Local Bonding Drivers of Frontier Orbital Integration in the QM9 Manifold via Conditional Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The computed properties of small organic molecules are typically analyzed as independent targets, which obscures their joint structural variances beneath dominant stoichiometric scaling and leaves frontier orbital energies seemingly decoupled from macroscopic thermodynamic states. To determine whether local bonding environments govern this electronic decoupling, we analyzed the QM9 dataset by isolating genuine chemical architecture from molecular size effects. Specifically, we employed non-linear gradient-boosted models to regress out atom-count baselines and defined a macroscopic vibrational-spatial-rotational manifold using the residuals of zero-point vibrational energy, heat capacity, electronic spatial extent, and the third rotational constant. By stratifying the dataset into robust functional group sub-populations, we computed the multidimensional unbiased distance correlation between frontier orbital residuals and this macroscopic manifold. Our results indicate that while the highest occupied molecular orbital (HOMO) is globally decoupled from the manifold, specific local motifs such as ketones and lactams force significant integration. Conversely, the lowest unoccupied molecular orbital (LUMO) exhibits higher global integration that is further amplified by ionic motifs like quaternary nitrogens. Furthermore, pairwise correlation analyses identify molecular shape and vibrational degrees of freedom as the primary hubs pulling these frontier orbitals into alignment. Validated via bootstrap null distributions and split-half stability testing, our findings demonstrate that specific local bonding environments systematically rewire the relationship between frontier electronic states and macroscopic molecular properties.

1 Introduction

The prediction and analysis of quantum chemical properties for small organic molecules have largely treated individual properties as independent targets. This perspective, while highly effective for optimizing specific molecular traits, obscures the joint structural variances inherent to molecular systems. A major confounding factor in uncovering these joint relationships is stoichiometric determinism; dominant size effects and atom-count scaling mask the underlying chemical architecture. When these macroscopic scaling laws are accounted for, frontier orbital energies—specifically the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO)—appear globally decoupled from macroscopic thermodynamic and spatial states.

To resolve this dataset-wide orthogonality, we investigate whether the apparent decoupling of frontier electronic states is a universal property of the chemical space or if it is governed by local bonding environments. We hypothesize that specific local motifs, rather than macroscopic topology, dictate the extent to which frontier orbitals integrate with a molecule's broader thermodynamic and spatial manifold.

To test this hypothesis, we analyzed the QM9 dataset by systematically isolating genuine chemical architecture from molecular size effects. We employed non-linear gradient-boosted models to regress out atom-count baselines, extracting the residuals for key molecular properties. Abandoning sparse global scaffolds, we defined a tightly bound macroscopic vibrational-spatial-rotational manifold comprising the residuals of zero-point vibrational energy (res_zpve), heat capacity (res_c_v), electronic spatial extent (res_r^2), and the third rotational constant (res_C). By stratifying the deduplicated dataset into robust functional group sub-populations, we computed the multidimensional unbiased distance correlation ($dCor$) between the frontier orbital residuals and this macroscopic manifold.

Property Pair	Mean $dCor$ across Strata
res_lumo vs. res_C	0.2486
res_lumo vs. res_c_v	0.2471
res_lumo vs. res_r^2	0.2428
res_homo vs. res_c_v	0.2157
res_homo vs. res_r^2	0.1888
res_homo vs. res_C	0.1731

Our conditional distance correlation analysis reveals that local bonding environments systematically rewire the relationship between frontier electronic states and macroscopic properties. While the HOMO energy is globally decoupled from the manifold, specific local motifs such as ketones and lactams force significant integration, breaking its isolation. Conversely, the LUMO exhibits a higher baseline of global integration that is further amplified by ionic motifs, notably quaternary nitrogens. Furthermore, pairwise correlation analyses across these functional group strata identify molecular shape and vibrational degrees of freedom as the primary hubs pulling these frontier orbitals into alignment, as summarized in Table 1.

Validated via bootstrap null distributions and split-half stability testing, our findings confirm that these structural rewirings are not artifacts of finite sampling but represent reproducible physical chemistry. By mapping the conditional dependencies of frontier orbitals, this work demonstrates that specific local bonding environments systematically govern the integration of quantum electronic states with macroscopic molecular properties.

2 Methods

2.1 Data ingestion and structural parsing

To investigate the joint structural variances of molecular properties, we utilized the QM9 dataset. The dataset was first deduplicated based on SMILES strings, removing 83 redundant structures and retaining only the first occurrence of each to prevent downstream data leakage. From the available data, we isolated 12 non-redundant properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), frontier orbital energies ($homo$, $lumo$), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity (c_v), and atomization energy at 0 K (u_{0_atom}). To explicitly account for molecular geometry prior to transformation, a binary indicator variable was created to flag linear molecules (where $A = 0$). Subsequently, a $\log(A+1)$ transformation was applied to the first rotational constant to stabilize its extreme variance.

Molecular structures were parsed from the deduplicated SMILES strings using RDKit. To establish a stoichiometric baseline, the exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted for each molecule. Furthermore, comprehensive functional group profiles were generated using RDKit's fragment analysis tools. The resulting fragment counts were binarized to create presence/absence masks for each functional group. Highly co-occurring motifs were grouped into broader, chemically distinct classes to mitigate confounding effects and address the multiple testing problem during subsequent stratification.

2.2 Non-linear stoichiometric baseline and residual extraction

Because the thermodynamic properties in the QM9 dataset exhibit extreme stoichiometric determinism, it was necessary to isolate genuine chemical architecture from molecular size effects. We regressed out the atom-count baseline for each of the 12 non-redundant properties using LightGBM regressors. The input features for these models were restricted to the five exact atom counts and the linear molecule indicator. Model hyperparameters were kept minimal and fixed (e.g., utilizing shallow trees) to ensure the regressors captured only macroscopic stoichiometric scaling rather than overfitting to structural noise or inadvertently learning specific chemical topologies.

A robust 5-fold cross-validation scheme was employed to generate out-of-fold predictions across the entire dataset. These stoichiometric predictions were subtracted from the actual property values to yield non-linear

gradient-boosted residuals (e.g., *res_homo*, *res_lumo*, *res_zpve*). The residuals for the linear molecules were verified to ensure they did not constitute extreme outliers that could disproportionately skew distance correlation metrics.

2.3 Manifold definition and global baseline computation

We defined a macroscopic target manifold representing the tightly bound vibrational-spatial-rotational state of the molecules. This manifold was constructed using the residuals of four properties: zero-point vibrational energy (*res_zpve*), heat capacity (res_c_v), electronic spatial extent (res_r^2), and the third rotational constant (*res_C*). The second rotational constant residual (*res_B*) was explicitly excluded to prevent multicollinearity with *res_C*.

To quantify the dataset-wide decoupling of frontier orbitals, we computed the global multidimensional unbiased distance correlation (*dCor*) between the HOMO energy residual (*res_homo*) and this 4D manifold across the entire deduplicated dataset. An identical global unbiased *dCor* computation was performed for the LUMO energy residual (*res_lumo*). These global baseline values served as the reference point for evaluating local structural rewirings.

2.4 Conditional distance correlation analysis

To determine whether frontier electronic decoupling is governed by local bonding environments, the dataset was stratified into distinct sub-populations based on the binarized functional group profiles. To ensure statistical robustness and minimize small-sample noise, any sub-population containing fewer than 1,000 molecules was filtered out.

Within each robust functional group stratum, the multidimensional unbiased *dCor* between the frontier orbital residuals (*res_homo* and *res_lumo*) and the 4D manifold was computed. We then calculated the integration shift ($\Delta dCor$) by subtracting the global baseline from these conditional *dCor* values. To rigorously test whether the observed shifts were statistically significant rather than artifacts of varying sample sizes, a null distribution was generated for each stratum. This was achieved by bootstrapping the unbiased *dCor* on random subsets of the global dataset that exactly matched the size of the corresponding functional group sub-population.

2.5 Pairwise analysis and split-half validation

To gain a granular view of the integration mechanics and determine whether the structural rewiring of frontier orbitals is symmetric, we systematically compared the conditional $\Delta dCor$ shifts of *res_homo* against those of *res_lumo*. Additionally, pairwise unbiased *dCor* values were computed between the individual frontier orbital residuals and the individual components of the manifold (*res_zpve*, res_c_v, res_r^2, *res_C*) within each sub-population. This pairwise analysis allowed us to identify which specific macroscopic properties act as hubs pulling the frontier orbitals into alignment.

Finally, the robustness of the structural claims was validated via a split-half stability test. Each functional group sub-population was randomly divided into two equal halves, and the conditional multidimensional unbiased *dCor* for both *res_homo* and *res_lumo* was recomputed independently for each half. Confidence intervals were calculated via bootstrapping within the sub-populations to verify that the observed deviations from the global baseline were consistent, reproducible, and not artifacts of specific dataset slices.

3 Results

3.1 Stoichiometric decoupling and manifold definition

To isolate genuine chemical architecture beyond molecular size effects, we utilized non-linear gradient-boosted residuals (LightGBM) to regress out the counts of C, N, O, F, and H atoms from 12 non-redundant properties. Following this stoichiometric decoupling, we established the global baseline integration of frontier orbitals. A macroscopic "vibrational-spatial-rotational" manifold was defined using the residuals of four

tightly bound properties: zero-point vibrational energy (res_zpve), heat capacity (res_c_v), electronic spatial extent (res_r^2), and the third rotational constant (res_C). Globally, the HOMO energy residual (res_homo) exhibits significant decoupling from this manifold, yielding a baseline unbiased distance correlation ($dCor$) of 0.228. In contrast, the LUMO energy residual (res_lumo) demonstrates a higher baseline integration, with a global $dCor$ of 0.317. These global baselines serve as the reference point for evaluating how specific local bonding environments perturb the relationship between electronic states and macroscopic properties.

3.2 Orbital rewiring and integration shifts

Stratifying the dataset into 47 robust functional group sub-populations ($n \geq 1000$) reveals that local chemical motifs can fundamentally rewire the relationship between frontier orbitals and the macroscopic manifold, either forcing integration or driving further decoupling (Figure 1).

The integration shifts ($\Delta dCor$) relative to the global baseline demonstrate distinct behaviors for occupied and unoccupied states (Figure 2). The isolation of the HOMO energy is most strongly broken by the presence of carbonyl and lactam groups. The largest positive integration shifts are observed for Topliss ketones ($\Delta dCor = +0.161$, $p < 0.001$), lactams ($\Delta dCor = +0.151$, $p < 0.001$), general ketones ($\Delta dCor = +0.130$, $p < 0.001$), N-dealkylation motif 2 ($\Delta dCor = +0.123$, $p < 0.001$), and morpholine rings ($\Delta dCor = +0.085$, $p < 0.001$). In these specific sub-populations, the HOMO energy ceases to act as an independent electronic parameter and instead strongly tracks the molecule's vibrational and spatial state. Conversely, aromatic systems tend to further decouple the HOMO; for instance, aryl hydroxyl groups (`frag_fr_Ar_OH`) exhibit a negative shift of $\Delta dCor = -0.185$.

The LUMO integration landscape is driven predominantly by ionic and specific heterocyclic motifs rather than carbonyls. Quaternary nitrogens (`quatN`) induce a massive integration shift ($\Delta dCor = +0.105$, $p < 0.001$), reaching an absolute conditional $dCor$ of 0.422 compared to the global 0.317. This suggests that the formal positive charge significantly tethers the unoccupied orbital energy to the molecule's overall spatial extent and vibrational modes. Other notable LUMO integration drivers include Topliss ketones ($\Delta dCor = +0.065$, $p < 0.001$), morpholines ($\Delta dCor = +0.064$, $p < 0.001$), and esters ($\Delta dCor = +0.057$, $p < 0.001$).

3.3 Hub property analysis

To determine which specific components of the manifold pull the frontier orbitals into alignment, we performed a pairwise $dCor$ analysis between the orbital residuals and the individual manifold properties across all 47 strata (Figure 3).

The mean distance correlations across all strata identify the third rotational constant (res_C) and heat capacity (res_c_v) as the primary hubs for integration (Table 1). The LUMO consistently exhibits higher mean coupling across all manifold components than the HOMO. The prominent role of res_C (reflecting molecular shape and inertia) and res_c_v (reflecting vibrational degrees of freedom) indicates that frontier orbital rewiring is primarily a spatial and vibrational phenomenon. Functional groups that integrate these orbitals do so by making their energies highly sensitive to the overall geometric and vibrational state of the molecule.

3.4 Statistical robustness and stability

To ensure the observed structural rewirings are not artifacts of finite sampling or varying sub-population sizes, we validated the results using bootstrap null distributions and split-half stability testing. For every stratum, the observed integration shift was compared to a size-matched null distribution of randomly sampled molecules. The top integration drivers, including ketones, lactams, and quaternary nitrogens, all yielded $p < 0.05$, confirming that the shifts are specific to the chemical motifs rather than sample size effects.

Furthermore, each functional group sub-population was randomly split into two halves to independently recompute the multidimensional $dCor$ (Figure 4).

The split-half stability analysis demonstrates excellent reproducibility. The correlation between $dCor$ values in the first and second halves across all strata is $r = 0.904$ for the HOMO (mean absolute deviation of 0.028) and $r = 0.867$ for the LUMO (mean absolute deviation of 0.034). This high level of agreement confirms

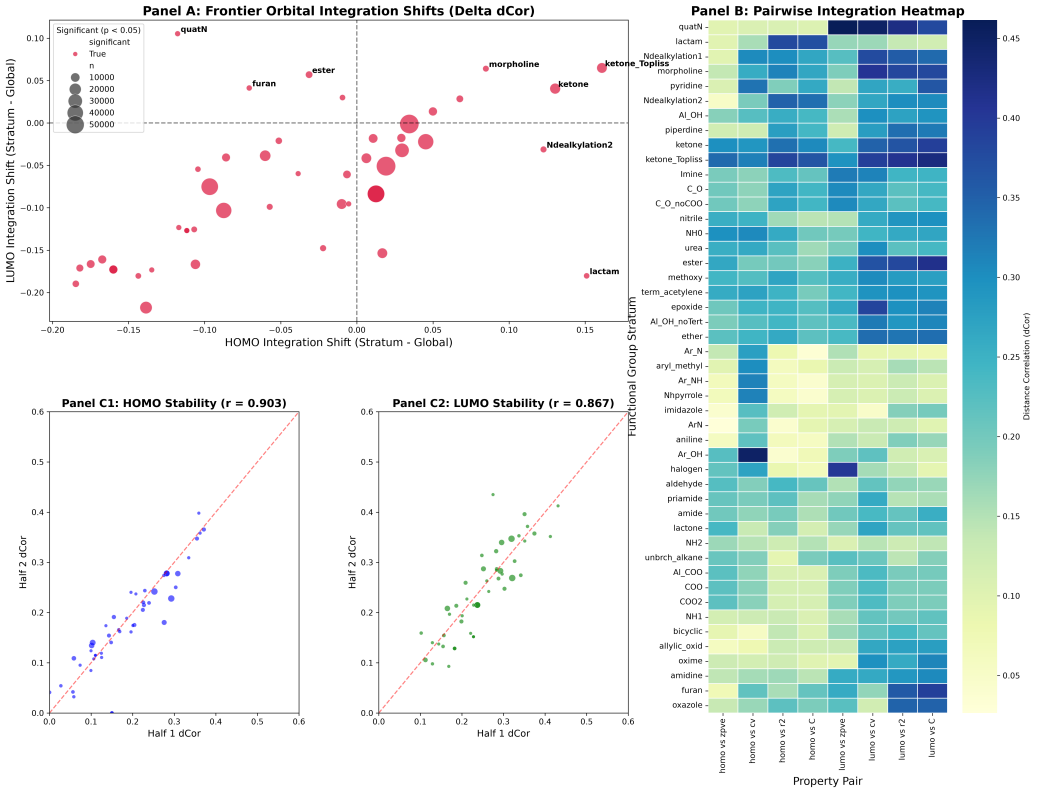

Figure 1: The figure illustrates the integration of frontier orbital energies (HOMO and LUMO) into the vibrational-spatial-rotational manifold across different local chemical environments in the QM9 dataset. Panel A displays the integration shifts ($\Delta dCor$) for HOMO and LUMO relative to the global baseline across 47 functional group strata, highlighting that motifs like ketones (e.g., `ketone_Topliss`) and lactams strongly drive HOMO integration, while quaternary nitrogens (`quatN`) predominantly drive LUMO integration. Panel B presents a pairwise distance correlation ($dCor$) heatmap, revealing that the third rotational constant and heat capacity act as primary hubs for this integration, with LUMO consistently exhibiting stronger coupling to the manifold than HOMO. Panels C1 and C2 demonstrate the statistical robustness of these findings through split-half stability analysis, showing high reproducibility for both HOMO ($r = 0.903$) and LUMO ($r = 0.867$) integration metrics. Overall, the data reveals that while frontier orbitals are globally decoupled from thermodynamic properties, specific local bonding environments force their re-integration with the molecule's macroscopic spatial and vibrational state.

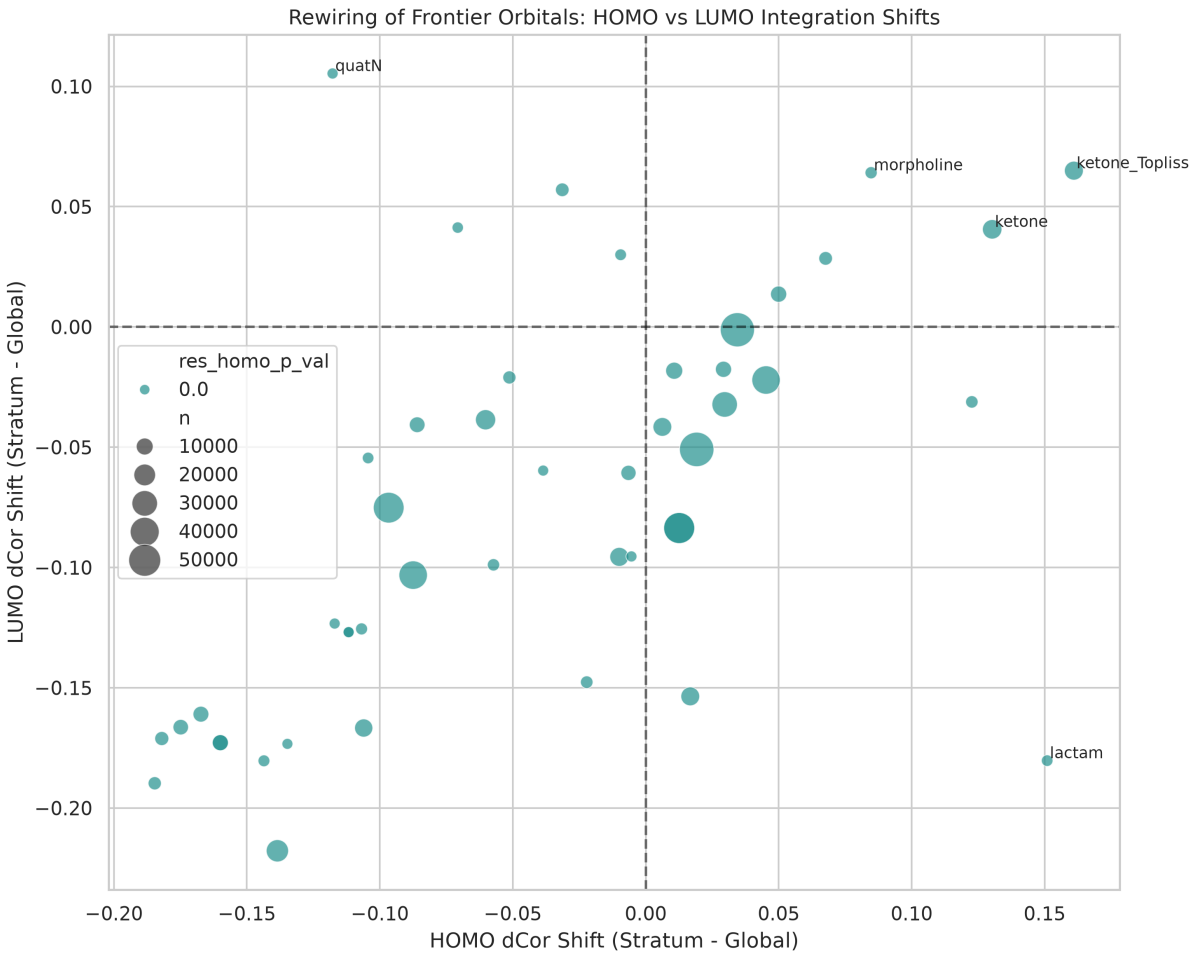

Figure 2: Scatter plot illustrating the integration shifts ($\Delta dCor$) of HOMO versus LUMO energies relative to the global baseline across 47 functional group sub-populations. Dashed lines indicate the global baseline of zero shift. Point sizes reflect the sub-population size (n), with all shown groups possessing highly significant shifts ($p < 0.001$). Labeled points highlight key chemical motifs that drive orbital rewiring: quaternary nitrogen (`quatN`) strongly integrates the LUMO while decoupling the HOMO, whereas lactam groups show the inverse behavior. Conversely, motifs like morpholine and Topliss ketones (`ketone_Topliss`) force the integration of both frontier orbitals simultaneously. This distribution demonstrates that specific local bonding environments, particularly those involving strong dipoles or formal charges, fundamentally rewire the relationship between frontier electronic states and the macroscopic vibrational-geometric manifold of the molecule.

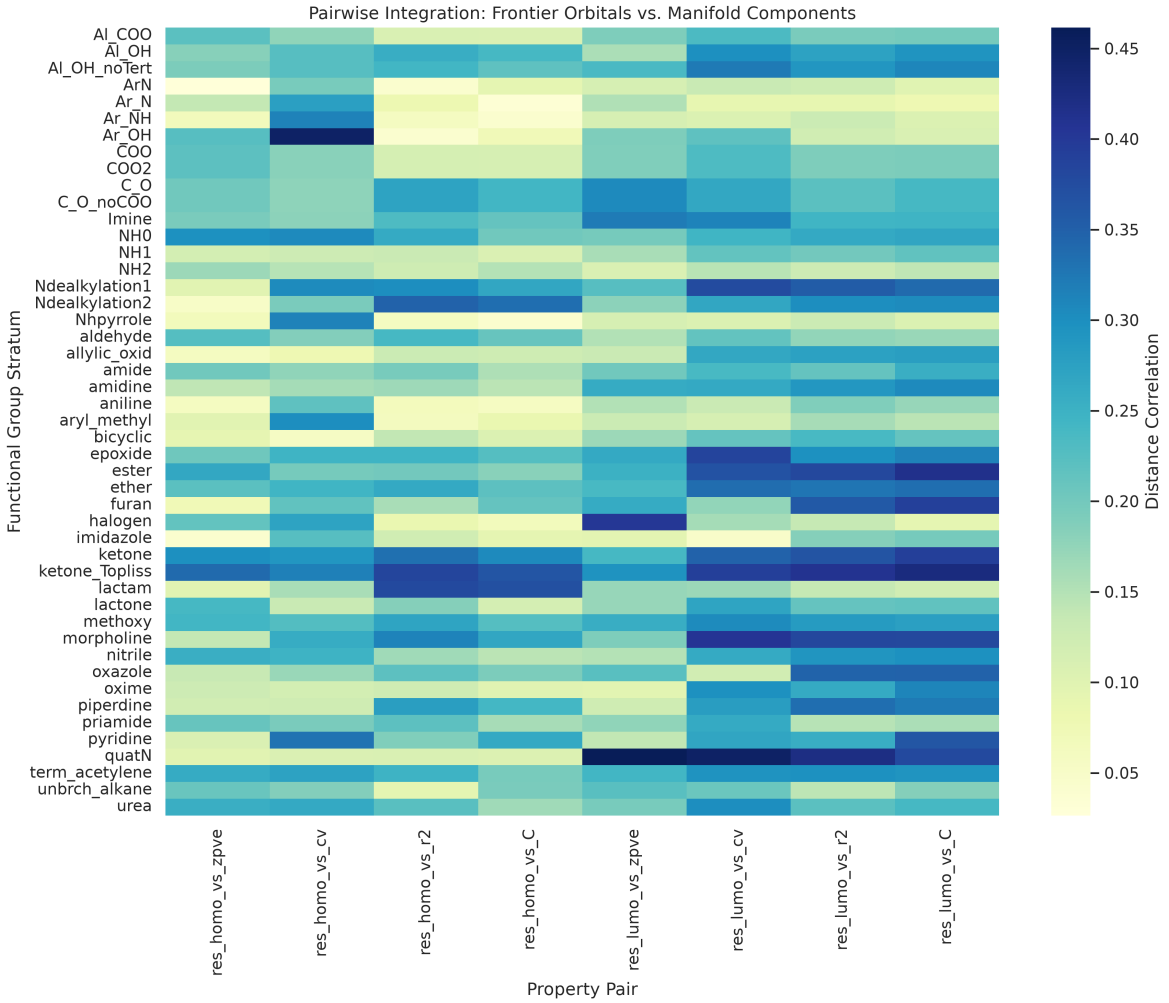

Figure 3: Heatmap of pairwise distance correlations ($dCor$) between frontier orbital energy residuals (res_homo, res_lumo) and components of the vibrational-spatial-rotational manifold (res_zpve, res_c_v, res_r^2, res_C) across 47 functional group strata. The visualization demonstrates how local bonding environments rewire the relationship between electronic states and macroscopic molecular properties. The LUMO energy exhibits consistently higher integration across the manifold components compared to the HOMO, as indicated by the generally higher correlations on the right side of the plot. Specific chemical motifs act as strong integration drivers; for example, quaternary nitrogens strongly tether the LUMO to the manifold, while ketones and lactams drive HOMO integration. Furthermore, the strong coupling observed with res_C and res_c_v identifies molecular shape and vibrational degrees of freedom as the primary hubs for this frontier orbital rewiring.

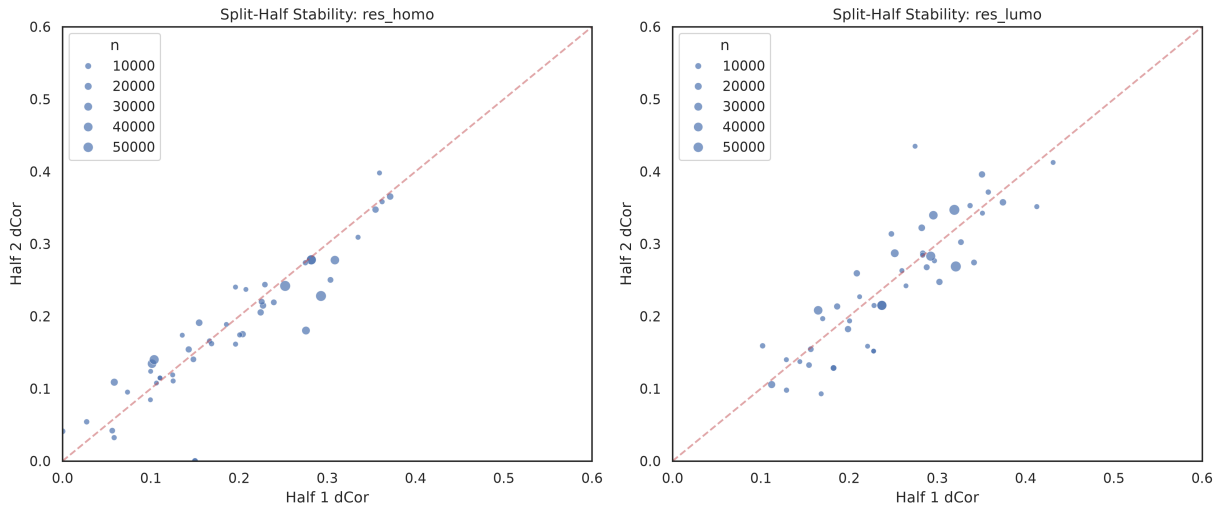

Figure 4: Split-half stability analysis of the multidimensional distance correlation ($dCor$) for the HOMO (`res_homo`) and LUMO (`res_lumo`) energy residuals across functional group sub-populations. Each sub-population was randomly split into two halves to independently recompute the $dCor$ values, with marker sizes indicating the number of molecules (n) in each stratum. The strong alignment of the data points along the identity line (dashed) demonstrates high reproducibility, yielding correlation coefficients of $r = 0.9035$ for the HOMO and $r = 0.8666$ for the LUMO. This high level of agreement confirms that the observed integration shifts between frontier orbitals and the macroscopic vibrational-geometric manifold are not artifacts of finite sampling, but represent reproducible physical chemistry inherent to specific local bonding environments.

that the reported shifts in joint structure represent reproducible physical chemistry inherent to specific functional group classes. Ultimately, the data demonstrates that while HOMO and LUMO energies are globally decoupled from thermodynamics, specific local bonding environments—particularly those involving strong dipoles or formal charges—force a systematic re-integration of frontier electronic states with the molecule's macroscopic vibrational and geometric manifold.

4 Conclusions

In this study, we addressed the apparent dataset-wide decoupling of frontier orbital energies from macroscopic thermodynamic and spatial states in small organic molecules. By systematically regressing out dominant stoichiometric scaling using non-linear gradient-boosted models, we isolated the genuine chemical architecture of the QM9 dataset. Our conditional distance correlation analysis across robust functional group sub-populations revealed that the orthogonality of frontier electronic states is not a universal property of chemical space. Instead, the degree to which these quantum states integrate with a molecule's macroscopic manifold is highly governed by local bonding environments.

We demonstrated that while the HOMO energy is globally isolated from the macroscopic vibrational-spatial-rotational manifold, specific local motifs—most notably carbonyls and lactams—force a significant re-integration. In these specific chemical environments, the HOMO ceases to act as an independent electronic parameter and strongly tracks the molecule's spatial and vibrational state. Conversely, the LUMO exhibits a higher baseline of global integration that is distinctly amplified by ionic and specific heterocyclic motifs, such as quaternary nitrogens. This asymmetry indicates that occupied and unoccupied electronic states are tethered to the macroscopic manifold through fundamentally different local structural drivers.

Furthermore, our pairwise correlation analysis identified the third rotational constant (res_C) and heat capacity (res_c_v) as the primary hubs mediating this integration. This finding highlights that frontier orbital rewiring is predominantly a spatial and vibrational phenomenon. Functional groups that successfully integrate these orbitals do so by coupling their electronic energies to the overall geometric inertia and

vibrational degrees of freedom of the molecule.

The statistical robustness of these structural rewirings was rigorously validated through bootstrap null distributions and split-half stability testing, confirming that the observed integration shifts represent reproducible physical chemistry rather than finite sampling artifacts. Ultimately, this work establishes that specific local bonding environments—particularly those characterized by strong dipoles or formal charges—systematically govern the coupling between quantum electronic states and macroscopic molecular properties. Recognizing these local drivers provides a more cohesive framework for understanding the joint structural variances of molecular property spaces, moving beyond global stoichiometric scaling to uncover the localized rules of electronic-thermodynamic integration.