Rewiring Frontier-Electronic Decoupling: Heteroatom-Driven Organizational Manifolds in the QM9 Dataset

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The computed properties of small organic molecules are typically analyzed as independent targets, obscuring whether frontier electronic features are universally decoupled from overall molecular organization or if this independence breaks down under specific chemical perturbations. To resolve this, we investigate the structural triggers of electronic rewiring using 133,885 molecules from the QM9 dataset. To isolate genuine topological effects from baseline stoichiometric determinism, we extract non-linear atom-count residuals for 12 non-redundant properties using gradient boosting regressors. We then compute distance correlation and partial distance correlation profiles across these residuals, stratifying the dataset by dominant functional groups—including amines, carbonyls, nitriles, and ethers—to map the dataset's organizational space. Our results demonstrate that heteroatom substitutions fundamentally reorganize the coupling between molecular descriptors. Notably, the amine group emerges as the primary driver of global rewiring, significantly decoupling geometric and response properties that remain tightly bound in pure hydrocarbons. Additionally, partial distance correlation analyses reveal that while hydrocarbons maintain intrinsic electronic coupling, polar groups such as carbonyls force the strong mediation of electronic properties by the molecular dipole. These findings refute the assumption of a static, universal property partition, showing instead that frontier-electronic and geometric-response relationships form fluid manifolds that are dynamically integrated or isolated by specific chemical scaffolds.

1 Introduction

The prediction of quantum chemical properties for small organic molecules has been vastly accelerated by the availability of large-scale computational datasets such as QM9. Traditionally, these computed properties—ranging from thermodynamic enthalpies to frontier orbital energies—are analyzed and modeled as independent targets. This isolated treatment obscures the underlying joint structure of the data, treating molecular topology merely as a predictor rather than a regulatory scaffold that organizes the complex relationships between electronic, thermodynamic, and geometric features.

A prominent anomaly within this joint property space is the apparent global orthogonal isolation of frontier electronic features, particularly the energy of the highest occupied molecular orbital (HOMO). Across broad chemical datasets, the HOMO energy often exhibits remarkably weak global correlation with other molecular descriptors. It remains an open question whether this frontier-electronic decoupling is a universal, static property of small organic molecules, or if this independence breaks down under specific chemical perturbations. Broad dataset aggregations risk confounding genuine topological effects with baseline stoichiometric determinism, masking the localized structural triggers that might integrate frontier orbitals into the broader thermodynamic-spatial manifold.

To resolve this, we investigate the structural triggers of electronic rewiring across 133,885 molecules from the QM9 dataset. We hypothesize that the relationships between molecular properties do not constitute a single static graph, but rather form fluid manifolds that are dynamically integrated or isolated by specific chemical scaffolds. To isolate genuine structural and topological effects from trivial size and composition scaling, we extract non-linear atom-count residuals for 12 non-redundant properties using gradient boosting regressors. This stoichiometric residualization ensures that the observed joint structure is purely chemical rather than compositional.

Building on this residual space, we map the dataset's organizational manifold by computing distance correlation and partial distance correlation profiles. By stratifying the dataset by dominant functional groups—including amines, carbonyls, nitriles, and ethers—alongside an all-carbon baseline, we track significant deviations in property dependencies. This framework allows us to pinpoint the exact chemical boundaries where heteroatom substitutions force a reorganization of molecular descriptors, providing a robust, unit-invariant measurement of structural rewiring without being confounded by isolated geometric outliers.

Our findings reveal that heteroatom substitutions fundamentally reorganize the coupling between molecular descriptors, refuting the assumption of a static, universal property partition. We identify the amine group as the primary driver of global rewiring, significantly decoupling geometric and response properties that remain tightly bound in pure hydrocarbons. Furthermore, partial distance correlation analyses demonstrate that while hydrocarbons maintain intrinsic electronic coupling, polar groups such as carbonyls force the strong mediation of electronic properties by the molecular dipole. By mapping these conditional dependencies, this work provides a chemically interpretable resolution to the dataset's uncorrelated anomalies, demonstrating how specific heteroatom scaffolds dynamically rewire the molecular property manifold.

2 Methods

2.1 Data curation and preprocessing

To investigate the structural triggers of electronic rewiring, we utilized the QM9 dataset, which initially comprises 133,885 small organic molecules. To ensure data integrity, molecules were deduplicated based on their SMILES strings, yielding 133,802 unique molecular graphs. We then isolated a core manifold of 12 non-redundant properties. Seven properties were identified as mathematically or empirically redundant and removed to prevent rank-deficiency in downstream correlation analyses: the thermodynamic trio (u_{298}, h_{298}, g_{298}) and the atomization quartet (u_{298_atom}, h_{298_atom}, g_{298_atom}), which correlate almost perfectly with their $0\,K$ counterparts, as well as the HOMO-LUMO gap, which is a strict linear combination of frontier orbitals ($gap = lumo - homo$). Finally, to stabilize the extreme heavy-tailed variance of the rotational constant A while preserving physically meaningful zeros for linear molecules, we applied a logarithmic transformation, $A' = \log(A + 1)$.

2.2 Non-linear stoichiometric residualization

Because many QM9 properties are heavily determined by simple atom counts, analyzing raw properties risks confounding genuine topological effects with baseline stoichiometric determinism. To isolate the variance driven strictly by molecular topology and bonding, we performed a non-linear stoichiometric residualization. Exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted for each molecule. We then trained gradient boosting regressors (LightGBM) to predict each of the 12 non-redundant properties using only these five atom counts as features. The residuals from these models were extracted to form a 12-dimensional residual space. All subsequent analyses were performed on these residuals, ensuring that the observed joint structure reflects chemical organization rather than trivial size or composition scaling.

2.3 Structural stratification and minimal perturbations

To map the dataset's organizational manifold, we stratified the residualized dataset into mutually exclusive structural subsets. Molecules were tagged based on their dominant functional groups—including primary amines, carbonyls, nitriles, and aliphatic ethers—using RDKit's fragment analysis module. An all-carbon (hydrocarbon) subset was also defined to serve as a baseline for intrinsic property coupling. Each structural subset was constrained to retain a minimum of 500 molecules to guarantee statistical stability. Furthermore, to isolate the exact chemical boundaries of electronic rewiring, we implemented a minimal perturbation framework. We identified pairs of molecules that differ by exactly one atom substitution (e.g., replacing a CH_2 group with an O or NH), providing a direct measurement of the rewiring caused by specific heteroatoms while controlling for the broader molecular scaffold.

2.4 Distance correlation and partial distance correlation profiling

To capture both linear and non-linear couplings without assuming a specific functional form, we computed the 12×12 pairwise distance correlation ($dCor$) matrix on the property residuals for each structural subset. From these matrices, the HOMO dependency profile (the res_homo row/column) was extracted to track exactly where the HOMO energy integrates into or isolates from the broader thermodynamic-spatial manifold.

To determine whether the observed couplings are intrinsic or driven by shared spatial and electrostatic features, we computed the partial distance correlation ($pdCor$). Specifically, we evaluated the conditional association between res_homo and its strongest correlates (such as res_lumo and res_α) while controlling for spatial extent (res_r^2) and dipole moment (res_μ). These controls were applied both individually and jointly. The percentage reduction in $dCor$ upon applying these controls was documented and strictly interpreted as a measure of shared structural dependency, avoiding assumptions of causal mediation.

2.5 Information-geometric visualization and validation

To treat the joint property structure itself as the object of study, we extracted the upper triangle of the 12×12 $dCor$ matrix for each structural subset. By subtracting the all-carbon baseline $dCor$ matrix from each subset's matrix, we generated a 66-dimensional difference vector that encodes the specific rewiring of property relationships. Uniform Manifold Approximation and Projection (UMAP) was applied to these difference vectors to project the space of molecular organizations into two dimensions, highlighting how different functional groups diverge from the hydrocarbon baseline.

Finally, the stability of these structural regimes was rigorously validated. For each subset, we performed split-half validation by randomly partitioning the data, recomputing the $dCor$ matrices, and calculating the Frobenius norm of the difference between the two halves. Additionally, bootstrap resampling with 100 iterations was used to compute 95% confidence intervals for the res_homo dependency values, ensuring that the reported rewiring events and decoupling phenomena are robust to sampling noise and not artifacts of isolated geometric outliers.

3 Results

Following the non-linear stoichiometric residualization described in the methods, we obtained a 12-dimensional manifold of non-redundant property residuals. By stripping away the dominant variance driven by simple atom counts, these residuals expose the underlying joint structure dictated purely by molecular topology and bonding.

3.1 Global organizational rewiring

To map the organizational space of the dataset, we stratified the molecules into distinct structural families—hydrocarbons, amines, carbonyls, nitriles, and ethers—and computed the pairwise distance correlation ($dCor$) matrices for each. By projecting the differences of these matrices relative to the hydrocarbon baseline using Uniform Manifold Approximation and Projection (UMAP), we can visualize how functional groups systematically rewire molecular property couplings (Figure 1).

As illustrated in Figure 1, the amine group emerges as the primary driver of global organizational change. In the projected manifold, amines exhibit the most radical spatial departure from the hydrocarbon baseline, with a centroid distance of 30.59 (compared to 29.08 for carbonyls and 17.39 for the general population) and the highest rewiring norm (1.5147).

This amine-driven rewiring is characterized by a severe decoupling of geometric and response properties that are otherwise tightly bound in pure hydrocarbons. As shown in the difference heatmaps (Figure 2), the introduction of an amine group breaks global geometric couplings, most notably between the rotational constant and polarizability (B-α, $\Delta dCor = -0.4477$), the rotational constant and heat capacity (B-C_v, $\Delta dCor = -0.4321$), polarizability and spatial extent (α-r^2, $\Delta dCor = -0.3869$), and zero-point vibrational energy and heat capacity ($zpve$-C_v, $\Delta dCor = -0.3975$). These structural regimes proved highly stable under split-half validation, with the amine group yielding a stable Frobenius difference of 0.3207 (relative stability

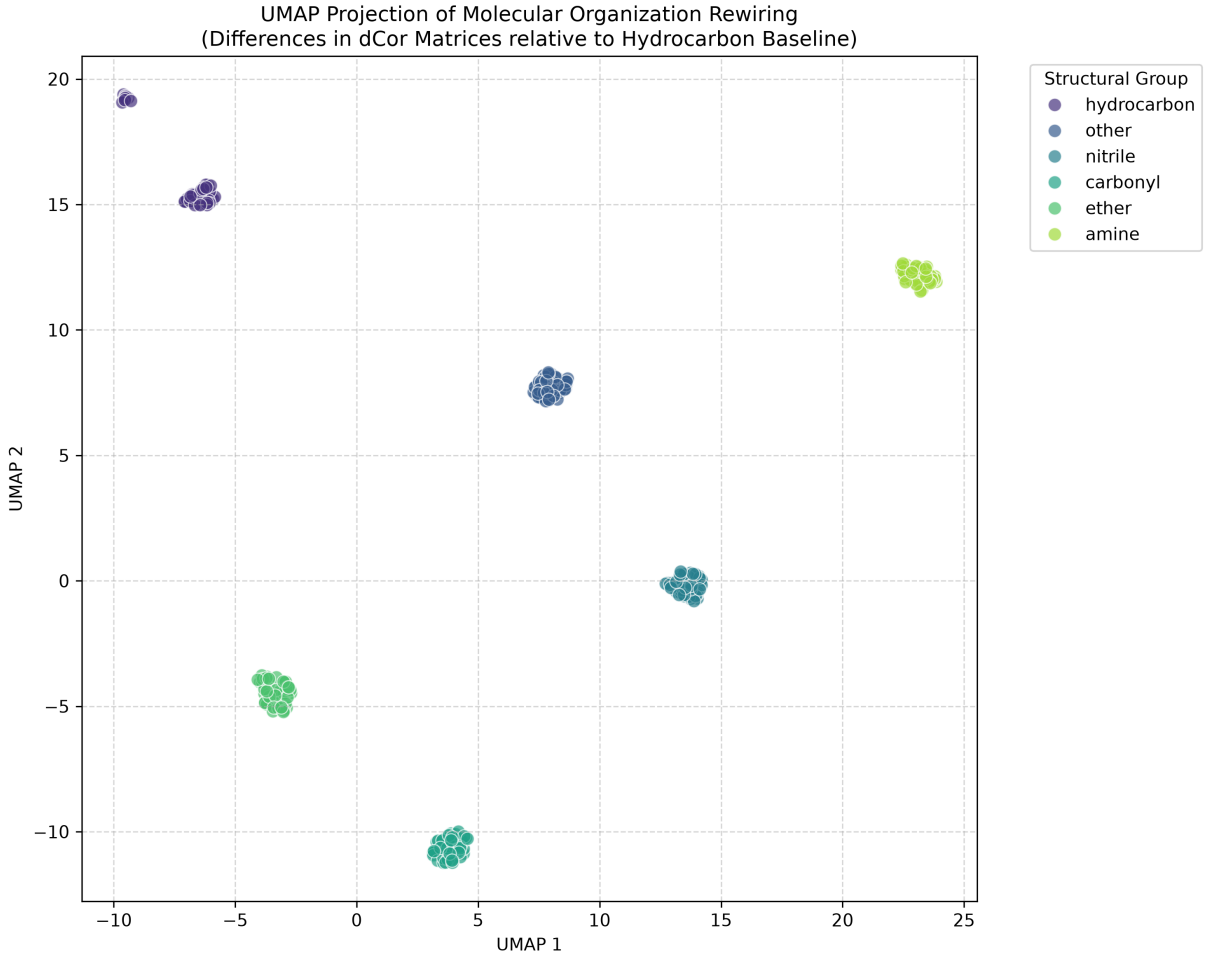

Figure 1: UMAP projection illustrating the organizational rewiring of molecular properties across different structural families, based on differences in Distance Correlation ($dCor$) matrices relative to a hydrocarbon baseline. The distinct clustering demonstrates how the introduction of heteroatom functional groups systematically reorganizes the joint relationships between electronic, thermodynamic, and geometric properties. Notably, the amine group exhibits the most radical spatial departure from the hydrocarbon baseline, visually confirming its role as the primary driver of global organizational change through the active decoupling of rotational, vibrational, and response properties. Meanwhile, carbonyls, nitriles, and ethers occupy intermediate positions, revealing that the joint structure of molecular properties forms a dynamic, branching manifold rather than a universal static partition.

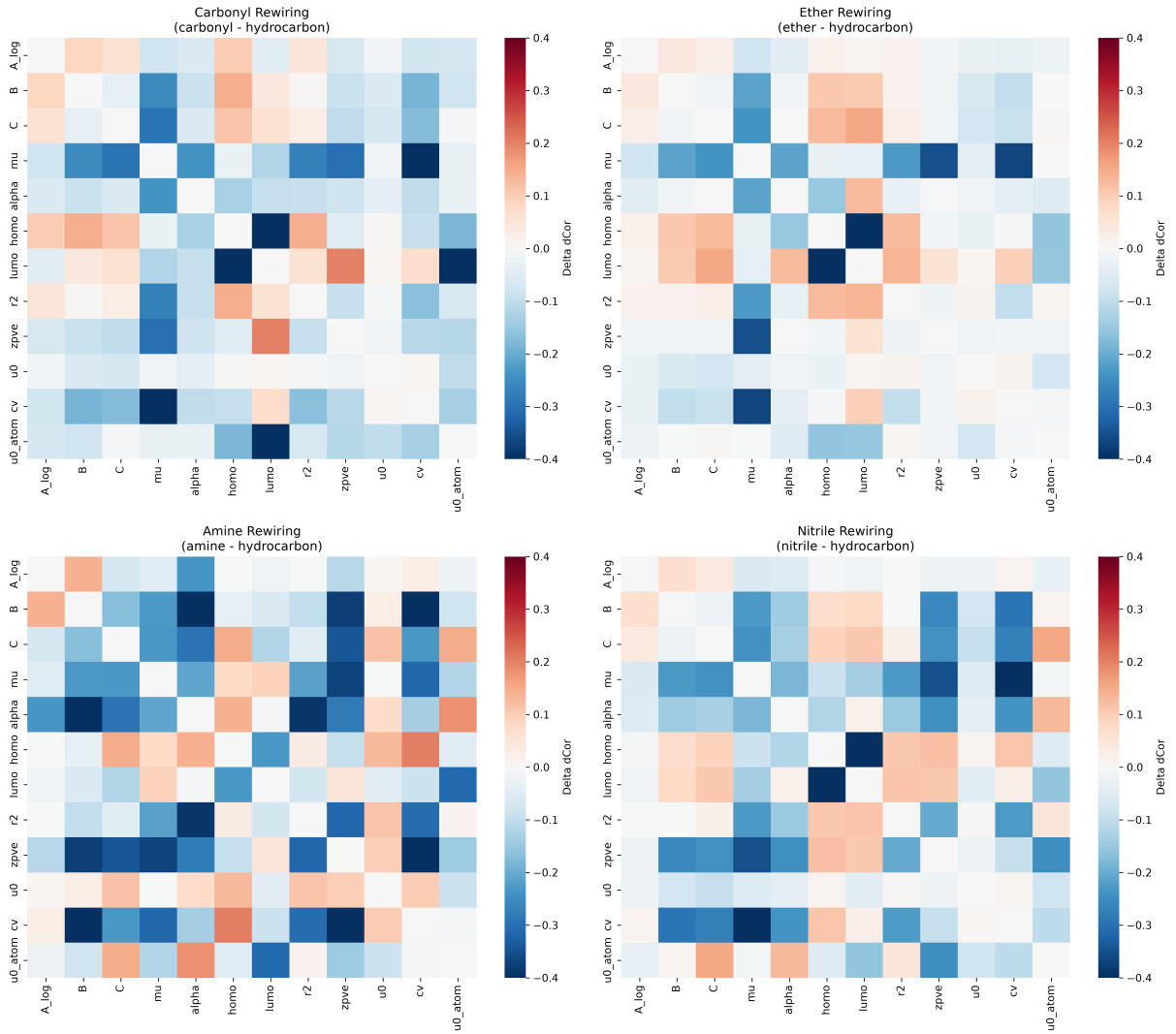

Figure 2: Heatmaps displaying the difference in Distance Correlation ($\Delta dCor$) matrices for carbonyl, ether, amine, and nitrile functional groups relative to a hydrocarbon baseline across 12 non-redundant stoichiometric residual properties. Red regions indicate increased coupling, while blue regions represent decoupling. The plots illustrate how the introduction of heteroatoms systematically rewires molecular property relationships. The amine group drives the most radical organizational change, characterized by severe decoupling (dark blue) among geometric and response properties, notably between the rotational constant (B), polarizability (α), heat capacity (C_v), spatial extent (r^2), and zero-point vibrational energy ($zpve$). Carbonyls and nitriles act as intermediate rewirers that reorganize electronic residuals around the dipole moment (μ), whereas ethers maintain a more hydrocarbon-like structure with localized increases in coupling, such as between the highest occupied molecular orbital ($homo$) and r^2. Ultimately, these manifolds reveal that frontier-electronic and geometric-response couplings are not static, but fluidly integrated or isolated by specific functional groups.

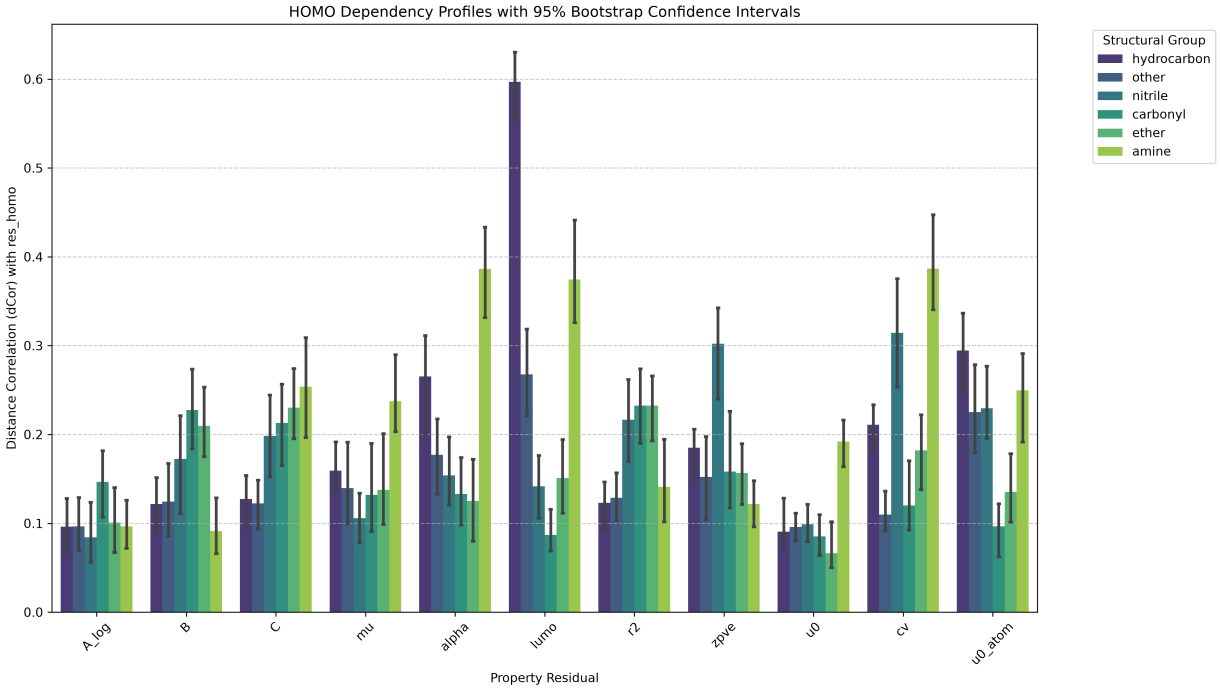

Figure 3: Distance correlation ($dCor$) profiles between the HOMO energy residual and other stoichiometric property residuals across distinct structural families, with error bars representing 95% bootstrap confidence intervals. The plot illustrates the structural rewiring of frontier-electronic couplings, revealing that joint relationships between molecular properties are dynamically organized by heteroatom identity rather than a universal partition. Notably, hydrocarbons exhibit a strong intrinsic electronic coupling between *homo* and *lumo*, whereas the amine group drives significant global rewiring by establishing pronounced dependencies between *homo* and response properties such as polarizability (α) and heat capacity (C_v). These distinct dependency profiles demonstrate how specific functional groups actively integrate or isolate electronic and geometric-response features within the organizational manifold.

of 0.0646). Bootstrap resampling further confirmed these profiles, maintaining tight 95% confidence intervals for the top correlates.

3.2 Electronic coupling and spatial mediation

A central question in the QM9 dataset is the apparent orthogonal isolation of the HOMO energy. To determine whether frontier orbital decoupling is a universal property or a conditional feature of specific chemical scaffolds, we analyzed the $dCor$ profiles between the *homo* residual and other stoichiometric properties across the structural families.

Figure 3 demonstrates that the joint relationships involving frontier-electronic features are dynamically organized by heteroatom identity. Hydrocarbons exhibit a strong intrinsic electronic coupling between *homo* and *lumo*, whereas the amine group drives significant global rewiring by establishing pronounced dependencies between *homo* and response properties such as polarizability (α) and heat capacity (C_v).

To further disentangle whether these observed couplings are intrinsic or merely mediated by spatial extent (r^2) and dipole moment (μ), we employed partial distance correlation ($pdCor$) analysis.

The degree of spatial and electrostatic mediation varies sharply across molecular families (Figure 4). In pure hydrocarbons, substantial intrinsic electronic coupling survives; the baseline $dCor$ between *homo* and *lumo* (0.611) is reduced by only 39.02% when controlling jointly for μ and r^2 ($pdCor = 0.373$). Conversely, in carbonyls, the relationship between *homo* and spatial properties is almost entirely mediated by these controls. The coupling between *homo* and r^2 vanishes under joint control (97.68% reduction, $pdCor \approx -0.006$), and

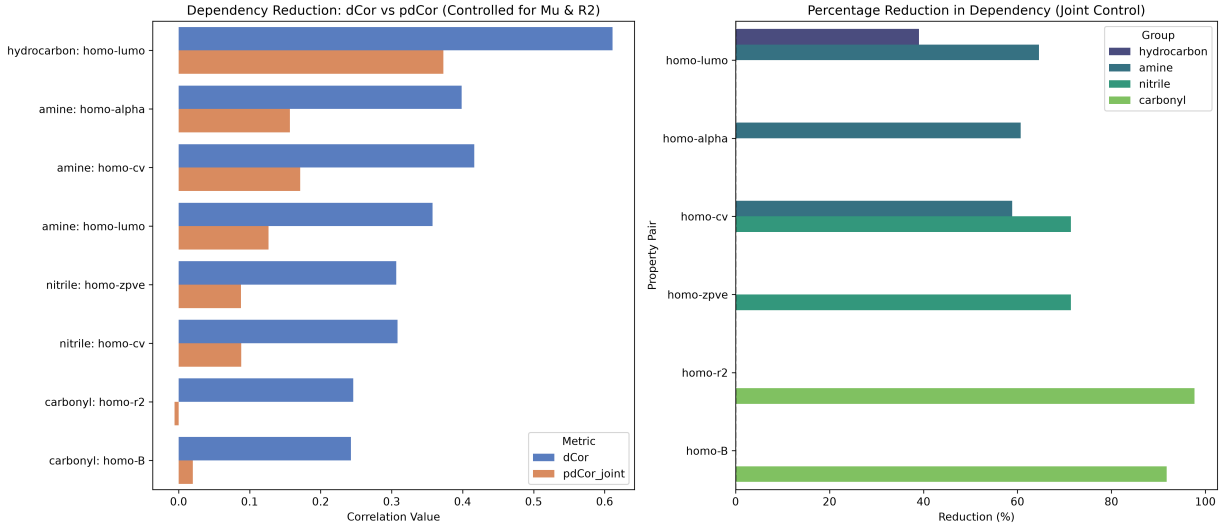

Figure 4: The figure illustrates the reduction in dependency between molecular properties when jointly controlling for spatial extent (r^2) and dipole moment (μ) across different structural families. The left panel compares the baseline distance correlation ($dCor$) with the partial distance correlation ($pdCor$) for selected property pairs, while the right panel displays the corresponding percentage reduction. These comparisons reveal the degree to which spatial and electrostatic features mediate electronic structure. Hydrocarbons exhibit strong intrinsic electronic coupling, evidenced by only a 39.02% reduction in the *homo-lumo* dependency. Conversely, carbonyls show almost complete mediation, with the *homo-*r^2 and *homo-B* relationships nearly vanishing under joint control (exceeding 90% reduction). Amines and nitriles represent intermediate cases of rewiring, demonstrating moderate mediation (approximately 60% to 65% reduction) for couplings such as *homo-lumo*, *homo-*α, and *homo-*C_v. Ultimately, the plot demonstrates that frontier-electronic couplings are not universally static but are fluidly integrated or isolated depending on the identity of heteroatom functional groups.

the *homo-B* coupling is similarly mediated (91.76% reduction). Amines and nitriles represent intermediate cases, where the *homo-lumo* coupling is reduced by approximately 64.58% ($pdCor = 0.127$), and couplings like *homo-*α and *homo-*C_v see reductions near 60%.

3.3 The oxygen anomaly and functional group divergence

Direct comparison of oxygen-containing groups further highlights that molecular properties do not adhere to a universal static partition. We specifically contrasted the rewiring modes of carbonyls (polar sp^2 oxygen) and ethers (lone-pair donating sp^3 oxygen) to resolve previously observed anomalies in oxygen-driven decoupling.

As detailed in Figure 5, these two functional groups drive fundamentally different organizational manifolds. Carbonyls force a stronger coupling between electronic residuals and the molecular dipole (μ), effectively using the dipole as a central hub to organize electronic properties. In contrast, ethers largely maintain a hydrocarbon-like organizational structure, though they exhibit localized increases in coupling, such as between *homo* and spatial extent (r^2, $dCor = 0.232$). Together, these results confirm that frontier-electronic and geometric-response relationships are fluid manifolds, dynamically integrated or isolated by the specific identity and bonding environment of the heteroatom scaffold.

4 Conclusions

In this study, we challenged the prevailing paradigm that treats computed quantum chemical properties as independent predictive targets. By isolating non-linear stoichiometric residuals across 133,885 molecules from the QM9 dataset, we demonstrated that molecular topology acts as a regulatory scaffold, dynamically organizing the joint relationships between electronic, thermodynamic, and geometric features. Our distance correlation profiling framework successfully mapped this organizational manifold, revealing that the apparent global orthogonal isolation of frontier orbitals is not a universal constant, but a conditional feature dictated by specific chemical perturbations.

The most striking evidence of this dynamic rewiring is driven by the amine group. Rather than merely shifting property baselines, the introduction of nitrogen-based lone pairs fundamentally fractures the geometric and response couplings—such as those between the rotational constant, polarizability, spatial extent, and heat capacity—that remain tightly bound in pure hydrocarbons. This identifies the amine scaffold as the primary driver of global organizational change within the dataset, actively decoupling spatial and thermodynamic responses.

Furthermore, our partial distance correlation analyses resolved the structural triggers of frontier-electronic decoupling. We found that pure hydrocarbons possess a robust, intrinsic electronic coupling between the HOMO and LUMO energies that persists independently of spatial or electrostatic mediation. In contrast, polar functional groups fundamentally alter this architecture. Carbonyls force the nearly complete mediation of electronic-spatial relationships through the molecular dipole, effectively utilizing it as a central organizational hub. Meanwhile, lone-pair donating ethers maintain a more hydrocarbon-like manifold, confirming that even within oxygen-containing molecules, the specific hybridization and bonding environment dictate the property network.

Ultimately, these findings refute the assumption of a static, universal property partition in small organic molecules. Frontier-electronic and geometric-response relationships form fluid manifolds that are dynamically integrated or isolated by specific heteroatom scaffolds. Recognizing this conditional joint structure is crucial for the future development of chemical machine learning models and data analysis frameworks, as it highlights the necessity of capturing the structural triggers that rewire molecular property spaces rather than treating them as isolated predictive tasks.

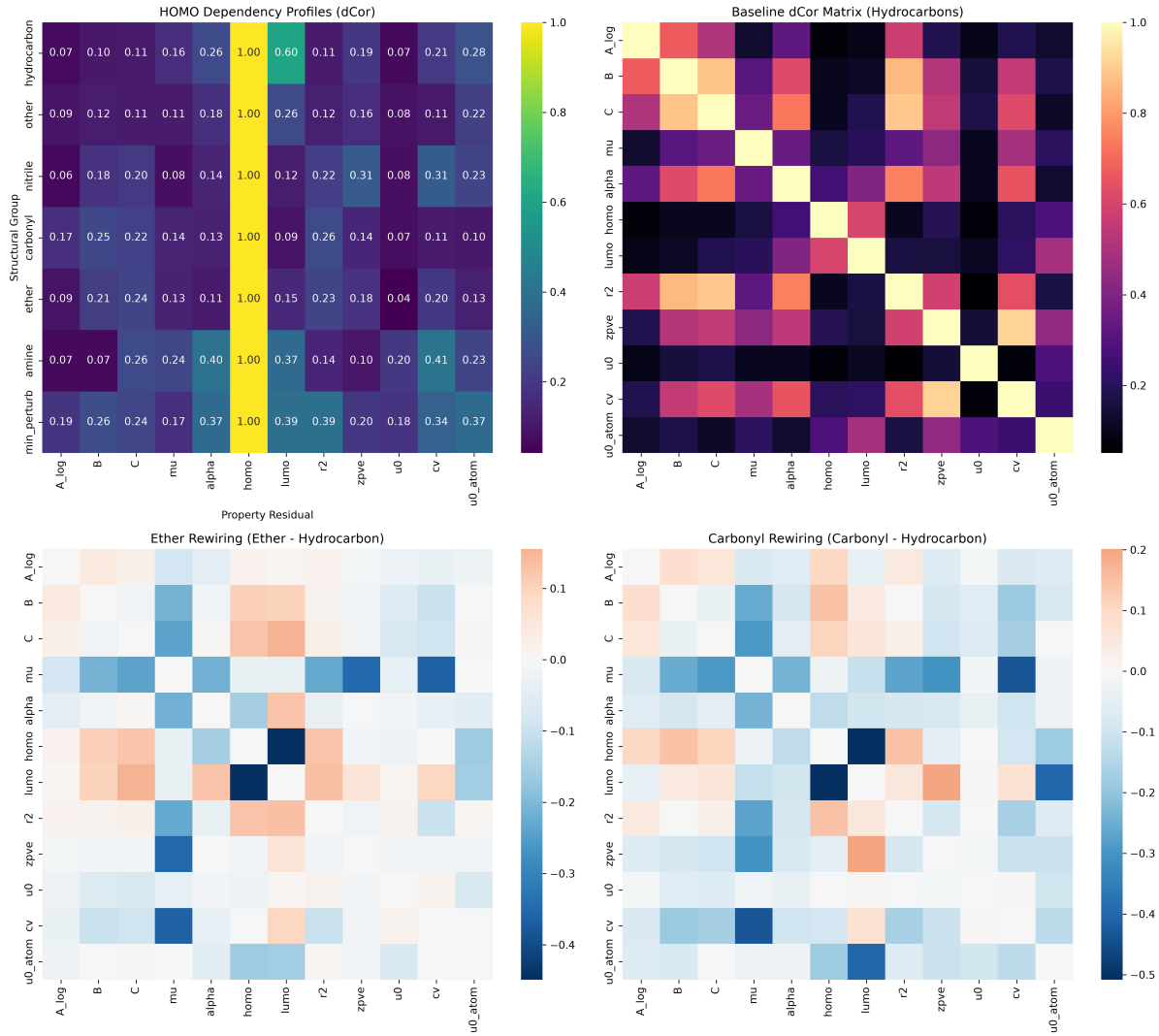

Figure 5: This figure illustrates the organizational rewiring of molecular property couplings across structural families in the QM9 dataset, based on distance correlation ($dCor$) matrices of non-redundant stoichiometric residuals. The top-right panel displays the baseline $dCor$ matrix for hydrocarbons, characterized by tightly bound geometric properties and strong intrinsic electronic coupling between *homo* and *lumo*. The top-left panel isolates the *homo* dependency profiles across various functional groups, demonstrating how heteroatoms decouple these baseline frontier orbital relationships. The bottom panels highlight the distinct rewiring modes of oxygen-containing groups by plotting the $dCor$ differences for ethers (bottom-left) and carbonyls (bottom-right) relative to the hydrocarbon baseline. These difference matrices reveal that while ethers largely maintain a hydrocarbon-like structure with increased coupling between *homo* and spatial extent (r^2), carbonyls force a stronger coupling between electronic residuals and the molecular dipole (μ). Together, these visualizations demonstrate that molecular properties do not adhere to a universal static partition, but instead form a dynamic manifold where specific functional groups actively integrate or isolate frontier-electronic and geometric-response features.