Gromov-Wasserstein Partitioning of Frontier-Electronic and Thermodynamic-Spatial Manifolds Across Molecular Architectures

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The extent to which frontier molecular orbitals decouple from global thermodynamic and spatial properties remains an open question, specifically regarding whether this decoupling is a universal constant or modulated by molecular architecture. To address this, we investigate the joint property structure of the QM9 dataset using a manifold partitioning approach. After extracting non-linear atom-count residuals to eliminate stoichiometric determinism, we construct property-dependency graphs based on distance correlation and partition the features into localized frontier-electronic and macroscopic thermodynamic-spatial sub-graphs. By applying Gromov-Wasserstein optimal transport, we compute the structural alignment cost between these sub-manifolds across distinct topological families and heteroatom densities. Our results indicate that the electronic spatial extent and rotational constant serve as the primary non-linear hubs organizing the global property manifold. Furthermore, we observe that while acyclic molecules exhibit tight manifold integration, the introduction of cyclic constraints significantly fragments the relationship between the electronic frontier and macroscopic spatial properties. In contrast, increased heteroatom density acts to bridge these manifolds, enhancing their overall alignment. Finally, we find that the highest occupied molecular orbital remains distinctly isolated across all topological families—driven primarily by local bonding environments rather than global architecture—and scaffold-based cross-validation confirms these architectural rewiring phenomena are robust, fundamental characteristics of the chemical space.

1 Introduction

The mapping of molecular architecture to physical and chemical properties is a central pillar of chemical physics and materials discovery. While individual properties—ranging from thermodynamic stability to frontier orbital energies—are routinely predicted with high accuracy, the underlying joint structure of these properties remains less explored. A critical open question is the extent to which localized electronic states, such as frontier molecular orbitals, decouple from global macroscopic features like spatial extent and thermodynamic capacity.

1.1 Motivation and background

It is well established that frontier orbitals dictate chemical reactivity and optical behavior, often depending heavily on local bonding environments. Conversely, thermodynamic and spatial properties scale largely with molecular size and global topology. However, it is unclear whether the decoupling between these two regimes is a universal constant across chemical space or a structurally dependent phenomenon modulated by specific architectural features, such as ring fusion or heteroatom density. Standard linear correlation analyses often fail to capture these complex dependencies, and raw property values are heavily confounded by stoichiometric determinism—the trivial scaling of properties with the number of atoms.

1.2 Manifold partitioning and optimal transport

To address this, we present a manifold partitioning analysis of the QM9 dataset to investigate the joint property structure across diverse molecular architectures. We first extract non-linear atom-count residuals

Property	Non-linear Hub Score ($dCor$)	Linear Hub Score (Partial Corr sum)
res_r2	0.8440	1.7159
res_C	0.8180	1.6095
res_B	0.6862	1.2764
res_alpha	0.6140	1.8305
res_cv	0.6050	2.2347
res_zpve	0.5626	2.4623

to rigorously eliminate stoichiometric determinism. We then construct robust property-dependency graphs using distance correlation, which captures both linear and non-linear associations. By partitioning the molecular features into a localized frontier-electronic sub-graph (comprising the highest occupied and lowest unoccupied molecular orbitals) and a macroscopic thermodynamic-spatial sub-graph, we mathematically quantify their structural alignment. We employ Gromov-Wasserstein optimal transport to compute the exact alignment cost between these sub-manifolds across distinct topological families and compositional gradients.

1.3 Overview of findings

Our analysis reveals that the global property manifold is not uniformly connected but organized around specific topographic hubs. By evaluating the distance correlation ($dCor$) between each property and the principal component of the remaining features, we identify the electronic spatial extent (res_r2) and the rotational constant C (res_C) as the primary non-linear hubs organizing the global property manifold (Table 1.3).

Furthermore, we demonstrate that molecular architecture fundamentally rewires the connection between the electronic and macroscopic manifolds. While acyclic molecules exhibit tight manifold integration, the introduction of cyclic constraints significantly fragments this relationship. Conversely, increased heteroatom density acts to bridge these sub-manifolds, enhancing their overall alignment. Finally, we show that the highest occupied molecular orbital remains distinctly isolated across all topological families, driven by local bonding rather than global architecture. Scaffold-based cross-validation confirms that these architectural rewiring phenomena are robust, fundamental characteristics of the chemical space.

2 Methods

2.1 Data initialization and preprocessing

To investigate the joint property structure across diverse molecular architectures, we utilized the QM9 dataset. Initial preprocessing involved deduplicating 83 SMILES strings that appeared multiple times to ensure a strictly non-redundant dataset. We isolated a core set of physical and chemical properties by discarding redundant thermodynamic identities (specifically, the atomization and standard state energies at 298 K) and the rank-deficient HOMO-LUMO gap, which is a linear combination of the individual frontier orbital energies. Furthermore, the rotational constant A was excluded entirely to prevent extreme heavy-tail distortions from biasing the global manifold topology, leaving rotational constants B and C to represent the rotational manifold.

Using RDKit, we parsed the SMILES strings to extract exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen. To facilitate downstream analysis of local bonding effects on frontier orbitals, we additionally computed structural descriptors for each molecule, including the maximum absolute partial charge and the number of aromatic rings.

2.2 Cross-validated non-linear residual extraction

Raw molecular properties are heavily confounded by stoichiometric determinism—the trivial scaling of extensive properties with molecular size. To rigorously isolate structural and architectural effects from this

baseline size scaling, we performed a cross-validated residual extraction. For each of the 11 retained properties, we trained a Gradient Boosting Regressor using K-fold out-of-sample prediction, providing only the five exact atom counts as independent predictors.

The out-of-sample residuals were extracted, and their orthogonality to the original atom counts was verified to ensure the size effect was fully eliminated. To robustify the continuous manifold against the dataset's latent geometric anomalies, we applied a quantile transformation independently to each property's residuals. This was followed by standardizing (z-scoring) the transformed residuals, ensuring all properties were on a comparable scale for subsequent distance-based topological metrics.

2.3 Topological and compositional stratification

To determine whether the rewiring of the joint property structure is a threshold effect driven by topology or a gradual drift driven by composition, we partitioned the dataset along two distinct axes. First, we defined discrete topological families based on structural connectivity: strictly acyclic molecules, isolated ring systems, and fused polycyclic architectures. Second, we established a continuous compositional metric defined by the ratio of heteroatoms (N, O, F) to carbon atoms. The dataset was subsequently binned into quantiles based on this heteroatom-to-carbon ratio, enabling a comparative analysis of manifold alignment across compositional gradients.

2.4 Manifold partitioning and optimal transport

We constructed robust property-dependency graphs for each topological family and compositional quantile using Distance Correlation ($dCor$), a metric capable of capturing both linear and non-linear associations between the standardized residuals. The property nodes were partitioned into two distinct sub-graphs: a localized frontier-electronic sub-graph (comprising res_homo and res_lumo) and a macroscopic thermodynamic-spatial sub-graph (comprising res_zpve, res_cv, res_r2, res_C, and other macroscopic features).

To mathematically quantify the structural alignment between these two sub-manifolds, we employed Gromov-Wasserstein (GW) optimal transport. The GW distance computes the minimum cost required to align the intra-manifold distance matrices of the electronic and macroscopic sub-graphs. Shifts in this transport cost across the stratified subsets provided a direct measure of where molecular architecture forces the isolated electronic state space to integrate with, or fragment from, the macroscopic thermodynamic manifold. Additionally, structural rewiring was quantified by computing the Frobenius norm of the difference between the $dCor$ matrices of distinct families, alongside the RV coefficient and the Mantel test statistic to formally test for significant differences in correlation structures.

2.5 Hub property identification and frontier isolation

To identify the central organizing nodes of the global property manifold, we computed both linear and non-linear hub scores. Linear hubs were defined quantitatively by computing the partial correlation matrix of the 11 standardized residuals and summing the absolute partial correlations for each property. To account for non-linear dependencies, we established a non-linear hub metric by calculating the $dCor$ between each individual property and the first principal component of the remaining ten properties.

Concurrently, we explicitly tested the isolation of the highest occupied molecular orbital (res_homo). We computed the correlation between the HOMO residuals and local bonding descriptors (such as the number of aromatic rings and maximum absolute partial charge) to determine if its decoupling from the macroscopic manifold is driven by sensitivity to the local bonding environment rather than global topology.

2.6 Scaffold-based bootstrap validation

To ensure that the observed differences in correlation structures and manifold alignments represent genuine chemical phenomena rather than sampling artifacts, we implemented a rigorous validation framework. We utilized bootstrap resampling with equal subset sizes to compute confidence intervals for the Frobenius norms of the matrix differences.

Furthermore, we executed a Murcko scaffold-based split, holding out 20% of the dataset comprising entire molecular families unseen during the initial manifold construction. We recomputed the non-linear

residuals, partial correlations, and $dCor$ matrices exclusively on the training set, and subsequently evaluated the stability of the topological alignments, GW transport costs, and rewiring metrics on the held-out scaffold set.

3 Results

Analysis of the joint structure of the QM9 dataset reveals that molecular properties are organized into coherent manifolds whose internal alignment and coupling are fundamentally modified by molecular architecture and heteroatom density. By conditioning properties on atom-count baselines as detailed in the Methods section, we move beyond stoichiometric determinism to expose a joint structure defined by topological hubs and the distinct isolation of frontier electronic states.

3.1 Topographic hubs and manifold centrality

The property manifold is not a flat collection of correlations but a structured network with clear "hub" properties that anchor other variables. Using distance correlation ($dCor$) to measure non-linear centrality—defined as the $dCor$ between a property and the first principal component of all other properties—we identify the electronic spatial extent (res_r2) and the rotational constant C (res_C) as the primary structural hubs ().

Property	Non-linear Hub Score ($dCor$)	Linear Hub Score (Partial Corr sum)
res_r2	0.8440	1.7159
res_C	0.8180	1.6095
res_B	0.6862	1.2764
res_alpha	0.6140	1.8305
res_cv	0.6050	2.2347
res_zpve	0.5626	2.4623

Table 1: Comparison of non-linear and linear hub scores for selected molecular properties.

The dominance of res_r2 and res_C suggests that the global spatial distribution of electron density is the central organizing principle for the thermodynamic and geometric response properties. While the zero-point vibrational energy (res_zpve) and heat capacity (res_cv) show high linear partial correlations, indicating tight coupling within the thermodynamic block, their lower non-linear hub scores suggest they are more peripheral to the global manifold compared to spatial-geometric variables. This contrast between linear and non-linear coupling, as well as the architectural rewiring of these hubs, is further illustrated in .

3.2 Architectural rewiring via topology and heteroatom density

Molecular architecture significantly modifies the alignment between the frontier-electronic manifold (comprising res_homo and res_lumo) and the macroscopic-spatial manifold (comprising res_zpve, res_cv, res_r2, res_C, res_alpha, and res_u0). This structural rewiring was quantified using the Gromov-Wasserstein (GW) distance, which measures the optimal transport cost required to align the intra-manifold distance matrices of these two sub-graphs ().

3.2.1 Topological modulation

Acyclic molecules exhibit the tightest manifold integration, with a GW alignment cost of 0.0919. This cost increases sharply in cyclic systems, reaching 0.2136 for single-ring molecules and 0.1999 for polycyclics. This suggests that the introduction of cyclic constraints fragments the relationship between a molecule's electronic frontier and its global physical extent. Interestingly, linear methods such as Principal Component Analysis (PCA) partially mask this effect. The variance explained by the first principal component (PC1) is nearly identical for acyclic (36.96%) and polycyclic (36.68%) families, yet the GW cost reveals that the underlying manifold structure is twice as fragmented in the latter.

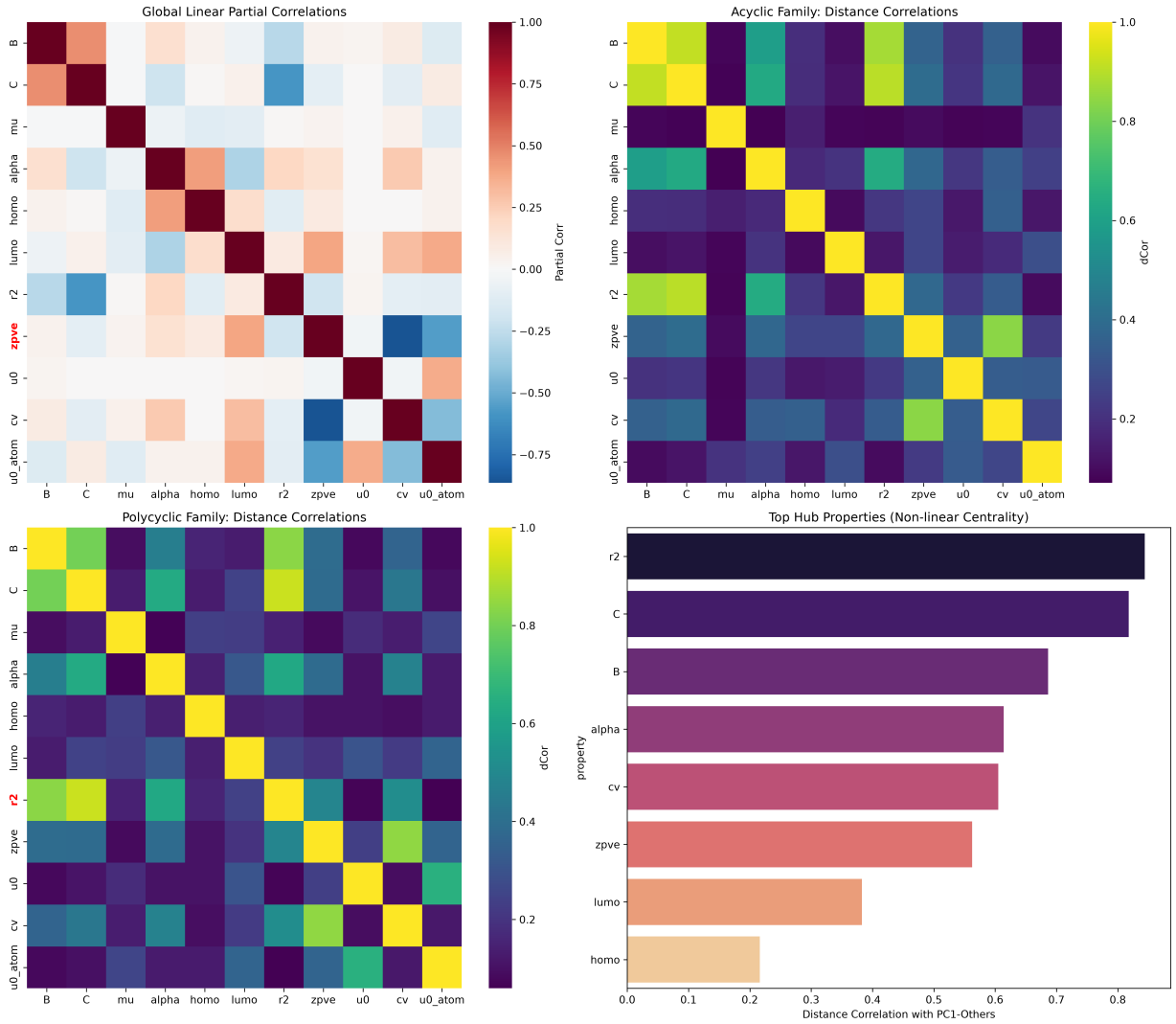

Figure 1: Comparison of linear partial correlations (top-left) and non-linear distance correlations ($dCor$) across acyclic (top-right) and polycyclic (bottom-left) molecular families, alongside non-linear hub scores (bottom-right). While global linear partial correlations indicate tight coupling within the thermodynamic block (e.g., res_zpve and res_cv), the non-linear centrality metric identifies spatial-geometric properties—namely the electronic spatial extent (res_r2) and rotational constant (res_C)—as the primary topographic hubs anchoring the property manifold. The contrast between acyclic and polycyclic $dCor$ matrices illustrates architectural rewiring, revealing that the introduction of cyclic constraints fragments the integration between frontier-electronic and macroscopic-spatial sub-manifolds. Finally, the consistently low non-linear hub score of the highest occupied molecular orbital (res_homo) highlights its distinct isolation from the joint manifold, demonstrating that it functions as a local property decoupled from global spatial and thermodynamic responses across all topological families.

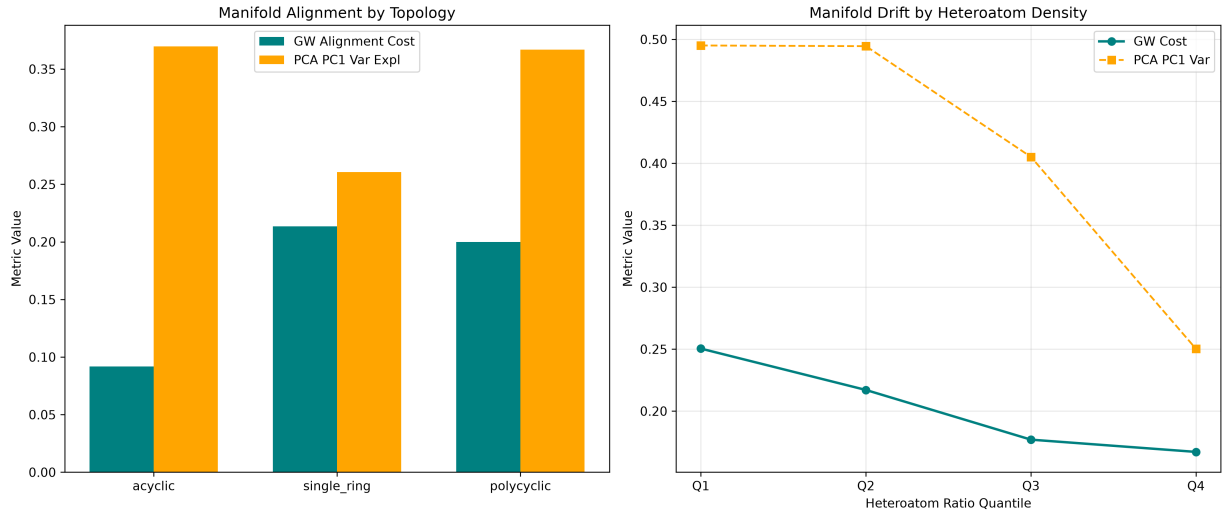

Figure 2: Comparison of manifold alignment between frontier-electronic and macroscopic-spatial properties across different molecular topologies (left) and heteroatom density quantiles (right). The left panel demonstrates that introducing cyclic constraints (single-ring and polycyclic structures) significantly increases the Gromov-Wasserstein (GW) alignment cost compared to acyclic molecules, indicating a fragmentation of the property manifolds. Notably, linear methods like Principal Component Analysis (PCA) mask this topological fragmentation, as the variance explained by the first principal component (PC1) remains nearly identical for acyclic and polycyclic families (approximately 37%). The right panel illustrates that increasing the density of heteroatoms from the lowest (Q1) to the highest (Q4) quantile decreases the GW cost from 0.2504 to 0.1671. This reveals that heteroatoms help bridge and align the electronic and macroscopic manifolds. Together, these metrics show that molecular architecture and heteroatom density fundamentally rewire the joint property structure, an effect captured by non-linear GW optimal transport but obscured by linear PCA.

3.2.2 Heteroatom influence

The density of heteroatoms (N, O, F) acts as a second axis of rewiring. Molecules in the lowest heteroatom-ratio quantile (Q1) show the highest manifold fragmentation (GW cost = 0.2504), despite having high PC1 variance (49.51%). As heteroatom density increases, the manifolds become more aligned, with the GW cost dropping to 0.1671 in the highest quantile (Q4). This indicates that nitrogen, oxygen, and fluorine atoms bridge the electronic and macroscopic manifolds, likely by introducing more uniform bonding patterns that couple local orbital energies to global spatial responses.

3.3 Isolation of the highest occupied molecular orbital

The energy of the highest occupied molecular orbital (*res_homo*) stands apart from the joint property structure. Its non-linear hub score is a mere 0.2162, significantly lower than other response properties such as polarizability (*res_alpha*, 0.6140). Analysis of structural descriptors reveals that the isolation of *res_homo* is driven by its sensitivity to local bonding rather than global architecture ().

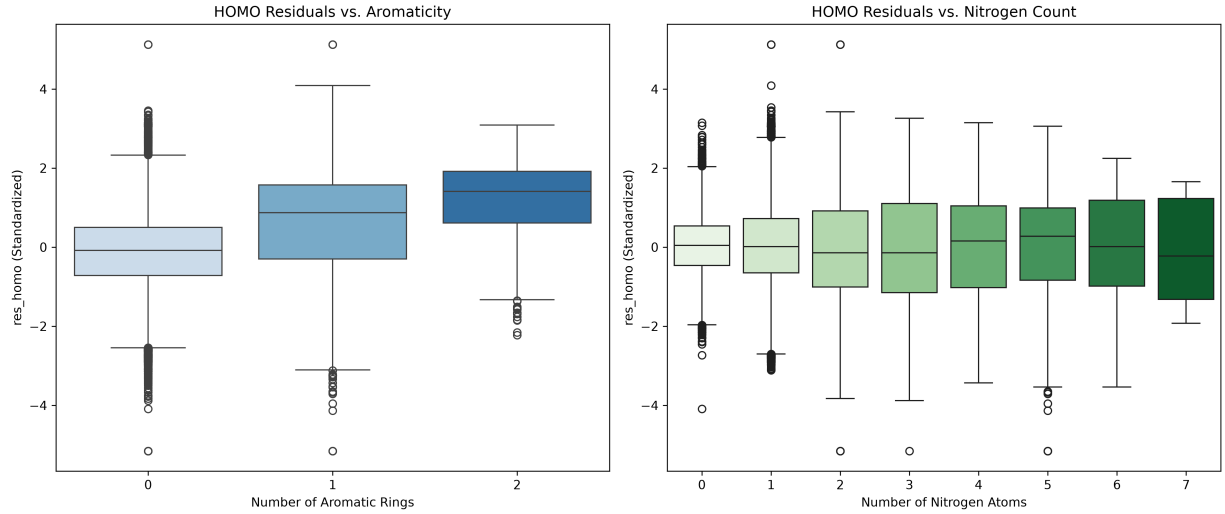

Figure 3: Boxplots displaying the distribution of standardized highest occupied molecular orbital residuals (*res_homo*) against the number of aromatic rings (left) and nitrogen atoms (right). These distributions illustrate that the HOMO level is driven by sensitivity to local bonding rather than global molecular architecture. While increasing aromaticity shifts the mean *res_homo* level, the relatively weak correlation (Pearson $r = 0.2940$) and broad variance across topological families and nitrogen counts highlight its decoupling from the global property manifold. This confirms that the HOMO energy acts as an isolated, local property that resists the joint organization governing the spatial and thermodynamic variables in the QM9 dataset.

The correlation between *res_homo* and the number of aromatic rings is relatively weak (Pearson $r = 0.2940$). While aromaticity shifts the mean HOMO level, it does not reintegrate it into the global manifold. Instead, *res_homo* remains decoupled from the spatial extent and thermodynamic variables across all topological families. This isolation identifies the HOMO level as a local property that resists the joint organization governing the rest of the QM9 property set.

3.4 Scaffold-based stability and validation

The observed joint structures are robust features of the molecular property space rather than artifacts of the specific molecules sampled. Validation using a Murcko scaffold-based holdout (80% train / 20% test) demonstrates high stability across structurally distinct families ().

The correlation between the distance correlation structures of the training and held-out scaffold sets is 0.9876. Furthermore, the alignment cost between sub-manifolds showed a variation of approximately 29.8%

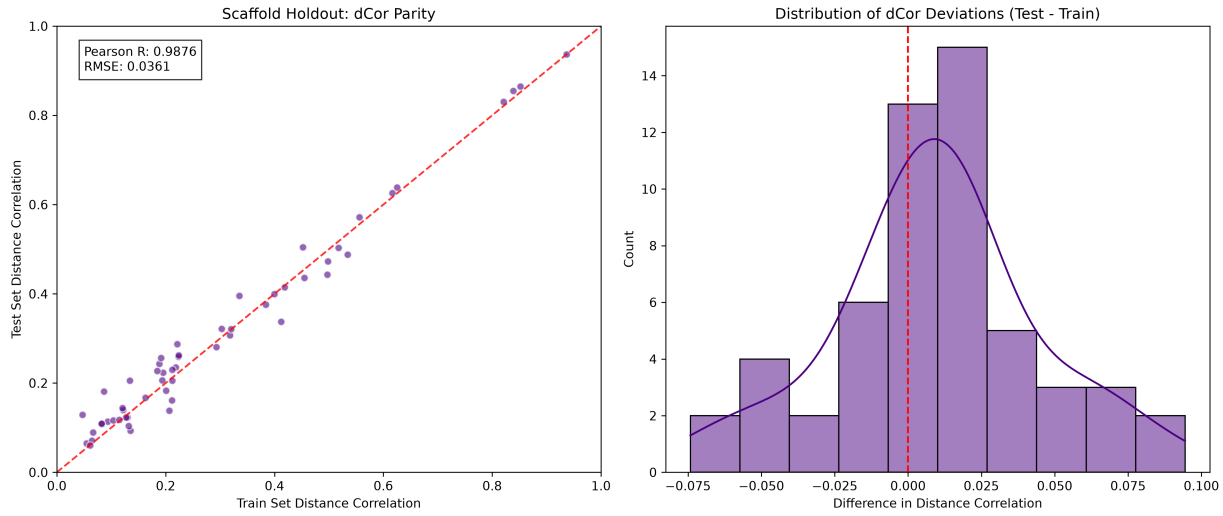

Figure 4: Validation of the distance correlation ($dCor$) matrix stability using an 80% train and 20% test Murcko scaffold-based holdout. The left panel illustrates the parity between training and test set $dCor$ values, revealing a strong linear relationship with a Pearson correlation of 0.9876. The right panel displays the distribution of $dCor$ deviations between the test and train sets, which is tightly centered near zero with a maximum absolute difference of 0.0944. These results demonstrate that the observed joint manifold structures are highly robust across structurally distinct molecular families, confirming that the spatial-geometric hub organization and architectural rewiring are fundamental characteristics of the QM9 chemical space rather than sampling artifacts.

(Train: 0.1377 vs Test: 0.1788), and the maximum absolute difference in any pairwise $dCor$ between train and test sets was tightly bounded at 0.0944. This level of stability confirms that the described manifold organization—including the spatial-geometric hub structure and the architectural rewiring—is a fundamental characteristic of the QM9 chemical space.

4 Conclusions

In this study, we investigated the joint property structure of the QM9 dataset to determine whether the decoupling of frontier molecular orbitals from global macroscopic features is a universal constant or a structurally dependent phenomenon. By extracting cross-validated, non-linear atom-count residuals, we successfully eliminated the confounding effects of stoichiometric determinism. This allowed us to construct robust property-dependency graphs and employ Gromov-Wasserstein optimal transport to quantify the structural alignment between localized frontier-electronic and macroscopic thermodynamic-spatial sub-manifolds.

Our manifold partitioning analysis revealed that the global molecular property space is not uniformly connected but is instead anchored by specific spatial-geometric hubs. The electronic spatial extent (res_r2) and the rotational constant C (res_C) emerged as the primary non-linear organizers of the global manifold. In contrast, thermodynamic properties such as zero-point vibrational energy and heat capacity, despite exhibiting strong linear intra-correlations, remain peripheral to this overarching non-linear architecture.

Crucially, we demonstrated that molecular architecture fundamentally rewires the integration of these property manifolds. While acyclic molecules display a tightly integrated joint property structure, the introduction of cyclic constraints significantly fragments the relationship between the electronic frontier and macroscopic spatial properties. This topological fragmentation is a non-linear phenomenon that is largely obscured by standard linear dimensionality reduction techniques such as Principal Component Analysis. Furthermore, we identified compositional gradients as a secondary axis of rewiring; increasing the density of heteroatoms acts as a structural bridge that realigns the electronic and macroscopic sub-manifolds, likely by coupling local orbital energies to global spatial responses through more uniform bonding patterns.

Amidst this architectural rewiring, the highest occupied molecular orbital (*res_homo*) stands out for its distinct and persistent isolation. Driven primarily by sensitivity to local bonding environments rather than global topology, the HOMO energy remains decoupled from the spatial and thermodynamic variables across all structural families. It functions as a strictly local property that resists the joint organization governing the rest of the dataset.

Finally, rigorous scaffold-based cross-validation confirmed that the observed spatial-geometric hub organization, architectural rewiring, and frontier isolation are highly stable. These phenomena are not sampling artifacts but fundamental characteristics of the chemical space. By mapping exactly where and how molecular architecture forces electronic states to integrate with or fragment from macroscopic properties, this work provides a more nuanced understanding of molecular property spaces, offering valuable structural context for future predictive modeling and rational molecular design.