Partial Distance Correlation Mapping of HOMO Re-entanglement via Local Structural Motifs

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In large-scale quantum chemical datasets, the energy of the highest occupied molecular orbital (HOMO) often appears statistically orthogonal to geometric and vibrational properties once stoichiometric baselines are removed. It remains unclear whether this global decoupling is an intrinsic physical invariant or a localized artifact that collapses under specific structural conditions. To address this, we investigate the localized re-entanglement of HOMO energy with the geometric-vibrational core using the QM9 dataset. After removing stoichiometric determinism via Ridge regression and applying Minimum Covariance Determinant filtering to the residual space, we compute local distance correlations within structural neighborhoods defined by UMAP embeddings of Morgan fingerprints and explicit ring counts. We then utilize XGBoost regressors and SHAP feature importance to identify the chemical architectures driving localized coupling. While global baselines confirm that HOMO residuals are largely decoupled from zero-point vibrational energy, heat capacity, and spatial extent, local non-parametric distance correlation analysis reveals that approximately four percent of structural neighborhoods exhibit high coupling. SHAP analysis demonstrates that this re-entanglement is systematically triggered by specific local atom-centered environments, predominantly nitrogen- and oxygen-containing heterocycles and associated ring strain. These findings indicate that the statistical orthogonality of HOMO is a global population property rather than a strict physical invariant, highlighting how specific structural motifs conditionally re-integrate frontier electronic behavior with macroscopic thermodynamic and spatial properties.

1 Introduction

Large-scale quantum chemical datasets have provided unprecedented opportunities to map the relationships between molecular structure and quantum mechanical properties. However, the variance in these datasets is often overwhelmingly dominated by stoichiometric baselines. Properties such as total energy, zero-point vibrational energy ($zpve$), and heat capacity (c_v) scale almost deterministically with molecular size and atomic composition. To uncover the intrinsic chemical logic governing molecular behavior, it is necessary to remove these macroscopic baseline effects and examine the residual property space.

Within this composition-adjusted residual space, a persistent joint structure typically emerges among geometric and vibrational properties. Yet, the frontier electronic properties—most notably the energy of the highest occupied molecular orbital (HOMO)—frequently appear statistically orthogonal to this geometric-vibrational core. Global correlation metrics suggest that HOMO energy varies independently of spatial extent (r^2), $zpve$, and c_v. This raises a fundamental question: is the statistical isolation of the HOMO an intrinsic physical invariant of small organic molecules, or is it a global population artifact that collapses under specific, localized structural conditions?

To address this, we investigate the conditional re-entanglement of HOMO energy with the geometric-vibrational core across the QM9 dataset. We hypothesize that the apparent global decoupling masks localized regions of chemical space where specific architectural constraints force the electronic frontier to re-integrate with macroscopic thermodynamic and spatial properties.

Our approach isolates these structural triggers by combining robust statistical filtering with non-parametric manifold analysis. We first remove stoichiometric determinism via Ridge regression and apply Minimum Covariance Determinant (MCD) filtering to the residual space to ensure that any observed coupling is not a

spurious artifact of latent density functional theory (DFT) convergence failures. We then define local structural neighborhoods using UMAP embeddings of Morgan fingerprints and explicit heteroatom ring counts. By computing local distance correlations within these neighborhoods, we bypass linear assumptions to map the exact regions where HOMO re-entangles with the $zpve$-c_v-r^2 core. Finally, we deploy XGBoost regressors and SHAP (SHapley Additive exPlanations) feature importance to extract the specific chemical motifs responsible for this localized coupling.

Through this framework, we demonstrate that the statistical orthogonality of HOMO is not a strict physical invariant but a conditional property. While global baselines confirm decoupling, our local distance correlation analysis reveals that a distinct subset of structural neighborhoods exhibits high coupling. Furthermore, we show that this re-entanglement is systematically driven by specific local atom-centered environments—predominantly nitrogen- and oxygen-containing heterocycles and their associated ring strain. These findings highlight how specific structural motifs can conditionally rewire frontier electronic behavior, bridging the gap between localized quantum states and global thermodynamic properties.

2 Methods

2.1 Dataset preparation and baseline removal

To investigate the conditional re-entanglement of frontier orbitals, we utilized the QM9 dataset, which provides a comprehensive mapping of quantum mechanical and thermodynamic properties for small organic molecules. Initial preprocessing involved deduplicating the dataset and dropping redundant thermodynamic columns to isolate 12 non-redundant target properties. The lowest unoccupied molecular orbital (LUMO) energy was retained as a control variable. Additionally, a binary indicator, is_linear, was introduced to explicitly flag linear molecules (where the rotational constant $A = 0$).

Because macroscopic properties such as total energy (u_0), zero-point vibrational energy ($zpve$), and heat capacity (c_v) scale almost deterministically with molecular size, we removed these stoichiometric baselines to prevent structural overfitting and to preserve non-linear chemical variance. We applied Ridge regression using explicit atom counts (C, N, O, F, H) as independent variables to predict the 12 properties. The resulting residual vectors—such as res_homo, res_zpve, res_c_v, and res_r^2—represent the property variance conditioned on atomic composition, effectively isolating the intrinsic chemical and structural effects from global stoichiometric determinism.

2.2 Robust filtering and structural representation

To ensure that any observed structural re-entanglement was not a spurious artifact of latent density functional theory (DFT) convergence failures or extreme outliers, we applied Minimum Covariance Determinant (MCD) filtering to the 12-dimensional residual space. This robust statistical filtering step isolates the core distribution of chemically well-behaved molecules.

Following the isolation of the residual space, we constructed explicit structural representations for each molecule to map the local chemical environments. We generated 2048-bit Morgan fingerprints to capture local atom-centered environments and computed explicit heteroatom-specific ring counts (including total rings, as well as nitrogen-, oxygen-, and fluorine-specific ring counts). To define local structural neighborhoods without succumbing to the curse of dimensionality, we applied Uniform Manifold Approximation and Projection (UMAP). For memory efficiency and computational scaling, the UMAP manifold was fitted on a representative subset of 50,000 molecules before transforming the full dataset into the reduced structural embedding.

2.3 Global and local distance correlation analysis

To establish a baseline for the statistical orthogonality of the frontier orbitals, we first computed global distance correlations ($dCor$) between the residual frontier orbital energies (res_homo, res_lumo) and the tightly bound geometric-vibrational core (res_zpve, res_c_v, res_r^2). Distance correlation was selected over Pearson correlation because it captures both linear and non-linear dependencies, evaluating to zero if and only if the variables are strictly independent.

To test the hypothesis that HOMO orthogonality collapses under specific structural conditions, we mapped the localized re-entanglement using a non-parametric Partial Distance Correlation approach. Within the UMAP structural embedding, we defined local chemical neighborhoods using a k-nearest neighbors algorithm with $k = 250$. For each neighborhood, we computed the local distance correlation between res_homo and the geometric-vibrational core (e.g., res_c_v). This localized approach bypasses global linear regression artifacts, allowing us to identify the exact regions of the structural manifold where the electronic frontier re-integrates with macroscopic thermodynamic and spatial properties.

2.4 Feature importance and motif extraction

To pinpoint the exact chemical architectures driving the localized coupling, we treated the local distance correlation (specifically between res_homo and res_c_v) as a target variable for supervised learning. We trained XGBoost regressors to predict this local coupling strength directly from the 2048-bit Morgan fingerprints and explicit ring counts.

Once the models were trained, we extracted SHAP (SHapley Additive exPlanations) values to determine feature importance. SHAP analysis provides a rigorous, game-theoretic quantification of how each specific structural motif (e.g., a specific nitrogen lone-pair configuration or constrained oxygen heterocycle) contributes to the breakdown of HOMO isolation. Finally, to ensure that the identified motifs were robust and independent of heavy-atom proxies, we validated the structural triggers using split-half resampling. The stability of the extracted motifs was confirmed by computing the Spearman rank correlation (ρ) of the SHAP feature importance rankings across the independent data subsets.

3 Results

3.1 Global decoupling of frontier orbitals

As outlined in our methodology, we first evaluated the extent to which macroscopic properties scale deterministically with molecular size and composition. Linear regression on explicit atom counts (C, N, O, F, H) yielded near-deterministic fits for thermodynamic properties, with an R^2 of 0.99999 for total energy (u_0), 0.99714 for zero-point vibrational energy ($zpve$), and 0.70272 for heat capacity (c_v). In contrast, frontier orbital energies exhibited significantly less stoichiometric dependence; the HOMO energy showed an R^2 of only 0.08547, while the LUMO energy was moderately constrained at 0.45852.

After removing these stoichiometric baselines, we analyzed the residual property space to isolate intrinsic chemical variance. While a persistent joint structure was observed among the geometric-vibrational core properties (res_zpve, res_cv, and res_r^2), the HOMO residuals (res_homo) remained globally decoupled from this core. Global distance correlation ($dCor$) analysis yielded low coupling metrics for HOMO against $zpve$ ($dCor = 0.23459$), c_v ($dCor = 0.19327$), and r^2 ($dCor = 0.23999$). Although LUMO residuals exhibited slightly higher global coupling (e.g., $dCor = 0.27886$ with res_zpve), both frontier electronic levels maintained a distinct axis of variation orthogonal to the spatial and thermodynamic organization of the general molecular population (Figure 1).

3.2 Localized re-entanglement mapping

To test whether this global orthogonality is a strict physical invariant or a population-level artifact, we mapped the statistical relationship between HOMO and the geometric-vibrational core across local structural neighborhoods. These neighborhoods were defined using a k-nearest neighbors approach ($k = 250$) within a UMAP embedding of the structural space (Figure 2).

The local distance correlation analysis revealed that the decoupling of HOMO is not uniform across chemical space. While the mean local $dCor$ for HOMO versus c_v was low (0.24119), the distribution exhibited a pronounced long rightward tail (Figure 3). Approximately 4.05% of the 15,000 sampled structural neighborhoods demonstrated high coupling ($dCor > 0.50$), with the maximum local correlation reaching 0.83600. This indicates that within specific, topologically localized chemical domains, the electronic frontier loses its independence and becomes tightly bound to the factors governing the molecule's heat capacity and vibrational modes.

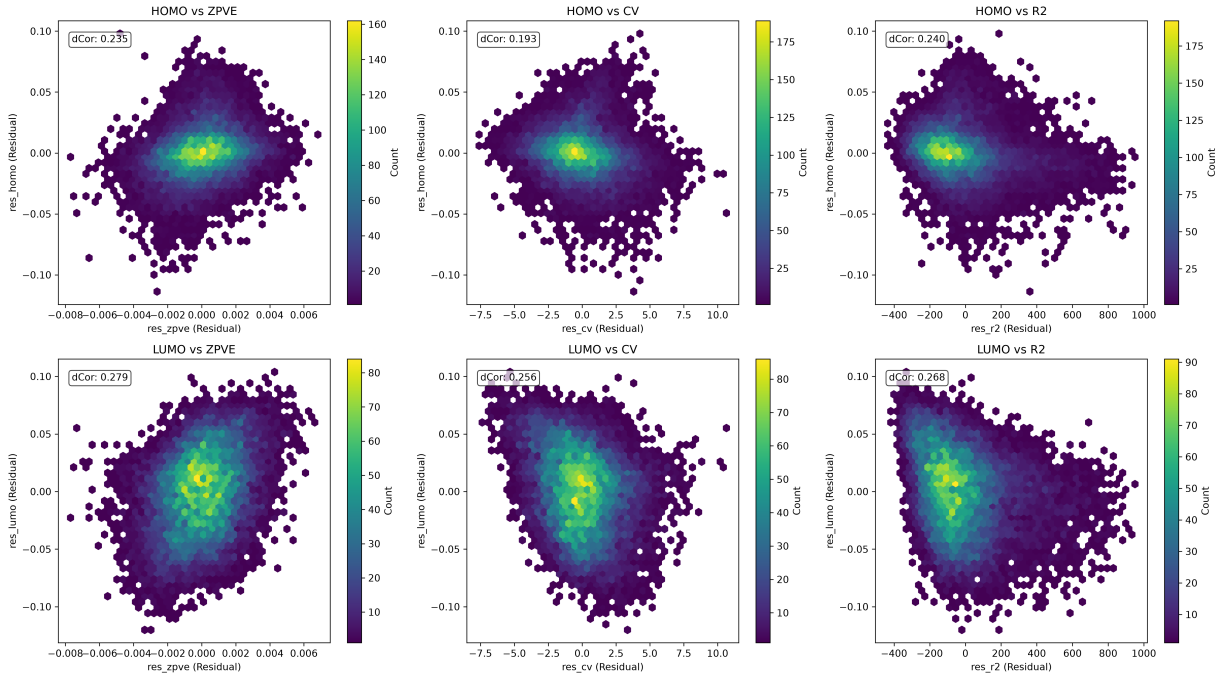

Figure 1: Global distribution of frontier orbital residuals (`res_homo` and `res_lumo`) plotted against geometric-vibrational core residuals (`res_zpve`, `res_cv`, and `res_r2`) after removing stoichiometric effects. The hexbin scatter plots and corresponding distance correlation ($dCor$) values demonstrate the global decoupling of frontier orbitals from thermodynamic and spatial properties. While LUMO residuals exhibit slightly higher global coupling than HOMO, both electronic levels maintain a distinct axis of variation that is largely orthogonal to the geometric-vibrational core ($zpve$, c_v, r^2) across the general molecular population.

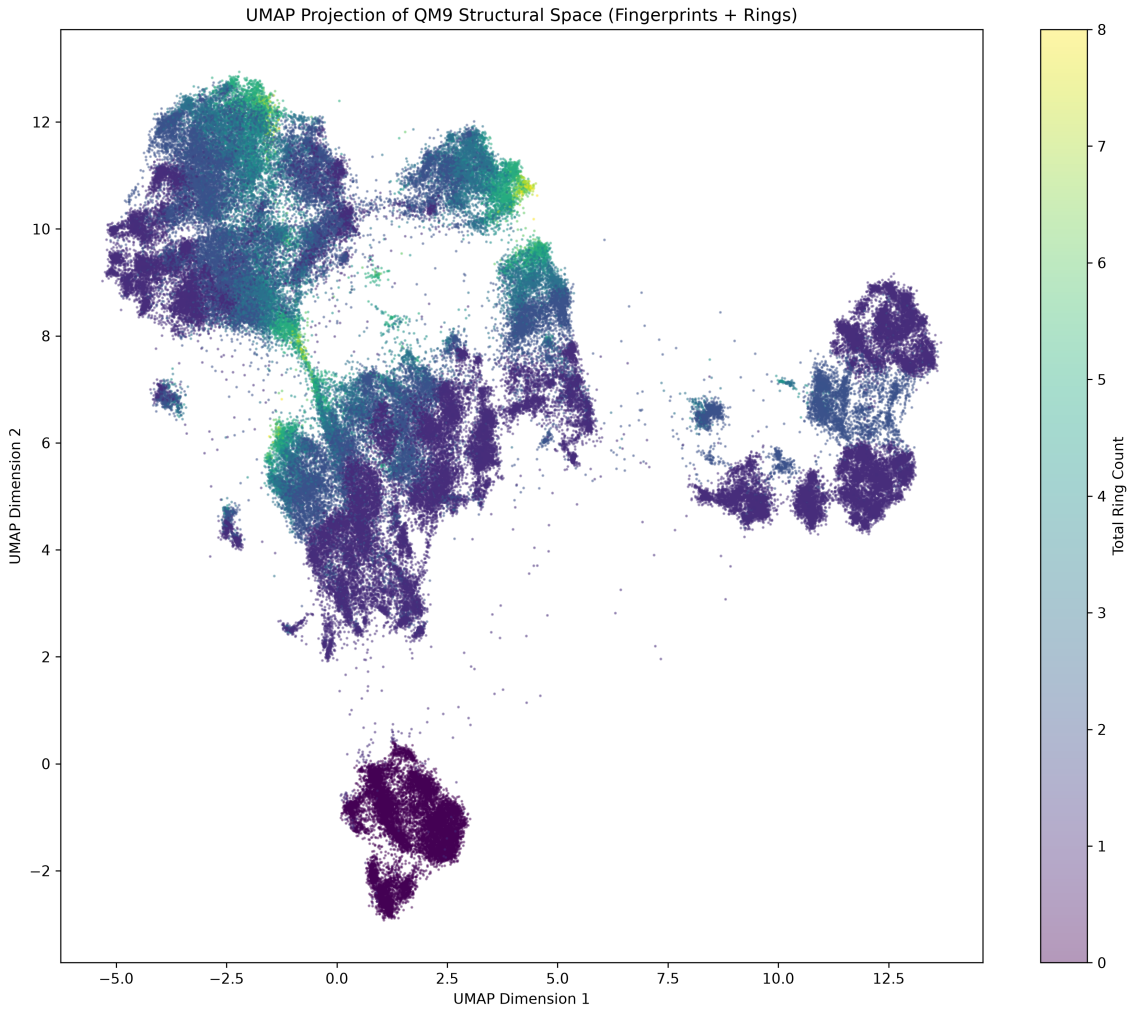

Figure 2: UMAP projection of the QM9 structural space based on Morgan fingerprints and explicit ring counts, colored by the total ring count. This embedding establishes the local structural neighborhoods used to analyze the distance correlation between frontier orbitals and the geometric-vibrational core. The visualization demonstrates that molecules cluster distinctly according to their `ring_count`, a feature identified as a primary architectural trigger for localized HOMO re-entanglement. From this spatial organization, we learn that the specific chemical domains where the globally decoupled HOMO energy becomes non-linearly entangled with thermodynamic properties, such as heat capacity (c_v) and zero-point vibrational energy ($zpve$), are topologically localized and strongly associated with regions of high ring density.

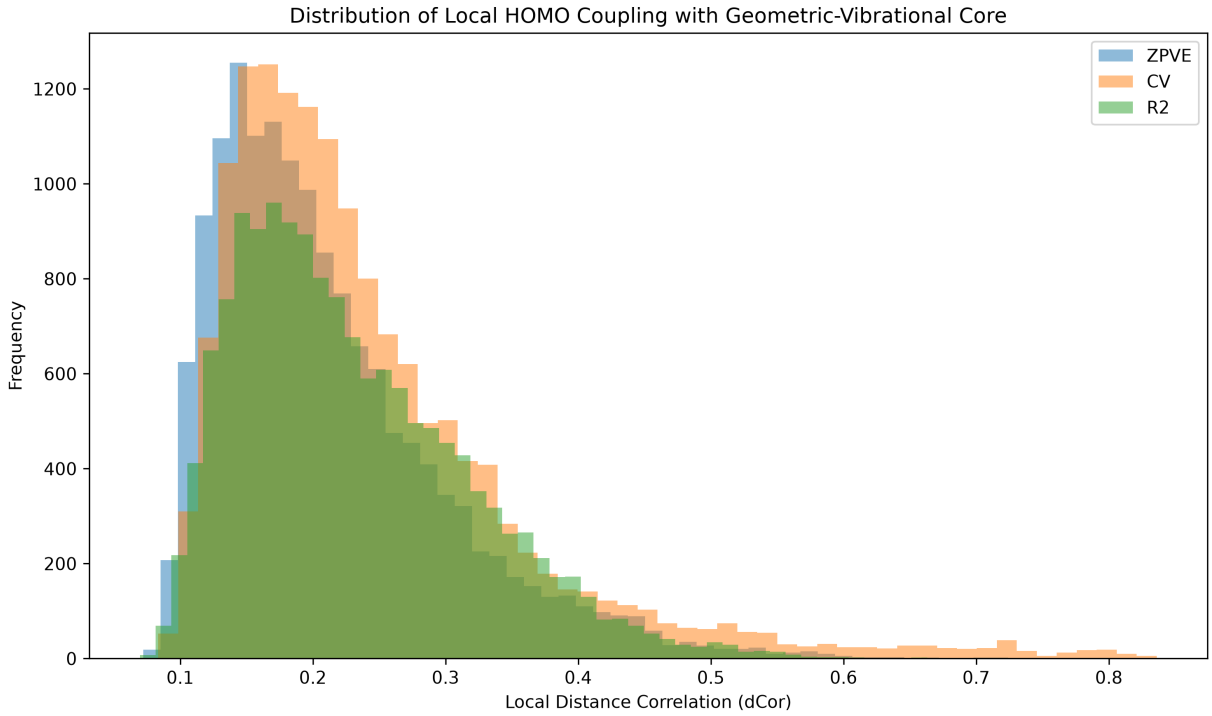

Figure 3: Distribution of local distance correlations ($dCor$) between the HOMO energy and properties of the geometric-vibrational core ($zpve$, c_v, and r^2) across sampled structural neighborhoods. Although HOMO is globally decoupled from these properties, the long rightward tail of the distributions illustrates localized re-entanglement. Notably, while the mean local $dCor$ for HOMO versus c_v is low, approximately 4.05% of the structural neighborhoods exhibit high coupling ($dCor > 0.50$), with maximum correlations reaching 0.83600. This long-tailed distribution demonstrates that for specific molecular architectures, the electronic frontier is no longer independent, but instead becomes tightly bound to the factors governing the molecule's heat capacity and vibrational modes.

3.3 Structural triggers of HOMO re-entanglement

Having established the existence of localized re-entanglement, we sought to identify the specific architectural causes. An XGBoost regressor trained to predict the local coupling strength between HOMO and c_v from 2048-bit Morgan fingerprints and explicit ring counts achieved an R^2 of 0.7780. This strong predictive performance confirms that the re-entanglement is a systematic function of molecular structure rather than random noise.

SHAP feature importance analysis (Figure 4) isolated the specific motifs driving this localized coupling. The most influential features were local atom-centered environments rather than global stoichiometric proxies. Specific molecular fragments, represented by Morgan fingerprint bits `mfp_378` (Rank 1, mean absolute SHAP 0.01490), `mfp_935` (Rank 2, 0.00981), and `mfp_1019` (Rank 3, 0.00945), emerged as the top three most important features. Furthermore, heteroatom ring strain proved to be a primary driver of re-entanglement. The total ring count (`ring_count`, Rank 7, mean absolute SHAP 0.00409) and nitrogen-specific ring count (`n_ring_count`, Rank 8, 0.00324) were highly influential. Oxygen heterocycles (`o_ring_count`) also contributed to this effect (Rank 20, 0.00137), whereas fluorine substitution (`f_ring_count`) showed zero importance (Rank 1643). These results suggest that nitrogen and oxygen heterocycles—likely through the interaction of lone pairs with strained ring orbitals—are the functional triggers that rewire the relationship between frontier orbitals and the vibrational core.

3.4 Stability and robustness of structural motifs

To ensure that the identified structural triggers were not artifacts of the specific molecules sampled, we validated our findings using split-half resampling. We compared the SHAP feature importance rankings across two independent subsets of the data. This split-half validation yielded a Spearman rank correlation of $\rho = 0.4215$ ($p = 7.56 \times 10^{-79}$), demonstrating a highly stable feature importance profile (Figure 5). The consistency of these rankings confirms that the identified motifs, particularly the constrained heterocycles and specific local environments, are genuine architectural triggers of HOMO re-entanglement.

4 Conclusions

In this study, we addressed a persistent ambiguity in the analysis of large-scale quantum chemical datasets: whether the statistical orthogonality of the highest occupied molecular orbital (HOMO) energy to geometric and vibrational properties is an intrinsic physical invariant or a global population artifact. By systematically removing stoichiometric baselines and applying robust non-parametric manifold analysis to the QM9 dataset, we demonstrated that the apparent decoupling of the electronic frontier is strictly a global phenomenon that collapses within well-defined chemical domains.

Our localized distance correlation mapping revealed that while the HOMO energy varies independently of the geometric-vibrational core (c_v, $zpve$, r^2) across the majority of the structural manifold, approximately four percent of local chemical neighborhoods exhibit strong re-entanglement, with local distance correlations exceeding 0.50. This confirms that the relationship between frontier electronic states and macroscopic thermodynamic properties is highly conditional and topologically localized.

Crucially, by treating this localized coupling strength as a target for supervised learning, we successfully isolated the specific architectural triggers responsible for this rewiring. SHAP feature importance analysis demonstrated that re-entanglement is not driven by global heavy-atom proxies, but rather by specific local atom-centered environments and heteroatom ring strain. Nitrogen- and oxygen-containing heterocycles emerged as the primary functional triggers, whereas fluorine substitution showed no influence. This suggests that the interaction of specific heteroatom lone-pair configurations with strained ring orbitals forces the electronic frontier to re-integrate with the spatial and vibrational organization of the molecule. The high stability of these extracted motifs, confirmed via split-half resampling, underscores the robustness of these structural triggers.

Ultimately, these findings refute the hypothesis that the HOMO energy acts as a universally independent axis of variation in small organic molecules. Instead, it must be characterized as a conditionally joint property. By pinpointing the exact structural motifs that bridge localized quantum states with global thermodynamic

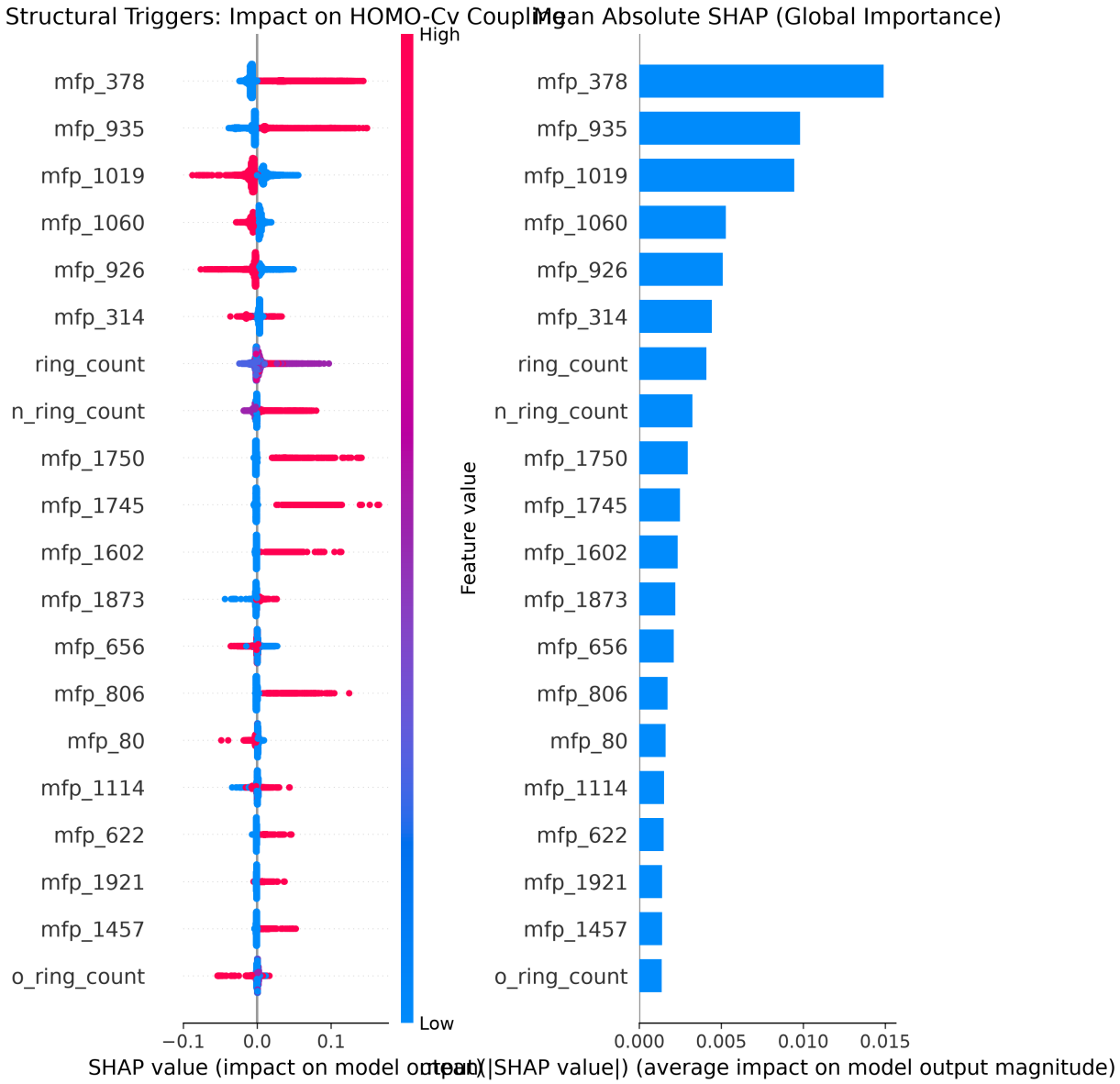

Figure 4: SHAP feature importance analysis for an XGBoost model predicting the local distance correlation between HOMO energy and heat capacity (c_v). The left panel displays the distribution of SHAP values for individual predictions, colored by feature magnitude, while the right panel ranks the top features by their mean absolute SHAP values. The plot reveals that specific molecular fragments, represented by Morgan fingerprint bits (such as `mfp_378`, `mfp_935`, and `mfp_1019`), along with heteroatom ring strain indicators (`ring_count`, `n_ring_count`, and `o_ring_count`), are the primary structural triggers for HOMO re-entanglement. This demonstrates that the localized coupling of the electronic frontier to the geometric-vibrational core is systematically driven by specific local atom-centered environments and nitrogen or oxygen heterocycles, rather than global stoichiometric proxies.

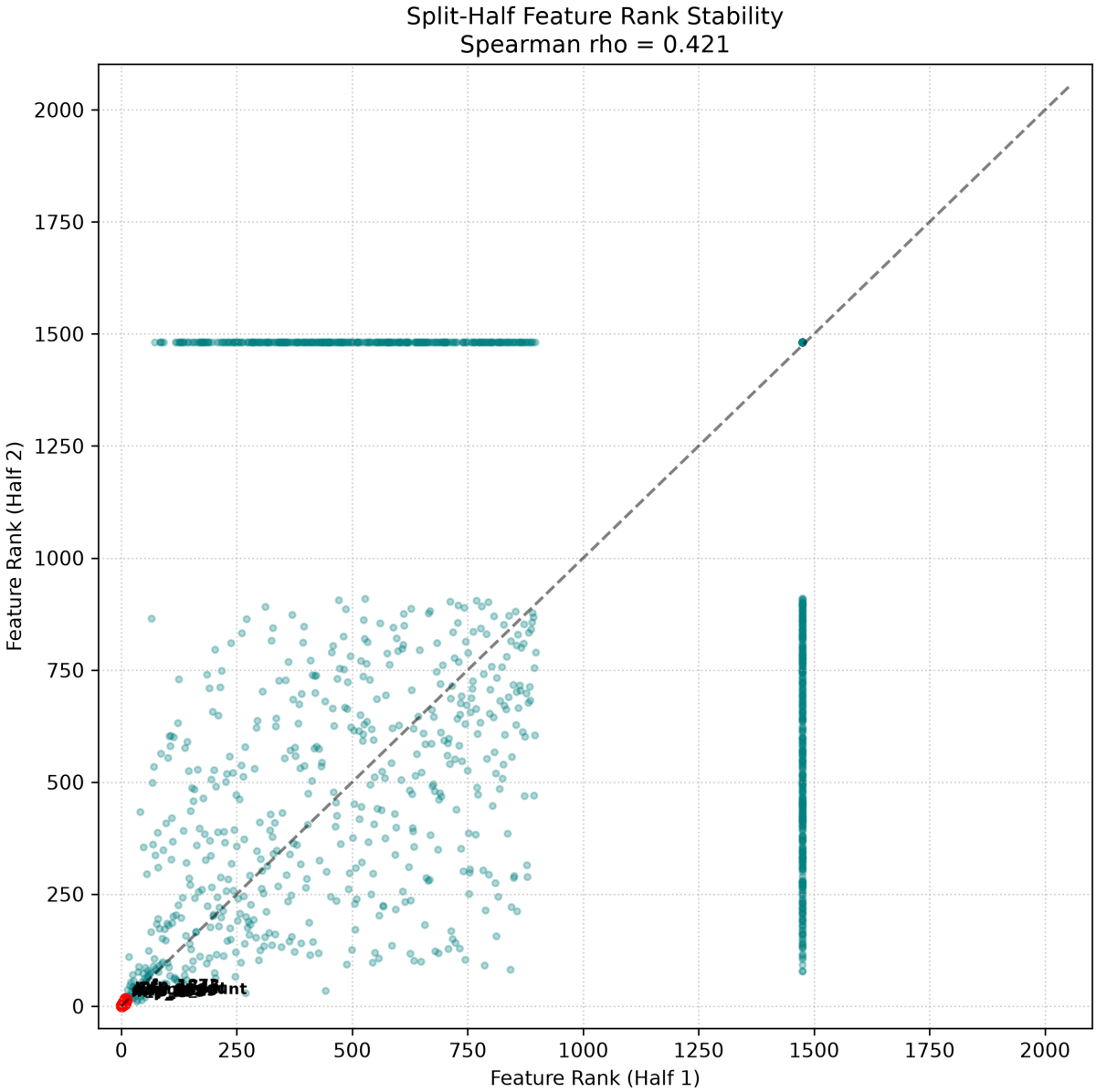

Figure 5: Split-half feature rank stability comparing SHAP importance rankings of structural motifs across two independent data subsets. The scatter plot exhibits a Spearman rank correlation of $\rho = 0.421$, demonstrating the robustness of the XGBoost model trained to predict the local coupling strength between the HOMO energy and the geometric-vibrational core, specifically heat capacity (c_v). The most influential features driving this re-entanglement, such as specific Morgan fingerprints (e.g., `mfp_378`) and ring counts (e.g., `ring_count`), are highlighted in red near the origin, indicating consistently high importance in both halves. This stability confirms that the identified structural motifs are genuine architectural triggers of HOMO re-entanglement rather than sampling artifacts.

behavior, this work provides a more nuanced understanding of molecular property spaces, highlighting the necessity of local, non-linear analytical frameworks in computational chemistry and molecular design.