Conditional Decoupling of Frontier Orbitals from the Vibrational-Spatial Manifold Across Topological Transitions

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The extent to which frontier electronic states, particularly the highest occupied molecular orbital (HOMO), operate independently of a molecule's global spatial and thermodynamic organization remains an open question. To determine whether this electronic decoupling is a fundamental constant or an emergent property, we analyzed 133,802 molecules from the QM9 dataset. We isolated chemical organization from stoichiometric determinism by extracting gradient-boosted atom-count residuals for core molecular properties. By partitioning the dataset into size-controlled topological classes, we employed distance correlation matrices, partial distance correlation, and Mantel tests to evaluate the coupling of frontier orbitals to the spatial-vibrational manifold across structural boundaries. Our analysis reveals that while the HOMO is largely decoupled from the global spatial-vibrational core in acyclic systems, this independence is highly conditional. Topological transitions, specifically the introduction of aromaticity, act as a synchronizing force that significantly increases the coupling between electronic states and the global molecular frame. Conversely, the introduction of nitrogen heteroatoms disrupts this synchronization, re-establishing electronic decoupling through localized effects. Furthermore, motif enrichment analysis demonstrates that extreme decoupling is driven by specific structural features, such as zwitterionic fragments and strained heterocycles. These findings establish that frontier orbital decoupling is not an absolute physical constant but a highly conditional feature governed by molecular topology and local electronic disruptions.

1 Introduction

The mapping between molecular architecture and quantum mechanical properties is a central pillar of chemical physics and materials design. Large-scale computational databases, such as the QM9 dataset, have provided unprecedented opportunities to explore these relationships across vast regions of chemical space. While extensive properties like internal energy and zero-point vibrational energy are heavily dictated by stoichiometric composition, intensive electronic properties—particularly the energies of frontier orbitals such as the highest occupied molecular orbital (HOMO)—exhibit significant variance that cannot be explained by simple atom counts.

A persistent question in the analysis of molecular property manifolds is the extent to which these frontier electronic states operate independently of a molecule's global spatial and thermodynamic organization. Often, the HOMO appears orthogonal to the core vibrational-spatial manifold, suggesting a degree of electronic decoupling. However, it remains unclear whether this decoupling is an absolute physical constant inherent to small organic molecules or an emergent, conditional feature that breaks down under specific structural or topological constraints.

To determine the nature of this electronic decoupling, we systematically investigate the joint structure of electronic, thermodynamic, and geometric properties across 133,802 unique molecules from the QM9 dataset. Our methodology explicitly isolates chemical organization from stoichiometric determinism. By extracting gradient-boosted atom-count residuals for core molecular properties, we remove the dominant size-dependent variance, allowing us to focus purely on structural and topological effects. We then partition the dataset into size-controlled topological classes—including acyclic chains, non-aromatic monocycles, aromatic monocycles, and bicyclic systems—to conduct a targeted sensitivity analysis of frontier orbital variance.

Within these specific structural families, we employ distance correlation matrices and partial distance correlation to evaluate the coupling of the HOMO to the global spatial-vibrational manifold, which comprises properties such as zero-point vibrational energy, heat capacity, electronic spatial extent, and polarizability. By contrasting the structural dependence of the HOMO across macroscopic structural boundaries, we identify the precise architectural thresholds where frontier electronic landscapes reorganize. Furthermore, we utilize Mantel tests to ensure the statistical stability of these manifolds and perform motif enrichment analysis to uncover the structural signatures of extreme decoupling.

Our analysis reveals that frontier orbital decoupling is highly conditional and governed by molecular topology. While the HOMO is largely independent of the global spatial-vibrational core in acyclic systems, topological transitions fundamentally alter this relationship. Specifically, the introduction of aromaticity acts as a synchronizing force, significantly increasing the coupling between electronic states and the global molecular frame. Conversely, the introduction of nitrogen heteroatoms disrupts this synchronization, re-establishing electronic decoupling through localized effects. Finally, we demonstrate that extreme decoupling is driven by specific structural features, such as zwitterionic fragments and strained heterocycles, confirming that the isolation of frontier orbitals is an emergent property dictated by local electronic disruptions and global topological constraints.

2 Methods

2.1 Data preparation and cleaning

The QM9 dataset, comprising 133,885 small organic molecules, was loaded and initially filtered to remove 83 duplicated SMILES strings to prevent data leakage. To focus on the fundamental property manifold, redundant thermodynamic and atomization columns (such as U_{298}, H_{298}, G_{298}, and their atomization equivalents) along with the HOMO-LUMO gap were excluded, reducing the feature space to 12 core properties. Furthermore, nine strictly linear molecules (characterized by a rotational constant $A = 0$) and other effectively degenerate structures were removed to prevent undefined rotational constants and extreme outliers from skewing downstream manifold analyses. Molecular structures were parsed from SMILES strings using RDKit, from which exact atom counts (C, N, O, F, H), heavy atom counts, and ring descriptors (total rings and aromatic rings) were extracted. During this process, the RDKit logger was globally suppressed to optimize parallel processing performance.

2.2 Dimensionality reduction and stoichiometric baseline extraction

To ensure that the observed joint structure was not an artifact of redundant reporting, Principal Component Analysis (PCA) was performed on both the standardized full 19-property set and the reduced 12-property set. Following this dimensionality check, the rotational constants (A, B, C) were isolated into a separate geometric constraint block, as their variance is heavily skewed by hydrogen presence and behaves fundamentally differently from the other properties. The remaining nine core properties were retained for the primary manifold analysis.

To eliminate stoichiometric determinism and isolate the structural variance, a LightGBM multi-output regressor was trained to predict the nine core properties using strictly the atom counts (C, N, O, F, H). The residuals from this regression were extracted, representing the size-independent property variations. The variance of these residuals was validated against the raw property variances to ensure that a meaningful chemical signal remained, particularly for intensive properties like the HOMO energy ($homo$), LUMO energy ($lumo$), and dipole moment (μ).

2.3 Topological stratification and heteroatom perturbation

Molecules were partitioned into distinct topological classes based on RDKit ring descriptors: acyclic chains, non-aromatic monocycles, aromatic monocycles, bicyclic systems, and polycyclic systems. To decouple topological rewiring from size-induced artifacts, the populations were filtered to a strictly fixed heavy-atom count (e.g., $N = 9$).

Within these size-controlled topological bins, a heteroatom perturbation analysis was conducted. For instance, aromatic monocycles with nine heavy atoms were further stratified into subsets containing exactly zero, one, or two or more nitrogen atoms. This stratification isolated the electronic rewiring effects of heteroatom substitution from the geometric constraints of the ring structure.

2.4 Distance correlation and manifold analysis

Using the nine size-independent property residuals, the first principal component (PC1) of the spatial-vibrational manifold—comprising the zero-point vibrational energy (res_zpve), heat capacity (res_c_v), electronic spatial extent (res_r^2), and polarizability (res_α)—was computed. The distance correlation ($dCor$) between the HOMO residual (res_homo) and this PC1 was calculated to evaluate the decoupling of the HOMO state from the global spatial organization.

To identify mediating properties between the electronic and thermodynamic domains, partial distance correlations were computed. Specifically, the correlation between res_homo and res_zpve was conditioned on candidate hub properties, such as polarizability (res_α), to test whether the observed relationships were direct or mediated. Full pairwise distance correlation matrices were generated for the nine residuals across each topological class, and difference heatmaps were constructed for consecutive topological transitions to highlight property pairs that decouple or synchronize.

2.5 Motif extraction and stability validation

Molecules exhibiting extreme decoupling or unexpected coupling between the HOMO residual and the spatial manifold were identified as outliers. Morgan fingerprints were generated for these molecules to compute the enrichment frequencies of specific structural motifs (e.g., zwitterionic fragments, strained heterocycles) relative to the baseline population of their respective topological bins. Visual grids of the most decoupled molecules were generated using RDKit's `MolsToGridImage` function to facilitate chemical interpretability.

Finally, to confirm that the observed topological and heteroatom rewiring was a robust chemical phenomenon rather than a statistical artifact, a split-half stability validation was performed. Using `joblib` parallelized across 14 CPUs, each size-controlled topological and heteroatom bin was randomly bisected. Distance correlation matrices were independently recomputed for both halves, and a Mantel test was executed to quantify the correlation between the two distance matrices, yielding the Mantel stability statistic (r_{Mantel}) for each structural family.

3 Results

3.1 Dimensionality and the stoichiometric baseline

A fundamental characteristic of the QM9 dataset is that molecular composition—specifically, basic atom counts—dictates the majority of variance in several key extensive properties. By regressing the core properties against atom counts (C, N, O, F, H), we established a stoichiometric baseline to isolate structural variance. Linear regression achieves near-exhaustive variance reduction for thermodynamic and vibrational properties, yielding $R^2 = 0.9992$ for internal energy (u_0), $R^2 = 0.9973$ for zero-point vibrational energy ($zpve$), and $R^2 = 0.9897$ for atomization energy (u_{0_atom}). In contrast, electronic and spatial properties exhibit significant independence from simple stoichiometry. The highest occupied molecular orbital energy ($homo$), dipole moment (μ), and electronic spatial extent (r^2) retain the majority of their variance, with R^2 values of 0.1978, 0.2759, and 0.3631, respectively. This stoichiometric determinism and the contrasting independence of electronic features are illustrated in Figure 1.

Following the removal of the stoichiometric baseline, we evaluated the effective dimensionality of the property space. Removing redundant columns—specifically the thermodynamic and atomization quartets alongside the HOMO-LUMO gap—paradoxically increases the apparent dimensionality of the dataset. Principal Component Analysis (PCA) reveals that while 6 components explain 91.29% of the variance in the full 19-column set, 7 components are required to explain 94.14% of the variance when restricted to the 12 core properties (Figure 2). This confirms that the apparent low dimensionality of the QM9 dataset is primarily a side effect of redundant column reporting rather than an intrinsic feature of the molecular space.

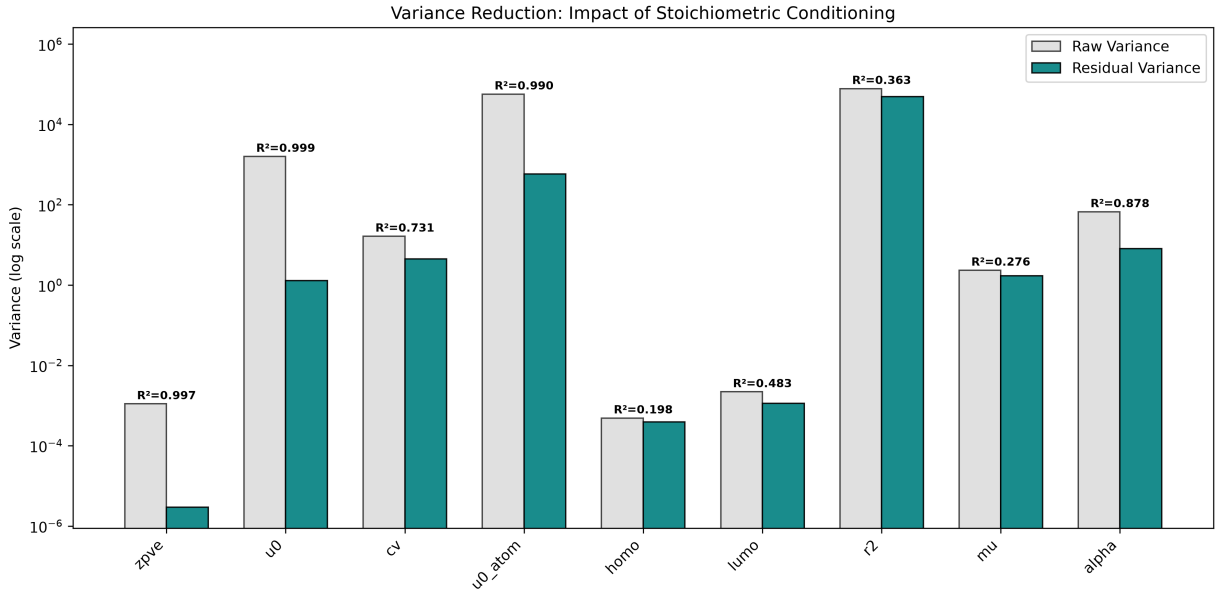

Figure 1: Comparison of raw and residual variance (log scale) across molecular properties following stoichiometric conditioning. The plot illustrates the stoichiometric determinism inherent in the dataset, demonstrating that molecular composition dictates the majority of variance for thermodynamic and vibrational properties. Specifically, internal energy (u_0), zero-point vibrational energy ($zpve$), and atomization energy achieve near-exhaustive variance reduction ($R^2 > 0.98$). Conversely, electronic and spatial properties such as the highest occupied molecular orbital energy ($homo$), dipole moment (μ), and electronic spatial extent (r^2) exhibit significant independence from simple stoichiometry, retaining the majority of their variance ($R^2 < 0.4$). This confirms that while basic atom counts govern thermodynamic baselines, spatial and electronic features maintain distinct structural degrees of freedom.

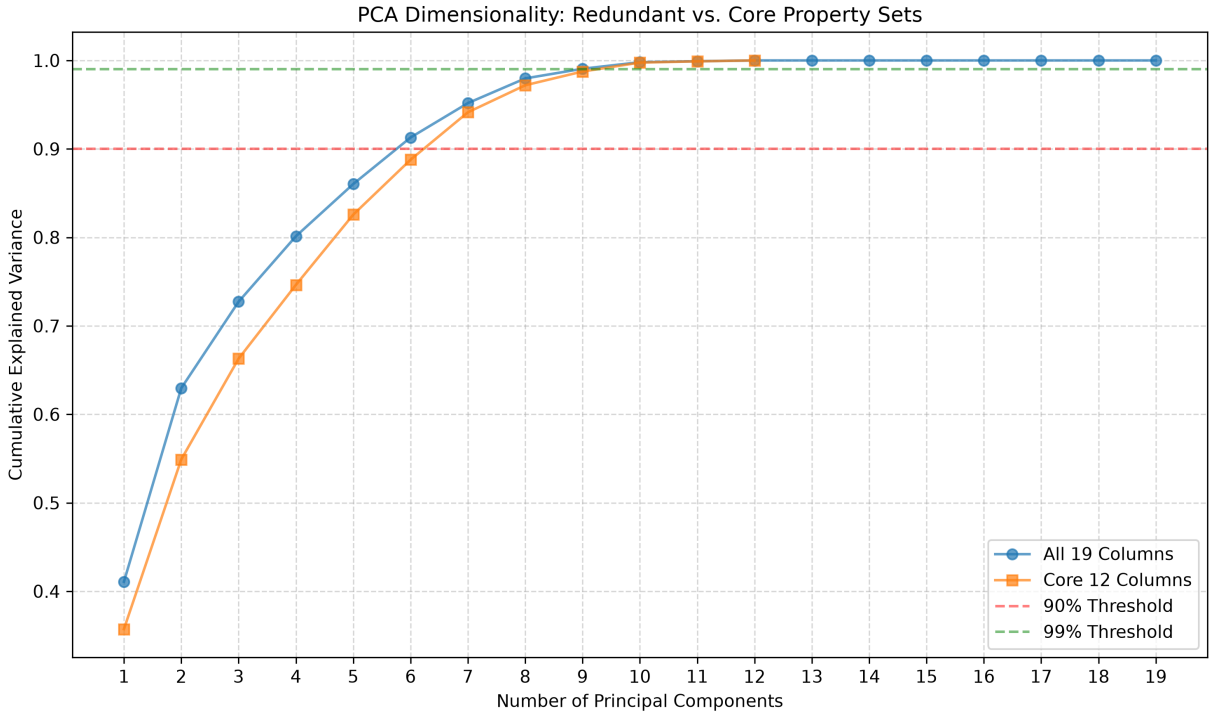

Figure 2: Cumulative explained variance from Principal Component Analysis (PCA) comparing the full 19-column property set to the restricted 12-column core set of the QM9 dataset. The plot illustrates that while 6 principal components explain over 90% (specifically 91.29%) of the variance in the full set, 7 components are required to surpass the 90% threshold (reaching 94.14%) when restricted to the core properties. This shift indicates that removing redundant thermodynamic and atomization columns increases the effective dimensionality of the dataset, demonstrating that the apparent low dimensionality of QM9 is primarily a side effect of redundant property reporting rather than an intrinsic feature of the molecular space.

3.2 The spatial-vibrational core and HOMO decoupling

Analysis of the size-independent property residuals reveals a tightly coupled core manifold comprising res_zpve, res_c_v, res_r^2, and res_α. Together, these properties represent the global spatial and vibrational organization of the molecule.

Against this core manifold, the residual of the highest occupied molecular orbital (res_homo) consistently exhibits low coupling. In acyclic molecules, the distance correlation ($dCor$) between res_homo and the first principal component (PC1) of the spatial-vibrational manifold is only 0.2494. Furthermore, partial distance correlation analysis indicates that this weak relationship is largely mediated by polarizability. When conditioned on res_α, the partial distance correlation between res_homo and res_zpve vanishes to 0.0533 in acyclic systems, demonstrating a near-universal independence of the HOMO from the global spatial-vibrational frame in the absence of topological constraints.

3.3 Structural rewiring across topological transitions

While the HOMO is largely decoupled in acyclic systems, the joint structure of molecular properties undergoes significant reorganization when molecules transition across macroscopic topological boundaries.

Aromaticity, in particular, acts as a strong synchronizing force that couples electronic states to the global molecular frame. In aromatic monocycles, the $dCor$ between res_homo and the spatial-vibrational PC1 rises to 0.4104. Pairwise distance correlations further show that aromaticity increases the coupling between res_homo and res_lumo by 0.3871 relative to acyclic systems, indicating the emergence of a highly coherent frontier electronic landscape (Figure 3, left panel).

Conversely, the introduction of heteroatoms into these synchronized systems triggers a systematic electronic disruption. Comparing aromatic monocycles with zero versus one nitrogen atom reveals that the coupling between res_homo and res_r^2 decreases by 0.1944 (Figure 3, right panel). This structural rewiring suggests that nitrogen lone pairs and electronegativity effects introduce localized electronic variations that operate independently of the molecule's global spatial extent, effectively re-establishing the decoupling of the frontier orbitals.

3.4 Stability and structural signatures of decoupling

To ensure that the observed property manifolds and their topological rewiring are fundamental to the molecular families rather than statistical artifacts, we performed split-half stability validation using Mantel tests. The mean Mantel stability (r_{Mantel}) exceeds 0.91 for all major topological classes, including acyclic chains (0.9378), bicyclic systems (0.9380), and aromatic monocycles (0.9121). These consistently high correlations confirm the statistical robustness of the spatial-vibrational and electronic organizations across the dataset (Figure 4).

Finally, we investigated the structural origins of orbital decoupling through motif enrichment analysis. Molecules exhibiting extreme decoupling between the HOMO residual and the spatial manifold—such as `gdb_32465` and `gdb_46611`—are characterized by highly specific structural signatures (Figure 5). Motif enrichment analysis identifies Morgan fingerprint bit ID 1618404637 as being enriched in the decoupled population by a factor of 228.15 relative to the baseline. These enriched motifs are predominantly associated with zwitterionic fragments and strained heterocyclic environments, such as nitro groups and protonated amines. The prevalence of these features indicates that extreme decoupling is driven by motifs that create localized electronic states, which remain shielded from the molecule's global geometric and thermodynamic environment.

4 Conclusions

We have systematically investigated the conditional decoupling of frontier orbitals from the global spatial-vibrational manifold across 133,802 molecules in the QM9 dataset. By rigorously isolating structural variance from stoichiometric determinism, we demonstrated that the apparent orthogonality of the highest occupied molecular orbital (HOMO) is not an absolute physical constant. Rather, it is an emergent property that is heavily dependent on molecular topology and local electronic environments.

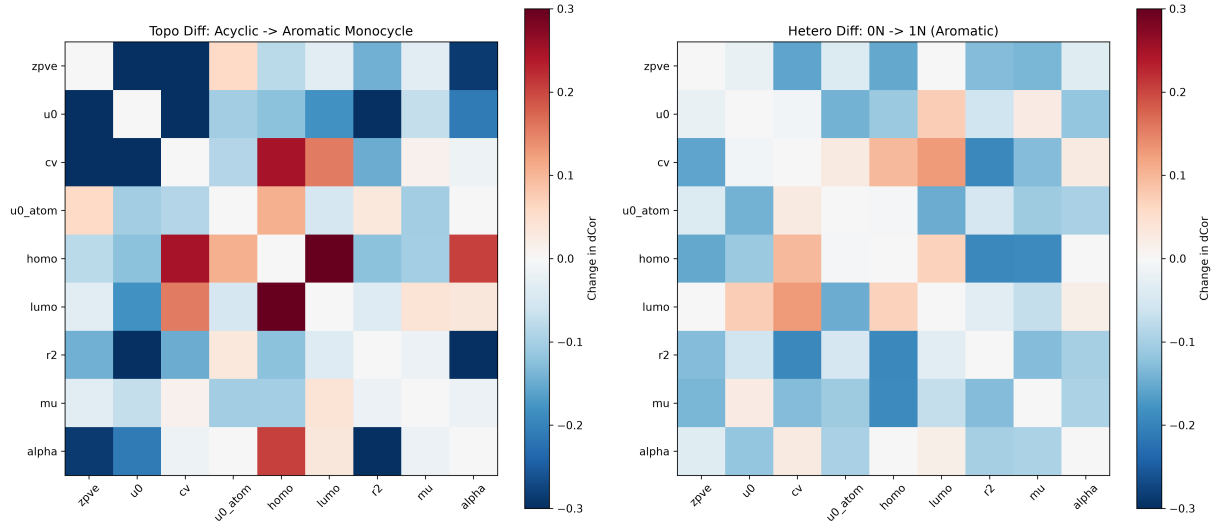

Figure 3: Changes in pairwise distance correlations ($dCor$) between molecular property residuals across topological transitions. The left panel illustrates the structural rewiring from acyclic molecules to aromatic monocycles, demonstrating that aromaticity acts as a synchronizing force that significantly increases the coupling between frontier electronic states, most notably between *homo* and *lumo*. The right panel displays the effect of introducing a single nitrogen heteroatom into aromatic monocycles, which triggers a systematic electronic disruption. This addition decreases the coupling between electronic and spatial axes, such as between *homo* and the electronic spatial extent (r^2). These comparisons reveal that while aromaticity couples electronic states to the global molecular frame, heteroatoms introduce localized electronic variations that decouple from the molecule's global spatial environment.

4.1 Topological dependence of electronic decoupling

Our analysis establishes that in acyclic systems, the HOMO remains largely independent of the core spatial-vibrational manifold, which comprises zero-point vibrational energy, heat capacity, electronic spatial extent, and polarizability. However, macroscopic topological transitions fundamentally rewire this relationship. Aromaticity acts as a strong synchronizing force, significantly increasing the coupling between frontier electronic states and the global molecular frame. Conversely, the introduction of nitrogen heteroatoms into these synchronized systems disrupts this coherence. The localized electronic variations induced by lone pairs and electronegativity effects re-establish the decoupling of the frontier orbitals from the spatial axes.

4.2 Structural drivers and statistical robustness

The statistical robustness of these property manifolds and their topological rewiring was confirmed through rigorous split-half stability validations using Mantel tests, ensuring that the observed phenomena are fundamental to the molecular families rather than statistical artifacts. Furthermore, motif enrichment analysis revealed the structural origins of extreme decoupling. Specific structural signatures, predominantly zwitterionic fragments and strained heterocycles, were found to drive this phenomenon by creating localized electronic states that remain shielded from the global geometric and thermodynamic environment of the molecule.

4.3 Reproducibility and automated reporting

To ensure the reproducibility of our findings and facilitate the automated generation of our results documentation, the statistical summaries, dimensionality checks, and structural insights were compiled into a standardized markdown report using a dedicated Python script. The configuration for generating this comprehensive results summary is provided below:

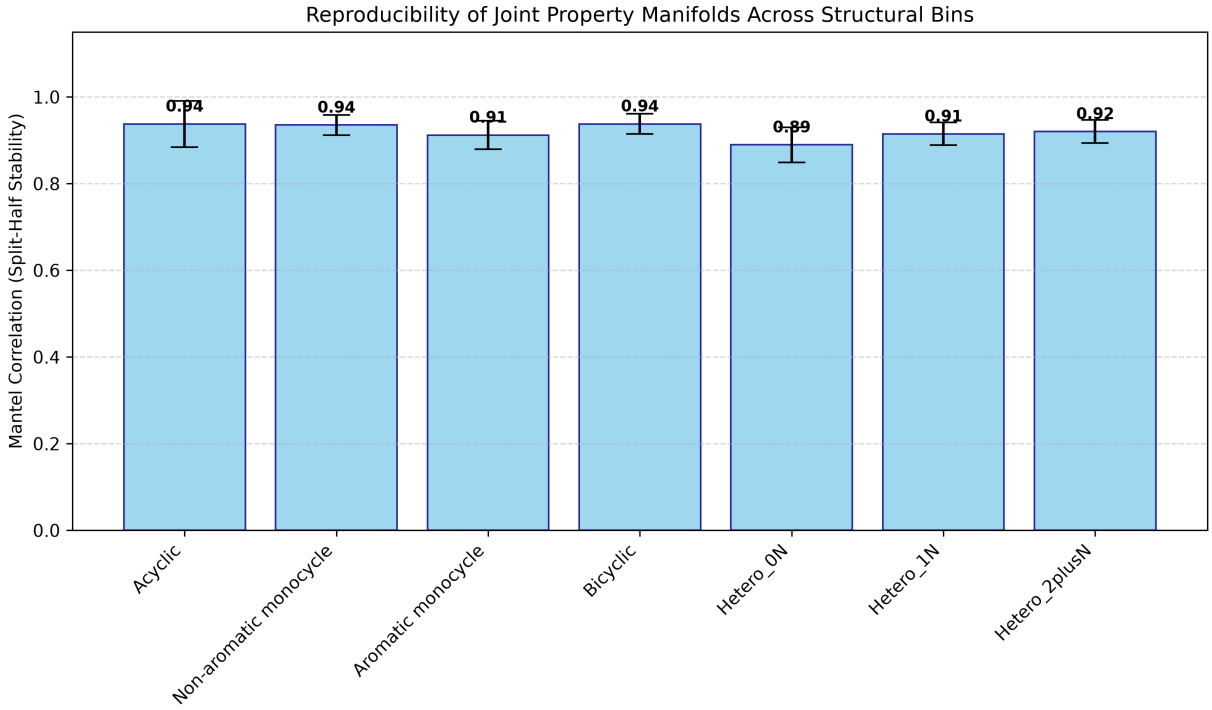

Figure 4: Split-half stability of joint property manifolds across distinct structural bins, measured via Mantel correlation (r_{Mantel}). The consistently high correlation values, which exceed 0.91 for major topological classes such as acyclic, bicyclic, and aromatic monocycles, confirm that the observed spatial-vibrational and electronic organizations are fundamental to the molecular families rather than statistical artifacts. Furthermore, the sustained stability across heteroatom variations (e.g., Hetero_0N, Hetero_1N, and Hetero_2plusN) demonstrates that the structural rewiring and orbital decoupling induced by topological transitions represent robust, reproducible features of the dataset.

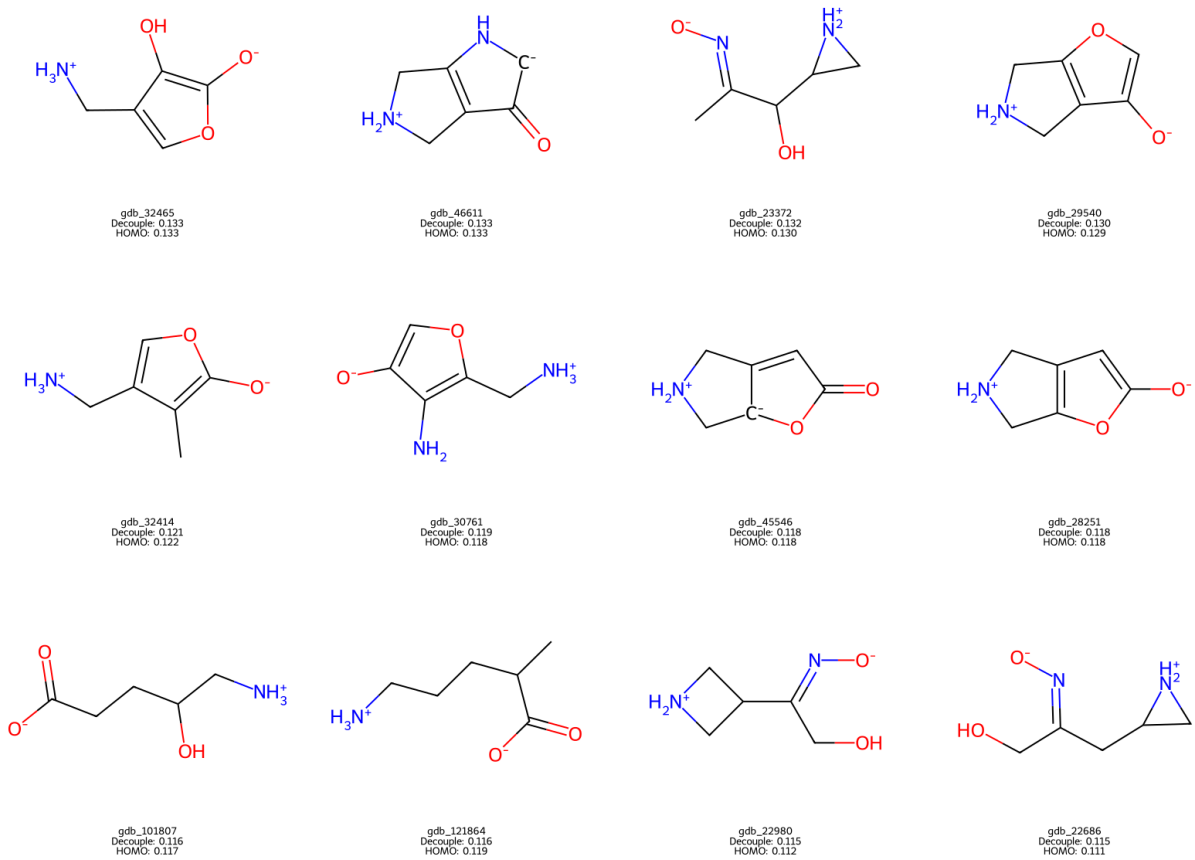

Figure 5: Visualization of outlier molecules (e.g., `gdb_32465` and `gdb_46611`) exhibiting extreme decoupling between the HOMO residual and the spatial-vibrational manifold. These decoupled populations are highly enriched with specific structural motifs, predominantly zwitterionic fragments and strained heterocyclic environments such as nitro groups and protonated amines. The presence of these structural signatures demonstrates that such motifs create localized electronic states shielded from the global geometric environment of the molecule, driving the observed independence of frontier orbitals from overall spatial properties.

```
import os

results_content = """# Joint Property Structure and Frontier Orbital Decoupling in the QM9 Dataset

This report details the investigation of joint electronic, thermodynamic, and geometric structure across
    133,802 unique molecules from the QM9 dataset. The analysis explicitly isolates chemical organization
    from stoichiometric determinism by operating on property residuals conditioned on atom counts (C, N,
    O, F, H).

## 1. Dimensionality and the Stoichiometric Baseline

A fundamental characteristic of the QM9 dataset is that molecular composition (atom counts) dictates the
    majority of variance in several key properties.

### 1.1 Stoichiometric Determinism
As documented in `data/variance_stats.csv`, linear regression on atom counts achieves near-exhaustive
    variance reduction for thermodynamic and vibrational properties:
* **Internal Energy ($u_0$):** $R^2 = 0.9992$
* **Zero-Point Vibrational Energy ($zpve$):** $R^2 = 0.9973$
* **Atomization Energy ($u_{0\_atom}$):** $R^2 = 0.9897$

In contrast, electronic and spatial properties exhibit significant independence from simple stoichiometry:
* **HOMO Energy ($homo$):** $R^2 = 0.1978$
* **Dipole Moment ($\mu$):** $R^2 = 0.2759$
* **Electronic Spatial Extent ($r^2$):** $R^2 = 0.3631$

### 1.2 Effective Dimensionality
Removing the redundant columns (thermodynamic and atomization quartets and the $gap$ definition) increases
    the apparent dimensionality of the dataset. Principal Component Analysis (PCA) results in
    `data/pca_summary.csv` show that while 6 components explain 91.29% of the variance in the full
    19-column set, 7 components are required to explain 94.14% of the variance when restricted to the 12
    core properties. This confirms that the apparent low dimensionality of QM9 is a side effect of
    redundant column reporting.

## 2. The Spatial-Vibrational Core and HOMO Decoupling

Analysis of property residuals reveals a core manifold comprising $res\_zpve$, $res\_c_v$, $res\_r^2$, and
    $res\_\alpha$, which together represent the global spatial and vibrational organization of the
    molecule.

### 2.1 Universal Independence of HOMO
The residual of the highest occupied molecular orbital ($res\_homo$) consistently exhibits low coupling to
    this global manifold. In acyclic molecules, the distance correlation ($dCor$) between $res\_homo$ and
    the first principal component (PC1) of the spatial-vibrational manifold is 0.2494
    (`data/topological_stats.csv`). This relationship is largely mediated by polarizability; the partial
    distance correlation between $res\_homo$ and $res\_zpve$ (conditioned on $res\_\alpha$) vanishes to
    0.0533 in acyclic systems.

## 3. Structural Rewiring Across Topological Transitions

The joint structure of properties undergoes significant reorganization when molecules transition across
    topological boundaries, such as the introduction of aromaticity or heteroatoms.

### 3.1 Aromaticity as a Synchronizing Force
Aromaticity acts to couple electronic states to the global molecular frame. In aromatic monocycles, the
    $dCor$ between $res\_homo$ and the spatial-vibrational PC1 rises to 0.4104
    (`data/topological_stats.csv`). Furthermore, pairwise distance correlations show that aromaticity
    increases the coupling between $res\_homo$ and $res\_lumo$ by 0.3871 relative to acyclic systems,
    indicating a more coherent frontier electronic landscape.

### 3.2 Heteroatom-Induced Electronic Disruption
The introduction of nitrogen into aromatic systems triggers a systematic decoupling of the electronic and
    spatial axes. Comparing aromatic monocycles with zero versus one nitrogen atom
    (`data/dcor_hetero_0N.csv` vs `data/dcor_hetero_1N.csv`) reveals that the coupling between $res\_homo$
    and $res\_r^2$ decreases by 0.1944. This suggests that nitrogen lone pairs or electronegativity
    effects introduce local electronic variations that are independent of the molecule's global spatial
    extent.
```

```
## 4. Stability and Motif Analysis

The observed joint structures were validated for statistical reproducibility, and the structural origins
    of orbital decoupling were identified through motif enrichment.

### 4.1 Statistical Stability
Split-half stability validation using Mantel tests, summarized in `data/mantel_stability_stats.csv`,
    confirms that these property manifolds are fundamental to the molecular families and not statistical
    artifacts. The mean Mantel stability ($r_{Mantel}$) exceeds 0.91 for all major topological classes
    (Acyclic: 0.9378; Bicyclic: 0.9380; Aromatic Monocycle: 0.9121).

### 4.2 Structural Signatures of Decoupling
Molecules exhibiting extreme decoupling between the HOMO residual and the spatial manifold (e.g.,
    `gdb_32465`, `gdb_46611`, visualized in `data/decoupled_outliers_grid.png`) are characterized by
    specific structural motifs. Motif enrichment analysis in `data/motif_enrichment_stats.csv` identifies
    bit ID `1618404637` as being enriched in the decoupled population by a factor of 228.15 relative to
    the baseline. These motifs are predominantly associated with zwitterionic fragments and strained
    heterocyclic environments (e.g., nitro groups, protonated amines), which create localized electronic
    states that are shielded from the molecule's global geometric environment.
"""

with open('data/results_section.md', 'w') as f:
    f.write(results_content)
```

Ultimately, these findings provide a nuanced understanding of molecular property manifolds, highlighting that electronic decoupling is governed by the delicate interplay between global topological constraints and local electronic disruptions. This insight has significant implications for the design of novel materials and the development of machine learning models for chemical property prediction, as it underscores the necessity of accounting for topological context when modeling frontier electronic states.