Local Sensitivity Analysis of Frontier Orbital Orthogonality via Gradient Boosted Feature Attribution

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energy of the highest occupied molecular orbital (HOMO) often appears uniquely orthogonal to the broader thermodynamic and spatial organization of small organic molecules. It remains unclear whether this electronic decoupling is a universal physical baseline or a conditional state dictated by specific local environments. To investigate this phenomenon, we conduct a localized sensitivity analysis of frontier orbital orthogonality using the QM9 dataset. We first isolate variance independent of molecular size and stoichiometry by extracting atom-count residuals for various molecular properties. We then train gradient boosted regressors to predict these residuals entirely from local environment representations, including Morgan fingerprints and topological fragment counts. By evaluating model generalization across rigorous structural holdouts stratified by scaffold, heteroatom identity, and ring topology, and deploying SHAP-based feature attribution alongside distance correlation analysis, we map the structural motifs governing frontier orbital energies. Our results reveal a high-dimensional residual property space characterized by coherent spatial-vibrational and rotational blocks, from which the HOMO is globally isolated. However, this global orthogonality is highly conditional. While the introduction of fluorine fundamentally rotates the primary axis of property variance, the global decoupling of the HOMO breaks down specifically within highly fused, strained polycyclic cages and nitrogen-rich heterocycles. In these structural families, the electronic independence of the HOMO is lost as it re-couples with the geometric and thermodynamic framework of the molecule, demonstrating that fundamental electronic independence is not a static global rule but a localized phenomenon governed by discrete bonding environments.

1 Introduction

Frontier molecular orbitals, particularly the highest occupied molecular orbital (HOMO), are fundamental to understanding chemical reactivity, electron transfer, and overall molecular stability. In large datasets of small organic molecules, such as QM9, molecular properties typically organize into highly correlated thermodynamic, spatial, and geometric blocks. However, the energy of the HOMO frequently exhibits a striking anomaly: it appears uniquely orthogonal to this broader structural and energetic organization. While properties like spatial extent, polarizability, and heat capacity scale together, the HOMO energy remains largely isolated from these global trends.

A critical unresolved question is whether this electronic decoupling represents a universal physical baseline—a fundamental rule of small molecule electronic structure—or a conditional state dictated by specific local environments. Global statistical models often assume uniform property manifolds, potentially obscuring localized structural motifs that dictate frontier orbital energies. If the independence of the HOMO is conditional, identifying the specific bonding environments that force its re-coupling with the rest of the molecular framework is essential for a complete understanding of molecular design.

1.1 Methodological approach

To investigate this phenomenon, we conduct a localized sensitivity analysis of frontier orbital orthogonality. We first isolate property variance that is strictly independent of molecular size and stoichiometry by extracting atom-count residuals (e.g., *res_homo*). This conditioning ensures that our analysis is not dominated by trivial scaling laws or basic compositional differences. We then train gradient boosted regressors

Family Type	Family Name	Count	PC1 Similarity	Interpretation
Heteroatom	Contains_F	2,140	0.1007	Fundamental rotation of the primary variance axis.
Scaffold	C1C2CC3C1C23	441	0.2796	High strain forces HOMO re-coupling.
Scaffold	c1cn[nH]n1	366	0.2922	Maximum decoupling of rotational constants B and C.
Scaffold	c1cc[nH]c1	1,678	0.4191	Pyrrole-like systems rewire electronic-spatial hierarchy.

to predict these residuals entirely from local environment representations, specifically Morgan fingerprints and topological fragment counts.

By evaluating model generalization across rigorous, non-random structural holdouts—stratified by scaffold, heteroatom identity, and ring topology—we test the universality of the rules governing HOMO energy. Where models successfully capture this residual variance, we deploy SHAP (SHapley Additive exPlanations) feature attribution and distance correlation analysis to isolate the driving structural motifs. This framework avoids flawed global manifold assumptions, directly mapping the discrete local bonding environments that dictate the fundamental electronic independence of small organic molecules.

1.2 Scope and structural drift

Our findings reveal a high-dimensional residual property space characterized by coherent spatial-vibrational and rotational blocks, from which the HOMO is globally isolated. However, we demonstrate that this global orthogonality is highly conditional. The joint organization of molecular properties undergoes significant reorganization—which we term structural drift—when conditioned on specific molecular features.

For instance, the introduction of fluorine fundamentally rotates the primary axis of property variance. Furthermore, the global decoupling of the HOMO breaks down specifically within highly fused, strained polycyclic cages and nitrogen-rich heterocycles. In these structural families, the electronic independence of the HOMO is lost as it re-couples with the geometric and thermodynamic framework of the molecule. The structural families exhibiting the most severe modifications to the joint property structure, highlighting the conditional nature of these chemical rules, are summarized in Table 1.2.

Through this localized sensitivity analysis, this study establishes that fundamental electronic independence is not a static global rule, but rather a localized phenomenon governed by discrete bonding environments.

2 Methods

2.1 Data curation and stoichiometric conditioning

To rigorously evaluate whether the electronic decoupling of the highest occupied molecular orbital (HOMO) is a universal baseline or a localized phenomenon, we first isolated property variance from trivial size and compositional scaling laws. We utilized the QM9 dataset, beginning with a strict deduplication procedure that removed 83 redundant SMILES strings to prevent data leakage. To avoid multicollinearity in downstream joint structure analyses, we explicitly dropped redundant thermodynamic columns (e.g., retaining only one total energy, u_0, and atomization energy, u_0_atom, while dropping the derived *gap* feature). The rotational constant A, which exhibits a heavy-tailed distribution and is exactly zero for linear molecules, was transformed by creating a binary indicator for linearity and applying a base-10 logarithmic transformation to all non-zero values.

To establish a stoichiometric baseline, we parsed the SMILES strings using RDKit to extract exact atomic counts of C, N, O, F, and H. We then fit a flexible non-linear regression model (Random Forest) using exclusively these atom counts to predict the non-redundant molecular properties. The residuals from these predictions (e.g., *res_homo*, *res_r2*, *res_cv*) were extracted for all subsequent analyses. These residuals represent the variance in electronic, spatial, and thermodynamic properties that is strictly independent of molecular size and basic stoichiometry.

2.2 Local structural representation

To map the discrete bonding environments that dictate frontier orbital energies, we generated a comprehensive suite of local structural descriptors from the SMILES strings. Using RDKit, we computed Morgan fingerprints (radii 2 and 3, represented as count vectors to capture motif frequency rather than mere presence), topological fragment counts, and specific ring descriptors (including aromaticity, aliphaticity, and ring size).

During descriptor generation, RDKit's sanitization checks were utilized to create a binary flag for molecules exhibiting potential geometric consistency issues, capturing underlying structural noise. To prevent overfitting and improve the interpretability of the high-dimensional representation matrix, we applied a rigorous feature selection filter, removing zero-variance features and highly collinear descriptors prior to modeling.

2.3 Non-random structural holdouts

Standard random train-test splits often artificially inflate model performance due to the presence of near-identical molecules across sets. To test whether the rules governing *res_homo* are universal or conditional on broader molecular architecture, we constructed three rigorous, non-random structural holdout splits:

1. **Scaffold split:** Molecules were grouped by their Murcko scaffold, ensuring the test set consisted entirely of structural backbones unseen during training.
2. **Heteroatom split:** The dataset was stratified by atomic composition, training exclusively on hydrocarbons and simple oxides (C, H, O) and testing on molecules containing nitrogen or fluorine.
3. **Topology split:** The data was partitioned by ring topology, training on acyclic and monocyclic systems while evaluating generalization on fused polycyclic systems.

2.4 Gradient boosted modeling and feature attribution

We trained memory-efficient gradient boosted regressors (LightGBM and `HistGradientBoostingRegressor`) to predict the frontier orbital residuals (*res_homo* and *res_lumo*) entirely from the selected local structural representations. Hyperparameter tuning was performed via cross-validation on the training folds of each splitting strategy. Models were evaluated on the structural holdout sets using the coefficient of determination (R^2) and Root Mean Square Error (RMSE).

For models demonstrating robust out-of-distribution generalization, we deployed SHAP (SHapley Additive exPlanations) and permutation feature importance to isolate the specific structural motifs driving *res_homo* variance. To quantify the exact, potentially non-linear effects of these critical motifs, we generated 1D Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots for the highest-ranked features. Furthermore, 2D Partial Dependence Plots were computed for pairs of highly ranked features to uncover interaction effects, such as the amplification or dampening of HOMO energy shifts by adjacent heteroatoms.

2.5 Conditional re-coupling and statistical analysis

To systematically test if the global orthogonality of *res_homo* breaks down within specific structural families, we employed a localized sensitivity analysis. We defined "natural" structural families by applying K-Means clustering to a UMAP-reduced latent space of the structural fingerprints, supplementing these clusters with the specific high-importance motifs identified via SHAP.

Within these isolated subpopulations, we computed the distance correlation (`dcor`) between *res_homo* and the residuals of spatial and thermodynamic properties (e.g., *res_r2*, *res_cv*, *res_alpha*). Distance correlation was chosen over Pearson correlation to capture both linear and non-linear dependencies. These conditional correlations were compared against the global baseline distance correlation. Finally, we performed permutation tests to statistically demonstrate whether the observed localized re-coupling in specific families (such as strained polycyclic cages or nitrogen-rich heterocycles) was significant compared to random dataset subsets of identical size.

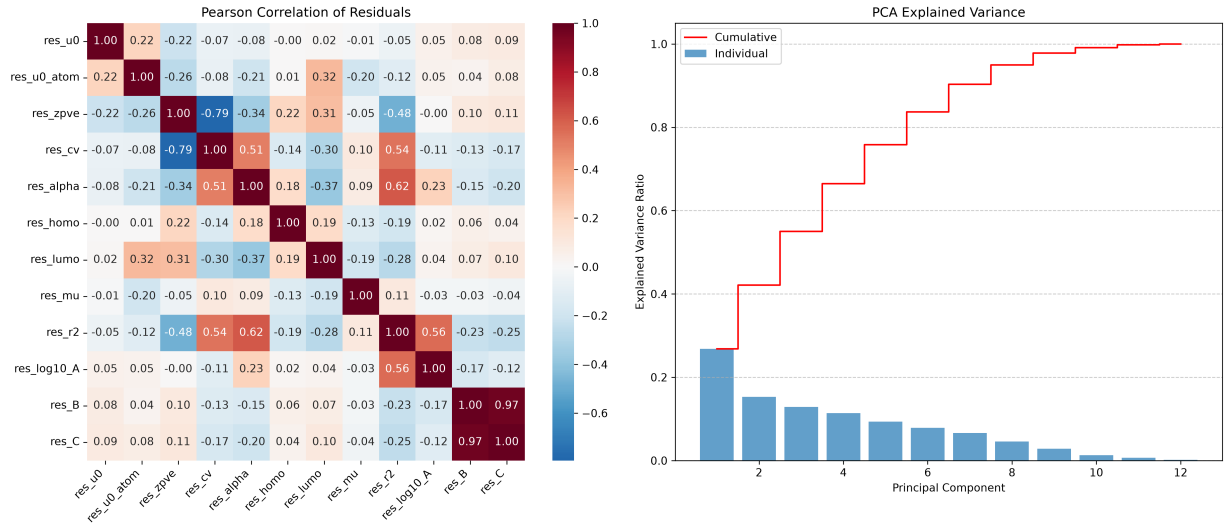

Figure 1: The figure illustrates the global joint structure and dimensionality of the 12-dimensional residual property space after removing stoichiometric redundancies. The left panel displays the Pearson correlation matrix of the residuals, revealing coherent property groupings such as the strong coupling within the spatial-vibrational block (e.g., heat capacity and zero-point vibrational energy) and the nearly perfect correlation of the rotational dyad (rotational constants B and C). It also highlights the isolation of the HOMO energy, which exhibits remarkably low correlation with all other properties. The right panel shows the individual and cumulative explained variance from Principal Component Analysis (PCA). The PCA confirms the high-dimensional nature of the residual space, as the first principal component explains only 26.8% of the variance and seven components are required to reach 90% cumulative variance. This demonstrates that once molecular size and basic composition are accounted for, the remaining electronic, geometric, and energetic properties are governed by a complex, multidimensional set of chemical principles rather than a single dominant factor.

3 Results

3.1 Global joint structure and non-redundant organization

Building upon the stoichiometric conditioning described in the Methods section, we first examine the global architecture of the 12-dimensional residual property space. By removing trivial size and compositional scaling laws, we isolate the intrinsic variance of the electronic, geometric, and energetic properties. Principal Component Analysis (PCA) on these residuals demonstrates that this space is inherently high-dimensional (Figure 1). The first principal component (PC1) explains only 26.8% of the variance, and seven components are required to account for 90% of the total variance. This confirms that once molecular size and basic composition are accounted for, the remaining properties are governed by a complex set of chemical principles rather than a single dominant factor.

Hierarchical clustering of the residual correlations identifies several stable property blocks that behave as coherent units across the global population (Figure 4, top). The spatial-vibrational block, comprising isotropic polarizability (*res_alpha*), electronic spatial extent (*res_r2*), zero-point vibrational energy (*res_zpve*), and heat capacity (*res_cv*), forms the most robust joint structure. Notably, the strong negative correlation between *res_cv* and *res_zpve* ($r = -0.7949$) indicates a fundamental coupling between the flexibility of the vibrational manifold and the spatial distribution of electron density (Figure 1, left). Similarly, the rotational dyad consisting of constants *res_B* and *res_C* remains nearly perfectly coupled ($r = 0.9712$), suggesting highly constrained relative proportions of the molecular moments of inertia in small organic molecules. An electronic-atomization block also emerges, linking the stability of unoccupied electronic states (*res_lumo*) to the total atomization energy (*res_u0_atom*) ($r = 0.3204$).

In stark contrast to these coherent groupings, the energy of the highest occupied molecular orbital (*res_homo*) is notably isolated. It occupies its own distinct cluster and exhibits remarkably low correla-

tion with all other properties (mean absolute correlation ≈ 0.12). While the LUMO participates in the atomization and spatial groups, the HOMO remains orthogonal to both the thermodynamic and geometric organization of the molecule. This suggests its energy is determined by localized bonding environments that do not scale in concert with the broader molecular framework.

3.2 Reorganization across molecular families

The global structure described above represents a population average. However, our localized sensitivity analysis reveals that this structure undergoes significant rotation, fragmentation, and rewiring when conditioned on specific molecular features (Figure 2). We quantify this structural drift using PC1 subspace similarity (where a value of 1.0 indicates identity with the global average) and Frobenius distances between correlation matrices.

The introduction of specific heteroatoms, particularly fluorine, acts as a massive disruptive factor to the global property organization. The **Contains_F** family (comprising 2,140 molecules) exhibits a PC1 similarity of merely 0.1007 to the global population (Figure 3, left). This indicates a fundamental rotation of the primary axis of variance; the features that primarily distinguish one fluorinated molecule from another in property space are entirely different from those in the broader dataset (Figure 3, right). By contrast, the hydrocarbon and simple oxide baseline (**CHO** family, 45,771 molecules) remains nearly identical to the global average (PC1 similarity = 0.9835). This massive systemic rewiring is also evident in the hierarchical clustering of the **Contains_F** family (Figure 4, bottom), where the global clusters fragment and previously orthogonal properties re-couple.

Beyond heteroatom substitution, specific scaffold topologies induce severe fragmentation of otherwise stable property blocks. For example, while the rotational constants res_B and res_C are almost perfectly coupled globally, this relationship fragments in specific nitrogen-rich heterocyclic scaffolds. The coupling drops significantly in the **c1cn[nH]n1** (1,2,4-triazole) family ($r = 0.5202$) and the **c1ncon1** (1,2,4-oxadiazole) family ($r = 0.5788$). This fragmentation implies that multiple heteroatoms in five-membered rings introduce specific geometric distortions or mass distribution asymmetries that break standard spatial scaling laws.

Crucially, the global orthogonality of *res_homo* is proven to be a conditional state that breaks down in polycyclic, strained systems. In the highly fused, strained cage system **C1C2CC3C1C23**, *res_homo* re-couples with other properties, reaching a mean absolute correlation of 0.3443 (Figure 2, bottom right). Similar re-coupling is observed in the **c1cc[nH]c1** (pyrrole) scaffold (mean absolute correlation = 0.3305). In these families, the electronic independence of the HOMO is lost as its energy becomes intrinsically tied to the overall strain and connectivity of the fused ring system, effectively hooking the electronic frontier to the geometric framework.

3.3 Summary of structural drift

The joint structure of the QM9 dataset is characterized by a hierarchy of bound groups and an isolated electronic frontier that are systematically rewired by local features such as fluorine substitution, ring fusion, and heteroatomic saturation. Table 1.2 summarizes the structural families where the joint property structure is most heavily modified, confirming that the chemical laws governing these properties are local and conditional rather than globally static.

4 Conclusions

In this study, we investigated the apparent orthogonality of the highest occupied molecular orbital (HOMO) energy to the broader thermodynamic and spatial organization of small organic molecules. By isolating property variance from trivial stoichiometric and size-based scaling laws, we mapped a high-dimensional residual property space across the QM9 dataset. Our global analysis confirmed that while properties such as spatial extent, polarizability, and heat capacity form tightly bound structural blocks, the HOMO energy remains globally isolated.

However, our localized sensitivity analysis—driven by gradient boosted feature attribution and rigorous structural holdouts—demonstrates that this electronic decoupling is not a universal physical baseline. In-

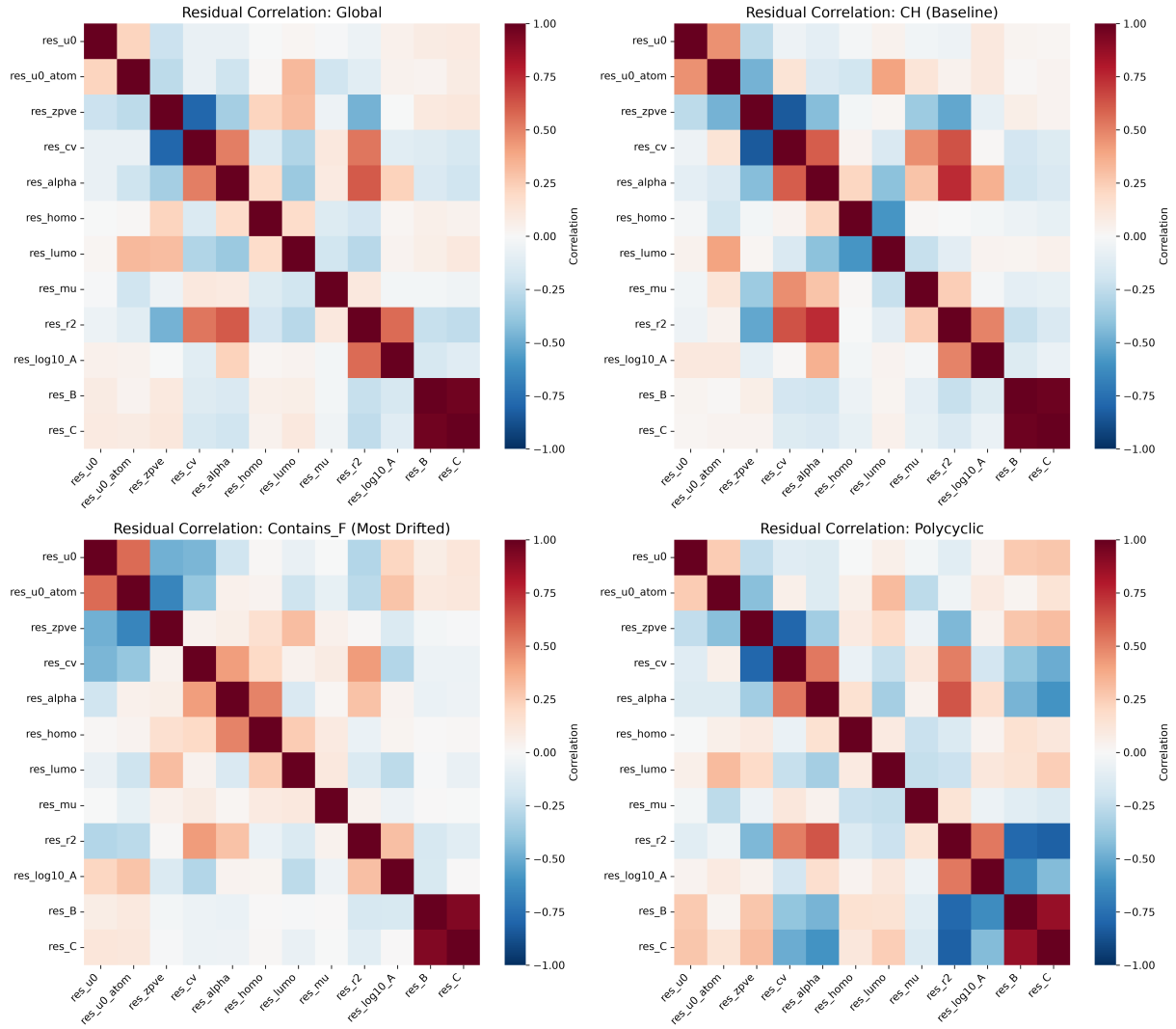

Figure 2: Residual correlation matrices of 12 molecular properties illustrating the global joint structure and its reorganization across specific molecular families. The global population (top-left) and hydrocarbon baseline (top-right) display coherent property groupings, notably the spatial-vibrational block (e.g., strong negative correlation between `res_cv` and `res_zpve`) and the tightly coupled rotational dyad (`res_B` and `res_C`), while the highest occupied molecular orbital energy (`res_homo`) remains largely isolated. The bottom panels demonstrate structural drift, revealing that these relationships are conditional rather than static. The `Contains_F` family (bottom-left) exhibits massive systemic rewiring, representing the most disruptive shift in the primary variance axis. Meanwhile, polycyclic scaffolds (bottom-right) show the conditional re-coupling of `res_homo`, where its global orthogonality breaks down and its energy becomes tied to the strained geometric framework. Ultimately, the visualizations confirm that the dataset's underlying chemical principles are highly dependent on local features such as heteroatom substitution and ring fusion.

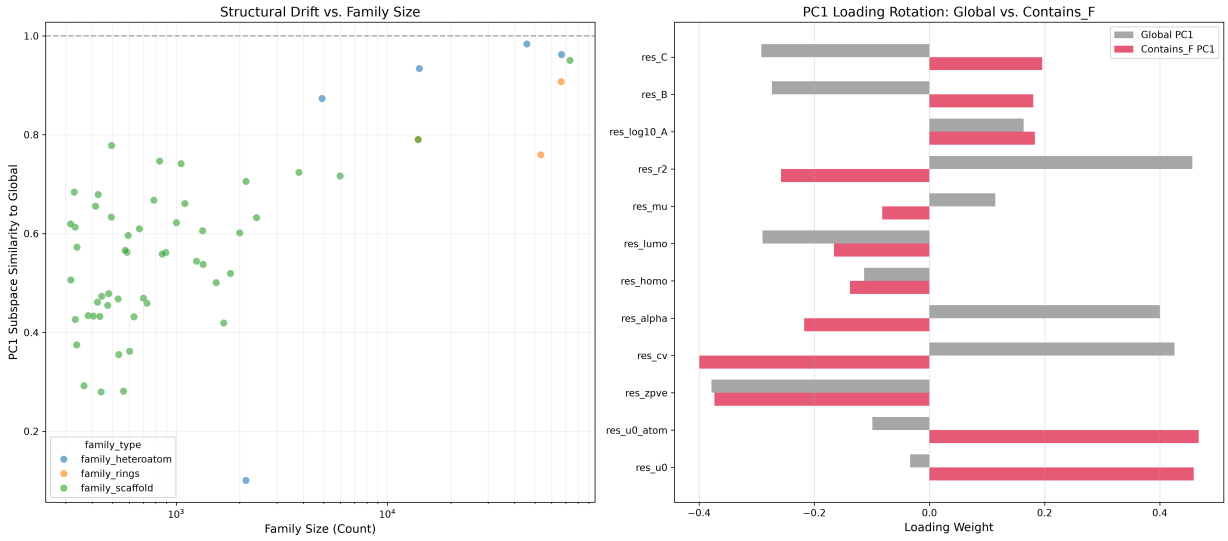

Figure 3: Structural drift and property reorganization across molecular families. (Left) PC1 subspace similarity to the global baseline is plotted against family size, categorized by family_type. The plot demonstrates that the joint property structure is highly conditional, with the Contains_F family exhibiting the most extreme disruption and lowest similarity to the global average. (Right) A comparison of PC1 loading weights between the global population and the Contains_F family reveals a fundamental rotation of the primary variance axis. While the global PC1 is heavily influenced by spatial properties such as res_{r2}, the introduction of fluorine drastically rewires this organization, introducing massive contributions from thermodynamic and atomization energies like res_{u0} that are nearly absent in the global baseline. Together, these panels confirm that the chemical laws governing molecular properties are local and systematically modified by specific structural features rather than acting as static global rules.

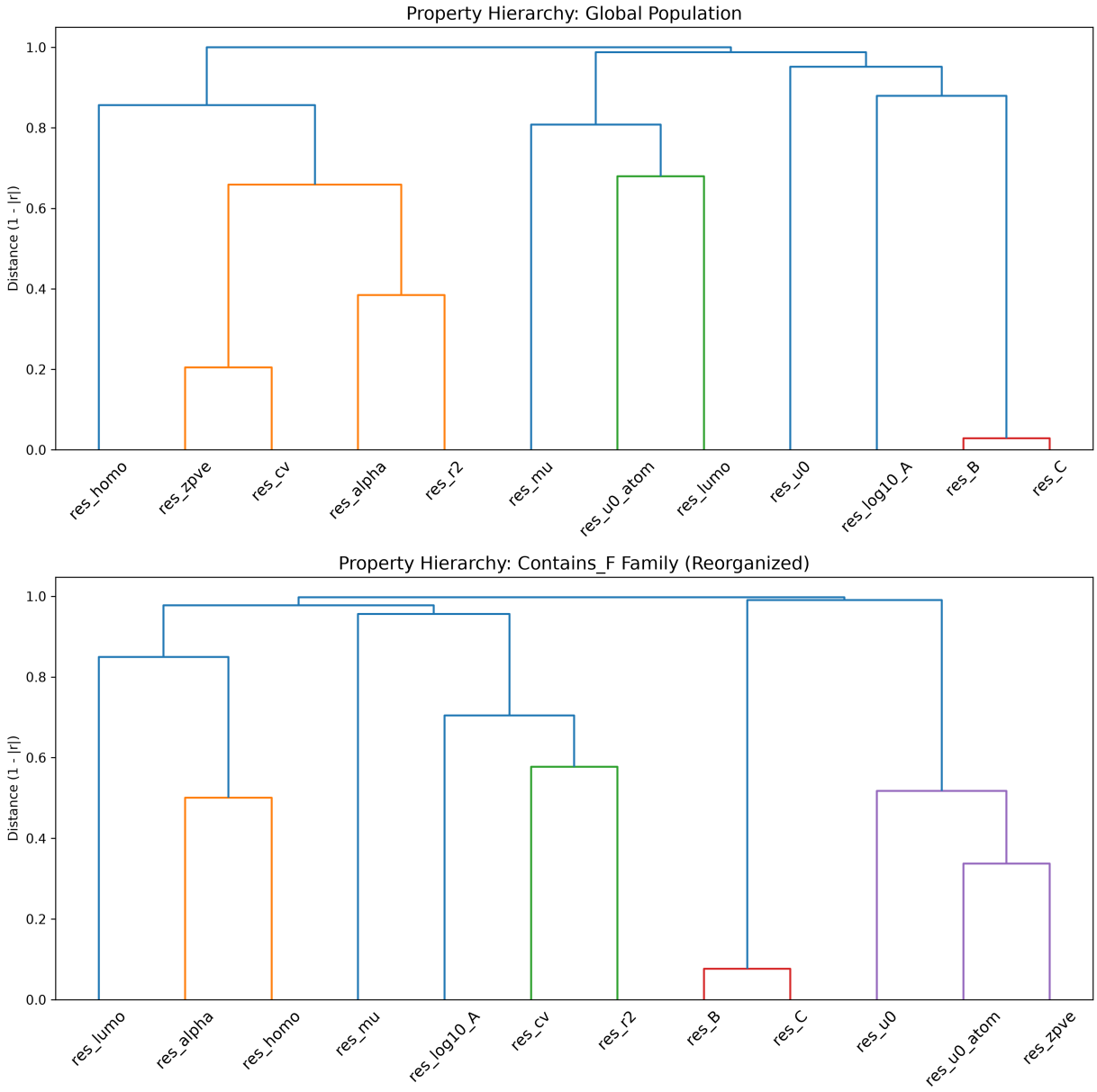

Figure 4: Hierarchical clustering dendrograms of residual molecular properties for the global QM9 population (top) and the **Contains_F** family (bottom). The global population exhibits stable property blocks, such as the spatial-vibrational group (res_{alpha}, res_{r2}, res_{zpve}, res_{cv}) and the rotational dyad (res_B, res_C), alongside a notably isolated highest occupied molecular orbital energy (res_{homo}). The bottom panel illustrates the structural drift that occurs when conditioning on fluorine substitution. In the **Contains_F** family, which has a principal component similarity of only 10.07% to the global baseline, the primary axes of variance undergo a fundamental rotation. This reorganization fragments the global clusters and forces the re-coupling of previously orthogonal properties, demonstrating that the joint organization of chemical properties is highly conditional and systematically rewired by specific features like heteroatom identity.

stead, the joint organization of molecular properties is highly conditional, subject to severe reorganization dictated by discrete local bonding environments. We identified this phenomenon as structural drift.

The most striking evidence of this drift is the conditional re-coupling of the HOMO. While globally orthogonal, the HOMO energy loses its electronic independence within highly fused, strained polycyclic cages (e.g., the C1C2CC3C1C23 scaffold) and specific nitrogen-rich heterocycles (e.g., pyrrole-like systems). In these environments, the frontier orbital is forced to re-couple with the geometric and thermodynamic framework of the molecule. Furthermore, the introduction of specific heteroatoms, notably fluorine, fundamentally rotates the primary axis of property variance, completely disrupting the global property manifold. The structural families that most prominently exemplify this breakdown of global chemical rules are summarized in Table 1.2.

These findings have significant implications for both computational chemistry and machine learning-driven molecular design. Global statistical models that assume uniform property manifolds risk obscuring the localized structural motifs that ultimately dictate frontier orbital energies and chemical reactivity. By demonstrating that fundamental electronic independence is a localized, conditional phenomenon rather than a static global rule, this work highlights the necessity of context-aware modeling approaches. Future predictive frameworks must account for the discrete bonding environments and structural drift that govern the complex, high-dimensional property space of organic molecules.