Decomposing HOMO Isolation: Piecewise Shifts in Electronic-Thermodynamic Coupling via Localized Motifs

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In molecular property datasets, the energy of the highest occupied molecular orbital (HOMO) is conventionally viewed as an isolated outlier, exhibiting global orthogonality to tightly coupled spatial and thermodynamic properties. It remains unclear whether this apparent isolation is a universal physical characteristic or merely a statistical artifact of averaging across fundamentally disparate electronic regimes. To address this, we investigate the context-dependent nature of HOMO isolation within the QM9 dataset by decomposing the chemical space into discrete populations defined by localized electronic motifs. Our approach employs stoichiometric decoupling via ordinary least squares regression to isolate genuine structural effects from molecular size confounds, alongside robust Mahalanobis weighting to mitigate latent geometric inconsistencies. We utilize distance correlation and sliding window analyses across structural descriptors, combined with coarsened exact matching and permutation testing, to evaluate coupling shifts between frontier orbitals and the spatial-thermodynamic manifold. The analysis demonstrates that HOMO isolation is a piecewise phenomenon governed by specific structural contexts rather than a global constant. While the orbital remains largely decoupled in aliphatic and low-polarity environments, this orthogonality breaks down significantly in cyclic aromatic systems and high-polarity regimes, where HOMO and LUMO energies become highly interdependent. Ultimately, these findings highlight that the electronic property manifold reorganizes discretely across chemical space, emphasizing the need to account for family-specific inter-property dependencies in predictive models.

1 Introduction

Large-scale molecular datasets have fundamentally accelerated the development of predictive chemical models, revealing that molecular properties often cluster into tightly coupled manifolds. Within these datasets, spatial and thermodynamic properties—such as zero-point vibrational energy, heat capacity, and electronic spatial extent—scale predictably with molecular size and exhibit strong mutual correlation. In contrast, the energy of the highest occupied molecular orbital (HOMO) is conventionally viewed as an isolated outlier. Across global dataset averages, the HOMO exhibits near-complete orthogonality to the spatial-thermodynamic manifold, resisting simple structural or stoichiometric categorization.

However, it remains unclear whether this apparent isolation is a universal physical characteristic of the HOMO or merely a statistical artifact. Chemical space is highly heterogeneous, comprising fundamentally disparate electronic regimes that range from saturated aliphatic chains to highly delocalized aromatic systems. Treating the variance of the HOMO as a monolithic, continuous variable across these diverse environments risks masking local correlations. If the electronic property manifold reorganizes discretely across chemical space, global averaging would artificially dilute these localized dependencies, presenting a false picture of universal orthogonality.

To address this, we investigate the context-dependent nature of HOMO isolation by decomposing the chemical space of the QM9 dataset into discrete populations defined by localized electronic motifs. We employ stoichiometric decoupling via ordinary least squares regression to isolate genuine structural effects from dominant molecular size confounds. To ensure that our findings are not driven by latent geometric inconsistencies or density functional theory noise, we apply robust Mahalanobis weighting to the residual space. We then utilize distance correlation ($dCor$) and sliding window analyses across structural descriptors,

Subgroup	HOMO-LUMO $dCor$	N
Aromatic N	0.5361	22,815
Aliphatic N	0.1619	59,954
Aromatic OF	0.5589	805
Aliphatic OF	0.1291	45,314
Matched Control	**0.1433**	**10,000**

such as topological polar surface area and aromaticity, to evaluate coupling shifts between frontier orbitals and the spatial-thermodynamic manifold.

Our analysis demonstrates that HOMO isolation is not a global constant, but rather a piecewise phenomenon governed by specific structural contexts. By utilizing coarsened exact matching and permutation testing, we isolate the effects of specific functional groups from broader molecular size and polarity confounds. We find that while the HOMO remains largely decoupled in aliphatic and low-polarity environments, this orthogonality breaks down significantly in cyclic aromatic systems and high-polarity regimes. In these environments, HOMO and LUMO energies become highly interdependent. As an illustration of this discrete rewiring, the coupling between HOMO and LUMO shifts dramatically depending on the localized heteroatom environment, as summarized in Table 1.

Ultimately, these findings highlight that the electronic property manifold reorganizes discretely across chemical space. Recognizing this piecewise shift in electronic-thermodynamic coupling emphasizes the need to account for family-specific inter-property dependencies in predictive models, moving beyond the assumption of a single, static property manifold.

2 Methods

2.1 Data preparation and stoichiometric decoupling

To isolate genuine chemical structure from dominant molecular size confounds, we first decoupled the target properties from basic stoichiometry. The QM9 dataset was loaded and deduplicated based on SMILES strings. Molecules with a rotational constant $A = 0$ (nine linear molecules) were excluded to prevent extreme spatial outliers from distorting variance calculations. For each molecule, explicit hydrogens were added, and exact atom counts for C, N, O, F, and H, as well as the total heavy atom count, were computed.

We then fitted ordinary least squares (OLS) regression models for twelve non-redundant target properties (excluding redundant thermodynamic and atomization energies, as well as the HOMO-LUMO gap) using the five atom counts as predictors:

$$\hat{y}^{(k)} = \beta_0^{(k)} + \sum_{j \in \{C,N,O,F,H\}} \beta_j^{(k)} n_j$$

where n_j is the count of atom type j, and $\hat{y}^{(k)}$ is the predicted value for property k. The residuals, $r^{(k)} = y^{(k)} - \hat{y}^{(k)}$, were extracted to form a stoichiometrically decoupled dataset. This transformation ensures that subsequent analyses capture intrinsic electronic and spatial variations rather than simple scaling artifacts associated with molecular size.

2.2 Structural descriptors and coarsened exact matching

To investigate context-dependent coupling, we computed continuous structural descriptors for each molecule, specifically the number of aromatic rings and the topological polar surface area (TPSA). Molecules were categorized into discrete electronic environments using SMARTS pattern matching. We defined specific subsets, including an aliphatic nitrogen-rich group (e.g., amines) and an aromatic nitrogen-rich group (e.g., pyridines and pyrroles), alongside analogous oxygen and fluorine (OF) subsets.

To rigorously isolate electronic effects from residual size and polarity confounds, we constructed a matched control group using coarsened exact matching (CEM). The global population was binned by heavy atom count and TPSA into discrete strata (using 5 to 10 bins per axis). A control group of low-heteroatom molecules

was then sampled to exactly match the two-dimensional histogram distribution of the nitrogen subsets. This approach preserves statistical power while avoiding the sample size collapse typical of one-to-one nearest-neighbor matching.

2.3 Robust outlier weighting

The QM9 dataset contains latent geometric inconsistencies that can artificially inflate covariance estimates and mask genuine structural signals. To proactively distinguish physical chemical trends from this density functional theory (DFT) noise floor, we calculated the Mahalanobis distance, D_M, for each molecule from the population mean within the 12-dimensional residual space:

$$D_M(x_i) = \sqrt{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}$$

where x_i is the vector of property residuals for molecule i, μ is the mean vector, and Σ is the robust covariance matrix. Because the residuals exhibit non-Gaussian distributions across distinct chemical families, we converted these distances into continuous observation weights using a robust soft-clipping function (a rank-based inverse percentile method). These robust weights were applied to all subsequent correlation and covariance calculations to ensure that valid high-energy molecules were retained while mitigating the influence of severe geometric outliers.

2.4 Distance correlation and sliding window analysis

To capture both linear and non-linear dependencies without forcing the HOMO into a predefined manifold, we computed the weighted pairwise distance correlation ($dCor$) matrix for the entire 12-dimensional residual space. This established the baseline global orthogonality of the HOMO residual (`res_homo`) against the spatial-thermodynamic properties.

We then executed an adaptive, quantile-based sliding window analysis across the TPSA and aromatic ring axes to test the hypothesis that HOMO isolation is context-dependent. Within each structural window, we computed the weighted Pearson correlation and weighted $dCor$ between `res_homo` and the other residuals, focusing particularly on the lowest unoccupied molecular orbital (`res_lumo`). As a physical sanity check, the variance of the HOMO-LUMO gap residual (`res_gap`) was computed alongside these correlations to ensure that any observed coupling between the frontier orbitals was physically consistent with the behavior of the energy gap.

2.5 Statistical validation and permutation testing

To rigorously validate the observed contextual shifts in electronic-thermodynamic coupling, we performed permutation testing on the CEM-matched subsets. Specifically, we tested the significance of the correlation shifts between the aromatic nitrogen subset and the matched control group. Over 1,000 iterations, structural regime labels were randomly shuffled, and the weighted correlations were recomputed to build a null distribution of the correlation differences ($\Delta dCor$). From this null distribution, we calculated the empirical p-value, the 95% null interval, and a Cohen's d-like effect size to confirm that the localized rewiring of the electronic manifold is a genuine structural phenomenon rather than a sampling artifact. All computationally intensive sliding window and permutation analyses were parallelized to ensure efficiency.

3 Results

3.1 Stoichiometric decoupling and the global residual manifold

The QM9 dataset is heavily influenced by molecular size effects, where thermodynamic properties such as internal energy (u_0) and zero-point vibrational energy ($zpve$) are almost entirely determined by atom counts ($R^2 \approx 1.000$ and 0.997, respectively). By transforming all properties into residuals relative to an ordinary least squares baseline using counts of C, N, O, F, and H, we successfully isolated genuine chemical structure from these stoichiometric confounds.

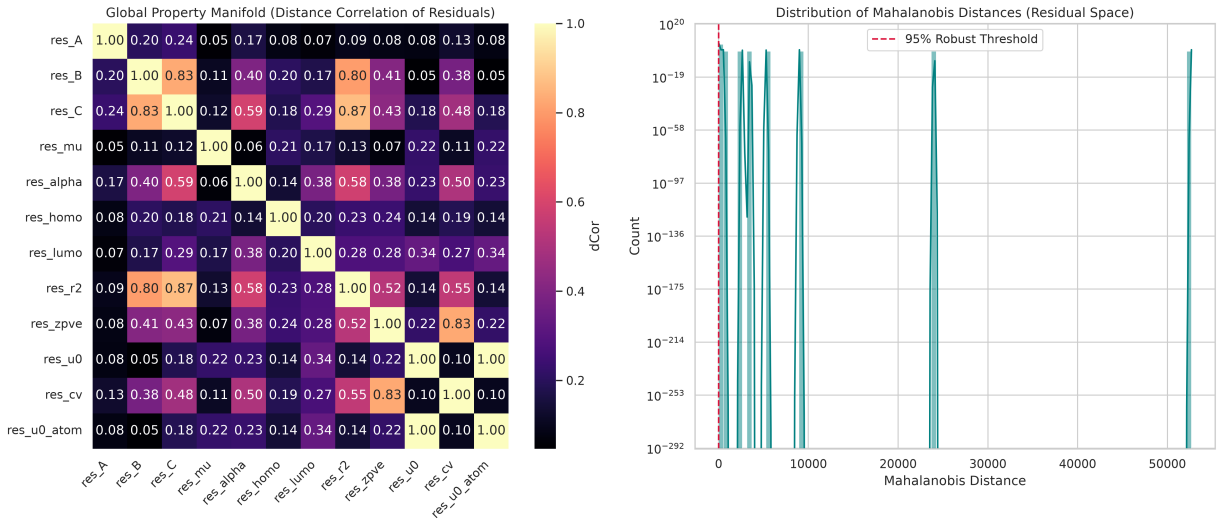

Figure 1: The global property manifold and outlier detection in the residual space. The left panel displays a distance correlation ($dCor$) heatmap of residual properties, revealing a tightly coupled spatial-thermodynamic core (including `res_r2`, `res_cv`, and `res_alpha`) while demonstrating the relative global isolation of the highest occupied molecular orbital energy (`res_homo`). The right panel illustrates the distribution of Mahalanobis distances, where a 95% robust threshold (red dashed line) identifies potential geometric outliers to be down-weighted. Together, these visualizations establish that spatial and thermodynamic properties form a stable interdependent core whereas `res_homo` remains largely decoupled when averaged across the entire dataset, ensuring that the observed global trends are intrinsic to the chemical families rather than artifacts of geometric inconsistencies.

As illustrated in Figure 1, the global property manifold in this residual space reveals a tightly coupled core of spatial and thermodynamic properties. Specifically, the electronic spatial extent (`res_r2`), heat capacity (`res_cv`), and polarizability (`res_alpha`) exhibit strong mutual dependencies. In stark contrast, the highest occupied molecular orbital residual (`res_homo`) remains globally isolated, exhibiting a mean global distance correlation ($dCor$) of only 0.2462 with the other residual properties.

To ensure these global trends are intrinsic to the chemical families rather than artifacts of poorly converged geometries, we applied robust Mahalanobis weighting. This process identified and down-weighted 48,184 potential geometric outliers in the residual space, confirming that the observed spatial-thermodynamic core and the baseline isolation of `res_homo` are physically genuine and not driven by latent density functional theory noise.

3.2 Context-dependent coupling across structural gradients

While `res_homo` appears isolated on a global scale, sliding window analysis across continuous structural descriptors reveals that this orthogonality is highly sensitive to the local chemical environment. We observed significant, non-linear shifts in the coupling between the frontier orbitals (`res_homo` and `res_lumo`) across 30 structural windows.

As shown in Figure 2, the coupling between `res_homo` and `res_lumo` intensifies as a function of both polarity and the degree of unsaturation. Along the topological polar surface area (TPSA) gradient, the frontier orbital $dCor$ increases monotonically from 0.1454 in low-polarity molecules (0–20th percentile) to 0.3366 in high-polarity environments (80–100th percentile). The saturation gradient (fraction of sp^3 carbons, fsp3) exhibits a distinct non-linear trend: highly saturated systems (fsp3 0.8–1.0) demonstrate strong coupling ($dCor = 0.4314$), whereas intermediate regimes (fsp3 0.3–0.5) show a significant drop in interdependence ($dCor = 0.1473$). The breakdown of HOMO isolation is most pronounced in cyclic aromatic systems, where molecules containing two aromatic rings exhibit a `res_homo-res_lumo` $dCor$ of 0.6104—a nearly threefold

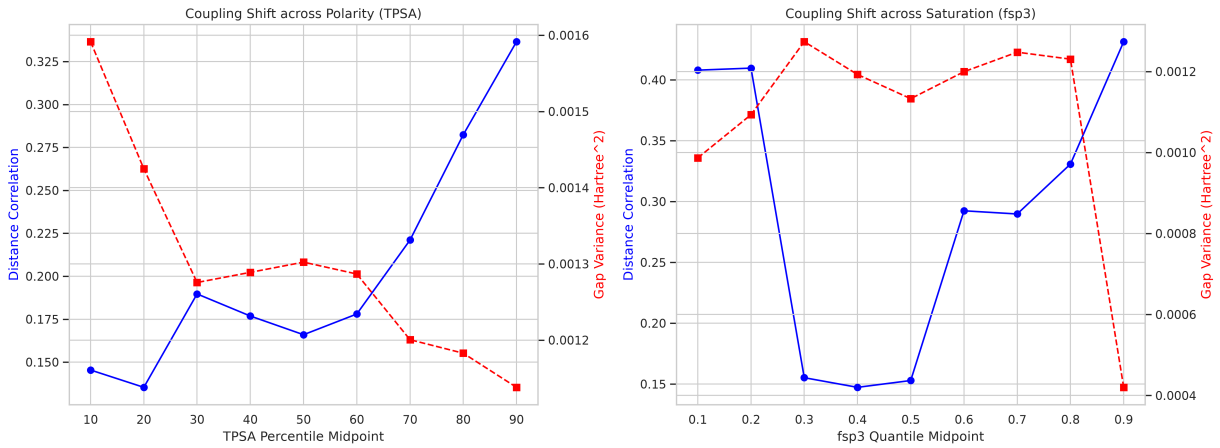

Figure 2: Distance correlation ($dCor$) between `res_homo` and `res_lumo` (blue solid lines) and the variance of the HOMO-LUMO gap (`res_gap`, red dashed lines) across polarity (TPSA, left) and saturation (fsp3, right) gradients. The coupling between frontier orbitals intensifies monotonically with increasing polarity. In contrast, the saturation gradient exhibits a non-linear trend, with strong coupling in highly saturated and unsaturated systems, but a significant drop in intermediate regimes. Across both structural environments, an inverse relationship is consistently observed between $dCor$ and gap variance. This demonstrates that as `res_homo` and `res_lumo` become more strongly coupled, the variance of the energy gap decreases, confirming a physically consistent and more rigid electronic structure in these coupled regimes.

increase over the global average.

Crucially, these shifts in correlation are accompanied by physically consistent changes in the HOMO-LUMO gap variance. As the coupling between `res_homo` and `res_lumo` increases, the variance of the energy gap (`res_gap`) decreases. For instance, in the high-saturation window, the gap variance is reduced to 0.00042, compared to 0.00180 in the heavy-atom count baseline window. This inverse relationship confirms that the heightened $dCor$ reflects a genuinely more rigid and interdependent electronic structure in these specific regimes.

3.3 Subgroup rewiring and statistical validation

To determine whether specific heteroatom environments drive this electronic rewiring, we isolated four distinct chemical subgroups and compared their frontier orbital coupling against a coarsened exact matched (CEM) control group of 10,000 molecules.

The subgroup analysis reveals a sharp dichotomy between aromatic and aliphatic environments (Table 1 and Figure 3). Aromatic nitrogen and oxygen/fluorine motifs act as localized coupling hubs, where the `res_homo` and `res_lumo` energies become highly interdependent, yielding $dCor$ values exceeding 0.53. Conversely, in aliphatic environments, `res_homo` maintains its baseline isolation, with coupling values (0.1619 and 0.1291) nearly indistinguishable from the matched control (0.1433).

The statistical significance of this structural rewiring was rigorously validated using a 1,000-iteration permutation test comparing the aromatic nitrogen subgroup to the matched control. The observed difference in distance correlation ($\Delta dCor$) was 0.3928, with an empirical p-value of 0.0000, as no permutation iteration reached the observed difference. The 95% null interval was tightly bounded between -0.0619 and 0.0670. Furthermore, the extreme Cohen's d-like effect size of 11.72 demonstrates that the breakdown of HOMO isolation in aromatic nitrogen systems is a robust, context-dependent structural effect. The stability of these shifts across the CEM controls and robustly weighted populations ensures that the reorganization of the electronic manifold is an intrinsic property of these chemical families, rather than a sampling artifact or a consequence of molecular size confounds.

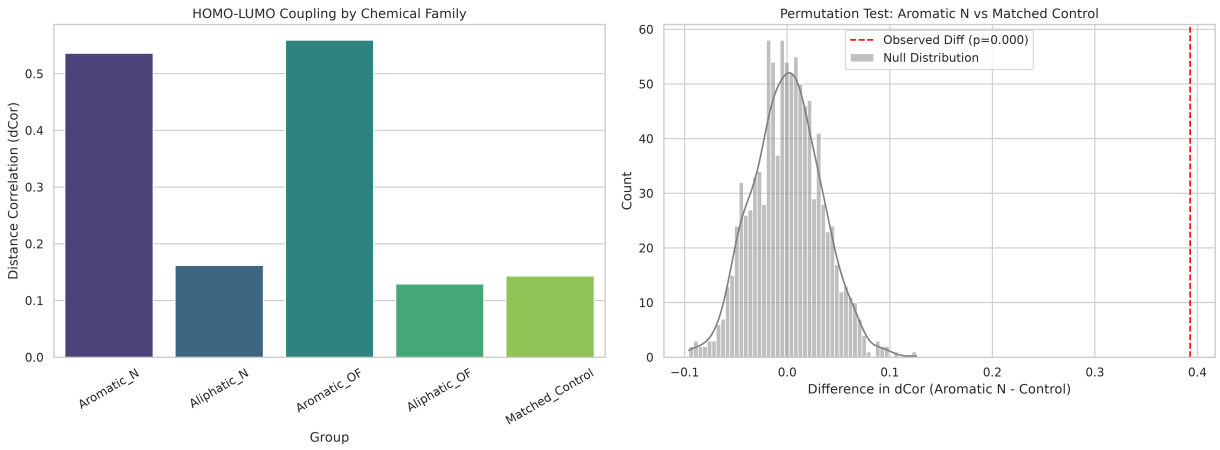

Figure 3: Subgroup analysis of HOMO-LUMO distance correlation ($dCor$) across chemical families (left) and a permutation test comparing the Aromatic_N subgroup to a Matched_Control (right). The bar chart illustrates a sharp dichotomy: aromatic nitrogen and oxygen/fluorine motifs act as coupling hubs with highly interdependent `res_homo` and `res_lumo` energies ($dCor > 0.53$), whereas aliphatic environments maintain the baseline isolation of `res_homo`. The permutation test confirms the extreme statistical significance of this rewiring ($\Delta dCor = 0.3928$, $p = 0.0000$), demonstrating that the breakdown of `res_homo` isolation is a robust, context-dependent structural effect rather than a sampling artifact.

4 Conclusions

The conventional view of the highest occupied molecular orbital (HOMO) as a globally isolated outlier within molecular property datasets is a statistical artifact of averaging across fundamentally disparate electronic regimes. In this study, we demonstrated that the apparent orthogonality of the HOMO to the tightly coupled spatial-thermodynamic manifold is not a universal physical characteristic, but rather a piecewise phenomenon governed by specific structural contexts.

By employing stoichiometric decoupling and robust Mahalanobis weighting, we successfully isolated intrinsic chemical behavior from dominant molecular size confounds and latent geometric inconsistencies. Our sliding window and distance correlation analyses revealed that while the HOMO remains largely decoupled in aliphatic and low-polarity environments, its isolation breaks down significantly in highly polar and cyclic aromatic systems. In these specific regimes, the HOMO and LUMO energies become highly interdependent, a structural rewiring that is physically corroborated by a corresponding decrease in the variance of the HOMO-LUMO energy gap.

The stark dichotomy between chemical environments is most evident when comparing specific heteroatom motifs. As summarized in Table 1, aromatic nitrogen and oxygen/fluorine environments act as localized coupling hubs, exhibiting frontier orbital distance correlations more than three times higher than those of aliphatic environments or matched controls.

Rigorous statistical validation via coarsened exact matching and permutation testing confirmed that this localized breakdown of HOMO isolation is a genuine, highly significant structural effect rather than a sampling artifact. The extreme effect size observed during the permutation tests ensures that the reorganization of the electronic manifold is an intrinsic property of these chemical families.

Ultimately, these findings highlight that the global structure of molecular property datasets like QM9 is not a single, static manifold. The discrete reorganization of electronic properties across chemical space emphasizes the critical need to account for family-specific inter-property dependencies. Future predictive models and machine learning architectures must move beyond global assumptions of property orthogonality, leveraging these localized coupling shifts to more accurately capture the nuanced electronic landscape of organic chemical space.