Localizing the Orthogonal Frontier: Explaining HOMO-LUMO Decoupling via Substructure-Based Distance Correlation

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The energies of the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) frequently exhibit an anomalous decoupling from the global thermodynamic and spatial properties of a molecule. To explain this orthogonality, we hypothesize that frontier orbitals are primarily governed by highly localized structural micro-environments rather than the macroscopic molecular topology that dictates thermodynamic behavior. We investigate this by isolating the structural drivers of these properties using the QM9 dataset of over 133,000 molecules. After removing stoichiometric size effects by extracting property residuals via gradient-boosted regression on exact atom counts, we encode local bonding environments using non-collinear structural descriptors and functional group counts. We then map these local features to the isolated property residuals using LightGBM regressors, distance correlation, and multi-dimensional scaling. The results demonstrate that thermodynamic and spatial properties are dictated by global topological features, whereas frontier orbital variances are driven by specific local motifs. Additionally, this local governance is shown to break down in highly conjugated systems due to the emergence of long-range electronic effects. Ultimately, topological mapping of the residual space confirms that molecular physical properties form weakly coupled modules, with frontier electronic states occupying a subspace largely orthogonal to the spatial and vibrational states.

1 Introduction

The relationship between molecular structure and physical properties is a foundational concept in chemistry and materials science. For many macroscopic and thermodynamic properties, such as internal energy, heat capacity, and spatial extent, molecular behavior is largely determined by global topology and stoichiometric composition. However, the frontier electronic states—specifically the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO)—frequently exhibit an anomalous decoupling from this global thermodynamic and spatial manifold. In large, diverse chemical spaces, the energies of these frontier orbitals often vary orthogonally to the rest of the physical state space. Understanding the structural drivers of this orthogonality is critical for rational molecular design, particularly in fields like organic electronics and photocatalysis, where tuning frontier orbitals independently of a molecule's thermodynamic stability is highly desirable.

To explain this anomalous global orthogonality, we hypothesize that frontier orbitals are primarily governed by highly localized structural micro-environments, rather than the macroscopic molecular topology that dictates thermodynamic behavior. Under this framework, specific atomic architectures and functional motifs force electronic decoupling, allowing frontier properties to drift independently of a molecule's global spatial state.

We investigate this hypothesis by isolating the structural drivers of these properties using the QM9 dataset, which contains over 133,000 diverse organic molecules. Because physical properties in such datasets are heavily dominated by molecular size and composition, we first establish a stoichiometric baseline. By extracting property residuals via gradient-boosted regression on exact atom counts, we eliminate stoichiometric determinism and isolate the variance driven purely by structural isomerism and connectivity. We then encode the local bonding environments of the deduplicated molecules using non-collinear structural descriptors and exact functional group counts.

By mapping these local features to the isolated property residuals using interpretable gradient-boosted regressors (LightGBM), distance correlation metrics, and multi-dimensional scaling, we contrast the structural features that drive HOMO and LUMO variance against those that govern the tightly coupled thermodynamic manifold.

Our results demonstrate a sharp divergence in structural governance: thermodynamic and spatial properties are dictated by global topological features, whereas frontier orbital variances are driven by specific local motifs. Furthermore, we identify the boundary conditions of this local governance, showing that it breaks down in highly conjugated systems due to the emergence of long-range electronic effects. Ultimately, topological mapping of the residual space confirms that molecular physical properties form weakly coupled modules, with frontier electronic states occupying a subspace that is largely orthogonal to the spatial and vibrational states of the molecule. This framework bridges the gap between global thermodynamic organization and local orbital mechanics, revealing the specific chemical substructures that enable frontier orbital decoupling.

2 Methods

2.1 Data preprocessing and stoichiometric residual extraction

To investigate the structural drivers of frontier orbital decoupling, we utilized the QM9 dataset, which contains the computed geometric, energetic, electronic, and thermodynamic properties of small organic molecules. We first deduplicated the dataset by removing 83 redundant SMILES strings, retaining only the first occurrence to prevent data leakage during cross-validation, yielding a final dataset of 133,802 unique molecules. We identified and flagged 9 linear molecules (where the rotational constant $A = 0$) and 45 highly symmetric molecules (where the dipole moment $\mu = 0$) to serve as edge cases for downstream stress-testing. To manage the extreme heavy-tailed distribution of the rotational constant A, we applied a log transformation with a small epsilon. Redundant thermodynamic columns evaluated at 298.15 K were removed, leaving a core set of 12 non-redundant physical properties.

Because macroscopic physical properties in QM9 are heavily dominated by molecular size and stoichiometric composition, we established a baseline to isolate the variance driven purely by structural isomerism and connectivity. Using RDKit, we parsed the SMILES strings and computed exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. We then trained a LightGBM regressor for each of the 12 properties using strictly these five atom counts as features. The stoichiometric residuals were extracted by subtracting the predicted values from the raw observed properties ($res_property = property_{observed} - property_{predicted}$). To maintain its strict physical definition, the energy gap residual was explicitly defined as $res_gap = res_lumo - res_homo$ rather than being predicted independently. All subsequent analyses were performed on this isolated residual space.

2.2 Structural feature engineering and collinearity reduction

To encode the local bonding environments and structural micro-environments of the molecules, we generated a suite of physically grounded 2D RDKit descriptors. These included molecular complexity ('BertzCT'), topological polar surface area (TPSA), and 'MolLogP'. We computed exact functional group counts using RDKit fragments and defined custom structural ratios, including a heteroatom-to-carbon ratio (the sum of N, O, and F atoms divided by C atoms, with a small epsilon added to the denominator) and a hydrogen-to-heavy-atom ratio. Additionally, we extracted the exact ring count and a proxy for conjugation length based on the count of sp2 and sp hybridized carbons.

Because many functional group fragments are highly correlated (e.g., specific carbonyl subtypes), we performed hierarchical clustering on all generated structural features using Spearman correlation as the distance metric. We selected one representative feature from each cluster to form a reduced, non-collinear feature set of 63 descriptors. This collinearity reduction ensures that downstream feature importance rankings are robust and reflect true chemical drivers rather than statistical artifacts.

Target Residual	OOF R^2	Mean Absolute Error (MAE)
'res_lumo'	0.8493	0.00947 Hartree
'res_gap'	0.8157	0.01126 Hartree
'res_homo'	0.7471	0.00737 Hartree
'res_zpve'	0.6228	0.00081 Hartree
'res_r2'	0.5435	97.914 Bohr^2

2.3 Predictive modeling of property residuals

To map the dependency between localized structural environments and the isolated property residuals, we trained LightGBM regressors using the reduced 63-feature set. The models were evaluated using a 5-fold GroupKFold cross-validation strategy, grouping by Murcko scaffolds to ensure that structurally similar molecules were kept within the same fold, thereby testing true structural generalization.

The performance of these models on key target residuals is summarized in Table 2.3. To identify the specific chemical classes or geometric complexities driving the variance in the frontier orbitals, we extracted the feature importances using the LightGBM gain metric. This ranking allowed us to contrast the local functional motifs governing the HOMO and LUMO residuals against the global topological features dictating the spatial and vibrational residuals.

2.4 Stress-testing the local environment hypothesis

We stress-tested the hypothesis that frontier orbitals are governed by local micro-environments by evaluating the model's performance across structural extremes. First, we isolated the flagged linear molecules to verify if their frontier residuals were systematically shifted in the exact direction predicted by the model's top features. Second, we tested the sensitivity of the frontier orbitals to long-range electronic effects by stratifying the dataset according to conjugation length (proxied by the count of sp2/sp carbons). We evaluated the LightGBM model's out-of-fold R^2 across these strata to identify the boundary where the local micro-environment hypothesis breaks down and long-range electronic communication re-couples the orbitals to the global molecular topology.

2.5 Variance asymmetry and structural decoupling analysis

To address the structural observation that the gap residual is primarily driven by the LUMO residual, we computed the variance ratio $Var(res_lumo)/Var(res_homo)$. We binned the dataset by the heteroatom-to-carbon ratio and evaluated this variance ratio across the bins to determine whether the LUMO-dominance of the gap is a static mathematical artifact or a property that undergoes a structural transition as heteroatom density increases.

Furthermore, we continuously mapped the transition between the global thermodynamic regime and the local frontier regime. Within each heteroatom-to-carbon bin, we computed the Distance Correlation ($dCor$) between res_homo and res_lumo, as well as their decoupling from the thermodynamic manifold (e.g., $dCor$ between res_homo and res_zpve). To ensure this transition was a true structural property and not a hidden size effect, we also computed the partial Distance Correlation ($pdCor$) within each bin, strictly controlling for the total heavy atom count.

2.6 Manifold learning and topological projection

Finally, to visually and topologically validate the decoupling of the frontier orbitals from the thermodynamic and spatial manifold, we utilized $1 - dCor$ as a distance metric to generate a 12×12 distance matrix of the property residuals. This matrix was computed on a robust subset of the data, excluding the linear and highly symmetric anomalies. We applied Multi-Dimensional Scaling (MDS) and hierarchical clustering to this distance matrix to project the properties into a 2D space. This manifold learning approach allowed us to visually assess whether the HOMO and LUMO residuals occupy a distinct, isolated region of the latent space compared to the tightly clustered thermodynamic and spatial variables.

3 Results

3.1 Stoichiometric baseline and residual extraction

The physical properties of the 133,802 unique molecules in the QM9 dataset (following the deduplication of 83 SMILES strings) are heavily dominated by molecular size and composition. To isolate the chemical structure's role in organizing these properties, we established a stoichiometric baseline using a five-feature linear regression on the counts of C, N, O, F, and H atoms. This baseline confirms that thermodynamic properties are essentially extensive: as shown in Figure 1, the total internal energy (u_0) and zero-point vibrational energy ($zpve$) exhibit R^2 values of 1.0000 and 0.9971, respectively, from atom counts alone.

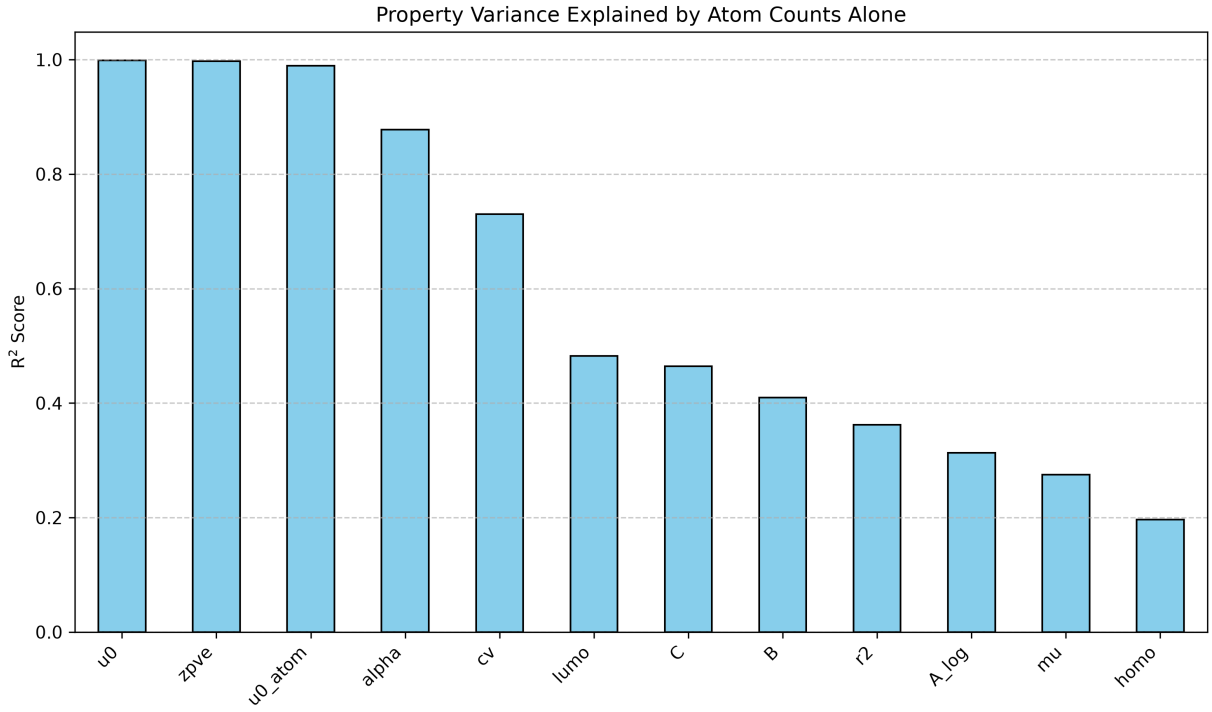

Figure 1: The bar chart displays the R^2 scores for various molecular properties predicted solely from a stoichiometric baseline of C, N, O, F, and H atom counts. Thermodynamic properties, such as the total internal energy (u_0) and zero-point vibrational energy ($zpve$), are essentially extensive and have their variance almost entirely explained by atom counts alone ($R^2 \approx 1.0$). Conversely, frontier orbital energies (LUMO and HOMO) and spatial metrics (such as r^2) exhibit significantly lower predictability from stoichiometry. This demonstrates that while thermodynamic properties are heavily dominated by molecular size and composition, electronic and spatial characteristics are governed by the underlying chemical structure, necessitating the extraction of property residuals to isolate structural organization from simple size effects.

We extracted a 12-property residual manifold by training LightGBM regressors on these five atom counts and calculating the residuals ($res_property = property_{observed} - property_{predicted}$). This procedure removes the dominant size effect, effectively raising the intrinsic dimensionality of the data: a Principal Component Analysis (PCA) on these residuals requires 8 components to explain 90% of the variance, whereas the raw columns require only 7. All subsequent analyses were performed on these residuals to ensure that the discovered structures reflect chemical organization rather than simple stoichiometry.

3.2 Local structural governance of frontier versus thermodynamic residuals

We modeled the relationship between molecular substructure and property residuals using a reduced set of 63 non-collinear RDKit descriptors and functional group counts. Model performance was validated using a

5-fold GroupKFold cross-validation strategy split by Murcko scaffolds to ensure structural generalization. As detailed in Table 2.3, the local structural environment provides a strong explanatory framework for frontier orbital residuals but is less coupled to spatial and vibrational residuals.

Feature importance rankings (Figure 2) reveal a sharp divergence in governance. Frontier orbitals (`res_homo`, `res_lumo`) are primarily driven by specific functional motifs. The `res_homo` is most sensitive to `fr_nitrile` (8.9e+01 gain), `fr_aniline` (4.0e+01), and `fr_NH0` (2.3e+01), indicating that local nitrogen environments are the primary determinants of the highest occupied level. Conversely, `res_lumo` is governed by the count of sp2/sp hybridized carbons (`n_sp2_sp_carbons`, 3.9e+02) and `fr_ketone` (1.5e+02). In contrast, `res_zpve` and the spatial extent `res_r2` are dictated by global topological features such as the number of rings (`n_rings`) and molecular complexity (`BertzCT`).

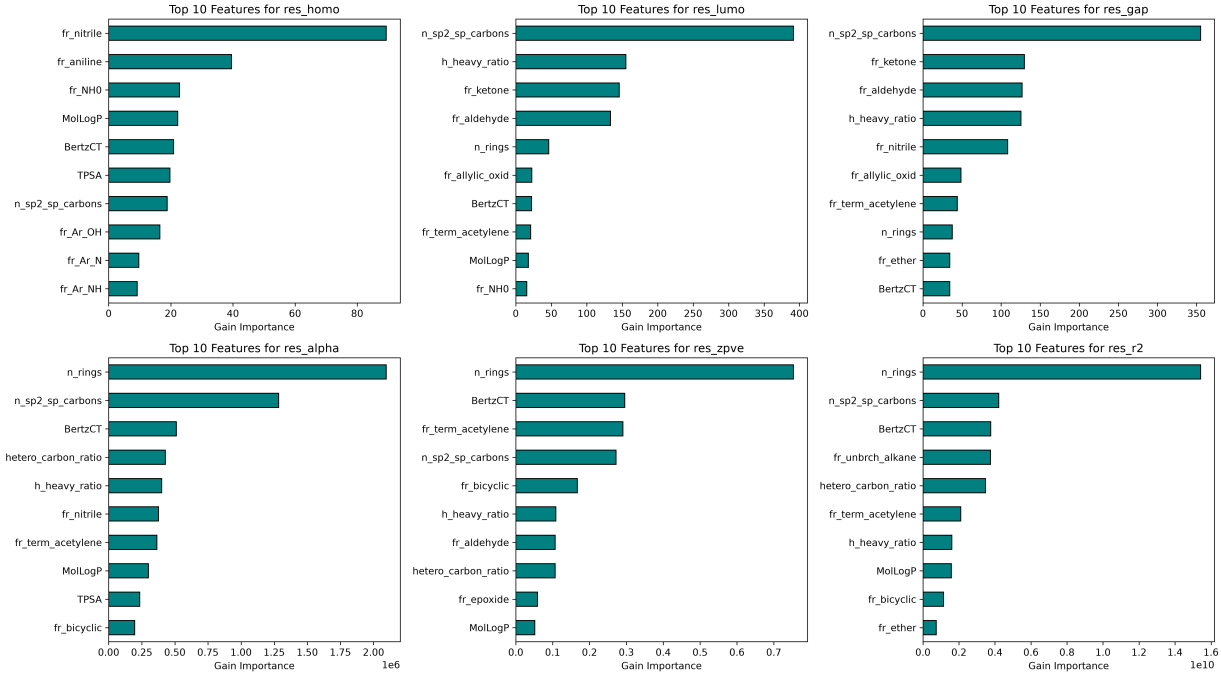

Figure 2: Feature importance rankings (measured by gain) for six molecular property residuals, illustrating a sharp divergence in structural governance. Frontier orbital residuals (`res_homo`, `res_lumo`, and `res_gap`) are primarily driven by specific local functional motifs; `res_homo` is most sensitive to local nitrogen environments like `fr_nitrile` and `fr_aniline`, while `res_lumo` is governed by sp2/sp hybridized carbon counts (`n_sp2_sp_carbons`) and `fr_ketone`. Conversely, thermodynamic and spatial residuals, such as zero-point vibrational energy (`res_zpve`) and spatial extent (`res_r2`), are dictated by global topological features like the number of rings (`n_rings`) and molecular complexity (`BertzCT`). This confirms that local structural environments strongly determine frontier orbitals, whereas spatial and vibrational properties depend on global molecular topology.

3.3 Breakdown of local predictability under conjugation

The "local micro-environment" hypothesis—that frontier properties are governed by immediate bonding—was tested by stratifying the dataset by conjugation length (proxied by the count of sp2/sp carbons). As shown in Figure 3, the predictive power (R^2) for `res_homo` and `res_lumo` remains high in moderately conjugated systems but collapses in the limit of high conjugation.

For molecules with zero sp2/sp carbons ($N = 30,190$), the model achieves an R^2 of 0.4979 for `res_homo` and 0.8193 for `res_lumo`. Performance for `res_homo` peaks at 0.7674 for 6 sp2/sp carbons ($N = 1,577$) but drops to 0.6057 as conjugation increases further to 7 carbons ($N = 268$). The `res_lumo` exhibits a more precipitous decline, falling from 0.8193 (0 sp2 carbons) to 0.5046 (7 sp2 carbons). This confirms that

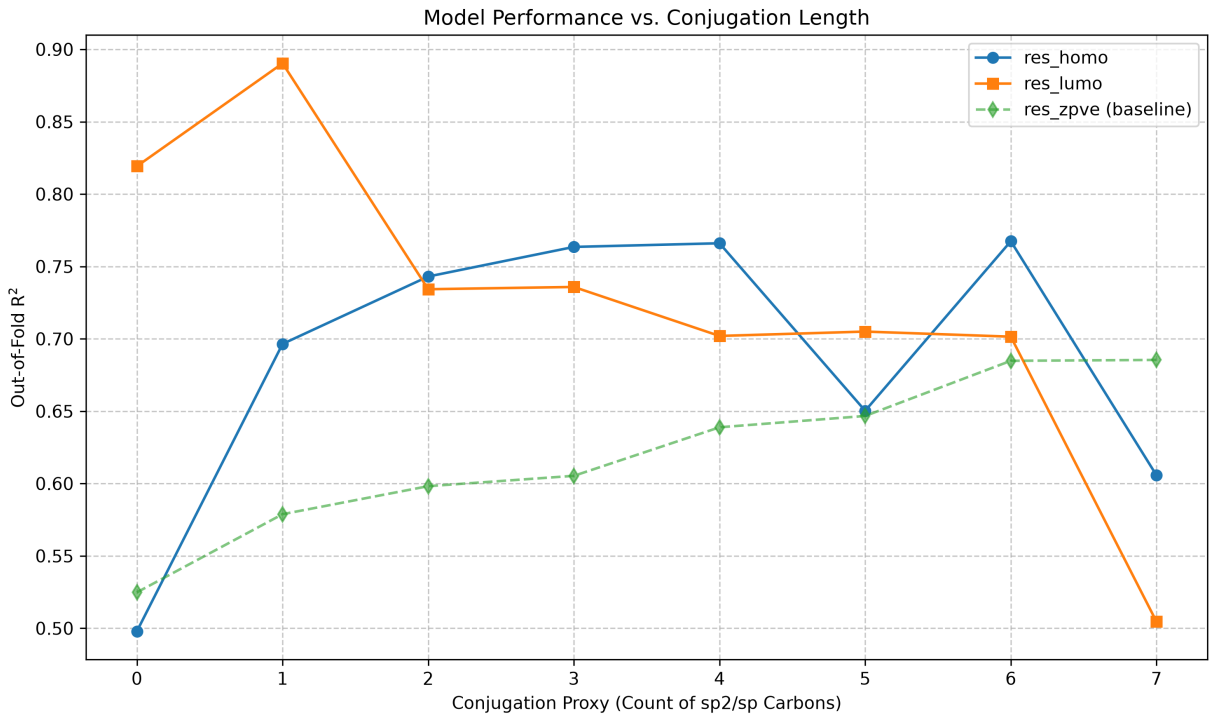

Figure 3: Out-of-fold R^2 model performance for predicting the `res_homo`, `res_lumo`, and `res_zpve` residuals as a function of conjugation length, proxied by the count of sp2/sp carbons. The predictive power for the frontier orbital residuals (`res_homo` and `res_lumo`) remains high in moderately conjugated systems but collapses in the limit of high conjugation (7 sp2/sp carbons), where R^2 drops to 0.6057 and 0.5046, respectively. Conversely, the baseline thermodynamic residual `res_zpve` exhibits a steady increase in predictability across the same range. This breakdown in local predictability reveals that while local micro-environments dominate the frontier orbitals for most molecules, long-range electronic effects emerge in highly conjugated systems, re-coupling these orbitals to the global molecular topology.

while local environments dominate the frontier orbitals of most QM9 molecules, long-range electronic effects emerge in highly conjugated systems, re-coupling the orbitals to the global topology.

3.4 LUMO-HOMO asymmetry and heteroatom decoupling

We investigated the structural decoupling of frontier orbitals from the thermodynamic manifold across gradients of heteroatom density. The variance of `res_lumo` is significantly higher than that of `res_homo`, with a global variance ratio (Var(`res_lumo`)/Var(`res_homo`)) of 2.8956. This ratio is not static; it is a function of heteroatom density (Figure 4). In low-heteroatom regimes (median ratio 0.125), the LUMO variance is 4.0011 times that of the HOMO. This ratio relaxes toward unity (1.4727) as the heteroatom-to-carbon ratio exceeds 0.8.

Using Distance Correlation (*dCor*) and Partial Distance Correlation (*pdCor* controlling for heavy atom count), we mapped the decoupling of properties across 6 heteroatom bins (Figure 5). In hydrocarbon-dominant molecules, the HOMO is almost entirely decoupled from both the LUMO ($pdCor = 0.0284$) and the thermodynamic manifold (`res_zpve`, $pdCor = 0.0286$). As heteroatom density increases, the coupling between the frontier orbitals rises, reaching a *pdCor* of 0.2017 in the highest heteroatom bin (median ratio 1.25). However, the HOMO remains strikingly decoupled from the thermodynamic manifold across the entire dataset, with $pdCor$(`res_homo`, `res_zpve`) never exceeding 0.0813.

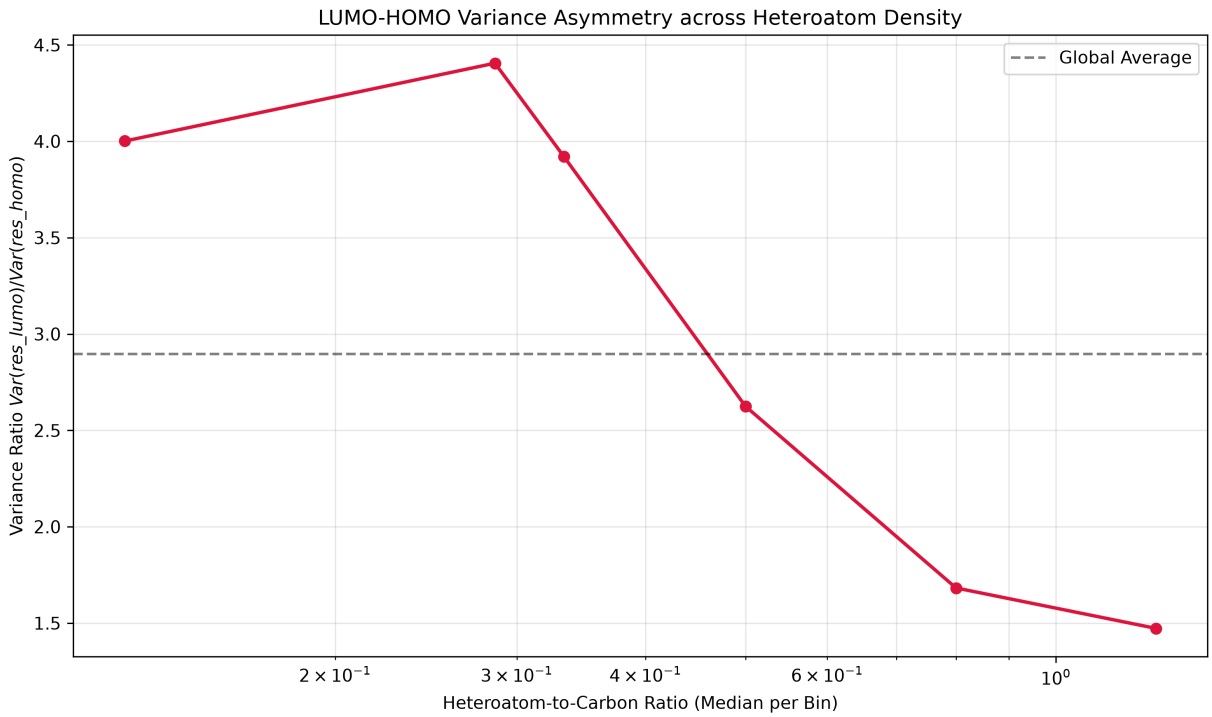

Figure 4: The plot illustrates the variance asymmetry between frontier orbital residuals, measured as the ratio Var(`res_lumo`)/Var(`res_homo`), across varying heteroatom densities. The dashed line indicates the global average variance ratio of 2.8956. In low-heteroatom regimes (median heteroatom-to-carbon ratio of 0.125), the LUMO variance is highly asymmetric, being 4.0011 times greater than that of the HOMO. However, as the heteroatom-to-carbon ratio exceeds 0.8, this ratio relaxes toward unity, dropping to 1.4727. This demonstrates that while LUMO properties generally exhibit significantly higher variance than HOMO properties, increasing heteroatom density systematically reduces this disparity and alters the structural decoupling of the frontier orbitals.

3.5 Non-linear topology of the 12-property manifold

The global organization of the property residuals was visualized using Multi-Dimensional Scaling (MDS) on a distance matrix defined by $1 - dCor$ (Figure 6).

Hierarchical clustering (Figure 7) identifies four distinct modules within the residual space:

1. **Spatial-geometric block:** The rotational constants B and C are tightly bound to the electronic spatial extent r^2. The pair (C, r^2) is the tightest in the dataset (Distance = 0.0848).

2. **Vibrational-thermal block:** `res_zpve` and `res_cv` form a coherent unit (Distance = 0.1899), relatively isolated from the spatial variables.

3. **Frontier orbital block:** `res_homo` and `res_lumo` are significantly isolated from the spatial-thermal manifold. The HOMO residual is particularly distant from the geometric constants A, B, and C (all Distances > 0.78).

4. **Isolated energies:** The total energy u_0 is the most isolated property in the residual manifold (Average Distance = 0.8314), indicating that once atom counts are removed, its remaining variance is largely independent of all other physical responses.

This topological map confirms that the QM9 property set is not a single joint structure but a collection of weakly coupled modules, with the frontier electronic states and total energies existing in largely orthogonal subspaces to the spatial and vibrational state of the molecule.

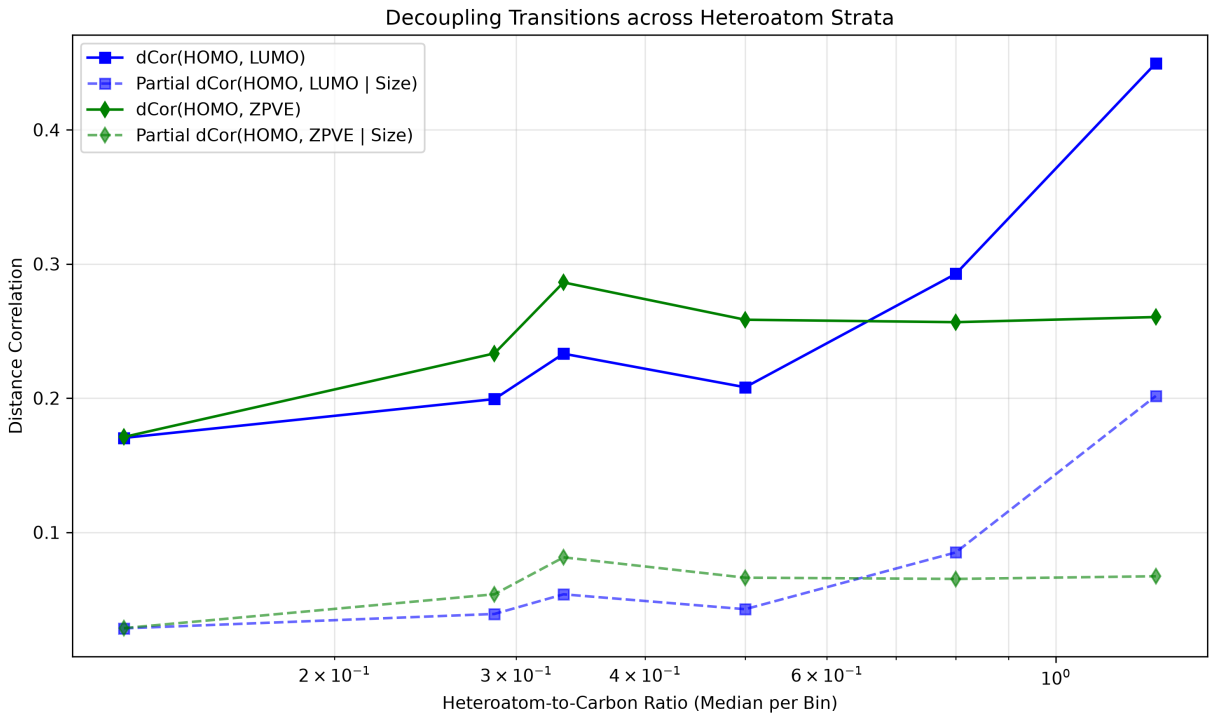

Figure 5: Distance correlation ($dCor$) and partial distance correlation ($pdCor$, controlling for heavy atom count) between the HOMO and both the LUMO and zero-point vibrational energy (ZPVE) across six heteroatom-to-carbon ratio bins. The plot illustrates the structural decoupling of frontier orbitals from the thermodynamic manifold. In hydrocarbon-dominant molecules, `res_homo` is almost entirely decoupled from both `res_lumo` ($pdCor = 0.0284$) and `res_zpve` ($pdCor = 0.0286$). As heteroatom density increases, the coupling between the frontier orbitals rises, reaching a $pdCor$ of 0.2017 in the highest heteroatom bin. However, `res_homo` remains strikingly decoupled from the thermodynamic manifold across the entire dataset, with the partial distance correlation between `res_homo` and `res_zpve` never exceeding 0.0813. This demonstrates that while increasing heteroatom density re-couples the frontier electronic states to each other, the HOMO remains fundamentally orthogonal to the vibrational state of the molecule.

4 Conclusions

In this study, we demonstrated that the anomalous decoupling of frontier orbital energies from the global thermodynamic and spatial manifold is fundamentally rooted in divergent structural governance. By establishing a stoichiometric baseline and extracting property residuals across the QM9 dataset, we successfully isolated the variance driven purely by structural isomerism and connectivity. This approach revealed that the physical state space of small organic molecules is not a single, tightly coupled structure, but rather a collection of weakly interacting modules.

Our predictive modeling confirms that frontier orbital residuals are highly sensitive to local structural micro-environments, whereas spatial and vibrational residuals are governed by macroscopic molecular topology. As summarized in Table 2.3, local functional motifs provide a strong explanatory framework for frontier orbitals, achieving high out-of-fold R^2 values for *res_lumo* and *res_homo*, but yield significantly lower predictive power for global properties like *res_zpve* and *res_r2*. Specifically, the HOMO is dictated by local nitrogen environments, and the LUMO by localized sp2/sp carbon counts and ketones. This contrasts sharply with the global topological features, such as ring counts and overall molecular complexity, that drive spatial extent and zero-point vibrational energy.

Crucially, we identified the boundary conditions where this local governance breaks down. In highly conjugated systems, the emergence of long-range electronic effects overrides local micro-environments, effectively

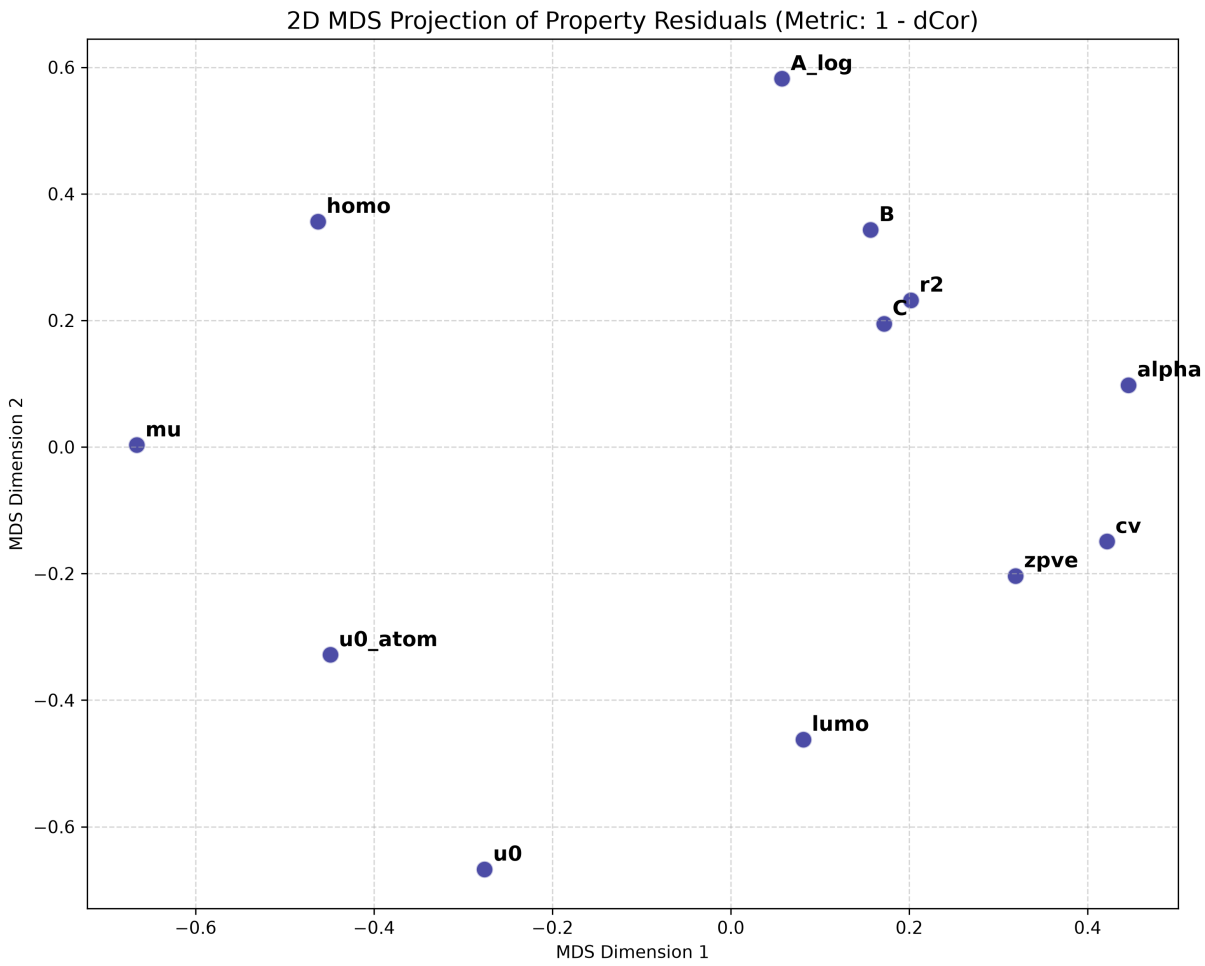

Figure 6: A 2D Multi-Dimensional Scaling (MDS) projection of the 12-property residual manifold based on a distance matrix defined by $1 - dCor$. The visualization reveals the modular organization of the molecular properties into weakly coupled blocks after removing stoichiometric size effects. A tight spatial-geometric cluster is formed by the rotational constants B and C and the electronic spatial extent r^2, while the zero-point vibrational energy and heat capacity form a distinct vibrational-thermal unit. The frontier orbitals are significantly isolated from the spatial-thermal manifold, and the total internal energy u_0 is the most isolated property overall. This topological map demonstrates that the frontier electronic states and total energies exist in largely orthogonal subspaces relative to the spatial and vibrational states of the molecules.

re-coupling the frontier orbitals to the global molecular topology. Furthermore, our analysis of heteroatom density highlighted a dynamic variance asymmetry between the LUMO and HOMO. While increasing heteroatom density reduces this asymmetry and re-couples the frontier orbitals to one another, the HOMO remains strictly orthogonal to the thermodynamic manifold across all structural regimes.

Ultimately, topological mapping of the 12-property residual space definitively shows that frontier electronic states and total internal energies occupy distinct, orthogonal subspaces relative to the spatial-geometric and vibrational-thermal blocks. This modularity provides a robust theoretical foundation for rational molecular design. Because frontier orbitals are governed by specific, localized atomic architectures rather than global topology, they can be tuned independently of a molecule's macroscopic spatial and thermodynamic state. This insight is particularly valuable for the development of organic electronics, photovoltaics, and photocatalysts, where decoupling electronic function from structural and thermodynamic stability is a primary design objective.

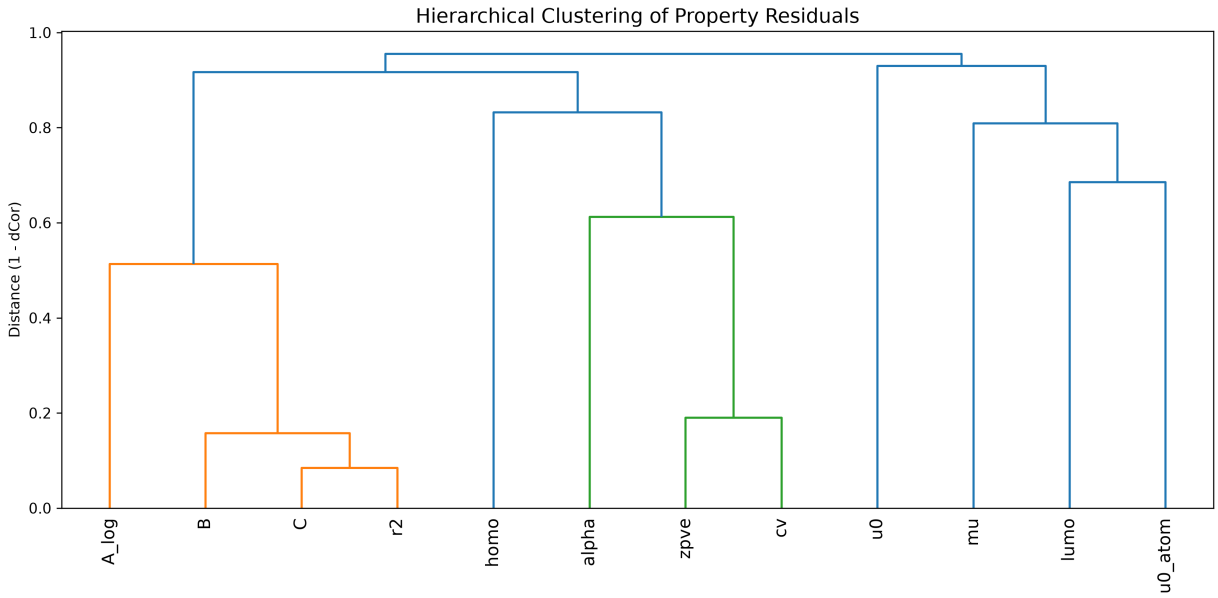

Figure 7: Hierarchical clustering dendrogram of the 12-property residual manifold based on a distance matrix defined by $1 - dCor$. The plot illustrates the modular organization of the dataset into weakly coupled blocks, revealing that the properties exist in largely orthogonal subspaces rather than a single joint structure. A tightly bound spatial-geometric block is visible, containing the dataset's closest pair, C and r^2 (distance of 0.0848), along with B and `A_log`. Adjacent to this is a coherent vibrational-thermal block formed by `res_zpve` and `res_cv`. In contrast, the frontier orbitals (`res_homo` and `res_lumo`) are significantly isolated from the spatial-thermal manifold. Finally, the total internal energy (u_0) is the most isolated property in the residual space, demonstrating that once atom counts are removed, its remaining variance is largely independent of all other physical responses.