Conditional Distance Correlation of the Vibrational-Spatial Axis Across Topological Complexity

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding how macroscopic spatial confinement and structural topology rewire the coupling between thermodynamic, vibrational, and spatial properties remains a significant challenge in molecular physics. To address this, we investigate the joint organizational structure of 133,802 unique molecules from the QM9 dataset by isolating intrinsic chemical variance from stoichiometric effects. We extract atom-count residuals using gradient-boosted regressors and apply conditional distance correlation to quantify non-linear dependencies between thermodynamic-vibrational and geometric-spatial axes. By stratifying the chemical space based on exact ring count and the fraction of sp^3 carbons, we reveal that the coupling between vibrational and spatial properties systematically strengthens with increasing cyclization and saturation, particularly in highly polycyclic systems. Conversely, global rotational shape and spatial extent are tightly bound in simple cyclic molecules but decouple in complex polycyclic scaffolds. Furthermore, bootstrap analysis of linear molecules demonstrates a near-deterministic coupling, highlighting a distinct physical regime that fundamentally drops in strength upon transition to three-dimensional acyclic structures. These findings quantitatively establish that a molecule's joint physical structure is not a global constant, but is instead dynamically reorganized by its topological architecture.

1 Introduction

Understanding how macroscopic spatial confinement and structural topology dictate the physical properties of molecules remains a central challenge in chemical physics. While the isolated prediction of thermodynamic, vibrational, and spatial properties has seen immense progress, the joint organizational structure—how these properties couple and constrain one another—is significantly less understood. Specifically, it is unclear how architectural features such as cyclization and saturation dynamically rewire the coupling between a molecule's vibrational energy landscape and its spatial extent.

A primary obstacle in quantifying these intrinsic physical couplings is the overwhelming influence of stoichiometry. Extensive properties scale dominantly with atom counts, which can mask the subtle, non-linear dependencies driven by molecular geometry. Furthermore, traditional linear covariance metrics are often inadequate for capturing complex dependencies across diverse chemical scaffolds. They can become unstable or rank-deficient when applied to sparsely populated structural classes, and they frequently conflate coupling strength with metric distance.

1.1 Isolating intrinsic chemical variance

To rigorously map how topological complexity rewires molecular physics, this study investigates the joint organizational structure of 133,802 structurally unique molecules from the QM9 dataset. We isolate intrinsic chemical variance from stoichiometric effects by extracting atom-count residuals using gradient-boosted regressors. By stripping away the dominant size-dependent scaling, the remaining residuals preserve the rich, intensive chemical information necessary to study joint structural organization without the confounding effects of molecular size.

1.2 Quantifying joint structural organization

Rather than relying on linear correlation, we deploy conditional distance correlation to quantify non-linear dependencies between distinct physical axes across the deduplicated dataset. This approach provides a robust, stoichiometry-independent measurement of structural organization. We specifically isolate two primary axes of joint structure:

- The coupling between the thermodynamic-vibrational axis and the geometric-spatial axis.
- The coupling between global rotational shape, quantified by Ray's asymmetry parameter (κ), and spatial extent.

By stratifying the chemical space based on exact ring count and the fraction of sp^3 hybridized carbons (F_{sp^3}), we evaluate how these partial distance correlations strengthen or decay as a continuous function of molecular rigidity.

Through this framework, we demonstrate that the coupling between vibrational and spatial properties systematically strengthens with increasing cyclization and saturation, particularly in highly polycyclic systems. Conversely, we reveal that while global rotational shape and spatial extent are tightly bound in simple cyclic molecules, they fundamentally decouple in complex polycyclic scaffolds. Finally, by isolating linear molecules, we identify a distinct physical regime characterized by near-deterministic coupling, highlighting a fundamental drop in coupling strength upon the transition to three-dimensional acyclic structures. Ultimately, this work quantitatively establishes that a molecule's joint physical structure is not a global constant, but is instead dynamically reorganized by its topological architecture.

2 Methods

2.1 Data curation and topological triage

To rigorously map how macroscopic spatial confinement rewires molecular physics, we utilized the QM9 dataset. We first ensured a structurally unique set of molecules by identifying duplicates in the SMILES representations and retaining only the first occurrence, yielding exactly 133,802 unique molecules. To prevent fundamentally different rotational physics from biasing global dependency metrics, we isolated the 9 linear molecules, identifiable by a rotational constant $A = 0$, into a distinct topological class. Finally, the feature space was filtered to remove redundant thermodynamic and electronic properties (specifically, u_{298}, h_{298}, g_{298}, u_{298_atom}, h_{298_atom}, g_{298_atom}, and the HOMO-LUMO gap). This triage prevents rank-deficient covariance structures in downstream multivariate analyses.

2.2 Structural parsing and descriptor generation

Deduplicated SMILES strings were parsed into molecular objects to compute exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen. To define our axes of molecular rigidity, we computed topological descriptors including the exact ring count, the fraction of sp^3 hybridized carbons (F_{sp^3}), and the number of rotatable bonds.

For the non-linear molecules, we quantified molecular shape using Ray's asymmetry parameter, κ, defined as:

$$\kappa = \frac{2B - A - C}{A - C}$$

where A, B, and C are the rotational constants strictly sorted such that $A \geq B \geq C$. Molecules exhibiting near-spherical geometries, defined by $|A - C| < 10^{-3}$, were assigned $\kappa = 0$. A small amount of Gaussian noise (jitter) was added to these assigned values to prevent point-mass artifacts in downstream distance correlation calculations. Linear molecules were assigned $\kappa = \text{NaN}$. Additionally, Morgan fingerprints (radius 2, 2048 bits) were generated for all molecules to serve as a high-dimensional structural baseline for sanity checks.

2.3 Stoichiometric isolation via gradient boosting

To isolate intrinsic chemical structure from extensive stoichiometric effects, we extracted property residuals using gradient-boosted regressors (LightGBM) via 5-fold cross-validation on the non-linear molecules. We trained two separate sets of models for each non-redundant property ($zpve$, c_v, r^2, α, μ, u_0, $homo$, $lumo$, u_{0_atom}):

- **Model A (Baseline):** Predictors were strictly restricted to the exact counts of C, N, O, F, and H.
- **Model B (Sensitivity):** Predictors included the atom counts alongside the number of rotatable bonds and F_{sp^3}.

Residuals were extracted by subtracting the out-of-sample predictions from the observed values. For the 9 linear molecules, values were predicted using an ensemble of the cross-validated models trained on the non-linear set. All resulting residuals were standardized using the mean and standard deviation derived strictly from the non-linear molecules. To validate that Model A residuals retained intensive structural variance rather than just noise, we regressed them against the Morgan fingerprints using a Ridge regression model and recorded the resulting R^2.

2.4 Multivariate grouping and topological stratification

To evaluate joint structural organization, we defined two multivariate coupling vectors for each molecule:

- **Coupling 1 (Thermodynamic-Vibrational vs. Geometric-Spatial):** Vector $X = [res_zpve, res_c_v]$ and Vector $Y = [res_r^2, res_\alpha]$.
- **Coupling 2 (Rotational Shape vs. Spatial):** Vector $X = [\kappa]$ and Vector $Y = [res_r^2, res_\alpha]$.

We stratified the non-linear molecules into a two-dimensional grid to decouple the effects of cyclization and saturation. The first axis was defined by exact ring count (0, 1, 2, 3, and 4+ rings), and the second axis by F_{sp^3}, which was divided into 8 quantile-based bins. Each non-linear molecule was mapped into one of the 40 resulting cells, and the variance of κ within each cell was calculated and stored.

2.5 Distance correlation and statistical validation

Within each of the 40 topological cells, we computed the multivariate distance correlation (dCor) for Coupling 1 and Coupling 2 using both Model A and Model B residuals. To address potential mathematical artifacts where κ is constrained by the geometry of polycyclic systems, Coupling 2 dCor values were normalized by dividing them by the variance of κ within their respective cells. This normalization distinguishes true topological masking effects from range-restriction artifacts. For the isolated linear molecules, we computed the dCor for Coupling 1 using a bootstrap approach, resampling the 9 molecules with replacement 1,000 times to generate a robust mean estimate and confidence interval.

To rigorously test the significance of the observed couplings, we implemented a global permutation-based null model. The property vectors (X and Y) were shuffled across the entire dataset 1,000 times while keeping the topological descriptors (ring count, F_{sp^3}) fixed to their original molecules. Empirical p-values were calculated for each cell by comparing the observed dCor to the distribution of permuted dCor values.

Finally, for each ring-count slice, we fit a linear regression of the significant observed dCor values against the median F_{sp^3} of their respective bins to extract the slope and R^2 of the trend. We explicitly compared the magnitude of the dCor coefficients between Model A and Model B to quantify how much of the joint structure acts as a proxy for specific chemical motifs versus a fundamental physical coupling. The dCor trends across F_{sp^3} were then evaluated against the bootstrapped confidence interval of the linear molecules to determine if the trends extrapolate to the extreme limit of spatial confinement or if the linear limit represents a distinct physical regime.

3 Results

3.1 Stoichiometric isolation and residual validity

To identify joint structural organization independent of molecular size, we first confirmed the efficacy of our stoichiometric isolation. The gradient-boosted baseline (Model A), trained exclusively on exact atom counts, accounted for the vast majority of variance in extensive properties (e.g., $R^2 = 0.9971$ for zero-point vibrational energy, *zpve*). To ensure that the extracted residuals retained meaningful intensive chemical information rather than mere statistical noise, we regressed these residuals against Morgan fingerprints (radius 2) using Ridge regression. The remaining structural variance in the residuals was substantial: $R^2 = 0.747$ for *zpve*, 0.803 for heat capacity (c_v), and 0.739 for electronic spatial extent (r^2). This validates that while atom counts dominate the raw property values, our residuals successfully preserve the rich chemical variance necessary to study intrinsic joint organization.

3.2 Topological stratification of joint organizational axes

We evaluated the joint structural organization across 34 valid topological cells defined by exact ring count and the fraction of sp^3 hybridized carbons (F_{sp^3}). Two primary axes were analyzed: Coupling 1 (thermodynamic-vibrational versus geometric-spatial properties) and Coupling 2 (rotational shape, κ, versus spatial extent). Global permutation testing confirmed the statistical significance of Coupling 1 in every analyzed cell ($p < 0.05$). The spatial distribution of these distance correlation (dCor) values across the topological grid is visualized in Figure 1.

3.3 Thermodynamic-vibrational and spatial coupling

For Coupling 1, the global average dCor under Model A is 0.4184. In purely acyclic molecules (0 rings), the average dCor is 0.3570. As cyclization increases, we observe a systematic strengthening of the relationship between vibrational energy profiles and spatial extent (Figure 2, left panel). This topological strengthening is most pronounced in highly polycyclic systems (4 or more rings), where the dCor exhibits a strong positive trend with increasing saturation (slope = 0.5228, $R^2 = 0.736$, $p = 0.0289$). This indicates that in complex architectural cages, the degree of saturation becomes a dominant regulator of the vibrational-spatial joint structure.

3.4 Geometric decoupling in complex polycyclic systems

Coupling 2, which measures the dependency between a molecule's global geometric asymmetry (κ) and its electronic spatial distribution, yields a global average dCor of 0.4332. The trends across F_{sp^3} for different ring counts are illustrated in Figure 2 (right panel). In simple cyclic systems (1 and 2 rings), Coupling 2 strengthens significantly as F_{sp^3} increases (slopes of 0.3286 and 0.3964, respectively, with $R^2 > 0.84$ and $p < 0.004$). This demonstrates that the transition from planar aromaticity to three-dimensional saturation tightly locks the molecule's rotational shape to its electronic extent. However, this relationship fundamentally collapses in highly complex polycyclic scaffolds (4+ rings), where the trend flattens and loses statistical significance (slope = -0.0764, $p = 0.586$). High topological complexity thus decouples global shape from local spatial residuals.

3.5 The linear limit and extreme spatial confinement

To contextualize these trends, we analyzed the 9 linear molecules as the extreme limit of topological simplicity. Using a bootstrap approach (1,000 resamples), the linear molecules exhibit a Coupling 1 dCor of 0.9265 (95% CI: [0.7130, 0.9867]), as highlighted in Figure 2 (left panel). This near-deterministic coupling is significantly higher than the average for acyclic molecules (0.3570). This stark contrast suggests a physical phase transition: the joint structure that is only moderately coupled in complex, three-dimensional molecules becomes tightly bound under extreme spatial confinement. The transition from a one-dimensional linear topology to even the simplest three-dimensional acyclic structure fundamentally rewires the relationship between vibrational energy and spatial distribution, dropping the coupling strength by more than half.

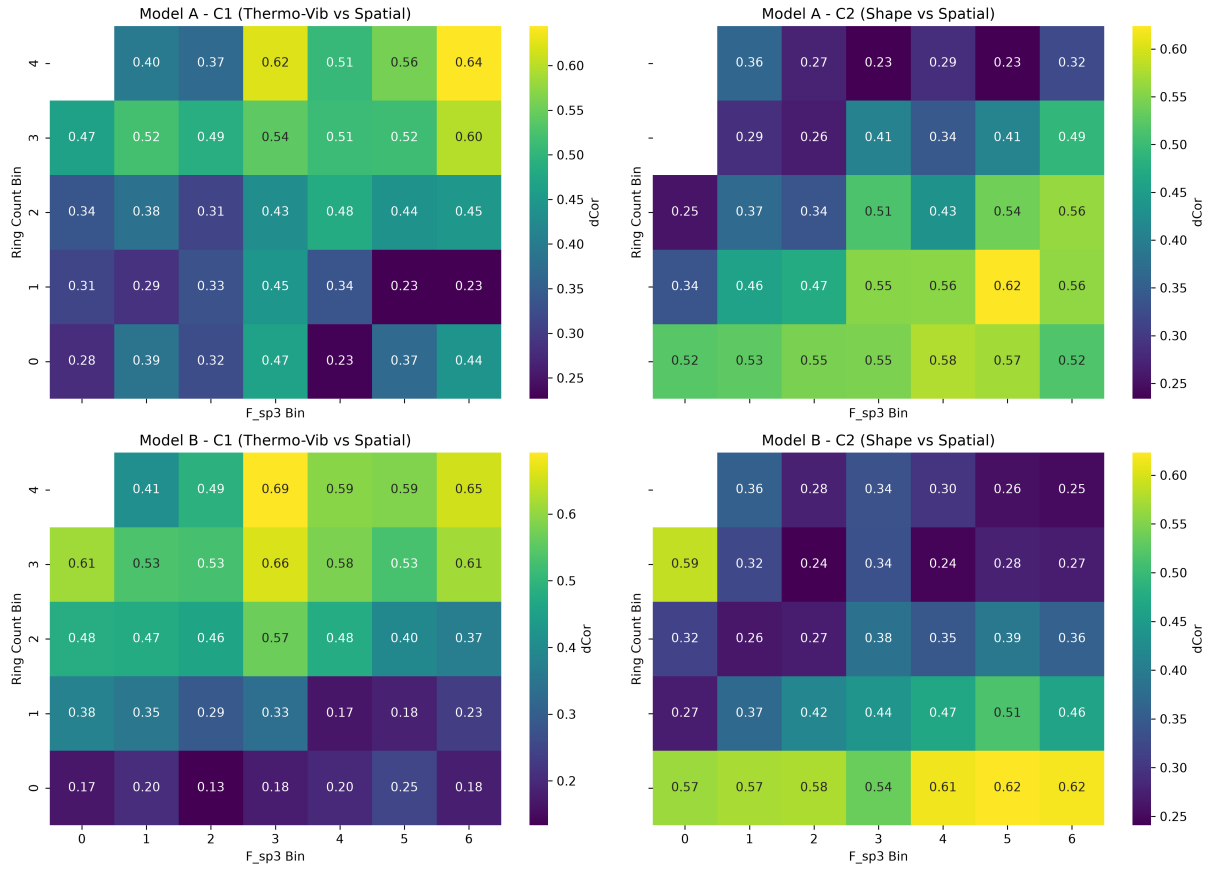

Figure 1: Heatmaps of distance correlation (dCor) illustrating the joint organizational structure of molecular properties across topological cells defined by ring count and saturation (F_{sp^3}). The left panels display Coupling 1 (thermodynamic-vibrational vs. spatial properties), demonstrating a systematic topological strengthening where coupling increases with cyclization, particularly in polycyclic systems (4+ rings). The right panels display Coupling 2 (rotational shape vs. spatial extent), revealing a geometric decoupling: global shape and spatial extent are tightly bound in simple 1- and 2-ring systems as F_{sp^3} increases, but this relationship collapses in highly complex 4+ ring systems. Furthermore, comparing residuals from Model A (atom counts only, top row) and Model B (incorporating F_{sp^3} and rotatable bonds, bottom row) shows that Coupling 1 remains robustly consistent, whereas Coupling 2 is highly sensitive to structural descriptors, which explain away a larger portion of the joint variance in complex scaffolds.

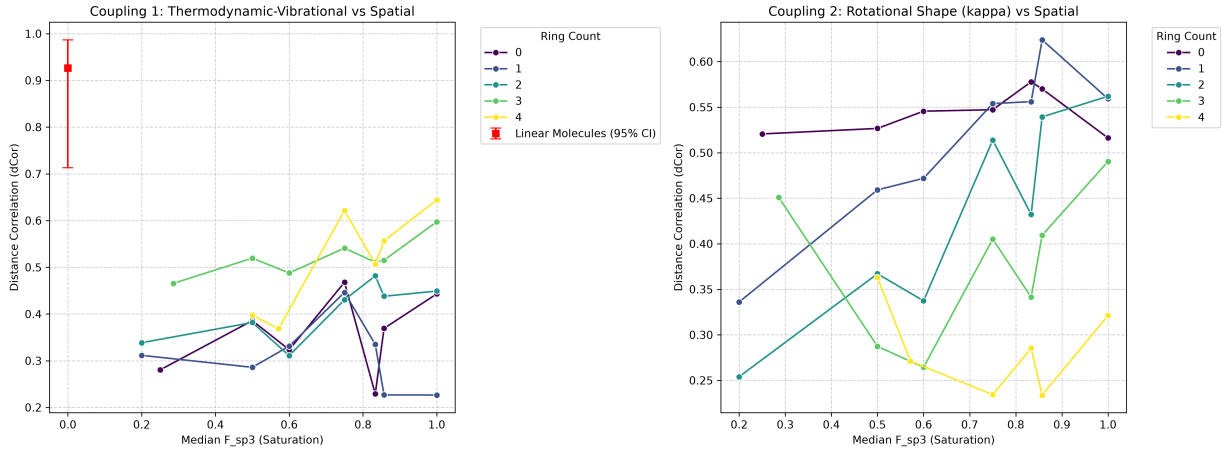

Figure 2: Distance correlation (dCor) of joint structural properties across topological cells defined by ring count and median saturation (F_{sp^3}). The left panel illustrates Coupling 1 (thermodynamic-vibrational versus spatial properties), revealing that this relationship strengthens with increasing cyclization and saturation, particularly in polycyclic systems (4+ rings). Notably, linear molecules (red square, 95% CI) exhibit an exceptionally high coupling ($dCor > 0.9$), representing a distinct physical regime of extreme spatial confinement where properties are nearly deterministically linked compared to their acyclic counterparts. The right panel displays Coupling 2 (rotational shape, κ, versus spatial extent), showing that while simple 1- and 2-ring systems tightly lock their rotational shape to electronic extent as saturation increases, this trend collapses in complex 4+ ring systems due to geometric decoupling. Overall, the plots demonstrate that a molecule's joint organizational structure is not a global constant but is dynamically rewired by its topological complexity.

3.6 Model sensitivity and chemical rewiring

Finally, we compared the dCor values derived from Model A (atom counts only) with those from Model B (atom counts, F_{sp^3}, and rotatable bonds) to determine if the observed joint structure is merely a proxy for specific saturation motifs (Figure 1). For Coupling 1, the mean difference in dCor ($dCor_A - dCor_B$) is minimal (0.0083), indicating that the thermodynamic-vibrational axis represents a robust physical coupling independent of the specific structural descriptors used in Model B. Conversely, the mean difference for Coupling 2 is 0.0380, reaching up to 0.2206 in specific highly saturated polycyclic cells (e.g., 3 rings, bin 6). This confirms that the rotational-spatial coupling is highly sensitive to structural descriptors, which effectively explain away a larger portion of the joint variance in complex scaffolds.

4 Conclusions

In this study, we established a quantitative framework to map how macroscopic spatial confinement and structural topology dynamically rewire the joint organizational structure of molecular properties. By isolating intrinsic chemical variance from dominant stoichiometric effects using gradient-boosted residuals, we bypassed the limitations of traditional linear covariance and size-dependent scaling. Our application of conditional distance correlation across 133,802 unique molecules from the QM9 dataset reveals that the coupling between physical properties is not a global constant, but rather a dynamic feature dictated by a molecule's architectural cage.

First, we demonstrated a systematic topological strengthening in the coupling between the thermodynamic-vibrational and geometric-spatial axes. As cyclization and saturation increase, particularly within highly polycyclic systems, these properties become increasingly interdependent. Crucially, this relationship represents a robust physical coupling that persists independently of specific structural descriptors, confirming that topological complexity fundamentally tightens the constraints between a molecule's vibrational energy landscape and its spatial extent.

Conversely, we identified a distinct geometric decoupling regarding global rotational shape. While the transition from planar aromaticity to three-dimensional saturation tightly locks rotational shape to electronic spatial extent in simple cyclic systems, this joint structure collapses in highly complex polycyclic scaffolds. In these complex systems, global shape and local spatial residuals decouple, and the joint variance is largely explained away by specific structural motifs such as the fraction of sp^3 carbons and bond rotatability.

Finally, our analysis of linear molecules highlights a distinct physical regime governed by extreme spatial confinement. In the one-dimensional limit, thermodynamic and spatial properties exhibit near-deterministic coupling. The transition to even the simplest three-dimensional acyclic structure induces a fundamental drop in this coupling strength, acting as a topological phase transition that drastically rewires the molecule's physical organization.

Ultimately, these findings demonstrate that molecular properties cannot be fully understood in isolation or through simple linear scaling. The topological architecture of a molecule—from the rigid confinement of linear systems to the complex decoupling in polycyclic networks—dynamically reorganizes its joint physical structure.