Stress-Testing HOMO Decoupling: Distance Correlation Networks Across Structural Strain and Latent Geometric Artifacts

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

It remains unclear whether the observed decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from core spatial-thermodynamic molecular properties is a fundamental physical invariant or an artifact of structural constraints and algorithmic noise. To address this, we investigate the joint property structure of organic molecules to mathematically isolate the geometric and architectural boundaries governing frontier-orbital independence. Utilizing the QM9 dataset, we employ gradient boosting regression to systematically remove stoichiometric and size-dependent variance, subsequently constructing distance correlation networks across acyclic, unstrained, and strained molecular families, alongside a statistically isolated cohort of multivariate outliers. By comparing network centrality metrics against permutation-based null models, we find that heat capacity and zero-point vibrational energy act as primary hubs in the spatial-thermodynamic manifold, whereas the HOMO residual remains fundamentally decoupled. Rather than global geometry, HOMO variance is governed by local electronic features such as nitrogen content, topological polar surface area, and carbon saturation. Finally, structural profiling demonstrates that this decoupling breaks down primarily in highly conjugated and aromatic systems, indicating that frontier orbital independence is a robust physical rule disrupted only by extreme electron delocalization or underlying density functional theory convergence limits.

1 Introduction

The mapping of molecular structures to their quantum mechanical properties is a central pillar of modern computational chemistry. Large-scale databases, most notably the QM9 dataset, have provided unprecedented opportunities to explore the joint distribution of spatial, thermodynamic, and electronic properties across chemical space. Within this multivariate property manifold, most macroscopic and thermodynamic features—such as heat capacity (C_v), zero-point vibrational energy (ZPVE), and spatial extent (R^2)—exhibit strong intercorrelations driven by molecular size and stoichiometric composition. However, the Highest Occupied Molecular Orbital (HOMO) energy consistently emerges as an anomaly, displaying near-total orthogonality to the core spatial-thermodynamic network.

1.1 The phenomenon of frontier orbital decoupling

It remains unclear whether the observed decoupling of the HOMO energy is a fundamental physical invariant of small organic molecules or an artifact of structural constraints and algorithmic noise. If frontier orbital independence is a robust physical rule, it should persist across diverse structural topologies, from flexible acyclic chains to highly strained small-ring scaffolds. Conversely, if the isolation of the HOMO energy is conditionally dependent on structural relaxation or underlying density functional theory (DFT) convergence limits, its orthogonality should break down in specific geometric regimes or among statistical outliers. Resolving this ambiguity requires moving beyond simple correlation surveys to mathematically isolate the exact geometric and architectural boundaries that govern frontier-orbital variance.

1.2 Methodological framework

To stress-test the hypothesis of HOMO decoupling, we conduct a topological sensitivity analysis of the QM9 property manifold. We systematically strip away stoichiometric and size-dependent variance using

Property Pair	Mean $dCor$	Acyclic	Unstrained	Strained
res_homo vs res_gap	0.486	0.468	0.425	0.567
res_homo vs res_lumo	0.197	0.090	0.380	0.120
res_homo vs res_u_0	0.050	0.067	0.045	0.038

gradient boosting regression, isolating the intrinsic chemical variations of 12 non-redundant properties. The dataset is subsequently partitioned into validated structural subsets—comprising acyclic, unstrained, and strained molecular families—alongside a statistically isolated cohort of multivariate outliers acting as a proxy for unflagged DFT-inconsistent geometries. By constructing distance correlation ($dCor$) networks for each partition and comparing network centrality metrics against permutation-based null models, we rigorously define the boundaries of the weak-coupling regime.

1.3 Scope and key findings

Our analysis reveals that while properties like heat capacity and zero-point vibrational energy act as primary hubs in the spatial-thermodynamic manifold, the HOMO residual remains fundamentally decoupled. Rather than being governed by global geometry, HOMO variance is dictated by local electronic features such as nitrogen content, topological polar surface area, and carbon saturation.

Furthermore, we investigate whether this decoupling is specific to the HOMO level or shared by frontier orbitals generally. As detailed in Table 1.3, the HOMO residual exhibits negligible coupling to the total internal energy (res_u_0) across all structural families, confirming its independence from the thermodynamic state. However, its relationship with the LUMO residual (res_lumo) is highly sensitive to structural topology, indicating that frontier orbitals do not always behave as a unified, decoupled block.

Finally, structural profiling of the outlier cohort demonstrates that HOMO decoupling breaks down primarily in highly conjugated and aromatic systems. This indicates that frontier orbital independence is indeed a robust physical rule in standard small organic molecules, disrupted only by extreme electron delocalization or latent geometric artifacts approaching DFT convergence limits.

2 Methods

2.1 Data preparation and stoichiometric baseline

To systematically isolate the intrinsic chemical variations governing molecular properties, we established a rigorous stoichiometric baseline using the QM9 dataset. The initial dataset of 133,885 molecules was deduplicated by SMILES strings. To prevent variance distortion during subsequent regressions, nine linear molecules lacking a defined rotational constant ($A = 0$) were removed. A logarithmic transformation, $\log(A + 1)$, was applied to the rotational constant A to normalize its distribution. We isolated 12 non-redundant spatial, thermodynamic, and electronic properties by excluding highly collinear thermodynamic and gap-derived features.

For each molecule, the exact stoichiometric counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were extracted using RDKit. We trained GPU-accelerated LightGBM regressors to predict the 12 non-redundant properties solely from these atom counts. The residuals from these models were extracted and standardized, yielding a stoichiometric baseline that removes the primary variance driven by atomic composition.

2.2 Aggressive size-effect removal and structural partitioning

Because macroscopic properties such as heat capacity (C_v) and zero-point vibrational energy (ZPVE) scale non-linearly with molecular size, we applied an aggressive secondary size correction to ensure that downstream network hubs were not artifacts of residual size scaling. Extended size descriptors were computed for all molecules, including the number of heavy atoms, the number of bonds, and the exact molecular weight. To eliminate multicollinearity, we extracted the first principal component (PC1) of these metrics to serve as a single, robust size-factor covariate. A second set of LightGBM regressors was trained to predict the initial

stoichiometric residuals from this PC1 size factor. The resulting "aggressively size-corrected" residuals were used for all subsequent topological analyses.

To evaluate the conditional dependence of property relationships on structural relaxation, the dataset was partitioned into three primary cohorts based on RDKit ring topology: acyclic molecules, unstrained molecules (containing exclusively 5- or 6-membered rings), and strained molecules (containing 3- or 4-membered rings).

2.3 Distance correlation networks and null model generation

To map the joint property manifold, we constructed Distance Correlation ($dCor$) networks for the 12 aggressively size-corrected properties within each structural partition. Distance correlation was selected over Pearson correlation to capture both linear and non-linear dependencies between complex quantum mechanical properties. The topological importance of each property was quantified by calculating its normalized betweenness centrality within the $dCor$-weighted network.

To rigorously define the "weak coupling" regime of the HOMO residual (res_homo), we established a statistical null baseline. For each structural partition, a "noise" node was generated using a permutation-based null model by randomly shuffling the values of the res_homo array. This procedure destroys true topological links while perfectly preserving the distributional characteristics (mean, variance, and range) of the residuals. The $dCor$ network was then recomputed for the 12 properties plus this 13th permuted noise node. By comparing the centrality of res_homo against the permuted noise node, we statistically validated whether the HOMO energy is genuinely decoupled from the spatial-thermodynamic manifold or if its network position is indistinguishable from random noise.

2.4 Outlier isolation and structural profiling

To investigate whether frontier orbital decoupling breaks down under extreme geometric or algorithmic conditions, we isolated a "stress test" cohort of 3,054 geometrically inconsistent molecules. These multivariate outliers were identified using an Isolation Forest anomaly score computed across the 12-dimensional residual space.

Rather than discarding these outliers as density functional theory (DFT) convergence failures, we performed a structural profiling of the cohort against the valid dataset. We compared the distributions of heavy atom counts, ring sizes, degree of saturation, and specific strained motifs. To determine if structural strain or algorithmic artifacts drive property re-coupling, we computed the partial Pearson correlation between res_homo and res_c_v within this outlier group and compared it to a structurally matched subset of valid molecules possessing identical strain profiles.

2.5 Conjugation-driven mechanism and local electronic descriptors

To investigate the mechanism behind observed shifts in HOMO-C_v correlations, we isolated the effect of electron delocalization. For all molecules in the 5- and 6-membered ring cohorts, we calculated the ratio of aromatic and conjugated atoms to total heavy atoms using RDKit. A partial regression of the res_homo and res_c_v correlation against the conjugation ratio was performed, explicitly partialling out the confounding effects of heavy atom count and ring count.

Pivoting from global network topology to the specific drivers of HOMO variance, we generated a diverse set of structural descriptors capturing local electronic environments. These included the fraction of specific heteroatoms (e.g., N/HeavyAtom), topological polar surface area (TPSA), the fraction of sp3-hybridized carbons, and Gasteiger partial charges. We computed the distance correlation between res_homo and these local descriptors across the structural partitions to identify the specific chemical features that define the independent HOMO axis.

2.6 Orbital versus ensemble decoupling verification

Finally, to determine if the observed weak coupling is unique to the HOMO energy level or shared by frontier orbitals generally, we evaluated the decoupling dynamics of the broader electronic block. We computed the distance correlation between res_homo and the total internal energy residual (res_u_0), the HOMO-LUMO

gap residual (*res_gap*), and the LUMO residual (*res_lumo*) separately. These metrics were calculated across all valid structural partitions to verify whether the frontier orbitals behave as a unified, decoupled block or if the HOMO and LUMO decouple independently from the macroscopic ensemble properties.

3 Results

3.1 Network hubs and the isolation of the HOMO residual

Following the aggressive removal of stoichiometric and size-dependent variance, the remaining property residuals reveal a stable joint manifold across the QM9 dataset. Distance correlation (*dCor*) networks constructed from these aggressively size-corrected residuals identify heat capacity (res_c_v) and zero-point vibrational energy (*res_zpve*) as the primary topological hubs within the spatial-thermodynamic manifold.

In contrast, network centrality analysis confirms that the HOMO energy residual (*res_homo*) remains fundamentally decoupled from this core manifold. Across both the acyclic and unstrained molecular partitions, the normalized betweenness centrality of *res_homo* evaluates to exactly 0.0. When compared against the permutation-based null models, the network position of the HOMO residual is statistically indistinguishable from random noise. This isolation highlights a robust physical feature of the property space: while spatial and thermodynamic properties such as polarizability (res_α) and spatial extent (res_r^2) organize hierarchically around thermodynamic hubs, the frontier orbital levels operate on an entirely independent axis.

3.2 Structural partners of the HOMO axis

Because the variance in *res_homo* is decoupled from global geometric and thermodynamic descriptors, we investigated whether it is instead modulated by local electronic environments. Analysis of structural descriptors across the different molecular families identifies specific chemical partners that track the HOMO residual, as illustrated in Figure 1 (top-left panel).

The strongest global partner for the HOMO residual is nitrogen content, which exhibits a mean *dCor* of 0.215. This relationship is most pronounced in acyclic molecules ($dCor = 0.291$), suggesting that in the absence of ring constraints, the availability of lone pairs and the electronegativity of nitrogen atoms act as primary determinants of HOMO energy variance. Similarly, the topological polar surface area (TPSA) exhibits a mean *dCor* of 0.203, with its influence peaking in acyclic ($dCor = 0.255$) and strained ($dCor = 0.239$) systems.

Conversely, the degree of carbon saturation (measured as the fraction of sp^3-hybridized carbons) shows a distinct topological dependence. While it shares a mean *dCor* of 0.203 globally, it becomes the uniquely dominant partner in unstrained molecules ($dCor = 0.382$). In these 5- and 6-membered ring systems, the local hybridization state significantly rewires the frontier orbital level. Together, these correlations indicate that *res_homo* is governed by the chemical identity of heteroatoms and local hybridization rather than the global spatial extent of the molecule.

3.3 Orbital decoupling dynamics

To determine whether this decoupling is a unique feature of the HOMO level or a general characteristic of frontier orbitals, we evaluated the coupling dynamics within the broader electronic block (Table 1.3 and Figure 1, top-right panel).

As expected, the HOMO residual is moderately coupled to the HOMO-LUMO gap ($dCor \approx 0.486$), given that the HOMO energy is a direct mathematical component of the gap. Crucially, however, the coupling between *res_homo* and the total internal energy (res_u_0) is negligible ($dCor \approx 0.050$) across all structural partitions. This confirms that the frontier orbitals form an electronic block that is largely orthogonal to the total thermodynamic state of the molecule.

Despite this shared orthogonality to the thermodynamic ensemble, the HOMO and LUMO levels do not always behave as a unified block. The coupling between *res_homo* and *res_lumo* is highly sensitive to structural family, reaching a moderate $dCor = 0.380$ in unstrained rings but dropping to a mere 0.090 in

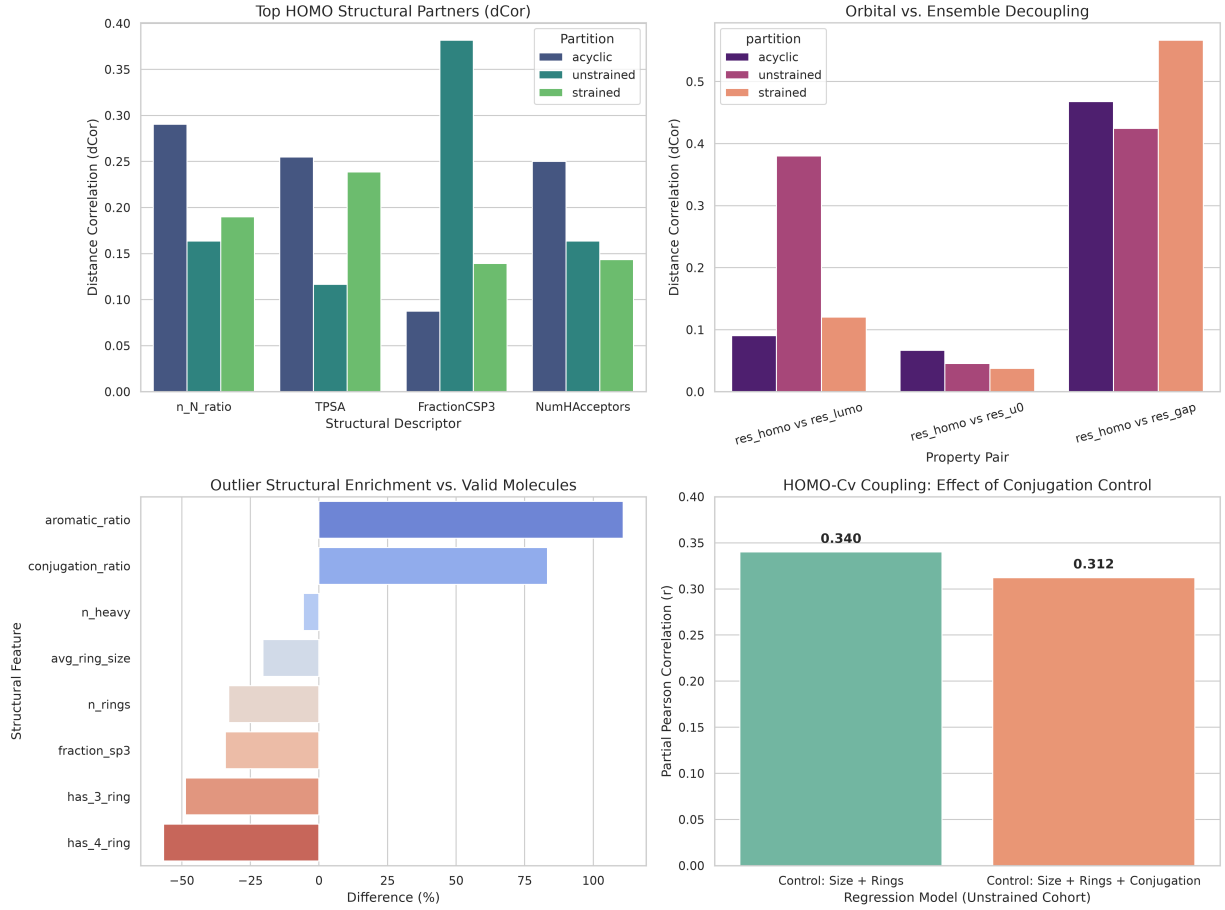

Figure 1: This figure illustrates the structural and electronic factors driving the decoupling of the Highest Occupied Molecular Orbital (HOMO) energy residual (*res_homo*) in the QM9 dataset. The top-left panel displays the distance correlation of *res_homo* with key structural descriptors, revealing that nitrogen content and polar surface area strongly influence acyclic and strained systems, while carbon saturation dominates in unstrained molecules. The top-right panel demonstrates orbital decoupling dynamics, showing that *res_homo* is moderately coupled to the HOMO-LUMO gap (*res_gap*) but remains largely orthogonal to total internal energy (res_u_0) across all families, whereas its coupling to *res_lumo* depends heavily on structural constraints. The bottom-left panel profiles statistical outliers, highlighting a significant enrichment in aromaticity (110.89%) and conjugation (83.35%) alongside a depletion in small rings and saturated carbons, indicating where HOMO decoupling is most likely to break down. Finally, the bottom-right panel shows the partial correlation between *res_homo* and heat capacity (res_c_v) in the unstrained cohort; the minimal change when controlling for conjugation (from $r = 0.340$ to $r = 0.312$) suggests that electron delocalization is not the sole mediator of property re-coupling.

acyclic chains. This suggests that structural constraints dictate the degree to which frontier orbitals couple with one another.

3.4 The influence of conjugation and structural strain

Given the structural dependence of orbital coupling, we utilized partial correlation analysis within the unstrained cohort to test the role of electron delocalization in mediating these relationships (Figure 1, bottom-right panel). When explicitly controlling for molecular size and ring count, the baseline correlation between *res_homo* and res_c_v is $r = 0.340$. When the conjugation ratio is introduced as an additional covariate, the correlation remains largely stable, shifting only slightly to $r = 0.312$. This minimal attenuation indicates that while conjugation is a defining chemical feature of these systems, it is not the sole mediator of the property re-coupling observed in specific structural subsets.

3.5 Statistical outlier profiling and the breakdown of decoupling

To identify the geometric and architectural boundaries where frontier-orbital independence breaks down, we profiled the "stress test" cohort of 3,054 multivariate outliers isolated via the Isolation Forest anomaly score (Figure 1, bottom-left panel). Structural profiling reveals that these outliers are not randomly distributed but are highly enriched in specific electronic and geometric motifs.

Compared to the valid molecular dataset, the outlier cohort exhibits a 110.89% increase in aromatic ratio and an 83.35% increase in conjugation ratio. Conversely, these molecules are significantly depleted in small rings (showing a 48.67% reduction in 3-membered rings and a 56.61% reduction in 4-membered rings) as well as saturated carbons (a 34.14% reduction in sp^3 fraction).

Within this highly conjugated and aromatic outlier group, the correlation between *res_homo* and res_c_v shifts to $r = -0.114$. This is a notable deviation from the near-zero correlation ($r = -0.001$) observed in a structurally matched valid population. This shift confirms that the fundamental decoupling of the HOMO energy is most likely to break down in systems characterized by extreme electron delocalization, which may simultaneously represent the convergence limits of the B3LYP/6-31G(2df,p) functional within the QM9 dataset.

4 Conclusions

We set out to determine whether the observed decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from the core spatial-thermodynamic manifold is a fundamental quantum invariant or an artifact of structural constraints and algorithmic noise. By systematically removing stoichiometric and size-dependent variance across the QM9 dataset, our topological sensitivity analysis confirms that frontier orbital independence is indeed a robust physical rule in standard small organic molecules. While macroscopic properties such as heat capacity and zero-point vibrational energy form the highly interconnected hubs of the property manifold, the HOMO residual operates on a statistically independent axis, exhibiting a network centrality indistinguishable from random noise in acyclic and unstrained systems.

Rather than being governed by global geometric constraints or spatial extent, the variance of the HOMO energy is dictated by local electronic environments. Our findings demonstrate that nitrogen content, topological polar surface area, and the degree of carbon saturation act as the primary structural partners of the HOMO axis. Crucially, the relative influence of these local features is heavily dependent on the underlying ring topology, with carbon saturation dominating in unstrained rings and heteroatom identity driving variance in acyclic chains.

Furthermore, we established that this decoupling is not a uniform feature of all frontier orbitals. As summarized in Table 1.3, while the HOMO residual remains strictly orthogonal to the total internal energy (res_u_0) across all structural families, its coupling to the LUMO residual (*res_lumo*) is highly sensitive to structural constraints. This reveals that frontier orbitals do not universally behave as a single decoupled block; instead, their internal coupling dynamics are dictated by the topological flexibility of the molecule.

Finally, by stress-testing this decoupling against a statistically isolated cohort of multivariate outliers, we identified the specific architectural boundaries where this physical rule breaks down. The re-coupling of the HOMO energy to the thermodynamic manifold occurs primarily in highly conjugated and aromatic

systems. This suggests that extreme electron delocalization disrupts frontier orbital independence, marking either a fundamental shift in the physical behavior of highly delocalized systems or the convergence limits of the underlying B3LYP density functional theory approximations.

Ultimately, this framework moves beyond simple correlation surveys to mathematically isolate the exact geometric and architectural boundaries of HOMO decoupling. These insights refine our understanding of joint molecular property manifolds and provide critical boundary conditions for the development of more robust, physics-aware machine learning models in computational chemistry.