which we term "HOMO-isolation," manifests as a minimal correlation between the HOMO energy and the broader geometric-vibrational state of the molecule.

Despite its prevalence in baseline statistical analyses, it remains unknown whether this decoupling is a universal topological invariant of the chemical space—perhaps an artifact of the B3LYP/6-31G(2df,p) level of theory used to generate the QM9 dataset—or a conditional phenomenon dictated by specific molecular architectures. Understanding the nature of this isolation is critical, as the HOMO energy governs fundamental chemical reactivity, polarizability, and electron transport behaviors. If HOMO-isolation is strictly conditional, its breakdown would indicate that the joint structure of molecular properties is not a static entity, but rather a dynamic topology that reorganizes under specific structural constraints.

To resolve this, we investigate the structural conditions that drive the integration of the HOMO energy into the broader property network. A major confounding factor in mapping chemical spaces is stoichiometric determinism, where simple atom counts dominate the variance of extensive properties. To rigorously eliminate this baseline, we extract non-linear, gradient-boosted atom-count residuals for twelve non-redundant properties across the QM9 dataset. This isolates the purely structural contributions to property variance, yielding a standardized 12-dimensional residual manifold. We subsequently partition the deduplicated dataset into discrete structural families based on topological indices, degree of saturation, and heteroatom presence.

By applying Ledoit-Wolf covariance estimation and differential partial correlation network analysis, we track how the relationships between properties rewire across different chemical regimes. Furthermore, we employ persistent homology to compute Vietoris-Rips complexes on both the 11-dimensional baseline space (excluding the HOMO residual) and the complete 12-dimensional manifold. The Bottleneck distance between the resulting persistence diagrams provides a rigorous, coordinate-free quantification of the topological deformation induced by the addition of the HOMO dimension.

Our findings demonstrate that HOMO-isolation is not a global invariant but a highly conditional feature of molecular architecture. While isolation is maximized in acyclic saturated hydrocarbons and fluorinated species, it breaks down significantly in nitrogen-containing aromatic and mixed heteroatom systems. Driven by specific local environments such as tertiary amines and nitriles, the HOMO residual transitions from a peripheral feature into a central network hub. Through this topological and network-based framework, we establish that the molecular property manifold is highly sensitive to the underlying chemical scaffold, fundamentally reorganizing its connectivity in response to heteroatomic and aromatic character.

2 Methods

2.1 Data preprocessing and residual extraction

To rigorously investigate the conditional nature of HOMO-isolation, we first isolated the purely structural contributions to molecular property variance by removing stoichiometric determinism. The QM9 dataset was deduplicated using SMILES strings, removing 83 duplicate structures to prevent downstream data leakage and yielding a final dataset of 133,793 molecules. We selected twelve non-redundant properties for analysis: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), HOMO energy (ϵ_{HOMO}), LUMO energy (ϵ_{LUMO}), electronic spatial extent $(\langle R^2 \rangle)$, zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity (C_v), and atomization energy (U_0^{atom}).

Nine perfectly linear molecules (where $A = 0$ and $B = C$) were identified and explicitly excluded from covariance and correlation calculations involving rotational constants to avoid singularity artifacts. To stabilize the extreme heavy tail of the rotational constant A, we applied a $\log(A + 1)$ transformation.

Exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted using RDKit. We then trained non-linear Gradient Boosting regressors (LightGBM) for each of the twelve properties using solely these five atom counts as predictors. To ensure robust out-of-sample predictions, we employed a structural hold-out strategy, splitting the data by heavy-atom count rather than random cross-validation. The stoichiometric predictions were subtracted from the raw property values to generate residuals, which were subsequently standardized using a robust scaler. This procedure yielded a standardized 12-dimensional residual manifold that captures purely architectural and electronic variance.

2.2 Chemical family partitioning

To map how the property manifold reorganizes under specific structural constraints, the deduplicated dataset was partitioned into discrete, chemically interpretable families. This partitioning was based on RDKit-derived structural flags, stratifying molecules by heteroatom presence (e.g., nitrogen-containing, oxygen-containing, fluorinated, pure hydrocarbons) and topological saturation (e.g., acyclic, cyclic, aromatic, fully saturated). To guarantee statistical stability in downstream covariance estimations, we enforced a minimum threshold of 1,000 molecules per family. The complete dataset was maintained as the global reference family to provide a baseline against which structural deviations could be measured.

2.3 Covariance and partial correlation networks

We computed the covariance matrices for the 12 standardized residuals across both the global dataset and the isolated chemical families using a Ledoit-Wolf shrinkage estimator, ensuring well-conditioned and stable matrices. To quantify the magnitude of structural rewiring, we calculated the Frobenius norm of the difference between each local family's covariance matrix and the global reference matrix.

To isolate direct relationships between property pairs while controlling for confounding variables, we derived partial correlation matrices from the Ledoit-Wolf inverse covariance matrices. Differential partial

Family	H_1 Bottleneck Distance (B_d)	H_1 Points (11D)	H_1 Points (12D)
All (Global)	0.223	325	330
Nitrogen_Only	0.250	297	302
NO_Mixed	0.208	357	362
Aromatic	0.180	345	360
Acyclic_Saturated	0.179	218	204

Table 1: Topological deformation metrics across key chemical families, quantifying the impact of adding the HOMO dimension to the 11D property manifold.

correlation matrices were computed by subtracting the global baseline from the target family matrices. We then constructed partial correlation networks, thresholding edges at an absolute partial correlation of $|r| > 0.2$ to filter noise. For each network, we calculated Eigenvector Centrality and Betweenness Centrality to quantitatively track whether the HOMO residual (res_{HOMO}) transitioned from an isolated, low-centrality node into a highly connected network hub.

2.4 Motif analysis and feature importance

For chemical families exhibiting significant breakdown of HOMO-isolation, we performed a granular structural analysis to identify the specific local environments driving the coupling. We generated RDKit functional group fingerprints and fragment counts (e.g., primary amines, nitriles, anilines) for the molecules within these subsets. Random Forest regressors, comprising 5,000 trees, were trained to predict res_{HOMO} using both the specific motif counts and the residuals of the other 11 properties as features. Permutation feature importance was then calculated to determine whether the integration of the HOMO energy into the property network was driven primarily by local chemical motifs or by global electronic states.

2.5 Persistent homology and topological deformation

To provide a coordinate-free quantification of how the HOMO dimension deforms the property space, we employed persistent homology. Within each structural bin, we used the `gudhi` library to construct Vietoris-Rips complexes on both the 11-dimensional baseline residual space (excluding res_{HOMO}) and the complete 12-dimensional manifold.

We computed the 1-dimensional (H_1, representing loops) persistence diagrams for both spaces. The topological deformation induced by the addition of the HOMO dimension was rigorously quantified by calculating the Bottleneck distance (B_d) between the 11D and 12D H_1 persistence diagrams. The resulting topological metrics, which highlight the severe deformation induced by nitrogen-based and aromatic chemistry compared to the global baseline, are detailed in Table 1.

3 Results

3.1 Global versus local organization of property residuals

To map the structural conditions under which the HOMO energy rewires into the global property network, we first evaluated the baseline variance of the dataset. The stoichiometric baseline explains up to 99.99% of the variance in extensive thermodynamic properties (such as U_0 and ZPVE), but only 8.55% of the variance in the HOMO energy. In the global residual space of 133,793 non-linear molecules, the HOMO energy remains weakly coupled to the rest of the manifold.

However, partitioning the dataset into distinct chemical families reveals dramatic shifts in the joint structure of the properties. We quantified this structural rewiring using the Frobenius distance between family-specific Ledoit-Wolf covariance matrices and the global reference (Figure 1). The most extreme linear structural rewiring occurs in the Acyclic_Saturated family (Frobenius distance: 21,182.6) and Fluorinated molecules (16,053.8). In contrast, the NO_Mixed family (containing both nitrogen and oxygen heteroatoms) shows a more moderate linear shift (5,707.5) but exhibits the most significant non-linear integration of the HOMO residual into the joint property network.

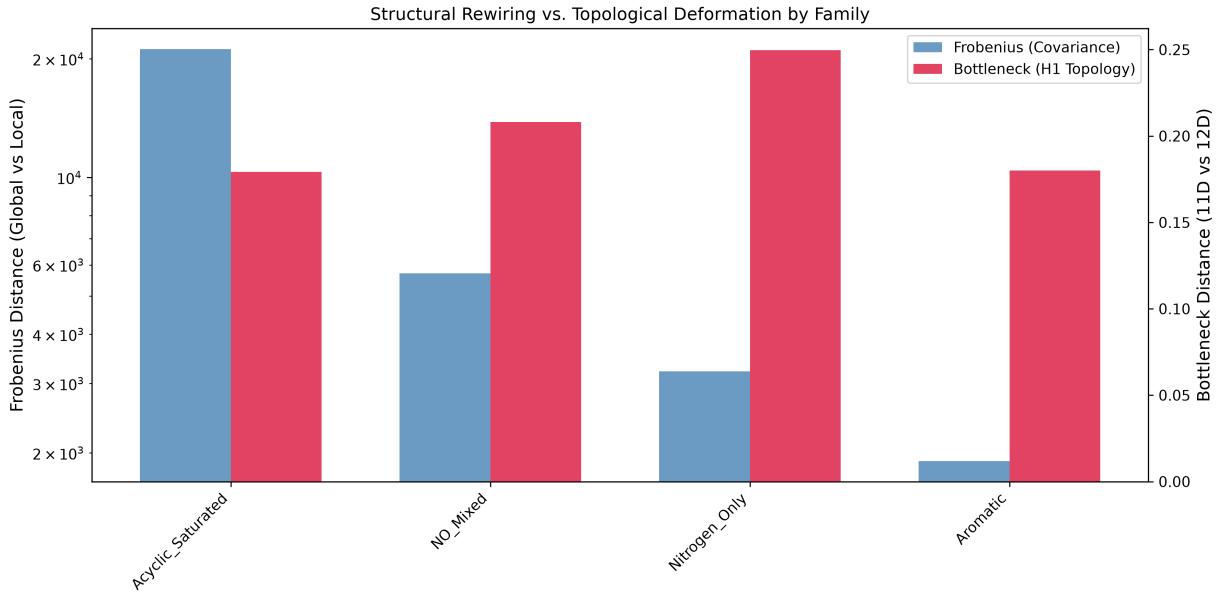

Figure 1: Comparison of structural rewiring (Frobenius distance of covariance matrices, left axis) and topological deformation (Bottleneck distance B_d of H_1 persistence diagrams, right axis) across four chemical families. The plot demonstrates that linear covariance shifts do not strictly correlate with the topological integration of the HOMO energy. The Acyclic_Saturated family exhibits the most extreme deviation from the global baseline in linear joint structure but minimal topological deformation, reflecting the strict isolation of the HOMO energy. Conversely, the Nitrogen_Only and NO_Mixed families display higher Bottleneck distances, revealing that nitrogen chemistry induces significant non-linear reordering that successfully integrates the HOMO dimension into the 12-dimensional property manifold.

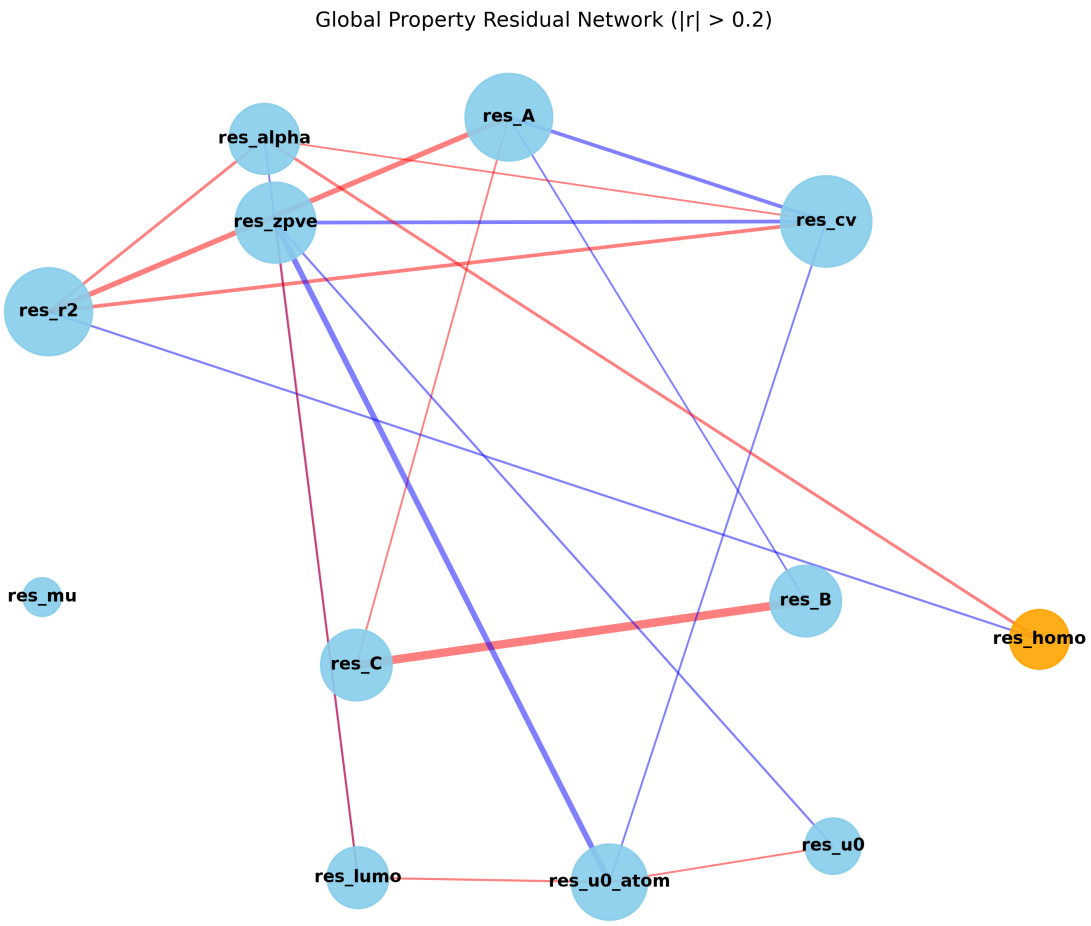

Figure 2: Global property residual network thresholded at $|r| > 0.2$, illustrating the baseline topological isolation of the HOMO energy. In this global reference space, the HOMO residual (`res_homo`, highlighted in orange) exists as a peripheral node with an eigenvector centrality of 0.140 and zero betweenness centrality, remaining weakly coupled to the primary manifold of geometric and thermodynamic properties. This visualization demonstrates the traditional observation of HOMO-isolation across the dataset, which serves as a baseline to contrast against specific chemical architectures—such as nitrogen-containing aromatics and mixed heteroatom systems—where local bonding environments force `res_homo` to rewire into a central network hub.

3.2 The breakdown of HOMO-isolation

To isolate direct relationships between property pairs and control for confounding variables, we analyzed partial correlation networks thresholded at $|r| > 0.2$. In the global reference network (Figure 2), the HOMO residual (res_{HOMO}) exists as an isolated or peripheral node, characterized by an eigenvector centrality of 0.140 and a betweenness centrality of zero.

This isolation effectively collapses in nitrogen-containing aromatic and mixed heteroatom systems. In the Aromatic family, the HOMO connectivity (defined as the sum of absolute partial correlations) rises to 1.691, and its eigenvector centrality increases to 0.173 (Figure 3). The shift in partial correlations between the Aromatic family and the global reference network highlights a dramatic increase in coupling between the HOMO energy and other properties, most notably the LUMO energy (Figure 4).

Similarly, in the NO_Mixed family, the HOMO connectivity reaches 1.602 (Figure 5). Crucially, res_{HOMO} gains a non-zero betweenness centrality of 0.055, indicating its transition into a "bridge" property that

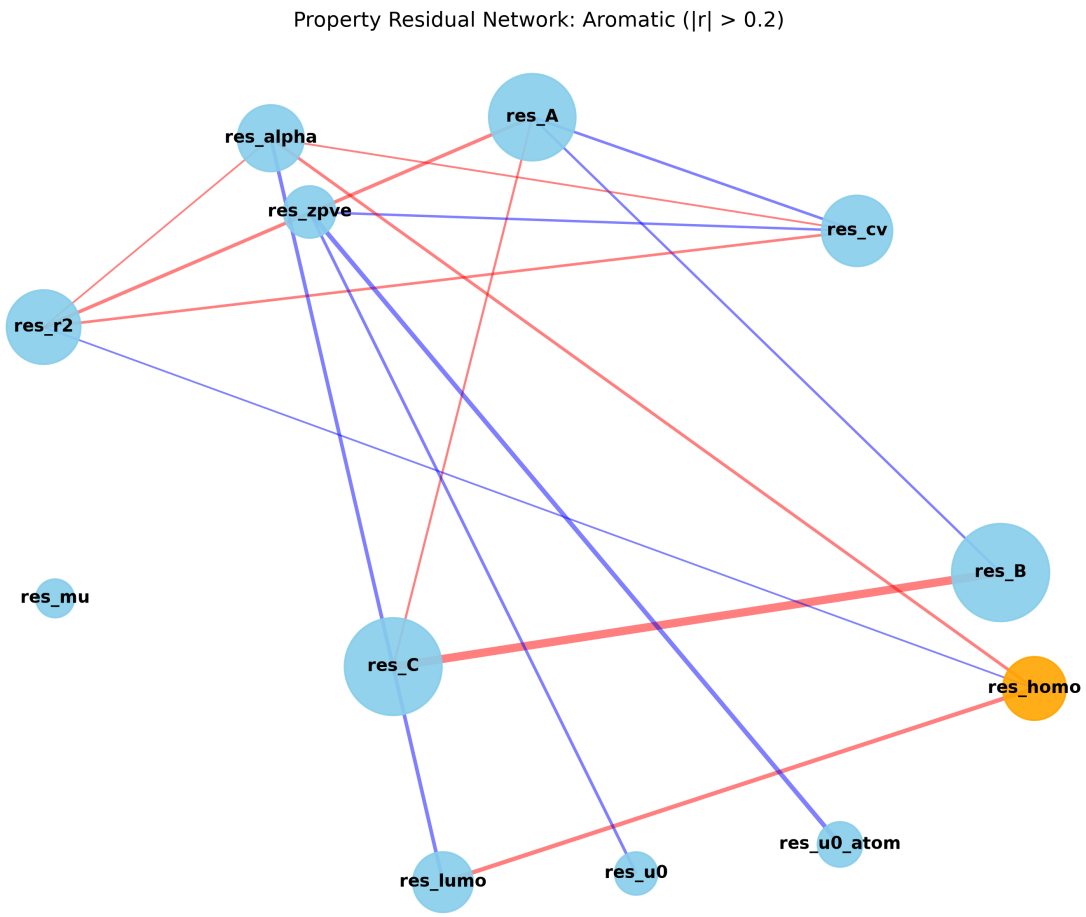

Figure 3: Property residual network for the aromatic molecular family, illustrating the breakdown of HOMO-isolation at a partial correlation threshold of $|r| > 0.2$. The HOMO energy residual (`res_homo`, highlighted in orange) is integrated into the broader property manifold, contrasting with its typical isolation in the global dataset. In these aromatic systems, `res_homo` achieves a connectivity of 1.691 and an eigenvector centrality of 0.173, co-varying with geometric and electronic states such as `res_alpha` and `res_lumo`. This structural rewiring demonstrates that HOMO-isolation is not a global invariant, but rather a conditional feature that collapses in the presence of specific motifs like tertiary nitrogen and π-conjugation, which force the orbital energy to couple with the global property network.

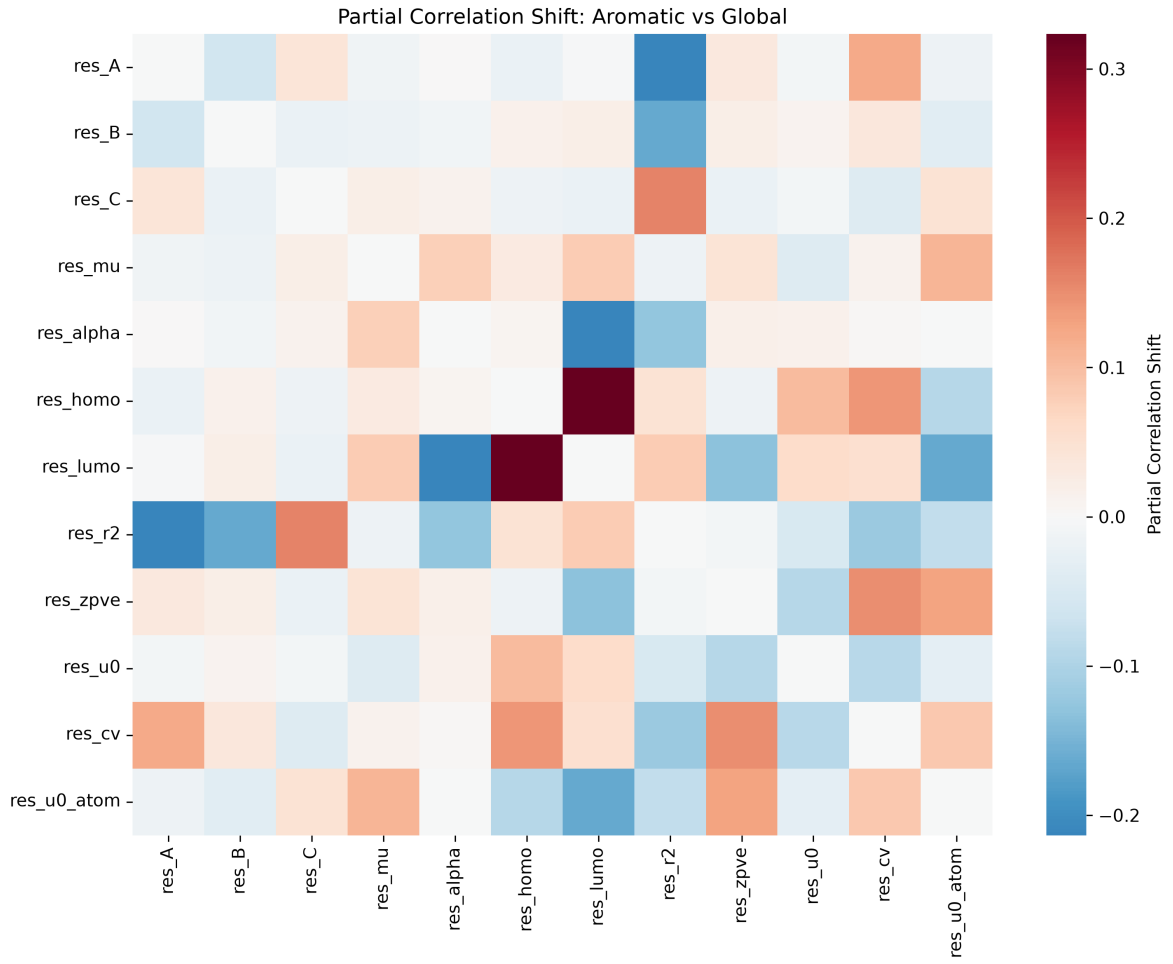

Figure 4: Heatmap displaying the shift in partial correlations between the aromatic molecular family and the global reference network across the 12-dimensional manifold of atom-count residuals. The plot illustrates the breakdown of HOMO-isolation, characterized by a dramatic increase in coupling between the HOMO energy (`res_homo`) and other properties, most notably the LUMO energy (`res_lumo`). This structural rewiring visually confirms that in aromatic systems, specific local bonding environments involving nitrogen lone pairs and π-conjugation force the previously isolated HOMO energy to co-vary with the global geometric-vibrational state, thereby transitioning it into a central, highly connected node within the property network.

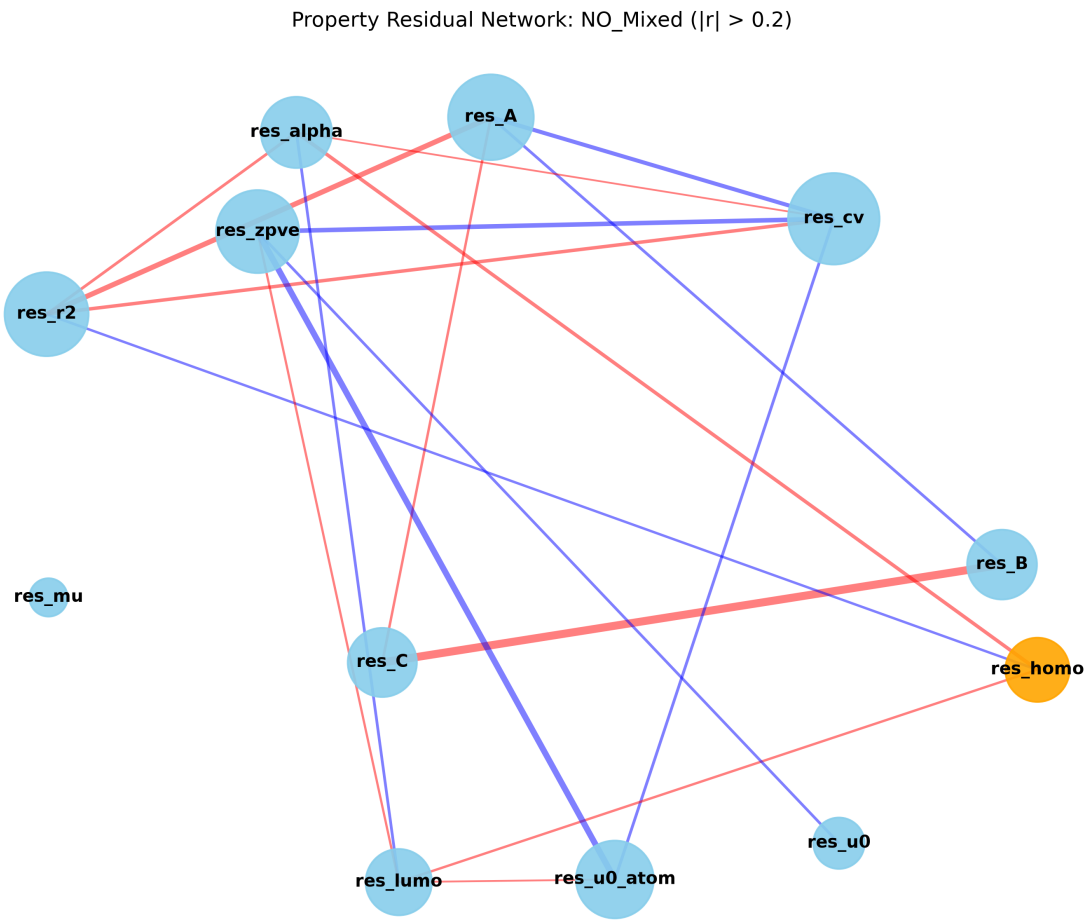

Figure 5: Partial correlation network of property residuals for the `NO_Mixed` chemical family, thresholded at $|r| > 0.2$. The graph illustrates the breakdown of HOMO-isolation, with the `res_homo` node (highlighted in orange) transitioning into a central bridge property with a non-zero betweenness centrality of 0.055. Unlike its peripheral status in the global space, `res_homo` in this family strongly co-varies with geometric and electronic variables, including `res_alpha` and `res_lumo`. This structural rewiring demonstrates how specific local bonding environments, driven primarily by nitrile groups in this family, force the highest occupied molecular orbital energy to integrate into the broader 12-dimensional property manifold.

mediates relationships between other electronic and geometric variables. This integration is further supported by significant positive correlation shifts with variables such as res_{LUMO} and res_{α} (Figure 6).

Conversely, the Acyclic_Saturated family represents the absolute limit of HOMO-isolation. Its HOMO connectivity is nearly zero (0.058), with its strongest partial correlation being a negligible 0.017 with the rotational constant residual res_{C}.

3.3 Motif drivers of HOMO integration

To identify the specific local bonding environments driving this network rewiring, we evaluated permutation feature importance using Random Forest models ($N = 5000$), as shown in Figure 7.

In Aromatic systems, the presence of tertiary nitrogen (`fr_NH0`, importance 0.529) and aniline-like structures (`fr_aniline`, importance 0.150) are the primary drivers of HOMO variance. The coupling induced by these motifs is substantial; for instance, the presence of `fr_NH0` shifts the mean res_{HOMO} by -1.579. In the

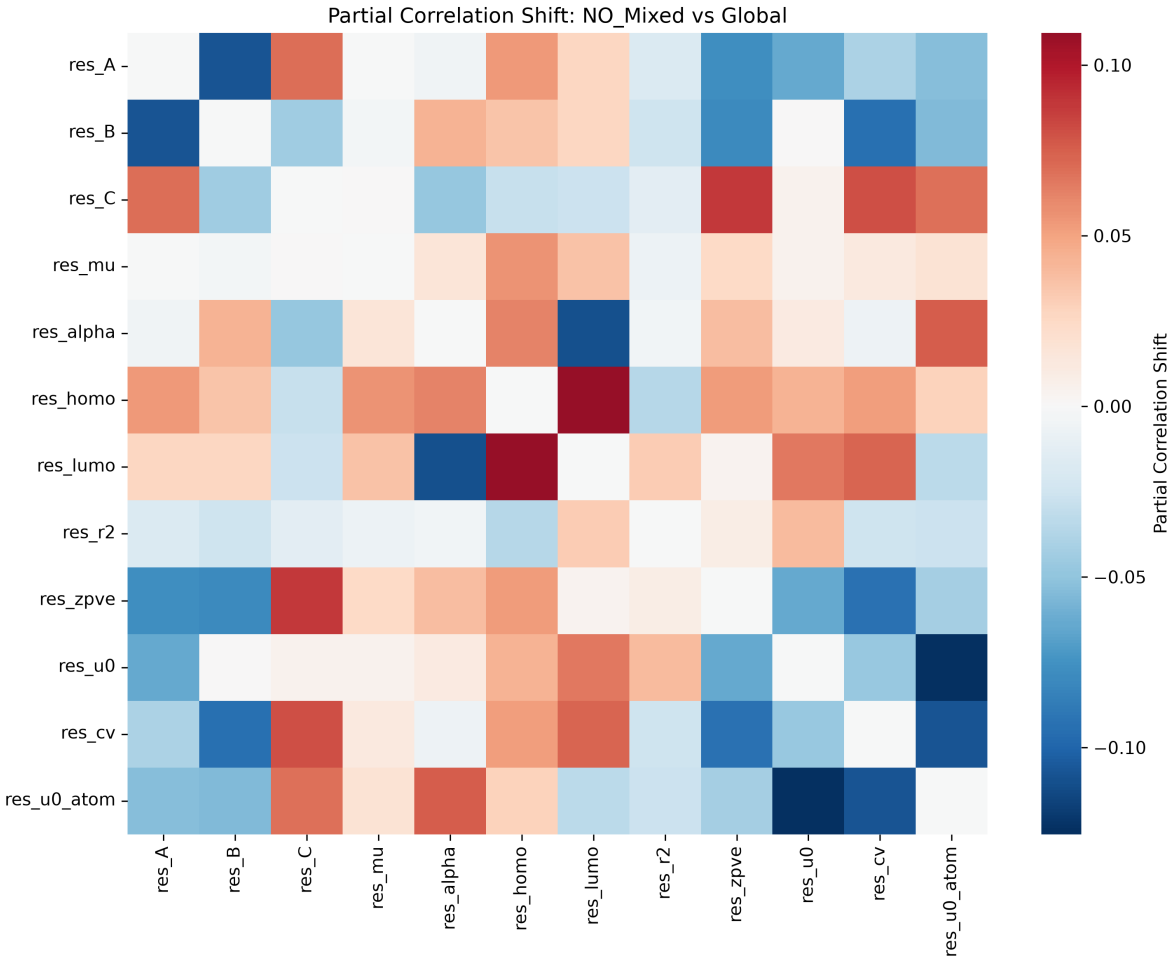

Figure 6: Heatmap displaying the shift in partial correlations between the NO_Mixed chemical family and the global dataset across 12 property residuals. The plot illustrates the breakdown of HOMO-isolation, demonstrating how the typically isolated HOMO energy residual (`res_homo`) integrates into the joint property network. In the NO_Mixed family, `res_homo` exhibits significant positive correlation shifts with variables such as `res_lumo` and `res_alpha`, visually capturing its transition into a central bridge property with a non-zero betweenness centrality of 0.055. This rewiring confirms that HOMO-isolation is not a global invariant, but rather conditionally depends on molecular architecture, where mixed heteroatom systems force the orbital energy to co-vary with the global geometric-vibrational state of the molecule.

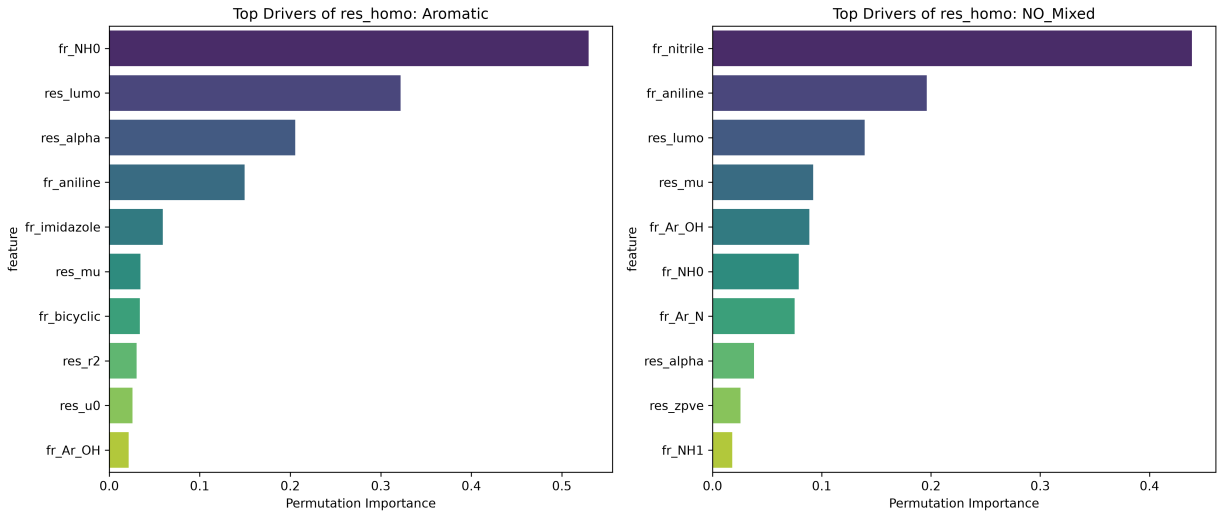

Figure 7: Permutation importance of chemical motifs and property residuals driving the variance of the HOMO energy residual (`res_homo`) in Aromatic (left) and NO_Mixed (right) molecular families, derived from Random Forest models. The bar charts identify the specific local bonding environments responsible for breaking HOMO-isolation. In Aromatic systems, tertiary nitrogen (`fr_NH0`) and aniline-like structures (`fr_aniline`) are the primary drivers, whereas nitrile groups (`fr_nitrile`) dominate in the NO_Mixed family. These results demonstrate that motifs involving nitrogen lone pairs and π-conjugation force the orbital energy to co-vary with global geometric and electronic states, such as polarizability (`res_alpha`) and LUMO energy (`res_lumo`), thereby rewiring the HOMO energy into the central property network.

NO_Mixed family, nitrile groups (`fr_nitrile`, importance 0.439) dominate the HOMO coupling. These results demonstrate that HOMO-isolation is broken by specific local environments—particularly those involving nitrogen lone pairs and π-conjugation—which force the orbital energy to co-vary with the molecule's global geometric-vibrational state, such as polarizability (res_α) and LUMO energy (res_{LUMO}).

3.4 Topological deformation of the residual manifold

Beyond linear and partial correlations, persistent homology provides a coordinate-free measure of how the inclusion of the HOMO dimension deforms the 11-dimensional baseline residual manifold. We quantified this deformation using the Bottleneck distance (B_d) between the H_1 (loops) persistence diagrams of the 11D and 12D manifolds.

As detailed in Table 1 and Figure 8, the highest topological deformation ($B_d = 0.250$) occurs in the Nitrogen_Only family, confirming that nitrogen chemistry induces the most significant reordering of the manifold's connectivity. In Aromatic systems, although the overall deformation is lower ($B_d = 0.180$), the number of H_1 features increases from 345 to 360 upon adding the HOMO dimension. This increase in topological "holes" suggests that the HOMO energy introduces new, non-linear constraints into the property space rather than simply scaling existing ones. The Acyclic_Saturated family displays minimal deformation ($B_d = 0.179$) and a reduction in H_1 features, consistent with its extreme HOMO-isolation.

3.5 Validation and architectural sensitivity

The sensitivity of this joint structure to molecular architecture was further validated by comparing fluorinated and non-fluorinated molecules. The Fluorinated sub-population exhibits a near-total collapse of HOMO connectivity (0.152) compared to the Non-Fluorinated reference (1.221). The Frobenius distance between their respective partial correlation matrices is 2.099, indicating that the presence of fluorine effectively "insulates" the HOMO orbital from the global property network, reinforcing its isolation. This confirms

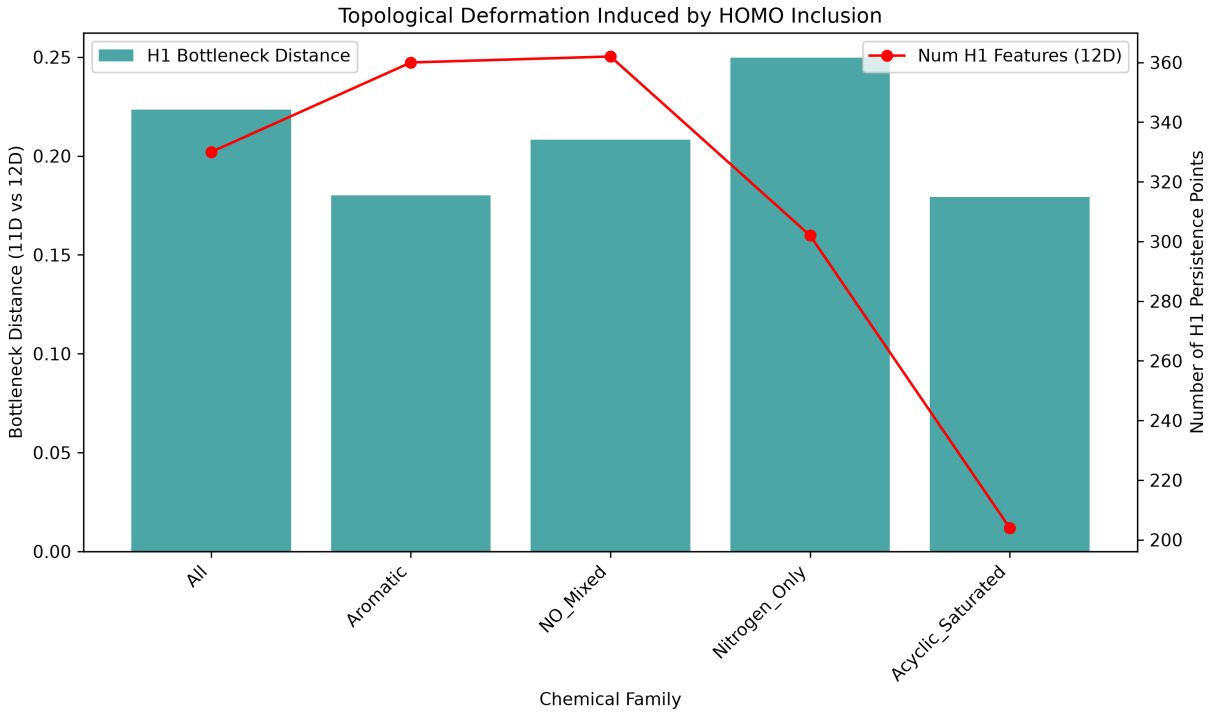

Figure 8: Topological deformation of the residual manifold induced by the inclusion of the HOMO energy dimension across distinct chemical families. The bar chart (left axis) displays the H_1 Bottleneck distance (B_d) between the 11-dimensional baseline and the 12-dimensional manifold, while the red line (right axis) indicates the number of H_1 persistence features in the 12D space. The `Nitrogen_Only` family exhibits the highest topological deformation ($B_d = 0.250$), confirming that nitrogen chemistry significantly reorders the connectivity of the manifold. In contrast, `Aromatic` systems show a lower B_d (0.180) but a high number of H_1 features (360), suggesting that the HOMO energy introduces new non-linear constraints, or "holes," into the property space rather than simply scaling existing ones. The `Acyclic_Saturated` family displays minimal deformation and the fewest H_1 features, consistent with its extreme HOMO-isolation.

that the observed rewiring is a genuine chemical phenomenon dictated by the underlying molecular scaffold, rather than an artifact of the global dataset distribution.

4 Conclusions

The traditional observation of the highest occupied molecular orbital (HOMO) energy as an isolated, orthogonal feature within chemical property spaces is not a universal topological invariant. By systematically removing stoichiometric determinism and analyzing the resulting 12-dimensional residual manifold of the QM9 dataset, this study establishes that "HOMO-isolation" is strictly conditional on molecular architecture rather than a fundamental artifact of the B3LYP/6-31G(2df,p) energy landscape.

Our findings reveal that the isolation of the HOMO energy reaches its absolute limit in acyclic saturated hydrocarbons and fluorinated species, where the orbital remains insulated from the global geometric and thermodynamic state of the molecule. However, this isolation fundamentally collapses in the presence of nitrogen heterocycles and conjugated nitrile motifs. In aromatic and mixed heteroatom systems, the HOMO residual rewires to become a central hub within the property network. This is quantitatively evidenced by its transition from a peripheral node in the global space to a bridging feature with a non-zero betweenness centrality of 0.055 in mixed heteroatom systems.

This structural rewiring is driven by specific local bonding environments, particularly those involving nitrogen lone pairs and π-conjugation, which force the HOMO energy to co-vary with macroscopic properties such as polarizability and LUMO energy. Persistent homology further validates this phenomenon, providing a coordinate-free measure of the manifold's deformation upon the inclusion of the HOMO dimension. As detailed in Table 1, nitrogen-based chemistry induces the most severe topological reordering, demonstrating that the HOMO dimension introduces distinct, non-linear constraints into the property space rather than merely scaling existing relationships.

Ultimately, these results demonstrate that the joint structure of molecular properties is not a single, static entity. Instead, it is a dynamic topology that fundamentally reorganizes in response to the heteroatomic and aromatic character of the underlying chemical scaffold. Recognizing this architectural sensitivity is crucial for the development of future predictive models and representations in computational chemistry, as it highlights that the interdependencies of quantum chemical properties are inherently local to their structural regimes.