Conditional Dependency and Heteroatom-Driven Recoupling of the HOMO Manifold in QM9

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy of small organic molecules frequently exhibits an anomalous global decoupling from broader thermodynamic and spatial properties. To investigate the chemical drivers of this orthogonality, we isolate variance independent of molecular size and stoichiometry to identify how local environments influence electronic states. Using the QM9 dataset, we train gradient boosting regressors on atom counts to extract property residuals, effectively removing compositional effects. We then employ principal component analysis and random forest regression to model the global physical state, enabling the computation of partial distance correlations between HOMO residuals and local bonding descriptors. While global distance correlations confirm that HOMO energy remains largely orthogonal to other molecular properties, stratifying the dataset by heteroatom signatures reveals significant, structure-specific recoupling events. Our results show that pure hydrocarbons exhibit a strong recoupling between HOMO and LUMO energies, and nitrogen-only systems demonstrate a pronounced recoupling between HOMO and the dipole moment. In contrast, oxygen-containing systems largely maintain the global decoupling. These findings illustrate that while HOMO energy is globally anomalous, specific heteroatom identities act as structural switches that reintegrate frontier orbitals with the global physical manifold of the molecule.

1 Introduction

The exploration of chemical space through large-scale computational datasets, such as QM9, has revealed that many molecular properties are heavily governed by fundamental compositional factors. Molecular size and stoichiometry dictate the vast majority of variance in thermodynamic properties, including internal energy and zero-point vibrational energy. However, frontier orbital energies—most notably the energy of the highest occupied molecular orbital (HOMO)—exhibit a striking independence from these macroscopic molecular descriptors.

This independence manifests as a global decoupling, where the HOMO energy remains largely orthogonal to the broader thermodynamic and spatial manifold of the molecule. While properties like heat capacity and isotropic polarizability participate in a tightly coupled joint structure, the HOMO energy frequently yields distance correlations below 0.30 against other physical properties. A critical open question is whether this global "HOMO anomaly" represents an inherent, irreducible orthogonality of electronic states, or if specific local bonding environments and functional groups can force these frontier orbitals to reintegrate with the global physical state of the molecule.

To resolve the chemical drivers of this orthogonality, it is necessary to isolate variance that is strictly independent of molecular size and stoichiometry. Traditional linear partial correlations and high-dimensional density estimates often fail to capture the complex, non-linear dependencies inherent in molecular property manifolds. In this work, we introduce a conditional dependency framework based on partial distance correlations to rigorously evaluate the structural drivers of HOMO recoupling. By training gradient boosting regressors on exact atom counts, we extract property residuals that effectively remove compositional and size effects. We then model the global physical state of the molecule using principal component analysis and random forest regression, allowing us to explicitly hold the thermodynamic-spatial manifold constant.

With this stoichiometry-independent baseline, we condition the isolated HOMO residuals against local bonding descriptors, such as the fraction of sp^3 hybridized carbons and aromatic ring counts. Furthermore,

by stratifying the dataset across distinct heteroatom signatures and Murcko scaffolds, we mathematically pinpoint the exact structural conditions that trigger recoupling events.

Our analysis reveals that the global HOMO anomaly is not uniform across chemical space. Instead, specific heteroatom identities act as structural switches. Pure hydrocarbons exhibit a profound recoupling between HOMO and LUMO energies, while nitrogen-only systems demonstrate a strong recoupling between the HOMO energy and the molecular dipole moment. Conversely, oxygen-containing systems largely preserve the global decoupling. By bypassing the confounding effects of molecular composition, this study provides a rigorous measurement of how localized functional groups dictate orthogonal electronic properties, illustrating that frontier orbitals are selectively wired into the global physical manifold by the underlying molecular architecture.

2 Methods

2.1 Data curation and structural descriptor extraction

To investigate the structural drivers of the highest occupied molecular orbital (HOMO) anomaly, we utilized the QM9 dataset. To prevent stereoisomers that collapse to the same 2D graph from leaking across data splits and artificially inflating structural relationships, the dataset was deduplicated based on the SMILES string, retaining only the first occurrence. Redundant thermodynamic and gap properties were removed, leaving a core set of 12 non-redundant physical and electronic properties. For molecules where the rotational constant A was exactly zero, a binary indicator variable, is_linear, was created. A $\log(A+1)$ transformation was subsequently applied to tame the extreme heavy-tailed distribution of the rotational constant.

Molecular parsing and descriptor extraction were performed using RDKit. Explicit hydrogens were added to each molecule before extracting the exact atom counts for C, N, O, F, and H. To capture the local electronic environment, we computed specific local bonding descriptors: the fraction of sp^3 hybridized carbons (F_{sp^3}), the number of aromatic rings, and the number of double bonds. Molecules were then classified into distinct heteroatom signatures—pure hydrocarbons, nitrogen-only (N-only), oxygen-only (O-only), fluorine-only (F-only), or mixed heteroatoms—and their Murcko scaffolds were extracted to facilitate structural grouping. Geometrically inconsistent molecules that failed the parsing stage were explicitly flagged and excluded from downstream manifold learning.

2.2 Isolation of stoichiometry-independent property residuals

As established, molecular size and composition dominate the variance of macroscopic thermodynamic properties. To isolate the variance that is purely structural, we rigorously removed these stoichiometric effects. For each of the 12 core properties, a Gradient Boosting Regressor was trained using the five exact atom counts and the is_linear flag as predictors.

To prevent overfitting and ensure robust residual estimation, we employed a 5-fold cross-validation strategy to generate out-of-fold predictions for every molecule. The out-of-fold predictions were subtracted from the true property values to compute the property residuals. These residuals represent the variation in each property that is strictly orthogonal to molecular stoichiometry and linear-molecule size effects, providing a purely structural baseline for subsequent analysis.

2.3 Manifold representation and conditional distance correlation

To formally validate the global HOMO anomaly, we calculated the pairwise distance correlation ($dCor$) matrix for the 12 property residuals across the entire deduplicated dataset. Distance correlation was selected over Pearson correlation to capture non-linear dependencies without assuming a specific functional form.

To understand the drivers of the HOMO residual independently of the broader molecular state, we developed a conditional dependency framework. We first extracted the principal components (PCs) of the 11-property residual manifold (excluding the HOMO residual) to construct a linear representation of the molecule's global physical state. We then employed a Random Forest regressor to regress out these PCs and the is_linear flag from both the HOMO residual and the local bonding descriptors (such as F_{sp^3} and aromatic ring count). Finally, we calculated the distance correlation between the resulting secondary

residuals. This procedure mathematically isolates the specific local environments that dictate HOMO energy, independent of the molecule's global thermodynamic and spatial organization.

2.4 Heteroatom stratification and statistical validation of recoupling

To determine how specific molecular architectures rewire the relationship between frontier orbitals and the global physical state, the dataset was stratified into the previously defined heteroatom signature groups. Within each specific heteroatom class, the distance correlation matrices of the 12 property residuals were recomputed. "Recoupling events" were identified by calculating the change in correlation strength ($\Delta dCor$) relative to the global baseline.

The statistical significance of these recoupling events was tested using a computationally efficient permutation test. Within each heteroatom family (stratified by the *is_linear* flag), the HOMO residuals were shuffled 1,000 times to generate a null distribution, thereby preserving the stoichiometric and structural distribution of that specific family. A recoupling event was deemed significant if the observed distance correlation exceeded the 99th percentile of this null distribution ($p < 0.01$).

To ensure that the observed recoupling was a robust chemical phenomenon rather than an artifact of outliers, we performed a split-half stability test. For families exhibiting significant HOMO recoupling, the data was randomly split in half 10 times, and the mean and variance of the effect size were recorded. Additionally, a leave-one-scaffold-out validation was performed within the broader heteroatom classes to ensure that the recoupling generalized across different structural motifs. Finally, differential correlation heatmaps and UMAP projections of the 12-property residual space were generated to visually confirm the structural regions where the HOMO manifold reintegrates with the global physical state.

3 Results

3.1 The stoichiometric baseline and residual variance

Molecular size, parameterized by the counts of C, N, O, F, and H atoms, dominates the variance of many QM9 properties. To isolate the joint structure that is purely "chemical" rather than "compositional," we trained gradient boosting regressors to predict each property from atom counts alone and evaluated their out-of-fold predictions. The baseline metrics confirm that thermodynamic properties are almost entirely determined by stoichiometry: internal energy (U_0) and zero-point vibrational energy (ZPVE) achieve R^2 values of 0.9987 and 0.9973, respectively. Isotropic polarizability (α) and heat capacity (C_v) also exhibit high stoichiometric dependence, with R^2 values of 0.8774 and 0.7298.

Conversely, the frontier orbitals and geometric response variables are significantly more independent of molecular size. The HOMO energy ($R^2 = 0.1928$), dipole moment (μ, $R^2 = 0.2697$), and electronic spatial extent (r^2, $R^2 = 0.3606$) retain over 60% of their variance after conditioning on stoichiometry. By utilizing the out-of-fold residuals (denoted with the prefix `res_`) from these regressions, we successfully isolated the structural organization of the molecules, providing a rigorous foundation for subsequent manifold learning.

3.2 Validation of the global HOMO anomaly

To formally investigate the decoupling of frontier orbitals, we computed the distance correlation ($dCor$) matrix of the 12 non-redundant property residuals across the entire dataset. The results confirm that the HOMO residual (`res_homo`) is remarkably orthogonal to the rest of the dataset. While properties such as C_v and ZPVE remain tightly coupled even in the residual space, `res_homo` exhibits distance correlations below 0.30 with all other variables globally.

This global anomaly, visualized in Figure 1, was further interrogated by decomposing the 11-property residual manifold (excluding HOMO) using Principal Component Analysis (PCA). We found that seven principal components were required to explain 93.39% of the manifold's variance. A Random Forest regression model demonstrated that this 11-dimensional global physical state explains only 41.38% of the variance in `res_homo`. For comparison, the same global state explains 50.60% of the variance in the fraction of sp^3 carbons (F_{sp^3}) and 41.11% of the aromatic ring count. This validates that E_{HOMO} is not merely noisy, but is fundamentally driven by factors other than the global thermodynamic-spatial organization.



Figure 1: Global distance correlation ($dCor$) heatmap of the 12 non-redundant property residuals from the QM9 dataset, illustrating the global HOMO anomaly. The matrix reveals that while thermodynamic and spatial properties remain tightly coupled in the residual space—such as the strong correlations between the zero-point vibrational energy (`res_zpve`) and heat capacity (`res_cv`), or the rotational constant (`res_C`) and electronic spatial extent (`res_r2`)—the highest occupied molecular orbital energy residual (`res_homo`) is remarkably orthogonal to the rest of the dataset. Specifically, `res_homo` exhibits distance correlations below 0.30 with all other variables globally. This demonstrates that E_{HOMO} is fundamentally decoupled from the global thermodynamic and spatial manifold of the molecules once the stoichiometric baseline variance is removed.

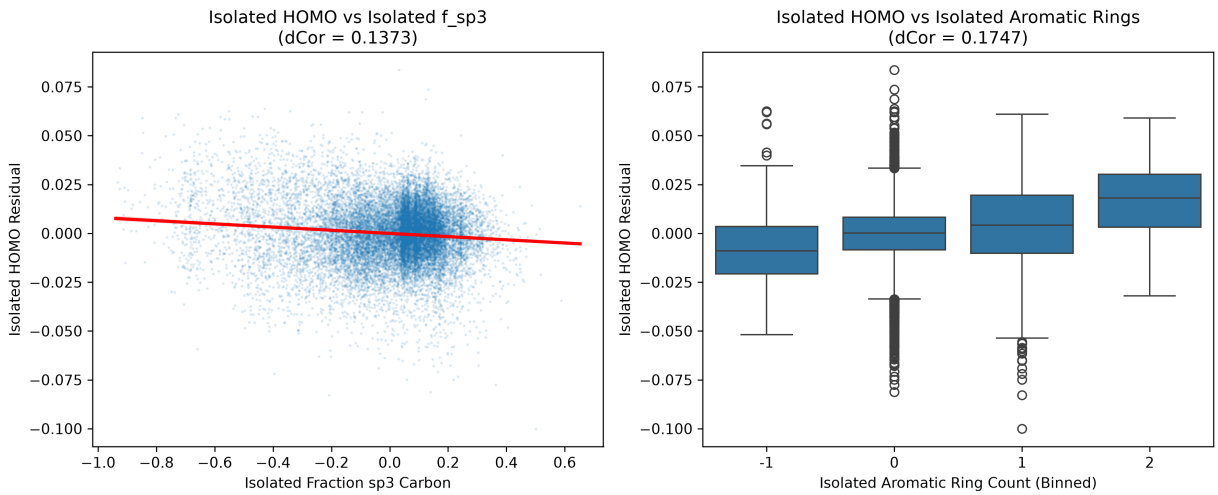

Figure 2: Conditional dependency of the isolated HOMO residual on local bonding descriptors after regressing out the 11-dimensional global physical state. The left panel displays a scatter plot of the HOMO residual against the isolated fraction of sp^3 carbons (F_{sp^3}), yielding a low distance correlation ($dCor = 0.1373$). The right panel shows a box plot of the HOMO residual across binned isolated aromatic ring counts, which also exhibits a weak correlation ($dCor = 0.1747$). These low conditional correlations demonstrate that the global HOMO anomaly is robust, indicating that the highest occupied molecular orbital energy (E_{HOMO}) is not simply a restatement of molecular saturation or aromaticity once the overall molecular size and shape are accounted for.

3.3 Local bonding and conditional dependency

To determine if local bonding environments dictate the HOMO residual independently of the global manifold, we calculated partial distance correlations. By regressing the 11-dimensional global state (PCs) out of both `res_homo` and the local descriptors (F_{sp^3} and aromatic ring count), we isolated their direct dependencies.

As shown in Figure 2, the resulting conditional correlations are remarkably low: `res_homo` versus F_{sp^3} yields a $dCor$ of 0.1373, and `res_homo` versus aromatic ring count yields a $dCor$ of 0.1747. These values indicate that the global HOMO anomaly is robust. The energy of the highest occupied molecular orbital is not simply a restatement of saturation or aromaticity once the overall molecular size and shape are accounted for. Consequently, the relationship between frontier orbitals and the rest of the molecule must be investigated at the level of specific chemical families.

3.4 Heteroatom-driven recoupling events

Stratifying the 133,802 molecules into heteroatom signature groups—Pure Hydrocarbons, N-only, O-only, and Mixed—reveals that the HOMO anomaly is not uniform across chemical space. We identified significant "recoupling events" where the distance correlation between `res_homo` and other properties substantially exceeds the global baseline.

Using a permutation test with 1,000 shuffles to establish a null distribution, we found several highly significant ($p < 0.01$) recoupling events, summarized in Figure 3. The strongest recoupling observed in the entire study occurs between `res_homo` and `res_lumo` in pure hydrocarbons, reaching a $dCor$ of 0.6025 compared to the global baseline of 0.1822. This massive shift ($\Delta dCor = 0.4203$) indicates that in the absence of heteroatoms, the frontier orbitals behave as a single coherent unit tightly linked to the electronic gap.

Furthermore, nitrogen-containing molecules (N-only) exhibit a strong recoupling between the dipole moment (`res_mu`) and `res_homo` ($dCor = 0.5219$ vs. global 0.1970). This $\Delta dCor$ of 0.3249 suggests that nitrogen functional groups—such as amines, nitriles, or pyridinic nitrogens—create a specific joint structure where the electronic orbital energy is physically bound to the molecule's electrostatic response. In contrast, oxygen-containing molecules (O-only) show the weakest recoupling, primarily with rotational constants and

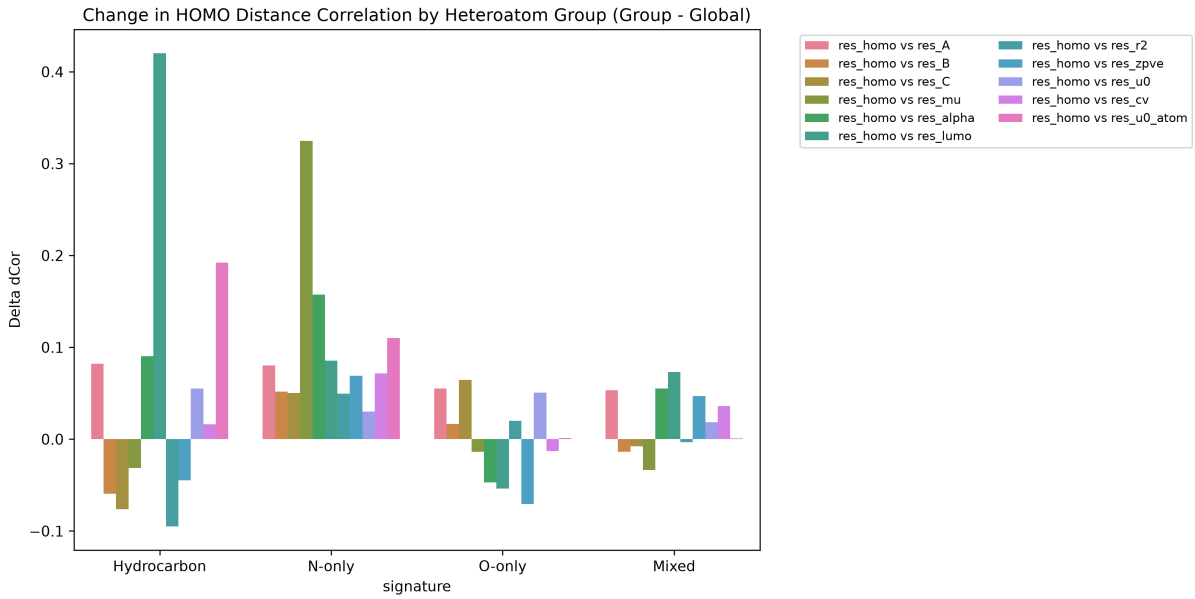

Figure 3: Change in distance correlation ($\Delta dCor$) between the HOMO residual (`res_homo`) and other property residuals across heteroatom signature groups relative to the global baseline. The plot illustrates significant heteroatom-driven recoupling events that overcome the global decoupling of the highest occupied molecular orbital energy (E_{HOMO}). Pure hydrocarbons display a pronounced recoupling between `res_homo` and the LUMO residual (`res_lumo`), indicating that frontier orbitals act as a tightly linked coherent unit in the absence of heteroatoms. Similarly, nitrogen-only systems exhibit a strong recoupling between `res_homo` and the dipole moment (`res_mu`), suggesting that nitrogen functional groups physically bind electronic orbital energy to the electrostatic response. In contrast, oxygen-only and mixed systems show minimal changes, effectively maintaining the global anomaly. This demonstrates that specific chemical families, particularly pure carbon skeletons and nitrogen-containing molecules, act as structural switches that rewire frontier orbitals into the global physical manifold.

spatial extent (e.g., `res_C`, $\Delta dCor = 0.0646$), suggesting that oxygen maintains the HOMO decoupling more effectively than nitrogen.

Split-half stability tests over 10 iterations confirmed the robustness of these findings. The hydrocarbon `res_homo`/`res_lumo` coupling showed a mean $dCor$ of 0.5945 ($\sigma = 0.0175$), and the N-only `res_homo`/`res_mu` coupling showed a mean $dCor$ of 0.4551 ($\sigma = 0.0205$). Because the critical values for the 99th percentile of the null distributions were consistently near 0.13, our observed recouplings fall well beyond the threshold of statistical significance.

3.5 Visualizing the rewired manifold

To contextualize these numerical findings, we generated differential correlation heatmaps that visually confirm the structural rewiring of the property manifold.

As depicted in Figure 4, the HOMO row and column brighten considerably in the Nitrogen-only and Hydrocarbon families relative to the global baseline, signaling the integration of E_{HOMO} into the broader property network.

UMAP projections of the residual manifold further clarify this phenomenon (Figure 5). In the Nitrogen-only group, the HOMO residuals display structured, continuous gradients across the UMAP embedding space. Conversely, in the Mixed and Oxygen-only groups, the HOMO residual appears randomly distributed across the physical manifold. This visual evidence corroborates our statistical tests, confirming that nitrogen and pure carbon skeletons provide the necessary structural scaffold for the frontier orbitals to organize jointly with the molecule's global physical state.

4 Conclusions

The investigation into the QM9 dataset reveals that the property manifold of small organic molecules is not a monolithic structure dictated solely by size and stoichiometry. By rigorously isolating variance orthogonal to compositional effects through gradient-boosted property residuals, we have formalized the existence of a global "HOMO anomaly." The energy of the highest occupied molecular orbital (E_{HOMO}) remains fundamentally decoupled from the broader thermodynamic and spatial state of the molecule. Furthermore, our conditional dependency framework demonstrates that this anomaly is robust against local bonding descriptors; neither the fraction of sp^3 hybridized carbons (F_{sp^3}) nor aromaticity can independently resolve the orthogonal nature of the frontier orbitals once the global physical state is held constant.

However, this global decoupling is not an irreducible property of chemical space. By stratifying the residual manifold across distinct heteroatom signatures, we identified significant, structure-specific recoupling events. In the absence of heteroatoms, pure hydrocarbon skeletons force a profound recoupling between the HOMO and LUMO energies, causing the frontier orbitals to behave as a tightly linked coherent unit. Similarly, the introduction of nitrogen acts as a specific structural switch, strongly binding the HOMO energy to the molecule's electrostatic response, as evidenced by its recoupling with the dipole moment. In contrast, oxygen-containing systems largely preserve the global decoupling, indicating that oxygen maintains the orthogonality of the HOMO manifold more effectively than nitrogen.

These findings demonstrate that the relationship between frontier electronic states and macroscopic molecular properties is highly conditional. Heteroatom identity dictates the underlying molecular architecture, selectively wiring or rewiring the frontier orbitals into the global physical manifold. By bypassing the confounding effects of molecular composition, this work provides a rigorous, stoichiometry-independent perspective on molecular property organization. Recognizing these heteroatom-driven recoupling events offers critical insights for future computational chemistry endeavors, suggesting that predictive models and targeted molecular design strategies must account for the conditional, switch-like nature of chemical functional groups rather than assuming a uniform property manifold.

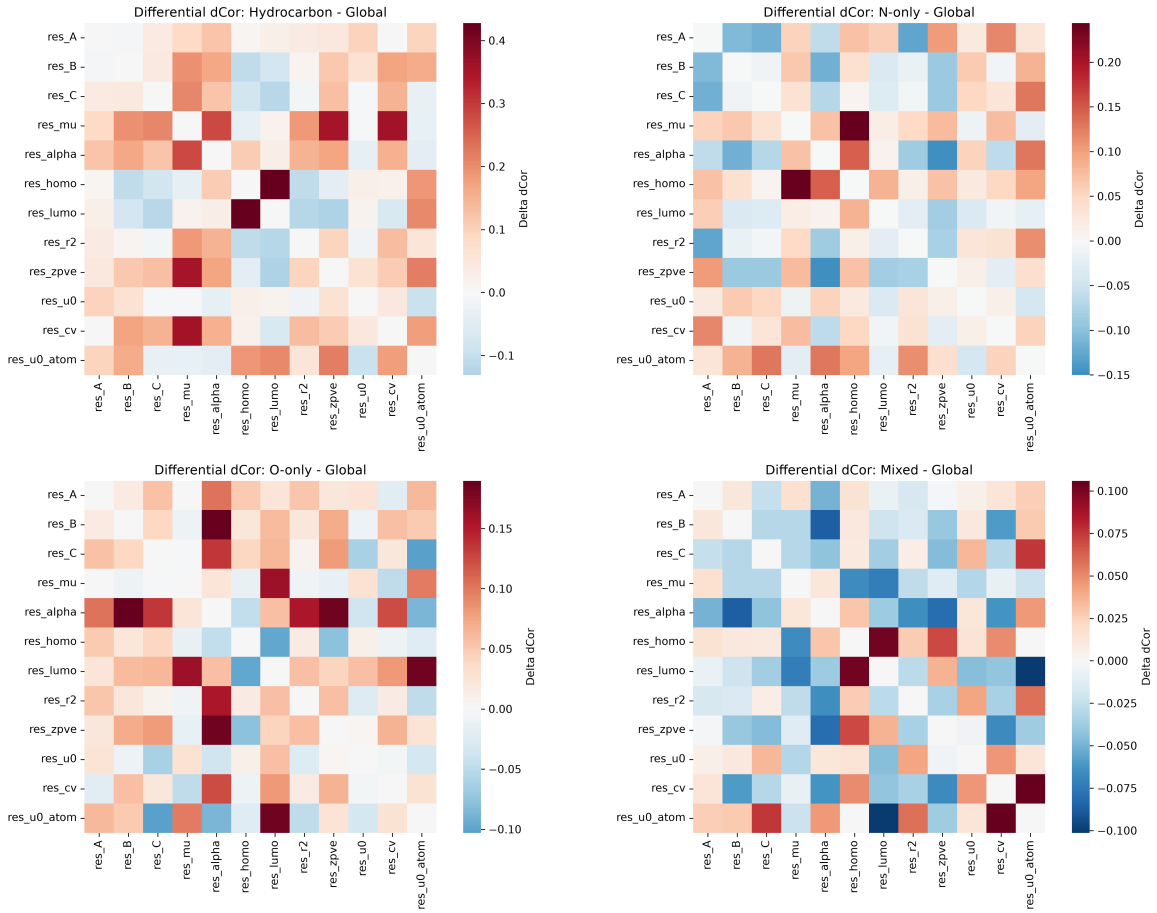

Figure 4: Differential distance correlation ($\Delta dCor$) heatmaps illustrating the rewiring of the molecular property manifold across four heteroatom signature groups (Hydrocarbon, N-only, O-only, and Mixed) relative to the global baseline. The matrices display correlations between out-of-fold property residuals (e.g., `res_homo`, `res_lumo`, `res_mu`) after conditioning on stoichiometry. Bright regions along the `res_homo` row and column highlight significant recoupling events where the globally anomalous HOMO energy (E_{HOMO}) integrates into the broader physical state. Specifically, pure hydrocarbons exhibit a strong recoupling between the frontier orbitals (`res_homo` and `res_lumo`), while nitrogen-only systems show a strong coupling between `res_homo` and the dipole moment (`res_mu`). Conversely, oxygen-only and mixed systems display weaker recoupling, indicating they maintain the E_{HOMO} decoupling more effectively. These visualizations reveal that heteroatom identity acts as a structural switch, selectively rewiring frontier electronic orbitals into the global physical manifold of the molecule.

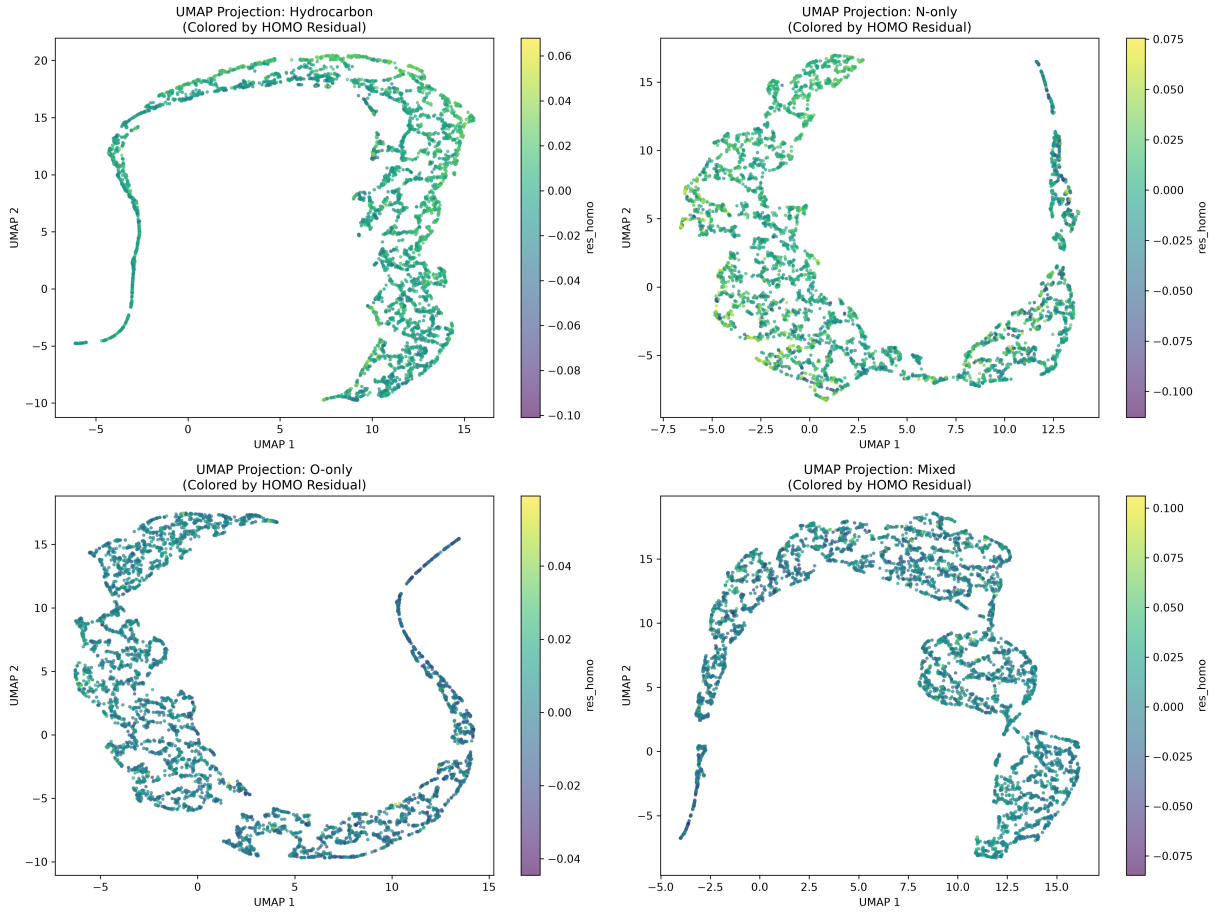

Figure 5: UMAP projections of the residual manifold stratified by heteroatom signature groups (Hydrocarbon, N-only, O-only, and Mixed), colored by the HOMO energy residual (`res_homo`). These embeddings visualize the heteroatom-driven recoupling of the frontier orbitals with the global physical state. In the N-only and Hydrocarbon families, `res_homo` exhibits structured gradients across the UMAP space, demonstrating that nitrogen and pure carbon skeletons provide the structural scaffold necessary for the HOMO energy to organize jointly with the broader property network. Conversely, in the O-only and Mixed groups, `res_homo` appears more randomly distributed, reflecting the persistent global decoupling of E_{HOMO} from the thermodynamic and spatial manifold.