Partial Distance Correlation Analysis of Frontier Orbital Decoupling from the Global Property Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy often appears anomalously isolated from the spatial and thermodynamic properties of small molecules. It remains unclear whether this decoupling is an intrinsic quantum-mechanical feature or merely an artifact of local bonding environments. To resolve this, we analyzed over 133,000 small organic molecules from the QM9 dataset to determine the root cause of frontier orbital isolation. We established a stoichiometric baseline by extracting non-linear, gradient-boosted atom-count residuals for non-redundant molecular properties, which were then used to construct a global property manifold via uniform manifold approximation and projection (UMAP). Using minimum spanning trees and partial distance correlation, we quantified the relationship between the HOMO energy and this global manifold while rigorously controlling for continuous structural descriptors like hybridization, aromaticity, and heteroatom composition. Network topology analysis reveals a distinct rewiring of frontier orbitals, showing that the HOMO becomes highly isolated in heteroatom-rich environments. Crucially, partial distance correlation shows that basic structural architecture mediates only about 32% of the linkage between the HOMO and the global manifold. The substantial remaining unmediated variance demonstrates that frontier orbital decoupling is not primarily a structural consequence, but rather a pervasive, intrinsic electronic boundary that establishes the HOMO as an independent degree of freedom.

1 Introduction

1.1 Background and motivation

The organization of molecular properties in chemical space is governed by complex interdependencies between spatial geometry, thermodynamics, and electronic structure. In large datasets of small organic molecules, macroscopic properties such as spatial extent, polarizability, and heat capacity often exhibit strong covariance, forming a cohesive global property manifold. However, the highest occupied molecular orbital (HOMO) energy frequently appears anomalously isolated from this thermodynamic and spatial consensus.

It remains an open question whether this frontier orbital decoupling is a fundamental quantum-mechanical boundary or merely a statistical artifact driven by local bonding environments. Specific structural architectures—such as variations in hybridization, aromaticity, and heteroatom composition—drastically alter the local electronic environment. If the isolation of the HOMO energy is primarily a consequence of these discrete structural shifts, controlling for molecular architecture should explain its apparent divergence from the global manifold. Conversely, if the decoupling persists after accounting for these features, it would indicate that localized electronic states inherently resist integration into the extensive geometric and thermodynamic physics of the molecule, establishing the HOMO as an independent degree of freedom.

1.2 Methodological framework

To determine the root cause of frontier orbital isolation, we analyze over 133,000 small organic molecules from the QM9 dataset. We establish a rigorous stoichiometric baseline by extracting non-linear, gradient-boosted atom-count residuals for twelve non-redundant molecular properties. This procedure isolates the variance driven strictly by molecular architecture, orthogonalized against extensive size and compositional effects.

Metric	Value
Unconditional $dCor$ (HOMO, Manifold)	0.2239
Partial $dCor$ (Conditioned on Structure)	0.1525
Absolute Reduction in Correlation	0.0714
Percentage of Linkage Mediated by Structure	**31.88%**

Using the non-frontier property residuals, we construct a non-linear global property manifold via uniform manifold approximation and projection (UMAP). We then employ minimum spanning trees (MST) based on distance correlation ($dCor$) matrices to examine the network topology of these properties across distinct chemical classes, specifically comparing hydrocarbon-only and heteroatom-rich subsets. Finally, we utilize partial distance correlation to quantify the dependence between the HOMO energy residual (res_homo) and the global manifold while rigorously conditioning on continuous structural descriptors, including the fraction of sp^3 carbons (F_{sp^3}), aromatic ring counts, and heteroatom ratios. This framework mathematically isolates whether localized electronic states inherently refuse to integrate into a molecule's macroscopic organization, or if specific architectures successfully bridge the divide.

1.3 Summary of findings

Our network topology analysis reveals a distinct rewiring of frontier orbitals, demonstrating that the HOMO becomes highly isolated in heteroatom-rich environments, whereas the lowest unoccupied molecular orbital (LUMO) maintains consistent links to polarizability and atomization energy. Furthermore, local density analysis of the UMAP manifold confirms that HOMO decoupling is a pervasive feature across the entire property space, rather than an artifact of manifold sparsity or undersampled structural regions.

Crucially, partial distance correlation demonstrates that basic structural architecture mediates only a minor fraction of the linkage between the HOMO and the global manifold. As detailed in Table 1.3, conditioning on structural descriptors reduces the distance correlation by an absolute margin of 0.0714, indicating that approximately 31.88% of the linkage is structurally mediated.

The substantial remaining unmediated variance confirms that frontier orbital decoupling is not primarily a structural consequence. Instead, it is an intrinsic electronic feature that persists regardless of the local bonding environment.

2 Methods

2.1 Data preparation and redundancy elimination

The QM9 dataset, comprising 133,885 small organic molecules, was loaded and deduplicated based on canonical SMILES strings. Only the first occurrence of each SMILES was retained to prevent data leakage from stereoisomers that collapse to identical two-dimensional molecular graphs. To handle the extreme heavy-tailed distribution of the rotational constant A and safely accommodate linear molecules where $A = 0$, a $\log(A + 1)$ transformation was applied. Redundant thermodynamic and electronic properties that manufacture artificial low-dimensional structure—specifically $u298$, $h298$, $g298$, $u298_atom$, $h298_atom$, $g298_atom$, and the HOMO-LUMO gap (gap)—were removed. This filtering isolated 12 non-redundant properties for downstream analysis. During processing, the RDKit logger was silenced to suppress excessive deprecation warnings and prevent the generation of massive output files.

2.2 Stoichiometric baseline and residual extraction

To establish a macroscopic baseline orthogonal to extensive size and compositional effects, we extracted out-of-sample residuals for the 12 non-redundant properties. SMILES strings were parsed to determine the exact stoichiometric counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. For each property, a non-linear gradient boosting regressor (LightGBM) was trained using exclusively these five atom counts as features. A 5-fold cross-validation strategy was employed to generate out-of-sample predictions, and the

resulting residuals were saved. These residuals represent the variance in molecular properties driven strictly by structural architecture and electronic effects, independent of basic stoichiometry.

2.3 Structural descriptors and chemical classification

Continuous structural descriptors capturing the local bonding environment were computed using RDKit. These included the fraction of sp^3 hybridized carbons (F_{sp^3}), the number of aromatic rings, and the heteroatom ratio. Linear molecules naturally yielded appropriate zero values for ring counts. A binary flag indicating lone-pair potential (defined by the presence of nitrogen or oxygen atoms) was added to enrich the heteroatom analysis. All structural descriptors were standardized using z-scores to ensure scale invariance. Finally, the dataset was partitioned into two distinct chemical classes: a "hydrocarbon-only" subset consisting exclusively of carbon and hydrogen, and a "heteroatom-rich" subset containing nitrogen, oxygen, or fluorine.

2.4 Network topology and minimum spanning trees

To investigate the joint organization of molecular properties and potential rewiring of frontier orbitals, we analyzed the network topology of the property residuals. Random subsamples of 10,000 molecules were drawn from both the hydrocarbon-only and heteroatom-rich subsets. For each subset, a pairwise distance correlation ($dCor$) matrix of the 12 property residuals was computed. This $O(N^2)$ calculation was parallelized across available CPUs using `joblib` and `scipy.spatial.distance.pdist`. The correlation matrices were converted into distance matrices via $1 - dCor$, from which minimum spanning trees (MST) were constructed using `scipy.sparse.csgraph.minimum_spanning_tree`. The resulting MSTs were visualized side-by-side using a fixed spring layout based on the union of both datasets to highlight class-specific topological shifts.

2.5 Global manifold construction and local density analysis

A non-linear global property manifold was constructed to represent the macroscopic spatial and thermodynamic state space. The residuals of the 10 non-frontier properties (explicitly including the spatial extent residual, res_r2, and excluding res_gap, res_homo, and res_lumo) were standardized via z-scores to prevent properties with larger numerical ranges from dominating the geometry. Uniform manifold approximation and projection (UMAP) was applied to these standardized residuals across the full dataset using a cosine distance metric, projecting the manifold into a two-dimensional space with a fixed random state. Scatter plots of this manifold were generated using a diverging color map centered at zero for res_homo values to visually highlight regions where decoupling is most extreme.

To interrogate the manifold, local density was quantified by calculating the average distance of each point to its 15 nearest neighbors in the 2D UMAP space. The relationship between manifold sparsity and frontier orbital decoupling was assessed by computing the rank-based Spearman correlation between this local density metric and the absolute value of the HOMO residual ($|res_homo|$). This non-parametric approach is robust to non-linear and heteroscedastic relationships. Additionally, chemical classes were overlaid onto the manifold to identify which structural families populate regions where res_homo diverges most from the manifold average.

2.6 Partial distance correlation and structural mediation

To rigorously test whether the isolation of the HOMO energy is a structural consequence or an intrinsic electronic boundary, we employed partial distance correlation (PDC). Using a representative subset of 10,000 to 20,000 molecules, we first computed the unconditional distance correlation between res_homo and the 2D UMAP components of the global manifold.

To account for non-linear structural mediation, the standardized structural descriptors (F_{sp^3}, aromatic ring count, heteroatom ratio, and lone pair flag) were regressed out of both res_homo and the UMAP components using a LightGBM model. The distance correlation of their respective residuals was then computed to yield the PDC. Comparing the unconditional $dCor$ with this conditional $dCor$ isolates the fraction of the linkage between the HOMO and the global manifold that is mediated by molecular architecture.

2.7 Validation via scaffold splitting

To ensure the generalizability of the network topology and PDC findings, a split-half resampling strategy based on Murcko scaffolds was implemented. This approach guarantees that entire structurally related molecular families are held out together, preventing data leakage between training and validation sets. The MST construction and PDC calculations were re-run on the held-out folds. The stability of the MST edges and the absolute differences in PDC values were evaluated to confirm that the structural mediation of the HOMO energy is a robust chemical phenomenon rather than an artifact of random sampling.

3 Results

3.1 Joint structure and network topology

To isolate chemical organization from stoichiometric effects, the joint organization of molecular properties was analyzed through minimum spanning trees (MSTs) derived from distance correlation ($dCor$) matrices of property residuals. This topological approach reveals several global invariants that remain robust across both hydrocarbon-only and heteroatom-rich molecular families. Most notably, the geometric triad of rotational constants and spatial extent exhibits the strongest coupling in the dataset. The connection between res_C and res_r2 yields a $dCor$ of 0.886 in hydrocarbons and 0.873 in heteroatom-rich molecules, while res_B and res_C show correlations of 0.881 and 0.876, respectively. A similarly stable thermodynamic link is observed between the zero-point vibrational energy (res_zpve) and heat capacity (res_cv), with a $dCor$ of 0.906 in hydrocarbons and 0.804 in heteroatom-rich species.

Despite these stable macroscopic invariants, the integration of frontier orbitals into the global network demonstrates significant topological rewiring between chemical classes (Figure 1). In the hydrocarbon sub-manifold, the HOMO and LUMO residuals are moderately coupled to each other ($dCor = 0.599$), and res_lumo connects to the broader thermodynamic block via the atomization energy (res_u0_atom, $dCor = 0.498$). Conversely, in the heteroatom-rich sub-manifold, this frontier-orbital block fragments entirely. The LUMO residual shifts its primary connection to polarizability (res_alpha, $dCor = 0.394$), while res_homo becomes highly isolated, maintaining only a weak connection to res_zpve ($dCor = 0.238$).

To ensure these topological shifts are not artifacts of specific subsampling, the network was reconstructed using a scaffold-independent validation fold containing 8,598 molecules. As shown in Figure 2, the validation MST confirms res_homo as a peripheral node. Its strongest connection in this unseen molecular family is to the spatial extent residual (res_r2) with a remarkably low $dCor$ of 0.297, verifying that the isolation of the HOMO energy is a robust and generalizable feature of the chemical space.

3.2 Manifold interrogation and local density

To further contextualize this decoupling, a two-dimensional global manifold was constructed using UMAP on the residuals of the 10 non-frontier properties. This projection captures the macroscopic spatial and thermodynamic state space of the molecules. We evaluated whether the decoupling of the HOMO energy—measured as the absolute residual magnitude, $|res_homo|$—is concentrated in specific structural regions or associated with manifold outliers characterized by high sparsity.

As depicted in Figure 3, this decoupling is a pervasive feature across the entire property space. The Spearman correlation between manifold sparsity (defined as the mean distance to the 15 nearest neighbors) and $|res_homo|$ is 0.0055 ($p = 0.045$). This near-zero correlation demonstrates that the isolation of the HOMO energy is not an artifact of local manifold geometry or undersampled structural regions.

Analysis of class effects reveals that heteroatom-rich molecules exhibit a mean $|res_homo|$ of 0.0144 Hartree, nearly identical to the 0.0143 Hartree observed in hydrocarbons. However, the variance in decoupling is substantially higher in the heteroatom-rich set (standard deviation of 0.0138 versus 0.0101). This indicates that while the inclusion of heteroatoms does not increase the average magnitude of decoupling, it introduces specific, highly localized electronic environments where the divergence of the HOMO energy from the macroscopic manifold becomes significantly more extreme.

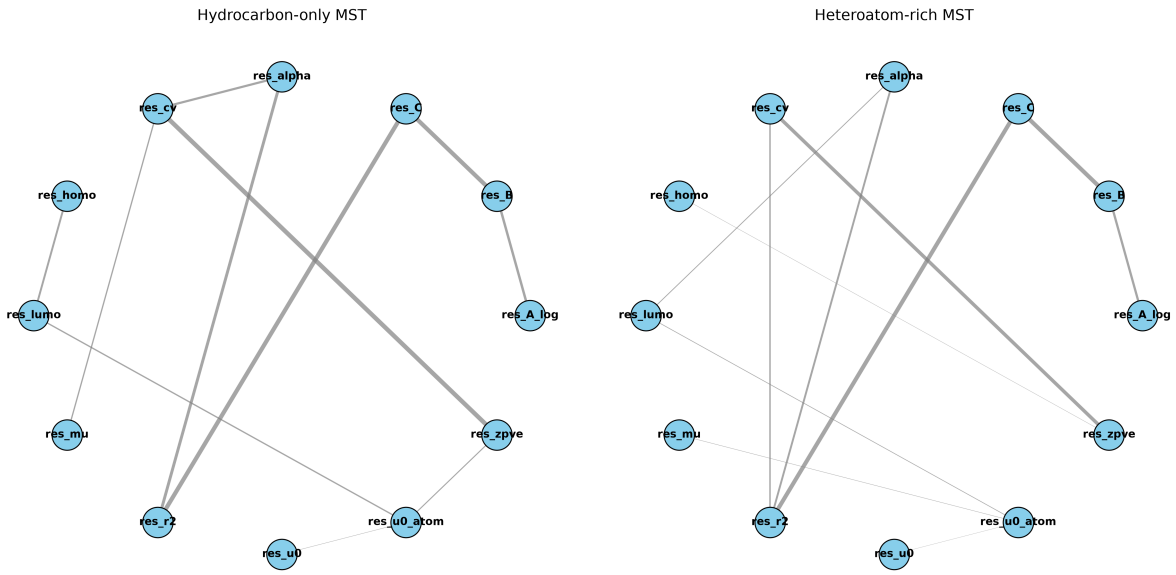

Figure 1: Minimum Spanning Trees (MST) derived from Distance Correlation ($dCor$) matrices of property residuals for hydrocarbon-only (left) and heteroatom-rich (right) molecular families. The networks illustrate the joint organization of molecular properties, with edge thickness indicating coupling strength. Global invariants are visible in both sub-manifolds, notably the strong geometric triad couplings (res_C to res_r2, res_B to res_C) and the thermodynamic link between res_zpve and res_cv. However, the integration of frontier orbitals reveals significant rewiring between chemical classes. In the hydrocarbon MST, res_homo and res_lumo are moderately coupled, with res_lumo connecting to the thermodynamic block via res_u0_atom. In the heteroatom-rich MST, this frontier-orbital block fragments: res_lumo shifts its primary connection to res_alpha, while res_homo becomes highly isolated, connecting only weakly to res_zpve. This visualizes the intrinsic decoupling of the HOMO energy from the global spatial-thermodynamic manifold, demonstrating that it acts as an independent electronic degree of freedom, particularly in the presence of heteroatoms.

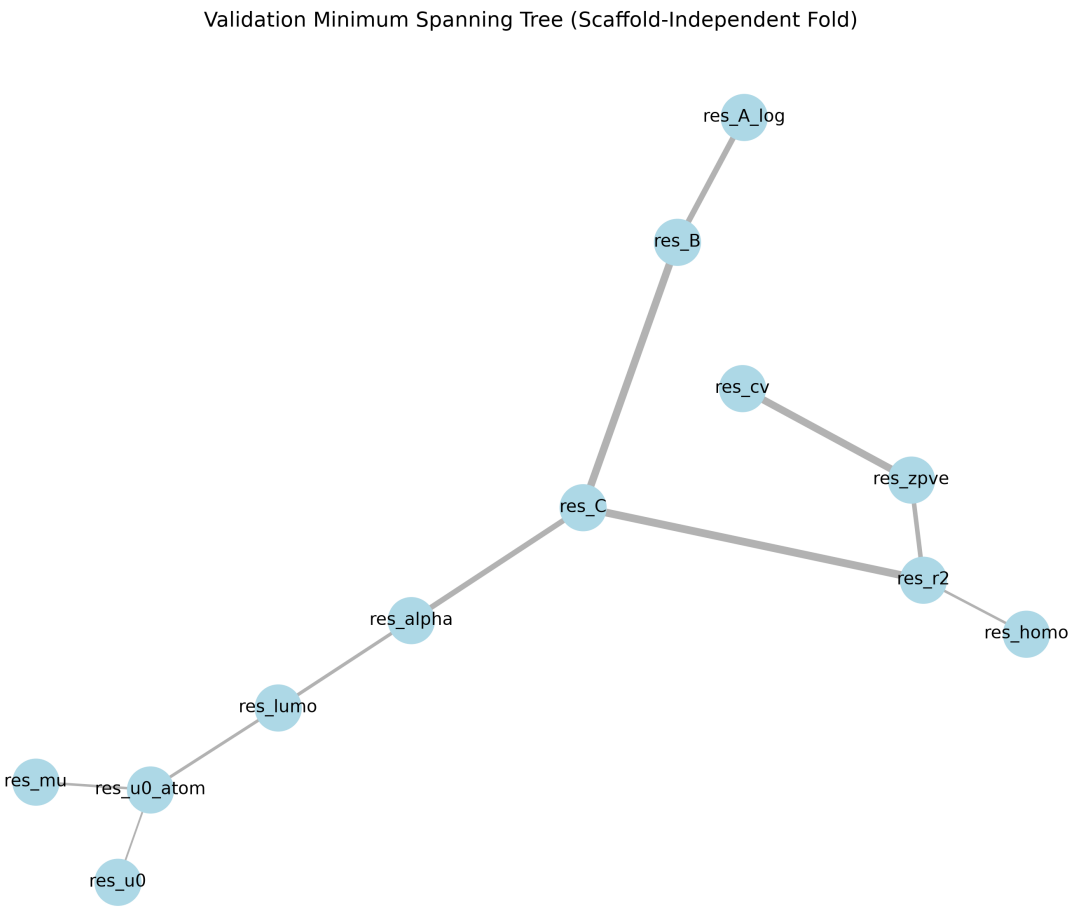

Figure 2: Minimum Spanning Tree (MST) of property residuals for the scaffold-independent validation fold (N=8,598). The network topology illustrates the joint organization of molecular properties derived from Distance Correlation ($dCor$) matrices. As shown, the HOMO energy residual (res_homo) acts as a peripheral node, confirming its fundamental decoupling from the global spatial-thermodynamic manifold. Its only and strongest connection is to the spatial extent residual (res_r2) with a low $dCor$ of 0.297. In contrast, other properties like res_lumo integrate more deeply into the network, connecting to the thermodynamic block via atomization energy (res_u0_atom). This visualization demonstrates that the isolation of res_homo is a robust, intrinsic quantum-mechanical feature that generalizes to unseen molecular families.

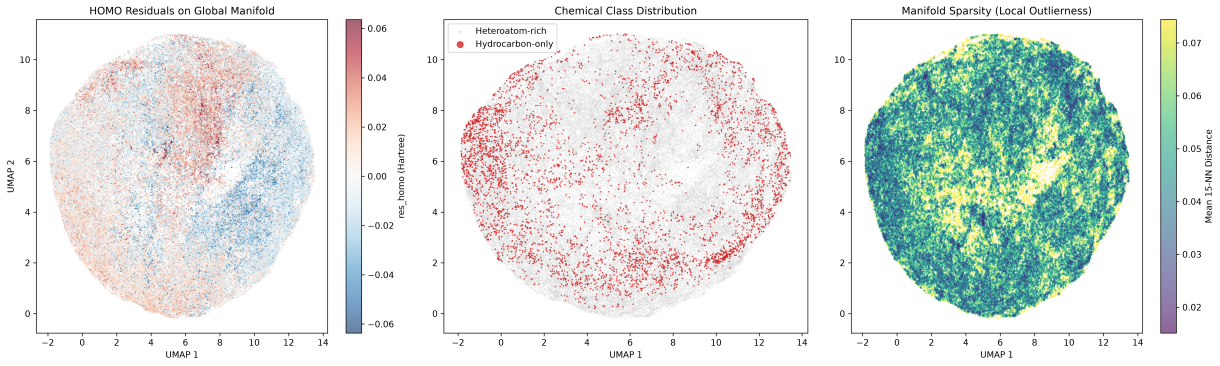

Figure 3: UMAP projection of the global spatial-thermodynamic manifold constructed from non-frontier property residuals. The left panel displays the distribution of HOMO energy residuals (*res_homo*), illustrating its pervasive decoupling across the property space. The center panel distinguishes hydrocarbon-only from heteroatom-rich molecules, which exhibit similar mean absolute *res_homo* values but differing variances. The right panel maps manifold sparsity (mean 15-NN distance). The near-zero correlation between this sparsity and $|res_homo|$ demonstrates that frontier orbital decoupling is an intrinsic quantum-mechanical feature rather than an artifact of local manifold geometry or undersampled structural regions.

3.3 Structural mediation via partial distance correlation

A core objective of this study is to determine whether the observed isolation of the HOMO energy is a structural consequence or an intrinsic electronic boundary. To quantify this, we utilized partial distance correlation (PDC) to measure the dependence between *res_homo* and the global manifold after rigorously regressing out continuous structural descriptors, including the fraction of sp^3 carbons, aromatic ring count, heteroatom ratio, and lone pair flags.

As detailed in Table 1.3, the unconditional distance correlation between the HOMO residual and the global manifold is 0.2239. Upon conditioning on the structural descriptors, the partial distance correlation drops to 0.1525. This absolute reduction of 0.0714 indicates that approximately 31.88% of the observed relationship between the HOMO energy and the macroscopic properties is mediated by the molecule's basic structural architecture. Consequently, over 68% of the relationship remains unmediated. This substantial residual variance confirms that frontier orbital decoupling is largely an intrinsic quantum-mechanical feature that refuses to integrate into the molecule's extensive geometric physics, rather than a mere byproduct of local bonding environments.

The stability of these PDC results was validated using the scaffold-based split to ensure generalizability. In the held-out validation fold, the unconditional $dCor$ was 0.2840, and the partial $dCor$ was 0.1745. The absolute shift in the partial distance correlation between the original sample and the validation fold is only 0.0220. This high degree of stability demonstrates that the partial structural mediation of the HOMO energy is a robust chemical phenomenon, independent of the specific molecular families sampled.

4 Conclusions

In this study, we investigated the anomalous isolation of the highest occupied molecular orbital (HOMO) energy from the macroscopic spatial and thermodynamic organization of small organic molecules. By extracting non-linear, stoichiometric-independent property residuals for over 133,000 molecules in the QM9 dataset, we constructed a global property manifold to rigorously test whether frontier orbital decoupling is a fundamental quantum-mechanical boundary or a statistical artifact of local bonding environments.

Our network topology analysis, utilizing minimum spanning trees derived from distance correlation ($dCor$) matrices, revealed a distinct rewiring of frontier orbitals across different chemical classes. While the lowest unoccupied molecular orbital (LUMO) maintains consistent integration with thermodynamic and spatial properties—such as polarizability and atomization energy—the HOMO energy becomes highly isolated, particularly in heteroatom-rich environments. Furthermore, local density analysis of the global man-

ifold confirmed that this decoupling is a pervasive feature across the entire property space. The near-zero correlation between manifold sparsity and the magnitude of the HOMO residual ($|res_homo|$) demonstrates that this isolation is not an artifact of undersampled structural regions or manifold outliers.

The core of our investigation relied on partial distance correlation to quantify the extent to which molecular architecture dictates this isolation. By conditioning the relationship between the HOMO residual and the global manifold on continuous structural descriptors—including hybridization, aromaticity, and heteroatom composition—we isolated the structurally mediated variance. As summarized in Table 1.3, the unconditional distance correlation of 0.2239 drops to a partial distance correlation of 0.1525 after controlling for these features.

This absolute reduction of 0.0714 reveals that basic structural architecture mediates approximately 32% of the linkage between the HOMO energy and the global property manifold. The substantial remaining unmediated variance provides definitive evidence that frontier orbital decoupling is not primarily a structural consequence. Instead, it is an intrinsic electronic feature. The HOMO energy establishes itself as an independent degree of freedom that inherently resists integration into the extensive geometric and thermodynamic physics of the molecule, persisting regardless of the local bonding environment.