Gromov-Wasserstein Alignment of Distance Correlation Networks to Quantify Heteroatom-Driven Property Rewiring

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular properties are typically modeled as independent targets, obscuring the highly structured joint manifold they form. Understanding how topological changes and heteroatom substitutions reorganize this underlying property network is challenging due to the dominant deterministic influence of stoichiometry. To address this, we propose a coordinate-invariant framework to quantify the deformation of molecular property manifolds across structural boundaries. Using the QM9 dataset, we first neutralize stoichiometric determinism via non-linear regression on exact atom counts to isolate intrinsic physical couplings. We then construct non-linear dependency networks for discrete structural classes, such as acyclic chains and heteroatom-doped cyclic systems, using Distance Correlation matrices. By applying Gromov-Wasserstein optimal transport to these scaffold-specific matrices, we compute exact manifold alignment costs to measure how electronegative doping and spatial confinement rewire the thermodynamic and geometric state space. Our results reveal that the most severe manifold reorganization occurs during transitions from large cyclic systems to acyclic chains coupled with fluorine substitutions. Decomposing the optimal transport plans identifies isotropic polarizability and HOMO energy as the most structurally volatile properties, whereas the dipole moment remains highly conserved. Additionally, specific structural shifts target distinct properties, such as acyclic-to-cyclic transitions rewiring the rotational constant and heteroatom doping altering internal energy couplings. Scaffold-based cross-validation confirms the statistical robustness of these topological deformations, providing a rigorous approach for understanding how molecular structure dictates joint property dependencies.

1 Introduction

Molecular properties are typically modeled as independent targets for prediction, an approach that obscures the highly structured joint manifold they form. In reality, a molecule's thermodynamic, geometric, and electronic characteristics are deeply interconnected, forming a tightly coupled vibrational-spatial-rotational state space. Understanding how this underlying property network reorganizes in response to topological changes—such as ring closure—and heteroatom substitutions is a fundamental challenge in chemical physics.

1.1 The challenge of stoichiometric determinism

Isolating the intrinsic physical couplings between molecular properties is notoriously difficult due to the dominant influence of stoichiometry. The exact counts of constituent atoms (such as carbon, nitrogen, oxygen, fluorine, and hydrogen) deterministically drive the bulk of the variance in extensive and semi-extensive properties. For example, in the QM9 dataset, stoichiometric variables explain near-total variance in internal energy ($R^2 = 1.0000$), zero-point vibrational energy ($R^2 = 0.9971$), and atomization energies. This stoichiometric determinism masks the subtle, non-linear dependencies that arise from structural isomerism and scaffold architecture.

To rigorously study how specific heteroatom substitutions and spatial confinements reorganize internal molecular physics, one must first neutralize this stoichiometric scaling without making unverified causal assumptions. Only by analyzing the residuals of these properties conditioned on exact atom counts can the true physical couplings be revealed, such as the strong non-linear dependency ($dCor = 0.8365$) between residual zero-point vibrational energy and residual heat capacity.

1.2 Coordinate-invariant manifold alignment

Once stoichiometric effects are removed, quantifying the deformation of the property manifold across different structural boundaries presents a secondary mathematical challenge. Traditional distance metrics operate within fixed global coordinate spaces, which are ill-suited for comparing distinct chemical families where the fundamental relational structure of the properties has shifted.

To address this, we propose a coordinate-invariant framework to quantify the topological deformation of molecular property manifolds. We partition the deduplicated QM9 dataset into discrete structural classes based on heavy-atom count, topology (acyclic chains versus cyclic systems), and heteroatom doping (e.g., nitrogen- or fluorine-doped systems). Within each chemical class, we construct non-linear dependency networks using Distance Correlation matrices. Because Distance Correlation captures both linear and non-linear associations, it provides a robust representation of the intra-group dependency structure.

We then deploy entropic regularized Gromov-Wasserstein optimal transport directly on these scaffold-specific correlation matrices. The Gromov-Wasserstein distance computes the exact manifold alignment cost across structural boundaries by comparing intra-space relational geometries rather than absolute coordinates. This provides a mathematically rigorous metric of how electronegative doping and spatial confinement fundamentally rewire the property state space. Furthermore, this relational approach inherently helps isolate chemically distinct regimes, including latent geometric consistency artifacts present in the dataset.

1.3 Overview of findings

By applying this framework, we systematically map the rewiring of the molecular property manifold. Our analysis reveals that the most severe manifold reorganization occurs during transitions from large cyclic systems to acyclic chains coupled with fluorine substitutions. By decomposing the optimal transport plans, we identify which specific properties are most susceptible to structural shifts. We find that isotropic polarizability (α) and HOMO energy are the most structurally volatile properties, whereas the dipole moment (μ) remains highly conserved across chemical families.

Additionally, we demonstrate that specific structural shifts target distinct physical properties. For instance, acyclic-to-cyclic transitions primarily rewire the rotational constant, reflecting dramatic shifts in molecular symmetry, while transitions between nitrogen- and fluorine-doped systems predominantly alter internal energy couplings. Scaffold-based cross-validation confirms the statistical robustness of these topological deformations, establishing Gromov-Wasserstein alignment as a powerful tool for quantifying how molecular structure dictates joint property dependencies.

2 Methods

2.1 Data curation and stoichiometric neutralization

To isolate the intrinsic physical couplings between molecular properties from the dominant effects of stoichiometry, we first established a rigorously filtered and neutralized dataset. The QM9 dataset was ingested and deduplicated based on SMILES strings, removing 83 duplicate structures to prevent data leakage. To ensure full-rank covariance structures and focus the analysis on genuine chemical relationships rather than thermodynamic identities, we reduced the 19 available numeric columns to 12 non-redundant properties. Specifically, we excluded redundant thermodynamic features at room temperature and atomic reference energies. The rotational constant A exhibited extreme skewness and exact zeros for certain geometries; this was corrected using a $\log(1 + A)$ transformation.

As motivated in the introduction, stoichiometric variables deterministically drive the bulk of the variance in extensive properties. To neutralize this stoichiometric determinism, we applied non-linear gradient-boosted regression to predict each of the 12 retained properties strictly from exact atom counts (C, N, O, F, H). The residuals from these regression models represent the purely structural and topological variance of the properties. These isolated residuals formed the basis for all subsequent manifold constructions.

2.2 Structural stratification and artifact flagging

To evaluate manifold deformation across distinct structural boundaries, the dataset was stratified into discrete chemical families. Using RDKit, we pre-computed exact atom counts, heavy atom counts, and Murcko scaffolds for all molecules. The dataset was first binned by heavy-atom count (ranging from 6 to 9). Within each size bin, molecules were partitioned into mutually exclusive topological families: acyclic chains, small rings (3 to 4 atoms), and larger rings (5 to 9 atoms). These topological classes were further subdivided by heteroatom doping into nitrogen-doped, fluorine-doped, and pure-CHO systems. Any sparse classes were merged to enforce a strict minimum sample size threshold of $N > 500$ per stratum, ensuring robust estimation of the dependency networks.

Additionally, we addressed latent geometric consistency failures present in the dataset. We approximated these artifacts by fitting a robust linear model predicting atomization energy from atom counts. Molecules exhibiting a robust Z-score > 3 on the residuals were flagged. Rather than discarding these approximately 3,054 molecules, we retained the inconsistency flag as a covariate to explicitly evaluate its impact on the joint property structure, treating geometric artifacts as a chemically distinct regime.

2.3 Construction of distance correlation networks

Within each structural stratum, we constructed non-linear dependency networks to capture the joint property manifold. To ensure that dependency metrics reflected true coupling strength rather than marginal distribution shapes, we applied a rank-based quantile normalization to the residual properties, mapping them to a Gaussian distribution strictly within their respective strata.

For linear molecules, zero-variance properties (such as the rotational constant A) or perfectly collinear features (where rotational constants B and C are identical) were dynamically dropped for that specific stratum to prevent numerical instability. We then computed the pairwise Distance Correlation ($dCor$) matrix among the transformed properties. Distance Correlation was selected because it captures both linear and non-linear associations, providing a comprehensive representation of the intra-group dependency structure. These computationally intensive matrix calculations were parallelized across all strata.

2.4 Gromov-Wasserstein manifold alignment

To quantify how the joint property structure rewires across different chemical families within the same heavy-atom size bin, we employed optimal transport. The intra-stratum $dCor$ matrices were first transformed into dissimilarity matrices, defined as $1 - dCor^2$.

We computed the entropic regularized Gromov-Wasserstein (GW) distance between the dissimilarity matrices of different structural strata. Because the GW framework operates on intra-space relational geometries rather than absolute coordinates, it provides a coordinate-invariant metric of manifold deformation. The entropic regularization parameter (ϵ) was tuned based on the scale of the cost matrix to ensure a stable transport plan. The resulting GW distance was normalized by the total variance of the properties in the target stratum, yielding a relative rewiring intensity that quantifies the global alignment cost across structural boundaries.

2.5 Transport plan decomposition and cross-validation

To identify which specific properties drive the topological deformation, we extracted and decomposed the optimal transport plan (the coupling matrix) from the GW computation. We calculated the marginal contribution of each property to the total GW distance. A "rewiring intensity" was defined for each property as the sum of the mass mapped off the diagonal in the transport plan, where a value of 1.0 indicates complete structural role-reversal and 0.0 indicates perfect conservation. For the top-ranked rewired properties, we calculated the signed shift in correlation ($\Delta dCor$) to determine the directionality and physical nature of the coupling changes.

Finally, the statistical robustness of the $dCor$ networks and GW transport plans was validated using a leave-k-scaffolds-out cross-validation strategy. During each iteration, 10% to 20% of the Murcko scaffolds within a stratum were held out. The intra-stratum $dCor$ matrices, GW distances, and transport plans were recomputed to establish 95% confidence intervals for the alignment costs and off-diagonal transport masses,

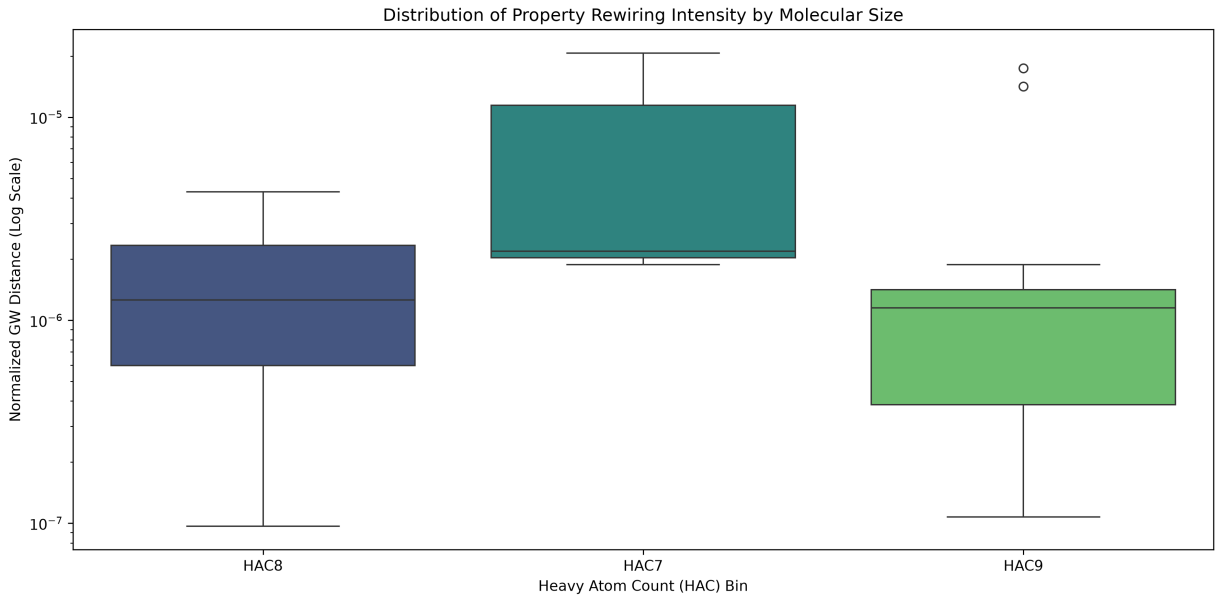

Figure 1: Distribution of property rewiring intensity, quantified by the normalized Gromov-Wasserstein distance (d_{GW}) on a logarithmic scale, across different molecular sizes denoted by Heavy Atom Count (HAC) bins. The box plots reveal the extent of manifold deformation resulting from structural transformations within each size category. Molecules with seven heavy atoms (HAC7) display the highest median rewiring intensity and broadest distribution, indicating that their property-dependency networks undergo more severe reorganization across structural boundaries compared to the HAC8 and HAC9 bins. This demonstrates that while deterministic stoichiometric effects are neutralized prior to alignment, the susceptibility of the underlying physical couplings to topological rewiring remains highly dependent on the overall molecular size.

ensuring that the observed manifold deformations were driven by fundamental structural shifts rather than idiosyncratic molecular scaffolds.

3 Results

3.1 Topological deformation of the property manifold

Following the neutralization of stoichiometric determinism detailed in the methods, the underlying physical couplings between molecular properties become quantifiable. The apparent joint structure in raw molecular data is heavily dominated by molecular size, with stoichiometric counts explaining near-total variance in internal energy (u_0, $R^2 = 1.0000$), zero-point vibrational energy (ZPVE, $R^2 = 0.9971$), and atomization energies ($R^2 \approx 0.989$). By analyzing the residuals of these properties conditioned on exact atom counts, we isolate the intrinsic non-linear dependencies. For instance, residual ZPVE and residual heat capacity (c_v) exhibit a strong non-linear dependency ($dCor = 0.8365$), reflecting a vibrational-spatial axis that persists independently of stoichiometric scaling.

To understand how this joint structure reorganizes across structural boundaries, we computed the Gromov-Wasserstein (GW) distance (d_{GW}) between the dependency networks of different chemical families. The distribution of property rewiring intensity across different molecular sizes (Figure 1) reveals that molecules with seven heavy atoms (HAC7) display the highest median rewiring intensity and the broadest distribution of manifold deformation. This indicates that the property-dependency networks of HAC7 systems undergo more severe reorganization across structural boundaries compared to larger HAC8 and HAC9 systems.

The magnitude of manifold deformation depends significantly on the specific nature of the structural shift (Figure 2B). The most severe reorganization occurs during the transition from large cyclic systems to

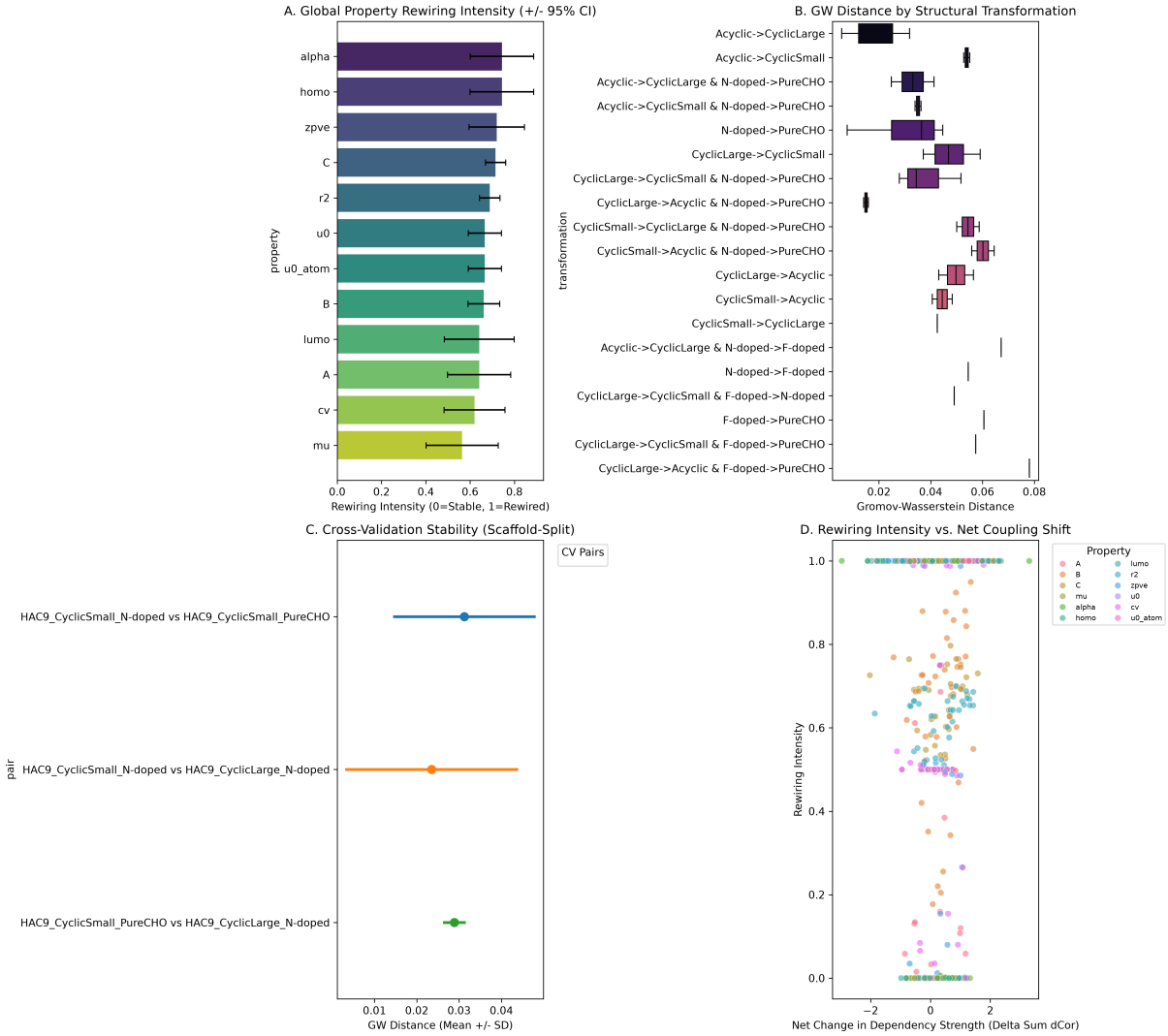

Figure 2: Quantification of molecular property manifold rewiring using Gromov-Wasserstein (GW) optimal transport. (A) Global property rewiring intensity across structural pairs shows that isotropic polarizability (α) and HOMO energy are the most volatile components of the manifold, whereas the dipole moment (μ) remains the most stable. (B) GW distances (d_{GW}) for various structural transformations demonstrate that maximum manifold deformation occurs during transitions from large cyclic to acyclic systems paired with F-doped to pure-CHO substitutions ($d_{GW} = 0.0780$), while purely heteroatom changes exhibit greater resilience. (C) Scaffold-split cross-validation confirms the stability of these GW distances across top strata, including `HAC9_CyclicSmall_N-doped` and `HAC9_CyclicSmall_PureCHO`. (D) Rewiring intensity compared to the net change in dependency strength ($\Delta dCor$) illustrates systemic shifts in property couplings across structural boundaries.

acyclic chains, particularly when paired with a shift from fluorine-doped to pure-CHO compositions, yielding a maximum alignment cost of $d_{GW} = 0.0780$. Purely topological changes, such as the transition from acyclic chains to small-ring cyclic systems, result in consistent manifold deformation with a mean GW distance of 0.0539 ± 0.0146 (95% confidence interval). Conversely, transformations involving purely heteroatom changes, such as nitrogen-doped to pure-CHO systems, exhibit a lower alignment cost ($d_{GW} = 0.0316 \pm 0.0129$). This suggests that while nitrogen doping significantly influences absolute property values, it is less disruptive to the global dependency network than major topological shifts like ring closure.

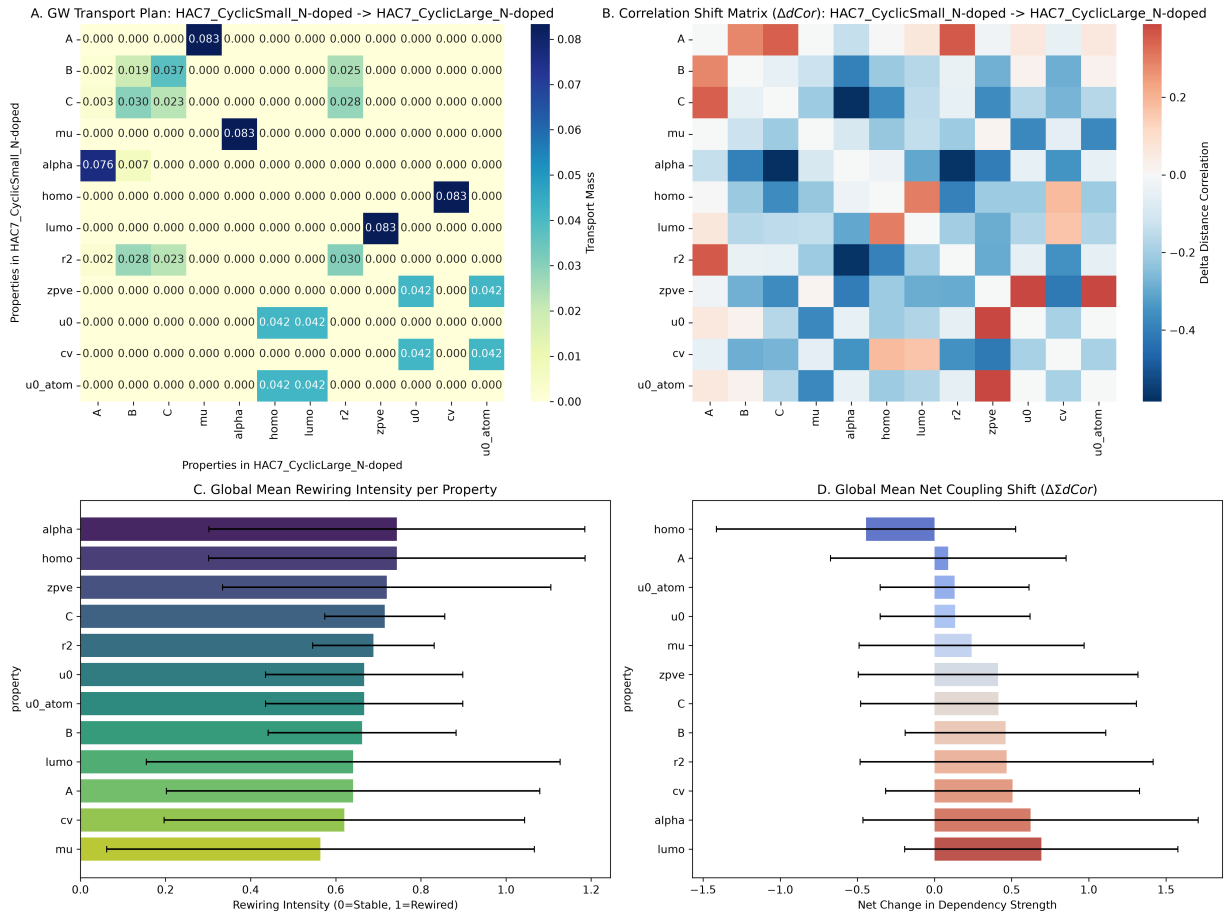

Figure 3: Decomposition of the Gromov-Wasserstein optimal transport plan and property rewiring across structural boundaries. Panels A and B detail the transport mass and correlation shift matrix ($\Delta dCor$) for the transition from `HAC7_CyclicSmall_N-doped` to `HAC7_CyclicLarge_N-doped`, revealing a systemic weakening of geometric-rotational coupling—specifically between spatial extent (r^2) and rotational constants (B, C)—as rings become larger and more flexible. Panels C and D display the global mean rewiring intensity and net coupling shifts across structural pairs, demonstrating that isotropic polarizability (α) and HOMO energy are the most volatile components of the manifold, whereas the dipole moment (μ) maintains the most stable relationship with the rest of the network.

3.2 Decomposing the transport plan and property stability

By decomposing the optimal transport plan between these manifolds, we can identify which specific properties are most susceptible to role-reversal or coupling shifts. We define "rewiring intensity" based on the off-diagonal mass in the transport plan, where a value of 1.0 indicates complete structural role-reversal and 0.0 indicates perfect conservation.

Aggregated across 39 structural pairs, isotropic polarizability (α) and HOMO energy emerge as the most volatile components of the manifold, both exhibiting a mean rewiring intensity of 0.744 ± 0.143 (Figure 2A, Figure 3C). ZPVE also shows high volatility with a mean intensity of 0.720 ± 0.125. In contrast, the dipole moment (μ) maintains the most stable relationship with the rest of the manifold, showing a mean intensity of 0.564 ± 0.163. This indicates that the structural role of the dipole moment is highly conserved across chemical families, whereas geometric and electronic response properties are highly sensitive to structural boundaries.

Different structural changes systematically target distinct physical properties. Transformations involving acyclic to cyclic transitions primarily rewire the rotational constant A. This reflects the dramatic shift in

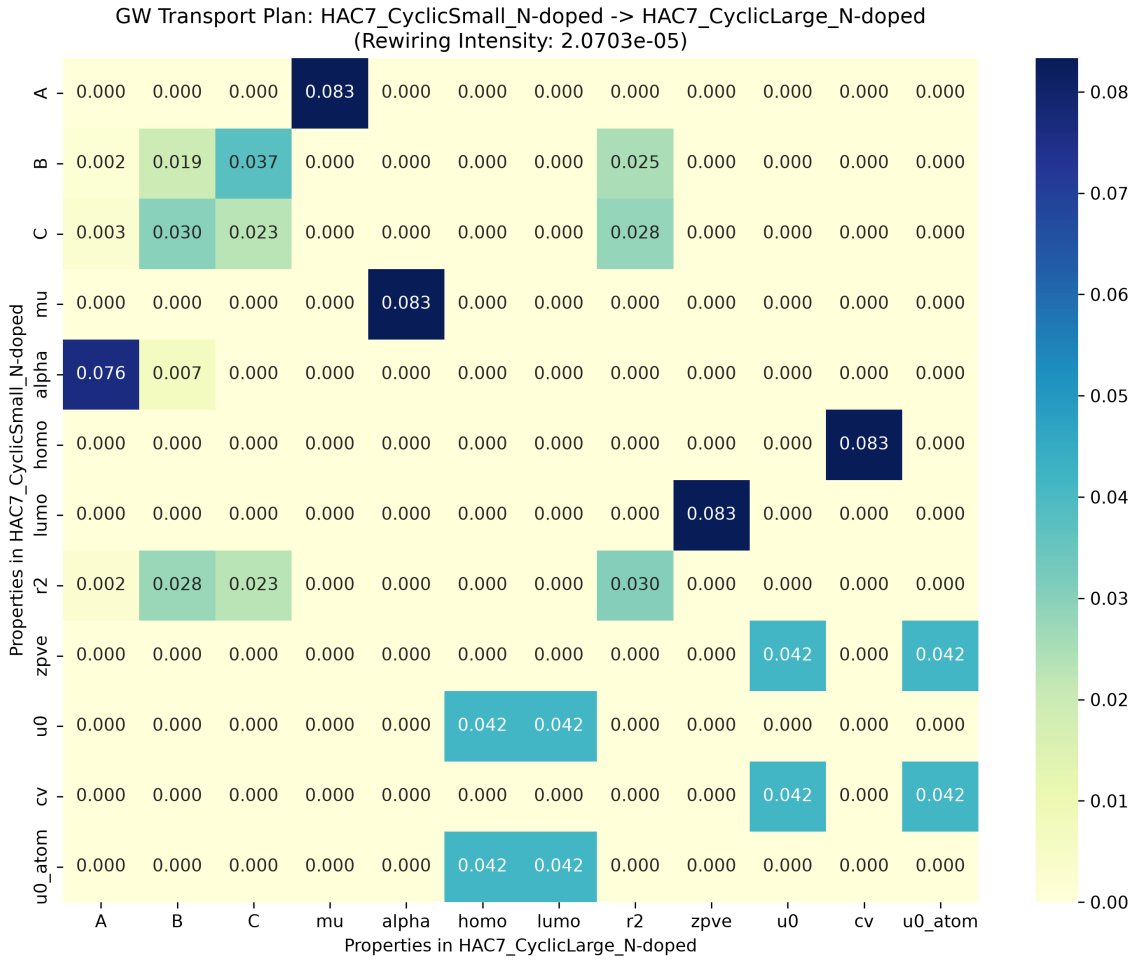

Figure 4: Heatmap of the Gromov-Wasserstein optimal transport plan detailing the property manifold alignment between `HAC7_CyclicSmall_N-doped` and `HAC7_CyclicLarge_N-doped` molecular systems. By decomposing the transport plan, off-diagonal mass reveals which properties are most susceptible to role-reversal and coupling shifts across structural boundaries. Specifically, the cross-mapping between spatial extent (r^2) and rotational constants (B, C) illustrates a systemic weakening of geometric-rotational coupling as molecular rings become larger and more flexible.

molecular symmetry and the principal moment of inertia when an open, flexible chain closes into a rigid ring. Meanwhile, transitions between nitrogen-doped and fluorine-doped systems are dominated by the rewiring of internal energy (u_0). This is likely driven by the fundamental shift from the lone-pair donation characteristic of nitrogen to the high electronegativity and distinct bonding environments introduced by fluorine.

Detailed analysis of the most severely rewired pair within the HAC7 bin (`HAC7_CyclicSmall_N-doped` transitioning to `HAC7_CyclicLarge_N-doped`) highlights a systemic weakening of geometric-rotational coupling (Figure 4, Figure 3A,B). The correlation shift matrix ($\Delta dCor$) demonstrates that the coupling between spatial extent (r^2) and rotational constants (B, C) is significantly reduced as rings become larger and more flexible, with $\Delta dCor$ shifts reaching nearly -2.0.

3.3 Statistical robustness and geometric artifacts

To ensure that the detected manifold rewiring is a fundamental physical feature rather than an artifact of specific molecular families, we validated the results using scaffold-based cross-validation. Holding out 15%

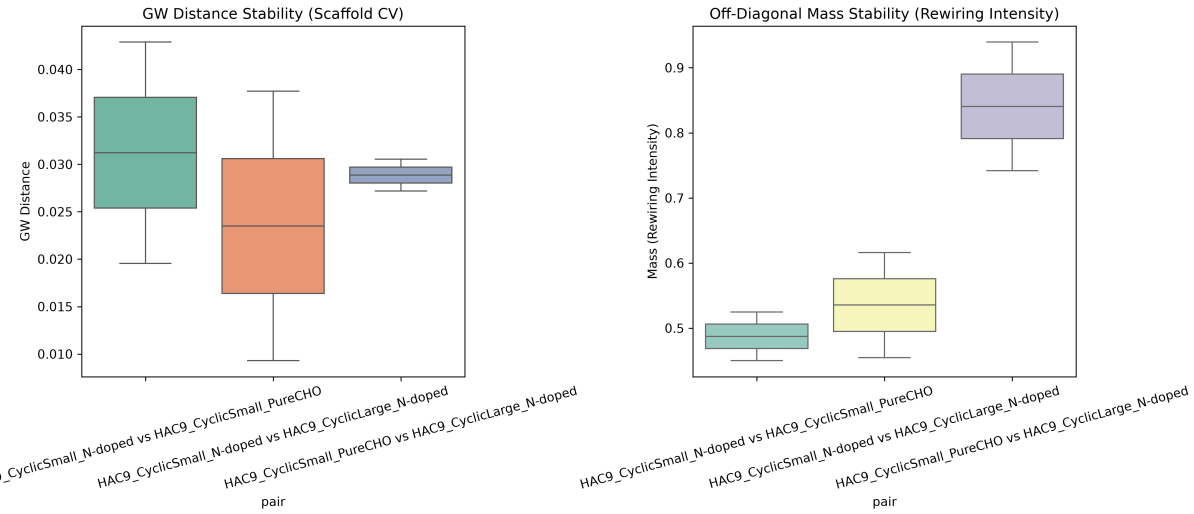

Figure 5: Scaffold-based cross-validation distributions demonstrating the stability of Gromov-Wasserstein (GW) distances (left) and off-diagonal mass, or rewiring intensity (right), when holding out 15% of Murcko scaffolds. The boxplots evaluate three structural transformation pairs among `HAC9_CyclicSmall_N-doped`, `HAC9_CyclicSmall_PureCHO`, and `HAC9_CyclicLarge_N-doped`. The left panel confirms the robustness of the manifold alignment with a global mean GW stability of 0.0279 ± 0.0122. The right panel reveals that the most substantial property rewiring consistently occurs between oxygen-heavy small rings and nitrogen-doped larger rings (`HAC9_CyclicSmall_PureCHO` vs. `HAC9_CyclicLarge_N-doped`), yielding a mean off-diagonal mass of 0.841 ± 0.140. These results confirm that the quantified property rewiring across structural boundaries is a stable physical feature rather than an artifact of specific molecular families.

of Murcko scaffolds per iteration confirms the stability of the computed GW distances (Figure 2C, Figure 5).

Across the top three strata—`HAC9_CyclicSmall_N-doped`, `HAC9_CyclicSmall_PureCHO`, and `HAC9_CyclicLarge_N-doped`—global mean GW stability was measured at 0.0279 ± 0.0122. The most substantial property rewiring was consistently observed between oxygen-heavy small rings and nitrogen-doped larger rings (`HAC9_CyclicSmall_PureCHO` versus `HAC9_CyclicLarge_N-doped`), yielding a mean off-diagonal mass of 0.841 ± 0.140. These tight confidence intervals demonstrate that the topological deformations captured by the optimal transport plan are statistically robust.

Finally, we evaluated the impact of latent geometric consistency artifacts, which were flagged during the initial data curation. A comparison of the distance correlation matrices between flagged and unflagged molecules revealed substantial structural divergence. The Frobenius norm differences between the dependency networks of these subsets were 1.640 for `HAC8_CyclicSmall_N-doped` and 1.249 for `HAC9_CyclicSmall_N-doped`. These large deviations confirm that molecules with inconsistent geometries do not merely represent random noise; rather, they constitute a distinct physical regime characterized by highly anomalous dependency patterns that fundamentally disrupt the expected joint property manifold.

4 Conclusions

In this study, we introduced a coordinate-invariant framework to quantify the topological deformation of molecular property manifolds across structural boundaries. By first neutralizing the deterministic influence of stoichiometry through non-linear regression on exact atom counts, we successfully isolated the intrinsic physical couplings that define a molecule's vibrational-spatial-rotational state space. Applying Gromov-Wasserstein optimal transport to scaffold-specific distance correlation networks allowed us to measure exact manifold alignment costs, providing a mathematically rigorous perspective on how molecular properties are jointly reorganized by topological and chemical shifts.

Our results demonstrate that the magnitude of manifold deformation is highly dependent on the nature of the structural transformation. The most severe reorganization of the property network occurs during transitions from large cyclic systems to acyclic chains, particularly when coupled with fluorine substitutions. In contrast, purely heteroatom-driven changes, such as the transition from nitrogen-doped to pure-CHO systems, exhibit significantly lower alignment costs. This indicates that while heteroatom doping alters absolute property values, major topological shifts like ring closure and spatial confinement are the primary drivers of global dependency network rewiring.

By decomposing the optimal transport plans, we identified the specific physical properties most susceptible to structural shifts. Isotropic polarizability (α) and HOMO energy emerged as the most structurally volatile components of the manifold, whereas the dipole moment (μ) remained highly conserved across diverse chemical families. Furthermore, we established that distinct structural transformations target specific physical properties: acyclic-to-cyclic transitions predominantly rewire the rotational constant due to dramatic shifts in molecular symmetry, while transitions between nitrogen- and fluorine-doped systems fundamentally alter internal energy couplings. We also observed a systemic weakening of geometric-rotational coupling as molecular rings increase in size and flexibility.

The statistical robustness of these findings was confirmed through rigorous scaffold-based cross-validation, ensuring that the observed manifold deformations reflect fundamental physical principles rather than idiosyncratic molecular artifacts. Additionally, our relational framework successfully isolated latent geometric consistency failures, demonstrating that such artifacts constitute a chemically distinct regime with anomalous dependency patterns rather than mere random noise.

Ultimately, this work highlights the limitations of modeling molecular properties as independent targets. By mapping the exact topological deformations of the joint property manifold, we provide a robust foundation for understanding how molecular structure dictates complex physical dependencies. This coordinate-invariant approach offers new avenues for structure-property modeling, ensuring that the deeply interconnected nature of molecular physics is preserved and rigorously quantified.