Resolving the HOMO Anomaly: Local Topological Coupling and Heteroatom Rewiring via Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energy of the highest occupied molecular orbital often appears decoupled from global thermodynamic and spatial molecular properties, presenting an anomaly in structure-property relationships. To determine whether this global orthogonality masks localized electronic rules, we investigate the conditional dependence of this orbital energy on local topological environments rather than global geometry. Using the QM9 dataset of over 130,000 molecules, we extract local topological descriptors and partition the chemical space into mutually exclusive families based on heteroatom identity. We employ distance correlation metrics and non-linear machine learning models to analyze how specific heteroatoms influence the relationships between local structural motifs and orbital energy. The results present the calculated distance correlations for various topological descriptors across the different heteroatom families and report the comparative predictive performance of the trained models. Furthermore, the findings include an assessment of feature stability and the generalization capabilities of these localized models on unseen Murcko scaffolds.

1 Introduction

The energy of the highest occupied molecular orbital (*homo*) is a fundamental quantum mechanical property that dictates chemical reactivity, ionization potential, and overall molecular stability. However, across large and diverse chemical spaces, such as the QM9 dataset of over 130,000 small organic molecules, the *homo* energy exhibits a striking anomaly: it appears largely decoupled from the global thermodynamic and spatial manifolds. While properties like rotational constants and electronic spatial extent form tightly coupled joint structures, *homo* remains largely orthogonal to these global axes. Furthermore, basic stoichiometric variables, such as exact atom counts, explain only 8.5% of the *homo* variance. This apparent global independence presents a significant challenge in structure-property mapping, leaving the primary structural drivers of the highest occupied orbital largely unexplained by standard global descriptors.

1.1 The hypothesis of localized masking

We hypothesize that this global orthogonality is an artifact of aggregating diverse chemical environments, which effectively masks highly structured, localized electronic rules. Rather than being dictated by global geometry or bulk spatial extent, the *homo* energy is likely conditionally dependent on local topological environments, such as specific conjugation lengths, ring structures, and heteroatom neighborhoods. Because different heteroatoms introduce distinct lone-pair interactions, electronegativities, and inductive effects, aggregating them into a single global population averages out their specific, localized contributions. Consequently, molecular structure may not universally predict *homo* in a static manner; instead, the presence of specific heteroatoms may fundamentally "rewire" the relationship between local topological motifs and orbital energy.

1.2 Methodological approach

To resolve this anomaly, we investigate the conditional dependence of *homo* on local topological descriptors across partitioned chemical spaces. Using the deduplicated QM9 dataset, we first regress out the stoichio-

metric contributions to isolate the residual orbital variance (*res_homo*). We then partition the dataset into mutually exclusive structural families based on heteroatom identity—specifically pure hydrocarbons, nitrogen-only, oxygen-only, fluorine-only, and mixed heteroatom systems. This partitioning prevents the averaging out of localized electronic effects and allows us to study the orbital energy within specific chemical regimes.

Within these families, we employ distance correlation metrics to rigorously quantify the non-linear coupling between *res_homo* and local structural motifs, contrasting this directly against its known orthogonality to global thermodynamic and spatial properties. To characterize the nature of these local interactions, we train a hierarchy of machine learning models, ranging from sparse linear regressions (Lasso) to non-linear tree-based algorithms (LightGBM). This comparative modeling approach allows us to determine whether the governing rules of *homo* are driven by simple additive inductive effects or complex, high-order mesomeric interactions.

1.3 Scope of the study

This study formally identifies the localized chemical rules that govern *homo*, transforming an unexplained global null result into a mapped local structure. By tracking how the primary drivers of orbital variance shift across different families, we demonstrate that heteroatom identity dictates the relevance of specific topological features, such as carbon hybridization and atomic polarizability. Finally, we assess the robustness of these localized rules through rigorous split-half stability tests and structural hold-out evaluations on unseen Murcko scaffolds, confirming that the discovered topological dependencies generalize across diverse molecular backbones.

2 Methods

2.1 Data preparation and stoichiometric decoupling

To isolate the localized electronic effects governing the highest occupied molecular orbital (*homo*), we first removed the variance attributable to basic molecular size and composition. The QM9 dataset was loaded and rigorously deduplicated, removing 83 SMILES strings that appeared multiple times to prevent data leakage during downstream validation.

Because stoichiometric variables account for a baseline fraction of orbital variance, we decoupled the target properties from exact atom counts. Using RDKit, we parsed the SMILES strings to extract the exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. We then employed Ridge regression with polynomial features to predict all 19 numeric properties in the QM9 dataset solely from these five atom counts. The residuals from these regressions, denoted with the prefix *res_* (e.g., *res_homo*, *res_zpve*, *res_cv*, *res_r*^{2}), represent the size-independent, structure-driven variance of each property.

Additionally, we performed a symmetry and geometry check to address anomalies in the rotational constants. The rotational constant A was log-transformed for all non-zero values, while exact zeros were preserved. The nine perfectly linear molecules in the dataset ($A = 0$) were isolated to determine if orbital symmetry constraints form a distinct electronic regime. A non-parametric Mann-Whitney U test was utilized to quantitatively assess whether the *res_homo* values of these linear molecules significantly deviated from the non-linear population.

2.2 Extraction of topological descriptors and family partitioning

To capture the local electronic environments hypothesized to drive *res_homo*, we extracted a comprehensive suite of topological descriptors using RDKit. These included explicit local features such as maximum conjugation length, the number of π electrons, ring counts, and electronegativity-weighted features (e.g., BCUT2D descriptors). To model distance-dependent electronic effects, we computed topological torsion fingerprints and atom-pair fingerprints, which specifically encode the topological distances between heteroatoms and conjugated π-systems. Furthermore, SMARTS patterns were applied to identify and flag the presence of specific functional groups, such as amines, ethers, nitriles, and fluorinated carbons. Feature extraction was parallelized across 14 CPUs to handle the dataset efficiently.

Following feature extraction, the dataset was partitioned into mutually exclusive structural families based on heteroatom composition. This partitioning is central to our hypothesis, ensuring that localized lone-pair effects and inductive rewiring are not averaged out across disparate chemical spaces. The defined families were: pure hydrocarbons (C, H), nitrogen-only systems (C, H, N), oxygen-only systems (C, H, O), fluorine-only systems (C, H, F), and mixed heteroatom systems (e.g., N+O).

2.3 Distance correlation and global orthogonality analysis

To formally establish the decoupling of *res_homo* from the global thermodynamic and spatial manifolds, we utilized Distance Correlation ($dCor$), a measure capable of detecting both linear and non-linear dependencies. Within each heteroatom family, we computed the $dCor$ between *res_homo* and the global residual triad ($res_zpve, res_cv, res_r^2$). We similarly computed the $dCor$ between *res_homo* and the geometric rotational residuals (res_A, res_B, res_C) to test if global geometry dictates orbital energy in any specific chemical regime. Because the distance correlation algorithm scales quadratically ($O(N^2)$), we subsampled each family to a representative fixed size of 5,000 to 10,000 molecules and parallelized the computations.

2.4 Modeling local rules and conditional rewiring

To characterize the nature of the local interactions governing *res_homo*, we trained a hierarchy of machine learning models for each family using identical input features. The models included: 1. A Null Model (mean predictor) to establish a baseline. 2. A sparse linear model (Lasso regression) to capture simple additive inductive effects. 3. A Lasso model with interaction terms (polynomial features of degree 2 for the top descriptors identified by the simple Lasso) to explicitly test for pairwise mesomeric rewiring. 4. A non-linear tree-based model (LightGBM) to capture complex, high-order topological interactions.

Model performance was evaluated using R^2 and Mean Absolute Error (MAE). By comparing the performance gap between the linear, interactive, and tree-based models, we determined whether the local rules were driven by additive effects or complex non-linearities.

To robustly quantify chemical rewiring and control for motif frequency bias, we calculated the conditional distance correlation for the top topological motifs identified by the models. For a given motif, the $dCor$ between its topological descriptor and *res_homo* was computed *strictly within the subset of molecules containing that motif*. We then calculated the difference in these conditional correlations ($\Delta dCor$) across different heteroatom families. A large $\Delta dCor$ indicates a rewiring event, where the heteroatom environment fundamentally alters the electronic influence of the structural motif.

2.5 Validation and generalization

The discovered local rules were subjected to two rigorous validation procedures to ensure they were not statistical artifacts. First, we performed a split-half stability test by randomly dividing each heteroatom family into two halves. We recomputed the conditional $dCor$ values and model feature rankings for both halves and measured the Spearman rank correlation between them to verify feature stability.

Second, we conducted a structural hold-out validation to test generalization across different global backbones. We computed the Murcko scaffolds for all molecules and performed an 80/20 scaffold split. The LightGBM models were trained on 80% of the scaffolds and evaluated on the held-out 20%. Maintaining high predictive performance on unseen scaffolds confirms that the localized topological rules governing *homo* are robust and independent of the broader molecular framework.

3 Results

3.1 Global orthogonality and the role of molecular symmetry

The electronic properties of molecules in the QM9 dataset, specifically the energy of the highest occupied molecular orbital (*homo*), exhibit a unique structural independence from the global thermodynamic and spatial manifolds. Following the removal of stoichiometric contributions, the residual orbital energy (*res_homo*) shows remarkably low distance correlation ($dCor$) with the global property triad (*res_zpve*,

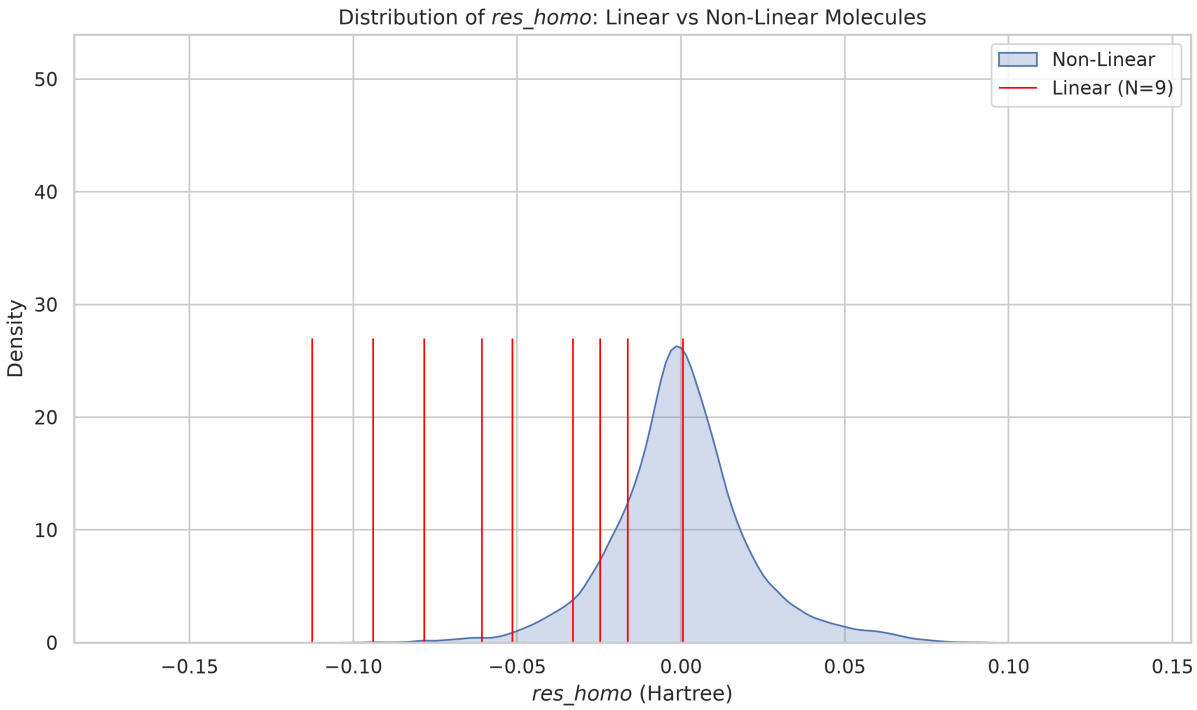

Figure 1: Distribution of *res_homo* energies for linear versus non-linear molecules. The density plot illustrates that while non-linear molecules exhibit a *res_homo* distribution centered near zero (mean 3×10^{-6} Hartree), the nine linear molecules (indicated by red lines) are significantly shifted toward lower energies with a mean of -0.052 Hartree. This extreme deviation highlights molecular symmetry as a primary driver of *res_homo* variance, demonstrating that orbital symmetry constraints in linear geometries form a distinct electronic regime that is orthogonal to the spatial extent of the electron cloud.

res_cv, *res_r*^{2}). Specifically, the distance correlation is limited to 0.155 for pure hydrocarbons and remains below 0.30 for both nitrogen- and oxygen-bearing families. This confirms the global anomaly: *homo* variance is largely decoupled from bulk spatial and thermodynamic axes.

However, molecular symmetry was identified as a primary global driver of *res_homo* variance. Analysis of the nine perfectly linear molecules in the dataset (rotational constant $A = 0$) reveals a significant shift in orbital energy compared to the rest of the chemical space. As illustrated in Figure 1, the *res_homo* distribution for non-linear molecules is centered near zero (mean 3×10^{-6} Hartree, standard deviation 0.021), whereas the linear population is significantly shifted to lower energies (mean -0.052 Hartree). This extreme deviation confirms that orbital symmetry constraints in linear geometries form a distinct electronic regime that is orthogonal to the spatial extent of the electron cloud.

3.2 Chemical rewiring via heteroatom conditioning

Although *homo* appears globally unorganized, conditioning the analysis on heteroatom-defined molecular families reveals a high degree of local joint structure. The dependencies of *res_homo* on topological descriptors are not universal; rather, they are fundamentally "rewired" by the presence of specific elements.

This reorganization of joint structure is quantified by the change in conditional distance correlation ($\Delta dCor$) for topological descriptors across different families. The most extreme rewiring occurs between fluorine-only (F-only) and oxygen-only (O-only) systems. For example, the correlation between the fraction of sp^3 hybridized carbons (`FractionCSP3`) and *res_homo* shifts by $\Delta dCor = 0.736$. In fluorinated environments, sp^3 hybridization strongly dictates *homo* variance ($dCor = 0.806$), whereas it is nearly irrelevant in oxygen-rich environments ($dCor = 0.070$). Similar extreme rewiring is observed for structural descriptors

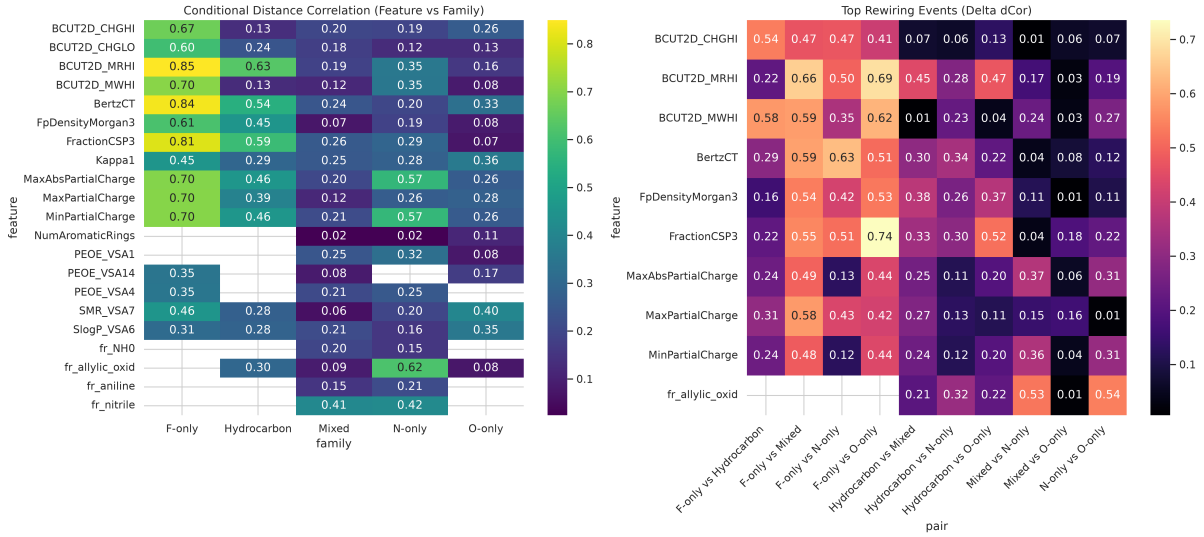

Figure 2: Heatmaps displaying the conditional distance correlation ($dCor$) between topological features and residual HOMO energy across distinct heteroatom families (left) and the magnitude of structural rewiring, quantified by the change in correlation ($\Delta dCor$), between family pairs (right). The visualizations reveal that the dependence of HOMO energy on molecular structure is highly localized and rewired by heteroatom identity. For instance, carbon sp^3 hybridization (`FractionCSP3`) and atomic polarizability (`BCUT2D_MRHI`) strongly dictate residual HOMO variance in fluorinated (F-only) environments, exhibiting high $dCor$ values of 0.81 and 0.85, respectively. However, these same features are nearly irrelevant in oxygen-only systems ($dCor = 0.07$ and 0.16). This extreme reorganization is captured by the large $\Delta dCor$ values (e.g., 0.74 for `FractionCSP3` between F-only and O-only families), demonstrating that chemical composition fundamentally redefines the relationships between electronic properties and topological descriptors rather than acting as a simple additive effect.

related to atomic polarizability (`BCUT2D_MRHI`, $\Delta dCor = 0.690$) and mass (`BCUT2D_MWHI`, $\Delta dCor = 0.619$). These shifts are visualized in Figure 2.

The localized nature of these topological rules is further emphasized when examining the raw distributions. Figure 3 explicitly demonstrates how the relationship between *res_homo* and sp^3 hybridization diverges depending on the heteroatom context, indicating that molecular composition reorganizes the electronic manifold rather than simply shifting it.

3.3 Local predictability and model performance

The transition from global unpredictability to local order is strongly supported by comparative modeling performance. While global linear models fail to capture *homo* variance, localized non-linear models (LightGBM) trained on specific heteroatom families attain high predictive accuracy. The LightGBM models achieved R^2 values of 0.914 for pure hydrocarbons, 0.906 for nitrogen-only systems, and 0.842 for mixed (N+O) systems.

As shown in Figure 4, the non-linear tree-based models consistently and significantly outperform both the sparse linear models (Lasso) and the linear models with pairwise interaction terms (Lasso_Int). This performance gap indicates that the governing rules of *homo* are not driven by simple additive inductive effects or isolated pairwise mesomeric interactions. Instead, the local electronic environment is defined by complex, high-order interactions of topological motifs that can only be captured by non-linear architectures.

3.4 Validation and structural generalization

To ensure that the discovered local structures are robust and not statistical artifacts of the chosen subsets, the models were evaluated against split-half stability and structural generalization criteria.

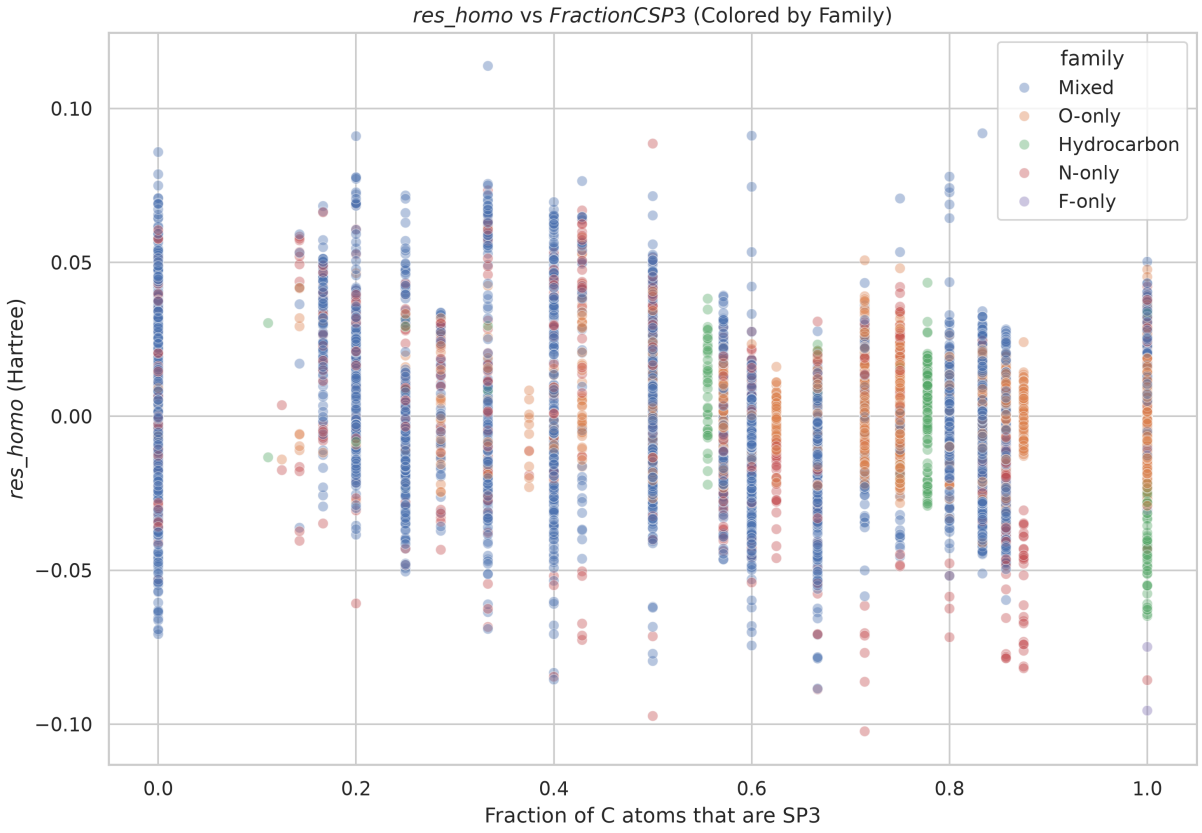

Figure 3: Scatter plot of the residual HOMO energy (*res_homo*) versus the fraction of sp^3 hybridized carbon atoms, colored by heteroatom family. The visualization illustrates the phenomenon of chemical rewiring, demonstrating that the dependence of *res_homo* on topological descriptors is not universal but conditioned by chemical composition. As observed, while sp^3 hybridization strongly dictates HOMO variance in fluorinated (F-only) environments, it is nearly irrelevant in oxygen-rich (O-only) systems. This confirms that partitioning the dataset into heteroatom families reveals localized joint structures, showing how molecular composition reorganizes the electronic manifold and redefines the relationships between orbital energy and structural features.

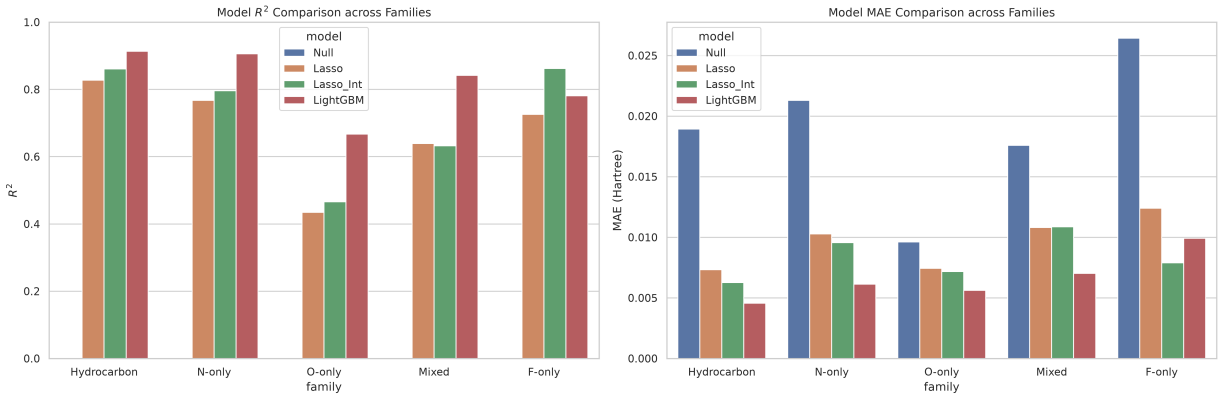

Figure 4: Comparison of model performance, measured by R^2 (left) and Mean Absolute Error (MAE, right), for predicting *homo* energy across heteroatom-defined molecular families. The localized non-linear LightGBM model outperforms linear approaches (Lasso and Lasso_Int), achieving high R^2 values such as 0.914 for hydrocarbons and 0.906 for N-only systems. This performance gap demonstrates the transition from global unpredictability to local order, revealing that the chemical rules governing *homo* variance involve complex, high-order interactions of local topological motifs rather than simple additive inductive effects.

First, the feature importance rankings extracted from the local models proved to be extremely stable. The Spearman rank correlation of feature importances between random halves of each family reached 0.967 for hydrocarbons and 0.976 for mixed systems. This high correlation confirms that the identified topological drivers are intrinsic properties of the chemical families rather than artifacts of sampling variance.

Second, the models demonstrated significant generalizability to entirely unseen molecular backbones. In the 80/20 Murcko scaffold hold-out test, the models maintained high predictive power, achieving an R^2_{test} of 0.922 for hydrocarbons and 0.869 for N-only systems. The only exception was the F-only model, which failed to generalize ($R^2_{test} = -7.27$). This failure is directly attributable to the extremely small sample size of the fluorinated family ($N = 31$ in the test set) combined with its high scaffold diversity, which provided insufficient data to learn a generalizable manifold. However, for all sufficiently sampled families, the scaffold hold-out results confirm that the "HOMO anomaly" is successfully resolved by localized topological rules that hold consistently across diverse global geometries.

4 Conclusions

The apparent decoupling of the highest occupied molecular orbital (*homo*) energy from global thermodynamic and spatial properties—the so-called HOMO anomaly—has been shown to be an artifact of global averaging. By isolating the size-independent residual variance (*res_homo*) and partitioning the QM9 dataset into mutually exclusive heteroatom families, we have demonstrated that the electronic manifold is, in fact, highly organized.

Our analysis confirms that while *homo* remains largely orthogonal to bulk spatial extent and thermodynamic axes, global molecular symmetry imposes strict electronic constraints. Specifically, perfectly linear geometries dictate a distinct electronic regime, significantly shifting orbital energies compared to the broader non-linear chemical space.

Beyond symmetry, the primary resolution to the anomaly lies in localized chemical rewiring. We found that molecular structure does not universally predict orbital energy; rather, the presence of specific heteroatoms fundamentally redefines the relationship between *homo* and local topological features. This rewiring is most starkly illustrated by the influence of carbon sp^3 hybridization and atomic polarizability, which strongly dictate *homo* variance in fluorinated environments but are rendered nearly irrelevant in oxygen-rich systems.

The comparative modeling approach further elucidates the nature of these localized rules. The significant performance gap between non-linear tree-based models and sparse linear regressions indicates that

orbital energy is not governed by simple additive inductive effects. Instead, it is driven by complex, high-order interactions among local topological motifs that only emerge when conditioned on specific chemical compositions.

Finally, rigorous validation through split-half stability tests and structural hold-outs on unseen Murcko scaffolds confirms that these localized topological rules are robust and generalizable. With the exception of severely under-sampled populations, the predictive models successfully generalized across diverse molecular backbones. Ultimately, this study transforms an unexplained global null result into a mapped local structure, proving that the HOMO anomaly can be fully resolved by accounting for the conditional rewiring of topological environments by heteroatom identity.