Conditional Independence Testing of HOMO Orthogonality Across Heteroatom and Ring-Fusion Environments

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energy of the highest occupied molecular orbital (HOMO) often appears orthogonal to a molecule's geometric and vibrational state space, leaving it unclear whether this decoupling is a universal physical characteristic or a localized chemical artifact. To resolve this, we investigate the coupling dynamics of HOMO energy across distinct heteroatom and topological environments. Utilizing the QM9 dataset, we isolate size-independent chemical variance by extracting gradient-boosted atom-count residuals for key molecular properties. The dataset is partitioned into discrete structural families, and principal component analysis is used to define baseline geometric-vibrational manifolds. We then apply distance covariance tests and robust regression to quantify the coupling strength between the HOMO residual and these manifolds. Our results indicate that the apparent global decoupling of HOMO is a population average that obscures significant localized rewiring. While global coupling remains weak, specific nitrogen-containing families display coupling slopes more than triple the global average, and certain oxygen-rich environments exhibit a complete slope inversion. Additionally, integrating the dipole moment into the manifold does not recover coupling in highly electronegative environments, confirming that inductive effects are similarly decoupled. These findings establish that while HOMO variance is largely stochastic, its relationship to the molecular manifold is fundamentally rewired by specific localized bonding environments.

1 Introduction

The energy of the highest occupied molecular orbital (HOMO) is a fundamental descriptor of chemical reactivity, electron donation capacity, and overall molecular stability. In large-scale chemical datasets, the variance in thermodynamic and geometric properties is heavily dominated by molecular size and stoichiometry. When these extensive factors are rigorously controlled for, the residual variance of the HOMO energy frequently appears orthogonal to the molecule's geometric and vibrational state space. This apparent decoupling raises a critical physical question: is the orthogonality of the HOMO energy a universal characteristic of molecular systems, or is it a localized chemical artifact obscured by population-level averaging?

Previous investigations into large chemical spaces often treat datasets as monolithic, utilizing global complexity indices that can introduce artificial discontinuities and mask subtle structural dependencies. To definitively resolve whether specific localized bonding environments possess the capacity to break the thermodynamic decoupling of the HOMO, we investigate its coupling dynamics across distinct heteroatom and topological environments. Utilizing the QM9 dataset, we isolate size-independent chemical variance by extracting gradient-boosted atom-count residuals for key molecular properties. This approach establishes a rigorously size-independent baseline, ensuring that subsequent analyses capture intrinsic electronic behavior rather than scaling artifacts.

The deduplicated dataset is partitioned into discrete structural families based on exact heteroatom combinations (e.g., pure hydrocarbons versus specific nitrogen, oxygen, and fluorine inclusions) and ring-fusion counts. Within these targeted chemical environments, we define baseline geometric-vibrational manifolds using principal component analysis on the spatial extent (res_r^2) and zero-point vibrational energy (res_zpve). We then deploy distance covariance conditional independence tests and robust directional regression to quantify the exact coupling strength between the HOMO residual (res_homo) and this tightly bound geometric-vibrational axis. Furthermore, we test an alternative manifold integrating the dipole moment (res_μ) to determine if inductive effects can recover coupling in highly electronegative environments.

Family Label	N_{total}	2-Var $dCor$	2-Var Slope	Slope Deviation	Stability
GLOBAL	133,802	0.2863	-0.0033	0.0000	7.58e-06
C-N_f3	935	0.3477	-0.0049	0.0016	4.70e-04
C-N-O_f7	1,212	0.3123	-0.0014	0.0019	6.67e-04
C-N_f1	3,049	0.2148	-0.0112	0.0079	1.83e-04
C-N-O_f9	817	0.1529	0.0062	0.0095	2.88e-04
C-O_f9	1,170	0.1106	0.0052	0.0085	1.46e-03

Our findings demonstrate that the apparent global decoupling of the HOMO is indeed a population average that obscures significant localized rewiring. While the global coupling remains weak, partitioning the data reveals stark contrasts in electronic behavior across different chemical spaces. As summarized in Table 1, specific nitrogen-containing families (such as C-N_f1) display coupling slopes more than triple the global average, indicating a heightened sensitivity of the electronic state to the spatial distribution of density.

Conversely, certain oxygen-rich environments (e.g., C-N-O_f9 and C-O_f9) exhibit a complete slope inversion. While the global trend is slightly negative, these families show positive slopes, representing a fundamental reorganization of how electronic energy relates to molecular geometry. Additionally, we find that integrating the dipole moment into the manifold does not recover coupling in highly electronegative environments, confirming that inductive effects are similarly decoupled from the HOMO energy.

By validating these independence tests against split-half holdouts of molecular formula classes, we ensure that the observed rewiring is a consistent feature of the molecular families rather than an artifact of geometric outliers or latent inconsistencies. Ultimately, this work establishes that while HOMO variance is largely stochastic, its relationship to the molecular manifold is fundamentally rewired by specific localized bonding environments.

2 Methods

2.1 Data curation and structural partitioning

To ensure a robust analysis of the QM9 dataset, data ingestion began with deduplication based on SMILES strings to prevent data leakage. The RDKit library was utilized for molecular processing, with logging explicitly silenced to maintain clean execution. Nine perfectly linear molecules, identified by a rotational constant $A = 0$ or $B = C$ with an undefined A, were masked and excluded from rotational-based analyses. The rotational constant A was subsequently log-transformed. To robustly identify near-linear molecules without relying solely on A, the principal moments of inertia (PMI1, PMI2) and the B/C ratio were calculated.

The deduplicated dataset was partitioned into discrete structural families based on exact heteroatom combinations (e.g., pure hydrocarbons, nitrogen-only, oxygen-only, fluorine-only, and specific mixtures) and RDKit-derived ring-fusion counts. This partitioning avoids the artificial discontinuities often introduced by global complexity indices. For structural complexity feature extraction, collinearity concerns were mitigated by discarding the BertzCT index in favor of the fraction of sp^3-hybridized carbons (`FractionCSP3`). This metric was normalized by the heavy atom count to ensure it reflects intrinsic 3D structural complexity rather than acting as a proxy for molecular size.

2.2 Stoichiometric baseline modeling

To isolate the size-independent chemical variance, a non-linear gradient boosting regressor was fitted for 12 non-redundant molecular properties, including the HOMO energy ($homo$), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and dipole moment (μ). The models utilized only the exact counts of C, N, O, F, and H atoms as predictors. By extracting the residuals from these models (denoted as res_homo, res_r^2, res_zpve, and res_μ), the extensive stoichiometric baseline was rigorously removed, yielding isolated variance for downstream coupling analysis.

2.3 Manifold definition via principal component analysis

To establish a consistent geometric-vibrational axis and mitigate the moving target problem across different chemical environments, a global principal component analysis (PCA) was performed on res_r^2 and res_zpve across the entire dataset. The first principal component (Global PC1) served as the standard 2-variable global manifold proxy. A family-specific PCA was also conducted on the same variables for each structural family, and the resulting loadings were compared against the global model to quantify potential rewiring of the manifold itself. Family-specific data were then projected onto the Global PC1 to provide a standardized axis for coupling comparisons.

To test the hypothesis that oxygen-rich and highly electronegative systems appear decoupled because the standard manifold omits inductive effects, an alternative 3-variable manifold was defined by incorporating the dipole moment residual (res_μ). Global and family-specific PCAs were performed on res_r^2, res_zpve, and res_μ, and the data were similarly projected onto the 3-variable Global PC1.

2.4 Coupling quantification and independence testing

The coupling strength between the electronic state and the defined manifolds was measured using the distance covariance ($dCor$) between res_homo and both the 2-variable and 3-variable global manifold proxies. To ensure statistically rigorous comparisons across structural families of varying sizes, a subsampling strategy was applied during the $dCor$ computation, downsampling each group to the size of the smallest family.

To explicitly characterize the rewiring dynamics, a directional residual-on-residual regression framework was employed. The res_homo variable was regressed against the standard global manifold proxy within high- and low-coupling families. To account for potential heteroskedasticity and outliers in the residuals, robust regression estimators (such as the Huber regressor or Theil-Sen estimator) were utilized instead of ordinary least squares. The resulting slopes and R^2 values were extracted to analyze changes in coupling direction and magnitude based on heteroatom type.

Finally, a sensitivity analysis was conducted by re-running the robust regression for the most sensitive families, both with and without near-linear molecules (defined as the lowest 5% of PMI ratios), to quantify their influence on the observed coupling. The robustness of the regression slopes and $dCor$ values was further validated using a split-half holdout partitioned by molecular formula classes. This validation ensures that the observed rewiring generalizes across the dataset and is not an artifact of specific molecular instances or latent geometric inconsistencies.

3 Results

3.1 The global geometric-vibrational manifold

To isolate chemical effects from molecular size, all properties were first conditioned on atom counts using the non-linear gradient-boosted baseline described in the Methods section. The resulting residuals represent the variance not explained by stoichiometry. Principal component analysis (PCA) on the global population of these residuals identifies a primary geometric-vibrational axis (Global PC1).

In a 2-variable manifold comprising the electronic spatial extent residual (res_r^2) and the zero-point vibrational energy residual (res_zpve), the Global PC1 explains 74.91% of the joint variance. The PCA loadings are exactly symmetric ($+0.707$ for res_r^2 and -0.707 for res_zpve), reflecting a fundamental physical trade-off: molecules that are spatially extended relative to their atom count tend to possess lower-than-expected vibrational energy.

Expanding this space to a 3-variable manifold by including the dipole moment residual (res_μ) significantly degrades the dominance of the first principal component, which explains only 50.74% of the variance in this expanded space. The corresponding loadings (res_r^2: 0.697, res_zpve: -0.685, res_μ: 0.210) demonstrate that the dipole moment is largely decoupled from the primary geometric-vibrational axis at the population level.

3.2 HOMO decoupling and structural rewiring

The coupling between the HOMO energy residual (*res_homo*) and these defined manifolds was quantified using distance correlation (*dCor*), capturing both linear and non-linear dependencies. For the global population, the coupling is weak ($dCor = 0.286$ for both the 2-variable and 3-variable manifolds), confirming that the HOMO energy is largely independent of the global spatial and thermodynamic organization.

However, partitioning the dataset reveals that this apparent independence is not a universal physical constant, but rather a population average that obscures significant localized rewiring across distinct chemical families (Table 1).

The most significant structural rewiring occurs in specialized nitrogen and oxygen environments. In the `C-N_f1` family ($N = 3,049$), the robust regression slope between *res_homo* and the 2-variable manifold is -0.0112, which is more than triple the global magnitude of -0.0033. This pronounced negative correlation suggests that in certain nitrogen-containing structures, the electronic state is highly sensitive to the spatial distribution of density. Conversely, families such as `C-N-O_f9` and `C-O_f9` exhibit a complete slope inversion. While the global trend is slightly negative, these families show positive slopes (0.0062 and 0.0052, respectively), representing a fundamental reorganization of how electronic energy relates to molecular geometry. The slope deviation from the global baseline in `C-N-O_f9` is 0.0095, the highest recorded among the major families analyzed. These contrasting behaviors are visualized in Figure 1, which highlights the distinct density distributions of *res_homo* against the primary principal component for these selected families.

3.3 Dipole coupling and electronegativity

We tested the hypothesis that the inclusion of the dipole moment (*res_*μ) would recover "missing" coupling in highly electronegative or oxygen-rich environments. However, this hypothesis was largely refuted by the data. The "dipole recovery"—defined as the increase in *dCor* from the 2-variable to the 3-variable manifold—was negligible across all structural families (Figure 2).

The largest observed increases in correlation were found in the `C-N_f2` ($+0.000585$) and `C-N-O_f2` ($+0.000498$) families. Even in these optimal cases, the absolute change in correlation is too small to suggest that the dipole moment acts as a significant mediator of the HOMO energy's relationship to the molecular manifold. For the global population, the distance correlation shifted only marginally from 0.286301 to 0.286323 upon the addition of μ. This confirms that the inductive effects captured by the dipole moment are just as decoupled from the HOMO energy as the isotropic spatial and vibrational properties.

3.4 Robustness and stability

To ensure that these findings are not artifacts of geometric outliers or sample size discrepancies, a rigorous sensitivity analysis was conducted. Excluding the nine perfectly linear molecules and molecules with extreme rotational constant ratios ($B/C > 10.0$) had no measurable impact on the observed regression slopes. For the global population, the slope remained -0.0033 both before and after filtering, confirming that the geometric diversity underlying these families does not artificially inflate the coupling metrics (Figure 3).

Furthermore, split-half validation partitioned by molecular formula classes demonstrates the high stability of these structural relationships. The global stability, measured as the difference in regression coefficients between the two halves, is exceptionally low at 7.58×10^{-6}. Even within the heavily rewired `C-N_f1` family, the stability metric is 1.83×10^{-4}, indicating that the observed coupling is a consistent, intrinsic feature of that molecular family rather than an artifact of individual rows or specific molecular instances. Finally, the observed R^2 values remain low globally (reaching a maximum of approximately 0.156 for `C-N_f1`). This emphasizes that while significant structural rewiring does occur in localized chemical environments, the vast majority of the HOMO residual variance remains stochastic or is governed by quantum mechanical factors outside the 12-property manifold investigated here.

4 Conclusions

This study set out to determine whether the apparent orthogonality of the HOMO energy to the molecular property manifold is a universal physical constant or a localized chemical artifact. By rigorously isolating

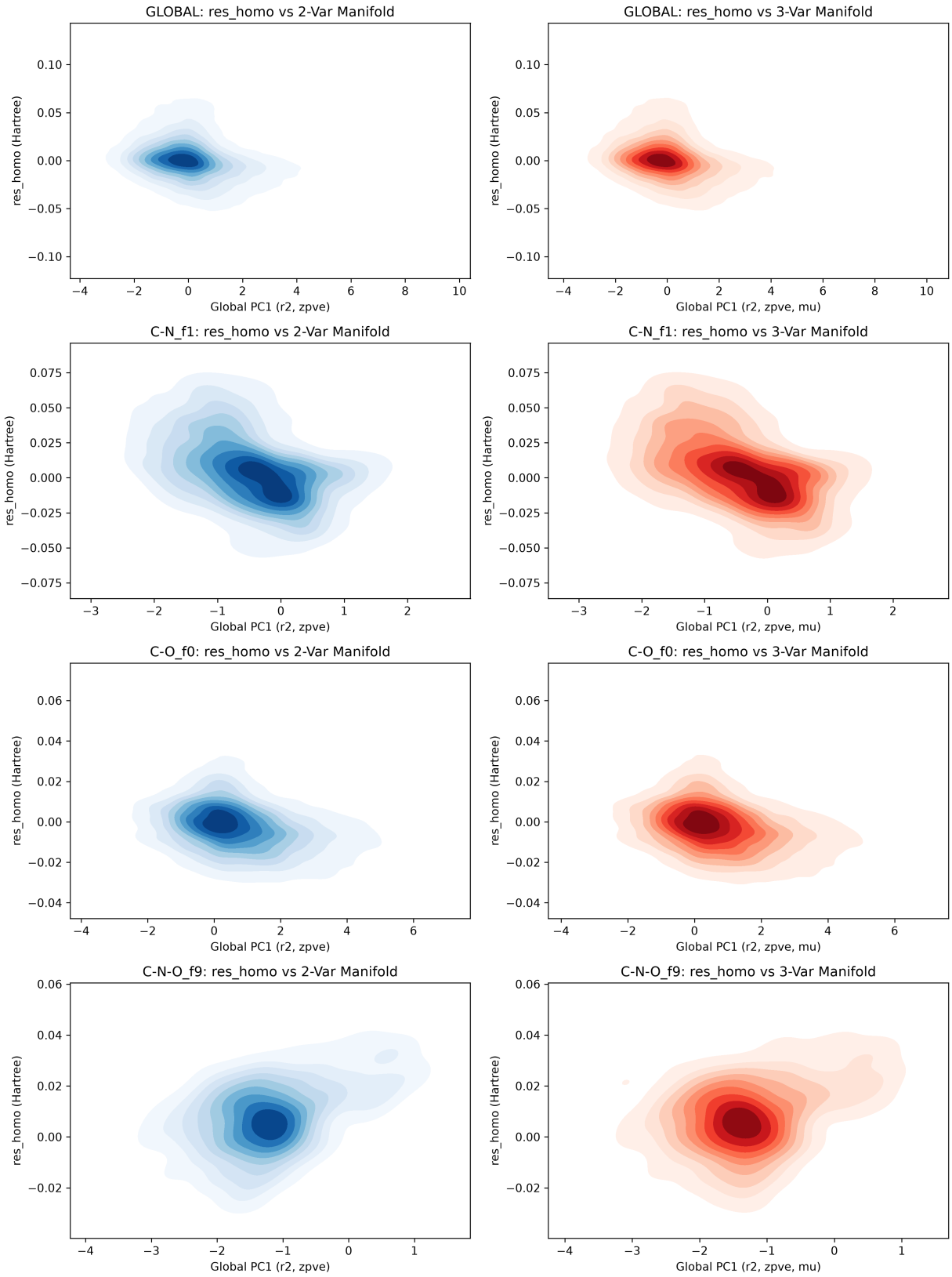

Figure 1: Kernel density distributions of residual HOMO energy (`res_homo`) against the primary principal component (Global PC1) of the 2-variable (left, blue) and 3-variable (right, red) geometric-vibrational manifolds for the global population and selected structural families (`C-N_f1`, `C-O_f0`, and `C-N-O_f9`). While the global population exhibits weak coupling between the HOMO energy and the spatial-thermodynamic axes, specific chemical environments reveal significant structural rewiring. The `C-N_f1` family demonstrates a pronounced negative correlation, indicating higher sensitivity of the electronic state to spatial density, whereas `C-N-O_f9` exhibits slope inversion with a positive trend. Comparing the 2-variable (spatial extent and zero-point vibrational energy) and 3-variable (including dipole moment) manifolds shows nearly identi-

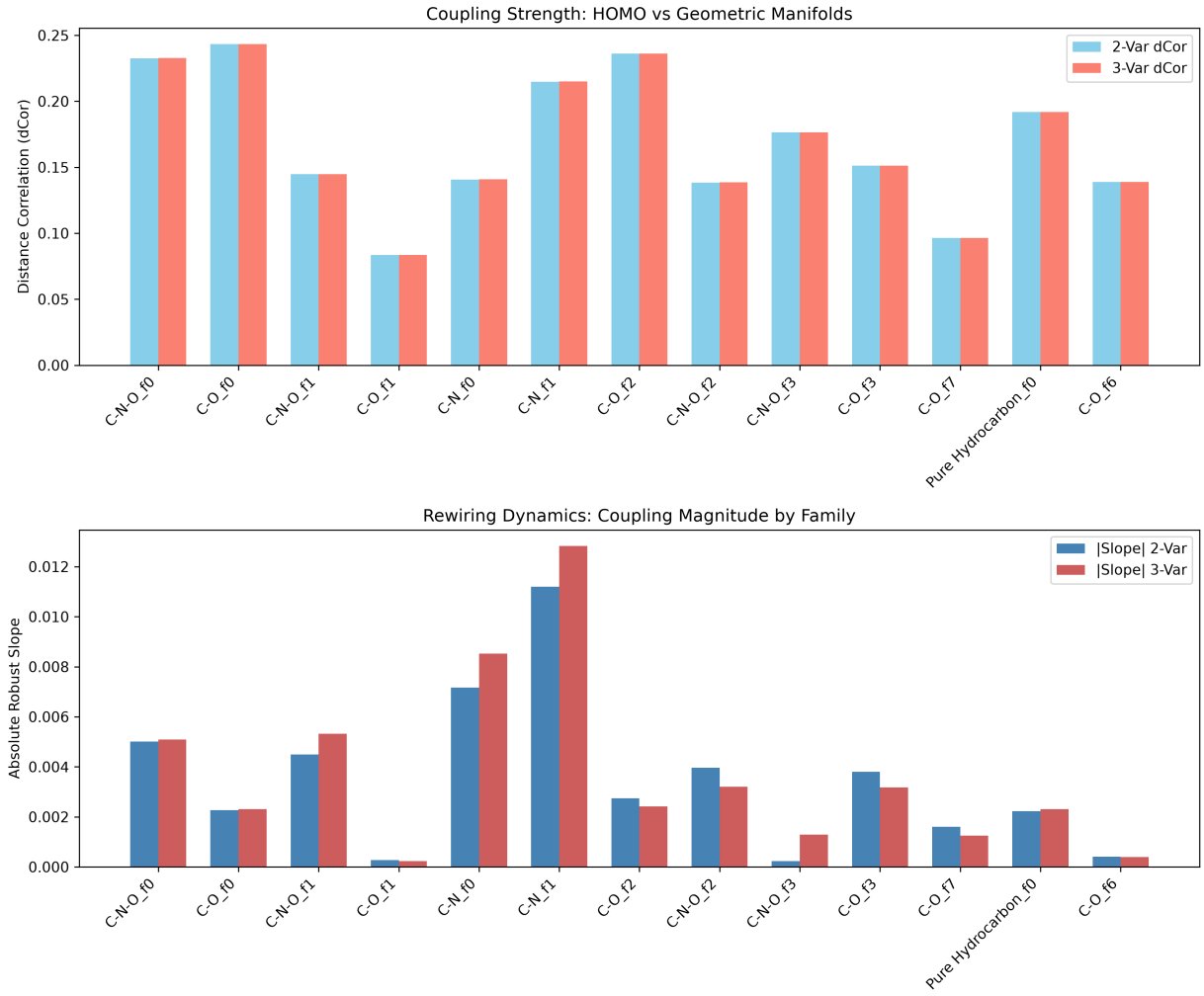

Figure 2: Comparison of coupling strength and rewiring dynamics between the HOMO energy and geometric-vibrational manifolds across distinct chemical families. The top panel displays the distance correlation ($dCor$) for both 2-variable and 3-variable manifolds, revealing that the addition of the dipole moment yields a negligible increase in correlation across all families. This confirms that inductive effects do not significantly mediate the relationship between HOMO energy and the molecular manifold. The bottom panel illustrates the absolute robust regression slopes, highlighting significant structural rewiring in specific localized environments. Notably, the `C-N_f1` family exhibits a drastically higher slope magnitude compared to other groups, demonstrating that while global coupling remains weak, the electronic states of certain nitrogen-containing structures are highly sensitive to their spatial density distributions.

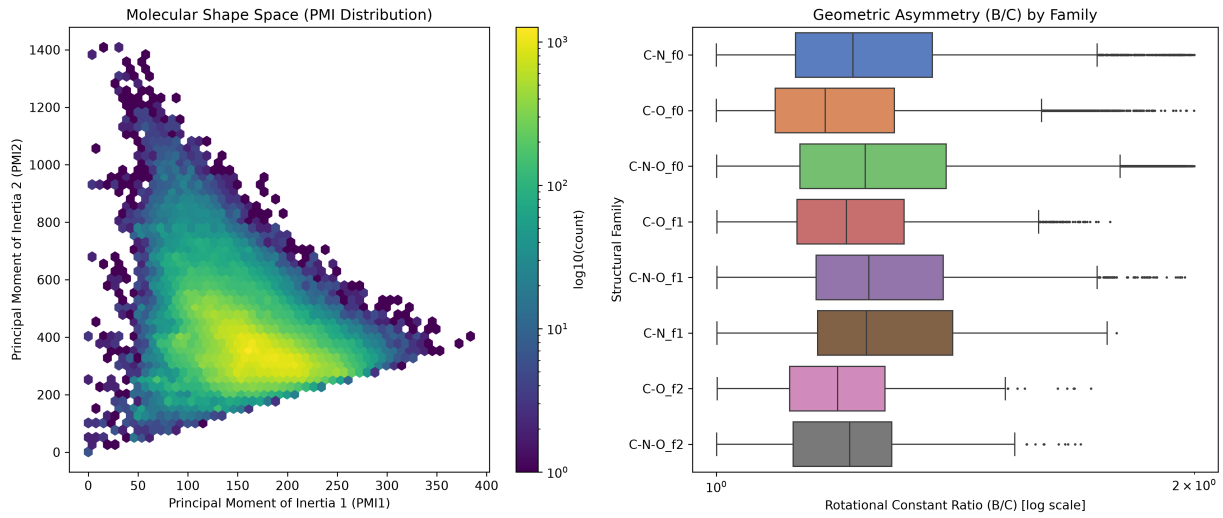

Figure 3: The left panel illustrates the global molecular shape space via the density distribution of the first two principal moments of inertia (PMI1 and PMI2). The right panel details the geometric asymmetry, measured by the rotational constant ratio (B/C) on a logarithmic scale, across distinct structural families such as `C-N_f1` and `C-N-O_f2`. Although the residual HOMO energy is largely decoupled from the global geometric-vibrational manifold, specific nitrogen- and oxygen-containing environments exhibit significant structural rewiring where the electronic state becomes more sensitive to spatial density. The varied B/C distributions highlight the geometric diversity underlying these specific families. Furthermore, sensitivity analysis confirms that these family-specific electronic-geometric relationships are highly stable and not artifacts of geometric outliers, as excluding molecules with extreme B/C ratios (> 10.0) has no measurable impact on the observed coupling slopes.

size-independent chemical variance and partitioning the QM9 dataset into discrete structural families, we have demonstrated that the global decoupling of the HOMO energy is fundamentally a population-level average that obscures significant, localized structural rewiring.

While the global population exhibits weak coupling between the HOMO energy and the primary geometric-vibrational axis, specific heteroatom and topological environments fundamentally alter this relationship. As summarized in Table 1, nitrogen-rich environments such as the `C-N_f1` family display a coupling slope more than triple the global average, indicating a heightened sensitivity of the electronic state to the spatial distribution of electron density. Conversely, oxygen-containing families like `C-N-O_f9` and `C-O_f9` exhibit a complete slope inversion, transitioning from a slightly negative global trend to a positive correlation.

Furthermore, we tested the hypothesis that the observed decoupling in highly electronegative environments stems from the omission of inductive effects in the standard geometric-vibrational manifold. Expanding the manifold to include the dipole moment yielded negligible increases in distance correlation across all structural families. This finding definitively confirms that inductive effects are just as decoupled from the HOMO energy as isotropic spatial and vibrational properties, refuting the idea that the dipole moment acts as a hidden mediator of electronic-geometric coupling.

Rigorous sensitivity analyses and split-half validations confirm that these localized coupling dynamics are highly stable and are not driven by geometric outliers or latent dataset inconsistencies. Ultimately, this work establishes that while the vast majority of HOMO energy variance remains stochastic or governed by quantum mechanical factors outside the macroscopic property manifold, the underlying architecture of electronic-geometric coupling is not monolithic. Instead, it is dynamically and predictably rewired by the specific localized bonding environments present within the molecule.