UMAP-SHAP Boundary Mapping of Frontier-Electronic Decoupling in the QM9 Property Manifold

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The extent to which the highest occupied molecular orbital (HOMO) energy decouples from global thermodynamic and spatial properties in small organic molecules remains poorly understood across diverse structural regimes. To elucidate the physical motifs governing this electronic-spatial decoupling, we analyzed 133,885 molecules from the QM9 dataset. We isolated structural variance by extracting gradient-boosted atom-count residuals for twelve non-redundant properties and projected this high-dimensional residual space into a low-dimensional manifold using UMAP. By computing local partial correlations within this manifold, we established a continuous global HOMO-coupling score for each molecule. We then trained an XGBoost regressor on structural descriptors and applied SHAP feature importance analysis to map the topological boundaries of electronic decoupling. Our results demonstrate that high-coupling and low-coupling regimes are robustly distinct and predictable across previously unseen scaffolds. Furthermore, we find that HOMO decoupling is a conditional structural regime rather than a universal property; simple acyclic and highly saturated molecules exhibit strong decoupling, whereas the introduction of five- and six-membered rings and increased unsaturation drive a significant reintegration of frontier orbitals with the global physical manifold.

1 Introduction

The relationship between frontier electronic states, such as the highest occupied molecular orbital (HOMO), and global molecular properties is a foundational concept in chemical physics. While frontier orbitals dictate reactivity and electronic behavior, their energetic coupling to the broader thermodynamic and spatial organization of a molecule varies significantly across chemical space. In large, diverse datasets, the extent to which HOMO energy decouples from these global properties remains poorly understood. It is currently unclear whether electronic-spatial decoupling is a universal baseline of small organic molecules or a conditional regime dependent on specific structural motifs.

To elucidate the physical rules governing this decoupling, we present an interpretable, information-geometric framework applied to 133,885 molecules from the QM9 dataset. Traditional statistical independence tests often fail to capture localized, non-linear relationships within complex property spaces. To overcome this, we isolate pure structural variance by extracting gradient-boosted atom-count residuals for twelve non-redundant thermodynamic, electronic, and spatial properties. By projecting this high-dimensional residual space into a low-dimensional manifold using Uniform Manifold Approximation and Projection (UMAP), we establish a continuous spatial graph of chemical properties. This approach inherently isolates latent density functional theory (DFT) geometric artifacts as topological noise, allowing for a robust analysis of the underlying physical manifold.

Within this manifold, we compute local partial correlations to derive a continuous global HOMO-coupling score for each molecule. This metric quantifies the degree to which the HOMO energy integrates with the entire global structure rather than just a single property. To identify the structural drivers of this integration, we train an XGBoost regressor on a comprehensive suite of topological descriptors and apply SHapley Additive exPlanations (SHAP) to the model's decision boundary. This allows us to map the exact molecular architectures—such as specific conjugation limits or spatial confinements—that force localized frontier orbitals to wire into the global physical manifold.

Our analysis reveals that HOMO decoupling is a highly conditional structural regime rather than a universal property. We demonstrate that high-coupling and low-coupling regimes are robustly distinct and predictable, even across previously unseen molecular scaffolds. While simple acyclic and highly saturated molecules exhibit strong electronic decoupling, the introduction of five- and six-membered rings and increased unsaturation drive a massive reintegration of frontier orbitals with the global physical manifold. By mapping the topological boundaries of this electronic-spatial decoupling, this work provides a precise structural threshold where the independent behavior of frontier orbitals physically collapses.

2 Methods

2.1 Data preprocessing and stoichiometric baseline

The QM9 dataset, comprising 133,885 small organic molecules, was ingested and deduplicated based on SMILES strings to prevent data leakage during downstream model evaluation. We isolated twelve non-redundant numeric properties representing a comprehensive thermodynamic, electronic, and spatial profile: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), HOMO and LUMO energies, spatial extent (r^2), zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity at 298.15 K (C_v), and atomization energy at 0 K (u_{0_atom}). To stabilize the extreme variance and heavy-tailed distribution of the rotational constant A, a $\log(A+1)$ transformation was applied.

To isolate structural variance from trivial size and compositional effects, we established a stoichiometric baseline. For each molecule, exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were parsed. A LightGBM regressor was trained for each of the twelve properties using exclusively these five atom counts as input features. The residuals—defined as the actual minus the predicted values—were extracted to capture the variance strictly independent of molecular stoichiometry. These twelve residual vectors were subsequently standardized to zero mean and unit variance.

2.2 Geometric artifact proxy

To ensure that downstream manifold learning was not confounded by known density functional theory (DFT) geometric inconsistencies within the QM9 dataset, we constructed a continuous geometric deviation proxy. Three-dimensional conformers were generated for each molecule using the ETKDGv3 algorithm with randomized coordinates and a fixed random seed. The principal moments of inertia were estimated in GHz, and a relative error metric, $|A_{\mathrm{DFT}} - A_{\mathrm{RDKit}}|/A_{\mathrm{DFT}}$, was computed. This continuous proxy score was retained as a control variable in subsequent structural regressions to quantify and isolate the influence of geometric artifacts.

2.3 Manifold projection and local neighborhood definition

To evaluate generalizability across distinct chemical families, the dataset was partitioned into training and testing sets using a Murcko scaffold split, yielding a held-out test set of 28,026 molecules. A Uniform Manifold Approximation and Projection (UMAP) model was fit on the 12-dimensional standardized residual space of the training set to project the data into a low-dimensional manifold. Rather than applying discrete clustering to force a binary bifurcation of chemical space, we established a continuous spatial graph. For each molecule in the UMAP projection, a local neighborhood was defined using its k-nearest neighbors ($k = 500$), enabling localized statistical calculations across the manifold.

2.4 Quantifying global HOMO-coupling

Within each defined local neighborhood, the partial correlation matrix of the twelve property residuals was computed. To ensure the precision matrix remained well-conditioned and the partial correlations stable despite the localized sample size, a Ledoit-Wolf shrinkage estimator was applied to the covariance matrix.

We defined a continuous metric, the Global HOMO-Coupling Score (S_{coupling}), to quantify the degree to which the HOMO residual (res_homo) integrates with the broader physical structure of the molecule.

For a given neighborhood, this score is calculated as the sum of the squared partial correlations between *res_homo* and the remaining eleven property residuals:

$$S_{\text{coupling}} = \sum_{i \neq \text{HOMO}} \rho_p(res_homo, res_i)^2 \tag{1}$$

where ρ_p denotes the partial correlation coefficient. This continuous score was mapped over the UMAP projection to visualize the topological boundaries where electronic decoupling breaks down.

2.5 Bootstrap validation of coupling shifts

To rigorously verify that the observed spatial variations in the Global HOMO-Coupling Score were not statistical artifacts driven by varying density or sample size within the UMAP projection, a bootstrap resampling procedure was implemented. The regions corresponding to the top 10% (high-coupling) and bottom 10% (low-coupling) of scores were identified. The low-coupling region was subsampled to exactly match the sample size and local density characteristics of the high-coupling region. Partial correlation networks were recomputed on these density-matched subsamples over 200 bootstrap iterations ($n = 5000$ per iteration) to quantify the robustness and effect size of the coupling shift.

2.6 Structural driver analysis via gradient boosting and SHAP

To pinpoint the specific molecular architectures that modulate the integration of frontier orbitals with the global physical manifold, we computed a targeted suite of topological descriptors. These included the Bertz complexity index (BertzCT), the number of aromatic rings, the fraction of sp^3 hybridized carbons (FractionCSP3), the topological polar surface area (TPSA), and a symmetry index defined as the ratio of rotational constants (A/B). The continuous geometric deviation proxy score was also included as an explicit feature to control for artifact-driven predictions.

An XGBoost regressor was trained on these structural features to predict the continuous Global HOMO-Coupling Score. The model's performance was evaluated on the held-out Murcko scaffold test set to confirm the generalizability of the learned topological rules. Finally, SHapley Additive exPlanations (SHAP) feature importance analysis was conducted on the trained model to map the exact structural thresholds—such as ring size, cyclicity, and saturation limits—that dictate the boundary between the HOMO-coupled and HOMO-decoupled regimes.

3 Results

3.1 The manifold of electronic-spatial coupling

Following the extraction of stoichiometry-independent property residuals, we projected the 12-dimensional residual space into a low-dimensional manifold using UMAP. Mapping the continuous Global HOMO-Coupling Score across this manifold reveals a highly structured distribution rather than a uniform background of electronic independence. As shown in Figure 1A, the dataset bifurcates into distinct structural regimes characterized by either strong integration or near-complete decoupling of the HOMO energy from the global thermodynamic and spatial properties.

3.2 Robustness of the coupling shift

To confirm that the observed spatial variations in the coupling score were not artifacts of local density or sample size distortions within the UMAP projection, we performed a rigorous bootstrap validation. Comparing density-matched subsamples ($n = 5000$) from the top 10% (high-coupling) and bottom 10% (low-coupling) regions over 200 iterations demonstrated a complete separation of the two regimes (Figure 1B).

The high-coupling region exhibited a mean score of 0.3464 ± 0.0204, whereas the low-coupling region averaged 0.0240 ± 0.0253. This mean coupling shift of 0.3224 corresponds to an exceptionally large effect size (Cohen's d = 14.02, $p < 0.0001$). The magnitude of this separation confirms that the decoupling of the

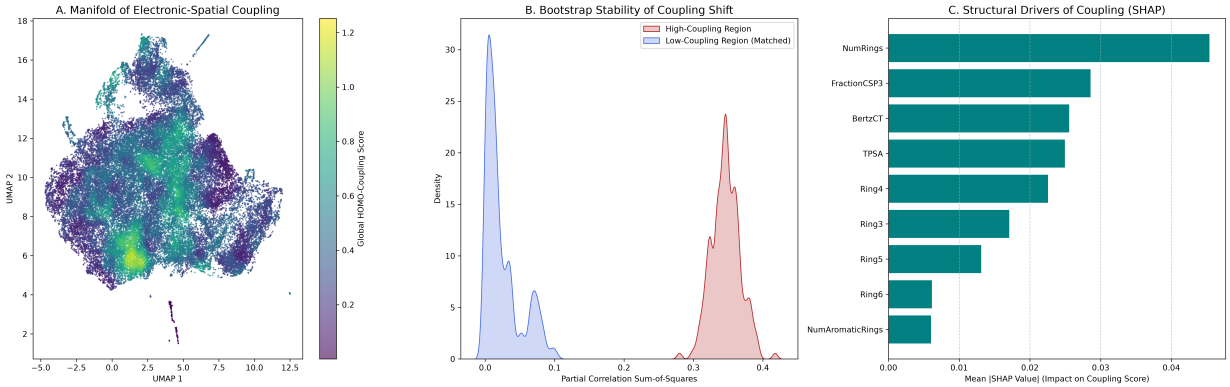

Figure 1: (A) A 2D UMAP projection of the 12-dimensional property residual space reveals distinct structural regimes of the Global HOMO-Coupling Score, demonstrating that the decoupling of the highest occupied molecular orbital (HOMO) energy is not uniform across the QM9 dataset. (B) Bootstrap validation ($n = 5000$) of the partial correlation sum-of-squares for density-matched top 10% (high-coupling) and bottom 10% (low-coupling) regions shows complete separation (Cohen's d = 14.02), confirming that this electronic-spatial coupling shift is a robust, conditional feature rather than a statistical artifact. (C) SHAP feature importance from an XGBoost regression model ($R^2 = 0.258$) identifies molecular topology as the primary driver of this reintegration. The total number of rings (NumRings) and the fraction of sp^3 hybridized carbons (FractionCSP3) strongly dictate the coupling boundary, indicating that polycyclic architectures and increased delocalization force the frontier-electronic state to integrate tightly with global thermodynamic and spatial properties, whereas simple acyclic or highly saturated molecules maintain decoupled electronic states.

frontier-electronic state is a conditional structural feature of the QM9 property space, rather than a universal baseline.

3.3 Structural drivers of frontier-orbital reintegration

To identify the specific molecular architectures responsible for this transition, we evaluated an XGBoost regressor trained to predict the continuous coupling score from topological descriptors. Evaluated on a held-out Murcko scaffold test set of 28,026 molecules, the model achieved an R^2 of 0.2579 and a root-mean-square error (RMSE) of 0.1689. This performance on previously unseen chemical families confirms that the degree of electronic-thermodynamic coupling is a predictable function of molecular topology.

Analysis of the model's feature importance, measured by normalized gain, highlights the dominance of cyclic architectures in driving HOMO reintegration. The total number of rings (`NumRings`) emerged as the primary driver (gain = 0.2289), followed closely by the presence of five-membered (0.1576) and six-membered (0.1547) rings. In contrast, three-membered rings exhibited a lower impact (0.1089), suggesting that the stability and delocalization potential inherent to standard ring sizes are critical for coupling.

SHapley Additive exPlanations (SHAP) value analysis (Figure 1C) further elucidates the directional impact of these structural features. Molecules with complex, polycyclic structures force the HOMO energy to integrate tightly with spatial and thermodynamic properties such as zero-point vibrational energy, spatial extent, and heat capacity. Furthermore, the fraction of sp^3 hybridized carbons (`FractionCSP3`) serves as a critical secondary threshold (gain = 0.1001). Lower values of `FractionCSP3`, indicative of higher unsaturation and conjugation, are strongly associated with elevated coupling scores. Conversely, simple acyclic and highly saturated molecules maintain decoupled electronic states. While heteroatom-driven polarity (`TPSA`) contributes to the model, its measurable impact is secondary (gain = 0.0512) compared to the carbon-skeleton topology. Together, these results map the precise structural boundaries where localized frontier orbitals cease to operate independently and wire into the global physical manifold.

4 Conclusions

In this study, we established that the energetic decoupling of the highest occupied molecular orbital (HOMO) from global thermodynamic and spatial properties is not a universal characteristic of small organic molecules. By mapping the 12-dimensional property residual space of the QM9 dataset, we demonstrated that electronic-spatial decoupling is instead a highly conditional structural regime.

Our information-geometric framework, which combined UMAP manifold projection with localized partial correlation networks, successfully quantified this behavior through a continuous Global HOMO-Coupling Score. The bifurcation between high-coupling and low-coupling regimes proved to be exceptionally robust (Cohen's d = 14.02) and independent of local density distortions or latent geometric artifacts. Furthermore, the ability to predict these coupling shifts on previously unseen molecular scaffolds ($R^2 = 0.258$) confirms that the reintegration of frontier orbitals is governed by generalizable topological rules rather than isolated anomalies.

Through gradient-boosted regression and SHAP feature importance analysis, we identified the precise molecular architectures that dictate this boundary. In simple acyclic and highly saturated systems, the frontier-electronic state operates independently of the global molecular framework. However, the introduction of polycyclic architectures—specifically five- and six-membered rings—and an increase in sp^2 character drive a massive reintegration of the HOMO energy with spatial and thermodynamic properties. While heteroatom-driven polarity plays a measurable role, the carbon-skeleton topology remains the dominant force in rewiring the molecule's physical profile.

These findings have significant implications for the predictive modeling of chemical properties. They suggest that for polycyclic or conjugated systems, predicting electronic behavior from thermodynamic or spatial proxies is physically grounded, as these properties are tightly coupled. Conversely, in acyclic and saturated molecules, frontier orbitals must be treated as independent degrees of freedom. Ultimately, this work provides a clear structural threshold mapping exactly where the independent behavior of frontier orbitals physically collapses, offering a more nuanced understanding of electronic organization across chemical space.