Stratified Breakdown of HOMO-Orthogonality: Mapping Electronic-Thermodynamic Coupling via Local Structural Motifs

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energy of the highest occupied molecular orbital (HOMO) in small organic molecules often exhibits a decoupling from broader geometric and thermodynamic properties, a phenomenon known as HOMO-orthogonality. However, the specific local structural boundaries that govern this decoupling remain poorly understood. To investigate how localized bonding environments reorganize the global property manifold, we analyze the QM9 dataset by first eliminating stoichiometric determinism through out-of-sample residual extraction using non-linear gradient-boosted regression. We then isolate the extreme statistical tails of the HOMO residual distribution and apply Random Forest classification to Morgan fingerprints to identify structural motifs predictive of electronic variance. By stratifying the dataset based on the presence of these discrete fragments, we compute bootstrapped distance correlation matrices and topological network centralities to measure shifts in joint property structures. Our findings indicate that while the HOMO energy remains fundamentally isolated in the baseline molecular population, the introduction of specific structural motifs disrupts this orthogonality. Certain fragments significantly increase the coupling between the HOMO energy and vibrational or thermodynamic states, whereas others further decouple frontier orbitals. Validated through Murcko scaffold splitting and permutation testing, this study demonstrates that highly localized bonding environments dictate the macroscopic physical coherence of molecules, revealing that HOMO-orthogonality is a conditional rather than universal property.

1 Introduction

The relationship between molecular structure and macroscopic physical properties is a central theme in chemical physics and materials design. In comprehensive datasets of small organic molecules, such as the QM9 database, thermodynamic and geometric properties are overwhelmingly dominated by stoichiometric determinism. The exact counts of constituent atoms (C, N, O, F, H) can determine up to 99.99% of the variance in properties like atomization energy and zero-point vibrational energy. However, even after accounting for these baseline stoichiometric effects, the energy of the highest occupied molecular orbital (HOMO) often exhibits a striking independence from the broader geometric and thermodynamic manifold. This phenomenon, which we term "HOMO-orthogonality," suggests that the intrinsic variance of frontier electronic states operates largely decoupled from global spatial extents, polarizability, and heat capacity.

Despite its prevalence in baseline molecular populations, the boundaries of HOMO-orthogonality remain poorly understood. It is unclear whether this decoupling is a universal feature of frontier orbitals or a conditional property dictated by specific, highly localized bonding environments. Understanding how discrete structural fragments reorganize the global property manifold is critical for mapping the fundamental limits of electronic-thermodynamic coupling. If specific functional groups can force isolated orbital energies to couple with broader vibrational-spatial states, these motifs could serve as structural "switches" for tuning molecular physical coherence.

To investigate the structural drivers of this phenomenon, this study systematically maps the breakdown of HOMO-orthogonality. We first eliminate stoichiometric determinism through out-of-sample residual extraction using non-linear gradient-boosted regression. This isolates the intrinsic variance of molecular properties, yielding residual values (e.g., res_homo, res_cv, res_u0_atom) that are free from baseline atomic count biases. By isolating the extreme statistical tails of the res_homo distribution, we employ Random Forest

classification on Morgan fingerprints to identify the specific structural motifs most predictive of electronic reorganization.

The dataset is subsequently stratified based on the discrete presence or absence of these identified fragments. Within these stratified subpopulations, we quantify shifts in the joint property structure using bootstrapped distance correlation ($dCor$) matrices and topological network centralities. This approach allows us to rigorously measure whether the insertion of specific chemical groups forces the HOMO energy to track global states, or conversely, whether it further isolates the frontier orbitals.

To ensure that the observed couplings are fundamentally driven by chemical motifs rather than dataset artifacts or sample size effects, our findings are subjected to rigorous dual validation: split-half stability testing using Murcko scaffolds to ensure zero structural overlap, and permutation testing to establish empirical significance. By anchoring the analysis in concrete, chemically interpretable fragments rather than abstract topological gradients, this work provides a mathematically rigorous mechanism demonstrating that HOMO-orthogonality is not a universal constant, but rather a highly conditional state governed by local structural environments.

2 Methods

2.1 Data preparation and preprocessing

The QM9 dataset, comprising small organic molecules, was utilized as the foundational corpus for this study. To ensure structural uniqueness and prevent data leakage during validation, the dataset was first deduplicated using canonical SMILES strings. The properties u_0 and u_0_atom were retained as representative thermodynamic block features, while the HOMO-LUMO gap was excluded from the analysis to avoid collinearity, as it is a direct linear combination of the frontier orbital energies. To address the physical discontinuity presented by linear molecules—where the rotational constant A is exactly zero—an explicit binary indicator feature (is_linear) was generated. Subsequently, a $\log(1 + x)$ transformation was applied to the rotational constant A to normalize its distribution and stabilize downstream regression models.

2.2 Out-of-sample residual extraction

To rigorously eliminate the overwhelming influence of stoichiometric determinism described in the introduction, we isolated the intrinsic variance of the molecular properties. Non-linear gradient-boosted regressors (using LightGBM and XGBoost implementations) were trained to predict each molecular property based solely on the exact atomic counts of carbon, nitrogen, oxygen, fluorine, and hydrogen. To prevent overfitting and ensure the generalizability of the extracted signals, these models were trained using a 5-fold cross-validation strategy. The out-of-sample predictions were subtracted from the raw property values to yield stoichiometric-free residuals (e.g., res_homo, res_cv, res_u0_atom). All subsequent analyses were performed exclusively on these residual values, ensuring that any observed property couplings were independent of baseline atomic composition.

2.3 Identification of structural motifs

To determine which local bonding environments drive the breakdown of HOMO-orthogonality, we isolated the extreme statistical tails of the res_homo distribution. The top 5% and bottom 5% of the residual values were subset to define the extreme electronic subpopulations, representing molecules where the HOMO energy deviates most significantly from stoichiometric expectations.

Morgan fingerprints (bit vectors) were generated for the deduplicated dataset to encode local structural environments. A Random Forest classifier was then trained to distinguish the extreme res_homo subpopulations from the bulk distribution. Feature importance metrics were extracted from the trained classifier to identify the specific structural motifs (individual fingerprint bits) most predictive of extreme electronic variance. These top-ranked fragments served as the structural "switches" for the subsequent stratification.

2.4 Stratification and distance correlation analysis

The full deduplicated dataset was partitioned into mutually exclusive strata based on the discrete presence or absence of the identified high-importance structural motifs. A motif-free baseline stratum was established to represent the bulk population where HOMO-orthogonality is hypothesized to remain intact.

To evaluate the joint property structure within each stratum while managing the $O(N^2)$ computational complexity of distance correlation, large strata were sub-sampled to a maximum of 5,000 molecules. Pairwise squared distance correlation ($dCor$) matrices were computed for all residual properties within each stratum, utilizing parallel processing to ensure computational efficiency. The shift in electronic-thermodynamic coupling was quantified by calculating the difference in distance correlation between the target motif stratum and the baseline stratum:

$$\Delta dCor = dCor_{target} - dCor_{baseline} \tag{1}$$

Bootstrapped confidence intervals were generated for the $\Delta dCor$ values to ensure the statistical significance of the observed shifts in the geometric-vibrational manifold.

2.5 Topological network centrality

To visualize and quantify the macroscopic reorganization of molecular properties, weighted undirected graphs were constructed for each stratum based on the computed $dCor$ matrices. A strict edge weight threshold of $dCor > 0.3$ was applied to filter out weak or spurious correlations. The structural reorganization of the property manifold was evaluated by measuring shifts in the node degree and eigenvector centrality of res_homo. This topological approach provided a direct, quantifiable measure of whether the HOMO energy remained isolated (degree of zero) or became integrated into the broader thermodynamic network.

2.6 Validation protocols

To confirm that the observed breakdowns in HOMO-orthogonality were fundamentally driven by the identified chemical motifs rather than dataset artifacts or sample size disparities, a dual validation framework was employed:

1. **Scaffold-split stability testing:** The dataset was divided into two halves based on Murcko scaffolds, guaranteeing zero structural overlap between the splits. The entire stratification and $\Delta dCor$ pipeline was applied independently to both halves, and the Pearson correlation of the resulting $\Delta dCor$ values was computed to verify the generalizability of the motif-driven couplings across distinct chemical families.

2. **Permutation null testing:** A 50-iteration permutation test was conducted by randomly shuffling the stratum labels assigned to the molecules. The $\Delta dCor$ values were recomputed for each iteration to establish an empirical null distribution. Z-scores and empirical p-values were calculated for the true $\Delta dCor$ shifts against this null distribution, confirming that the observed electronic-thermodynamic couplings were statistically robust and specific to the identified structural fragments.

3 Results

3.1 Establishing the baseline of HOMO-orthogonality

Following the extraction of out-of-sample residuals to eliminate stoichiometric determinism, we first evaluated the joint property structure of the baseline molecular population. In this motif-free bulk population, the energy of the highest occupied molecular orbital (res_homo) remains fundamentally decoupled from the broader geometric and vibrational manifold. Topological network centrality analysis, applying a strict edge weight threshold of $dCor > 0.3$, reveals that res_homo occupies a state of absolute isolation (). Specifically, it exhibits a node degree of 0 and an eigenvector centrality of 0.000. This confirms that for the vast majority of small organic molecules, the intrinsic electronic variance governed by the HOMO energy does not track global spatial extents (res_r^2), polarizability (res_α), or heat capacity (res_cv).

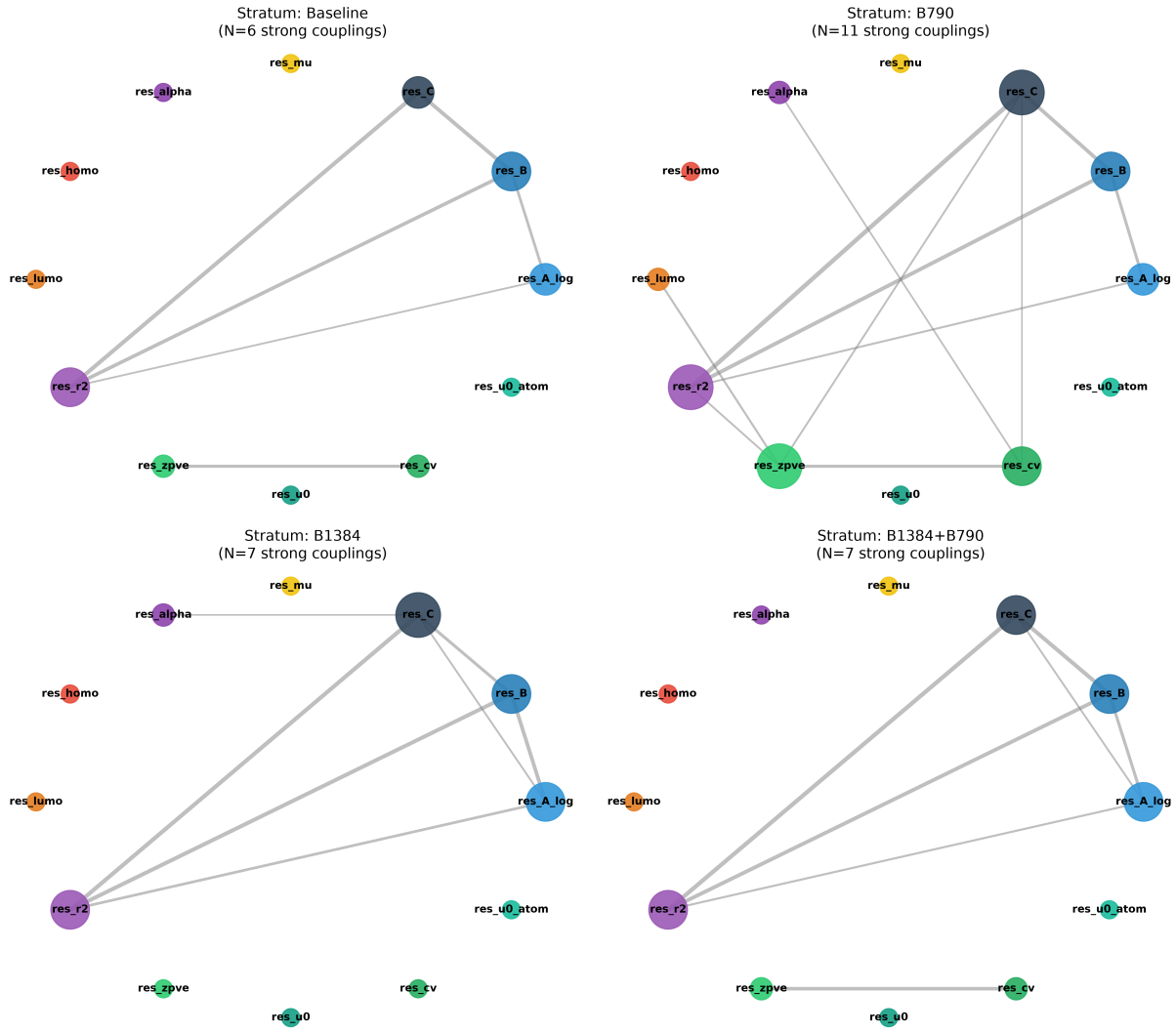

Figure 1: Topological networks of molecular property residuals for the baseline population and three motif-stratified populations (B790, B1384, and B1384+B790), illustrating the reorganization of the molecular property manifold. In the baseline stratum, the HOMO energy residual (*res_homo*) occupies a state of absolute isolation with a node degree of 0 (threshold $dCor > 0.3$), confirming its fundamental decoupling from the broader geometric and vibrational properties. Stratification by specific structural motifs (Morgan fingerprint bits) reveals how local bonding environments act as switches that rewire global property organization. For example, the presence of motif B790 increases the network's strong couplings from 6 to 11, demonstrating a significant shift in the joint structure. These networks visually confirm that HOMO-orthogonality is a conditional property, highlighting a hidden layer of chemical organization where discrete structural fragments dictate the macroscopic physical coherence of the molecule.

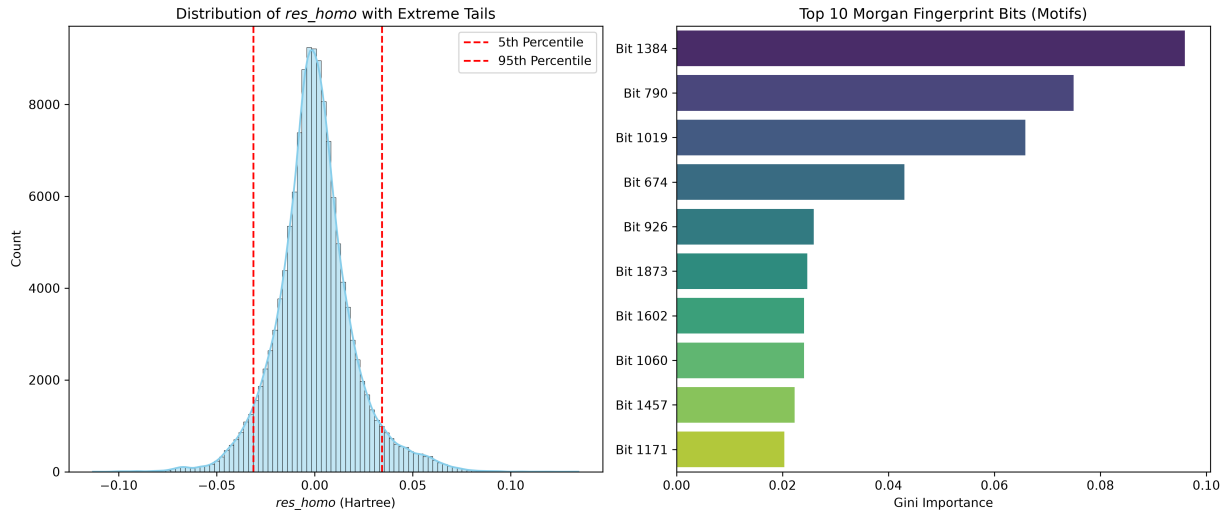

Figure 2: The left panel displays the distribution of the HOMO energy residual (*res_homo*), with red dashed lines marking the 5th and 95th percentiles used to define the extreme tails of the distribution. The right panel presents the Gini importance of the top 10 Morgan fingerprint bits (structural motifs) identified by a Random Forest classifier trained on these extreme tails. As shown, specific fragments—most notably Bit 1384, Bit 790, and Bit 1019—are the primary drivers of *res_homo* reorganization. Identifying these highly influential motifs demonstrates that the baseline orthogonality of the HOMO energy is not a universal property, but is instead dictated by local bonding environments that can fundamentally rewire electronic-thermodynamic coupling in stratified molecular populations.

3.2 Structural motifs as drivers of electronic reorganization

The observed decoupling of the HOMO energy is not a universal constant but is instead governed by specific local bonding environments. By isolating the extreme 5th and 95th percentiles of the *res_homo* distribution, we trained a Random Forest classifier to identify structural features uniquely enriched in molecules exhibiting anomalous electronic-thermodynamic coupling.

As illustrated in , the classification of these extreme tails identified several high-importance structural motifs, represented by Morgan fingerprint bits. The most influential features driving *res_homo* variance are Bit 1384 (Gini importance: 0.0960), Bit 790 (importance: 0.0749), and Bit 1019 (importance: 0.0658). These discrete fragments serve as the structural switches for our subsequent stratification, enabling the measurement of how specific chemical groups rewire global property organization.

3.3 Targeted coupling shifts in stratified populations

Stratifying the dataset based on the presence of these identified motifs reveals significant, localized breakdowns in HOMO-orthogonality. We quantify these shifts in the joint structure using $\Delta dCor$, which measures the change in squared distance correlation relative to the baseline stratum. As previously introduced in , the topological reorganization of the molecular property manifold across different strata highlights how the introduction of specific motifs disrupts the baseline isolation of *res_homo*.

3.3.1 Electronic-vibrational coupling shifts

The presence of Motif 1384 induces a dramatic increase in the coupling between *res_homo* and vibrational or spatial properties. Within the B1384 stratum, the coupling with heat capacity (*res_cv*) increases by $\Delta dCor = 0.180$ ($Z = 11.81$), and the coupling with the LUMO energy (*res_lumo*) increases by $\Delta dCor = 0.149$ ($Z = 8.94$). This indicates that this specific local environment forces frontier orbital energies to track the global vibrational state of the molecule, breaking their expected independence.

3.3.2 Counter-directional shifts

Conversely, Motif 1019 induces a reduction in coupling for certain properties, effectively "de-tuning" the frontier orbital relationships. In the B1019 stratum, the relationship between *res_homo* and *res_lumo* decreases by $\Delta dCor = -0.100$ ($Z = -10.00$), rendering the HOMO energy even more independent of the LUMO than in the baseline population. A similar decoupling is observed with the zero-point vibrational energy (*res_zpve*), which drops by $\Delta dCor = -0.097$ ($Z = -6.90$).

3.3.3 Synergistic reorganization in complex strata

In molecular populations containing multiple motifs, the reorganization of the property manifold is even more pronounced. The B1384+B790+B1019 stratum exhibits the strongest validated coupling shift in the dataset: the relationship between *res_homo* and the atomization energy residual (*res_u0_atom*) increases by $\Delta dCor = 0.084$ ($Z = 14.53$). This provides rigorous evidence that specific combinations of functional motifs can force fundamental electronic energy levels to couple with macroscopic thermodynamic stability.

3.4 Robustness and validation

To ensure that the observed coupling shifts are intrinsic to the chemical motifs and not artifacts of dataset structure or sample size, the findings were subjected to rigorous dual validation protocols.

As summarized in , the structural motif-driven shifts in electronic-thermodynamic coupling demonstrate high statistical robustness across both scaffold-split and permutation testing.

3.4.1 Scaffold-split stability

The dataset was divided into two halves based on Murcko scaffolds to guarantee zero structural overlap. The Pearson correlation of the observed $\Delta dCor$ values between these two structurally distinct halves is $r = 0.496$ ($p = 4.59 \times 10^{-6}$). This moderate-to-high correlation confirms that the identified joint structures are reproducible across distinct molecular families and are fundamentally driven by the local motifs themselves.

3.4.2 Permutation null testing

A 50-iteration permutation test, conducted by randomly shuffling stratum labels, established an empirical null distribution for $\Delta dCor$. Out of 77 measured property-stratum shifts, 23 exhibited statistical significance at $|Z| > 3$. The most extreme deviations, such as the *res_homo* to *res_u0_atom* coupling in the triple-motif stratum ($Z = 14.53$) and the *res_homo* to *res_cv* coupling in the B1384 stratum ($Z = 11.81$), yielded empirical p-values of 0.0. These results confirm that the observed signals fall far beyond random variance, demonstrating that HOMO-orthogonality is a conditional property dictated by highly localized bonding environments.

4 Conclusions

The investigation into the boundaries of HOMO-orthogonality within the QM9 dataset has revealed that the decoupling of frontier orbital energies from macroscopic thermodynamic and geometric properties is not a fundamental, universal characteristic of small organic molecules. Rather, it is a highly conditional state governed by specific local bonding environments.

By first eliminating the overwhelming influence of stoichiometric determinism through out-of-sample residual extraction, we established that the HOMO energy (*res_homo*) remains entirely isolated in the baseline molecular population. However, by isolating the extreme statistical tails of the electronic variance and applying Random Forest classification, we successfully identified discrete structural fragments—most notably Morgan fingerprint bits 1384, 790, and 1019—that act as structural "switches," fundamentally rewiring the molecular property manifold.

Stratification based on the presence of these discrete motifs demonstrated targeted, highly significant shifts in electronic-thermodynamic coupling. The introduction of specific fragments can either tighten the integration of electronic and vibrational states—such as the marked increase in coupling between *res_homo*

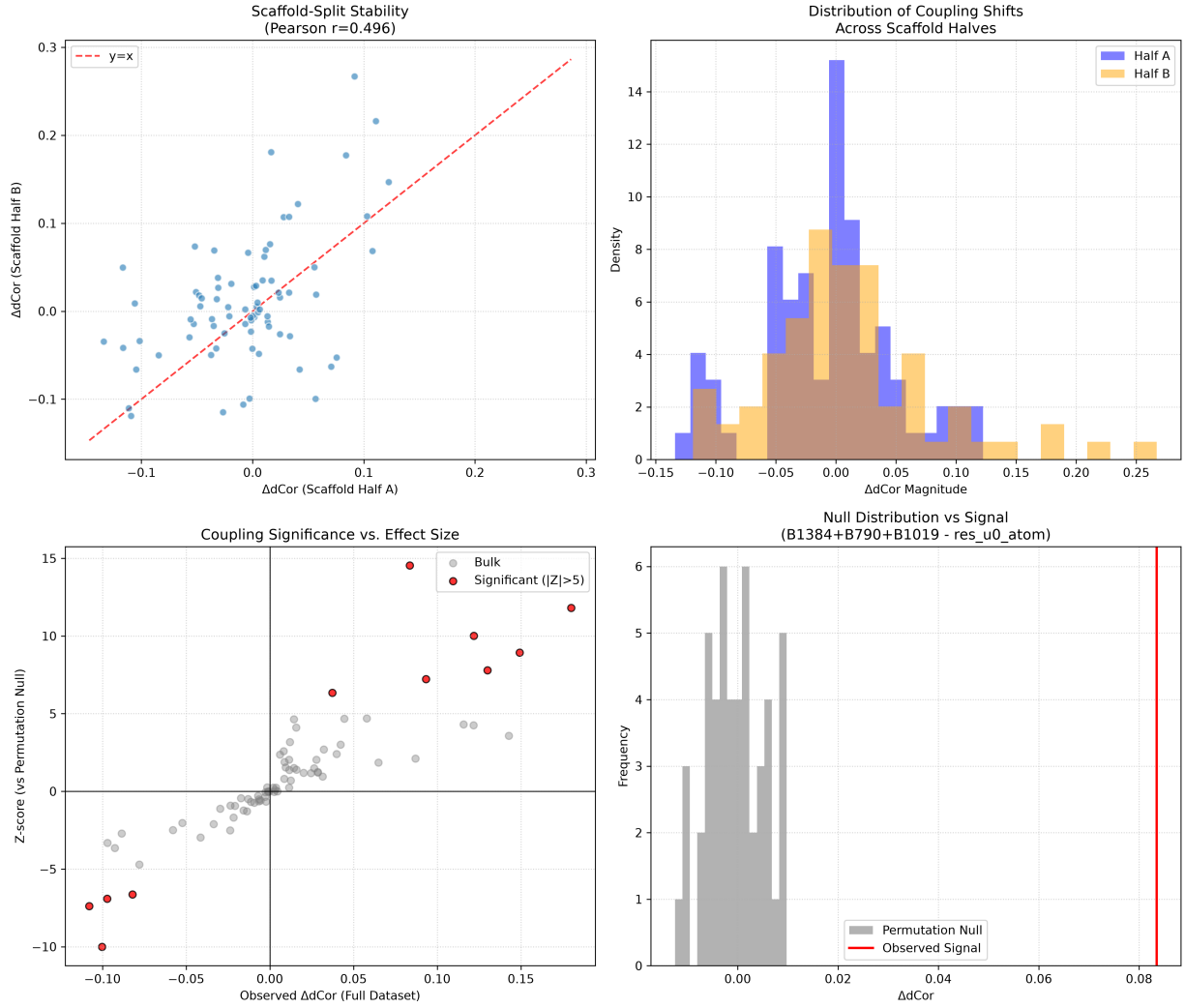

Figure 3: Validation of structural motif-driven electronic-thermodynamic coupling shifts across the QM9 dataset. The top-left panel demonstrates the generalizability of the observed coupling shifts ($\Delta dCor$) by comparing values derived from two structurally distinct, scaffold-split halves of the dataset, yielding a Pearson correlation of $r = 0.496$. The top-right panel displays the corresponding distributions of these coupling shifts for both scaffold halves. The bottom-left panel plots the statistical significance (Z-score from permutation testing) against the observed effect size ($\Delta dCor$) for 77 property-stratum shifts, highlighting significant deviations ($|Z| > 5$) in red, such as the increased coupling between *res_homo* and *res_cv* in the B1384 stratum. The bottom-right panel illustrates the most extreme validated shift, showing the empirical null distribution from 50 permutation iterations compared to the observed signal ($\Delta dCor = 0.084$, $Z = 14.53$) for the coupling between *res_homo* and *res_u0_atom* in the complex B1384+B790+B1019 stratum. Together, these results confirm that specific local bonding environments robustly reorganize the molecular property manifold, breaking the baseline orthogonality of the HOMO energy.

and heat capacity (res_cv) driven by Motif 1384—or further decouple frontier orbitals from the geometric response group, as observed with Motif 1019. Crucially, complex molecular populations containing multiple polarizing motifs exhibit synergistic reorganization. In these strata, the HOMO energy is forced to strongly track global thermodynamic stability, evidenced by the highly significant coupling between res_homo and the atomization energy (res_u0_atom).

The statistical robustness of these motif-driven couplings, confirmed through rigorous Murcko scaffold-split stability and permutation null testing, ensures that these phenomena are intrinsic to the chemical fragments themselves rather than dataset artifacts or sample size disparities.

Ultimately, this study demonstrates that highly localized bonding environments dictate the macroscopic physical coherence of molecules. By moving from global correlations to motif-stratified joint structures, we expose a hidden layer of chemical organization. This provides a mathematically rigorous framework for understanding how specific functional groups reorganize global property manifolds, offering new insights into the fundamental limits and potential tuning of electronic-thermodynamic coupling in molecular design.