Topological Reorganisation of the Electronic-Thermodynamic Manifold in QM9

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In molecular property spaces, frontier electronic states frequently decouple from global thermodynamic and spatial features, resulting in a bifurcated state space that complicates joint property modeling. To quantify this physical bifurcation, we analyze the QM9 dataset by extracting atom-count residuals to isolate stoichiometric effects, partitioning the molecular state space into an electronic sub-manifold and a thermodynamic-spatial hub. By representing molecular families as distance correlation graphs, we employ Gromov-Wasserstein optimal transport to compute the topological distance separating these sub-manifolds across distinct chemical architectures. Our results demonstrate that acyclic linear chains possess the highest topological mismatch between electronic and spatial organization, whereas rigid cyclic scaffolds and branched structures enforce tighter geometric reintegration. Additionally, leave-one-property-out ablation reveals that this manifold coupling is a collective effect of the hub's dimensionality rather than being driven by a single dominant property. Finally, machine learning classification highlights specific architectural constraints—most notably ring count and dense heterocyclic environments—as the primary chemical drivers that force isolated frontier orbitals to align with a molecule's global spatial and thermal extent.

1 Introduction

In the computational exploration of chemical space, molecular properties are frequently modeled under the implicit assumption of a unified, continuous manifold. However, the physical reality of molecular state spaces is often more complex. Frontier electronic states, such as the energies of the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO), frequently decouple from global thermodynamic and spatial features. This physical bifurcation results in a fragmented state space that complicates joint property modeling and obscures the underlying physical rules governing molecular behavior. Understanding the topological relationship between these divergent property sets is essential for developing more robust, physics-aware representations of chemical space.

To rigorously quantify this bifurcation, we analyze the QM9 dataset, a benchmark collection of small organic molecules. A primary challenge in mapping intrinsic property manifolds is the overwhelming confounding effect of molecular size and stoichiometry. To isolate the underlying physical topology, we extract exact atom-count residuals for non-redundant properties, applying a $\log(A+1)$ transformation to stabilize extreme rotational variance. This stoichiometric baseline extraction allows us to partition the molecular state space into two distinct domains: a bipolar electronic sub-manifold and a thermodynamic-spatial hub comprising zero-point vibrational energy, heat capacity, electronic spatial extent, and rotational constants.

With the state space partitioned, we represent distinct molecular families—defined by their structural scaffolds—as distance correlation graphs. To measure the structural divergence between the electronic and thermodynamic-spatial domains, we employ Gromov-Wasserstein optimal transport. Unlike standard metric comparisons, the Gromov-Wasserstein distance is uniquely suited for comparing metric measure spaces of different dimensions. This mathematical framework allows us to compute the topological cost of aligning the electronic sub-manifold with the thermodynamic-spatial hub across distinct chemical architectures.

In this work, we systematically track the expansion and collapse of this intra-manifold transport cost across diverse structural populations, ranging from acyclic chains to complex polycyclic systems. We demonstrate that acyclic linear chains exhibit the highest topological mismatch between electronic and spatial or-

Molecular Family	GW Distance (Electronic-Hub)	Transport Complexity
Acyclic Linear	1.1985	High
Acyclic Branched	0.9899	Moderate
Other Cyclic	0.9055	Low
Cyclopropane (C_1CC_1)	0.8951	Low
Aziridine (C_1CN_1)	1.0148	Moderate

ganization, whereas rigid cyclic scaffolds and branched structures enforce a tighter geometric reintegration. Through leave-one-property-out ablation, we further reveal that this manifold coupling is a collective effect of the hub's dimensionality rather than being dominated by any single property. Finally, by applying machine learning classification to the transport costs, we identify the primary chemical drivers—most notably ring count and dense heterocyclic environments—that force isolated frontier orbitals to align with a molecule's global spatial and thermal extent.

2 Methods

2.1 Data preparation and stoichiometric baseline extraction

To isolate the intrinsic topological features of the molecular state space from the overwhelming confounding effects of molecular size and stoichiometry, we first established a rigorous baseline using the QM9 dataset. Following the deduplication of 83 redundant SMILES strings, we extracted 12 non-redundant molecular properties. For properties exhibiting extreme rotational variance, a stabilizing transformation of $y' = \log(y+1)$ was applied. We then performed Ordinary Least Squares (OLS) regression against exact atom counts (C, N, O, F, H) to regress out macroscopic size dependencies. The resulting residuals (e.g., res_homo, res_zpve) represent the size-independent, intrinsic variations of the molecular properties, allowing us to partition the state space into a bipolar electronic sub-manifold (res_homo, res_lumo) and a thermodynamic-spatial hub (res_zpve, res_cv, res_r^2, res_C). The rotational constant C was selected as a single representative for the highly correlated B/C rotational pair to prevent overweighting its contribution to the hub.

2.2 Architectural partitioning and variance normalization

To evaluate how structural constraints influence manifold coupling, the dataset was partitioned into broad architectural families using Murcko Scaffolds generated via RDKit. To resolve topological instability within open-chain systems, the acyclic population was further sub-partitioned into high-branching and linear-chain sub-groups based on the fraction of sp^3 hybridized carbons and the Wiener index. Prior to any transport computation, the property residuals were standardized by their respective standard deviations strictly within each structural family. This intra-family variance normalization ensures that subsequent distance metrics capture the true geometric shape of the manifolds rather than being skewed by residual noise scales.

2.3 Manifold representation and Gromov-Wasserstein transport

We represented the state space of each molecular family as a distance correlation ($dCor$) graph, explicitly capturing the non-linear dependencies between the electronic and thermodynamic-spatial domains. To quantify the topological divergence between these two sub-manifolds, we employed Gromov-Wasserstein (GW) optimal transport using the POT library. The GW distance is uniquely suited for this task as it evaluates the relational cost of aligning metric measure spaces of differing dimensions.

The ground metric was defined consistently using Euclidean distance. To ensure computational tractability while preserving topological accuracy, the POT computations were parallelized across 14 CPUs, employing a random sampling strategy of up to 5,000 molecules per structural family. This optimal transport framework allowed us to compute the topological mismatch and transport complexity across distinct chemical architectures, as summarized in Table 2.3.

2.4 Leave-one-property-out ablation and geometric consistency

To determine whether the manifold coupling is driven by a single dominant property or is a collective effect of the hub's dimensionality, we performed a leave-one-property-out (LOPO) ablation analysis. We systematically removed individual properties from the thermodynamic-spatial hub (res_r^2, res_C, res_zpve, res_cv) and recomputed the GW distances. A collapse in the GW distance upon the exclusion of a specific property (such as res_r^2) would classify it as a primary manifold driver rather than a passive member of the hub.

Additionally, to address potential artifacts from 3,054 geometrically inconsistent molecules lacking explicit labels in the dataset, we isolated a high-consistency subset. This subset was defined by the lowest 10% of total energy OLS residuals, explicitly excluding extreme outliers in rotational constants. After verifying that the heavy-atom count distribution of this subset matched the full population to avoid size bias, we recomputed the GW distances to confirm that the macroscopic topological trends are invariant to underlying geometric noise.

2.5 Machine learning classification of manifold reintegration

Finally, to identify the chemical origins of the electronic-thermodynamic bifurcation, we isolated the specific molecular sub-populations where the electronic sub-manifold geometrically reintegrates with the hub. We defined this "collapse" cohort strictly as molecules falling into the bottom 10th percentile of point-wise GW distances.

Using RDKit, we generated Morgan fingerprints and extracted global structural descriptors, including ring count, VSA_EState5, AvgIpc, and maximum partial charge. A LightGBM classifier was trained to distinguish the high-reintegration cohort from the decoupled population. By analyzing the feature importances and substructural enrichment of specific Morgan bits, we mapped the local chemical environments—such as dense heterocyclic motifs and specific branching patterns—that force isolated frontier orbitals to align with the global spatial organization of the molecule.

3 Results

3.1 The electronic-thermodynamic bifurcation

The joint structure of molecular properties in the QM9 dataset is characterised by a profound bifurcation between frontier-electronic states and the global thermodynamic-response hub. By isolating stoichiometric effects through five-feature atom-count residuals (C, N, O, F, H), we establish a foundational state space that removes macroscopic size dependencies. The varied statistical dispersion of these residuals defines the intrinsic property manifolds (Figure 1).

Standardizing these residuals into Z-scores strictly within each structural family accounts for architectural variations, ensuring that subsequent optimal transport computations measure true manifold shape rather than residual noise scale (Figure 2).

The global organisation of these molecules is partitioned into an electronic sub-manifold (res_homo, res_lumo) and a thermodynamic-spatial hub (res_zpve, res_cv, res_r^2, res_C). Analysis of distance correlations ($dCor$) confirms a persistent non-linear decoupling between these groups across all molecular families (Figure 3). The Electronic-Hub $dCor$ remains consistently low, ranging from 0.1068 in heterocyclic aziridine-like systems (`C1CN1`) to 0.2186 in complex cyclic systems (`Other_Cyclic`).

This decoupling is a defining feature of the QM9 state space. While the intra-hub $dCor$ reaches 0.6075 in `Other_Cyclic` families, the connection to electronic properties never exceeds 0.22. This indicates that frontier orbital energies are governed by local bonding motifs that are largely orthogonal to the molecule's global spatial and thermal extent.

3.2 Architectural rewiring of manifold transport

Molecular architecture dictates the topological "cost" of aligning these two sub-manifolds. Using the Gromov-Wasserstein (GW) distance as a measure of topological discrepancy, we observe that cyclic systems generally possess a more integrated joint structure than acyclic chains (Figure 4).

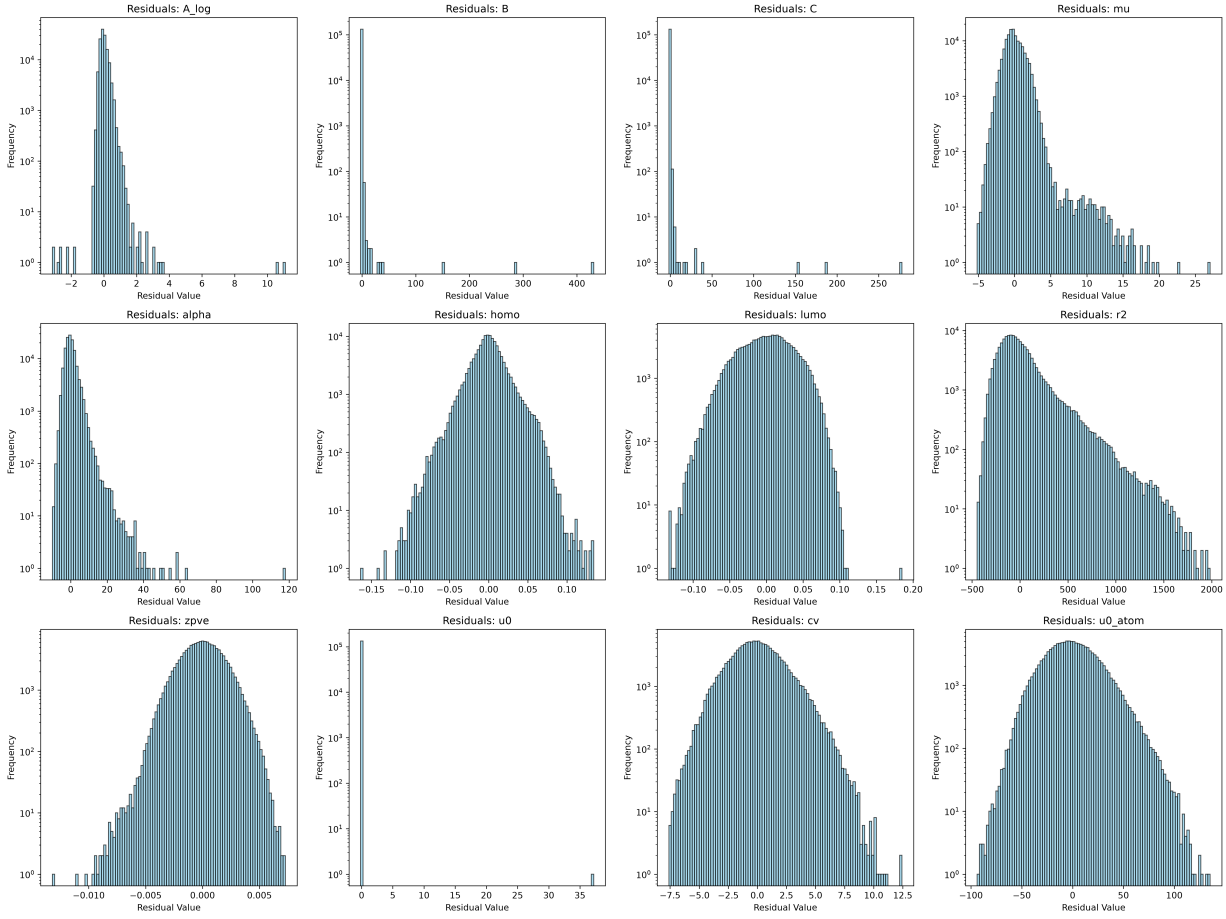

Figure 1: Distributions of molecular property residuals for the QM9 dataset after isolating stoichiometric effects via five-feature atom-count regression (C, N, O, F, H). The histograms, displayed with logarithmic frequency scales, illustrate the varied statistical dispersion across the feature space. While properties such as rotational constants (B, C) and internal energy (u_0) exhibit highly concentrated residuals, the features constituting the electronic sub-manifold (HOMO, LUMO) and the thermodynamic-spatial hub ($zpve$, c_v, r^2) display broader, more continuous distributions. These distinct residual profiles define the foundational state space used to quantify the topological bifurcation and compute the Gromov-Wasserstein transport costs between the decoupled electronic and thermodynamic manifolds across diverse chemical architectures.

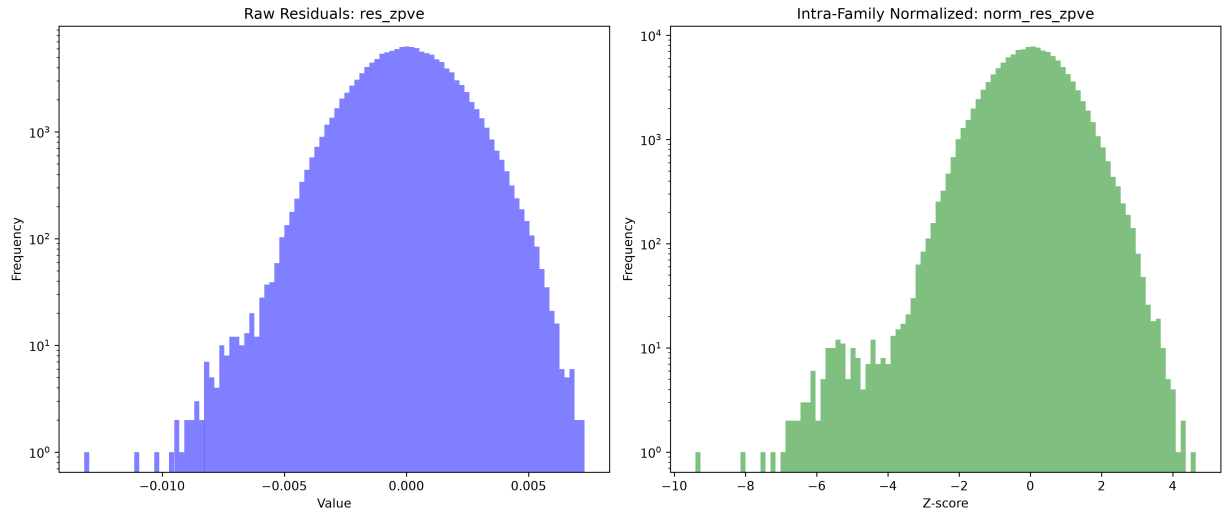

Figure 2: Distribution of the zero-point vibrational energy residuals (`res_zpve`), a core component of the thermodynamic-spatial hub, before (left) and after (right) intra-family normalization. The raw residuals isolate stoichiometric effects through five-feature atom counts. Standardizing these values into Z-scores across distinct molecular families accounts for architectural variations, enabling the Gromov-Wasserstein optimal transport analysis to accurately map the topological decoupling between the electronic and thermodynamic sub-manifolds.

As detailed in Table 2.3, the `Acyclic_Linear` family exhibits the highest GW distance (1.1985), representing the greatest topological mismatch between electronic and spatial organisation.

Conversely, the `C1CC1` (cyclopropane) scaffold shows the lowest transport cost (0.8951), suggesting that the rigid, strained ring architecture forces a tighter, more coherent mapping between the electron density distribution and the thermodynamic state. Furthermore, the transition from linear to branched acyclic structures reduces the GW distance from 1.1985 to 0.9899, a 17.4% decrease in topological mismatch. This identifies branching as a primary mechanism for electronic-spatial reintegration in open-chain molecules.

3.3 Collective dimensionality and the role of the hub

Leave-one-property-out (LOPO) ablation of the GW transport distance reveals that no single property acts as a lone driver of the manifold structure. Instead, the joint organisation is a collective effect of the hub's dimensionality (Figure 5).

Across all families, the removal of any single property from the hub systematically reduces the GW distance, typically to a range of 0.53 to 0.72. For the `Other_Cyclic` family, the baseline GW distance of 0.9055 drops to 0.5695 when res_r^2 is removed and to 0.5979 when res_C is excluded. This systematic collapse indicates that the electronic sub-manifold is easier to align to a lower-dimensional hub.

The relative importance of each property—calculated as the ratio of the ablated distance to the full baseline distance—is remarkably uniform. In the `Acyclic_Branched` family, the importance values for C, c_v, r^2, and $zpve$ are 0.631, 0.678, 0.641, and 0.696, respectively. This uniformity suggests that the thermodynamic-spatial hub functions as a truly joint unit where r^2 (electronic spatial extent) and C (rotational constant) are exchangeable representatives of the molecule's physical scale.

3.4 Chemical drivers of high reintegration

To identify the chemical origins of this manifold coupling, we applied a machine learning classification to the high-reintegration cohort, defined as molecules in the bottom 10th percentile of point-wise GW cost. A LightGBM classifier achieved an Area Under the Curve (AUC) of 0.7944, indicating strong structural patterns that force the electronic sub-manifold to align with the global hub.

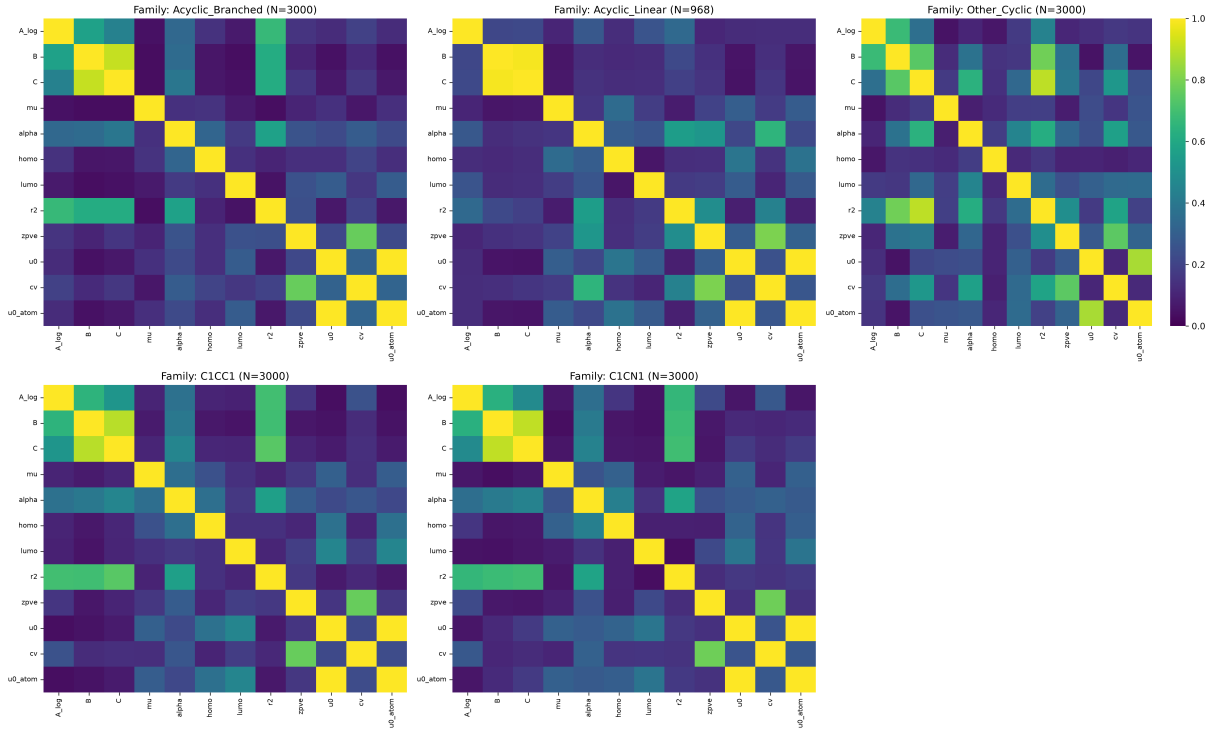

Figure 3: Distance correlation ($dCor$) heatmaps for five molecular families (Acyclic_Branched, Acyclic_Linear, Other_Cyclic, C1CC1, and C1CN1) illustrating the joint structure of molecular properties in the QM9 dataset. The matrices reveal a profound bifurcation between the electronic sub-manifold (res_{homo}, res_{lumo}) and the global thermodynamic-spatial hub (res_{zpve}, res_{cv}, res_{r^2}, res_C). Across all architectures, the correlation between these two sub-manifolds remains persistently low, ranging from 0.1068 in C1CN1 to 0.2186 in Other_Cyclic systems, visually represented by the dark purple off-diagonal blocks. This decoupling demonstrates that frontier electronic states are governed by local bonding motifs that are largely orthogonal to the molecule's global spatial and thermal extent. Conversely, the strong intra-hub correlations (reaching up to 0.6075 in Other_Cyclic systems) highlight a highly integrated thermodynamic state, the topological alignment of which with the electronic manifold is heavily dictated by molecular architecture, such as ring strain in cyclopropane or branching in acyclic chains.

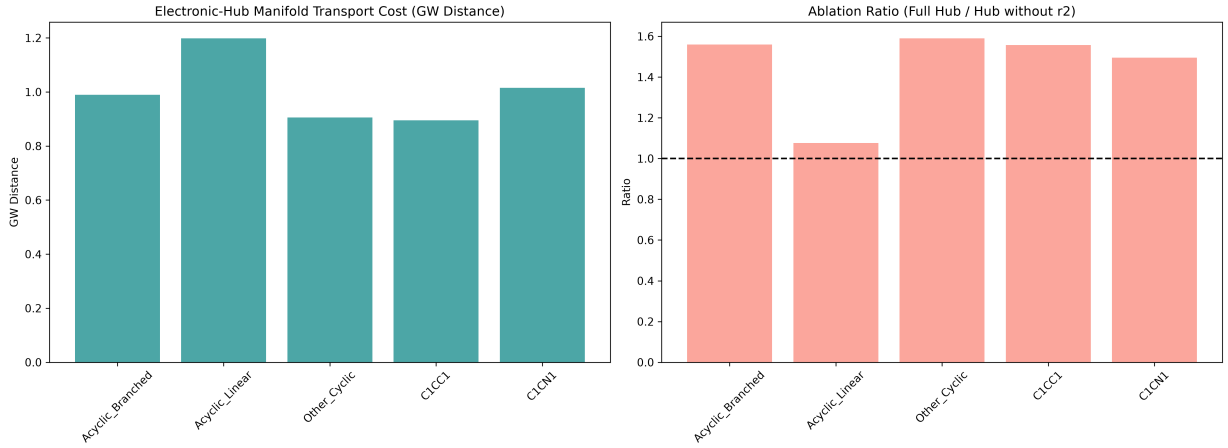

Figure 4: Gromov-Wasserstein (GW) distance between the electronic and thermodynamic-spatial sub-manifolds across distinct molecular families (left) and the ablation ratio comparing the full hub to the hub without the electronic spatial extent, r^2 (right). The left panel demonstrates that molecular architecture dictates manifold alignment; the `Acyclic_Linear` family exhibits the highest topological mismatch, whereas cyclic systems like cyclopropane (`C1CC1`) show lower transport costs, indicating a tighter, more coherent mapping between electron density and thermodynamic state. The right panel illustrates the effect of leave-one-property-out ablation, where the ratio of the full hub distance to the ablated hub distance consistently exceeds 1.0 (dashed line). This systematic reduction in GW distance upon removing r^2 reveals that the electronic sub-manifold is easier to align to a lower-dimensional hub, confirming that the joint manifold structure is driven by the collective dimensionality of the thermodynamic-spatial hub rather than any single property.

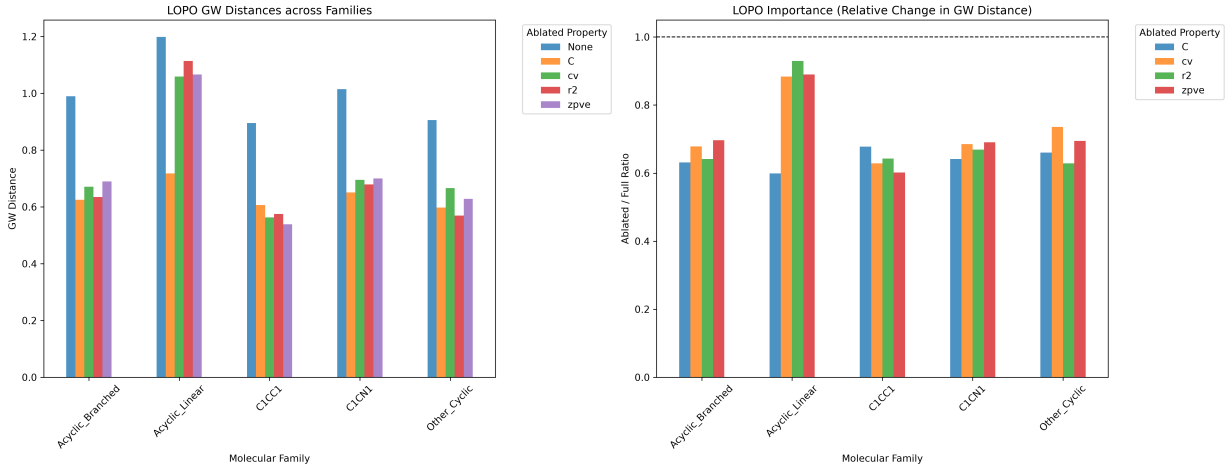

Figure 5: Leave-one-property-out (LOPO) ablation analysis of the Gromov-Wasserstein (GW) transport distance across distinct molecular families. The left panel displays the absolute GW distances when individual properties (C, c_v, r^2, $zpve$) are removed from the thermodynamic-spatial hub compared to the full baseline. The right panel illustrates the relative importance of each property, calculated as the ratio of the ablated distance to the full baseline distance. Across all families, removing any single property systematically reduces the GW distance, indicating that the electronic sub-manifold is easier to align to a lower-dimensional hub. The remarkably uniform importance ratios demonstrate that no single property acts as a lone driver of the manifold structure; instead, the thermodynamic-spatial hub functions as a collective, joint unit where properties are largely exchangeable representatives of the molecule's global physical scale.

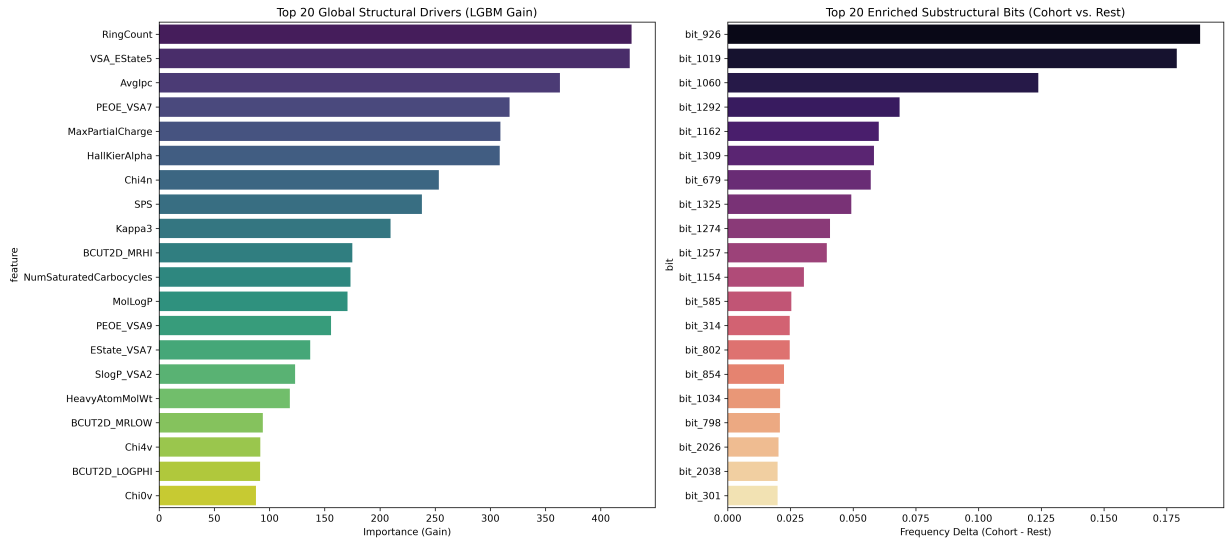

Figure 6: Identification of chemical drivers for high electronic-thermodynamic reintegration. The left panel displays the top 20 global structural features ranked by LightGBM importance gain, highlighting RingCount, `VSA_EState5`, AvgIpc, and MaxPartialCharge as the primary constraints forcing the alignment of the electronic sub-manifold with the global thermodynamic-spatial hub. The right panel illustrates the top 20 enriched substructural Morgan bits in the high-reintegration cohort compared to the rest of the dataset. The highest frequency deltas are observed for `bit_926` ($\Delta = 0.188$) and `bit_1019` ($\Delta = 0.179$), which correspond to dense, heterocyclic environments and specific branching patterns that effectively bridge the topological gap between isolated frontier orbitals and the global spatial response.

The most significant global drivers of reintegration (Figure 6) include RingCount (importance gain: 427.85), which acts as the primary architectural constraint; `VSA_EState5` (426.25), reflecting electronic state distributions across the molecular surface; AvgIpc (362.93), an information-theoretic measure of molecular complexity; and MaxPartialCharge (309.09), which links local electronic extremes to global manifold alignment.

Substructural enrichment analysis further highlights specific Morgan bits associated with manifold collapse. Bit 926 ($\Delta = 0.188$) and Bit 1019 ($\Delta = 0.179$) are the most enriched in the high-reintegration cohort. These features correspond to dense, heterocyclic environments and specific branching patterns that effectively bridge the gap between the isolated frontier orbitals and the molecule's global spatial response. In contrast, the persistent decoupling observed in the `Acyclic_Linear` family is driven by the absence of these reintegrating motifs, leaving the electronic states to vary independently of the spatial backbone.

4 Conclusions

In this study, we have rigorously quantified the physical bifurcation between frontier electronic states and global thermodynamic-spatial features within the QM9 dataset. By isolating stoichiometric effects through exact atom-count residuals and employing Gromov-Wasserstein optimal transport, we mapped the topological decoupling of these property sub-manifolds across diverse chemical architectures. Our findings challenge the implicit assumption of a unified molecular property manifold, demonstrating instead that frontier orbital energies are often governed by local bonding motifs that remain largely orthogonal to a molecule's global spatial and thermal extent.

We established that molecular architecture fundamentally dictates the topological cost of aligning these divergent sub-manifolds. As summarized in Table 2.3, acyclic linear chains exhibit the highest topological mismatch, reflecting a state space where electronic properties vary independently of the spatial backbone.

Conversely, rigid cyclic scaffolds, such as cyclopropane, and branched acyclic structures enforce a tighter

geometric reintegration. The transition from linear to branched chains alone reduces the topological mismatch by over 17%, identifying branching and ring strain as primary structural mechanisms for electronic-spatial coupling.

Furthermore, leave-one-property-out ablation revealed that the joint organization of the molecular state space is a collective effect of the thermodynamic-spatial hub's dimensionality. No single property acts as a lone driver of the manifold structure; rather, features such as electronic spatial extent and rotational constants function as exchangeable representatives of the molecule's physical scale. The systematic collapse of the Gromov-Wasserstein distance upon the removal of any single hub property underscores that the electronic sub-manifold is inherently easier to align to a lower-dimensional representation of the thermodynamic state.

Finally, machine learning classification of the high-reintegration cohort identified the specific chemical drivers responsible for manifold collapse. Global architectural constraints—most notably ring count and molecular complexity—alongside dense heterocyclic environments and specific branching patterns, force isolated frontier orbitals to align with the global spatial response. By mapping these topological constraints, this work provides a foundational framework for developing more robust, physics-aware representations of chemical space, ensuring that future joint property models account for the intrinsic geometric bifurcations of molecular state spaces.