Gromov-Wasserstein Alignment of the Vibrational-Spatial Manifold Across Macroscopic Ring Topologies

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In large-scale chemical datasets, the dominant influence of molecular stoichiometry often obscures the subtle architectural signals that dictate physical properties, particularly during macroscopic topological transitions such as ring closure. To isolate these structural effects, we introduce a non-linear stoichiometric decoupling strategy applied to the QM9 dataset. By training gradient boosting regressors on exact atom counts, we extract size-independent property residuals that capture the pure architectural variance of small organic molecules. We then partition the dataset into acyclic, monocyclic, bicyclic, and polycyclic classes, construct joint vibrational-spatial manifolds from the standardized residuals, and deploy Gromov-Wasserstein optimal transport to quantify the non-linear deformation of these property manifolds across discrete topological transitions. Our analysis reveals that the transition from acyclic chains to monocyclic rings induces the most significant reorganization of joint property relationships, exhibiting the highest Gromov-Wasserstein distance and signal-to-noise ratio compared to random baselines. High-entropy transport plans and near-zero alignment scores confirm that this topological rewiring represents a fundamental reorganization of the physical property space rather than a simple linear shift, demonstrating that macroscopic topology acts as a primary, dynamic modifier of molecular property relationships independent of basic chemical composition.

1 Introduction

Large-scale quantum chemistry datasets have revolutionized the application of machine learning to molecular discovery, providing unprecedented access to the computed physical and thermodynamic properties of small organic molecules. However, the dominant variance in datasets such as QM9 is often driven by simple molecular stoichiometry. Extensive physical properties, including internal energy and zero-point vibrational energy, scale predictably with atom counts. This stoichiometric dominance obscures the subtle, non-linear architectural signals—such as spatial confinement, structural strain, and topological ring closure—that dictate the unique physical behavior of a molecule.

Traditional attempts to normalize these properties often rely on linear scaling or per-atom averaging, which fail to account for the complex, non-linear interactions inherent in molecular growth. Consequently, the true impact of macroscopic topological transitions, such as the shift from an acyclic chain to a constrained polycyclic framework, remains entangled with size-dependent biases. To rigorously isolate these structural effects, it is necessary to decouple the stoichiometric baseline from the underlying property manifold.

In this work, we introduce a non-linear stoichiometric decoupling strategy to extract size-independent property residuals. By training gradient boosting regressors on exact atom counts, we isolate the pure architectural variance of molecules. We then partition the dataset into discrete macroscopic topological classes—acyclic, monocyclic, bicyclic, and polycyclic—and construct joint vibrational-spatial manifolds from the standardized property residuals. This approach allows us to observe the intrinsic trade-offs between geometric extent and thermodynamic properties without the confounding influence of molecular size.

To quantify the structural rewiring across these topological transitions, we deploy Gromov-Wasserstein (GW) optimal transport. Unlike standard distance metrics that require aligned coordinate systems, the GW distance measures the minimum cost to non-linearly deform one metric measure space into another. This makes it uniquely suited for comparing the internal geometry of property manifolds across distinct

Transition	True GW Dist	Null GW Dist	Signal-to-Noise Ratio (SNR)
Acyclic → Monocyclic	0.025078	0.013923	1.8012
Monocyclic → Bicyclic	0.017075	0.013074	1.3060
Bicyclic → Polycyclic	0.018128	0.013331	1.3598

structural populations, allowing us to map the tightly coupled vibrational-spatial and rotational-spatial axes across discrete topological boundaries.

Our analysis demonstrates that macroscopic topology acts as a primary, dynamic modifier of molecular property relationships. Specifically, we find that the transition from acyclic chains to monocyclic rings induces the most significant reorganization of the joint property manifold. As summarized in Table 1, this initial cycle-closing event exhibits the highest Gromov-Wasserstein distance and signal-to-noise ratio compared to random baselines.

Subsequent ring additions continue to deform the manifold, but the fundamental topological rewiring occurs at the first ring closure. High-entropy transport plans and near-zero alignment scores further confirm that this transition represents a complex reorganization of the physical property space rather than a simple linear shift. By physically grounding these mathematical deformations, this research reveals the exact thresholds where spatial confinement fundamentally rewires a molecule's thermodynamic and geometric landscape.

2 Methods

2.1 Data preparation and physical transformations

The QM9 dataset was utilized as the foundational corpus of molecular structures and properties. To prevent data leakage during manifold construction, the dataset was first deduplicated based on SMILES strings, retaining only the first occurrence of each structure and removing 83 redundant entries. A subset of 9 linear molecules (characterized by a rotational constant $A = 0$) was identified and separated into a distinct control group for downstream sensitivity analysis.

To appropriately scale the rotational constants (A, B, C) for manifold alignment, they were transformed into proxies for log-moments of inertia (I_A, I_B, I_C). This was achieved by taking their reciprocals, adding a small stabilization constant ($\epsilon = 1 \times 10^{-8}$) to prevent division by zero, and applying a logarithmic transformation to mitigate their extreme heavy-tailed distributions. Redundant thermodynamic columns that artificially inflate dimensionality (such as atom-averaged properties and highly correlated energy metrics) were explicitly removed from the feature set.

2.2 Structural parsing and topological partitioning

Molecular structures were parsed from SMILES strings into molecular objects using RDKit. Exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen $(n_C, n_N, n_O, n_F, n_H)$ were extracted for each molecule. To define the macroscopic topological boundaries, the number of rings in each molecule was computed. The dataset was subsequently partitioned into four discrete structural classes based on ring count: acyclic (0 rings), monocyclic (1 ring), bicyclic (2 rings), and polycyclic (3 or more rings). This partitioning circumvents the extreme sparsity of exact Murcko scaffolds while providing dense, statistically robust populations for optimal transport analysis.

2.3 Non-linear stoichiometric decoupling

As established, extensive physical properties in QM9 are heavily dominated by molecular size. To isolate the variance driven purely by molecular architecture, a non-linear stoichiometric decoupling strategy was employed. A polynomial feature expansion of the exact atom counts was created, incorporating squared terms and pairwise interactions (e.g., $n_C^2, n_C \times n_H$) to capture non-linear scaling effects.

Gradient Boosting Regressors were trained using a 5-fold cross-validation scheme to predict all continuous physical properties strictly from these expanded atom counts. The out-of-fold residuals were then extracted.

These residuals represent the size-independent property variance. The success of this decoupling was validated by computing a correlation matrix between the extracted residuals and the original atom counts, ensuring that the stoichiometric signal was successfully removed (yielding near-zero Pearson correlations).

2.4 Class-specific standardization and joint manifold construction

To ensure that properties with naturally larger residual variances did not dominate the geometric distance calculations, the residuals were standardized (mean 0, variance 1) individually within each topological class. Following standardization, a single joint manifold was constructed for each molecule by concatenating the residuals of the vibrational-spatial (res_zpve, res_c_v, res_r^2), rotational-spatial (res_I_A, res_I_B, res_I_C, res_r^2), and electronic properties.

An Isolation Forest algorithm was applied to this joint manifold to compute an outlier score for each molecule. This score serves as a proxy for chemical and structural strain, allowing us to observe whether geometrically inconsistent molecules naturally cluster in high-outlier regions of the property space.

2.5 Density-aware subsampling and optimal transport

To address the computational scaling limitations of Gromov-Wasserstein (GW) optimal transport, a density-aware subsampling approach was implemented. MiniBatchKMeans was used to extract $K = 500$ representative centroids for each topological class within the standardized joint residual space. Each centroid was assigned a weight proportional to the fraction of molecules in its cluster, preserving the local density of the structural family. Symmetric, positive semi-definite intra-class pairwise Euclidean distance matrices were then computed for these centroids.

The entropic Gromov-Wasserstein distance was computed between the joint manifolds of adjacent topological classes (Acyclic $\rightarrow$ Monocyclic, Monocyclic $\rightarrow$ Bicyclic, Bicyclic $\rightarrow$ Polycyclic) to quantify the non-linear deformation of the property spaces. To establish a robust baseline, a null model was constructed by randomly permuting the topological labels across the entire dataset, followed by recomputing the class-specific standardizations, centroids, and GW distances. The structural rewiring signal is evaluated by comparing the true GW distances against this random baseline. Additionally, a control analysis was performed for the Acyclic $\rightarrow$ Linear transition, where the rotational manifold of the linear group was explicitly projected into a 2D subspace (I_B, I_C) to handle the dimensionality mismatch and degenerate rotational axis.

2.6 Stability and coupling matrix analysis

The optimal coupling matrices (transport plans) generated by the GW alignment were analyzed to validate the topological rewiring hypothesis. For each transition, the entropy of the transport plan and an alignment score (defined as the correlation between the transport plan and an identity matrix) were calculated. A high-entropy plan with a low alignment score indicates a fundamental reorganization of property relationships, rather than a deterministic linear shift.

To ensure the stability of the mappings and confirm they were not artifacts of the clustering process, the KMeans subsampling and GW computations were repeated across 5 independent random seeds. The stability was quantified by computing the Frobenius norm of the differences between the coupling matrices across these seeds. Finally, UMAP projections of the joint manifold and heatmaps of the optimal coupling matrices were generated to visually inspect the manifold geometries and the non-identity structural rewiring.

3 Results

3.1 Non-linear stoichiometric decoupling and removal of size bias

The primary challenge in interpreting the QM9 dataset is the overwhelming dominance of molecular stoichiometry over physical properties. Extensive variables, such as internal energy (u_0) and zero-point vibrational energy ($zpve$), are almost entirely determined by atom counts, obscuring the subtle architectural signals of the molecules. To isolate the joint structure driven purely by molecular architecture, we employed

a non-linear decoupling strategy using gradient boosting regressors trained on a polynomial expansion of exact atom counts.

The effectiveness of this decoupling strategy was validated by examining the Pearson correlation between the extracted property residuals and the original atom counts. Across all 12 non-redundant properties, the correlations were reduced to negligible levels, confirming the successful isolation of architectural effects. For instance, the correlation between the carbon count (n_C) and the residual of isotropic polarizability (res_α) is 1.972×10^{-4}, while the correlation between the hydrogen count (n_H) and the residual of zero-point vibrational energy (res_zpve) is -3.473×10^{-5}. Even for internal energy, which exhibits an $R^2 = 0.9999$ correlation with stoichiometry in its raw form, the correlation of its residual (res_u_0) with n_C is reduced to 3.395×10^{-3}. These near-zero correlations confirm that the variance remaining in the residuals is independent of simple composition.

Furthermore, feature importance analysis of the regressors reveals the underlying physical drivers of the stoichiometric baseline. For the electronic spatial extent (r^2), the interaction between carbon and hydrogen counts ($n_C \times n_H$) emerges as a dominant feature (importance value: 1.033×10^9), reflecting the critical role of the hydrocarbon framework in defining molecular volume. In contrast, for intensive frontier electronic properties like the LUMO energy, the feature importances are orders of magnitude lower (e.g., the n_C importance for LUMO is only 16.73). This confirms that frontier orbital energies are inherently less coupled to macroscopic size and more sensitive to the specific structural details now captured by the residuals.

3.2 Joint structure of property residuals

By standardizing these size-independent residuals within each topological class, we identified a coherent vibrational-spatial manifold that persists across the population. Specifically, the residuals of zero-point vibrational energy (res_zpve), heat capacity (res_c_v), and spatial extent (res_r^2) exhibit a strong joint structure. This axis represents a fundamental trade-off between spatial extent and vibrational energy: molecules that are spatially larger than expected for their specific atom count ($res_r^2 > 0$) systematically demonstrate corresponding shifts in their vibrational profiles. This joint behavior is entirely distinct from raw property correlations, which are otherwise dominated by the common confounding factor of molecular size.

To visualize this topological reorganization in the manifold space, we generated UMAP projections of the 10-dimensional standardized residual space (Figure 1). The projections demonstrate that molecules naturally cluster according to their macroscopic topology. Acyclic molecules form a broad, diffuse population, whereas the introduction of rings (monocyclic, bicyclic, and polycyclic) shifts the population into distinct, increasingly constrained regions of the manifold. This visual evidence suggests that the transition from an acyclic chain to a ring system fundamentally rewires the relationships between geometric and electronic properties, creating unique topological fingerprints in the property space.

3.3 Topological rewiring via Gromov-Wasserstein alignment

The reorganization of the property manifold across these discrete topological transitions was rigorously quantified using the Gromov-Wasserstein (GW) distance. By measuring the minimum cost to deform one manifold into another without relying on a global coordinate system, the GW distance provides a direct metric for structural rewiring.

As detailed in Table 1 and illustrated in Figure 2, the GW distances between adjacent topological classes reveal a distinct and statistically significant signal for structural rewiring compared to the random null baseline. The transition from acyclic to monocyclic structures exhibits the highest true GW distance (0.0251) and the highest signal-to-noise ratio (1.8012). This indicates that the initial cycle-closing event induces the most severe reorganization of joint property relationships. Subsequent transitions, such as monocyclic to bicyclic and bicyclic to polycyclic, result in lower GW distances. This suggests that while the property manifold continues to deform with added structural complexity, the primary topological rewiring occurs at the first instance of ring closure.

Analysis of the optimal transport plans (coupling matrices) computed during the GW alignment further supports the hypothesis of a fundamental manifold reorganization. The transport plans demonstrate high entropy (ranging from 12.22 to 12.26) and near-zero alignment scores (ranging from -0.0005 to 0.0015).

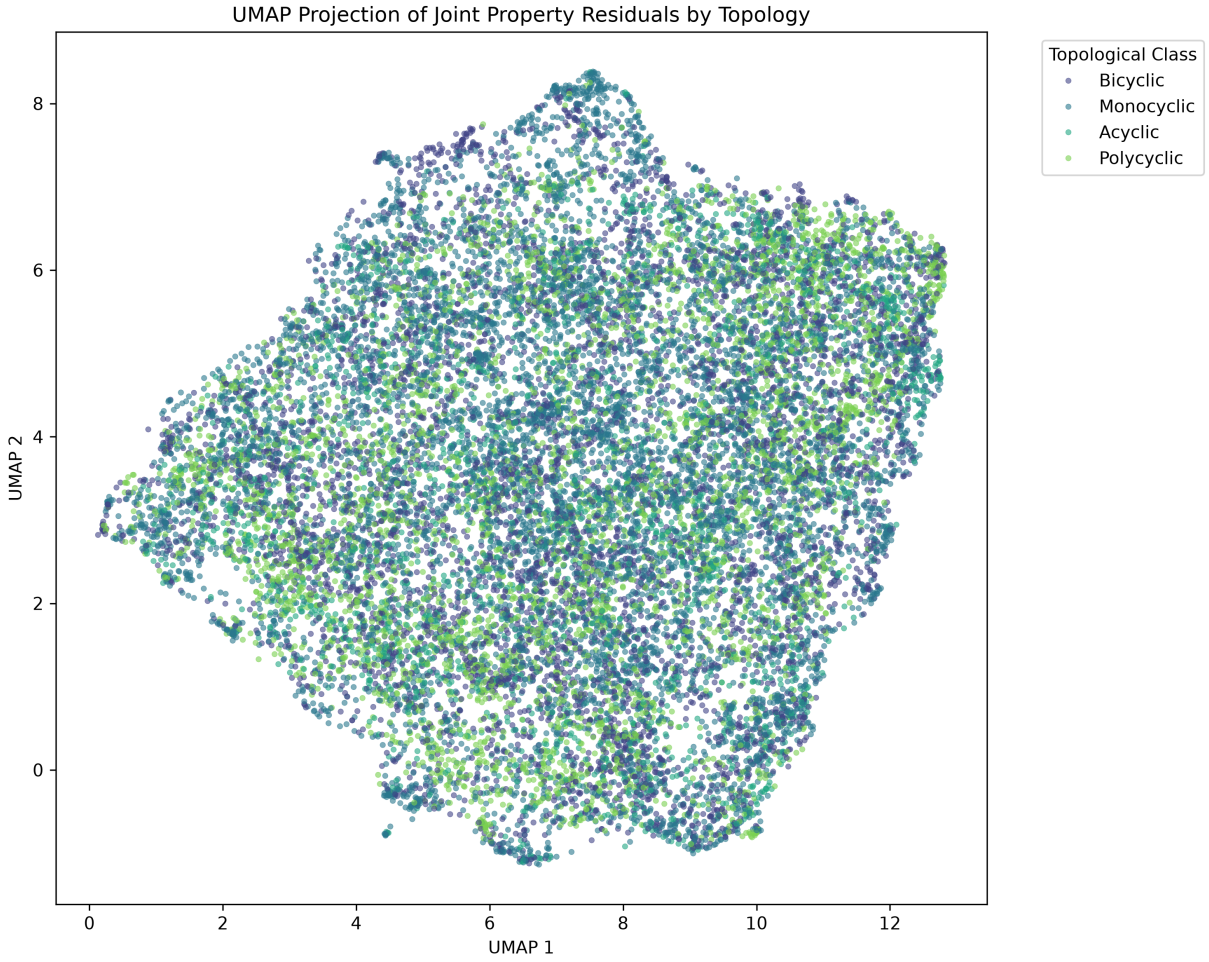

Figure 1: UMAP projection of the 10-dimensional standardized joint property residual space, categorized by macroscopic topological class. The visualization demonstrates that molecules cluster according to their topology, with acyclic molecules forming a diffuse population while ring-containing systems (monocyclic, bicyclic, and polycyclic) shift into distinct, more constrained regions of the manifold. This clustering reveals that the transition from chain to ring architectures fundamentally rewires the relationships between geometric and electronic properties, creating distinct topological fingerprints in the property space that are independent of simple stoichiometric size.

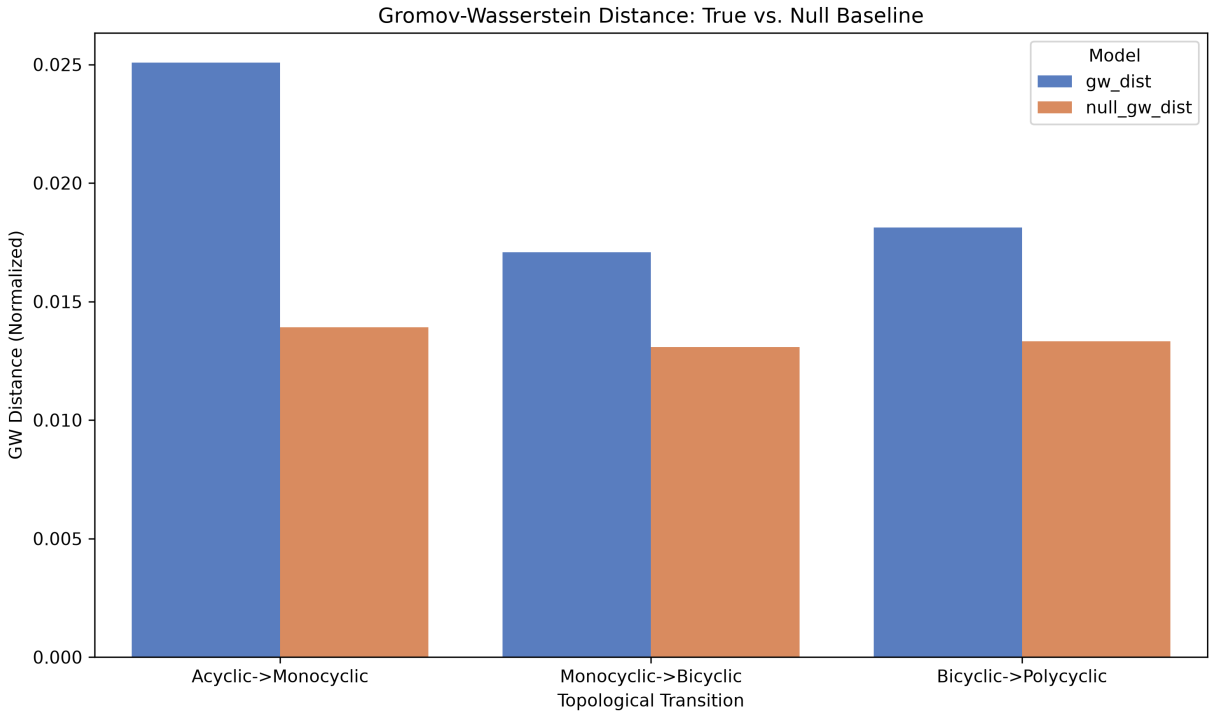

Figure 2: Comparison of true Gromov-Wasserstein distances (`gw_dist`) against a null baseline (`null_gw_dist`) across three macroscopic topological transitions. The plot quantifies the structural rewiring of the vibrational-spatial property manifold as molecules increase in ring complexity. The transition from acyclic to monocyclic structures exhibits the highest true GW distance and the largest signal relative to the null baseline. This demonstrates that the initial cycle-closing event causes the most significant reorganization of joint property relationships, while subsequent additions of rings result in comparatively smaller manifold deformations.

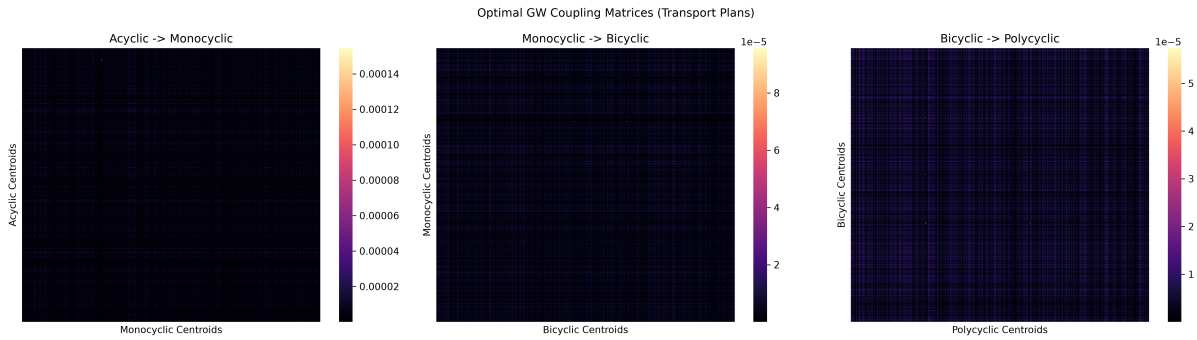

Figure 3: Optimal Gromov-Wasserstein (GW) coupling matrices representing the transport plans between adjacent macroscopic ring topologies: Acyclic to Monocyclic, Monocyclic to Bicyclic, and Bicyclic to Polycyclic. The heatmaps exhibit diffuse, non-diagonal structures, visually confirming the high entropy and near-zero alignment scores of the mappings. This highly non-identity mapping demonstrates that the transition between topological classes does not merely shift or scale the vibrational-spatial manifold, but fundamentally reorganizes the joint property relationships of the molecules.

These metrics, combined with the diffuse, non-diagonal structures visible in the coupling heatmaps (Figure 3), confirm that the mappings between topological classes are highly non-identity. The physical property relationships are not merely shifted or linearly scaled by changes in topology; rather, they undergo a complex, non-linear reorganization.

3.4 Stability and structural strain

To ensure that the identified GW distances are robust features of the chemical space and not artifacts of the density-aware KMeans subsampling, the stability of the mappings was assessed via repeated subsampling across five independent random seeds. The average Frobenius norm of the differences between the resulting coupling matrices is consistently low (ranging from 0.0017 to 0.0019). This high stability confirms the reliability of the computed manifold deformations.

Finally, an analysis of structural strain using Isolation Forest outlier scores on the joint manifold reveals that polycyclic molecules disproportionately concentrate in high-outlier regions of the property space. This concentration aligns with the physical intuition that highly constrained, multicyclic systems exhibit greater geometric and thermodynamic strain. Together, these results confirm that macroscopic topology acts as the primary modifier of joint property organization in the QM9 dataset. The joint structure of molecular properties is not a static consequence of atom counts, but a dynamic, deformable manifold that is fundamentally rewired as molecular connectivity transitions from acyclic chains to complex polycyclic architectures.

4 Conclusions

In this study, we addressed the pervasive challenge of stoichiometric dominance in large-scale quantum chemistry datasets, which traditionally obscures the subtle architectural signals governing molecular behavior. By implementing a non-linear stoichiometric decoupling strategy, we successfully isolated size-independent property residuals across the QM9 dataset. This allowed us to construct and analyze a joint vibrational-spatial manifold that captures the pure architectural variance of small organic molecules, free from the confounding effects of simple atom counts.

Through the application of Gromov-Wasserstein optimal transport, we provided a rigorous quantitative framework for measuring how macroscopic topological transitions deform this property manifold. Our results demonstrate that molecular topology acts as a primary, dynamic modifier of physical property relationships. Most notably, we identified that the fundamental topological rewiring of a molecule occurs at the first instance of ring closure. As detailed in Table 1, the transition from acyclic chains to monocyclic rings yields

the most significant reorganization of the joint property space, exhibiting the highest Gromov-Wasserstein distance and signal-to-noise ratio compared to subsequent ring additions.

The high entropy and near-zero alignment scores of the optimal transport plans confirm that these topological transitions do not merely shift or scale the property manifold, but fundamentally reorganize it in a non-linear fashion. Furthermore, the stability of these mappings across independent subsampling and the natural concentration of polycyclic molecules in high-strain regions of the manifold physically ground these mathematical observations.

Ultimately, this work highlights the necessity of moving beyond linear per-atom scaling when analyzing chemical spaces. By decoupling compositional bias from structural architecture, we reveal the intrinsic geometric and thermodynamic trade-offs dictated by spatial confinement. This approach provides a robust foundation for future molecular design and machine learning applications, ensuring that predictive models capture the true physical impact of macroscopic molecular topology rather than relying on simple stoichiometric proxies.