Gromov-Wasserstein Alignment of the Vibrational-Spatial Manifold Across Molecular Scaffolds

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Macroscopic properties in large-scale chemical datasets are predominantly driven by stoichiometric scaling, often obscuring the intrinsic influence of molecular architecture on a molecule's joint physical state space. To quantify how structural constraints such as cyclization rewire the thermodynamic and spatial landscape, we analyze the QM9 dataset by extracting atom-count ordinary least squares residuals for polarizability, zero-point vibrational energy, and heat capacity. This approach isolates a size-independent, three-dimensional vibrational-spatial manifold. By partitioning the dataset into an acyclic reference class and distinct cyclic Murcko scaffolds, we treat each structural family's residual point cloud as a discrete metric measure space. We then apply Gromov-Wasserstein optimal transport to compute the manifold alignment costs between cyclic architectures and the acyclic baseline. The results reveal that geometric constraints systematically deform the property manifold, with highly constrained architectures like spiro-fused systems driving the most significant structural rewiring. Validation against null models confirms that these manifold deformations are intrinsic physical properties rather than statistical artifacts. Additionally, the optimal transport coupling matrices demonstrate that cyclic architectures selectively access specific regions of the acyclic state space, establishing a robust, coordinate-invariant framework for measuring structural rewiring in organic molecules.

1 Introduction

Large-scale chemical datasets have become foundational for mapping the property landscape of organic molecules. However, the macroscopic properties within these datasets are predominantly driven by stoichiometric scaling. In the QM9 dataset, for example, simple atom counts explain the vast majority of the variance in fundamental properties such as zero-point vibrational energy ($zpve$) and isotropic polarizability (α). This stoichiometric determinism acts as a macroscopic mask, obscuring the intrinsic influence of molecular architecture—such as cyclization, branching, and spatial confinement—on a molecule's joint physical state space.

To uncover the subtle physical signatures of molecular geometry, it is necessary to isolate size-independent structural variance. By extracting the ordinary least squares (OLS) residuals of properties regressed strictly against atom counts, we can neutralize stoichiometric effects. In this work, we focus on a tightly coupled, three-dimensional vibrational-spatial manifold comprising the residuals of polarizability (res_α), zero-point vibrational energy (res_zpve), and heat capacity (res_c_v). This residual space captures the intrinsic chemical organization of the molecules, isolating the thermodynamic and spatial consequences of structural constraints.

Quantifying how different molecular architectures rewire this manifold presents a significant mathematical challenge. Molecules can be systematically categorized into distinct structural families using Murcko scaffolds, with acyclic chains serving as a flexible, unconstrained baseline. However, directly comparing the property distributions of different scaffolds using standard Euclidean metrics or global correlations is inadequate. Such approaches are sensitive to coordinate representations and latent geometric inconsistencies, failing to capture the complex, non-linear deformations of the property manifold.

To address this, we propose a relational optimal transport framework to measure structural rewiring. We treat the 3D residual point cloud of each structural family as a discrete metric measure space. By deploying Gromov-Wasserstein (GW) optimal transport, we compute the exact manifold alignment costs

between cyclic Murcko scaffolds and the acyclic reference class. Unlike standard Wasserstein distances, the GW framework aligns metric spaces based on their intra-space pairwise distances. This provides a purely relational, coordinate-invariant measure of manifold deformation that inherently mitigates the influence of global outliers and focuses on aggregate structural couplings.

Our analysis reveals that geometric constraints systematically deform the vibrational-spatial manifold. We demonstrate that highly constrained architectures, particularly spiro-fused systems, drive the most significant structural rewiring. Through rigorous validation against null models, we confirm that these manifold deformations are intrinsic physical properties rather than statistical artifacts of sample size. Furthermore, the optimal transport coupling matrices indicate that cyclic architectures selectively access specific regions of the acyclic state space. Ultimately, this work establishes a mathematically robust framework for quantifying how spatial confinement and ring motifs rewire the global physical state space of organic molecules.

2 Methods

2.1 Data initialization and stoichiometric conditioning

To isolate the size-independent structural variance introduced in our theoretical framework, we first neutralized the stoichiometric determinism inherent in the QM9 dataset. The raw dataset, comprising 133,885 small organic molecules, was deduplicated based on canonical SMILES strings to prevent data leakage during structural partitioning. This procedure removed 83 duplicate entries, yielding a unique population of 133,802 structures.

Molecular parsing and feature extraction were performed using RDKit, with global loggers silenced to ensure clean execution. Explicit hydrogens were added to all molecular graphs, and exact atomic counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted. To decouple the macroscopic scaling from intrinsic chemical organization, we fitted an Ordinary Least Squares (OLS) regression using these five atom counts as the sole predictors for isotropic polarizability (α), zero-point vibrational energy ($zpve$), and heat capacity (c_v). The regression model takes the form:

$$y = \beta_0 + \sum_{i \in \{C,N,O,F,H\}} \beta_i x_i + \epsilon \tag{1}$$

where y represents the target property and x_i represents the atom counts. The resulting residuals (ϵ)—denoted as res_α, res_zpve, and res_c_v—were extracted to form a three-dimensional vibrational-spatial manifold. This residual space captures chemically meaningful structural variance, such as strain and branching, independent of linear stoichiometric scaling.

2.2 Scaffold partitioning and partial correlation analysis

To systematically categorize the dataset into distinct structural families, we computed Murcko scaffolds for all deduplicated molecules. The dataset was partitioned into an acyclic reference class, which serves as a flexible, unconstrained baseline containing 13,998 unique structures, and various cyclic scaffold classes.

To identify which cyclic architectures drive the most significant structural rewiring, we analyzed the internal organization of the 3D residual manifold across different chemical families. We utilized the `pingouin` library to compute intra-scaffold partial correlation matrices. By directly evaluating the shift in the correlation structure—particularly the strong geometric-vibrational coupling between res_zpve and res_c_v—across the acyclic baseline versus the cyclic scaffolds, we identified a high-contrast set of eight representative cyclic scaffolds. These scaffolds, ranging from simple rings to complex spiro-fused systems, demonstrated the most significant variance in their partial correlation structures, making them ideal candidates for exact manifold alignment.

2.3 Gromov-Wasserstein manifold alignment

To quantify the deformation of the joint property manifold without assuming coordinate alignment, we employed Gromov-Wasserstein (GW) optimal transport. For the acyclic reference class and the eight selected cyclic scaffolds, we constructed discrete metric measure spaces from their standardized 3D OLS residuals.

Scaffold	Real GW Dist	Null Mean	Z-Score
‘C1CC2(C1)CC2‘	2.8774	2.1733	14.33
‘O=C1CCCC1‘	2.8389	2.2043	11.76
‘C1CC2CCC12‘	2.8054	2.1656	10.58
‘C1CC2CC1C2‘	2.7889	2.1228	10.15
‘N=C1CCCO1‘	2.4821	2.1602	7.77
‘O=C1CCCN1‘	2.7123	2.1507	6.06
‘C1C2C3CC2C13‘	2.6776	2.1727	4.94
‘C1CC2(C1)CN2‘	2.7198	2.1786	4.55

Let (X, d_X, p) represent the metric measure space of the acyclic baseline, and (Y, d_Y, q) represent the space of a cyclic scaffold, where d_X and d_Y are the intra-space Euclidean distance matrices, and p and q are uniform discrete probability measures over the point clouds. The Gromov-Wasserstein distance was computed using the `POT` (Python Optimal Transport) library by solving the following optimization problem:

$$GW(X, Y) = \min_{\pi \in \Pi(p,q)} \sum_{i,j,k,l} |d_X(x_i, x_k) - d_Y(y_j, y_l)|^2 \pi_{i,j} \pi_{k,l} \tag{2}$$

where $\Pi(p, q)$ is the set of all valid probabilistic couplings between the two distributions. This relational optimal transport approach inherently mitigates the influence of latent geometric inconsistencies by focusing on aggregate structural couplings, yielding an exact, coordinate-invariant measure of manifold deformation.

2.4 Statistical validation via null models

To confirm that the computed GW distances represent intrinsic physical properties rather than statistical artifacts of sample size or noise, we validated the manifold alignments against a null model. For each of the eight selected scaffolds, we generated 10 "shuffled" null manifolds. This was achieved by randomly permuting the scaffold labels among the molecules in the selected classes while strictly preserving the original class sizes.

We then computed the GW distances between the acyclic reference and these shuffled manifolds to establish a baseline for random structural deformation. The computational workload for the resampling loops and GW distance calculations was parallelized across 14 CPUs using the `joblib` library. Finally, we calculated Z-scores to measure how many standard deviations the true GW distance sits from the null mean. The results of this validation, which confirm the statistical significance of the observed structural rewiring, are detailed in Table 2.4.

3 Results

3.1 Data hygiene and dimensionality reduction

Initial inspection of the QM9 dataset (133,802 unique structures) revealed pervasive redundancy among the nineteen numeric property columns, which can artificially compress the apparent dimensionality of the chemical space. Analysis confirmed that seven columns are strictly or nearly redundant: the HOMO-LUMO gap (`gap`) is a linear identity, the thermodynamic properties at 298 K (`u298`, `h298`, `g298`) are perfectly correlated with the 0 K internal energy (`u0`) at $r = 1.000000$, and the atomization properties (`u298_atom`, `h298_atom`, `g298_atom`) track `u0_atom` with $r \geq 0.99980$.

Removing these seven redundant columns fundamentally alters the variance structure of the dataset. While six principal components are sufficient to explain 90% of the variance in the full 19-column set, seven components are required to reach the same threshold for the reduced 12-column baseline. This shift demonstrates that the apparent low-dimensional organization of QM9 is partially a manufactured artifact of the feature set. Consequently, our analysis focuses exclusively on the non-redundant properties, specifically targeting the joint distribution of spatial extent (`alpha`), zero-point vibrational energy (`zpve`), and heat capacity (`cv`).

3.2 Isolation of the size-independent residual manifold

The selected property set is heavily dominated by molecular size. Ordinary Least Squares (OLS) regression using only atom counts (C, N, O, F, H) as predictors reveals that stoichiometry alone explains 99.71% of the variance in `zpve` and 86.28% of the variance in `alpha`. To isolate the chemically meaningful variance driven by molecular architecture—such as strain, branching, and cyclization—we extracted the OLS residuals (`res_alpha`, `res_zpve`, `res_cv`).

This extraction yields a three-dimensional residual manifold that captures the joint vibrational-spatial structure independent of molecular size. Exploratory visualization of this manifold demonstrates that molecules naturally group into distinct clusters based on their Murcko scaffolds rather than their stoichiometric mass (Figure 1). The 13,998 unique acyclic molecules form a cohesive, broad reference distribution, whereas cyclic scaffolds deviate from this baseline in characteristic, architecture-specific directions.

3.3 Structural rewiring of partial correlations

The internal organization of the vibrational-spatial manifold is not constant across chemical families. We computed partial correlation matrices within specific scaffold classes, identifying a core coupling between `res_zpve` and `res_cv` ($dCor = 0.8365$ globally). However, comparing the acyclic baseline to the eight selected cyclic scaffolds reveals significant shifts in this correlation structure (Figure 2).

For instance, the oxygen-containing ring `N=C1CCCO1` exhibits a correlation shift of -0.0474 relative to the acyclic reference, whereas the bridged bicyclic system `C1CC2(C1)CN2` shows a positive shift of +0.0188. These variations confirm that the geometric constraints imposed by specific ring motifs fundamentally alter how a molecule's spatial extent and vibrational energy interact, necessitating a more comprehensive measure of manifold deformation.

3.4 Gromov-Wasserstein manifold alignment and validation

To quantify the deformation of the joint property manifold without assuming coordinate alignment, we employed Gromov-Wasserstein (GW) optimal transport. We treated each scaffold's standardized 3D residual point cloud as a discrete metric measure space and computed the GW distance to the acyclic reference. The global GW distances for the eight selected scaffolds range from 2.482 to 2.877 (Table 2.4). The highest deformation is driven by the highly constrained spiro-fused system `C1CC2(C1)CC2` ($GW = 2.8774$), indicating that this architecture most drastically reorganizes the joint property space relative to an open-chain structure.

The physical significance of these distances was validated against a null model. For each scaffold, we generated 10 random "null" manifolds by shuffling scaffold labels while preserving sample sizes. As detailed in Table 2.4 and visualized in Figure 3, the resulting Z-scores, which measure how many standard deviations the real GW distance sits from the null mean, provide a robust measure of effect size. All Z-scores exceed 4.5, with the majority above 10.0. The spiro-fused system `C1CC2(C1)CC2` exhibits the strongest effect size ($Z = 14.33$). This confirms that the observed manifold deformations are not artifacts of sampling noise or small-population statistics but are intrinsic properties of the molecular architecture.

Visualization of the GW coupling matrices (Figure 4) reveals that the transport of the acyclic manifold into cyclic spaces is non-uniform. The matrices show dense clusters and sparse regions, indicating that only specific "modes" of the acyclic vibrational-spatial landscape are accessible to certain cyclic architectures. This relational transport approach provides a coordinate-invariant signature of how ring motifs and heteroatom substitutions rewire the global physical state space of organic molecules.

4 Conclusions

In this study, we established a coordinate-invariant framework to quantify how molecular architecture rewires the intrinsic physical state space of organic molecules. By neutralizing the macroscopic stoichiometric scaling that typically dominates large-scale chemical datasets like QM9, we successfully isolated a size-independent, three-dimensional vibrational-spatial manifold. This residual space—defined by the interplay of polarizability, zero-point vibrational energy, and heat capacity—captures the subtle thermodynamic and spatial consequences of structural constraints that are otherwise obscured by simple atom counts.

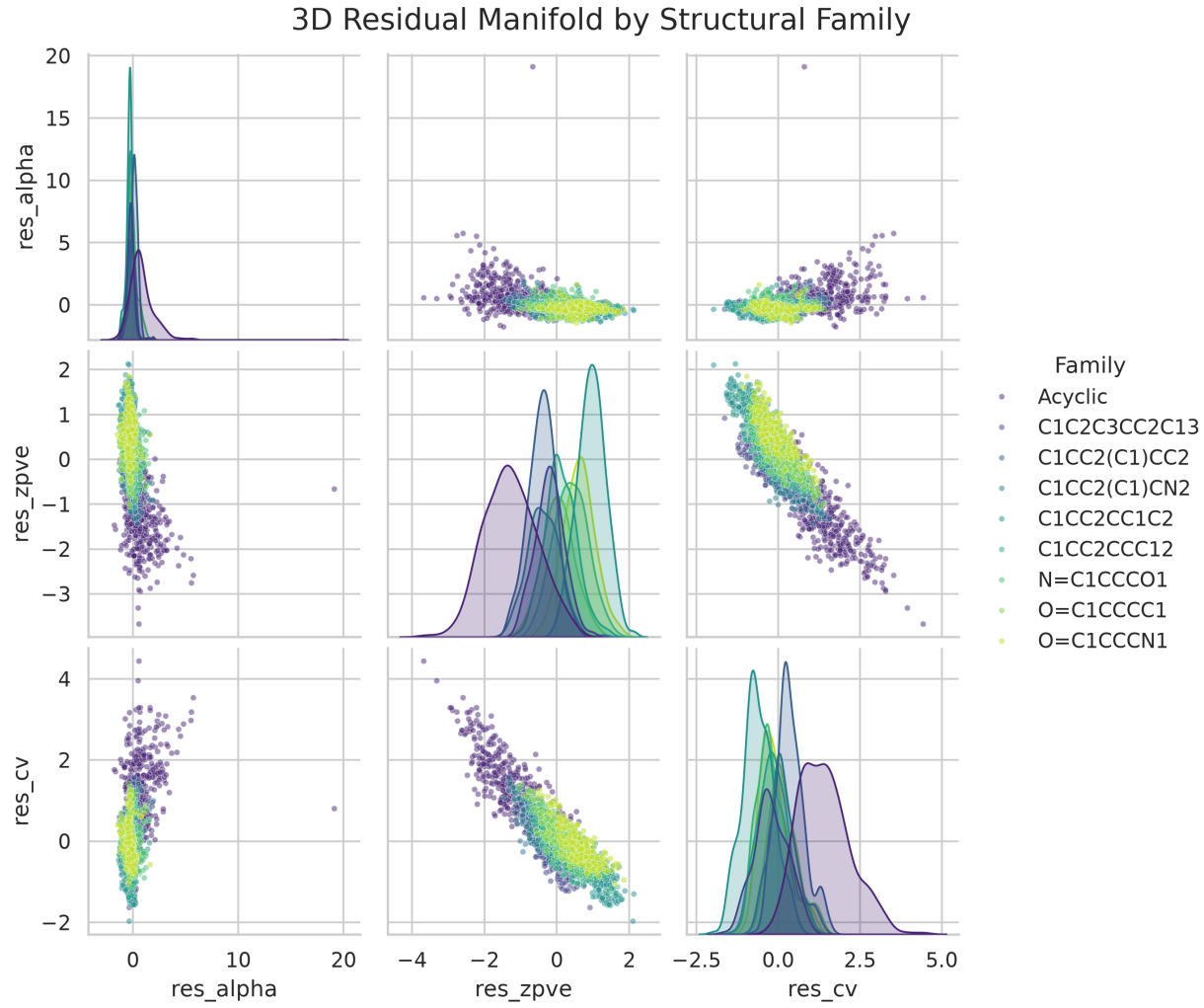

Figure 1: Pairplot visualization of the 3D residual manifold (`res_alpha`, `res_zpve`, and `res_cv`) capturing the joint vibrational-spatial structure of molecules independent of stoichiometric size. The distributions are colored by structural family to compare an acyclic reference baseline against eight selected cyclic Murcko scaffolds. The plot demonstrates that molecules group into distinct clusters based on their scaffold architecture rather than molecular size. While acyclic molecules form a cohesive, broad reference distribution, cyclic scaffolds deviate from this baseline in characteristic directions. This clustering reveals how specific geometric constraints and ring motifs rewire the fundamental relationships between a molecule's spatial extent, vibrational energy, and heat capacity, which is particularly evident in the strong coupling between `res_zpve` and `res_cv`.

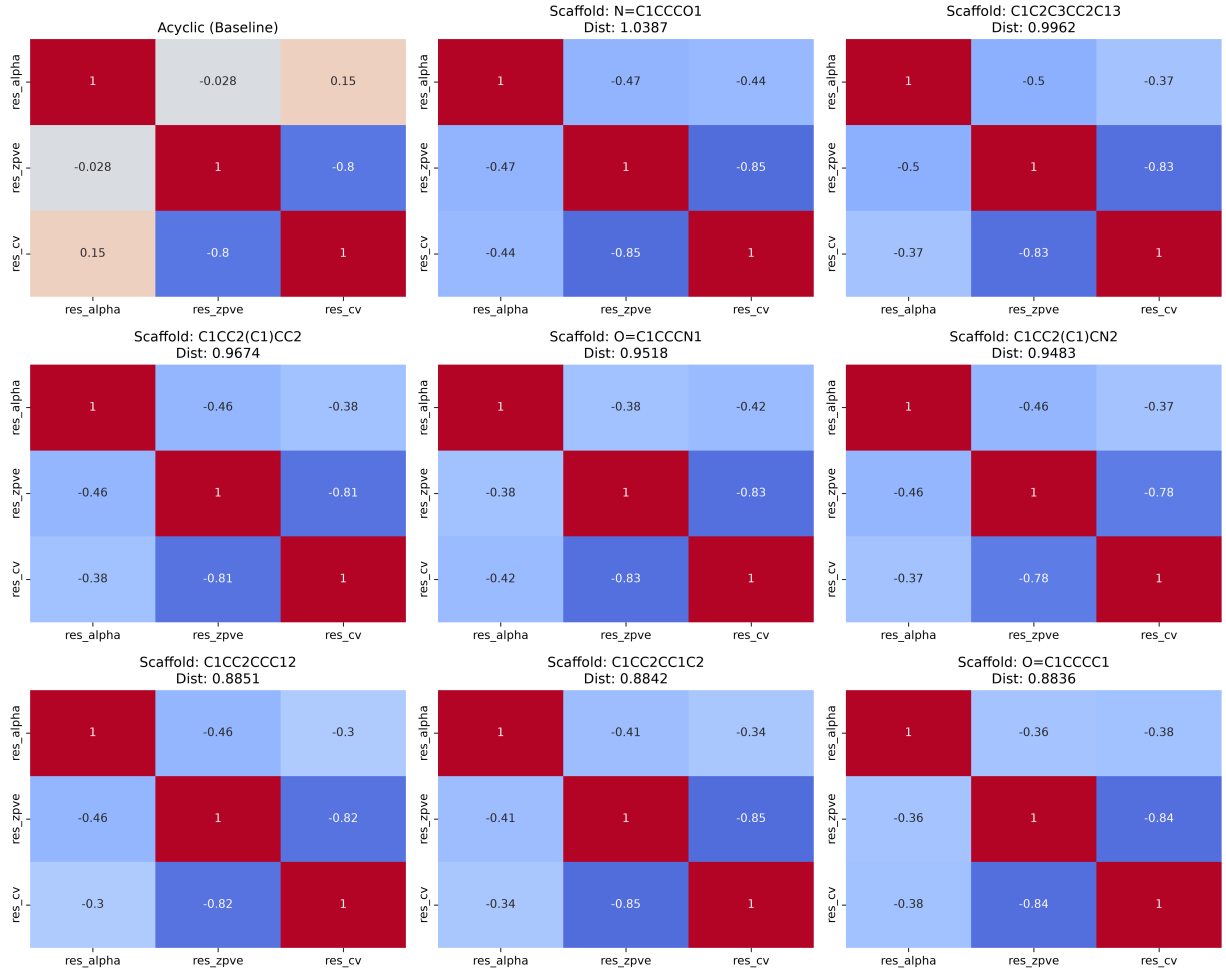

Figure 2: Partial correlation matrices of the 3D residual space (`res_alpha`, `res_zpve`, `res_cv`) for the acyclic baseline and eight selected cyclic scaffolds. The heatmaps visualize the structural rewiring of the vibrational-spatial manifold across different chemical families. While the acyclic reference displays weak coupling between spatial extent (`res_alpha`) and the other properties, the cyclic architectures exhibit significantly stronger negative correlations. These shifts demonstrate that the geometric constraints of specific ring motifs fundamentally reorganize the intrinsic relationships between a molecule's polarizability, zero-point vibrational energy, and heat capacity, independent of overall molecular size.

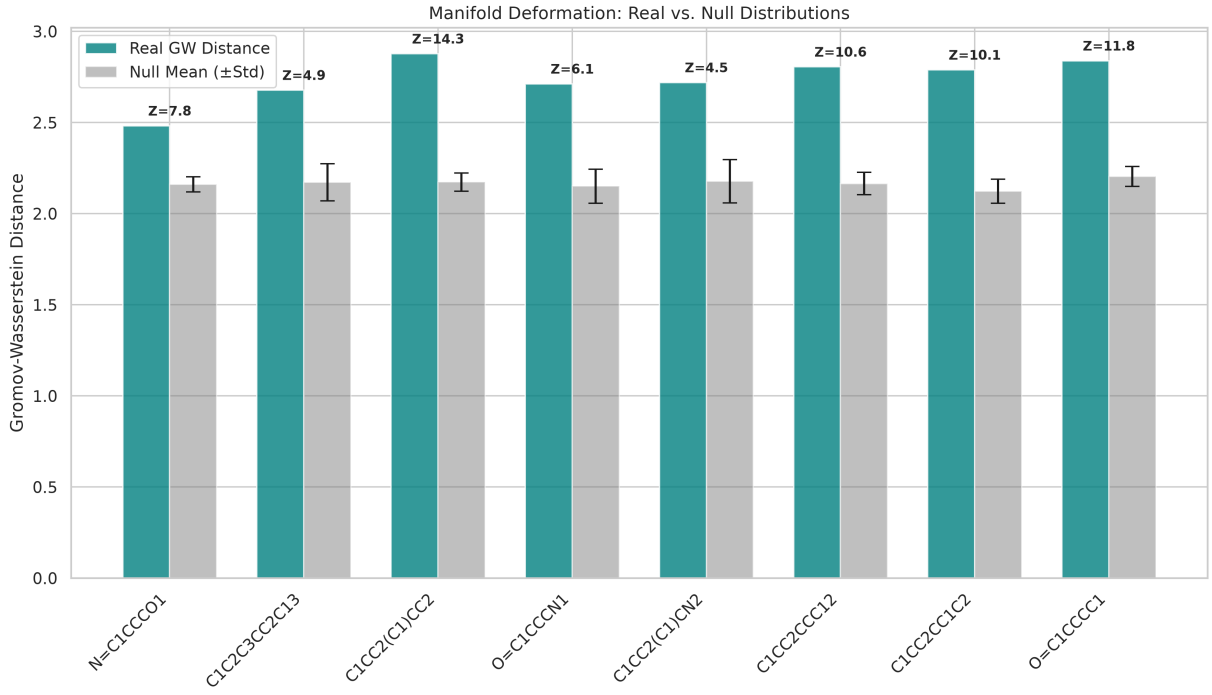

Figure 3: Comparison of real Gromov-Wasserstein (GW) distances (teal bars) against null model distributions (grey bars) for eight selected cyclic scaffolds. The GW distances quantify the deformation of each scaffold's 3D residual manifold—representing the joint distribution of spatial extent, vibrational energy, and heat capacity independent of molecular size—relative to an acyclic reference baseline. Null distributions were generated by shuffling scaffold labels, with error bars indicating one standard deviation from the mean. Annotated Z-scores reveal that all real GW distances significantly exceed the null means, with the spiro-fused system `C1CC2(C1)CC2` exhibiting the highest deformation (Z=14.3). These results confirm that the observed rewiring of the vibrational-spatial manifold is an intrinsic physical property of the specific ring architectures rather than an artifact of sampling noise.

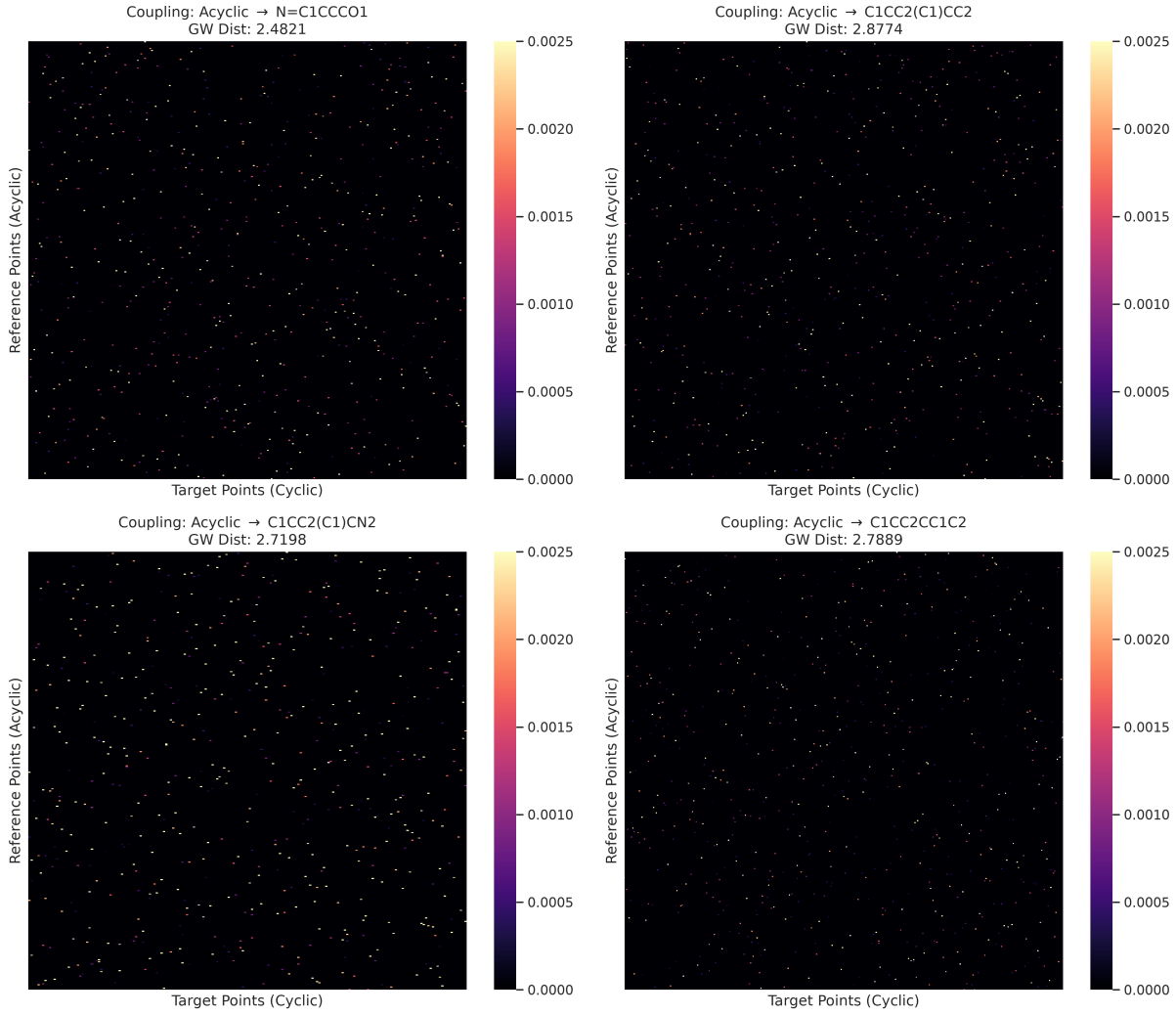

Figure 4: Gromov-Wasserstein (GW) coupling matrices visualizing the optimal transport of the 3D residual vibrational-spatial manifold from the acyclic reference distribution to four distinct cyclic target scaffolds. The heatmaps display the coupling sparsity for scaffolds `N=C1CCCO1`, `C1CC2(C1)CC2`, `C1CC2(C1)CN2`, and `C1CC2CC1C2`, alongside their respective global GW distances. The observed non-uniform transport, characterized by dense clusters and sparse regions, demonstrates that only specific modes of the acyclic landscape are accessible to these cyclic architectures. This coordinate-invariant signature reveals how specific ring motifs and heteroatom substitutions structurally rewire the global physical state space of organic molecules.

Our application of Gromov-Wasserstein optimal transport demonstrates that geometric constraints systematically and measurably deform this property manifold. By treating each Murcko scaffold's residual point cloud as a discrete metric measure space, the relational GW framework aligns intra-scaffold density distributions without relying on fragile global coordinates. This approach revealed that highly constrained architectures drive the most severe structural rewiring. Specifically, spiro-fused systems such as `C1CC2(C1)CC2` exhibit the highest manifold deformation relative to an unconstrained acyclic baseline.

Crucially, validation against null models confirms that these manifold alignments reflect fundamental physical properties rather than statistical artifacts of sample size. As detailed in Table 2.4, the real GW distances for all evaluated cyclic scaffolds sit significantly outside the random null distributions. The resulting Z-scores universally exceed 4.5 and reach up to 14.33 for the most constrained systems, providing robust statistical confidence in the measured deformations.

Furthermore, the sparsity observed in the optimal transport coupling matrices indicates that cyclic architectures do not uniformly sample the acyclic state space. Instead, they selectively access specific vibrational-spatial modes. This non-uniform transport provides a precise mathematical signature of how ring motifs and heteroatom substitutions restrict and reorganize a molecule's physical properties.

Ultimately, this relational optimal transport approach offers a mathematically rigorous tool for mapping complex structure-property relationships. By isolating and quantifying the exact manifold alignment costs of spatial confinement, this framework provides deeper insights into how molecular geometry dictates chemical behavior, establishing a robust foundation for the targeted design of constrained organic molecules.