Topological Rewiring of the Vibrational-Spatial Property Axis in Small Organic Molecules

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The physical state of a molecule is defined by a manifold of interdependent properties, yet how these properties jointly couple across distinct chemical architectures remains poorly understood beyond simple size effects. To isolate structural signals from stoichiometric determinism, we investigate the joint structure of electronic spatial extent, zero-point vibrational energy, and heat capacity. Using the QM9 dataset of small organic molecules, we establish a linear stoichiometric baseline based on atom counts and analyze the resulting property residuals using Distance Correlation and Pearson metrics. To examine structural dependencies, the dataset is stratified by topological complexity using the Bertz index and separated into aliphatic and aromatic subsets, while an Isolation Forest is applied to ensure robustness against geometric anomalies. Our results reveal a robust global vibrational-spatial coupling that is dynamically rewired by molecular complexity. Specifically, as topological complexity increases, aliphatic systems exhibit strengthened spatial-vibrational coherence, whereas aromatic systems maintain a decoupled state. Additionally, rotational geometry tracks spatial extent but remains largely independent of the vibrational manifold. Ultimately, these findings demonstrate that molecular topology acts as a primary organizer of joint physical behavior, fundamentally reorganizing property relationships as molecular complexity increases.

1 Introduction

The physical state of a molecule is defined by a complex manifold of interdependent properties. Historically, computational chemistry and cheminformatics have focused on predicting individual molecular properties from structural descriptors. Yet, how these diverse physical properties jointly couple across distinct chemical architectures remains poorly understood. This gap in understanding is largely due to the confounding influence of molecular size: extensive properties such as zero-point vibrational energy (ZPVE), heat capacity (C_v), and electronic spatial extent (r^2) scale predictably with the number of atoms. This stoichiometric determinism often masks the underlying structural signals, obscuring how the specific topological arrangement of atoms dictates the joint physical behavior of the molecule.

To uncover these fundamental relationships, it is necessary to isolate the variance in physical properties that arises purely from molecular topology rather than simple atom counts. In this work, we investigate the joint structure of the vibrational-spatial property axis using the QM9 dataset of small organic molecules. We establish a linear stoichiometric baseline based on exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms to residualize ZPVE, C_v, and r^2. By analyzing these property residuals, we remove the trivial size effects and expose the underlying structural coupling. To rigorously quantify these relationships without assuming multivariate normality, we employ both Pearson correlation and Distance Correlation ($dCor$), which captures both linear and non-linear dependencies within this property triad.

To determine if this joint structure is a static physical constant or a dynamic feature of chemical architecture, we stratify the dataset by topological complexity using the Bertz index and separate the molecules into fully aliphatic and aromatic-containing subsets. Furthermore, we integrate the rotational constant to evaluate whether rotational geometry tracks this vibrational-spatial axis. Because the QM9 dataset contains highly strained and geometrically inconsistent molecules, we apply an Isolation Forest algorithm to the residual space, ensuring that our observed correlations are robust against structural anomalies.

Our analysis reveals a robust global coupling between vibrational energy and spatial extent that is fundamentally rewired as molecular complexity increases. Specifically, we demonstrate a structural divergence: as topological complexity grows, aliphatic systems exhibit strengthened spatial-vibrational coherence, whereas aromatic systems maintain a decoupled state. Ultimately, these findings establish that molecular topology acts as a primary organizer of joint physical behavior, fundamentally reorganizing property relationships as molecular complexity increases.

2 Methods

2.1 Data preparation and preprocessing

The analysis was conducted on the QM9 dataset of small organic molecules. To prevent data leakage during subsequent statistical analyses, the dataset was first deduplicated by removing 83 redundant SMILES strings. We identified nine linear molecules characterized by a rotational constant A of exactly zero. A binary indicator was created to explicitly flag this geometric class. For the remaining non-linear molecules, a log-transformation was applied to the rotational constant A to stabilize its extreme variance, yielding $\log A$. The nine linear molecules were assigned undefined (NaN) values for $\log A$ and were strictly excluded from any calculations involving rotational geometry.

2.2 Stoichiometric residualization

As outlined in the introduction, extensive properties such as zero-point vibrational energy (ZPVE), heat capacity (C_v), and electronic spatial extent (r^2) are heavily confounded by molecular size. To isolate the variance attributable purely to chemical architecture, we established a linear stoichiometric baseline. For each molecule, exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were extracted from the SMILES strings. A multiple linear regression model was then fitted to predict each property $y \in \{r^2, \text{ZPVE}, C_v\}$ based on these atom counts:

$$y = \beta_0 + \beta_C N_C + \beta_N N_N + \beta_O N_O + \beta_F N_F + \beta_H N_H + \epsilon \tag{1}$$

The resulting residuals (res_r2, res_zpve, res_cv) represent the structural signal independent of stoichiometric determinism. A separate regression was performed for $\log A$ strictly on the non-linear subset to generate res_logA. To prevent extreme outliers from skewing the distribution, all residuals were standardized using a robust scaler based on the median and interquartile range (IQR):

$$x_{\text{scaled}} = \frac{x - \text{median}(x)}{\text{IQR}(x)} \tag{2}$$

2.3 Structural descriptors and anomaly detection

To quantify topological and electronic complexity, the Bertz complexity index (BertzCT) and the Kier-Hall alpha value (HallKierAlpha) were computed for each molecule. Molecules were further partitioned into fully aliphatic or aromatic-containing subsets based on the extraction of their Murcko scaffolds.

Because the uncleaned QM9 dataset contains geometrically inconsistent molecules that lack explicit flags, we applied an Isolation Forest algorithm directly to the 3D residual space (res_r2, res_zpve, res_cv). This approach generated continuous anomaly scores for each molecule, allowing us to robustly identify highly strained or anomalous structures without assuming multivariate normality.

2.4 Correlation metrics and null permutation validation

The global vibrational-spatial coupling was quantified using both Pearson correlation (r) and Distance Correlation ($dCor$), with the latter employed to capture non-linear dependencies within the property triad.

To rigorously prove that the observed joint structure is a genuine chemical phenomenon rather than a statistical artifact of the residualization process, a permutation test was conducted. The values of res_cv were randomly shuffled strictly *within* their respective structural families (i.e., independently within the

aliphatic and aromatic groups). This stratified shuffling preserves the marginal distributions of the subsets while destroying the local joint structure. The correlation matrices were then recomputed to verify that the coupling strength collapsed to near-zero under the null hypothesis.

2.5 Stratification and rotational geometry integration

To test the hypothesis that the thermodynamic-spatial axis is dynamically rewired by molecular complexity, the dataset was binned into deciles based on the BertzCT index. Within each decile, the correlation matrices for the res_zpve, res_r2, and res_cv triad were computed separately for the aliphatic and aromatic subsets.

To ensure that the observed correlation shifts were not driven by geometric artifacts, a sensitivity analysis was performed. A "clean" subset was defined by excluding the top 5% most anomalous molecules based on the Isolation Forest scores. The stratified correlation shifts were then recalculated on this filtered subset to confirm the robustness of the structural rewiring.

Finally, the rotational geometry (res_logA) was analyzed independently. Pearson and Distance correlations of res_logA against the primary triad were computed exclusively for the non-linear subset. This property was kept separate from the higher-dimensional joint calculations to prevent its sensitivity to the linear molecule mask and extreme variance from destabilizing the primary triad analysis.

3 Results

3.1 Global vibrational-spatial coupling

Following the stoichiometric residualization of the 133,802 unique molecules in the QM9 dataset, we evaluated the global joint structure of the property triad. Across the full dataset, a robust coupling exists between vibrational energy, thermal response, and spatial extent that is independent of molecular size. The triad of residuals exhibits high global distance correlations ($dCor$), which capture both linear and non-linear dependencies. Specifically, the coupling between zero-point vibrational energy (res_zpve) and heat capacity (res_cv) is exceptionally strong ($dCor = 0.8334$, Pearson $r = -0.8761$). Furthermore, res_zpve and electronic spatial extent (res_r2) demonstrate significant coherence ($dCor = 0.5391$, Pearson $r = -0.5525$), as do res_r2 and res_cv ($dCor = 0.5529$, Pearson $r = 0.5719$).

To verify that these observed couplings are chemically meaningful and not statistical artifacts of the residualization process, we performed a null permutation test. When res_cv values were randomly shuffled strictly within their respective aromaticity groups, the joint structure collapsed entirely. For instance, the $dCor$ between res_zpve and res_cv dropped precipitously from 0.8334 to 0.0245. This confirms that the observed vibrational-spatial coupling is a genuine physical manifestation of local chemical architecture.

3.2 Rotational geometry integration

To determine if this joint structure extends to rotational geometry, we analyzed the log-transformed rotational constant (res_logA) for the 133,793 non-linear molecules. The results indicate that rotational geometry exhibits a specific, targeted coupling to the spatial extent ($dCor = 0.5334$ against res_r2) but remains largely decoupled from the vibrational energy manifold ($dCor = 0.1483$ against res_zpve). This divergence confirms that while rotational constants accurately track the spatial distribution of mass and electrons, they do not inherently capture the energetic vibrational state of the molecule, justifying their exclusion from the primary property triad.

3.3 Complexity-driven rewiring of the property axis

The most significant finding of this study emerges when the dataset is stratified by topological complexity. The joint structure of the property triad is not a static physical constant but is dynamically rewired as molecular complexity increases, as illustrated in Figure 1.

In the lowest complexity bin (BertzCT 0.0–102.21), fully aliphatic and aromatic-containing molecules behave similarly, both exhibiting a baseline spatial-vibrational coupling of $dCor \approx 0.50$ for res_zpve versus res_r2. However, a stark structural divergence occurs as topological complexity grows. For aliphatic

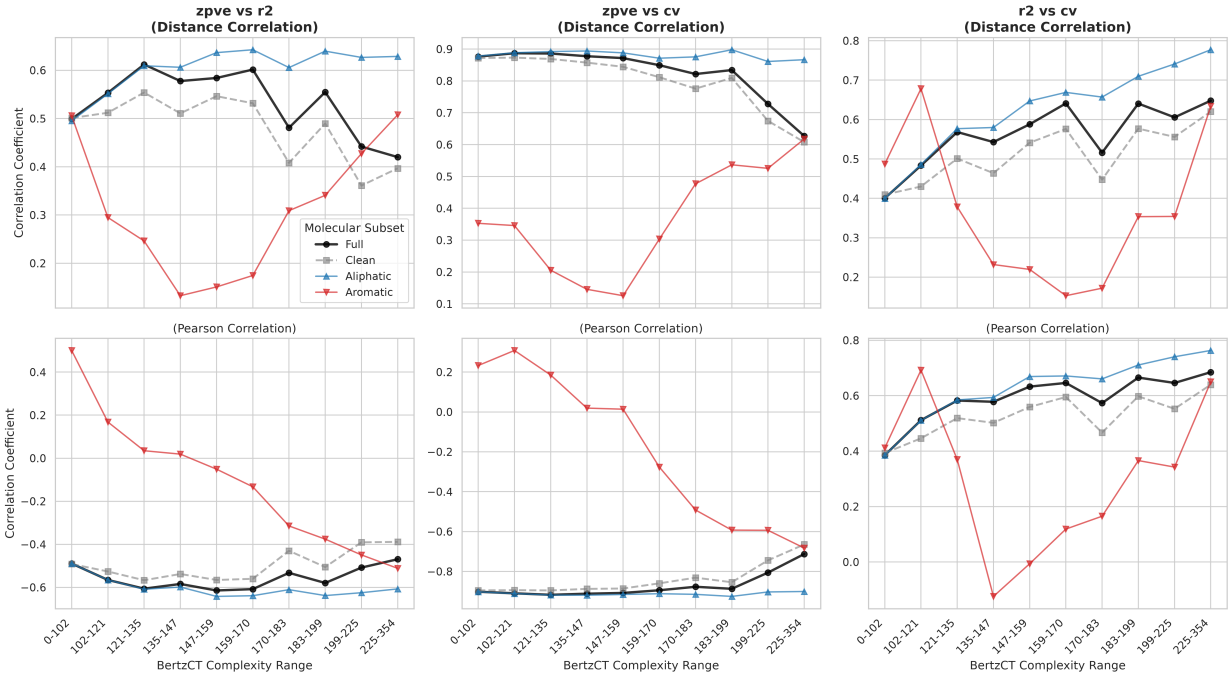

Figure 1: Distance (top row) and Pearson (bottom row) correlation coefficients between residualized molecular properties (ZPVE, r^2, and C_v) across deciles of BertzCT topological complexity. The data is stratified into the Full dataset, a Clean subset (excluding the top 5% of anomalies), and fully Aliphatic versus Aromatic molecules. The plots illustrate the complexity-driven rewiring of the vibrational-spatial axis. At low complexity, aliphatic and aromatic systems exhibit similar coupling behavior. However, as topological complexity increases, aliphatic molecules strengthen their spatial-vibrational coherence, whereas aromatic systems exhibit significant decoupling. This aliphatic-aromatic divergence drives the overall weakening of property correlations in the Full and Clean datasets at high complexity, demonstrating that molecular topology acts as a primary, dynamic organizer of joint physical behavior independent of absolute molecular size.

molecules, the vibrational-spatial coupling strengthens systematically, reaching $dCor = 0.6287$ in the highest complexity decile—a substantial positive drift of $+0.1334$. Conversely, aromatic molecules maintain a relatively stable, decoupled state across all complexity bins ($dCor \approx 0.5076$).

Interestingly, within the unstratified full dataset, the overall coupling between res_zpve and res_r2 weakens at the highest complexity levels, dropping from $dCor = 0.4990$ in the first decile to 0.4200 in the tenth. This global trend is driven by the increasing prevalence of aromatic systems in high BertzCT bins. The unique electronic constraints of extended aromatic networks effectively decouple the spatial extent from the vibrational energy modes, masking the underlying consolidation occurring within aliphatic systems of similar complexity.

3.4 Robustness against structural anomalies

To ensure that this rewiring effect is not an artifact driven by highly strained geometries or geometrically inconsistent molecules within the QM9 dataset, we evaluated the correlation shifts within the "Clean" subset. By excluding the top 5% most anomalous molecules identified by an Isolation Forest in the residual space, we isolated the behavior of structurally well-behaved molecules.

The clean subset exhibits an even sharper decline in global coupling at high complexity, with the res_zpve versus res_r2 $dCor$ falling from 0.5010 in Decile 0 to 0.3968 in Decile 9 (a drift of -0.1042). The persistence and amplification of this trend in the filtered data confirm that the observed structural reorganization is a fundamental characteristic of the chemical manifold. Molecular topology acts as a primary switch for property coupling, fundamentally altering the joint response of physical properties as molecules grow more complex.

4 Conclusions

The physical state of a molecule is governed by a manifold of interdependent properties, yet the underlying structural signals are frequently obscured by stoichiometric determinism. By establishing a linear stoichiometric baseline to residualize zero-point vibrational energy (ZPVE), heat capacity (C_v), and electronic spatial extent (r^2) across the QM9 dataset, we successfully isolated the variance attributable purely to chemical architecture. Our analysis demonstrates that vibrational energy and electronic extent are intrinsically coupled, forming a robust vibrational-spatial axis that is entirely independent of absolute molecular size. Furthermore, we established that while rotational geometry accurately tracks spatial extent, it remains largely decoupled from the vibrational manifold, functioning primarily as a spatial rather than energetic descriptor.

The most significant outcome of this study is the discovery that the joint structure of these physical properties is not a static constant, but rather a dynamic entity modulated by molecular topology. As topological complexity increases, the property axis undergoes a fundamental rewiring characterized by a stark aliphatic-aromatic divergence. Aliphatic systems systematically consolidate their spatial-vibrational coherence as they grow more complex. In contrast, aromatic systems maintain a decoupled state, and their increasing prevalence at higher complexity levels drives an overall global weakening of property correlations. The persistence of these trends within a rigorously filtered subset of structurally well-behaved molecules confirms that this reorganization is a fundamental characteristic of the chemical space.

Ultimately, these findings challenge the paradigm of treating molecular properties as isolated prediction targets or assuming they exhibit fixed correlations. Instead, property sets must be viewed as structurally dependent units. Future efforts in computational chemistry and targeted molecular design must account for the fact that modifying a molecule's topological complexity or aromaticity will fundamentally alter the joint response of its physical properties.