Structural Drivers of HOMO Decoupling in the QM9 Electronic State Space

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In molecular property datasets, the Highest Occupied Molecular Orbital (HOMO) energy exhibits an unexplained decoupling from the tightly bound thermodynamic and spatial property manifold. To uncover the physical drivers of this isolation, we analyzed the QM9 dataset of 133,802 molecules by extracting non-linear, composition-independent property residuals using gradient-boosted trees. After confirming the statistical orthogonality of the HOMO residual through Principal and Independent Component Analysis, we quantified its isolation and trained an XGBoost regressor to predict this decoupling using a pruned suite of topological descriptors. We then deployed SHapley Additive exPlanations (SHAP) to perform a global sensitivity and interaction analysis across distinct molecular architectures. Our results reveal that Topological Polar Surface Area (TPSA) and maximum conjugation length are the primary, universally stable drivers of HOMO isolation. Furthermore, interaction analyses demonstrate that heteroatom-driven decoupling is strongly modulated by the conjugation of the carbon skeleton. Finally, family-specific partitioning shows that while acyclic and heteroatom-rich systems are predominantly governed by conjugation and polar surface area, geometric constraints—such as four-membered rings—emerge as secondary decoupling mechanisms in strained molecules.

1 Introduction

Large-scale molecular datasets, such as the QM9 database of 133,802 molecules, have fundamentally advanced our understanding of chemical spaces and accelerated the development of predictive machine learning models. Within these datasets, the majority of thermodynamic and spatial properties—such as internal energy, polarizability, and spatial extent—form a tightly bound, low-dimensional manifold. This global organization is heavily dominated by basic stoichiometric variables, molecular size, and baseline composition.

However, the Highest Occupied Molecular Orbital (HOMO) energy exhibits an unexplained decoupling from this tightly bound manifold. The HOMO energy uniquely resists global molecular property organization, demonstrating near-total statistical orthogonality to the thermodynamic-spatial triad. While the fundamental HOMO-LUMO gap and its implications for chemical reactivity are well-established in molecular orbital theory, the specific local structural motifs and topological constraints that drive the isolated variance of the HOMO energy across diverse chemical spaces remain poorly understood.

To uncover the physical drivers of this electronic isolation, this study shifts the focus from merely observing the orthogonality of the HOMO to discovering the structural mechanics that control it. We isolate genuine chemical signals by extracting non-linear, composition-independent property residuals using gradient-boosted trees. This stoichiometric baseline extraction explicitly retains geometrically strained molecules to preserve extreme electronic variations without introducing artificial artifacts. After confirming the statistical orthogonality of the HOMO residual through Principal and Independent Component Analysis (PCA and ICA), we quantify its isolation on a per-molecule basis.

To map the topological triggers that rewire a molecule's occupied electronic states, we train a parallelized XGBoost regressor to predict this quantified HOMO decoupling directly from a comprehensive, rigorously pruned suite of topological descriptors. These descriptors capture key electronic environments, including conjugation length, ring strain, and heteroatom connectivity. By deploying SHapley Additive exPlanations (SHAP), we perform a global sensitivity and interaction analysis to extract precise feature importances and identify which specific local structural motifs govern the orthogonal behavior of the HOMO.

As highlighted by our primary global sensitivity results (Table 1), Topological Polar Surface Area (TPSA) and maximum conjugation length emerge as the dominant, universally stable drivers of HOMO isolation.

Beyond these primary main effects, our interaction analyses demonstrate that heteroatom-driven decoupling is strongly modulated by the conjugation of the carbon skeleton. Finally, by partitioning the dataset into distinct structural families, we show that these structural rules are not entirely static; while acyclic and heteroatom-rich systems are predominantly governed by conjugation and polar surface area, geometric constraints—such as four-membered rings—emerge as secondary decoupling mechanisms in highly strained molecules. Through this approach, we provide a mathematical and structural framework for understanding how specific architectural constraints force the HOMO to detach from, or reintegrate into, the global physical state space.

2 Methods

2.1 Data preparation and stoichiometric baseline extraction

To isolate genuine chemical signals from trivial size-dependent scaling, we first extracted non-linear, composition-independent property residuals. The QM9 dataset was loaded and deduplicated based on canonical SMILES strings, yielding 133,802 unique molecules. We isolated 12 non-redundant numeric properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), orbital energies (ϵ_{HOMO}, ϵ_{LUMO}), electronic spatial extent ($\langle R^2 \rangle$), zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity (C_v), and atomization energy (U_0^{atom}).

To handle the extreme skewness of the rotational constant A, a $\log(1+x)$ transformation was applied. A binary feature, *is_linear*, was created to explicitly flag the nine linear molecules in the dataset (where $A = 0$). For these molecules, the transformed A value was left as zero rather than applying mean imputation, allowing downstream models to learn the structural offset naturally without introducing artifacts.

SMILES strings were parsed using RDKit to explicitly add hydrogens and compute exact stoichiometric counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. A multi-output, parallelized XGBoost regressor was then trained using only the five atom counts and the *is_linear* flag to predict the 12 raw properties. These baseline predictions were subtracted from the raw values to yield 12 non-linear, composition-independent residuals (e.g., res_{homo}, res_{u0}).

2.2 Latent manifold construction and isolation quantification

To evaluate the statistical orthogonality of the HOMO energy, the 12 residuals were standardized using a robust scaler. We performed Principal Component Analysis (PCA) and Independent Component Analysis (ICA) on the standardized residuals to identify the latent manifold structure. We specifically examined the loading vectors to confirm the emergence of a "HOMO-orthogonal component"—an independent source where res_{homo} exhibited a dominant loading while thermodynamic and spatial residuals maintained near-zero loadings. To ensure this orthogonal axis was a stable global feature rather than an artifact of specific property inclusions, sensitivity checks were performed by repeating the PCA and ICA on subsets of the residuals (e.g., systematically excluding geometric-response or thermodynamic properties).

To quantify this decoupling on a per-molecule basis, we trained a strictly constrained, shallow XGBoost regressor (maximum depth of 3, high L2 regularization λ) to predict res_{homo} from the remaining 11 property residuals. The absolute error of this prediction was defined as the *homo_isolation_score*:

$$homo_isolation_score_i = |res_{\text{homo},i} - f(res_{\text{other},i})| \tag{1}$$

where f represents the shallow regressor. This score explicitly represents the residual variance of the HOMO energy that cannot be explained by the global property manifold. Permutation feature importance was calculated for the 11 input residuals to model-agnostically confirm the decoupling. A control isolation score was generated using res_{lumo} following the identical procedure for comparative baseline analysis.

2.3 Topological feature engineering and pruning

To map the physical drivers of HOMO isolation, we generated a targeted suite of 2D structural descriptors from the SMILES strings using RDKit, focusing strictly on topology and the electronic environment. Extracted features included exact ring counts by size (3-, 4-, 5-, and $\geq$6-membered rings), the fraction of sp^2 and sp^3 hybridized carbons, the number of rotatable bonds, Topological Polar Surface Area (TPSA), exact heteroatom counts, and conjugation indicators such as the maximum conjugated chain length.

To prevent feature attribution instability and signal splitting during downstream SHAP analysis, we aggressively pruned highly collinear features. We iteratively calculated the Variance Inflation Factor (VIF) and dropped redundant descriptors (e.g., removing heteroatom ratios in favor of TPSA and exact counts) until all remaining features possessed a VIF < 5. The final pruned descriptor set was then standardized. Additionally, the Murcko scaffold was computed for each molecule to facilitate structurally rigorous cross-validation.

2.4 Predictive modeling of HOMO isolation

A predictive modeling task was established using the *homo_isolation_score* as the target variable. To ensure the model learned generalizable chemical rules rather than memorizing specific structural families, the dataset was split into an 80% training set and a 20% holdout set using Group K-Fold cross-validation, grouped by the Murcko scaffold ID. A parallelized XGBoost regressor was trained to predict the isolation score from the pruned topological descriptors. Hyperparameter tuning was conducted via cross-validation strictly within the training set. The final model was evaluated on the scaffold-holdout set, with performance quantified via R^2 and Root Mean Square Error (RMSE).

2.5 Global sensitivity and interaction analysis

To extract the precise structural motifs governing HOMO decoupling, we deployed `shap.TreeExplainer` on the trained XGBoost model using the `feature_perturbation="interventional"` setting. SHAP (SHapley Additive exPlanations) values were computed for the entire dataset to determine the main structural drivers.

To statistically confirm that these drivers represent universal chemical rules rather than dataset-specific quirks, a split-half stability test was performed. The dataset was randomly partitioned into two equal halves, and the mean absolute SHAP values were computed independently for both. The Pearson correlation coefficient between the feature importance rankings of the two halves was calculated to verify stability.

Furthermore, to map how structural features synergistically drive HOMO decoupling, we computed the full SHAP interaction values matrix. To ensure computational tractability and stability, this was performed on a representative background subsample of 5,000 molecules, stratified by structural family and the *homo_isolation_score* distribution. We explicitly extracted and analyzed the off-diagonal interaction terms between key electronic descriptors, specifically focusing on the interplay between conjugation length and heteroatom content (TPSA).

2.6 Family-specific reintegration analysis

Finally, to test if the structural mechanics governing the HOMO shift under different architectural constraints, the dataset was partitioned into three distinct structural families:

1. **Acyclic molecules:** Purely linear or branched systems with no ring structures.
2. **Strained molecules:** Systems containing at least one 3- or 4-membered ring.
3. **Heteroatom-rich molecules:** Molecules in the top quartile of total heteroatom count.

For each sub-population, the mean absolute SHAP values and key interaction effects were recalculated. Bootstrap confidence interval estimation was performed for the SHAP importances between these subsets to rigorously determine whether the structural drivers of HOMO decoupling remain universal or significantly shift in response to specific architectural constraints.

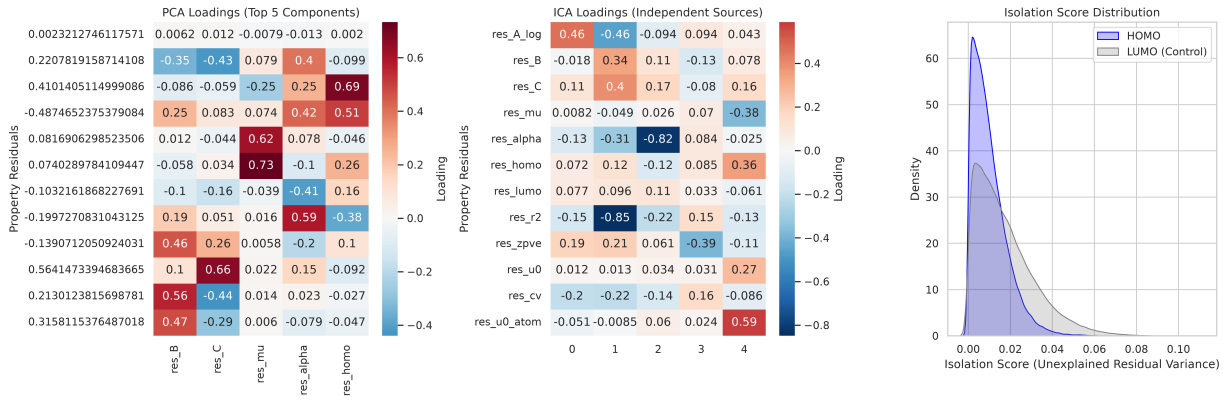

Figure 1: The figure illustrates the statistical isolation and structural decoupling of the Highest Occupied Molecular Orbital (HOMO) energy residual (res_{homo}) from the global property manifold in the QM9 dataset. The left and middle panels display the Principal Component Analysis (PCA) and Independent Component Analysis (ICA) loadings, respectively, demonstrating that res_{homo} remains largely orthogonal to tightly bound thermodynamic and spatial properties. In these linear decompositions, res_{homo} dominates specific isolated components while exhibiting negligible overlap with major properties like internal energy (res_{u0}) and spatial extent (res_{r2}). The right panel presents the distribution of the isolation score—defined as the unexplained residual variance when predicted by the other eleven property residuals—comparing HOMO to a LUMO control. This distribution visually confirms the modeling result that over 83% of the variance in res_{homo} cannot be recovered from the other properties, demonstrating that the HOMO energy is governed by a distinct set of physical drivers and effectively detaches from the global thermodynamic-spatial triad.

3 Results

3.1 The global manifold and HOMO orthogonality

Following the extraction of composition-independent property residuals, initial analysis of the 133,802 unique molecules in the QM9 dataset confirmed that while thermodynamic and spatial properties form a tightly bound manifold, the HOMO energy residual (res_{homo}) remains largely orthogonal to this structure. Principal Component Analysis (PCA) on the twelve non-redundant residuals revealed that the first two principal components explain only approximately 55% of the variance, a decrease from the 63% observed in the raw, unconditioned data. This reduction confirms that the apparent low-dimensional structure in QM9 is partially driven by baseline molecular size. Within the PCA loadings, res_{homo} exhibits its highest loading in the third principal component (0.688), where it emerges as the dominant feature.

Independent Component Analysis (ICA) further clarifies this isolation. As illustrated in , res_{homo} isolates as a near-singleton in the sixth independent component (loading $= -0.780$). In contrast, other major properties such as internal energy (res_{u0}) and spatial extent (res_{r2}) exhibit negligible loadings on this axis (0.112 and -0.017, respectively). This statistical independence confirms that the variance in HOMO energy is governed by a distinct set of physical drivers compared to the thermodynamic-spatial triad.

To quantify this decoupling on a per-molecule basis, the *homo_isolation_score* was computed using a constrained XGBoost regressor trained to predict res_{homo} from the other eleven property residuals. The model achieved a low R^2 of 0.164, indicating that over 83% of the variance in res_{homo} cannot be recovered from the global manifold (, right panel). Permutation feature importance analysis on this model showed that polarizability (res_{α}, importance $= 0.853$) and the second rotational constant (res_{B}, importance $= 0.483$) are the most related features; however, they ultimately fail to collapse the isolation of the HOMO energy.

3.2 Topological drivers of electronic decoupling

To explain the physical mechanisms driving this isolation, we modeled the *homo_isolation_score* using a pruned suite of topological descriptors. Global SHAP (SHapley Additive exPlanations) analysis was deployed

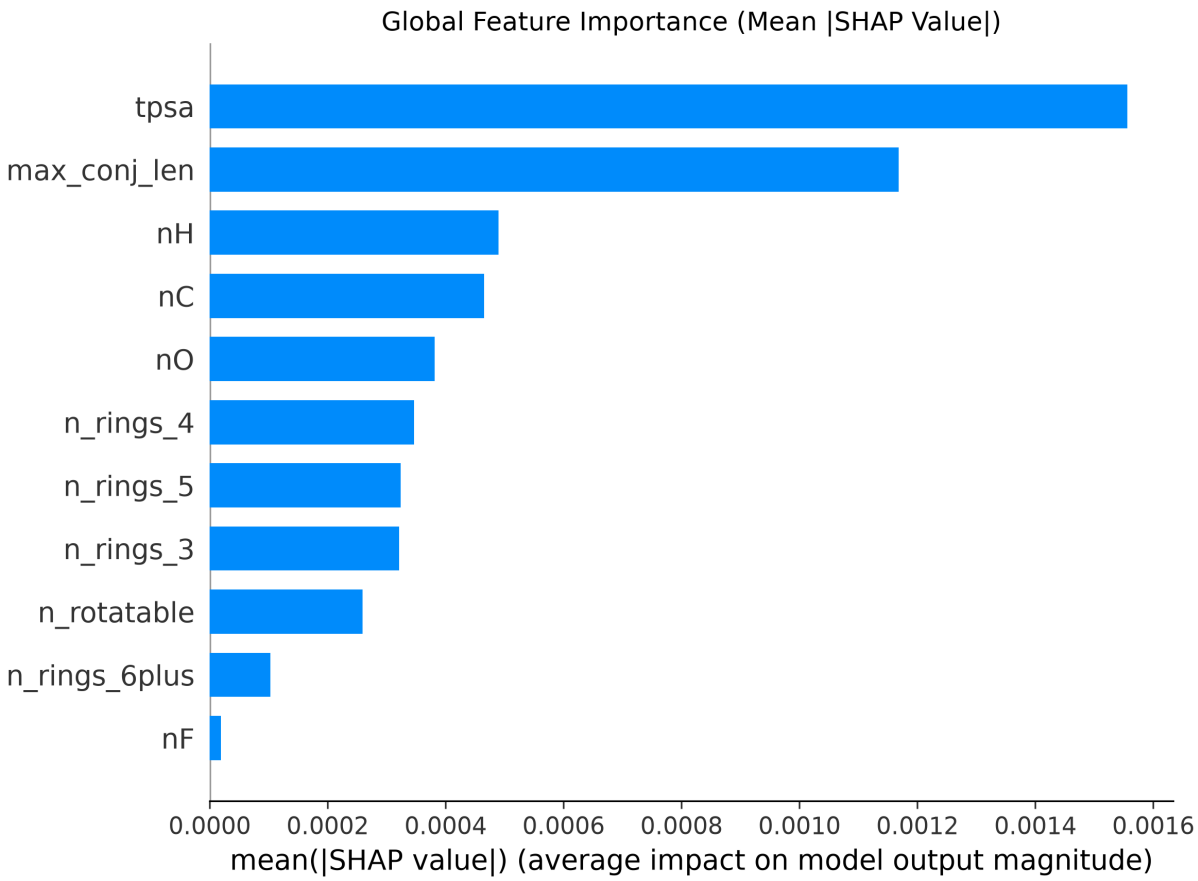

Figure 2: Global feature importance for predicting the HOMO isolation score, measured by mean absolute SHAP values. The plot demonstrates that Topological Polar Surface Area (TPSA) and maximum conjugation length are the dominant topological drivers of HOMO decoupling. This indicates that the isolation of the HOMO energy from the global thermodynamic and spatial manifold is primarily governed by heteroatom density and electronic delocalization, while atomic counts (n_H, n_C) and ring structures exert only a secondary influence on the detachment of the electronic state space.

Table 1: Mean absolute SHAP values for the primary topological features driving HOMO isolation.

Feature	Mean Absolute SHAP Value
TPSA	0.001556
Max Conjugation Length	0.001168
Hydrogen Count (n_H)	0.000490
Carbon Count (n_C)	0.000465
4-Membered Rings (n_rings_4)	0.000346

to identify the specific structural motifs that control whether the HOMO energy remains decoupled or integrates into the global manifold.

The SHAP main effects (and) reveal that Topological Polar Surface Area (TPSA) and maximum conjugation length are the primary drivers of HOMO isolation, significantly outweighing basic stoichiometric counts and ring features.

To ensure these topological rules are universal across the dataset rather than artifacts of specific sampling, a split-half stability test was conducted. As shown in , the feature importances are exceptionally stable, yielding a Pearson correlation of $R = 0.99999$ ($p = 4.08 \times 10^{-22}$) between the two independent dataset

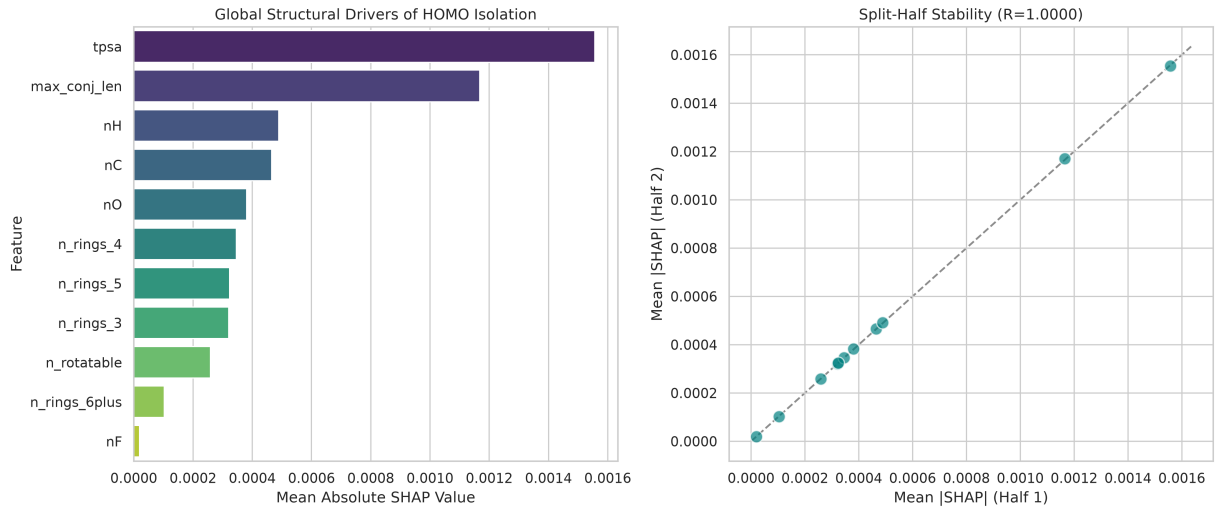

Figure 3: Global SHAP analysis of topological descriptors driving HOMO isolation. The left panel displays the mean absolute SHAP values, identifying Topological Polar Surface Area (TPSA) and Maximum Conjugation Length as the primary structural features controlling the decoupling of the HOMO energy from the global property manifold, followed by atomic counts (n_H, n_C) and small ring motifs. The right panel illustrates the split-half stability of these feature importances, demonstrating a near-perfect correlation ($R = 1.0000$). This extreme stability confirms that the topological rules governing electronic isolation are universal across the dataset and not artifacts of sampling.

halves.

Beyond main effects, SHAP interaction analysis revealed that the decoupling effect of heteroatoms is strongly modulated by the conjugation of the carbon skeleton (, left panel). The top interaction pair was TPSA × Carbon Count (n_C) with a mean absolute interaction value of 0.000543, closely followed by TPSA × Maximum Conjugation Length (0.000311). This synergy suggests that the isolation of the HOMO is most pronounced in heteroatom-rich systems where conjugation allows for extensive electronic delocalization. In such systems, the HOMO energy is rewired by the extended π-system, further distancing it from the local geometric and thermodynamic constraints that bind the other properties.

3.3 Family-specific reintegration

While TPSA and conjugation length are global drivers, partitioning the dataset into distinct structural families reveals that the structural mechanics of HOMO decoupling shift under different architectural constraints (, right panel).

In purely acyclic systems (linear or branched molecules with no rings), TPSA (mean absolute SHAP = 0.001545) and maximum conjugation length (0.001222) remain the dominant drivers. The absence of ring strain allows these purely electronic features to dictate the isolation score without interference from geometric frustration.

However, in highly strained systems containing 3- or 4-membered rings, the importance of 4-membered rings rises significantly as a decoupling driver (mean absolute SHAP = 0.000461). In these molecules, ring strain acts as a secondary mechanism that perturbs the HOMO energy, forced by geometric constraints that are not adequately captured by the thermodynamic-spatial triad.

Finally, in heteroatom-rich systems, maximum conjugation length (0.001467) overtakes TPSA (0.001146) as the primary driver of isolation. This inversion indicates that once a critical heteroatom threshold is met, the spatial extent of the conjugated system becomes the limiting factor in determining the HOMO's electronic decoupling. Together, these family-specific shifts demonstrate that while the HOMO energy is uniquely sensitive to the topological connectivity of the π-system, local geometric strain can force secondary pathways of electronic isolation.

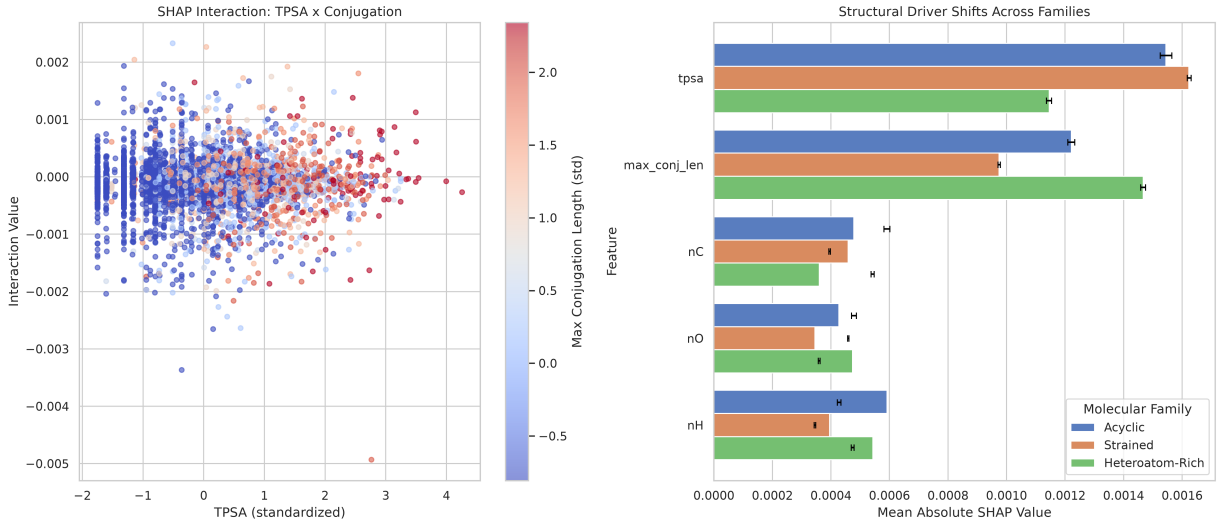

Figure 4: The figure illustrates the topological drivers of HOMO energy decoupling. The left panel displays the SHAP interaction between Topological Polar Surface Area (TPSA) and maximum conjugation length, revealing that HOMO isolation is strongly driven by the synergy between heteroatom density and extended π-system delocalization. The right panel compares the mean absolute SHAP values of structural features across Acyclic, Strained, and Heteroatom-Rich molecular families. This demonstrates that the structural mechanics of decoupling shift based on molecular context: while TPSA is the primary driver in acyclic and strained systems, maximum conjugation length overtakes it as the dominant factor in heteroatom-rich molecules. Ultimately, we learn that once a heteroatom threshold is met, the extent of the conjugated system becomes the critical factor in determining the electronic decoupling of the HOMO from the global thermodynamic manifold.

4 Conclusions

In this study, we investigated the unexplained statistical decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from the tightly bound thermodynamic and spatial property manifold in the QM9 dataset. By extracting non-linear, composition-independent property residuals and quantifying the isolation of the HOMO energy, we successfully mapped the topological triggers responsible for this phenomenon using gradient-boosted trees and SHAP sensitivity analysis.

Our results demonstrate that the decoupling of the HOMO energy is a highly stable, universal chemical rule rather than a dataset-specific artifact. Topological Polar Surface Area (TPSA) and maximum conjugation length emerged as the primary structural drivers of this isolation. Furthermore, interaction analyses revealed a strong synergy between heteroatom density and the conjugation of the carbon skeleton. In heteroatom-rich systems, the extended π-system facilitates electronic delocalization, effectively rewiring the occupied electronic states and distancing the HOMO energy from the local geometric and thermodynamic constraints that bind other molecular properties.

Crucially, we established that the structural mechanics governing this decoupling adapt to specific architectural constraints. While acyclic systems are predominantly controlled by TPSA and conjugation, the introduction of geometric frustration—such as in molecules with three- or four-membered rings—forces secondary pathways of electronic isolation driven by local ring strain. Conversely, in heteroatom-rich molecules, the spatial extent of the conjugated system overtakes TPSA as the limiting factor for HOMO decoupling.

Ultimately, this study provides a rigorous mathematical and structural framework for understanding the isolated variance of the HOMO energy. While most molecular properties organize around baseline size and spatial extent, the HOMO energy is uniquely sensitive to the topological connectivity of the π-system and local geometric strain. These insights shift the perspective from merely observing electronic orthogonality to understanding the fundamental structural mechanics that control it, offering valuable implications for the design of targeted electronic materials and the development of more robust, physics-aware molecular machine learning models.