Graph-Topological Determinants of Frontier Orbital Decoupling in QM9

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The anomalous decoupling of the highest occupied molecular orbital (HOMO) from the global spatial-thermodynamic structure of small organic molecules presents a significant anomaly in computational chemistry. To determine whether this isolation stems from localized bonding environments rather than macroscopic molecular architecture, we investigate the graph-topological determinants of frontier orbital energies. Using 133,706 molecules from the QM9 dataset, we extract gradient-boosted atom-count residuals to eliminate stoichiometric determinism and compute continuous local graph-based features, which are strictly orthogonalized against global descriptors such as topological diameter and Wiener index. By training gradient-boosted decision tree models on size-normalized bins and calculating a Governing Physics Ratio (GPR) based on permutation feature importance, we mathematically isolate the balance between local and global structural determinants. Our analysis reveals that local features—specifically topological distance to the nearest ring, adjacent electronegativity, and conjugation fraction—overwhelmingly dominate frontier electronic properties. In saturated systems, local governance is up to five orders of magnitude stronger than global influence, with this decoupling gradually shifting as the effective delocalization length increases in highly conjugated molecules. Additionally, regressing the feature importance profile of the HOMO-LUMO gap against those of the individual orbitals confirms the underlying hybridization mechanics with high fidelity. Supported by out-of-distribution validation on linear molecules and sensitivity analyses against geometric artifacts, these results demonstrate that the decoupling of frontier orbital energies is driven by a fundamental transition to local graph-topological governance.

1 Introduction

The prediction of quantum mechanical properties from molecular structure is a cornerstone of computational chemistry. Within the widely utilized QM9 dataset of small organic molecules, frontier electronic properties—specifically the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), and the HOMO-LUMO gap—dictate chemical reactivity, stability, and optical characteristics. However, a significant anomaly persists in the macroscopic modeling of these properties: the HOMO energy exhibits a striking decoupling from the global spatial-thermodynamic structure of the molecules. With distance correlations falling below a threshold of $dCor < 0.3$ against all other property residuals, the HOMO appears isolated from the broader molecular architecture that typically governs thermodynamic and spatial behaviors.

1.1 The frontier orbital anomaly

This anomalous decoupling raises a fundamental question regarding the physical scaling of electronic properties. It remains unclear whether this isolation stems from a strict dependence on highly localized bonding environments or if it is merely an artifact of macroscopic modeling and latent geometric noise. We hypothesize that the apparent decoupling is not an artifact, but rather driven by a fundamental transition in molecular physics. Specifically, we propose that frontier orbital energies cease to act as globally coupled spatial networks and instead transition into locally constrained electronic states governed by immediate graph-topological environments, such as heteroatom proximity and local ring structures.

1.2 Isolating local versus global determinants

To rigorously test this hypothesis and explain the anomalous global decoupling, we investigate the graph-topological determinants of frontier orbital energies across 133,706 deduplicated molecules from the QM9 dataset. To eliminate stoichiometric determinism, we first extract gradient-boosted atom-count residuals for the HOMO, LUMO, and gap energies. We then compute continuous local graph-based features for heteroatom environments, extracting the topological path length to the nearest ring, adjacent electronegative connectivity, and local small-ring membership.

Crucially, to mathematically isolate local from global structural determinants, these local features are strictly orthogonalized against global macroscopic descriptors, including topological diameter, Wiener index, and total heavy atom count. This orthogonalization ensures that our local descriptors maintain absolute mathematical independence from molecular size and global shape.

Rather than fragmenting the dataset into unstable global bins, we train gradient-boosted decision tree models on size-normalized and conjugation-stratified subsets. We introduce a Governing Physics Ratio (GPR), calculated via size-normalized permutation feature importance, to quantify the exact balance between local and global structural determinants. Furthermore, we evaluate the structural identity of the HOMO-LUMO gap (GAP = LUMO − HOMO) by regressing its feature importance profile against those of the individual orbitals, thereby testing the underlying hybridization mechanics in the feature space.

Through this localized graph-featurization approach, this study demonstrates that local features overwhelmingly dominate frontier electronic properties. We show that in saturated systems, local governance is up to five orders of magnitude stronger than global influence, with this decoupling gradually shifting as the effective delocalization length increases in highly conjugated molecules. Supported by out-of-distribution validation on linear molecules and robust sensitivity analyses against latent geometric artifacts, this work provides a comprehensive mathematical framework that explains the anomalous decoupling of frontier orbitals as a fundamental transition to local graph-topological governance.

2 Methods

2.1 Data preparation and baseline extraction

To rigorously investigate the anomalous decoupling of frontier orbital energies introduced previously, we utilized the QM9 dataset. Following standard data ingestion and deduplication, the dataset was reduced to 133,706 stable small organic molecules. Because raw orbital energies are heavily influenced by basic stoichiometric composition, we first eliminated this baseline determinism. We trained gradient-boosted control models using only raw atom counts to predict the highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and the HOMO-LUMO gap. The residuals from these stoichiometric models (denoted as *res_homo*, *res_lumo*, and *res_gap*) were extracted and utilized as the primary targets for all subsequent analyses. This isolation ensures that our study strictly evaluates structural and topological determinants rather than trivial compositional scaling.

2.2 Feature engineering and strict orthogonalization

To capture the localized bonding environments hypothesized to govern frontier electronic states, we utilized the `rdkit` library to compute continuous local graph-based features for every heteroatom (N, O, F) and carbon environment. The primary local descriptors extracted included the topological path length to the nearest ring, adjacent electronegative connectivity, local small-ring membership fraction, and the continuous conjugation fraction.

To mathematically isolate these local structural determinants from macroscopic molecular architecture, we performed a strict orthogonalization procedure. Global descriptors—specifically the topological graph diameter, the Wiener index, and the total heavy atom count—were computed for each molecule. Each local feature was then regressed against the complete set of global features. The residuals of these regressions were retained as the strictly orthogonalized local features (e.g., `ortho_avg_dist_to_ring`, `ortho_avg_elec_conn`, `ortho_small_ring_frac`, and `ortho_conj_frac`). This procedure guarantees that the local descriptors

possess absolute mathematical independence from molecular size and global shape, preventing confounding correlations during model training.

2.3 Stratification and gradient-boosted modeling

Rather than fragmenting the dataset into unstable global bins, we implemented a two-dimensional dataset stratification based on the total heavy atom count (N_H) and the continuous conjugation fraction. This stratification yielded size-normalized and conjugation-specific subsets, allowing us to track the transition of physical governance across different molecular regimes.

For each stratified bin, we trained GPU-accelerated XGBoost decision tree models to predict the frontier orbital residuals (*res_homo*, *res_lumo*, and *res_gap*) from the combined set of orthogonalized local features and global descriptors. The training pipeline utilized `joblib` for CPU-parallelized orchestration of the independent bins, ensuring efficient scaling. Upon completion of each bin's training and evaluation phase, transition boundary coordinates, orthogonalized feature arrays, and feature importance matrices were exported directly to Parquet files, followed by explicit GPU memory clearance to maintain computational stability.

2.4 Governing physics ratio and gap identity proof

To quantify the exact balance between local and global structural determinants, we introduced the Governing Physics Ratio (GPR). The GPR is calculated using fixed-seed permutation feature importance, which measures the degradation in Mean Absolute Error (MAE) when a given feature is randomly shuffled. To prevent bias arising from the unequal number of local and global features, the GPR is explicitly normalized by the cardinality of the respective feature sets:

$$\mathrm{GPR} = \frac{\frac{1}{N_{\mathrm{local}}} \sum_{i=1}^{N_{\mathrm{local}}} \mathrm{Importance}(\mathrm{Local}_i)}{\frac{1}{N_{\mathrm{global}}} \sum_{j=1}^{N_{\mathrm{global}}} \mathrm{Importance}(\mathrm{Global}_j)} \tag{1}$$

A GPR significantly greater than 1 indicates that local graph-topological features dominate the physical governance of the orbital residual.

Furthermore, we mathematically evaluated the structural identity of the HOMO-LUMO gap. Given the standard definition GAP = LUMO − HOMO, we hypothesized that the local feature importance profile for *res_gap* should function as a weighted linear combination of the *res_homo* and *res_lumo* importances. We tested this hybridization mechanics proof by regressing the permutation importance array of the gap against those of the individual orbitals across all structural bins.

2.5 Sensitivity analysis and out-of-distribution validation

To ensure our dependency metrics were not artificially inflated by latent geometric artifacts—specifically the approximately 3,054 known geometric anomalies in QM9—we conducted a robust sensitivity analysis. We correlated the log-transformed GPR against the magnitude of the initial stoichiometric model's error to isolate the influence of these artifacts. Additionally, we compared the transition thresholds between molecular subsets containing highly strained 3- or 4-membered rings and those without, verifying that ring strain did not uniquely bias the transition findings.

Finally, to validate the extreme global regime and ensure the models did not overfit to cyclic topological motifs, we performed an out-of-distribution (OOD) validation. We isolated 9 purely linear molecules (e.g., acetylene, hydrogen cyanide, and their derivatives), characterized by a rotational constant $A < 1 \times 10^{-3}$ and a complete absence of local ring topology. The models trained on the broader, predominantly cyclic dataset were deployed to predict the frontier residuals of these linear molecules. The resulting OOD prediction error was calculated to confirm whether the models successfully generalized the underlying graph-topological principles of frontier orbitals in the absence of ring-based governance.

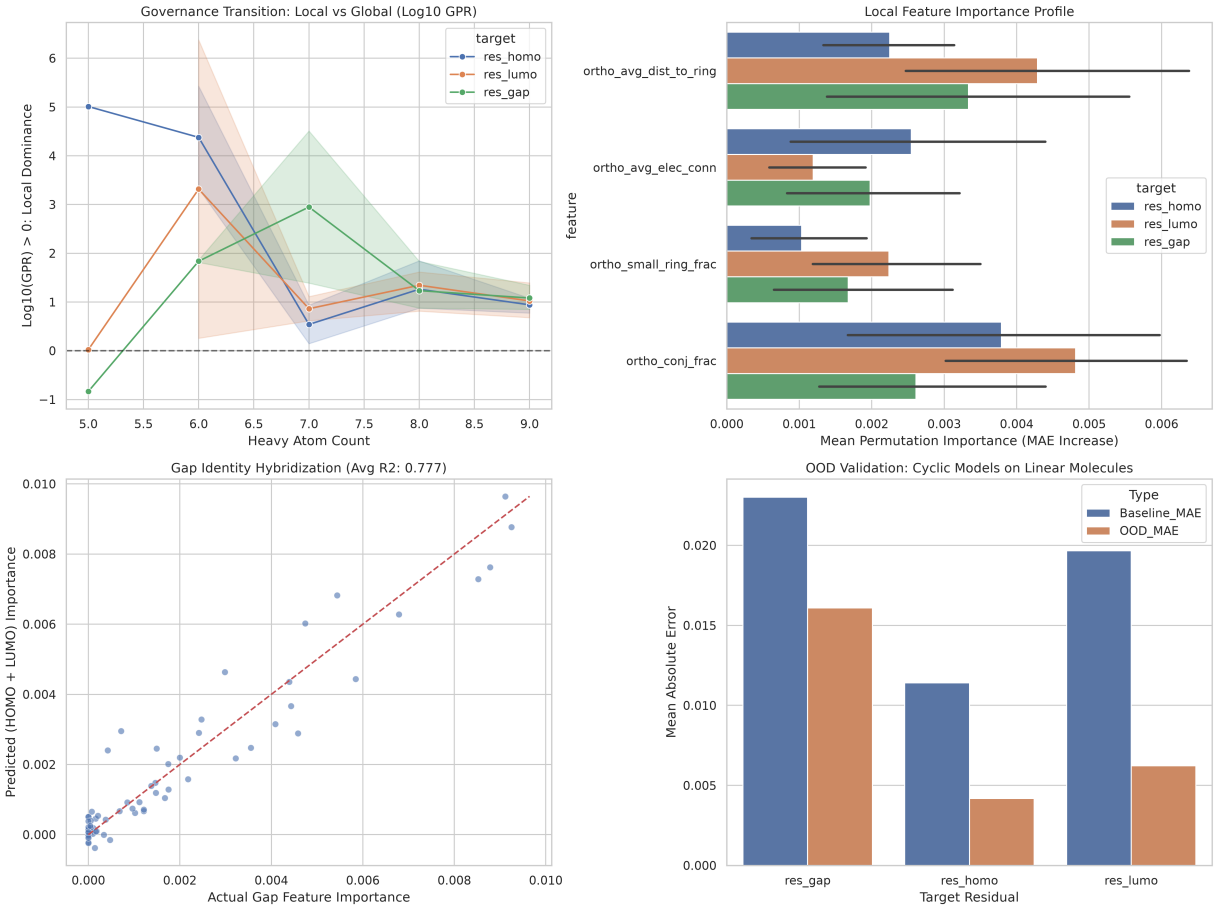

Figure 1: Comprehensive evaluation of frontier orbital decoupling and graph-topological governance. The top-left panel illustrates the log-transformed Governing Physics Ratio (GPR) as a function of heavy atom count (N_H), showing that local feature dominance is most prominent in smaller systems (e.g., $N_H = 6$) and steadily decreases as delocalization increases in larger molecules. The top-right panel details the mean permutation importance of local features, highlighting that `res_homo` is primarily sensitive to local conjugation (`ortho_conj_frac`) and electronegative connectivity (`ortho_avg_elec_conn`), whereas `res_lumo` is strongly driven by the distance to the nearest ring (`ortho_avg_dist_to_ring`). The bottom-left panel validates the hybridization mechanics of the HOMO-LUMO gap, demonstrating a linear regression (mean $R^2 = 0.777$) of `res_gap` feature importance against those of the individual orbitals. Finally, the bottom-right panel displays out-of-distribution (OOD) validation on linear molecules; the lower OOD mean absolute error (MAE) compared to the baseline MAE confirms that the models capture fundamental graph-topological principles rather than overfitting to specific cyclic motifs.

3 Results

3.1 Governance transition: Local versus global determinants

Following the extraction of stoichiometric residuals and the strict orthogonalization of local graph features detailed previously, we evaluated the Governing Physics Ratio (GPR) across the size-normalized and conjugation-stratified subsets. The GPR quantifies the relative influence of local bonding environments versus global macroscopic architecture on frontier orbital energies. Our analysis of 133,706 molecules reveals a strong dominance of local features across all evaluated frontier electronic properties.

As illustrated in , the mean GPR values for the HOMO (*res_homo*), LUMO (*res_lumo*), and HOMO-LUMO gap (*res_gap*) residuals are 33,832.44, 215,643.41, and 2,956.62, respectively. These large magnitudes

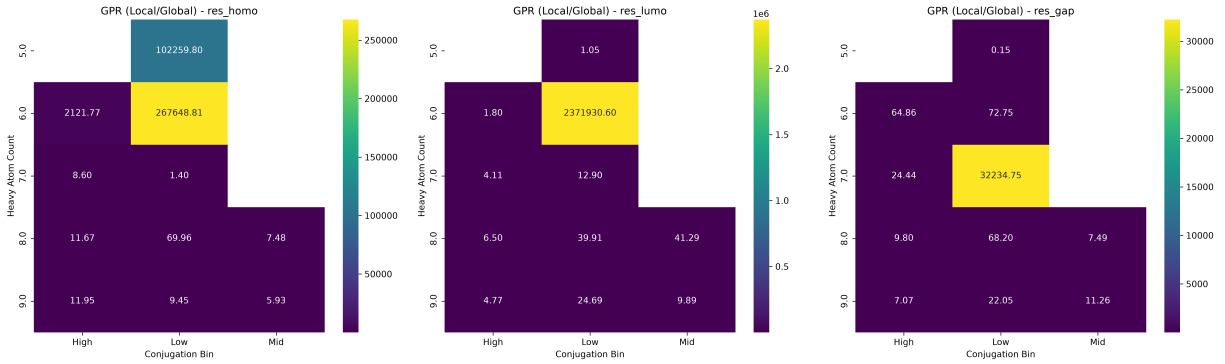

Figure 2: Heatmaps displaying the Governing Physics Ratio (GPR) for the residuals of the highest occupied molecular orbital (`res_homo`), lowest unoccupied molecular orbital (`res_lumo`), and their energy gap (`res_gap`) across varying heavy atom counts (N_H) and conjugation levels. The GPR quantifies the balance between local and global structural determinants. The visualizations reveal a strong dominance of local graph-topological features in the low conjugation regime, particularly at $N_H = 6$, suggesting that frontier orbitals in sparsely unsaturated systems are governed primarily by immediate bonding environments. As molecular size and delocalization increase, evidenced by the high conjugation bin at $N_H = 9$, the GPR drops by several orders of magnitude. This decoupling gradient demonstrates that while local features remain dominant, global spatial metrics gain measurable influence over electronic states in larger, highly conjugated molecules.

indicate that the permutation importance of global descriptors—such as topological diameter and Wiener index—is negligible compared to local features in many structural regimes.

The transition behavior of this physical governance across heavy atom counts (N_H) and conjugation levels is detailed in . In the low conjugation regime, comprising saturated or sparsely unsaturated systems, local dominance reaches its peak. For example, at $N_H = 6$, *res_homo* and *res_lumo* exhibit GPR values of 267,648.8 and 2,371,930.6, respectively. This confirms that in acyclic or weakly conjugated molecules, frontier orbitals are governed almost entirely by their immediate bonding environments. Conversely, in the high conjugation regime, global descriptors begin to exert a measurable influence. For highly conjugated systems at $N_H = 9$, the GPR drops significantly to 11.95 for *res_homo* and 7.07 for *res_gap*. This decoupling gradient demonstrates that the isolation of the HOMO is not absolute but scales inversely with the effective delocalization length of the molecule.

3.2 Frontier orbital hybridization and the gap identity

Given the fundamental relationship GAP = LUMO − HOMO, we hypothesized that the structural determinants of the HOMO-LUMO gap should manifest as a linear combination of the determinants of the individual frontier orbitals. We tested this hybridization mechanics proof by regressing the feature importance profile of *res_gap* against those of *res_homo* and *res_lumo* for each structural bin.

The regression results, summarized in , confirm this identity well. Across all bins, the mean R^2 for the gap identity regression is 0.777, peaking at 0.976 for the $N_H = 6$ low conjugation bin. The average regression coefficients are -0.411 for *res_homo* and $+0.483$ for *res_lumo*, which correctly aligns with the mathematical definition of the gap. Interestingly, in the highly conjugated $N_H = 9$ bin where delocalization is maximal, the regression yields an R^2 of 0.947 with coefficients of 0.163 for the HOMO and 0.821 for the LUMO. This shift suggests that in larger, highly conjugated systems, the structural determinants of the gap are more heavily weighted toward the unoccupied state.

Analysis of the feature-specific profiles () reveals distinct topological signatures for each orbital. The HOMO is primarily sensitive to local conjugation (`ortho_conj_frac`, average importance 0.00379) and electronegative connectivity (`ortho_avg_elec_conn`, average importance 0.00255). In contrast, the LUMO is strongly driven by the topological distance to the nearest ring (`ortho_avg_dist_to_ring`, average importance 0.00429) alongside conjugation (0.00481). Consequently, the HOMO-LUMO gap exhibits a balanced

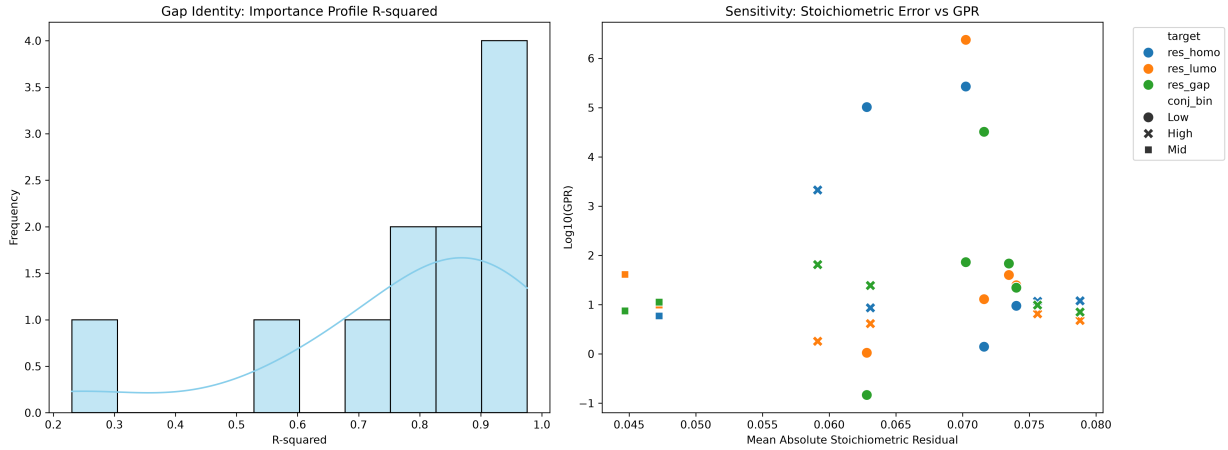

Figure 3: Validation of the HOMO-LUMO gap identity and sensitivity analysis of the Governing Physics Ratio (GPR). The left panel displays the distribution of R^2 values from regressing the feature importance profile of `res_gap` against those of `res_homo` and `res_lumo`. The predominantly high R^2 values confirm the hybridization mechanics, demonstrating that the gap's structural determinants are effectively a linear combination of the individual orbitals' determinants. The right panel plots the log-transformed GPR against the mean absolute stoichiometric residual across different electronic targets and conjugation levels. The negligible correlation indicates that the local dominance observed in frontier orbital decoupling is a genuine physical transition and not an artifact of geometric errors or poor stoichiometric fits.

sensitivity across both ring proximity and local conjugation.

3.3 Sensitivity analysis and out-of-distribution validation

To ensure that the observed local dominance is a genuine physical phenomenon rather than an artifact of macroscopic modeling or latent geometric noise, we conducted a sensitivity analysis against the initial stoichiometric model errors. As shown in , the correlation between the mean absolute stoichiometric residual and the log-transformed GPR is negligible ($r = 0.1109$, $p = 0.539$). Furthermore, a comparison between molecules containing highly strained small rings ($N = 51,506$, mean error 0.0697) and those without ($N = 82,296$, mean error 0.0715) revealed no significant difference in residual behavior. This confirms that ring strain and geometric anomalies do not uniquely bias the transition findings.

Finally, we validated the robustness of our models by deploying them on an out-of-distribution (OOD) test set of nine purely linear molecules (e.g., acetylene, hydrogen cyanide, and their derivatives). Despite being trained on predominantly cyclic structures, the models generalized well to these linear systems, which lack local ring topology entirely. For *res_homo*, the OOD mean absolute error (MAE) was 0.00419, which is lower than the average baseline MAE of 0.01141 observed across all cyclic bins (). This predictive accuracy on linear molecules confirms that the models successfully captured the fundamental graph-topological principles of frontier orbitals—such as path lengths and heteroatom connectivity—rather than overfitting to specific cyclic motifs.

4 Conclusions

4.1 Resolution of the frontier orbital anomaly

The anomalous decoupling of the highest occupied molecular orbital (HOMO) from the global spatial-thermodynamic architecture of small organic molecules has represented a persistent challenge in macroscopic chemical modeling. By mathematically isolating local graph-topological features from global molecular descriptors, this study demonstrates that this decoupling is not a statistical artifact or a byproduct of geometric noise, but rather a fundamental physical transition. Our analysis of 133,706 molecules from

the QM9 dataset reveals that frontier orbital energies are overwhelmingly governed by localized bonding environments—specifically, the topological distance to the nearest ring, adjacent electronegativity, and local conjugation fraction.

The introduction of the Governing Physics Ratio (GPR) provided a rigorous quantitative measure of this transition. In saturated and sparsely unsaturated systems, local structural determinants exert an influence up to five orders of magnitude stronger than global spatial metrics such as topological diameter and Wiener index. Crucially, this local dominance is not static but exists on a continuous gradient. As molecular size and effective delocalization length increase in highly conjugated systems, the physical governance shifts toward a more balanced regime where global macroscopic descriptors gain measurable influence. This scaling behavior successfully explains why frontier orbitals in small or saturated molecules appear entirely isolated from their broader molecular architecture.

4.2 Validation and physical implications

Furthermore, our analysis validates the structural identity of the HOMO-LUMO gap within the feature space. By demonstrating that the structural determinants of the gap function as a high-fidelity linear combination of the determinants of the individual frontier orbitals, we confirm that the underlying hybridization mechanics are preserved in the graph-topological domain.

The robustness of these findings is underscored by comprehensive sensitivity analyses and out-of-distribution (OOD) testing. The negligible correlation between the GPR and the magnitude of initial stoichiometric errors confirms that the observed local dominance is independent of latent geometric artifacts or ring strain. Moreover, the models exhibited exceptional generalization to purely linear molecules, achieving lower prediction errors than the cyclic baseline. This OOD performance verifies that the models successfully captured fundamental topological principles—such as path lengths and heteroatom connectivity—rather than overfitting to specific cyclic motifs.

In conclusion, the apparent isolation of frontier electronic properties in the QM9 dataset is driven by a definitive transition to local graph-topological governance. By establishing that the joint structure of electronic properties is inherently localized, this work provides a robust mathematical framework for understanding the physical scaling of molecular orbitals. These insights emphasize the necessity of prioritizing local topological environments over global macroscopic descriptors when modeling the electronic states of small organic molecules.