Bootstrapped Wasserstein Alignment of the 12-Dimensional Residual Manifold to Map Topological Rewiring

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how molecular topology influences physical properties is challenging because stoichiometric determinism and geometric artifacts often obscure these intrinsic effects. To isolate and quantify the true impact of structural rewiring, such as cyclization, we decouple the molecular property manifold from stoichiometry using gradient-boosted regression on the QM9 dataset. By extracting and robustly scaling the residuals of twelve non-redundant physical properties, and filtering out geometric inconsistencies via cross-validated reconstruction errors, we establish a clean, stoichiometry-independent state space. We then apply optimal transport theory, computing bootstrapped Wasserstein-2 distances between the twelve-dimensional residual point clouds of acyclic baseline molecules and specific cyclic scaffolds. The results demonstrate that cyclization induces a significant, multi-axial deformation of the property manifold. Cost decomposition reveals that this topological rewiring is overwhelmingly driven by geometric and spatial response properties—most notably the electronic spatial extent—which account for 60 to 70 percent of the alignment cost, while frontier orbital energies remain largely unaffected. Additionally, principal component analysis of the transport displacement vectors confirms that this structural transition manifests as a complex, high-dimensional reorganization rather than a simple shift along a single axis, providing a mathematically rigorous framework for mapping molecular physical states.

1 Introduction

The relationship between molecular architecture and physical properties is a central pillar of chemical physics. However, isolating the pure effect of topological rewiring—such as the transition from an acyclic chain to a cyclic scaffold—is notoriously difficult. The primary confounding factor is stoichiometric determinism: macroscopic thermodynamic and electronic properties are overwhelmingly dictated by simple atom counts and molecular weight. In large-scale chemical spaces such as the QM9 dataset, the variance in fundamental properties like internal energy (U_0) or zero-point vibrational energy (ZPVE) is almost entirely explained by molecular size and composition. This stoichiometric dominance masks the subtle, intrinsic physical effects of spatial confinement and structural connectivity.

Furthermore, direct comparisons of molecular properties across structural families are frequently obscured by geometric artifacts and extreme distributional skewness. For instance, unflagged geometric inconsistencies in computational datasets can masquerade as genuine topological shifts, while heavy-tailed distributions in features like the rotational constant A can destabilize distance-based metrics. Consequently, traditional linear correlations and low-dimensional projections fail to capture how a molecule's joint physical state space fundamentally deforms under topological constraints.

To rigorously quantify the impact of structural rewiring, we propose a framework that decouples the molecular property manifold from stoichiometry. By employing gradient-boosted regression to predict physical properties solely from atom counts and basic unsaturation metrics, we extract the residuals for twelve non-redundant physical properties. This procedure strips away the stoichiometric baseline, revealing a clean, stoichiometry-independent state space. Within this residual manifold, tightly coupled vibrational-spatial axes and orthogonal electronic features can be analyzed directly. To ensure stability, we apply robust rank-based scaling and a $\log(A + 1)$ transformation to neutralize extreme kurtosis, alongside a cross-validated reconstruction error filter to eliminate geometric outliers.

To measure the topological shift within this 12-dimensional residual space, we apply optimal transport theory. Specifically, we compute exact, bootstrapped Wasserstein-2 (W_2) distances between the residual point clouds of an acyclic baseline and distinct cyclic scaffolds. Unlike simple centroid shifts, the Wasserstein metric captures the full distributional deformation required to map one structural family onto another, providing a mathematically rigorous quantification of manifold rewiring.

In this work, we demonstrate that cyclization induces a significant, multi-axial deformation of the property manifold. Through a marginal decomposition of the optimal transport cost, we reveal that this topological rewiring is overwhelmingly driven by geometric and spatial response properties—most notably the electronic spatial extent (r^2)—which account for 60 to 70 percent of the alignment cost. In contrast, frontier orbital energies remain largely unaffected by these specific topological transitions. Finally, principal component analysis of the transport displacement vectors confirms that structural cyclization manifests as a complex, high-dimensional reorganization rather than a simple shift along a single axis, offering a robust new paradigm for mapping molecular physical states.

2 Methods

2.1 Data curation and structural parsing

To establish a rigorous foundation for manifold analysis, we utilized the QM9 dataset, initially deduplicating entries based on SMILES strings to isolate 133,802 unique molecules. To eliminate artificial low-dimensional structures caused by definitional redundancies, we discarded derived thermodynamic and atomization columns, as well as the HOMO-LUMO gap. We retained twelve non-redundant physical properties to define our state space: the rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), frontier orbital energies (HOMO, LUMO), electronic spatial extent (r^2), zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity at 298.15 K (C_v), and atomization energy at 0 K ($U_{0,\text{atom}}$).

Because the rotational constant A exhibits extreme skewness and a heavy-tailed distribution (kurtosis $> 100,000$), we applied a $\log(A+1)$ transformation to stabilize its variance. Nine perfectly linear molecules ($A = 0$) were isolated and excluded from the primary dataset to prevent them from acting as extreme outliers during manifold alignment, though they were retained for subsequent projection to verify their boundary positions in the residual space.

Structural parsing was performed to extract exact atom counts (C, N, O, F, H) and the total heavy-atom count. We computed the Murcko scaffold for each molecule to partition the dataset into an acyclic baseline (defined by an empty Murcko scaffold) and distinct cyclic populations. For the acyclic baseline, the distributions of the B and C rotational constants were verified to ensure the absence of highly strained or effectively linear artifacts. Finally, we calculated continuous ring complexity descriptors (number of rings, average ring size, aromatic rings, degree of unsaturation) alongside topological compactness descriptors that do not rely on 3D coordinates, specifically Balaban's J index and the Bertz CT index, serving as proxies for spatial extent and branching complexity.

2.2 Stoichiometric decoupling and outlier exclusion

To isolate genuine chemical architecture from stoichiometric determinism, we trained gradient-boosted decision tree (GBDT) regression models to predict each of the twelve physical properties. The predictors were strictly limited to compositional and basic structural metrics: exact atom counts, heavy-atom count, and degree of unsaturation. By extracting the residuals from these models, we stripped away the stoichiometric baseline, yielding a 12-dimensional residual manifold. To ensure the observed rewiring signals were not merely proxies for specific structural features (the "stoichiometric decoupling fallacy"), we also performed a leave-one-feature-out sensitivity analysis, training parallel sets of GBDT models with iteratively removed predictors.

Computational datasets like QM9 contain unflagged geometric inconsistencies that can masquerade as topological rewiring. To address this, we calculated the aggregate reconstruction error (mean squared error) for each molecule using a cross-validated, out-of-fold prediction strategy. This approach prevents circularity

wherein a model might "learn" an outlier's error. The top 2.5% of molecules with the highest cross-validated reconstruction errors (amounting to 3,054 molecules) were flagged as potential geometric outliers.

2.3 Residual scaling and matched-pair sampling

To ensure that properties with naturally larger residual variances did not dominate the downstream distance calculations, we applied a robust rank-based scaling to the residuals. A quantile transformer, mapped to a standard normal distribution, was fitted on the entire dataset to guarantee normalization stability and ensure perfect comparability between the acyclic baseline and cyclic scaffolds. We generated two parallel datasets for all subsequent computations: a "Full" dataset containing all scaled residuals, and a "Clean" dataset excluding the top 2.5% of flagged geometric outliers.

For each highly populated cyclic scaffold, we identified a matched subset of acyclic molecules to serve as a direct comparative baseline. Matching was performed first by enforcing exact equivalence in heavy-atom count and atom-type composition. Subsequently, a nearest-neighbor approach was applied in the topological compactness space (using the scaled Balaban's J and Bertz CT indices) to minimize branching divergence. This isomer-aware matching yielded subsets of approximately 1,000 to 2,000 molecules per scaffold, ensuring that any measured distance reflects pure topological rewiring rather than residual compositional drift.

2.4 Bootstrapped Wasserstein alignment

To quantify the deformation of the property manifold, we computed the exact optimal transport cost between the 12-dimensional residual point clouds of the matched acyclic and cyclic subsets. We utilized the Wasserstein-2 (W_2) distance, defined with a squared Euclidean cost metric, which captures the minimal effort required to morph one probability distribution into another.

To rigorously assess the stability of the structural rewiring signal and isolate genuine manifold deformation from sample-driven noise, we executed a subsampling bootstrap procedure. For each matched set, we sampled $N = 250$ molecules over 500 iterations, computing the W_2 distance for each iteration. This bootstrapped alignment was performed in parallel across both the Full and Clean datasets to quantify the inflationary effect of geometric artifacts on the perceived topological shift.

2.5 Cost decomposition and displacement analysis

Because the W_2 distance was computed using a squared Euclidean metric, the total transport cost is additively separable across the twelve dimensions. Utilizing the optimal transport coupling matrix (the transport plan), we calculated the marginal contribution of each physical property to the total W_2 cost. This decomposition allowed us to pinpoint exactly which physical dimensions—such as geometric spatial responses versus frontier orbital energies—drive the manifold deformation during cyclization.

Finally, to determine the directionality and coherence of the manifold shift (i.e., whether the property space stretches along a single axis or undergoes a complex rotation), we extracted the 12-dimensional displacement vectors ($y - x$) for the mass transported between the matched samples. We performed Principal Component Analysis (PCA) on the covariance structure of these displacement vectors. The variance explained by the principal components was analyzed to assess the dimensionality of the structural transition, providing a mathematical description of how spatial confinement reorganizes the joint physical state space.

3 Results

3.1 Stoichiometric decoupling and residual manifold isolation

The joint structure of molecular properties in the QM9 dataset is fundamentally governed by stoichiometric determinism, which masks the intrinsic physical effects of chemical architecture. As established in our methodology, we isolated the residuals of twelve non-redundant properties from a gradient-boosted stoichiometric baseline. The removal of these stoichiometric effects significantly reorganizes the property space. While raw thermodynamic properties such as internal energy (U_0) and zero-point vibrational energy (ZPVE) are near-perfectly determined by atom counts ($R^2 > 0.99$), their residuals reveal a hidden 12-dimensional

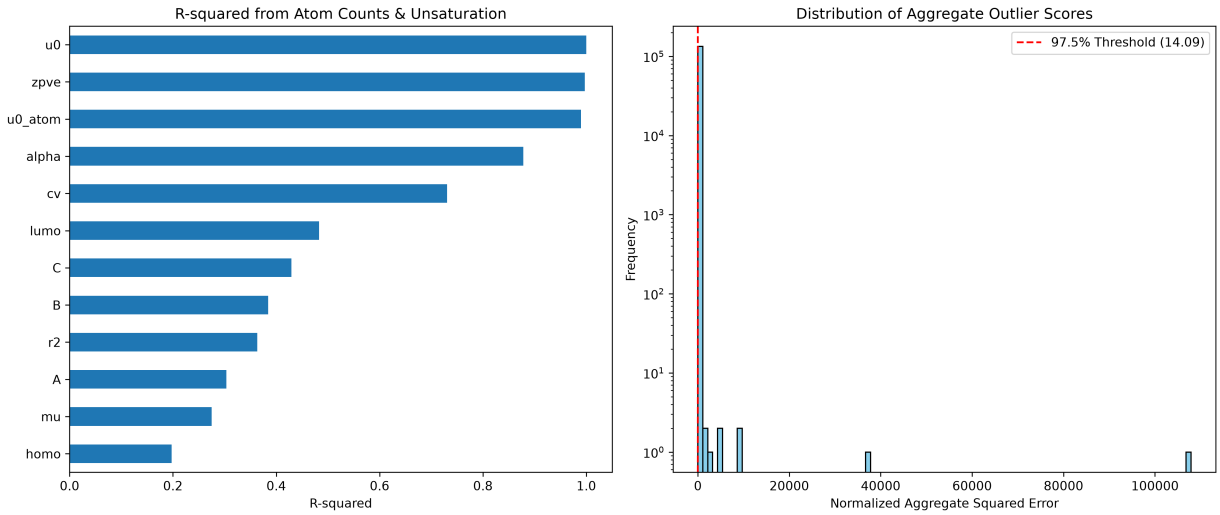

Figure 1: The left panel displays the R^2 values for molecular properties predicted from atom counts and unsaturation, illustrating that while thermodynamic properties like internal energy (U_0) and zero-point vibrational energy (ZPVE) are near-perfectly determined by stoichiometry ($R^2 > 0.99$), properties such as isotropic polarizability (α), heat capacity (C_v), and electronic spatial extent (r^2) retain variance that forms a distinct 12-dimensional residual manifold. The right panel shows the frequency distribution of normalized aggregate squared errors, with a red dashed line marking the 97.5% threshold (14.09). Excluding the top 2.5% of geometric outliers beyond this cutoff generates the "Clean" dataset, which prevents unflagged geometric inconsistencies from artificially inflating the perceived topological manifold deformation during optimal transport alignment.

manifold where properties like isotropic polarizability (α), heat capacity (C_v), and electronic spatial extent (r^2) remain coupled (Figure 1, left panel). To ensure that unflagged geometric inconsistencies did not artificially inflate the perceived topological shifts, we excluded the top 2.5% of molecules with the highest cross-validated reconstruction errors, generating a "Clean" dataset for downstream alignment (Figure 1, right panel).

Following extraction, the residuals were standardized using a robust rank-based quantile transformation mapped to a standard normal distribution. This scaling is essential to ensure that properties with naturally larger residual variances do not dominate the multidimensional space, thereby allowing for a balanced comparison of the intrinsic physical effects of chemical architecture (Figure 2).

Within this scaled residual space, the transition from acyclic to cyclic scaffolds (e.g., cyclopropane `C1CC1`, cyclobutane `C1CCC1`) induces a clear shift in the property centroids. Parallel coordinates analysis demonstrates that acyclic molecules occupy a distinct region of the 12-dimensional space, characterized particularly by lower mean ZPVE residuals (-1.071 standard deviations) and higher C_v residuals (1.090) compared to their cyclic counterparts (Figure 3). Cyclization systematically shifts these values toward the global mean; for instance, the cyclopropane scaffold (`C1CC1`) shifts the ZPVE and C_v residuals to -0.886 and 1.038, respectively.

3.2 Quantifying topological rewiring via Wasserstein alignment

To rigorously measure the total deformation required to align the property distribution of a cyclic scaffold with its stoichiometrically matched acyclic baseline, we computed the exact Wasserstein-2 (W_2) distance. Across the top five cyclic scaffolds in the Clean dataset, the alignment costs are both stable and significant. The cyclobutane scaffold (`C1CCC1`) demonstrates the largest manifold deformation with a mean W_2 distance of 4.009 ± 0.109 (standard deviation across bootstrap iterations). This is followed by the oxetane scaffold (`C1COC1`) at 3.745 ± 0.103. In contrast, the smaller oxirane scaffold (`C1CO1`) exhibits the smallest relative shift at 3.054 ± 0.100. The remarkable consistency of these costs across bootstrap iterations indicates that

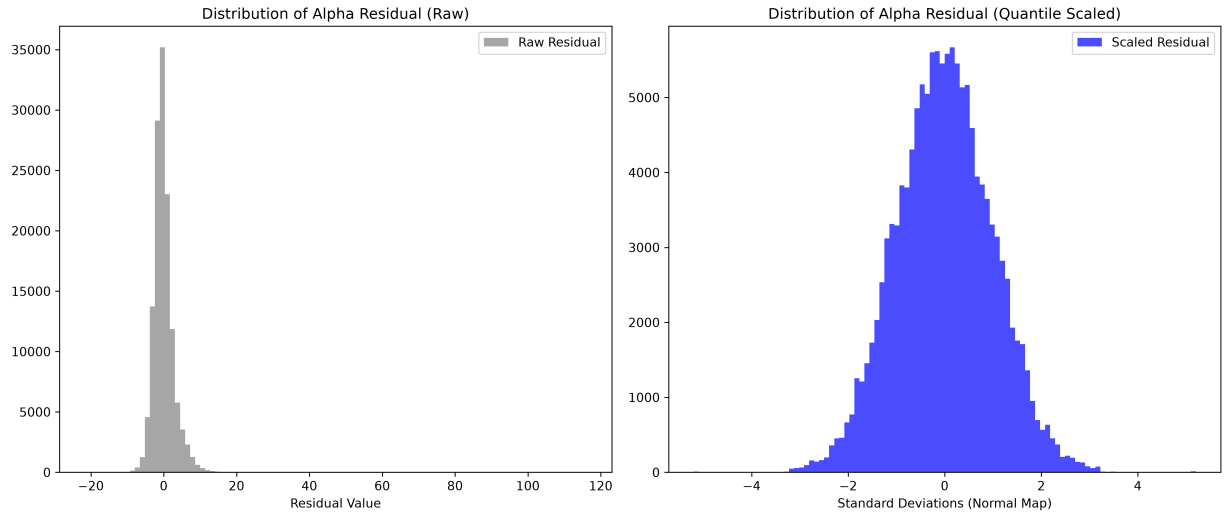

Figure 2: Distribution of the isotropic polarizability (α) residuals before (left) and after (right) quantile scaling. The left panel displays the raw residuals isolated after decoupling stoichiometric effects from the QM9 dataset. The right panel demonstrates the transformation of these raw values into quantile-scaled residuals mapped to a standard normal distribution. This standardization into standard deviations is essential for isolating the hidden 12-dimensional property manifold. By normalizing the distributions of properties like α, the intrinsic physical effects of chemical architecture are revealed, enabling the rigorous quantification of topological rewiring between acyclic and cyclic molecular scaffolds via Wasserstein alignment.

the observed rewiring is a robust feature of the chemical families rather than an artifact of specific molecular samples.

3.2.1 Impact of geometric inconsistency

The comparison between the Clean and Full datasets highlights the critical impact of geometric inconsistencies. In the Full dataset, mean W_2 costs are systematically inflated, ranging from 4.108 for `C1CO1` to 5.167 for `C1CCC1`. This inflation confirms that unflagged geometric artifacts can masquerade as topological shifts. However, the underlying structural trend—where larger rings like `C1CCC1` exhibit greater manifold deformation than smaller rings like `C1CC1`—is preserved even when outliers are included.

3.3 Dimensional drivers of manifold deformation

By decomposing the W_2 cost into the marginal contributions of each property, we identified the physical dimensions most sensitive to topological change. The deformation is overwhelmingly driven by geometric and spatial response properties rather than frontier orbital energies (Figure 4).

3.3.1 Geometric vs. electronic contributions

Across all evaluated scaffolds, geometric properties (rotational constants A, B, C, polarizability α, and spatial extent r^2) account for approximately 60% to 70% of the total W_2 cost. For example, in the `C1CCC1` scaffold, geometric properties contribute 68.8% to the total cost, while electronic properties contribute only 12.1%. Similarly, for the `C1CN1` scaffold, geometric properties account for 59.5% of the cost, compared to 21.9% for electronic properties. The electronic spatial extent (r^2) emerges as the single most dominant driver of manifold rewiring, contributing 23.1% to the cost for `C1CCC1` and 19.3% for `C1CN1`. This indicates that the primary physical consequence of cyclization is a fundamental reorganization of the electronic density distribution and the molecular moment of inertia.

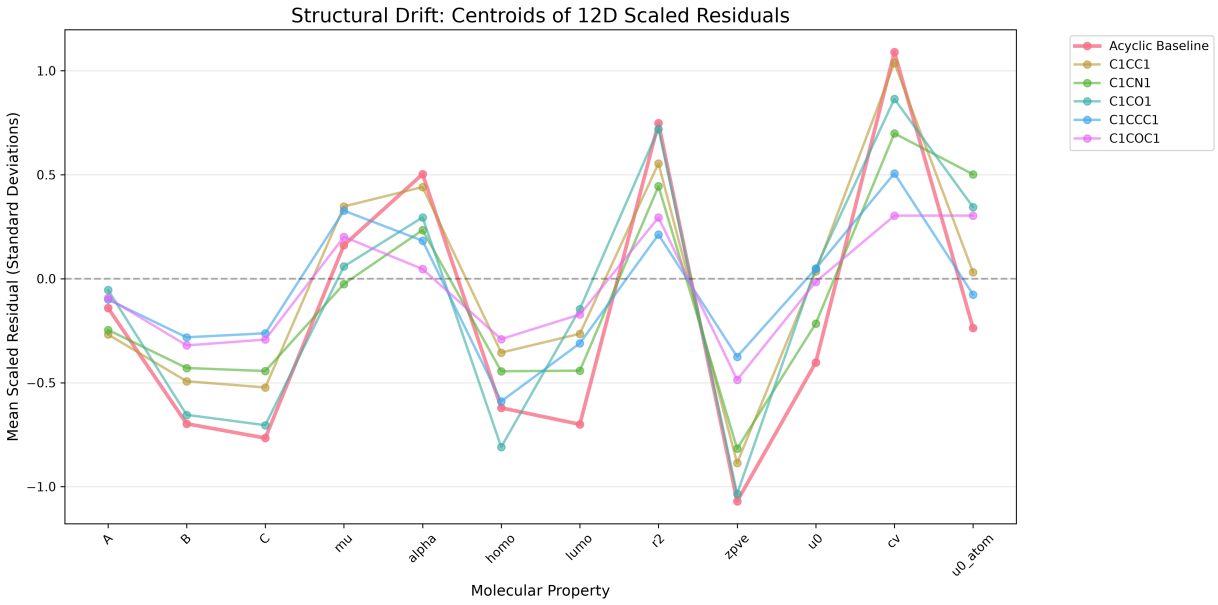

Figure 3: Parallel coordinates plot illustrating the structural drift of 12-dimensional scaled property residuals for five cyclic scaffolds relative to an acyclic baseline. Following stoichiometric decoupling, the transition from acyclic to cyclic structures induces a topological rewiring of the molecular property manifold. The plot reveals that acyclic molecules occupy a distinct region of the property space, characterized by lower mean zero-point vibrational energy (ZPVE) residuals and higher heat capacity (C_v) residuals compared to their cyclic counterparts. Cyclization shifts these property centroids toward the global mean; for instance, the cyclopropane scaffold (`C1CC1`) shifts ZPVE and C_v residuals from -1.071 and 1.090 standard deviations to -0.886 and 1.038, respectively. This demonstrates how topological changes fundamentally reorganize the coupled property space.

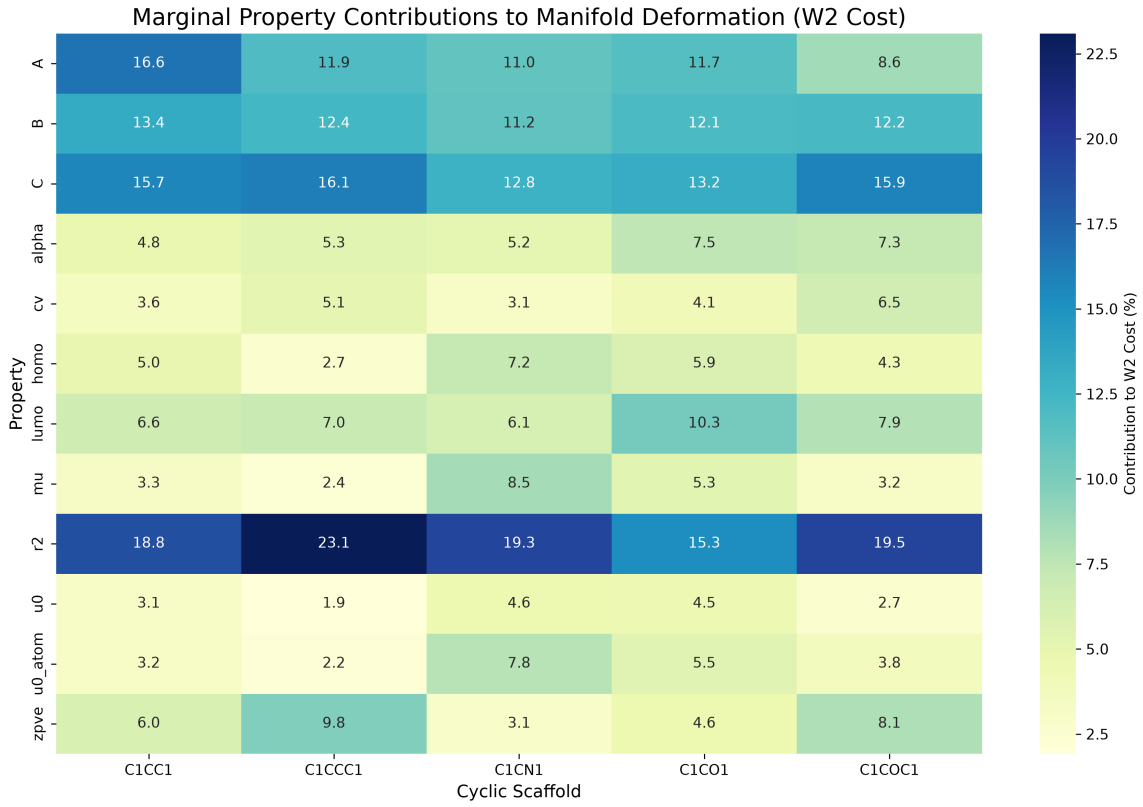

Figure 4: Heatmap detailing the marginal percentage contributions of twelve molecular properties to the total Wasserstein-2 (W_2) manifold deformation cost across five cyclic scaffolds. The decomposition reveals that topological rewiring from acyclic to cyclic structures is overwhelmingly driven by geometric and spatial response properties, which account for approximately 60–70% of the total W_2 cost. The electronic spatial extent (r^2) emerges as the single most dominant driver, contributing up to 23.1% for the cyclobutane scaffold (`C1CCC1`), followed closely by rotational constants (A, B, C). Conversely, frontier orbital energies (e.g., HOMO, LUMO) and thermodynamic properties contribute significantly less. This demonstrates that the primary physical effect of cyclization is the reorganization of the electronic density distribution and molecular moment of inertia, rather than shifts in electronic energy levels.

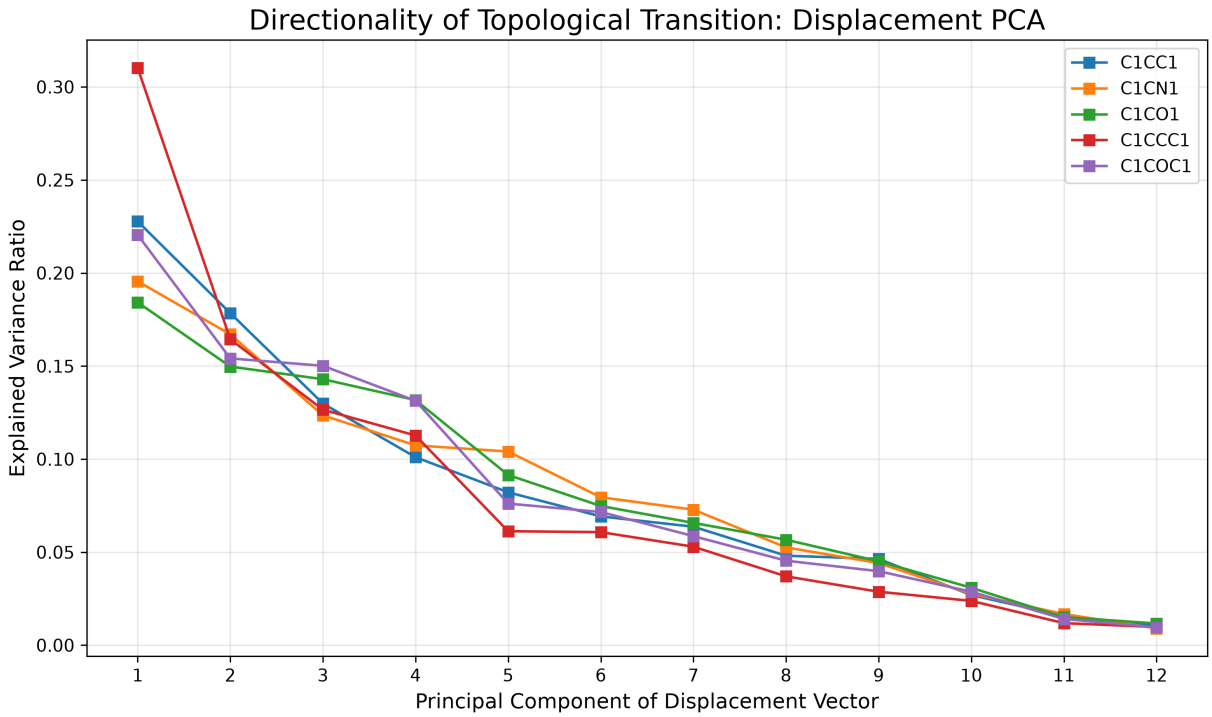

Figure 5: Scree plot detailing the principal component analysis (PCA) of optimal transport displacement vectors for five cyclic scaffolds (`C1CC1`, `C1CN1`, `C1CO1`, `C1CCC1`, and `C1COC1`). The explained variance ratio across the 12 principal components reveals the degree of coordination in the property manifold deformation during topological transitions from acyclic to cyclic structures. The absence of a single dominant axis demonstrates that topological rewiring is a complex, high-dimensional rotation and expansion rather than a simple scaling. Specifically, the first principal component explains only 31.0% of the variance for `C1CCC1` and 18.4% for `C1CO1`, with at least three components required to capture approximately 50% of the variance across all scaffolds. This multi-axial shift confirms that cyclization fundamentally rewires the relationships between several molecular properties simultaneously.

3.3.2 Frontier orbital stability

In contrast to the spatial properties, electronic features such as HOMO and LUMO energy residuals contribute significantly less to the topological shift. Even in nitrogen-containing cycles (`C1CN1`), where electronic effects are expected to be more pronounced, the combined contribution of HOMO, LUMO, and the dipole moment (μ) is only 21.9%. This confirms that frontier orbital energies are largely decoupled from the global spatial and thermodynamic reorganization induced by cyclization.

3.4 Directionality and coherence of the shift

To determine whether the manifold deformation represents a simple scaling or a more complex structural reorganization, we performed a Principal Component Analysis (PCA) on the optimal transport displacement vectors (Figure 5).

The PCA reveals that the displacement is not dominated by a single axis, indicating that topological rewiring is a complex, high-dimensional rotation and expansion rather than a simple shift along a size-related vector. For the `C1CCC1` scaffold, the first principal component (PC1) explains only 31.0% of the displacement variance, while for `C1CO1`, PC1 explains a mere 18.4%. Across all scaffolds, at least three principal components are required to capture approximately 50% of the variance. This multi-axial shift confirms that cyclization does not merely translate the property space; rather, it fundamentally rewires the coupling between several properties simultaneously, particularly the relationship between spatial extent and

rotational response.

4 Conclusions

In this study, we addressed the longstanding challenge of isolating the intrinsic physical effects of molecular topology from the overwhelming influence of stoichiometric determinism. By decoupling the molecular property manifold from basic compositional metrics using gradient-boosted regression, we established a clean, stoichiometry-independent 12-dimensional state space. Within this residual manifold, we deployed optimal transport theory—specifically, bootstrapped Wasserstein-2 (W_2) alignment—to rigorously quantify the structural rewiring induced by the transition from acyclic baselines to cyclic scaffolds in the QM9 dataset.

Our results demonstrate that cyclization induces a significant and robust deformation of the molecular property manifold. Through a marginal decomposition of the optimal transport costs, we identified that this topological rewiring is overwhelmingly driven by geometric and spatial response properties. Specifically, rotational constants, polarizability, and the electronic spatial extent (r^2) account for 60 to 70 percent of the total manifold deformation. The electronic spatial extent emerged as the single most dominant driver, indicating that the primary physical consequence of spatial confinement is a fundamental reorganization of the electronic density distribution and the molecular moment of inertia. In contrast, frontier orbital energies such as the HOMO and LUMO remain largely decoupled from this global spatial reorganization.

Furthermore, principal component analysis of the optimal transport displacement vectors revealed that topological transitions do not merely scale the property space along a single size-related axis. Instead, cyclization manifests as a complex, high-dimensional rotation and expansion, requiring at least three principal components to capture half of the displacement variance. This multi-axial shift confirms that structural rewiring fundamentally alters the coupled relationships between multiple physical properties simultaneously.

Finally, our comparative analysis between the full dataset and a rigorously filtered subset highlighted the critical importance of data integrity in computational chemistry. Unflagged geometric inconsistencies were shown to systematically inflate perceived topological shifts, underscoring the necessity of cross-validated reconstruction error filtering when mapping delicate manifold deformations.

Ultimately, this work provides a mathematically rigorous, distribution-based framework for quantifying molecular physical states. By stripping away stoichiometric artifacts and utilizing exact optimal transport metrics, we offer a robust new paradigm for understanding how chemical architecture and spatial confinement dictate the joint physical properties of molecules.