Distance Correlation Mapping of the HOMO Decoupling Anomaly Across Local Electronic Perturbations

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The Highest Occupied Molecular Orbital (HOMO) energy in small organic molecules often exhibits a decoupling anomaly, appearing largely independent of global spatial and thermodynamic properties. To determine whether this orthogonality is a universal constant or structurally modulated, we evaluate the sensitivity of HOMO energies to local electronic perturbations across distinct functional families. Using the QM9 dataset, we isolate chemical variance by extracting XGBoost-derived atom-count residuals and filter geometric artifacts using Minimum Covariance Determinant estimators. We then partition the molecules into mutually exclusive structural classes based on the presence of polarizing motifs. Sparse Canonical Correlation Analysis (SCCA) is applied to define a robust spatial-thermodynamic manifold, and Distance Correlation (dCor) is utilized to map the coupling strength of frontier orbitals to this manifold. Our analysis reveals that while the Lowest Unoccupied Molecular Orbital (LUMO) consistently maintains a strong coupling to the global manifold, the HOMO remains largely orthogonal in the presence of highly polarizing local motifs such as nitriles and carbonyls. Conversely, in pure hydrocarbons and heteroaromatic systems, the HOMO integrates significantly with the global physical state, bridged primarily by isotropic polarizability and electronic spatial extent. These findings demonstrate that HOMO decoupling is not a universal property but is highly dependent on functional context, behaving as a localized property in the presence of strong dipoles.

1 Introduction

The mapping of molecular properties in chemical space is a central pursuit in computational chemistry and cheminformatics. Large-scale datasets, such as QM9, have provided unprecedented opportunities to analyze the joint distribution of thermodynamic, spatial, and electronic properties across small organic molecules. While many macroscopic properties exhibit strong covariance driven by molecular size and stoichiometric composition, isolating the underlying structural variance reveals complex, non-linear relationships. Among these, the behavior of frontier molecular orbitals—specifically the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO)—is of particular interest due to their governing role in chemical reactivity, stability, and electronic transitions.

1.1 The frontier orbital decoupling anomaly

Despite the interconnected nature of molecular physical states, the HOMO energy often exhibits a striking decoupling anomaly. When stoichiometric determinism and size scaling are removed, the HOMO energy appears largely orthogonal to the global spatial and thermodynamic organization of the molecule. This notoriously low signal-to-noise baseline raises a fundamental question: is this orthogonality a universal constant inherent to the quantum mechanics of small organic molecules, or is it structurally modulated by local electronic perturbations? We hypothesize that HOMO decoupling is not a global invariant but rather a localized phenomenon driven by specific functional groups that pin the orbital density to local environments, thereby isolating it from macroscopic topological strain.

Property Residual	Canonical Weight
‘res_alpha‘	-0.8789
‘res_r2‘	0.3942
‘res_cv‘	0.2084
‘res_zpve‘	0.1653
‘res_mu‘	-0.0376

Table 1: Canonical weights defining the spatial-thermodynamic manifold.

1.2 Methodological overview

To determine the sensitivity of HOMO energies to local electronic perturbations, we systematically evaluate the QM9 dataset by isolating chemical variance from geometric artifacts. We extract non-linear atom-count residuals using XGBoost regressors to eliminate stoichiometric bias, and we apply Minimum Covariance Determinant (MCD) estimators to robustly filter out latent geometric artifacts that could masquerade as structural rewiring. The deduplicated molecules are then partitioned into mutually exclusive structural classes based on the presence of highly polarizing local motifs, such as nitriles, carbonyls, and heteroaromatic rings, which directly govern local electronic environments.

To quantify the coupling between the frontier-electronic state and the macroscopic physical state, we apply Sparse Canonical Correlation Analysis (SCCA) to define a robust spatial-thermodynamic manifold comprising zero-point vibrational energy (‘res_zpve‘), heat capacity (‘res_cv‘), electronic spatial extent (‘res_r2‘), isotropic polarizability (‘res_alpha‘), and dipole moment (‘res_mu‘). The resulting canonical weights, which define this global manifold, demonstrate that the joint organization of these residuals is primarily dominated by isotropic polarizability and electronic spatial extent (Table 1).

Using Distance Correlation ($dCor$) to map the non-linear coupling strength of frontier orbitals to this one-dimensional spatial-thermodynamic manifold, we reveal a stark contrast between the HOMO and LUMO. While the LUMO consistently maintains a strong coupling to the global manifold across all structural partitions, the HOMO remains largely orthogonal in the presence of highly polarizing local motifs. Conversely, in pure hydrocarbons and heteroaromatic systems, the HOMO integrates significantly with the global physical state. Through leave-one-out component analysis, we further demonstrate that isotropic polarizability and electronic spatial extent act as the primary bridges for this integration. These findings establish that HOMO decoupling is highly dependent on functional context, behaving as a localized property in the presence of strong dipoles.

2 Methods

2.1 Data preparation and stoichiometric baseline extraction

The QM9 dataset, comprising 133,802 unique small organic molecules, was initially preprocessed to remove 83 redundant SMILES strings, retaining only the first occurrence to prevent data leakage. To isolate the core physical properties, redundant thermodynamic and energetic features (‘u298‘, ‘h298‘, ‘g298‘, ‘u298_atom‘, ‘h298_atom‘, ‘g298_atom‘, and ‘gap‘) were excluded. The rotational constant A exhibits an extreme heavy-tailed distribution; thus, a $\log(A+1)$ transformation was applied, with the nine linear molecules ($A = 0$) explicitly masked to prevent distortion of the transformed space.

To eliminate stoichiometric determinism and size scaling—which account for the vast majority of variance in extensive properties like zero-point vibrational energy ($zpve$)—we extracted non-linear atom-count residuals. An XGBoost regressor was trained using 5-fold cross-validation to predict the remaining 12 properties based solely on the counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. The out-of-sample residuals were then extracted to represent the isolated chemical variance. Because frontier orbital residuals exhibit a notoriously low signal-to-noise baseline, a Minimum Covariance Determinant (MCD) estimator was applied to these residuals to robustly identify and filter out extreme multivariate outliers. This step ensured that latent geometric artifacts did not masquerade as structural rewiring. Finally, a rank-based normalization was applied to the filtered residuals using a quantile transformer to prepare the data for correlation analysis.

2.2 Structural partitioning by local electronic perturbations

To evaluate the sensitivity of the HOMO decoupling anomaly to local electronic environments, the filtered dataset was segmented into mutually exclusive structural families. This partitioning was implemented as a strict hierarchical 'if-elif' chain using RDKit SMARTS pattern matching to ensure reproducibility. The molecules were classified based on the presence of highly polarizing local motifs in the following order of precedence:

1. Heteroaromatic systems (aromatic rings containing at least one N, O, or F atom explicitly within the ring).
2. Nitrile-containing compounds (containing a C≡N group).
3. Carbonyl-containing compounds (containing a C=O group).
4. Ether, alcohol, or amine groups (containing N or O, excluding the above).
5. Pure hydrocarbons (strictly containing only C and H).

This partitioning yielded five distinct functional categories, allowing for the isolated analysis of frontier orbital behavior within specific local electronic contexts.

2.3 Sparse canonical correlation analysis (SCCA)

To identify the linear combinations of properties that maximize the correlation between the frontier-electronic state and the macroscopic physical state, Sparse Canonical Correlation Analysis (SCCA) was applied to the rank-normalized residuals. The variables were divided into two sets: Set X comprised the frontier orbitals ('res_homo' and 'res_lumo'), and Set Y comprised the spatial-thermodynamic response group ('res_zpve', 'res_cv', 'res_r2', 'res_alpha', and 'res_mu').

The SCCA was fit on the global filtered dataset to extract the canonical weights for Set Y, defining a robust one-dimensional spatial-thermodynamic manifold. The resulting canonical weights demonstrate the relative contribution of each spatial and thermodynamic property to the joint manifold (Table 1). Set Y was then projected into its first canonical variate to create a 1D representation of the macroscopic physical state for subsequent correlation mapping.

2.4 Distance correlation mapping and leave-one-out analysis

To quantify the non-linear coupling strength between the frontier orbitals and the spatial-thermodynamic manifold, Distance Correlation ($dCor$) was computed within each structural partition. The raw $dCor$ was calculated between the isolated frontier-electronic states ('res_homo' and 'res_lumo') and the 1D spatial-thermodynamic canonical variate. The LUMO coupling was computed alongside the HOMO to establish a baseline for frontier orbital behavior across the different functional families.

To determine whether the HOMO decoupling is orthogonal to the entire physical state or only to specific components, a leave-one-out (LOO) ablation analysis was performed on Set Y. Iteratively, each property in Set Y was removed, the SCCA canonical variate was recomputed for the remaining properties, and the $dCor$ with 'res_homo' was recalculated within each structural partition. The resulting change in coupling strength ($\Delta dCor$) was recorded to measure the information contribution of each property, identifying which spatial-thermodynamic variables act as the primary bridge between the HOMO and the macroscopic physical state.

Finally, to validate the statistical robustness of the observed coupling strengths and ensure that rewiring effects were not driven by a few highly leveraged molecules, bootstrap resampling was employed. For each structural class and LOO configuration, 100 iterations of 50% bootstrap resampling were performed to generate 95% confidence intervals for the $dCor$ values.

3 Results

3.1 Geometric outlier filtering and structural partitioning

Following the extraction of stoichiometric baselines, Minimum Covariance Determinant (MCD) estimators were applied to filter multivariate outliers and ensure that subsequent correlation mappings were not driven by latent geometric artifacts. The exclusion rates varied by structural family, ranging from 1.61% in heteroaromatics to 5.16% in pure hydrocarbons (Figure 1A).

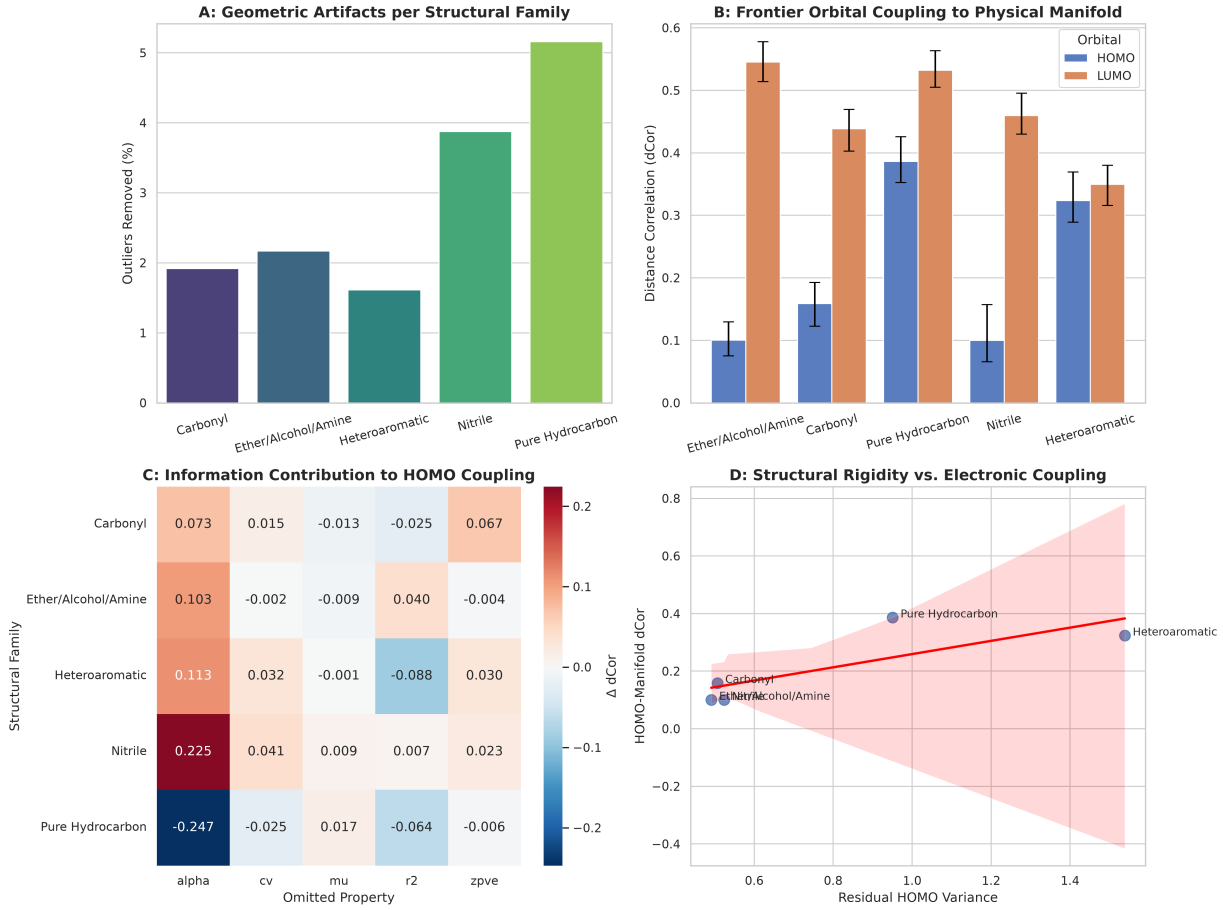

Figure 1: Structural modulation of the HOMO decoupling anomaly across five functional families in the QM9 dataset. (A) Percentage of geometric outliers removed during preprocessing, ranging from 1.61% in heteroaromatics to 5.16% in pure hydrocarbons. (B) Distance correlation ($dCor$) of frontier orbitals to the spatial-thermodynamic manifold, revealing that while the LUMO remains strongly coupled globally, the HOMO is significantly decoupled in families with localized polar groups but integrates more closely in pure hydrocarbons and heteroaromatics. (C) Leave-one-out information contribution ($\Delta dCor$) to HOMO coupling, identifying isotropic polarizability (`res_alpha`) as the primary coupling bridge for pure hydrocarbons and electronic spatial extent (`res_r2`) for heteroaromatics, whereas `res_alpha` acts as an orthogonal distractor in polarized groups like nitriles. (D) Positive relationship between residual HOMO variance and manifold coupling, confirming that delocalized or non-polar systems integrate the highest occupied state into the global physical manifold, while strong local dipoles functionally isolate it.

The final partitioned dataset yielded five mutually exclusive functional categories based on local electronic environments: ether/alcohol/amine ($N = 50,535$), carbonyl ($N = 39,674$), heteroaromatic ($N = 23,482$), nitrile ($N = 15,204$), and pure hydrocarbon ($N = 4,907$).

3.2 Composition of the spatial-thermodynamic manifold

To establish the baseline physical state of the molecules, Sparse Canonical Correlation Analysis (SCCA) was applied to the global filtered dataset. The resulting canonical weights define the latent spatial-thermodynamic manifold (Table 1).

The manifold is heavily dominated by isotropic polarizability (`res_alpha`, weight -0.8789) and electronic spatial extent (`res_r2`, weight 0.3942). This confirms that the primary axis of joint organization among the property residuals is the spatial distribution and deformability of the electron density, rather than purely thermodynamic variance.

3.3 Comparative coupling of frontier orbitals

Mapping the non-linear coupling of frontier orbitals to this one-dimensional spatial-thermodynamic manifold reveals a stark contrast between the HOMO and LUMO (Figure 1B). The LUMO residuals exhibit substantial and consistent coupling to the manifold across all structural partitions, with Distance Correlation ($dCor$) values ranging from approximately 0.460 to 0.545. In contrast, the HOMO exhibits a pronounced decoupling anomaly that is highly sensitive to the local electronic environment. In the presence of strongly polarizing motifs, such as ether/alcohol/amine and nitrile groups, the HOMO coupling is minimal ($dCor = 0.101$, 95% CI [0.075, 0.130] and $dCor = 0.100$, 95% CI [0.066, 0.157], respectively).

However, this decoupling is not a universal constant. In pure hydrocarbons, the HOMO coupling nearly reaches parity with the LUMO ($dCor = 0.386$, 95% CI [0.353, 0.426]). Similarly, heteroaromatic systems demonstrate significant HOMO integration ($dCor = 0.324$, 95% CI [0.289, 0.369]), suggesting that delocalized π-systems force a closer alignment between the highest occupied state and the global spatial properties. Furthermore, as demonstrated in Figure 1D, there is a positive relationship between residual HOMO variance and manifold coupling. This confirms that delocalized or non-polar systems integrate the HOMO into the global physical manifold, whereas strong local dipoles functionally isolate it.

3.4 Information contribution and structural rewiring

To determine which specific physical properties bridge the frontier electronic states to the manifold, a leave-one-out (LOO) ablation analysis was performed (Figure 1C). The change in coupling strength ($\Delta dCor$) upon removing individual properties highlights distinct structural rewiring mechanisms across the functional families.

In pure hydrocarbons, the HOMO coupling is extremely sensitive to isotropic polarizability (`res_alpha`). Removing this property collapses the coupling ($\Delta dCor = -0.247$), identifying polarizability as the primary bridge connecting the HOMO to the macroscopic state. In heteroaromatics, electronic spatial extent (`res_r2`) serves as the dominant bridge ($\Delta dCor = -0.086$).

Conversely, in highly polarized groups such as carbonyls and nitriles, removing `res_alpha` increases the correlation between the HOMO and the remaining manifold ($\Delta dCor = +0.073$ and $+0.225$, respectively). This indicates that in the presence of strong local dipoles, global polarizability acts as an orthogonal distractor. The HOMO becomes pinned to the local electronegative environment, refusing to organize alongside the thermodynamic and geometric-response properties that govern the rest of the molecule.

4 Conclusions

4.1 Contextual dependence of the HOMO anomaly

Our investigation into the joint structure of molecular properties within the QM9 dataset resolves the HOMO decoupling anomaly by demonstrating that it is not a universal quantum mechanical constant. Instead, the orthogonality of the HOMO energy to the macroscopic physical state is highly dependent on the local functional context. By isolating chemical variance from stoichiometric determinism and filtering latent geometric artifacts, we established a robust spatial-thermodynamic manifold. The canonical weights derived from Sparse Canonical Correlation Analysis (SCCA) confirm that this manifold is heavily dominated by isotropic polarizability and electronic spatial extent (Table 1).

Mapping the non-linear coupling of frontier orbitals to this manifold via Distance Correlation ($dCor$) revealed a fundamental dichotomy between the HOMO and LUMO. While the LUMO maintains a strong, consistent coupling to the global physical state across all structural families, the HOMO behaves as a localized property in the presence of strong dipoles. Highly polarizing motifs, such as nitriles, carbonyls, and ethers/alcohols/amines, effectively pin the HOMO density to specific local environments. This pinning functionally isolates the HOMO from the thermodynamic and geometric-response properties that govern the rest of the molecule, thereby manifesting the observed decoupling anomaly.

4.2 Physical bridges and orthogonal distractors

In contrast to polarized systems, molecules lacking strong local dipoles—namely pure hydrocarbons and heteroaromatics—exhibit significant integration of the HOMO into the global physical state. Our leave-one-out ablation analysis identified the specific physical properties that bridge this integration. In pure hydrocarbons, isotropic polarizability acts as the primary bridge connecting the HOMO to the macroscopic state. In heteroaromatic systems, the delocalized π-network forces a closer alignment with the electronic spatial extent, which serves as the dominant coupling bridge.

Crucially, in highly polarized groups, global polarizability acts as an orthogonal distractor. Removing polarizability from the manifold actually increases the correlation between the HOMO and the remaining physical state in nitrile and carbonyl partitions. This structural rewiring confirms that the HOMO decoupling anomaly is driven by a fundamental misalignment between the localized nature of the highest occupied state in polar molecules and the global deformability of the electron density.

4.3 Summary

Ultimately, these findings establish that the HOMO energy cannot be universally treated as a global molecular property in the same manner as the LUMO. For a significant portion of organic chemical space, the HOMO is functionally isolated by local electronegative perturbations, whereas the LUMO remains globally integrated. Recognizing this structural modulation is essential for the accurate interpretation of frontier orbital behavior. By demonstrating that HOMO decoupling is a localized phenomenon rather than a global invariant, this work provides a more nuanced foundation for predictive modeling, feature engineering, and property mapping in computational chemistry and cheminformatics.