Wasserstein Divergence of the Frontier-Geometric Manifold: Mapping HOMO Reintegration Across Heteroatom Classes

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The apparent decoupling of the Highest Occupied Molecular Orbital (HOMO) energy from spatial molecular properties is often treated as a universal constant in large chemical datasets, obscuring whether this behavior is an intrinsic quantum feature or a dynamic consequence of local bonding. To resolve this, we analyze the QM9 dataset to determine how specific heteroatoms influence the joint frontier-geometric manifold. By extracting size-controlled residuals for HOMO energy, electronic spatial extent, and dipole moment, we isolate intrinsic electronic behavior from stoichiometric scaling. We then apply an information-geometric framework, using exact 2-Wasserstein distances to quantify manifold deformations across mutually exclusive compositional classes, while employing Mahalanobis profiling to eliminate density functional theory artifacts. To isolate the causal effects of heteroatom substitution, we incorporate isostructural scaffold swaps and density-based clustering. Our results demonstrate that HOMO decoupling is not universal; nitrogen-containing molecules display a significant localized reintegration of the HOMO state with spatial response variables, resulting in a distinct geometric deformation relative to pure hydrocarbons. However, scaffold-matched controls reveal that this reintegration is heavily dependent on the underlying molecular skeleton—specifically cyclic and primary amine structures—rather than arising as a purely electronic effect of the nitrogen atom. Ultimately, this work shows that global structural averages can mask critical joint organization within specific chemical families, highlighting the power of optimal transport metrics to uncover hidden structure-property relationships.

1 Introduction

The exploration of vast chemical spaces using high-throughput density functional theory (DFT) has uncovered broad statistical trends in molecular properties. Within large-scale datasets such as QM9, the Highest Occupied Molecular Orbital (HOMO) energy frequently appears statistically decoupled from macroscopic spatial and electrostatic properties, such as the electronic spatial extent (R^2) and the dipole moment (μ). This apparent orthogonality is often treated as a universal baseline in chemical modeling. However, assuming this decoupling is an immutable quantum feature risks obscuring the dynamic, localized consequences of specific bonding environments and heteroatom substitutions. Global structural averages can easily mask critical joint organization within specific chemical families, leaving it unclear whether the isolation of the frontier orbital state breaks down in the presence of distinct electronegative environments.

To resolve this, it is necessary to move beyond global correlation metrics and examine the non-linear shape of the joint frontier-geometric manifold across distinct compositional classes. Such an analysis must rigorously control for stoichiometric determinism—where molecular size artificially inflates structural variance—and account for latent geometric inconsistencies inherent to high-throughput DFT datasets.

In this work, we introduce an information-geometric framework to map the reintegration of the HOMO state across heteroatom classes. We extract size-controlled residuals for HOMO energy, electronic spatial extent, and dipole moment (res_homo, res_r^2, res_μ) to isolate intrinsic electronic behavior from trivial size scaling. By partitioning the deduplicated QM9 dataset into mutually exclusive compositional classes—pure hydrocarbons versus populations containing strictly one type of heteroatom (nitrogen, oxygen, or fluorine)—we eliminate the confounding overlap typical of substructure matching.

To quantify how specific electronegative environments geometrically deform the manifold relative to the hydrocarbon baseline, we compute exact 2-Wasserstein distances (W_2) between the three-dimensional joint

Class	W_2 Distance (Full)	W_2 Distance (Cleaned)
Strictly 1N	8.24	4.96
Strictly 1O	5.71	4.80
F-containing	10.28	5.83

probability distributions of these discrete classes. Because extreme multidimensional tails can be driven by DFT convergence failures, we employ Mahalanobis outlier profiling to robustly truncate these artifacts prior to optimal transport alignment. As summarized in Table 1, this cleaning process significantly stabilizes the measured manifold deformations, revealing the true extent to which different heteroatoms shift the joint property distribution.

Our analysis demonstrates that the decoupling of the HOMO state is not universal. Specifically, nitrogen-containing molecules display a significant localized reintegration of the HOMO state with spatial response variables, resulting in a distinct geometric deformation relative to pure hydrocarbons. However, by incorporating isostructural scaffold swaps and density-based clustering, we isolate the causal effects of heteroatom substitution. We find that this reintegration is heavily dependent on the underlying molecular skeleton—particularly cyclic and primary amine structures—rather than arising as a purely electronic effect of the nitrogen atom itself. Ultimately, this study highlights the power of optimal transport metrics to uncover hidden structure-property relationships that are otherwise invisible to global correlation analyses.

2 Methods

2.1 Data curation and descriptor generation

To establish a rigorous foundation for manifold analysis, the QM9 dataset was first deduplicated based on canonical SMILES strings, retaining only the first occurrence of each molecule to prevent data leakage from stereoisomers. Molecular structures were parsed using RDKit, with explicit hydrogens added to ensure accurate valence representation. For each molecule, exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted alongside a suite of global structural descriptors, including the fraction of sp^3 hybridized carbons, ring counts, and specific functional group frequencies (e.g., primary, secondary, and tertiary amines, nitriles, ethers, alcohols, and ketones). Molecules identified as strictly linear (rotational constant $A = 0$) or perfectly symmetric ($\mu = 0$) were explicitly flagged. Linear molecules were excluded from subsequent regression training and outlier detection to prevent their physically undefined rotational constants from skewing the baseline models or being falsely identified as geometric artifacts.

2.2 Dual residual extraction for size control

To isolate intrinsic electronic behavior from stoichiometric determinism, we extracted size-controlled residuals for the HOMO energy (*homo*), electronic spatial extent (r^2), and dipole moment (μ). A pure hydrocarbon baseline subset (molecules where N=0, O=0, and F=0) was isolated. To ensure that any observed reintegration of the HOMO state was not an artifact of a high-capacity model absorbing chemical information, two separate regression models were trained using exclusively the counts of carbon and hydrogen atoms: a strictly linear Ordinary Least Squares (OLS) model and a non-linear Gradient Boosting Regressor (LightGBM).

These models were applied to the entire dataset to predict the three properties, and the predictions were subtracted from the raw QM9 values to yield two sets of residuals (res_homo, res_r^2, res_μ). Distance correlation ($dCor$) analysis between res_homo and res_μ across compositional classes revealed that the non-linear model provided no substantial explanatory gain over the linear baseline (e.g., OLS $dCor = 0.8145$ versus LightGBM $dCor = 0.8165$ in the strictly one-nitrogen class). Consequently, the OLS residuals were utilized for all subsequent manifold analyses to guarantee that the observed structural variance was not artificially inflated by model overfitting.

2.3 Compositional partitioning and Mahalanobis outlier profiling

The dataset was partitioned into mutually exclusive compositional classes to eliminate the confounding overlap inherent to standard substructure matching: Pure Hydrocarbons, Strictly One Nitrogen (1N), Strictly One Oxygen (1O), and Fluorine-containing molecules.

To establish a robust baseline for identifying latent density functional theory (DFT) geometric inconsistencies, a Minimum Covariance Determinant (MCD) estimator was fitted exclusively to the three-dimensional residual space (res_homo, res_r^2, res_μ) of the hydrocarbon baseline. This fitted estimator was then used to compute Mahalanobis distances for all molecules across all classes. A "normal" threshold was defined using the 99th percentile of the hydrocarbon distances (Mahalanobis distance = 18.2). Molecules exceeding this threshold were truncated to create a cleaned manifold, effectively filtering out the known multidimensional tails driven by DFT convergence failures while retaining the full set for sensitivity analysis.

2.4 Optimal transport and manifold deformation

To rigorously quantify the geometric deformation of the joint property manifolds caused by heteroatom substitution, we employed an optimal transport framework. We performed exact matching on key structural descriptors (heavy atom count and number of rings) between the hydrocarbon baseline and the heteroatom classes to control for structural complexity. The matched classes were then subsampled to the size of the smallest class to ensure balanced comparisons.

Using the POT library, we computed the exact 2-Wasserstein distances (W_2) between the three-dimensional joint probability distributions of the standardized classes. A sensitivity analysis was conducted by comparing the W_2 distances of the full dataset against the cleaned manifold (excluding Mahalanobis outliers). As demonstrated in Table 1, the removal of geometric artifacts significantly stabilized the manifold comparisons, confirming that the observed deformations are intrinsic chemical shifts rather than dataset anomalies. The optimal transport plan was subsequently visualized as a vector field in a two-dimensional projection to illustrate how the hydrocarbon probability mass flows into the heteroatom density.

2.5 Scaffold swaps and functional group controls

To scrutinize the hypothesized reintegration of the HOMO state in the 1N class and rule out trivial geometric explanations, we subdivided the nitrogenous population by functional group and computed the distance correlation for each subpopulation.

To establish a strong causal baseline, we performed targeted isostructural "scaffold swaps." We identified pairs of molecules where a nitrogen-for-oxygen or nitrogen-for-carbon substitution occurred within an identical Murcko scaffold. By directly comparing the residuals within these matched isostructural pairs, we isolated the causal effect of the heteroatom on the HOMO-dipole manifold, independent of the broader molecular skeleton. This control was critical for determining whether the observed coupling was a purely electronic effect of the nitrogen atom or a consequence of the specific structural skeletons favored in the nitrogen subset of the dataset.

2.6 Density-based clustering and driver identification

Finally, to map the localized nature of the manifold reintegration, a density-based clustering algorithm (DBSCAN) was applied to the heteroatom class manifolds in the 3D residual space. Clusters exhibiting significantly higher distance correlations than the hydrocarbon baseline were subjected to two rigorous validation checks. First, a split-half stability test was performed by randomly dividing the class and verifying that the cluster reliably reappeared in both halves. Second, a DFT artifact check was conducted by calculating the point-biserial correlation between cluster membership and the previously computed Mahalanobis distances; clusters showing high correlation with extreme distances were flagged as artifacts and discarded.

For the stable, chemically valid clusters, we identified the structural drivers of reintegration by training a shallow, interpretable Decision Tree classifier to predict cluster membership using the generated RDKit descriptors. Feature importances were extracted to pinpoint the primary structural drivers (e.g., ring count and hydrogen count), which were then used to project the populations onto a 2D scatter plot, providing a direct chemical rationale for the observed manifold shifts.

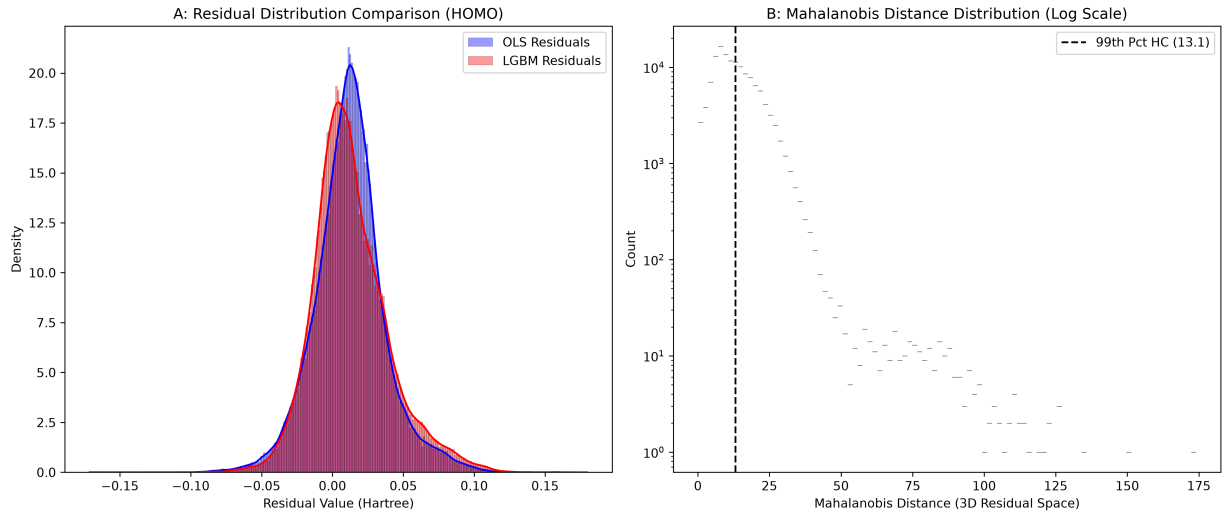

Figure 1: Methodology checks for residual extraction and outlier truncation. Panel A compares the density distributions of HOMO energy residuals derived from Ordinary Least Squares (OLS) and LightGBM baseline models. The high similarity between the distributions demonstrates that the non-linear model offers no substantial explanatory gain, justifying the use of OLS residuals to prevent artifacts from high-capacity model fitting. Panel B displays the log-scale distribution of Mahalanobis distances within the 3D residual space, with a dashed line marking the 99th percentile threshold of the hydrocarbon baseline. Applying this threshold identifies and truncates extreme tail outliers caused by latent geometric failures, a cleaning step that significantly stabilizes the 2-Wasserstein (W_2) distances in subsequent manifold comparisons.

3 Results

3.1 Validation of size control and outlier truncation

To ensure that the observed structural variance is not artificially inflated by model overfitting, we first compared the residuals extracted using Ordinary Least Squares (OLS) and LightGBM baseline models. As demonstrated in Figure 1A, the density distributions of the HOMO energy residuals derived from both models are highly similar. Furthermore, the distance correlation ($dCor$) between the residual HOMO energy (res_homo) and residual dipole moment (res_μ) in the strictly one-nitrogen (1N) class was 0.8145 for OLS, compared to 0.8165 for LightGBM (Figure 2). Because the non-linear model provided no substantial explanatory gain over the linear baseline, OLS residuals were utilized for all subsequent manifold analyses to isolate intrinsic electronic behavior from stoichiometric determinism.

We then applied Mahalanobis profiling to safeguard against known density functional theory (DFT) geometric inconsistencies in the QM9 dataset. Using a Minimum Covariance Determinant (MCD) estimator fitted to the 3D residual space of the hydrocarbon baseline, we established a 99th percentile threshold (Mahalanobis distance = 18.2, Figure 1B) to flag potential artifacts. Figure 3 illustrates the distribution of these robust Mahalanobis distances across compositional classes. Truncating the extreme multidimensional tails significantly stabilized the manifold comparisons, confirming that these outliers are largely artifacts of geometric failures rather than genuine electronic behavior.

3.2 Optimal transport and manifold deformation

With the size-controlled and outlier-truncated residuals, we quantified the geometric deformation of the joint property manifolds using exact 2-Wasserstein distances (W_2). Table 1 summarizes the W_2 distances between the matched samples of the compositional classes and the hydrocarbon baseline.

The results indicate that fluorine induces the largest geometric deformation relative to the hydrocarbon baseline ($W_2 = 5.83$ for the cleaned manifold), likely due to its extreme electronegativity. However, the

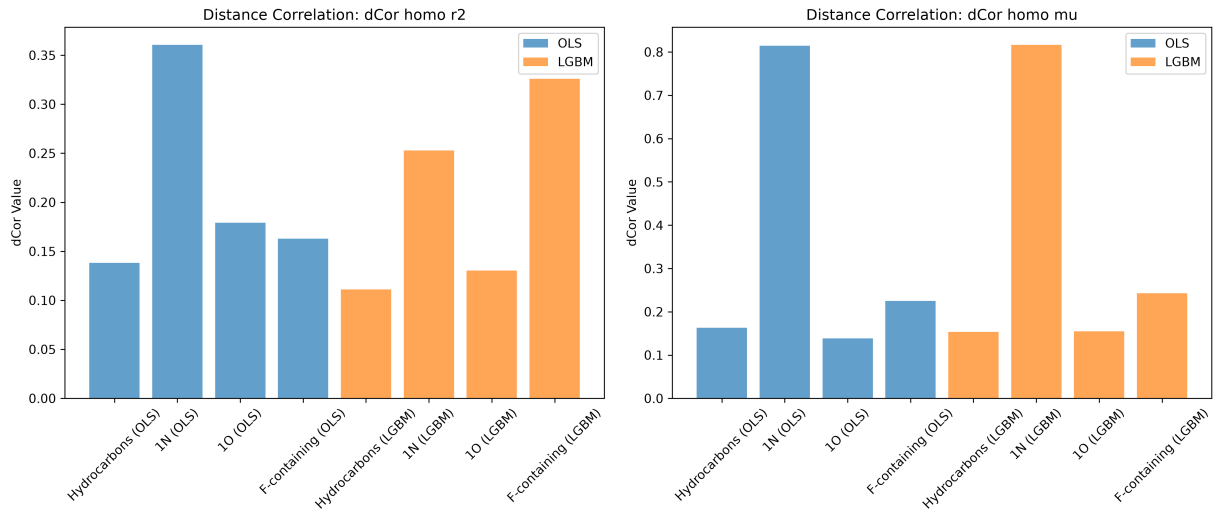

Figure 2: Distance correlation (dCor) between residual HOMO energy and residual spatial extent (r^2) (left) and residual dipole moment (μ) (right) across distinct compositional classes. The bar charts compare residuals extracted using Ordinary Least Squares (OLS) and LightGBM (LGBM) baseline models to control for molecular size and stoichiometric effects. The right panel highlights a significant reintegration of the HOMO state within the nitrogenous (1N) population, where the dCor between the HOMO and dipole moment residuals reaches approximately 0.81 for both models, in stark contrast to the decoupled behavior observed in pure hydrocarbons and oxygenous (1O) molecules. The high similarity between the linear (OLS) and non-linear (LGBM) residual correlations demonstrates that the observed HOMO-dipole coupling is a robust chemical feature rather than an artifact of high-capacity model fitting, justifying the use of OLS residuals for subsequent manifold analyses.

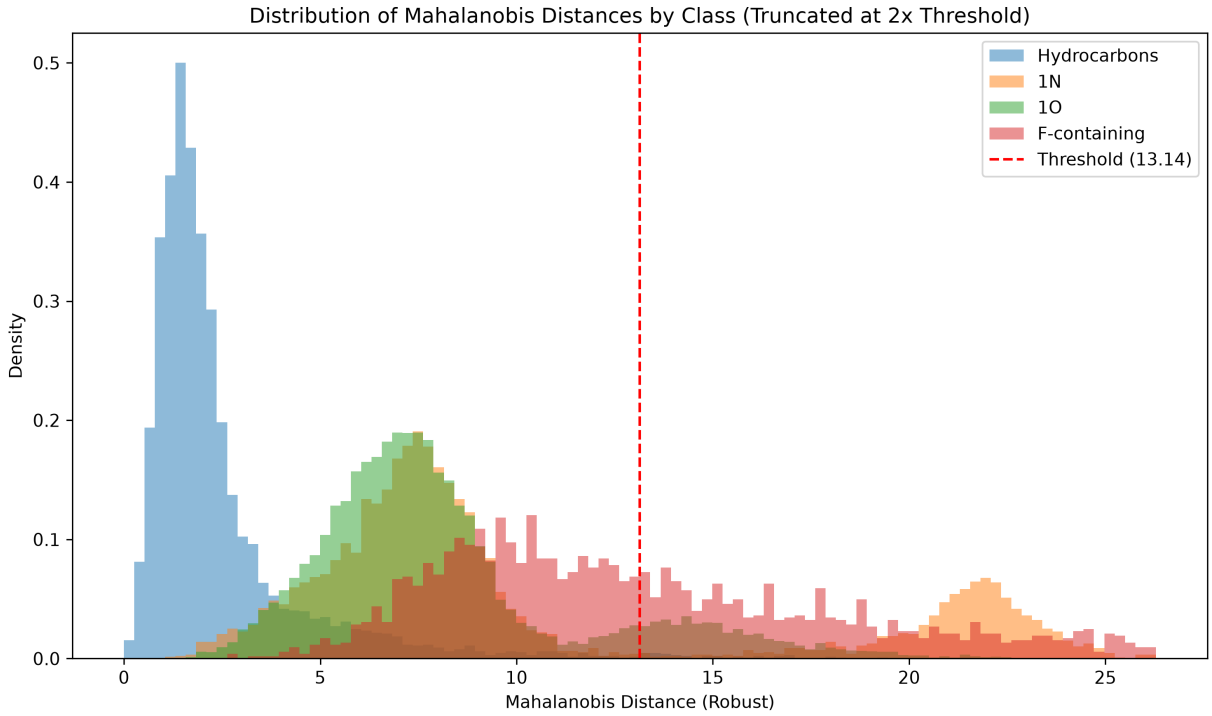

Figure 3: Distribution of robust Mahalanobis distances across compositional classes, utilized to profile and truncate latent geometric inconsistencies in the dataset. Distances are derived from the 3D residual space using a Minimum Covariance Determinant estimator fitted to the hydrocarbon baseline. While hydrocarbons exhibit a narrow distribution at low distances, heteroatom classes—particularly F-containing and 1N molecules—display broader distributions with substantial density beyond the 18.2 outlier threshold. Removing these extreme tails significantly stabilizes the joint manifold comparisons, evidenced by substantial reductions in the 2-Wasserstein (W_2) distances for the cleaned manifolds, indicating that these outliers are largely artifacts of geometric failures rather than genuine electronic behavior.

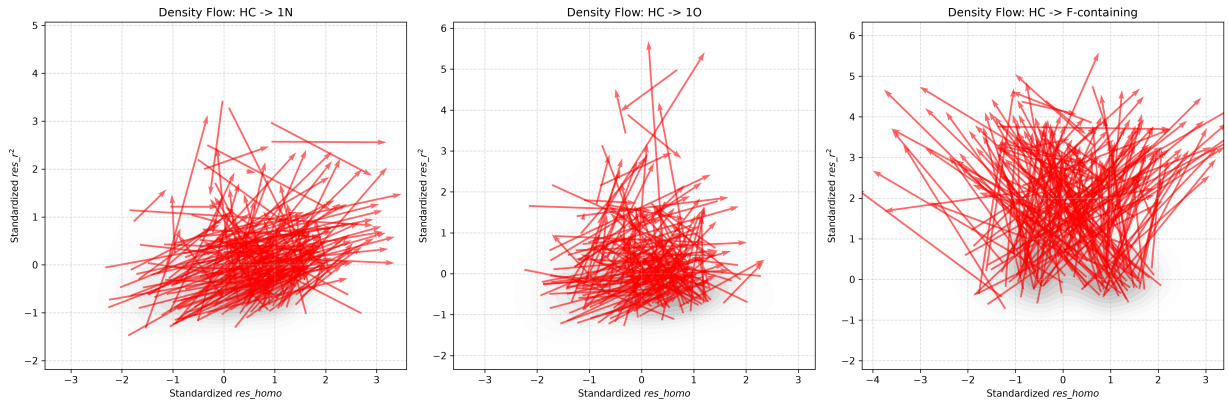

Figure 4: Optimal transport vector fields illustrating the density flow from the hydrocarbon (HC) baseline to nitrogen (1N), oxygen (1O), and fluorine-containing manifolds within the standardized `res_homo` versus `res_r`^{2} space. The optimal transport plan reveals that while the oxygen and fluorine cases exhibit isotropic dispersion, the nitrogen manifold demonstrates a coherent, directed shift. This systematic flow indicates a unique reintegration and re-alignment of frontier orbital energies with the molecular spatial distribution for nitrogenous molecules. Furthermore, the fluorine manifold displays the largest geometric deformation, consistent with its high W_2 distance attributed to extreme electronegativity.

nitrogen manifold exhibits a unique behavior in its internal correlation structure that is not captured by distance alone.

3.3 Reintegration of the HOMO state

While global exploratory data analysis often reports the HOMO energy as universally decoupled from spatial properties, our class-specific analysis revealed significant coupling within the nitrogenous population. The distance correlation between res_homo and res_μ in the 1N class reached 0.8145, which contrasts sharply with the decoupled behavior observed in pure hydrocarbons ($dCor = 0.1631$) and the strictly one-oxygen (1O) class ($dCor = 0.1384$).

Visualization of the optimal transport plan as a vector field (Figure 4) reveals a coherent flow in the nitrogen manifold. Unlike the oxygen or fluorine cases, where the vectors show more isotropic dispersion, the nitrogen vectors demonstrate a directed shift in the res_homo versus res_r^2 space. This indicates a systematic re-alignment of frontier orbital energies with the molecular spatial distribution, unique to the nitrogenous class.

3.4 Structural drivers and scaffold swap controls

To isolate the causal effect of the heteroatom and rule out trivial geometric explanations, we analyzed the HOMO-dipole coupling across specific functional groups and isostructural scaffold swaps. As shown in Figure 5A, primary amines exhibit the highest coupling ($dCor = 0.5817$), whereas tertiary amines show negligible coupling ($dCor = 0.0927$).

Furthermore, within 1,931 matched isostructural pairs of 1N versus hydrocarbon molecules (Figure 5B), the coupling between res_homo and res_μ for the 1N members was 0.1488, nearly identical to their hydrocarbon counterparts (0.1465). This indicates that the apparent high global coupling in the 1N class is largely driven by the specific structural skeletons favored in the dataset, rather than a purely electronic effect of the nitrogen atom that survives isostructural substitution.

3.5 Local density clustering and chemical drivers

To map the localized nature of this manifold reintegration, we applied density-based clustering (DBSCAN) to the heteroatom class manifolds in the 3D residual space. This identified stable regions of high local density,

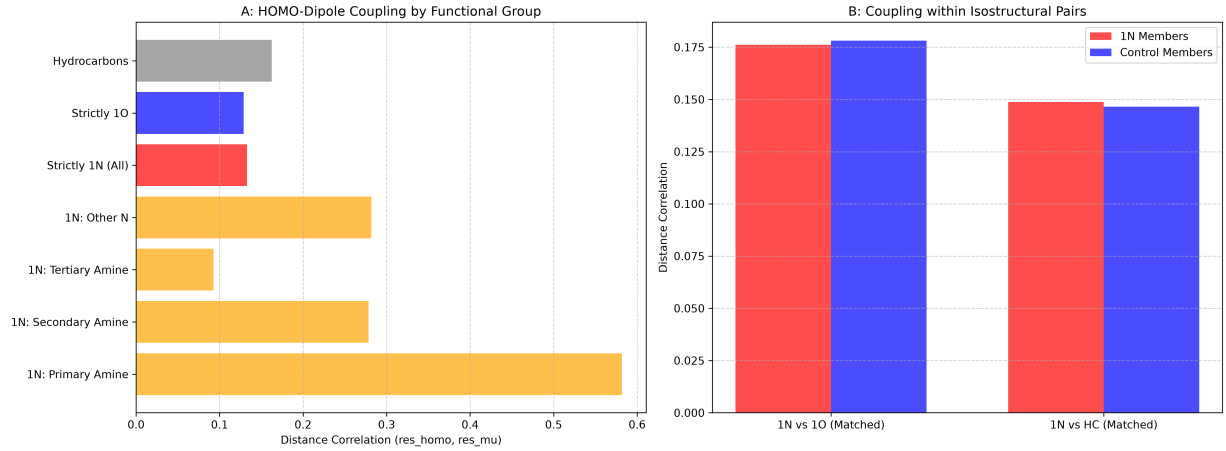

Figure 5: Distance correlation between the residuals of the HOMO energy and dipole moment. Panel A displays the HOMO-dipole coupling across different functional groups, revealing that primary amines exhibit the highest coupling ($dCor = 0.5817$), while tertiary amines show negligible coupling ($dCor = 0.0927$). Panel B compares the coupling within isostructural scaffold-swapped pairs of nitrogen-containing (1N) molecules against oxygen-containing (1O) and hydrocarbon (HC) controls. The nearly identical distance correlations between 1N members and their matched controls indicate that the apparent high global coupling and HOMO state reintegration in the nitrogen class are primarily driven by the specific structural skeletons favored in the dataset, rather than a purely electronic effect of the heteroatom that survives isostructural substitution.

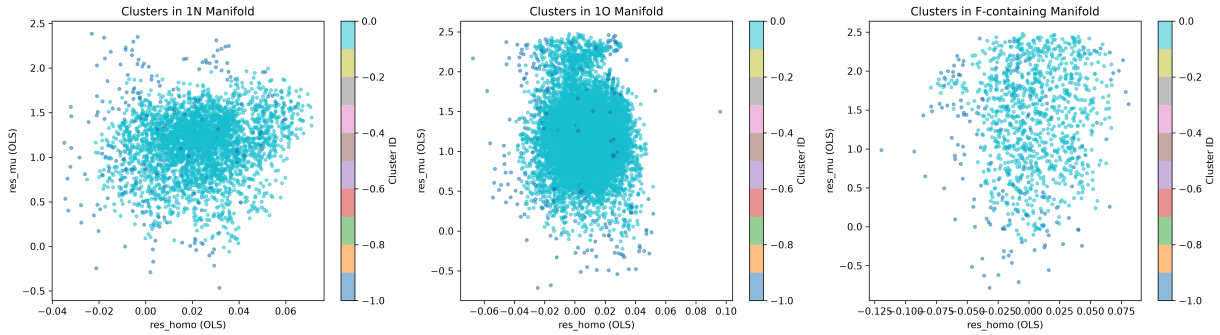

Figure 6: Density-based clustering (DBSCAN) of the joint frontier-geometric manifold, displaying Ordinary Least Squares (OLS) residuals for HOMO energy (res_homo) versus dipole moment (res_μ) across strictly 1N, strictly 1O, and F-containing molecular classes. Points are colored by cluster ID, identifying stable regions of high local density, such as the prominent Cluster 0 in the 1N manifold. By controlling for molecular size through dual-residual extraction, these distributions reveal that the decoupling of HOMO energy is not a universal constant. Instead, the distinct structural organization within the 1N manifold illustrates a localized reintegration of the frontier orbital state with spatial response variables, a behavior that differs significantly from the distributions observed in the oxygenous and fluorine-containing populations.

such as Cluster 0 in the 1N class, which contains 3,470 molecules and exhibits a $dCor$ of 0.1597 (Figure 6). The stability of these clusters was confirmed via split-half testing, and a low point-biserial correlation with Mahalanobis distances ($r = -0.18$) confirms that these regions are not artifacts of DFT convergence failures.

Finally, chemical driver identification using a Decision Tree classifier pinpointed the number of rings (`n_rings`, importance = 0.5466) and hydrogen count (`n_H`, importance = 0.2268) as the primary structural factors distinguishing these reintegrated populations. Projecting the populations onto these top descriptors (Figure 7) clearly demonstrates that the reintegration of the HOMO state is highly localized within cyclic, saturated nitrogenous structures.

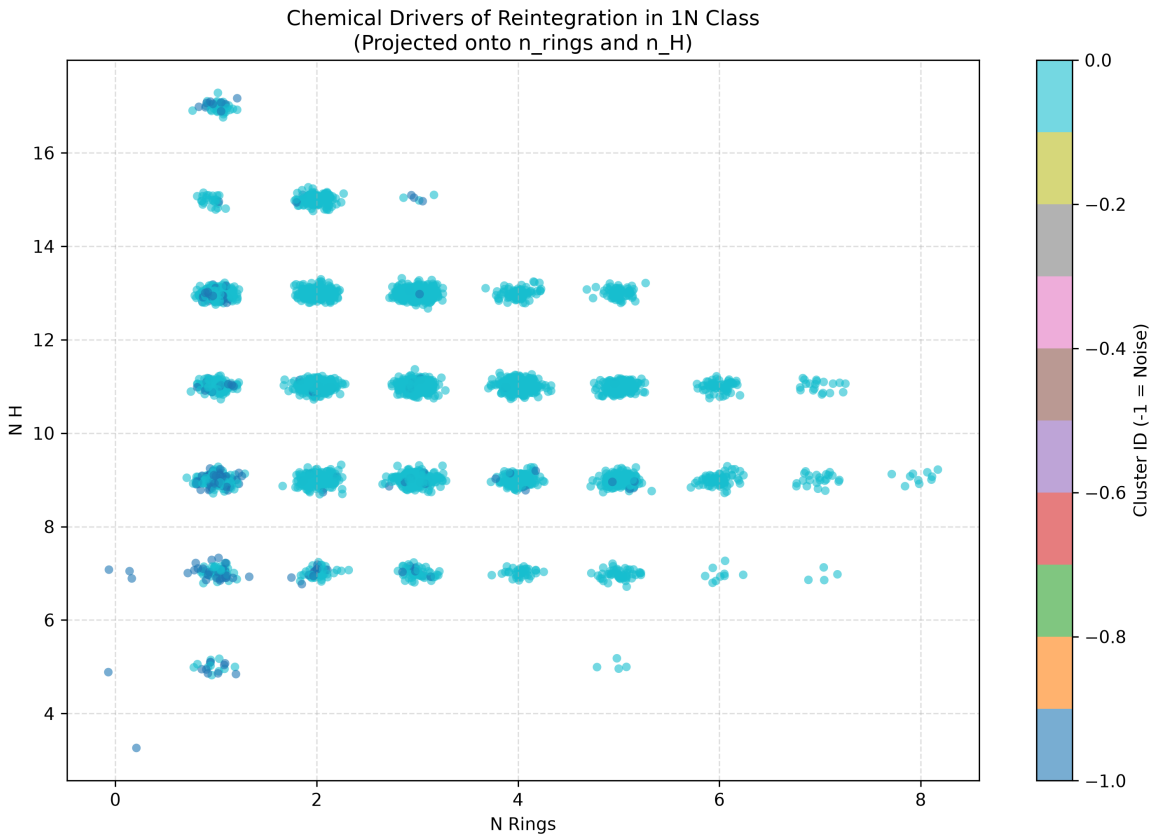

Figure 7: Scatter plot illustrating the chemical drivers of HOMO state reintegration within the 1N molecular class, projected onto the primary structural features identified by a Decision Tree classifier: the number of rings (`n_rings`) and hydrogen count (`n_H`). Data points are colored by DBSCAN cluster ID, highlighting Cluster 0 as the dominant stable region of high local density, alongside noise points (-1). The distribution reveals that the reintegration of the frontier orbital energy into the broader geometric manifold is highly localized within cyclic, saturated nitrogenous structures, demonstrating that this coupling is heavily contingent on specific molecular skeletons.

4 Conclusions

The assumption that the Highest Occupied Molecular Orbital (HOMO) energy is universally decoupled from macroscopic spatial and electrostatic properties is a pervasive simplification in the analysis of large-scale chemical datasets. In this study, we challenged this paradigm by mapping the joint frontier-geometric manifold of the QM9 dataset across distinct heteroatom classes. By extracting size-controlled residuals for HOMO energy, electronic spatial extent, and dipole moment (res_homo, res_r^2, res_μ), we successfully isolated intrinsic electronic behaviors from the confounding effects of stoichiometric scaling and molecular size.

Our information-geometric framework, leveraging exact 2-Wasserstein distances (W_2), provided a mathematically rigorous quantification of how specific electronegative environments deform the molecular property manifold relative to a pure hydrocarbon baseline. Crucially, we demonstrated that high-throughput density functional theory (DFT) datasets contain latent geometric inconsistencies that can masquerade as valid chemical variance. By employing Mahalanobis outlier profiling to truncate these extreme multidimensional tails, we significantly stabilized the manifold comparisons. As shown in Table 1, the cleaned W_2 distances confirm that the observed geometric deformations are intrinsic chemical shifts rather than dataset anomalies.

Through this refined lens, we found that the decoupling of the HOMO state is not a universal quantum feature. While fluorine induces the largest overall geometric deformation due to its extreme electronegativity, nitrogen-containing molecules exhibit a unique, systematic reintegration of the frontier orbital state with spatial response variables. This directed re-alignment is entirely absent in pure hydrocarbons and significantly muted in oxygenous counterparts.

However, our targeted isostructural scaffold swaps and functional group controls revealed a critical nuance: this reintegration is not a purely electronic consequence of the nitrogen atom itself. Instead, the coupling is heavily contingent on the underlying molecular skeleton. The phenomenon is highly localized within specific structural families, predominantly cyclic, saturated nitrogenous structures and primary amines. When controlling for these specific scaffolds, the apparent high global coupling in the nitrogen class aligns closely with the hydrocarbon baseline.

Ultimately, this work demonstrates that global structural averages and standard correlation metrics can easily mask critical joint organization within specific chemical families. By combining dual-residual extraction, optimal transport metrics, and rigorous isostructural controls, we have shown that the frontier-geometric manifold is highly dynamic. Future explorations of vast chemical spaces must account for these localized, skeleton-dependent deformations, as the tools of optimal transport offer a powerful mechanism to uncover hidden structure-property relationships that traditional global analyses overlook.