Optimal Transport Quantification of Topological Rewiring in Molecular Vibrational-Spatial Distributions

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular properties are heavily dominated by stoichiometric scaling, which obscures the intrinsic influence of chemical architecture—such as cyclic topology and ring strain—on a molecule's spatial and thermodynamic states. To isolate these structural effects, we analyze the QM9 dataset by training gradient-boosted models on acyclic molecules to predict electronic spatial extent, zero-point vibrational energy, and heat capacity solely from atom counts. Extracting and standardizing these residuals removes stoichiometric determinism, establishing an unconstrained acyclic baseline. We then employ optimal transport, computing the exact Wasserstein distance to measure the macroscopic deformation between the three-dimensional residual distributions of cyclic scaffolds and their stoichiometry-matched acyclic counterparts. Additionally, we evaluate the Frobenius norm of covariance difference matrices to quantify the internal rewiring of property correlations. Our results demonstrate that removing stoichiometric effects reveals a high-dimensional chemical manifold where cyclic topology systematically displaces the joint property distribution. We identify ring count as the primary driver of this manifold displacement, while local geometric constraints in highly strained or heteroatom-rich rings fundamentally rewire the natural correlations between vibrational energy and heat capacity, often inverting their relationship. Ultimately, this framework provides a precise, quantitative measure of how spatial confinement reorganizes the internal thermodynamic and spatial states of molecules.

1 Introduction

The physical and thermodynamic properties of molecules are heavily dominated by stoichiometric scaling. Extensive properties, such as internal energy and zero-point vibrational energy ($zpve$), are almost entirely determined by simple atom counts, often exhibiting near-perfect linear correlations with molecular size. While this stoichiometric determinism is a fundamental aspect of chemistry, it obscures the intrinsic influence of chemical architecture—such as cyclic topology, spatial confinement, and ring strain—on a molecule's spatial and thermodynamic states. Understanding how these structural features independently modulate molecular physics requires disentangling the effects of pure topology from the overwhelming signal of molecular composition.

1.1 Isolating topological effects

To isolate the genuine effects of chemical architecture, it is necessary to establish a baseline of unconstrained molecular behavior. Acyclic molecules, free from the geometric restrictions of ring closures, provide a natural reference state. By modeling the properties of acyclic molecules solely as a function of their atom counts, we can predict the expected stoichiometric contribution to any molecule's properties. The residuals of these predictions—the deviations from the acyclic baseline—represent the isolated variance attributable to structural connectivity and spatial confinement.

This study focuses on a tightly coupled triad of properties: electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and heat capacity (c_v). Together, these define a three-dimensional vibrational-spatial probability distribution. In unconstrained acyclic chains, these properties exhibit natural, extensive-property-driven couplings. However, the introduction of cyclic topology imposes geometric constraints that fundamentally reorganize this joint distribution, altering both its location in the chemical manifold and its internal correlation structure.

1.2 Optimal transport and covariance rewiring

Quantifying the deformation of this multidimensional distribution requires robust mathematical tools capable of capturing both macroscopic shifts and internal structural changes. We employ optimal transport to compute the exact Wasserstein distance between the standardized 3D residual distributions of cyclic scaffolds and their stoichiometry-matched acyclic counterparts. By measuring the minimum cost required to transform the unconstrained acyclic distribution into the constrained cyclic state, the Wasserstein distance provides a direct, quantitative metric of macroscopic manifold displacement driven by topological strain.

Furthermore, we evaluate the Frobenius norm of covariance difference matrices to capture the internal rewiring of property correlations. While the Wasserstein distance measures the global shift of the property manifold, the covariance differential quantifies how the relationships between spatial and thermodynamic states are fundamentally altered—and sometimes inverted—by local geometric constraints.

Through this framework, we demonstrate that removing stoichiometric effects reveals a high-dimensional chemical manifold where cyclic topology systematically displaces the joint property distribution. We identify macroscopic topological features, primarily ring count, as the main drivers of this displacement, while local geometric constraints in highly strained or heteroatom-rich rings are shown to fundamentally rewire the natural correlations between vibrational energy and heat capacity. Ultimately, this approach provides a precise measure of how spatial confinement reorganizes the internal thermodynamic and spatial states of molecules.

2 Methods

2.1 Data preparation and baseline initialization

To isolate the intrinsic effects of chemical architecture from stoichiometric determinism, we established an unconstrained baseline using the acyclic population of the QM9 dataset. Initial data ingestion included the suppression of RDKit logging to ensure clean processing. We trained gradient-boosting models, specifically LightGBM, exclusively on the 13,998 acyclic molecules to predict three tightly coupled properties: electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and heat capacity (c_v). The models utilized only the exact atom counts of carbon, nitrogen, oxygen, fluorine, and hydrogen as input features.

By applying these trained models to the entire dataset, we extracted the residuals for each property (res_r2, res_zpve, res_cv). These residuals represent the true deviations from expected acyclic behavior. To ensure that subsequent optimal transport costs reflect genuine shifts in distribution shape and location rather than artifactual variance, the residuals were globally standardized using the mean and standard deviation of the entire dataset prior to any scaffold partitioning.

2.2 Scaffold stratification and composition matching

Molecules were deduplicated and partitioned by their Murcko scaffolds. To establish a rigorous unconstrained reference manifold for each cyclic scaffold, we matched cyclic populations to acyclic baselines using a stratified sampling approach. The primary matching constraint was the exact total number of heavy atoms. For any remaining discrepancies, we minimized the Euclidean distance in the five-dimensional composition space defined by the atom counts:

$$d_{comp} = \sqrt{\Delta N_C^2 + \Delta N_N^2 + \Delta N_O^2 + \Delta N_F^2 + \Delta N_H^2} \tag{1}$$

This strict matching ensures that for every cyclic scaffold, the corresponding acyclic baseline possesses a nearly identical stoichiometric distribution, preventing downstream transport costs from being dominated by simple size differences.

As a data sanitization and sensitivity step, we evaluated the skewness and kurtosis of the residual distributions to rule out higher-order moment biases. Additionally, for the 9 linear molecules in the dataset with an undefined rotational constant A, we imputed the value using the maximum observed A in the dataset. This confirmed manifold stability at the physical limit of zero-dimensional rotation without introducing artificial clustering.

2.3 Optimal transport and Wasserstein displacement

To quantify the macroscopic deformation of the vibrational-spatial distribution under cyclic strain, we employed optimal transport. Using the `POT` library, we computed the exact Wasserstein distance (W) between the globally standardized 3D residual distributions of the cyclic scaffolds (P) and their matched acyclic baselines (Q). The Wasserstein distance is defined as:

$$W(P,Q) = \inf_{\gamma \in \Pi(P,Q)} \int_{\mathbb{R}^3 \times \mathbb{R}^3} \|\mathbf{x} - \mathbf{y}\| \, d\gamma(\mathbf{x}, \mathbf{y}) \tag{2}$$

where $\Pi(P,Q)$ denotes the set of all joint distributions γ with marginals P and Q, and $\mathbf{x}, \mathbf{y}$ are vectors in the 3D residual space. Computations were parallelized across 14 CPUs by scaffold. For the largest scaffolds where the exact Sinkhorn solver approached memory limits, `ot.sliced_wasserstein_distance` was utilized as a fallback.

To isolate the physical direction of the manifold shift, the resulting Wasserstein displacement vectors were projected onto the first three principal components of the acyclic baseline. This projection provides an interpretable coordinate system to distinguish between spatial expansion and vibrational stiffening.

2.4 Covariance rewiring and topological regression

Beyond macroscopic displacement, we quantified the internal rewiring of property correlations by computing the direct difference between the covariance matrices of the cyclic sets (Σ_{cyclic}) and their matched acyclic baselines ($\Sigma_{acyclic}$). We evaluated the Frobenius norm of this covariance difference matrix to serve as a single scalar summary of the rewiring magnitude:

$$||\Delta\Sigma||_F = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\Sigma_{cyclic}^{(i,j)} - \Sigma_{acyclic}^{(i,j)}\right)^2} \tag{3}$$

This metric complements the Wasserstein distance by explicitly capturing how cyclic topology introduces competing constraints that dilute or invert the extensive-property-driven couplings found in acyclic chains.

To identify the structural drivers of these deformations, we regressed the unexplained variance of the Wasserstein costs against explicit topological covariates, utilizing RDKit-derived metrics such as the number of rings and the number of aromatic rings as proxies for local geometry and ring strain. Finally, we isolated the top percentile of scaffolds exhibiting the highest Wasserstein costs. For these highly strained outliers, we performed a targeted extraction of their c_v to *zpve* residual ratios to chemically validate the macroscopic deformation and ground the optimal transport metrics in concrete thermodynamic anomalies. All displacement vectors, covariance differentials, and outlier metrics were exported to structured Parquet files for downstream visualization and analysis.

3 Results

Molecular properties in the QM9 dataset are fundamentally constrained by stoichiometry. Extensive thermodynamic properties such as internal energy (u_0) and zero-point vibrational energy (*zpve*) are almost entirely determined by atom counts ($R^2 = 0.9957$ and $R^2 = 0.9995$, respectively). This analysis identifies the joint structure of molecular properties that persists after these stoichiometric effects are removed, specifically examining how cyclic topology and ring strain reorganize the coupled distributions of spatial and vibrational states.

3.1 Residual joint structure and dimensionality

By conditioning 12 non-redundant QM9 properties on an atom-count baseline (C, N, O, F, H) using gradient-boosted models trained on the 13,998 acyclic molecules, we isolate the variance attributable to chemical architecture. The resulting residuals exhibit a higher effective dimensionality than the raw properties; global principal component analysis (PCA) on these residuals shows that the first five components explain approximately 74% of the variance (PC1: 0.2589, PC2: 0.1702, PC3: 0.1220, PC4: 0.1073, PC5: 0.0835).

This suggests that removing the dominant stoichiometric scaling reveals a more complex, high-dimensional chemical manifold. Within this manifold, a primary joint structure exists between the residuals of electronic spatial extent (r^2), *zpve*, and heat capacity (c_v). These properties form a vibrational-spatial triad that is highly sensitive to topological constraints.

3.2 Quantifying manifold deformation via optimal transport

Using the exact Wasserstein distance (W) between the 3D residual distributions (`res_r2`, `res_zpve`, `res_cv`) of 500 cyclic scaffolds and their stoichiometry-matched acyclic baselines, we capture the total cost of deforming the unconstrained acyclic manifold into the constrained cyclic state. Regression analysis of the Wasserstein distance against topological covariates reveals that the number of rings is the primary driver of manifold displacement (coefficient $= 0.3456$, $t = 13.666$, $p < 0.001$, $R^2 = 0.3853$), as shown in . In contrast, the number of bridgehead atoms does not independently increase the Wasserstein distance ($p = 0.638$), suggesting that the macro-topology of ring count is more influential than local connectivity patterns in driving global distributional shifts.

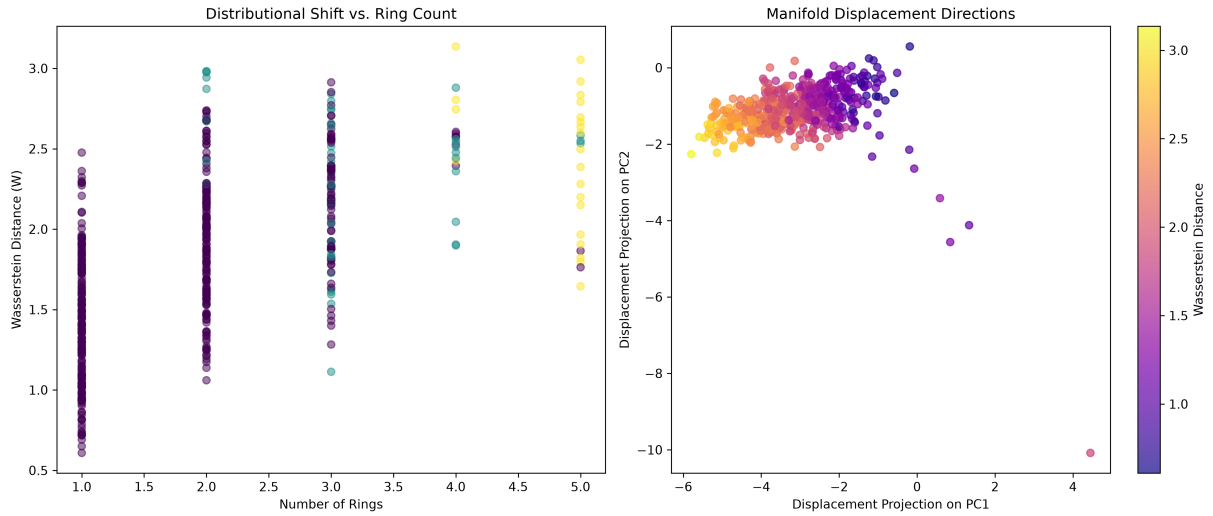

Figure 1: The left panel illustrates the relationship between the Wasserstein distance (W) and the number of rings, demonstrating that ring count is a primary driver of distributional shift between cyclic scaffolds and their stoichiometry-matched acyclic baselines. The right panel displays the projection of the 12D mean displacement vectors onto the first two principal components (PC1 and PC2), with points colored by W. Together, these visualizations reveal that cyclic topology induces a systematic manifold displacement; highly strained, multi-ring systems exhibit the largest distributional shifts ($W > 3.0$), and this topological constraint consistently drives the joint property distribution toward a distinct region of the chemical manifold characterized by negative PC1 and PC2 displacements.

The scaffolds exhibiting the largest distributional shifts ($W > 3.0$) are characterized by high ring counts and intense topological strain. Notable examples include the tetrahedrane-like scaffold `C1C2CC3C1C3C2` ($W = 3.136$), `C1C2CC3C1CC23` ($W = 3.054$), and `C1CC2CC1CO2` ($W = 2.983$).

The displacement of cyclic scaffolds relative to acyclic baselines is not random. Projecting the 12D mean displacement vectors onto the global PCA axes reveals a systematic migration (, right panel). For example, the cyclobutane scaffold (`C1CCC1`) exhibits a PC1 displacement of -1.819 and a PC2 displacement of -0.528. This consistent directional shift indicates that cyclic topology systematically moves the joint property distribution toward a distinct region of the chemical manifold characterized by negative PC1 and PC2 displacements.

3.3 Covariance rewiring and internal correlation shifts

Beyond macroscopic shifts in the mean, cyclic topology fundamentally alters the internal correlation structure of property sets. The Frobenius norm of the covariance difference matrix ($||\Delta\Sigma||_F$) between a cyclic scaffold and its matched acyclic baseline serves as a scalar measure of this rewiring. illustrates how both the global displacement and the internal covariance rewiring scale with ring count, highlighting that even single-ring systems can experience significant internal correlation shifts.

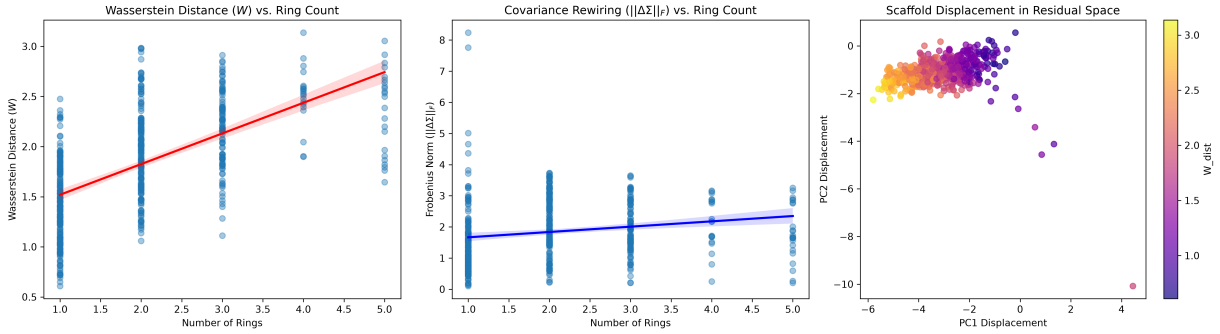

Figure 2: Topological drivers of manifold displacement and covariance rewiring in cyclic molecular scaffolds relative to stoichiometry-matched acyclic baselines. The left panel demonstrates that the Wasserstein distance (W) between the 3D residual distributions of spatial and vibrational properties (r^2, $zpve$, c_v) scales positively with ring count, identifying macro-topology as the primary driver of global distributional shifts. The middle panel illustrates internal covariance rewiring measured by the Frobenius norm ($||\Delta\Sigma||_F$), revealing that while rewiring generally increases with ring count, extreme internal correlation shifts occur even in single-ring systems like cyclohexane. The right panel projects the mean displacement vectors onto the first two principal components (PC1 and PC2), showing a systematic migration of cyclic scaffolds toward a distinct region of the chemical manifold characterized by negative displacements, with the color gradient indicating that larger displacements correspond to higher W. Ultimately, the figure reveals that cyclic topology reorganizes molecular physics through both predictable global displacement and complex internal rewiring of property correlations.

The global mean Frobenius norm across the analyzed scaffolds is 1.848 ± 1.024. Highly rewired scaffolds, such as cyclohexane (`C1CCCCC1`, $||\Delta\Sigma||_F = 8.237$), show significant reorganization of the property triad (). While the acyclic baseline typically shows tight positive coupling between vibrational and spatial residuals, cyclic scaffolds often exhibit shifts in these relationships. In the tetrazole scaffold (`n1nn[nH]n1`), the $zpve$-c_v covariance shifts by +2.608 relative to the acyclic baseline, indicating an intensification of this coupling under nitrogen-rich cyclic confinement.

3.4 Thermodynamic anomalies in outlier regimes

The top outliers for rewiring magnitude show extreme shifts in their property relationships. The tetrazole scaffold exhibits an average ratio of c_v to $zpve$ residuals of -0.961, indicating a stark decoupling or inversion relative to the positive correlation seen in typical acyclic chains. Similar inversions are observed in highly strained systems like `C1C2CC3C1C3C2` (ratio: -0.928). These shifts are visually apparent in the joint density distributions (), where cyclic distributions are compressed and rotated compared to the diffuse acyclic baselines.

3.5 Summary of joint structure reorganization

The transition from acyclic to cyclic topology in QM9 reorganizes molecular physics in two distinct ways. First, cyclic topology induces a global displacement, moving the entire joint distribution of electronic and vibrational properties toward a unique region of the residual manifold, a shift that scales primarily with ring count. Second, local geometric constraints—especially in small, strained, or heteroatom-rich rings—drive

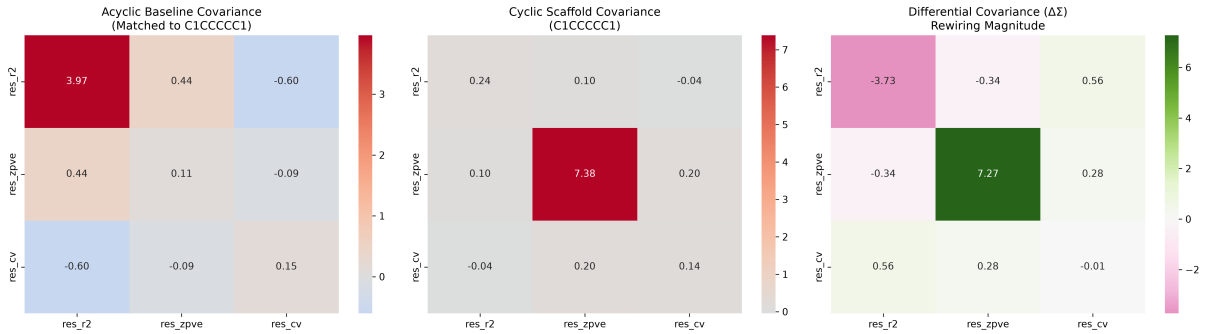

Figure 3: Covariance matrices of the residual distributions (`res_r2`, `res_zpve`, `res_cv`) for electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and heat capacity (c_v), comparing the cyclic scaffold cyclohexane (`C1CCCCC1`) to its stoichiometry-matched acyclic baseline. The differential covariance matrix ($\Delta\Sigma$) illustrates the magnitude of internal rewiring caused by cyclic topology. The plot reveals a significant reorganization of the vibrational-spatial property triad, where the cyclic constraint drastically increases the variance of $zpve$ while reducing the variance of r^2. This substantial shift, quantified by a high Frobenius norm ($||\Delta\Sigma||_F = 8.237$), demonstrates how local geometric constraints in cyclic structures fundamentally alter the internal correlation structure and thermodynamic couplings relative to unconstrained acyclic chains.

internal rewiring, fundamentally altering the natural correlations between vibrational energy and heat capacity. The magnitude of this rewiring represents a form of chemical strain that quantitatively differentiates cyclic families from their unconstrained acyclic counterparts.

4 Conclusions

The stoichiometric determinism inherent in molecular datasets often obscures the nuanced physical effects of chemical architecture. In this study, we introduced a quantitative framework to isolate and measure the impact of cyclic topology and spatial confinement on the coupled vibrational-spatial states of molecules. By establishing an unconstrained acyclic baseline and extracting standardized property residuals, we successfully removed the dominant scaling effects of atom counts, revealing a complex, high-dimensional chemical manifold that is highly sensitive to structural connectivity.

Applying optimal transport to this residual space allowed us to rigorously quantify the macroscopic deformation of the joint distribution of electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and heat capacity (c_v). Our computation of the exact Wasserstein distance between cyclic scaffolds and their stoichiometry-matched acyclic counterparts demonstrated that cyclic topology systematically displaces the property manifold. We identified macro-topology, specifically ring count, as the primary driver of this global distributional shift, consistently directing the joint property distribution toward a distinct, constrained region of the chemical space.

Furthermore, by evaluating the Frobenius norm of covariance difference matrices, we captured the profound internal rewiring of property correlations induced by cyclic confinement. While unconstrained acyclic chains exhibit natural, extensive-property-driven couplings, local geometric constraints in cyclic structures fundamentally alter these relationships. In highly strained polycyclic systems and heteroatom-rich rings, this topological strain is severe enough to invert the natural correlations, leading to a stark decoupling between vibrational energy and heat capacity.

Ultimately, the dual application of optimal transport for global displacement and covariance differentials for internal rewiring provides a precise, scalar measure of chemical strain. This approach successfully disentangles pure topological effects from raw compositional variance, offering a robust mathematical lens through which to understand how spatial confinement reorganizes the fundamental thermodynamic and spatial states of molecules.

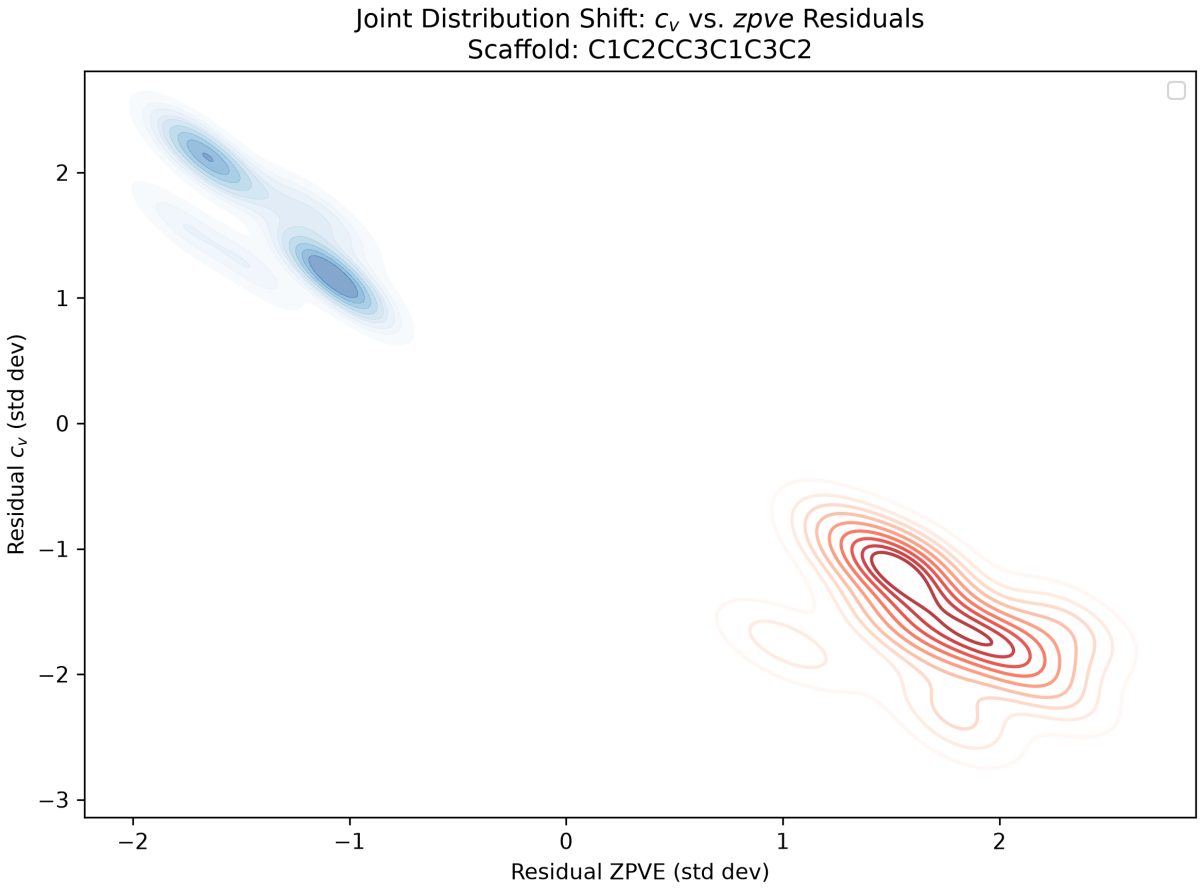

Figure 4: Joint density distributions of residual zero-point vibrational energy (*zpve*) and heat capacity (c_v) for the highly strained tetrahedrane-like cyclic scaffold `C1C2CC3C1C3C2` and its stoichiometry-matched acyclic baseline. The plot illustrates an extreme distributional shift (Wasserstein distance $W = 3.136$) driven by topological constraints, demonstrating both global manifold displacement and internal covariance rewiring. The compressed and shifted density contours reveal a stark inversion in the relationship between the c_v and *zpve* residuals (average ratio of -0.928) relative to the positive correlation typically observed in diffuse acyclic chains, highlighting how high ring counts and intense topological strain fundamentally reorganize coupled thermodynamic properties.