Structural Drivers of Frontier Orbital Orthogonality and Contextual Rewiring in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The electronic properties of small organic molecules are typically dominated by macroscopic size and stoichiometry, yet the highest occupied molecular orbital (HOMO) energy exhibits an anomalous orthogonal variance that remains poorly understood. To determine whether this decoupling is driven by localized chemical environments or represents a fundamentally independent manifold, we isolate frontier orbital energy residuals from stoichiometric baselines using the QM9 dataset. We train multi-output gradient-boosted trees on interpretable topological descriptors to predict these residuals and decompose the underlying structural drivers using SHapley Additive exPlanations (SHAP) combined with Canonical Correlation Analysis (CCA). Our models reveal a sharp contrast in predictability between the frontier orbitals, establishing the HOMO state as highly sensitive to fine-grained local bonding rather than standard topological features. Despite this global decoupling, canonical correlation of the feature importances identifies a strong joint structural organization between HOMO and LUMO residuals, driven primarily by conjugated fluorine content and aromaticity. Furthermore, distance correlation analysis demonstrates that these specific structural motifs act as switchboards that significantly rewire the fundamental relationships between electronic states and geometric response properties, such as strongly coupling the HOMO energy to isotropic polarizability in the presence of conjugated fluorine.

1 Introduction

The electronic and thermodynamic properties of small organic molecules are predominantly governed by macroscopic spatial variables and stoichiometry. In large-scale chemical spaces such as the QM9 dataset, fundamental properties like total internal energy and zero-point vibrational energy are almost entirely explained by simple atom counts. However, the energy of the highest occupied molecular orbital (HOMO) exhibits an anomalous orthogonal variance, remaining largely decoupled from these bulk molecular scales. This decoupling raises a fundamental chemical question: is the HOMO energy manifold driven by highly localized, fine-grained chemical environments, or does it represent a fundamentally independent electronic state that evades standard topological description?

To address this, we isolate the structural drivers of frontier orbital energies by decoupling them from macroscopic size effects. We establish a strictly compositional baseline to extract the residuals of the HOMO and LUMO energies, as well as isotropic polarizability. By removing the dominant stoichiometric variance, we expose the underlying structural logic governing these electronic states. We then employ multi-output gradient-boosted trees to predict these residuals exclusively from interpretable, chemically meaningful topological descriptors, such as ring aromaticity, hybridization fractions, and specific heteroatom environments.

To pinpoint the exact structural hubs dictating these frontier orbital energies, we decompose the predictive models using SHapley Additive exPlanations (SHAP). By coupling SHAP with Canonical Correlation Analysis (CCA), we evaluate whether the structural motifs driving the HOMO and LUMO residuals act in tandem or orthogonally. Finally, we introduce a contextual rewiring analysis using distance correlation metrics to test the hypothesis that specific structural motifs do not merely shift orbital energies, but fundamentally alter the coupling between electronic states and geometric response properties.

Our analysis reveals a stark contrast in the predictability of frontier orbitals, establishing the HOMO state as uniquely sensitive to local bonding environments compared to the LUMO. Despite this global decoupling, canonical correlation identifies a strong joint structural organization between the HOMO and

LUMO residuals, primarily coordinated by conjugated fluorine content and aromaticity. Furthermore, we demonstrate that these specific structural motifs act as chemical switchboards. Rather than acting as simple additive features, motifs such as conjugated fluorine significantly rewire the fundamental relationships between electronic states, strongly coupling the HOMO energy to isotropic polarizability in ways that are absent in the broader chemical space.

2 Methods

2.1 Data preprocessing and baseline extraction

To isolate the structural drivers of frontier orbital energies from macroscopic size effects, we first established a strictly compositional baseline using the QM9 dataset. The dataset was deduplicated based on SMILES strings to prevent data leakage during cross-validation. We retained 12 non-redundant numeric properties, treating thermodynamic and atomization blocks as single variables. Nine linear molecules (where the rotational constant $A = 0$) were excluded to prevent geometric edge cases from skewing the analysis. The 3,054 molecules failing standard geometric consistency checks were retained; we relied on the robustness of tree-based models and non-parametric statistical tests in downstream steps to mitigate their potential outlier effects.

We defined a macroscopic baseline feature set consisting of atom counts (C, N, O, F, H), total heavy atom count, total electron count, and simple connectivity metrics (number of rings, double bonds, and triple bonds). A shallow gradient-boosted tree (GBT) was trained to predict the molecular properties from this baseline. By restricting the model complexity, we prevented it from absorbing the fine-grained structural signals we aimed to isolate. The residuals from these baseline predictions—specifically for the HOMO energy (res_homo), LUMO energy (res_lumo), and isotropic polarizability (res_alpha)—were extracted for all subsequent analyses.

2.2 Structural feature engineering

To characterize the local chemical environments driving the isolated residuals, we computed a targeted set of interpretable topological descriptors using RDKit. Black-box representations such as Morgan fingerprints were avoided to maintain direct chemical interpretability. The engineered features captured three primary structural domains:

- **Conjugation and aromaticity:** The fraction of sp2 hybridized carbons (f_sp2_c) and the total number of aromatic rings ($n_aromatic_rings$).
- **Heteroatom environments:** The counts of nitrogen, oxygen, and fluorine atoms directly bonded to aromatic systems or participating in conjugated networks (e.g., n_F_conj, n_N_conj).
- **Strain and topology:** The number of spiro atoms, bridgehead atoms, and small (3- or 4-membered) rings.

All engineered features were standardized to ensure stable gradient calculations during the modeling phase.

2.3 Multi-output gradient boosting and SHAP decomposition

We formulated a multi-output regression task targeting the joint residual vector $[res_homo, res_lumo]$. A multi-output XGBoost model was trained using the standardized structural features as inputs. Tree depth was constrained to maintain interpretability and prevent overfitting to highly correlated structural motifs. To rigorously evaluate the model's ability to generalize structural rules across distinct chemical families, we implemented a 5-fold cross-validation strategy that split the dataset by molecular formula rather than by random row sampling.

To pinpoint the exact structural hubs dictating the frontier orbital energies, we extracted SHapley Additive exPlanations (SHAP) value matrices from the trained model for both res_homo and res_lumo. To mitigate the "hub fallacy"—where SHAP values might artificially split importance among highly correlated

Feature	HOMO CV1	LUMO CV1	HOMO CV2	LUMO CV2
Conjugated Fluorine (n_F_conj)	0.9528	0.9652	-0.1367	-0.1764
Aromatic Rings ($n_aromatic_rings$)	0.2929	0.3118	0.7719	1.0269
sp2 Carbon Fraction (f_sp2_c)	-0.2882	-0.1529	-0.6093	-0.4725
Conjugated Nitrogen (n_N_conj)	0.1927	-0.0425	0.5175	-0.2152

features—we calculated the Pearson correlation matrix of the SHAP values and grouped highly correlated features. The resulting grouped SHAP matrices were then normalized via column-wise standardization.

2.4 Canonical correlation analysis of feature importances

To determine whether the structural motifs driving the HOMO and LUMO residuals act in tandem or orthogonally, we performed Canonical Correlation Analysis (CCA) on the normalized SHAP matrices. This approach identifies linear combinations of feature importances that maximize the correlation between the structural drivers of the two frontier orbitals. The resulting canonical variables (CVs) and their feature loadings highlight which specific chemical environments coordinate the joint variance of the frontier manifold. The primary structural loadings on the first two canonical variables are detailed in Table 2.4.

Additionally, we applied a joint Uniform Manifold Approximation and Projection (UMAP) embedding on the SHAP matrices. This provided a visual diagnostic to identify non-linear clusters of molecules governed by fundamentally different frontier orbital logic.

2.5 Contextual rewiring and statistical validation

To test the hypothesis that specific structural motifs fundamentally alter the coupling between electronic states and geometric response properties, we conducted a contextual rewiring analysis. We first calculated the baseline distance correlation ($dCor$) between res_homo and res_alpha across the entire deduplicated dataset. The dataset was then partitioned based on the intensity of the top structural motifs identified via CCA (e.g., molecules with high versus low aromatic-fluorine conjugation). We recomputed the $dCor$ between res_homo and res_alpha within these specific partitions to quantify how the structural motif modulates the coupling.

The robustness of the observed rewiring effects was validated through split-half stability testing within each structural partition, recalculating the $dCor$ values on each half to confirm the persistence of the effect. Finally, to rigorously assess statistical significance, we performed a permutation test on the $dCor$ differences. The motif intensity labels were shuffled 1,000 times to generate a null distribution of the difference in $dCor$. The observed changes in property coupling were then compared against this null distribution to confirm that the rewiring is a genuine chemical consequence of the structural families rather than a statistical artifact.

3 Results

3.1 Isolation of frontier electronic residuals

As outlined in our methodology, we first established a stoichiometric baseline to decouple the structural drivers of frontier orbital energies from macroscopic size effects. In the QM9 dataset, bulk molecular scale and stoichiometry overwhelmingly dominate standard properties; atom counts alone explain over 99.9% of the variance in total internal energy and nearly 99.7% of the zero-point vibrational energy. When applying this baseline to the frontier orbitals, we found that it explains 45.1% of the LUMO variance and 40.5% of the HOMO-LUMO gap variance. In stark contrast, the stoichiometric baseline explains only 8.6% of the HOMO energy variance. This confirms that the HOMO energy is fundamentally decoupled from the bulk molecular scale.

To investigate the local chemical environments driving these isolated residuals, we trained multi-output gradient-boosted trees (XGBoost) to predict the residuals from interpretable structural features, including aromaticity, conjugation, ring strain, and heteroatom environments. Evaluating the models using a 5-fold cross-validation strategy split by molecular formula revealed a sharp contrast in predictability between the

two frontier orbitals. The model successfully captured a significant portion of the LUMO residual variance ($R^2 = 0.391 \pm 0.052$). However, it largely failed to predict the HOMO residuals ($R^2 = 0.028 \pm 0.089$). This relative failure establishes the frontier electronic state—particularly the HOMO—as a manifold that is significantly more sensitive to fine-grained local bonding details and non-linear interactions than can be captured by standard topological descriptors.

3.2 Joint structural drivers via CCA and SHAP decomposition

Despite the global decoupling of the HOMO energy from standard structural features, Canonical Correlation Analysis (CCA) applied to the SHAP values of the joint multi-output model reveals a high degree of coordination between the structural drivers of the HOMO and LUMO. The first canonical variable (CV1) achieves a strong linear correlation of 0.966 between the HOMO and LUMO SHAP profiles (Figure 1). This indicates that while the frontier orbitals may evade simple global prediction, the specific structural motifs that do influence one orbital are strongly coordinated with the other.

The structural loadings on these canonical variables identify the primary drivers of this joint organization. As detailed in Table 2.4, the first canonical axis (CV1) is almost entirely dominated by conjugated fluorine content (n_F_conj), which acts as the strongest joint coordinator of the frontier orbital residuals.

Conversely, the second canonical axis (CV2) represents the aromatic and unsaturation network of the molecules. On this axis, the number of aromatic rings and the fraction of sp^2 hybridized carbons exhibit opposing loadings. This orthogonal structure suggests a hierarchical organization: while fluorine conjugation shifts the frontier manifold globally along CV1, the degree of aromaticity and unsaturation modulates the internal relationship between the HOMO and LUMO along CV2.

3.3 Contextual rewiring of molecular property relationships

To determine whether these identified structural hubs merely shift orbital energies or fundamentally alter the coupling between electronic states and geometric response properties, we conducted a contextual rewiring analysis. Across the general dataset, the global distance correlation ($dCor$) between the HOMO residual (res_homo) and the isotropic polarizability residual (res_alpha) is relatively low at approximately 0.28. However, partitioning the molecules into structural families based on the motifs identified via CCA reveals a significant rewiring of these property relationships.

In molecules characterized by the presence of conjugated fluorine atoms ($N = 1,693$), the $dCor$ between res_homo and res_alpha increases dramatically to 0.636, compared to 0.286 in the remainder of the dataset. This difference of 0.350 is highly significant ($p < 0.001$, permutation test). Similarly, the coupling between res_homo and res_lumo is drastically enhanced in aromatic systems ($N = 23,643$), where the $dCor$ reaches 0.595, compared to a mere 0.097 in non-aromatic systems—a difference of 0.499 ($p < 0.001$). Furthermore, molecules with an above-median fraction of sp^2 carbons exhibit a moderate but significant increase in res_homo-res_alpha coupling ($dCor = 0.389$ vs 0.253, $p = 0.02$).

These rewiring effects are robust and not artifacts of sample size or statistical noise. Split-half stability testing within each structural partition yielded mean absolute errors in $dCor$ ranging from 0.033 to 0.066, which is an order of magnitude smaller than the observed differences for fluorine conjugation and aromaticity. As illustrated by the permutation tests in Figure 2, the observed differences in property coupling lie far outside the null distributions generated by randomly shuffling structural labels. These findings confirm that specific structural motifs act as chemical switchboards, locking the HOMO energy to the broader molecular response and the LUMO state in ways that are nearly non-existent in the general population of small organic molecules.

4 Conclusions

The electronic properties of small organic molecules are typically dominated by macroscopic size and stoichiometry. However, by isolating frontier orbital energies from these bulk effects, this study demonstrates that the highest occupied molecular orbital (HOMO) occupies a fundamentally distinct electronic manifold. While the LUMO energy retains a moderate dependence on standard topological descriptors, the HOMO

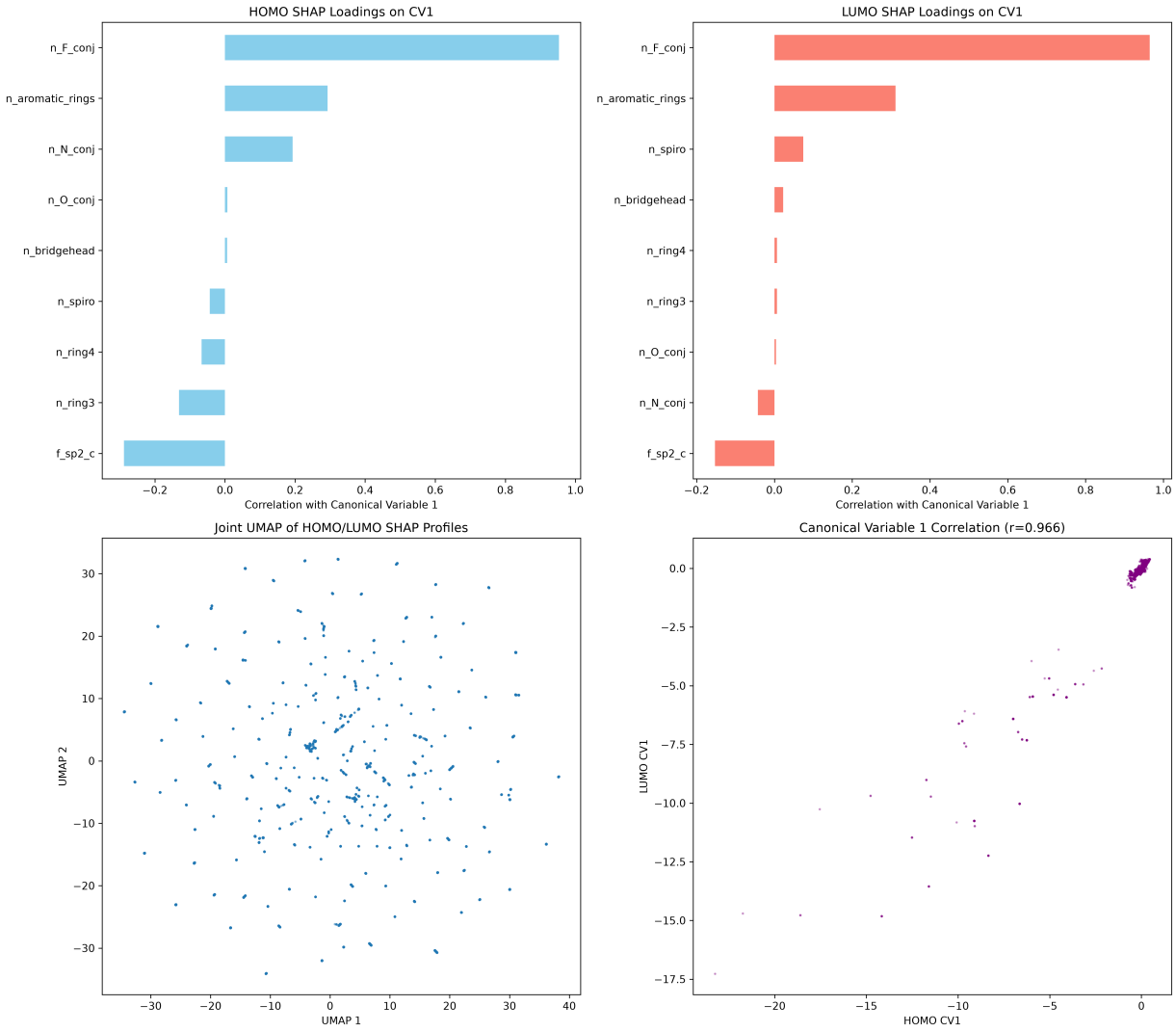

Figure 1: Joint structural drivers of frontier orbital residuals identified via Canonical Correlation Analysis (CCA) of SHAP values. The top panels display the feature loadings on the first canonical variable (CV1) for the HOMO (left) and LUMO (right) models, revealing that conjugated fluorine (`n_F_conj`) is the dominant structural motif coordinating both orbitals. The bottom left panel presents a joint UMAP projection of the HOMO and LUMO SHAP profiles, mapping the overall structural manifold. The bottom right panel illustrates the strong linear relationship between the HOMO and LUMO on CV1 ($r = 0.966$). Collectively, these results demonstrate that while frontier orbital energies are globally decoupled from bulk molecular size, their structural drivers are highly coordinated, with specific motifs like fluorine conjugation acting to jointly shift the frontier electronic manifold.

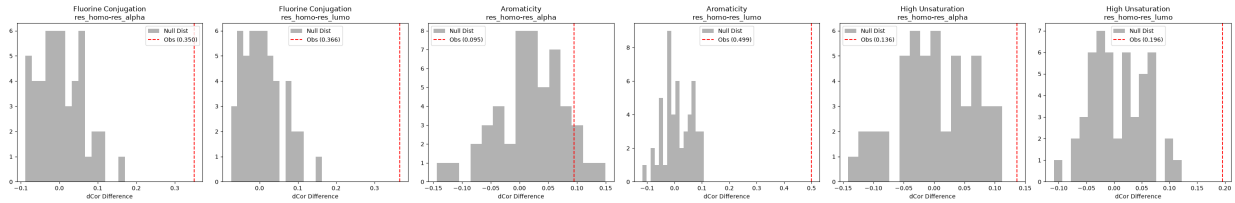

Figure 2: Permutation test results demonstrating the contextual rewiring of molecular property relationships across different structural families. The histograms display the null distributions of differences in distance correlation ($dCor$) for couplings between HOMO, LUMO, and isotropic polarizability residuals, generated by randomly shuffling structural labels. The red dashed lines indicate the observed $dCor$ differences for molecules characterized by fluorine conjugation, aromaticity, and high unsaturation. The observed differences—such as the 0.350 increase in HOMO-polarizability coupling for conjugated fluorine and the 0.499 increase in HOMO-LUMO coupling for aromatic systems—lie far outside their respective null distributions. This confirms that these specific structural motifs act as switchboards that significantly reorganize the fundamental relationships between frontier electronic states and geometric responses, demonstrating that the enhanced property couplings are statistically significant and not artifacts of sample size.

energy is globally decoupled, exhibiting a high sensitivity to fine-grained local bonding environments that evade simple structural prediction.

Despite this global decoupling, our decomposition of feature importances reveals a profound joint structural organization between the frontier orbitals. The structural drivers of the HOMO and LUMO residuals are not independent; rather, they are highly coordinated by a specific hierarchy of chemical motifs. The frontier electronic manifold is primarily organized along two axes: a dominant fluorine-conjugation axis and a secondary aromaticity and unsaturation axis.

Crucially, these structural hubs do more than merely shift orbital energies—they act as chemical switchboards that fundamentally rewire the relationships between electronic states and geometric response properties. In the broader chemical space of the QM9 dataset, the HOMO energy exhibits minimal coupling with isotropic polarizability or the LUMO state. Yet, the introduction of specific structural motifs drastically alters this behavior. The presence of conjugated fluorine effectively locks the HOMO energy to the broader molecular response, while aromatic systems strongly couple the HOMO and LUMO states.

These findings establish that the relationships between fundamental electronic properties are not universally static across chemical space. Instead, they are highly context-dependent and modulated by localized structural logic. Recognizing and modeling these contextual rewirings is essential for advancing our understanding of molecular property manifolds. It highlights the limitations of relying solely on global topological or stoichiometric descriptors in predictive chemical modeling, underscoring the need to account for the specific, local environments that dictate frontier orbital behavior.