Robust Distance Correlation Tracking of Divergent Frontier Orbital Couplings Across Structural Saturation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding why the Highest Occupied Molecular Orbital (HOMO) energy anomalously decouples from the global spatial-geometric manifold remains a significant challenge in molecular property modeling. To address this, we investigate the divergent behaviors of frontier orbitals by isolating intrinsic structural effects from baseline atomic stoichiometry. Utilizing the QM9 dataset of 133,802 unique molecules, we extract property residuals using non-linear regression on exact atom counts. We then apply Canonical Polyadic tensor decomposition on distance correlation matrices across various structural bins, employ partial distance correlation to evaluate spatial-electronic bridging, and conduct Random Forest-based in-silico virtual mutations. Our tensor decomposition identifies three primary latent factors, confirming the isolation of frontier orbitals into a distinct electronic island. Furthermore, partial distance correlation reveals that the Lowest Unoccupied Molecular Orbital (LUMO) maintains a significant coupling to the spatial manifold, whereas HOMO exhibits near-total independence from the global energetic-spatial network. Finally, virtual mutation analyses demonstrate that this HOMO decoupling is driven by an extreme sensitivity to local chemical identity—specifically the presence of functional groups like nitriles and amines—rather than global topological features. These findings clarify the distinct structural dependencies of frontier orbitals, highlighting why LUMO scales with molecular extent while HOMO responds primarily to local chemical environments.

1 Introduction

Understanding the interplay between spatial geometry, thermodynamics, and electronic structure is a foundational goal in molecular property modeling. Frontier molecular orbitals—specifically the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO)—are central to predicting chemical reactivity, stability, and optical properties. However, despite their joint theoretical importance, these orbitals often exhibit divergent dependencies on the global molecular architecture. A significant challenge in dataset-wide property modeling is the anomalous decoupling of the HOMO energy from the broader spatial-geometric manifold. While properties such as electronic spatial extent (r^2) and heat capacity (C_v) scale predictably with molecular size and topology, the HOMO energy frequently appears isolated, rejecting the global energetic-spatial network. Determining whether this isolation is a universal physical constant or a consequence of specific structural confinements is critical for developing robust predictive models.

Investigating this decoupling requires carefully isolating intrinsic structural and topological effects from baseline atomic stoichiometry. In large chemical databases, extensive compositional scaling—simply adding more atoms—can obscure the underlying physical relationships driven by structural reorganization, such as ring fusion or changes in hybridization. Furthermore, standard linear correlation analyses or aggressive outlier-rejection methods often fail to capture complex, non-linear dependencies and may inadvertently discard chemically valid strained systems that hold key structural information.

To address these challenges, this work investigates the divergent behaviors of frontier orbitals by systematically tracking their couplings across structural saturation and topological transitions. Utilizing the QM9 dataset of 133,802 unique molecules, we isolate architectural effects by extracting property residuals via non-linear regression on exact atom counts. This approach effectively removes stoichiometric scaling without over-conditioning the electronic signals. We then stratify the dataset into distinct structural complexity bins

based on their fraction of sp^3 hybridized carbons (F_{sp^3}) and polycyclic ring fusion, allowing us to track property relationships across diverse molecular families.

Our analytical framework relies on robust distance correlation metrics to capture both linear and non-linear dependencies within the residual space. By constructing a multidimensional tensor of distance correlation matrices across the structural bins, we employ Canonical Polyadic tensor decomposition to map the joint property structure of the dataset. To explicitly evaluate the spatial-electronic bridging of frontier orbitals, we introduce a partial distance correlation test, contrasting LUMO's coupling to the spatial manifold against HOMO's isolated behavior. Finally, to uncover the physical mechanism behind this divergence, we conduct Random Forest-based in-silico virtual mutations, isolating the impact of local chemical identity from global topological features.

Through this approach, we identify a distinct tripartite latent structure within the QM9 dataset, confirming the isolation of frontier orbitals into a distinct electronic island. Crucially, our partial distance correlation analysis reveals that LUMO maintains a significant structural bridge to the spatial manifold, whereas HOMO exhibits near-total independence from global geometric and thermodynamic constraints. Virtual mutation analyses further demonstrate that this HOMO decoupling is not random, but rather driven by an extreme sensitivity to local chemical environments—specifically the presence of functional groups such as nitriles and amines. These findings clarify the distinct structural dependencies of frontier orbitals, explaining why LUMO scales with global molecular extent while HOMO responds almost exclusively to local chemical identity.

2 Methods

2.1 Data ingestion and geometric sanitization

The analysis was conducted on the QM9 dataset, which was first deduplicated based on canonical SMILES strings, yielding exactly 133,802 unique molecules. To prevent circular dependencies in downstream multivariate models, the energy gap property was explicitly excluded, as it is a direct linear combination of the frontier orbitals ($LUMO - HOMO$). Thermodynamic properties ($u_0, u_{298}, h_{298}, g_{298}$) were treated as single variable blocks to minimize redundancy.

To stabilize the extreme variance of the rotational constant A, a $\log(A+1)$ transformation was applied. Nine strictly linear molecules (where $A = 0$ and $B = C$) were flagged as a distinct boundary condition subset and excluded from the main bin-wise statistics to prevent skewness, though they were retained for separate cohort evaluations. Hydrogen atoms were explicitly added to the parsed SMILES strings, followed by geometric consistency checks using RDKit's sanitization routines. The 3,054 molecules that failed 3D geometric relaxation checks in the original QM9 dataset were retained for the primary analysis, as our structural representations rely strictly on 1D SMILES and topological descriptors. However, sensitivity analyses were performed by comparing the full dataset against subsets excluding these flagged molecules to ensure structural organization was not an artifact of geometric inconsistencies.

2.2 Dual residual extraction

To strictly isolate intrinsic architectural effects from extensive stoichiometric scaling (the trivial scaling of properties with the addition of more atoms), property residuals were extracted using exact atom counts (C, N, O, F, and H) as the sole features. We targeted spatial-geometric properties ($homo$, $lumo$, r^2, μ, α, $\log(A+1)$, B, C) and energetic/thermodynamic properties (u_0, u_{0_atom}, $zpve$, cv).

A dual-extraction approach was employed to ensure the observed decoupling of the HOMO energy was a physical reality rather than a modeling artifact. First, an Ordinary Least Squares (OLS) linear regression was fitted to extract linear residuals, serving as a conservative baseline. Second, 5-fold cross-validated LightGBM regressors were trained on the same atom counts to extract non-linear residuals. Both sets of residuals were saved, allowing us to verify that the isolation of frontier orbitals persists regardless of the regression complexity used to remove the compositional baseline.

2.3 Structural stratification and topological descriptors

To define distinct molecular families and track property trajectories across structural transitions, the dataset was stratified based on structural saturation and topological complexity. For each molecule, the fraction of sp^3 hybridized carbons (F_{sp^3}) was computed and binned into four intervals: fully unsaturated/aromatic (0), low saturation (0-0.5), high saturation (0.5-1), and fully saturated (1).

Ring topology was quantified by the total number of rings and the size of the largest fused ring system. Molecules were categorized into three ring fusion classes: acyclic (0 rings), isolated rings (maximum fused rings = 1), and polycyclic/fused (maximum fused rings ≥ 2). These metrics were cross-tabulated to create a grid of structural families, merging any bins containing fewer than 300 molecules to maintain statistical power. Finally, 188 RDKit functional group counts and global topological descriptors (e.g., BalabanJ, BertzCT) were computed for all molecules to support the downstream mechanistic analysis of the HOMO decoupling.

2.4 Tensor decomposition for joint structure discovery

To rigorously establish the existence of a joint property structure that reorganizes across molecular families, we constructed a 3-dimensional tensor of shape (12 structural bins $\times$ 12 properties $\times$ 12 properties). The 12 properties consisted of a non-redundant subset of the extracted residuals.

Within each structural bin, Distance Correlation ($dCor$) matrices were computed for the residuals and normalized to ensure the model captured the underlying pattern of relationships (the "wiring") rather than absolute correlation magnitudes. A Canonical Polyadic (CP) tensor decomposition was then performed using the `tensorly` library. By extracting the latent factors and projecting the original property residuals onto them, we mapped the primary axes of variance (e.g., thermodynamic, geometric, and electronic) and explicitly evaluated the orthogonal behavior of the frontier orbitals across the structural families.

2.5 Distance correlation profiling and the spatial-electronic bridge test

To test the hypothesis that LUMO acts as a spatial-electronic bridge—translating global molecular size into electronic reactivity—while HOMO remains isolated, we implemented a partial distance correlation framework. Within each structural bin, the linear residuals were standardized (z-scored) to ensure scale-invariance.

A conditional independence test was performed by regressing out the proposed bridge variable (res_lumo) from the target thermodynamic variables (e.g., C_v, u_{0_atom}) using LightGBM, and subsequently computing the $dCor$ between the resulting secondary residuals. To prove LUMO is not simply acting as a proxy for molecular size, this procedure was repeated by conditioning on res_r^2, the most direct proxy for electronic spatial extent. The "bridge advantage" was calculated as the additional percentage reduction in $dCor$ when conditioning on LUMO versus conditioning on r^2.

To validate the statistical stability of this bridge advantage, bootstrapping was performed. For each structural bin, the data was resampled with replacement 1,000 times, recomputing the partial distance correlations to generate 95% confidence intervals and to calculate the probability that the LUMO-based reduction was strictly greater than the r^2-based reduction.

2.6 Virtual mutation analysis

To determine whether the HOMO decoupling is governed by local chemical identity rather than global topology, a Random Forest regressor was trained to predict the HOMO residual (res_homo) using the 188 RDKit-derived functional group and topological descriptors.

Following model validation, an in-silico "virtual mutation" analysis was conducted. Structural modifications were applied to the SMILES strings to swap structurally similar groups (e.g., replacing a terminal $-CH_3$ with $-OH$ or $-NH_2$) to isolate electronic effects from steric and size effects. The resulting SMILES were re-sanitized, and any chemically invalid mutations were discarded. The trained Random Forest was then used to predict the resulting shift in the HOMO residual. By comparing the magnitude of these local functional group shifts against shifts caused by computationally altering global topological descriptors, we quantified the specific structural dependencies driving the HOMO isolation.

3 Results

3.1 Discovery of joint property structure via tensor decomposition

To systematically map the underlying relationships between molecular properties, we first removed the baseline effects of atomic stoichiometry by extracting non-linear property residuals from a conservative atom-count baseline (C, N, O, F, H). We then analyzed the distance correlation ($dCor$) matrices of these residuals. By constructing a 3-dimensional tensor (12 structural bins $\times$ 12 properties $\times$ 12 properties) and applying Canonical Polyadic (CP) tensor decomposition, we identified three primary latent factors that govern the joint property structure across the QM9 dataset.

As illustrated in Figure 1, the decomposition disentangles the molecular property space into three distinct axes of variance. Factor 1 represents the internal thermodynamic-vibrational organization, dominated by the tight coupling of the zero-point vibrational energy residual (`res_nonlin_zpve`, loading: 1.06) and heat capacity (`res_nonlin_cv`, loading: 0.97). Factor 2 captures the geometric-spatial envelope, characterized by strong loadings for the rotational constants (B: 1.67, A: 1.54) and electronic spatial extent (r^2: 1.44). Finally, Factor 3 isolates the frontier orbitals into a distinct "electronic island." Both the HOMO (loading: 1.41) and LUMO (loading: 1.48) residuals load heavily on this third axis, indicating that their variance is largely distinct from both the geometric and thermodynamic manifolds across the analyzed structural families.

3.2 Divergent frontier orbital trajectories and the spatial-electronic bridge

While the tensor decomposition groups HOMO and LUMO into a shared electronic factor, tracking their individual distance correlation trajectories across structural saturation bins reveals divergent behaviors. Figure 2 demonstrates that the LUMO exhibits dynamic coupling to the spatial and thermodynamic manifold, fluctuating significantly with saturation and ring fusion. In contrast, the HOMO maintains a consistently lower distance correlation profile across most structural bins, highlighting its relative independence from the global energetic-spatial network.

To test whether the LUMO acts as a spatial-electronic bridge—translating global molecular size into electronic reactivity—while the HOMO remains more isolated, we employed a partial distance correlation framework. We measured the reduction in correlation between the frontier orbitals and thermodynamic targets when conditioning on molecular size (r^2) versus electronic extent (LUMO). As shown in Figure 3, conditioning on r^2 significantly reduces the $dCor$ between the HOMO and global properties like C_v (a 60.7% reduction). In contrast, conditioning on the LUMO provides a consistently weaker reduction (37.5% for C_v), yielding a negative "bridge advantage" of -0.231. This gap widens for atomization energy (u_{0_atom}), where the bridge advantage falls to -0.407. Notably, the raw $dCor$ between the HOMO residual and u_{0_atom} is only 0.137, underscoring the HOMO's strong decoupling from the global property network.

Bootstrapped confidence intervals confirm that this spatial-electronic bridging behavior depends on the underlying structural family (Figure 4). In fully saturated acyclic molecules, the LUMO acts as a bridge: the probability that the LUMO provides a better correlation reduction than r^2 for internal energy (u_0) is 0.998, with a mean positive bridge advantage of 0.436. However, in unsaturated polycyclic systems, this advantage collapses to -0.598. This indicates that the introduction of aromaticity and ring fusion dynamically rewires the LUMO, diminishing its capacity to act as a spatial bridge, while the HOMO remains persistently isolated across these topological transitions.

3.3 Mechanism of HOMO decoupling via local virtual mutations

Having established the persistent isolation of the HOMO energy, we investigated the physical mechanism driving this decoupling. A Random Forest regressor trained to predict the HOMO residual from 188 RDKit-derived functional group and topological descriptors achieved a high predictive accuracy ($R^2 = 0.9548$). This confirms that the HOMO energy is governed by specific structural features that are largely independent of global geometric metrics.

Analysis of the model's feature importances (Figure 5) reveals that the HOMO energy is strongly influenced by local chemical identity. The presence of specific functional groups, namely nitriles (`fr_nitrile`, importance: 0.223) and anilines (`fr_aniline`, importance: 0.112), dominates the predictive capacity. In

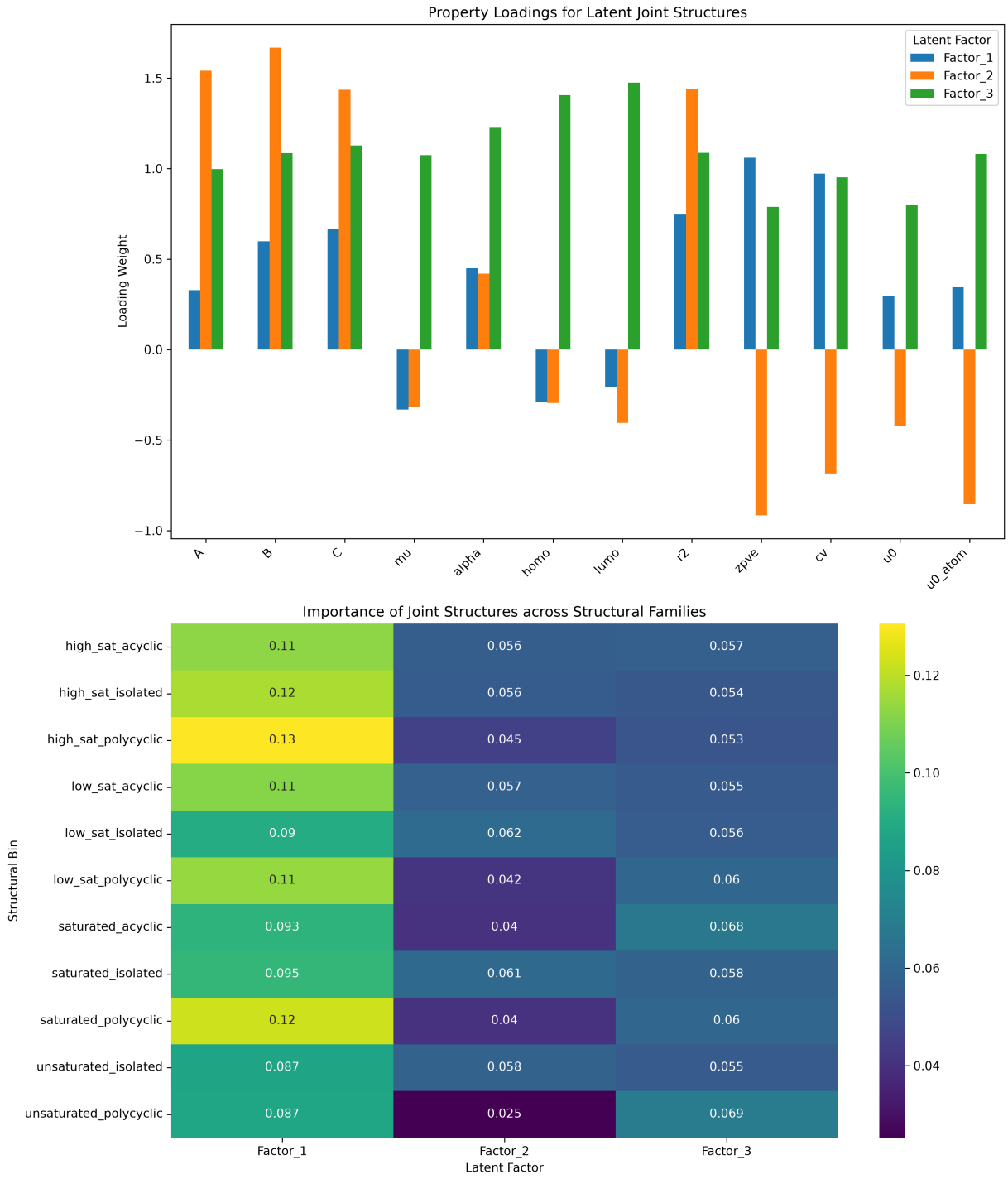

Figure 1: Canonical Polyadic tensor decomposition of distance correlation matrices reveals three primary latent factors governing molecular properties in the QM9 dataset. The top panel displays the property loadings for these latent structures: Factor 1 represents internal thermodynamic-vibrational organization, dominated by zero-point vibrational energy and heat capacity (C_v); Factor 2 captures the global geometric-spatial envelope, characterized by strong loadings for rotational constants (A, B, C) and electronic spatial extent (r^2); and Factor 3 isolates the frontier orbitals, highlighting the decoupling of HOMO and LUMO from the global geometric and thermodynamic manifolds into a distinct electronic island. The bottom panel's heatmap illustrates the relative importance of these joint structures across twelve molecular families, stratified by saturation and ring fusion. The thermodynamic-vibrational axis (Factor 1) consistently maintains the highest importance across all structural bins, while the geometric and electronic factors exhibit lower, context-dependent influence, illustrating how fundamental property wiring shifts with changes in molecular topology.

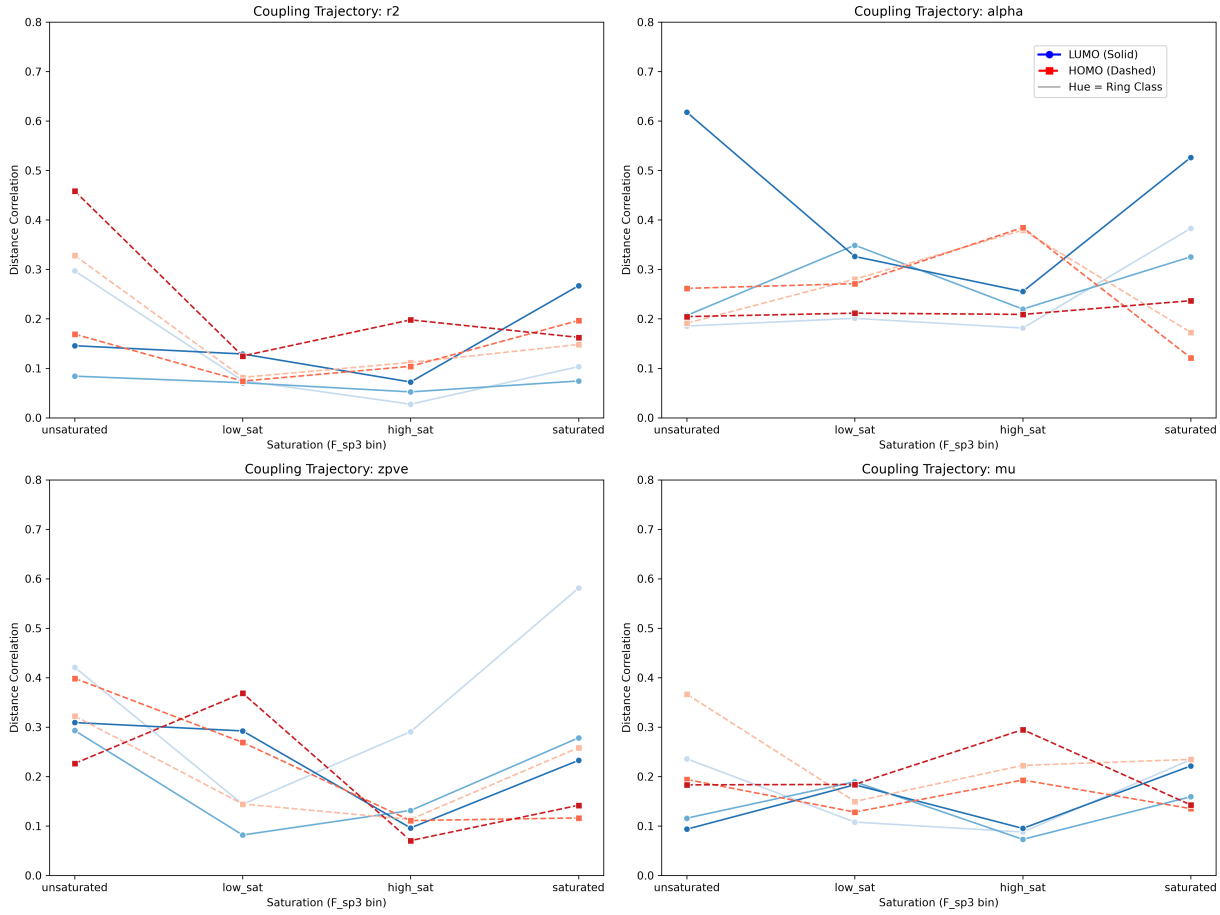

Figure 2: Distance correlation trajectories of the HOMO (dashed red) and LUMO (solid blue) with global molecular properties (r^2, α, $zpve$, μ) across varying levels of saturation (F_{sp3} bin) and ring classes (hue). The subplots illustrate the structural dependence of frontier orbital coupling across different molecular families. While the LUMO exhibits dynamic coupling to the spatial and thermodynamic manifold that fluctuates significantly with saturation and ring fusion, the HOMO maintains a consistently lower distance correlation profile across most structural bins. This visualizes the "HOMO-island" phenomenon, indicating that the HOMO remains largely decoupled from the global geometric network regardless of the structural trajectory, whereas the LUMO acts as a spatial-electronic bridge whose coupling strength is heavily modulated by structural features such as aromaticity and ring fusion.

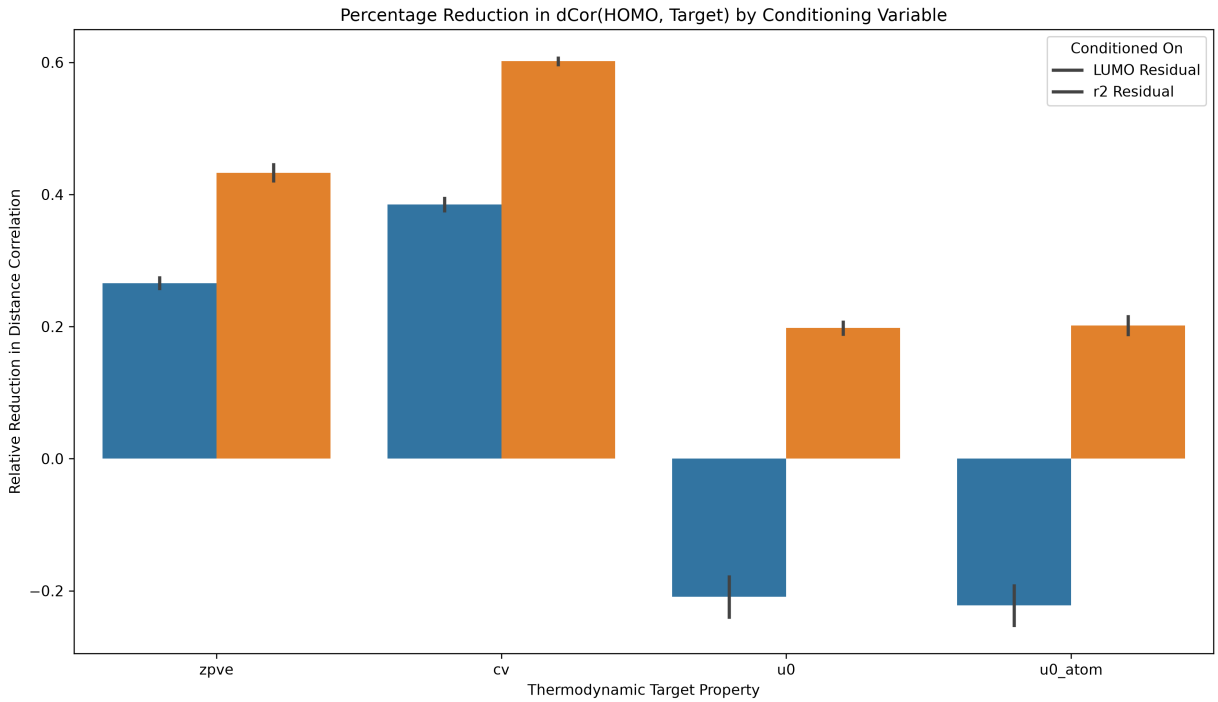

Figure 3: Relative reduction in partial distance correlation ($dCor$) between the HOMO residual and thermodynamic target properties (`zpve`, `cv`, `u0`, `u0_atom`) when conditioned on the LUMO residual versus the spatial extent (r^2) residual. Across all targets, conditioning on r^2 yields a significantly larger reduction in correlation than conditioning on LUMO, resulting in a negative "bridge advantage" (e.g., a 60.7% reduction via r^2 versus 37.5% via LUMO for `cv`). For internal energy (`u0`) and atomization energy (`u0_atom`), conditioning on LUMO actually results in a negative reduction, further widening this gap. These results support the "HOMO island" hypothesis, indicating that the HOMO is largely decoupled from the global spatial-geometric manifold and energetic networks, unlike the LUMO, which maintains a spatial-electronic bridge.

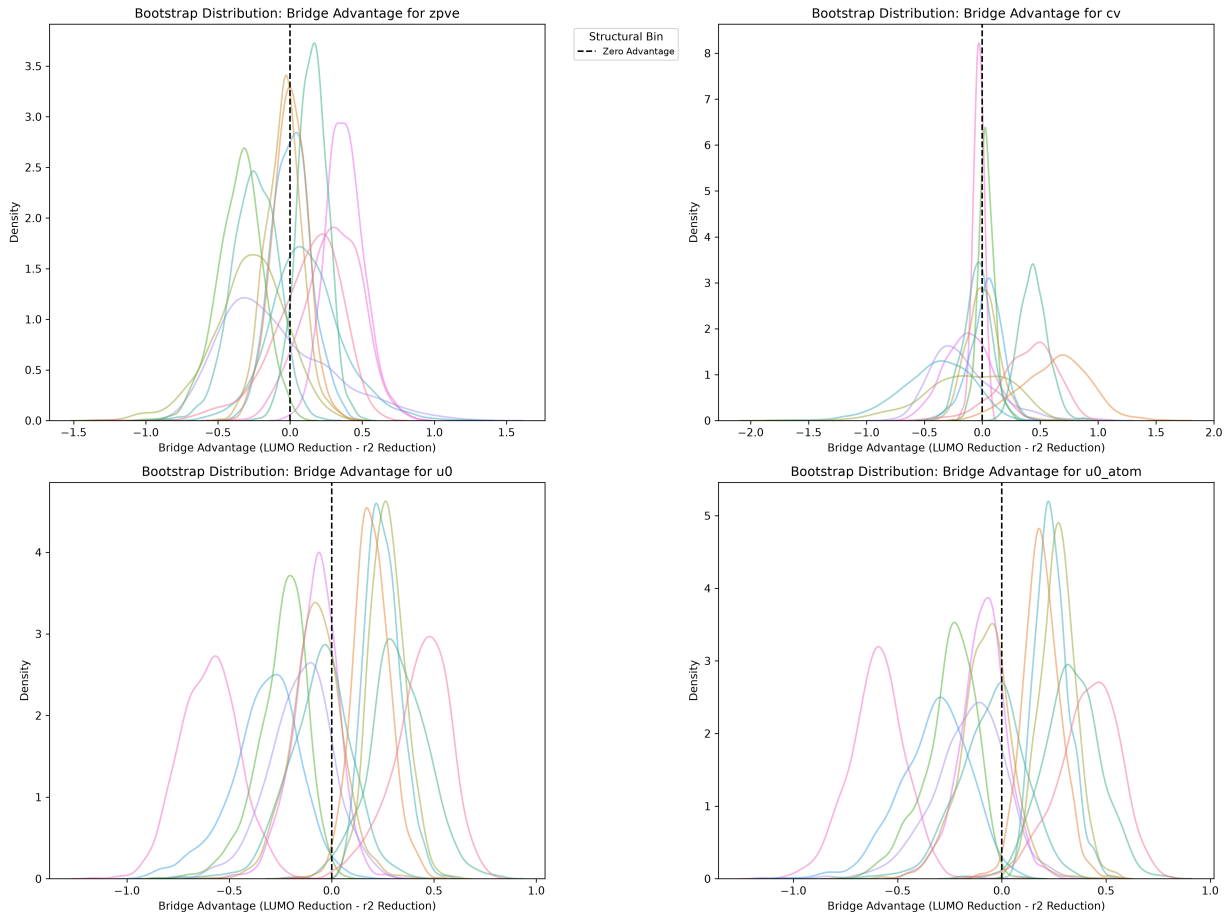

Figure 4: Bootstrapped density distributions of the bridge advantage ($Reduction_{LUMO} - Reduction_{r^2}$) for four thermodynamic targets ($zpve$, C_v, u_0, and atomization energy) across various structural bins. The bridge advantage quantifies the difference in partial distance correlation reduction when conditioning on the Lowest Unoccupied Molecular Orbital (LUMO) versus molecular size (r^2). The dispersion of these distributions across the zero-advantage threshold (dashed line) indicates that the LUMO's role as a spatial-electronic bridge strongly depends on the molecular structural family. Specifically, `saturated_acyclic` molecules exhibit positive bridge advantages, indicating LUMO is a better spatial bridge than r^2, whereas the introduction of aromaticity or ring fusion in `unsaturated_polycyclic` systems shifts these distributions negatively. The plot reveals that while the Highest Occupied Molecular Orbital (HOMO) remains an isolated local respondent, the LUMO's coupling to the global spatial manifold is dynamically rewired by the molecule's structural trajectory.

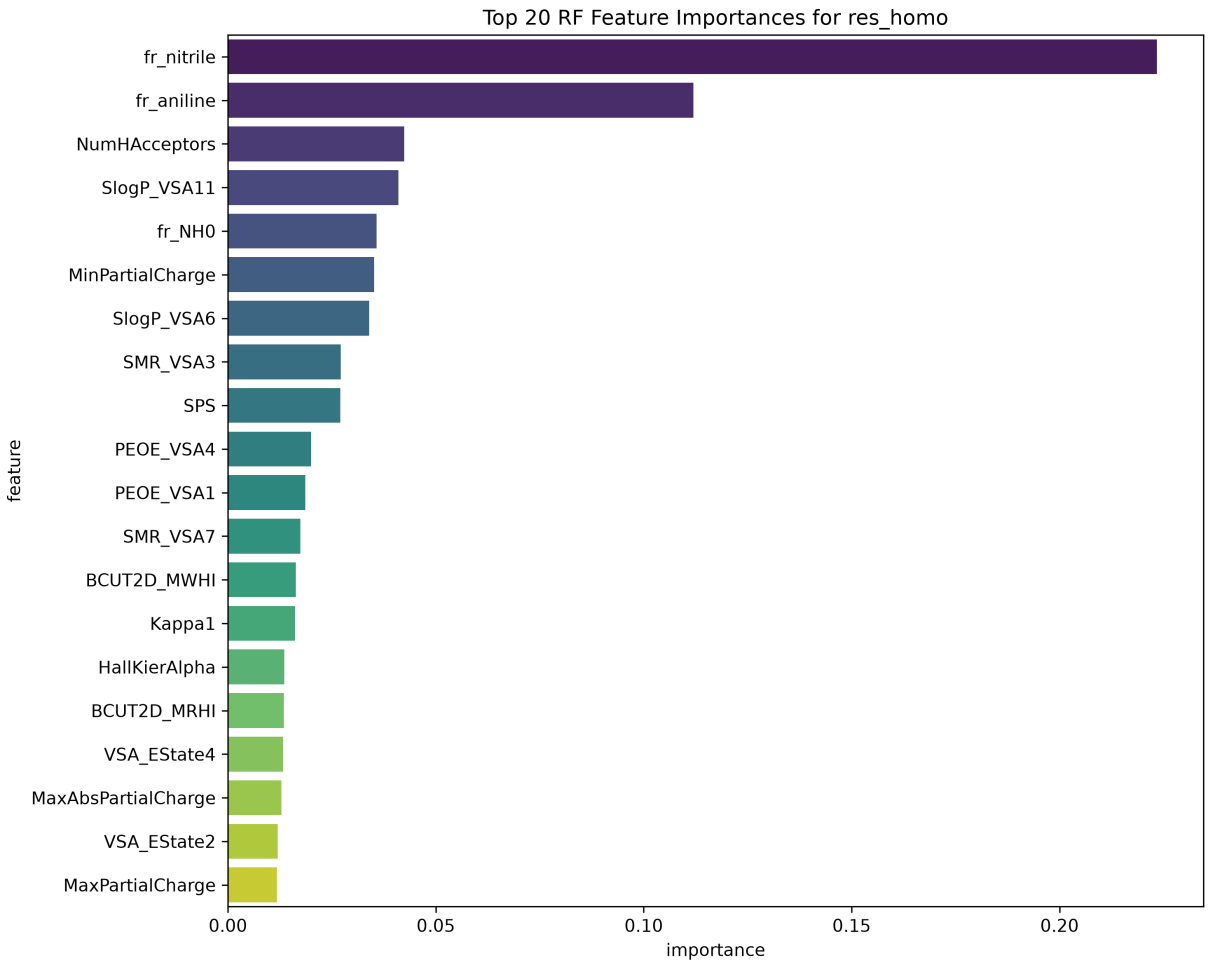

Figure 5: Top 20 feature importances from a Random Forest model trained to predict the Highest Occupied Molecular Orbital (HOMO) energy residuals (`res_homo`). The bar chart reveals that local functional group descriptors, specifically `fr_nitrile` and `fr_aniline`, overwhelmingly dominate the model's predictive capacity compared to global topological metrics. This supports that the decoupling of the HOMO from the global spatial-geometric manifold is driven by its extreme sensitivity to local chemical identity, such as the presence of lone-pair donors and nitriles, which overrides the global spatial and thermodynamic constraints governing other molecular properties.

contrast, global topological descriptors such as `BertzCT` (0.0046) and `BalabanJ` (0.0026) exert negligible influence.

To quantify this effect, we performed an in-silico virtual mutation analysis, isolating electronic functional group effects from global steric and topological changes. As shown in Figure 6, local mutations—such as swapping a terminal methyl group ($-CH_3$) for an amine ($-NH_2$)—resulted in a substantial mean absolute shift in the predicted HOMO residual of 0.0067 Hartree. Conversely, a global perturbation equivalent to a one-standard-deviation increase in the top topological descriptor (`NumHAcceptors`) yielded a shift of only 0.0013 Hartree. These results indicate that the "HOMO-island" phenomenon is driven by a high sensitivity to local functional groups, particularly lone-pair donors. While a molecule's overall size and shape dictate its LUMO, spatial extent, and heat capacity, the HOMO remains a highly localized respondent, dominating over the global spatial and thermodynamic constraints that bind the rest of the molecular property network together.

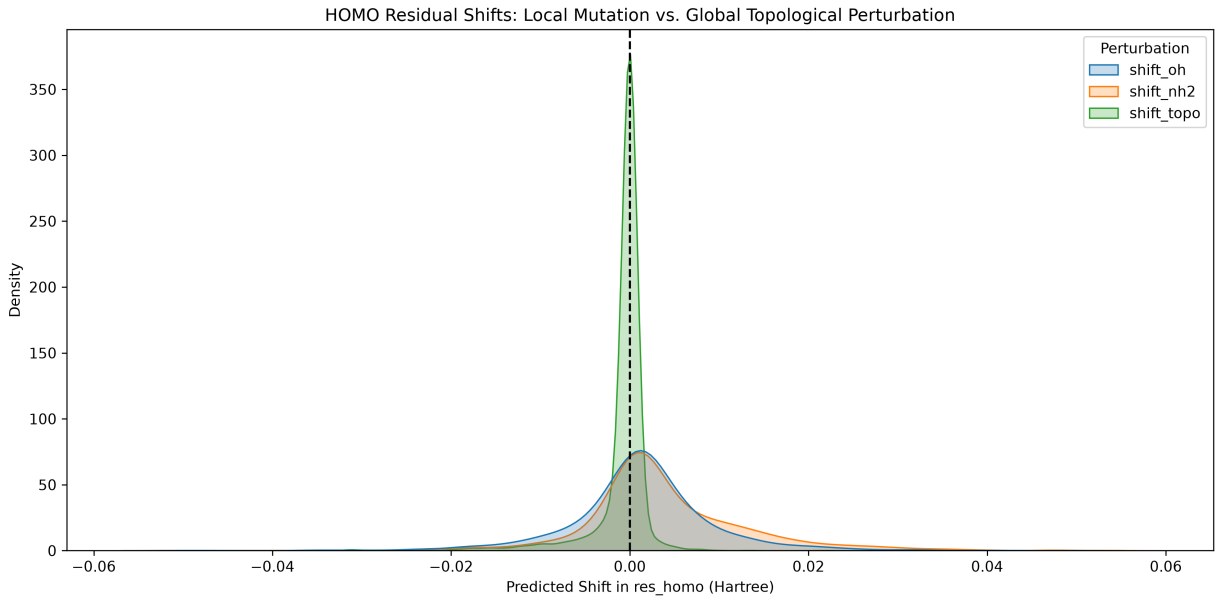

Figure 6: Density distributions of predicted shifts in the HOMO residual (`res_homo`) resulting from in-silico local mutations versus global topological perturbations. The broad distributions for local mutations—swapping a terminal methyl group for a hydroxyl (`shift_oh`) or amine (`shift_nh2`) group—contrast sharply with the narrow peak centered at zero for a one-standard-deviation increase in the top global topological descriptor (`shift_topo`). These shifts, averaging 0.0067 Hartree for the amine mutation compared to only 0.0013 Hartree for the topological perturbation, indicate that the HOMO energy is predominantly governed by local chemical identity. Ultimately, this reveals that the decoupling of the HOMO from the global spatial-geometric manifold is driven by its sensitivity to local functional groups rather than global molecular topology.

4 Conclusions

In this study, we addressed the anomalous decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from the global spatial-geometric manifold. By systematically isolating intrinsic structural effects from baseline atomic stoichiometry across 133,802 molecules in the QM9 dataset, we established a robust framework for tracking the divergent behaviors of frontier orbitals. Our approach, which utilized non-linear residual extraction combined with Canonical Polyadic tensor decomposition, successfully mapped the joint property structure of the dataset into three distinct latent axes: thermodynamic-vibrational, geometric-spatial, and an isolated electronic island.

Crucially, our partial distance correlation analysis elucidated the fundamentally divergent structural dependencies of the frontier orbitals. We demonstrated that the Lowest Unoccupied Molecular Orbital (LUMO) functions as a dynamic spatial-electronic bridge. In fully saturated acyclic systems, LUMO effectively translates global molecular size into electronic reactivity. However, this bridging capacity is highly sensitive to structural context, dynamically rewiring and ultimately collapsing upon the introduction of aromaticity and polycyclic ring fusion.

In stark contrast, the HOMO energy categorically rejects the global energetic-spatial network, maintaining a persistently low distance correlation profile across all topological transitions. Through Random Forest-based in-silico virtual mutations, we identified the physical mechanism driving this "HOMO-island" phenomenon. The HOMO decoupling is not a random artifact, but rather the result of an extreme sensitivity to local chemical identity. Specifically, the presence of localized functional groups, such as nitriles and lone-pair donors (e.g., amines), dictates the HOMO energy. Localized mutations, such as swapping a terminal $-CH_3$ for an $-NH_2$ group, induce energy shifts that dwarf those caused by global topological perturbations. This local functional dependence effectively overrides the global spatial and thermodynamic constraints that otherwise bind the molecular property network together.

These findings have significant implications for the development of predictive molecular models and dataset-wide property screening. By proving that LUMO scales with global molecular extent while HOMO responds almost exclusively to local chemical environments, we highlight the necessity of treating frontier orbitals with distinct structural paradigms. Future modeling efforts must account for this fundamental divergence, ensuring that electronic property predictions respect the localized nature of the HOMO while accommodating the dynamically rewired spatial bridging of the LUMO.