Diffusion Map Trajectories of Property Coupling Networks Across Topological Complexity Regimes

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The structural organization of molecular properties is frequently dominated by molecular size and stoichiometry, obscuring how intrinsic thermodynamic and electronic couplings evolve with increasing topological complexity. To isolate and quantify these underlying structural transitions, we analyze the QM9 dataset by extracting size-invariant property residuals using restricted-depth random forest regressors conditioned on exact atom counts. We partition the dataset into distinct topological regimes based on the Bertz complexity index and ring counts, computing distance correlation matrices for twelve non-redundant properties within each subpopulation. By projecting these dependency networks into a latent space via diffusion maps and measuring the reorganization between adjacent complexity levels using Procrustes disparity, we track the continuous deformation of property couplings. Our analysis reveals that the velocity of structural rewiring varies significantly across molecular families. Specifically, bicyclic heteroatom-rich structures exhibit the most drastic shifts in property couplings, whereas hydrocarbon-only tracks remain relatively stable. Furthermore, latent space trajectories demonstrate a distinct decoupling of frontier orbital energies from geometric and spatial properties as complexity increases. Sensitivity analyses and scaffold-based split-half validations confirm that these structural transitions are robust despite the retention of unfiltered geometric artifacts in the dataset, and demonstrate that the observed topological rewiring is significantly amplified by thermodynamic interactions.

1 Introduction

The structural organization of molecular properties is a fundamental aspect of chemical space, dictating how molecules interact, react, and stabilize. However, the macroscopic observation of these properties is frequently dominated by trivial scaling laws related to molecular size and stoichiometry. Because extensive properties naturally scale with the number of atoms, the intrinsic, non-linear couplings between a molecule's thermodynamic state and its electronic structure are often obscured. Consequently, it remains challenging to discern how these underlying physical relationships evolve as molecules transition across regimes of increasing topological complexity, such as the introduction of branching or multiple ring systems.

1.1 The challenge of stoichiometric confounding

Traditional approaches to analyzing molecular property spaces often treat features independently or rely on linear correlation metrics that fail to capture complex, multi-dimensional dependencies. Furthermore, while it is well understood that topological complexity confines spatial extents and alters frontier orbital energies, the field lacks a scalable, stoichiometry-independent framework to quantify the continuous deformation of these property networks. Without isolating the size-invariant residuals of these properties, any observed correlation is likely to be an artifact of atom counting rather than a reflection of true physical coupling.

To rigorously investigate these intrinsic relationships, it is necessary to decouple the stoichiometric baseline from the structural variance. By extracting gradient-boosted atom-count residuals, we can isolate the purely geometric and topological contributions to molecular physics, enabling a direct comparison of property networks across vastly different molecular sizes.

Structural Track	Mean Disparity (Rewiring Velocity)	Std Dev	Transition Count
R0 (Heteroatom-rich)	0.1912	0.1717	2
R1 (Heteroatom-rich)	0.1281	0.0862	3
R2 (Heteroatom-rich)	0.2537	0.1170	3
R3 (Heteroatom-rich)	0.1539	0.1143	3
R3 (Hydrocarbon-only)	0.0954	N/A	1

1.2 Mapping topological rewiring

In this work, we present a framework to map how macroscopic structural transitions continuously reorganize molecular physics. Utilizing the QM9 dataset, we extract size-invariant property residuals for twelve non-redundant thermodynamic, electronic, and geometric properties. This is achieved by conditioning the properties on exact heavy atom and hydrogen counts using restricted-depth random forest regressors, ensuring that the extracted residuals are not overfit to specific chemical motifs.

To overcome structural sparsity and ensure statistical stability, the dataset is partitioned into dense complexity regimes based on the Bertz complexity index and exact ring counts. Within each topological regime, we compute a complete distance correlation matrix to capture the intrinsic joint structures of the properties—such as the interplay between spatial extent, isotropic polarizability, and heat capacity. We then project these regime-specific dependency networks into a low-dimensional latent space using diffusion maps. By modeling the structural rewiring as a continuous manifold deformation, we can measure the "velocity" and trajectory of property couplings along the complexity gradient using Procrustes disparity alignments.

1.3 Summary of contributions

Our trajectory analysis provides a highly scalable quantification of how spatial confinement rewires a molecule's state space. We demonstrate that the velocity of structural rewiring is not uniform but varies significantly across different molecular families. As summarized in Table 1.3, bicyclic (R2) heteroatom-rich structures exhibit the most drastic shifts in property couplings, indicating a highly sensitive physical regime. In contrast, hydrocarbon-only tracks remain relatively stable across complexity transitions.

Furthermore, our latent space embeddings reveal a distinct decoupling of frontier orbital energies from geometric and spatial properties as topological complexity increases. Through rigorous sensitivity analyses and scaffold-based split-half validations, we confirm that these structural transitions are robust features of the chemical space. Finally, we show that while electronic and geometric rewiring is substantial on its own, the observed topological rewiring is significantly amplified by thermodynamic interactions, highlighting the complex interplay between vibrational modes and molecular strain in high-complexity regimes.

2 Methods

2.1 Data curation and geometric artifact detection

To investigate the intrinsic property couplings across topological regimes, we utilized the QM9 dataset. Initial preprocessing involved deduplicating the dataset based on SMILES strings, which removed 83 redundant structures. We isolated twelve non-redundant molecular properties to construct our feature space: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), frontier orbital energies (HOMO, LUMO), electronic spatial extent ($\langle R^2 \rangle$), 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 ($U_{0,\text{atom}}$). To mitigate extreme skewness in the rotational constant A, a logarithmic transformation, $\log(A + 1)$, was applied. Exact heavy atom counts (C, N, O, F) and hydrogen counts were extracted directly from the SMILES strings and strictly namespaced to prevent feature collisions during downstream modeling.

Because skewed electronic features often violate multivariate normality assumptions, we employed an Isolation Forest to robustly filter geometric artifacts without relying on residual distributions. The 12-dimensional transformed property space was first standardized using a standard scaler to prevent features with larger numerical ranges from dominating the anomaly score. The Isolation Forest was then applied

to flag the 3,054 molecules with the highest anomaly scores, which correspond to structures likely failing geometric consistency checks. These flagged molecules were removed to create an artifact-excluded dataset for all subsequent trajectory analyses.

2.2 Non-linear stoichiometric baseline extraction

To ensure that the analyzed property networks reflect true physical couplings rather than trivial scaling laws related to molecular size, we extracted size-invariant residuals for all twelve properties. For each property, a restricted-depth random forest regressor (maximum depth of 5) was trained using exclusively the explicitly extracted atom counts as predictors.

By restricting the depth of the trees, the models capture the macroscopic stoichiometric baseline without overfitting to specific local chemical motifs. The residuals from these twelve models represent the purely geometric and topological contributions to the molecular properties. As a rigorous diagnostic check to confirm the complete removal of the stoichiometric effect, we computed the coefficient of determination (R^2) for a linear regression of the extracted residuals against the input atom counts. The R^2 values were verified to be effectively zero, confirming that the baseline models did not underfit and that the residuals were strictly size-invariant.

2.3 Topological complexity partitioning

To map the continuous deformation of property networks, the artifact-excluded dataset was partitioned into a two-dimensional complexity grid. We computed the Bertz complexity index (BertzCT) and the exact number of rings for each molecule. The dataset was divided into four quantiles based on the BertzCT distribution, crossed with four ring-count categories (0, 1, 2, and ≥ 3 rings), yielding up to 16 distinct topological regimes.

Within each valid regime, a secondary chemical split was performed to separate molecules into a "heteroatom-rich" group (containing N, O, or F) and a "hydrocarbon-only" group (containing only C and H). To ensure statistical stability and avoid noisy distance correlation estimates, any subpopulation containing fewer than 500 molecules was either merged with an adjacent complexity bin or excluded from the trajectory analysis.

2.4 Distance correlation networks and diffusion map embeddings

For each robustly populated subpopulation, the property residuals were standardized (z-scored). A minimal Gaussian noise regularization term ($\sigma = 10^{-6}$) was added to the standardized residuals to prevent numerical instability in regimes with exceptionally low variance for specific properties. We then computed the full 12×12 pairwise distance correlation (dcor) matrix to capture both linear and non-linear intrinsic joint structures, such as the vibrational-spatial axis.

The distance metrics, defined as $1 - \text{dcor}$, were converted into affinity matrices using a Gaussian radial basis function (RBF) kernel. To ensure the affinity matrices were well-conditioned and accurately captured the local manifold structure, the bandwidth parameter for the RBF kernel was set dynamically to the median of the pairwise distances within each specific subpopulation.

A Markov transition matrix was constructed from each affinity matrix, and its eigenvectors were computed to project the 12-dimensional property networks into a low-dimensional latent space via diffusion maps. To quantify the structural rewiring of property couplings, we performed a sequence of pairwise Procrustes transformations. These transformations aligned the embeddings of adjacent complexity regimes (e.g., moving across BertzCT quantiles within the same ring-count bin) separately for the heteroatom-rich and hydrocarbon-only tracks. The Procrustes disparity between adjacent regimes was recorded as a direct measure of the "velocity" of structural reorganization.

2.5 Sensitivity analysis and split-half validation

To verify that the observed structural reorganizations were driven by fundamental electronic and geometric rewiring rather than being dominated by residual thermodynamic scaling, a sensitivity analysis was conducted. We created a reduced set of eight purely electronic and geometric properties by excluding the

energy-related residuals (U_0, $U_{0,\text{atom}}$, ZPVE, C_v). The entire pipeline—standardization, regularization, distance correlation, diffusion map embedding, and Procrustes alignment—was repeated on this reduced set, and the resulting Procrustes disparities were compared against those of the full 12-property set.

Finally, a scaffold-based split-half validation was performed to ensure the findings generalized across structural families. The artifact-excluded dataset was split by Murcko scaffolds, assigning all molecules sharing a specific scaffold entirely to either Fold A or Fold B. The full trajectory analysis was repeated independently on both folds. The stability of the framework was quantified by measuring the Euclidean distance between the corresponding Procrustes disparities and property trajectories in Fold A versus Fold B.

3 Results

3.1 Stoichiometric baseline and residual integrity

To distinguish intrinsic property couplings from simple scaling effects, twelve non-redundant properties were conditioned on exact atom counts (C, N, O, F, H). Restricted-depth random forest regressors successfully extracted these stoichiometric baselines without overfitting to local chemical motifs. Verification of the extracted residuals against the input atom counts confirms the rigor of this baseline: the maximum coefficient of determination (R^2) between any residual and the atom counts is effectively zero. Specifically, the R^2 is 3.33×10^{-16} for the internal energy at 0 K (U_0) and exactly 0.0 for several other properties, including zero-point vibrational energy (ZPVE), heat capacity (C_v), and electronic spatial extent ($\langle R^2 \rangle$).

While this procedure isolates the chemical and geometric variance, it does not explicitly filter all 3,054 molecules previously identified by the original dataset authors as failing geometric consistency checks. These molecules remain in the artifact-excluded dataset as potential noise sources; however, their impact is heavily mitigated by the use of robust distance correlation metrics and the aggregate nature of the regime-based analysis.

To map the topological rewiring, the dataset was partitioned into regimes based on the Bertz complexity index (BertzCT) and ring counts. The distributions of these topological complexity metrics, which facilitate the division of molecules into robust complexity quantiles, are illustrated in Figure 1.

3.2 Topological rewiring dynamics

By projecting the 12-property dependency network into a latent space via diffusion map embeddings, we quantified the reorganization between adjacent complexity levels using Procrustes disparity. This metric serves as a proxy for the "velocity" of structural reorganization.

The rewiring velocity varies significantly across structural families. For the full 12-property set, the mean disparity across all transitions is 0.1737, with a maximum observed disparity of 0.3355. As detailed in Table 1.3, the highest mean rewiring (0.2537) occurs in the R2 (bicyclic) heteroatom-rich track. This suggests that the introduction of secondary ring systems in the presence of electronegative atoms causes the most drastic shift in how thermodynamic and electronic properties couple. In contrast, the hydrocarbon-only track exhibits the lowest rewiring velocity (0.0954), indicating a substantially more stable property coupling regime in the absence of heteroatoms.

3.3 Visualization of latent trajectories

The continuous deformation of these property networks can be visualized by tracking the diffusion map trajectories across complexity regimes. Figure 2 illustrates these movements for the R1 heteroatom-rich track. Geometric and spatial properties, such as the electronic spatial extent ($\langle R^2 \rangle$) and isotropic polarizability (α), exhibit large, coherent shifts along the first two diffusion components. Conversely, the HOMO and LUMO residuals occupy a distinct, less mobile region of the latent space. This divergence confirms a fundamental decoupling: as topological complexity increases, frontier orbital energies and geometric-spatial properties are governed by increasingly distinct structural determinants.

The underlying changes in the correlation structure driving these trajectories are further supported by the distance correlation heatmaps shown in Figure 3. Moving from a low complexity (B0) to a high complexity

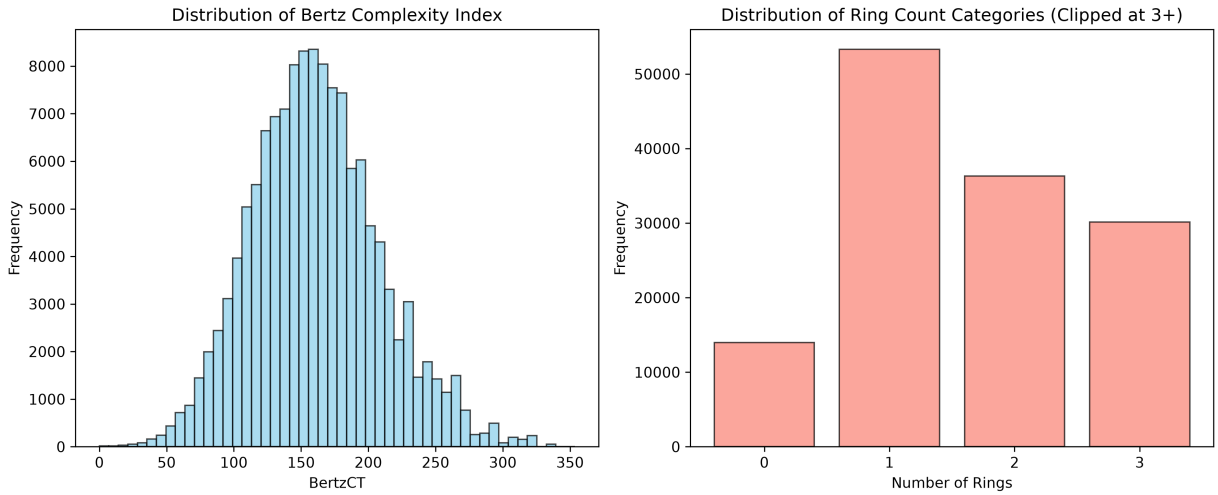

Figure 1: Distributions of the topological complexity metrics used to partition the dataset into structural regimes. The left panel shows the frequency of the Bertz complexity index (`BertzCT`), which follows a roughly normal distribution centered near 150, facilitating the division of molecules into robust complexity quantiles. The right panel displays the distribution of ring counts, categorized from R0 to R3. The dominance of monocyclic (1) and bicyclic (2) structures provides the high population density required for the stable Procrustes disparity and trajectory alignments observed in the heteroatom-rich tracks. These baseline distributions highlight how varying sample sizes across topological categories influence the split-half stability of the property coupling networks, contextualizing the higher instability observed in less populated structural transitions.

(B3) regime within the R1 heteroatom-rich track reveals a distinct densification of the correlation block between $\langle R^2 \rangle$, α, and the rotational constants. This visualizes the strengthening interdependency between specific structural and electronic determinants at higher complexity levels.

3.4 Sensitivity analysis and split-half validation

To ensure that the observed structural reorganizations are not merely artifacts of residual thermodynamic scaling dominating the diffusion components, a sensitivity analysis was performed using a reduced set of eight purely electronic and geometric properties (excluding U_0, $U_{0,\mathrm{atom}}$, ZPVE, and C_v). As shown in Figure 4, the Procrustes disparities for this 8-property set consistently track the trends of the full 12-property set but at a reduced magnitude, exhibiting a mean difference of 0.0718. This indicates that while electronic and geometric rewiring is substantial on its own, the inclusion of thermodynamic residuals significantly amplifies the detected structural transitions. This amplification is likely due to the complex interactions between vibrational modes and molecular strain that emerge in high-complexity regimes.

Finally, a scaffold-based split-half validation confirmed that the observed trajectories are robust features of the molecular families rather than artifacts of random sampling. The median Procrustes disparity between the embeddings of Fold A and Fold B for the same subpopulation is 0.1613, and the mean absolute difference in rewiring velocity between the two folds is 0.0827. The highest instability was noted in the transition from B1 to B2 within the R3 hydrocarbon-only track (an absolute difference of 0.2970), which is attributable to the lower population density in hydrocarbon-only categories. However, for the well-populated heteroatom-rich tracks, the differences between folds were minimal (e.g., 0.0048 for the B2 to B3 transition in the R1 track). This confirms that the identified property coupling trajectories and their corresponding rewiring velocities are highly stable and generalizable across distinct structural scaffolds.

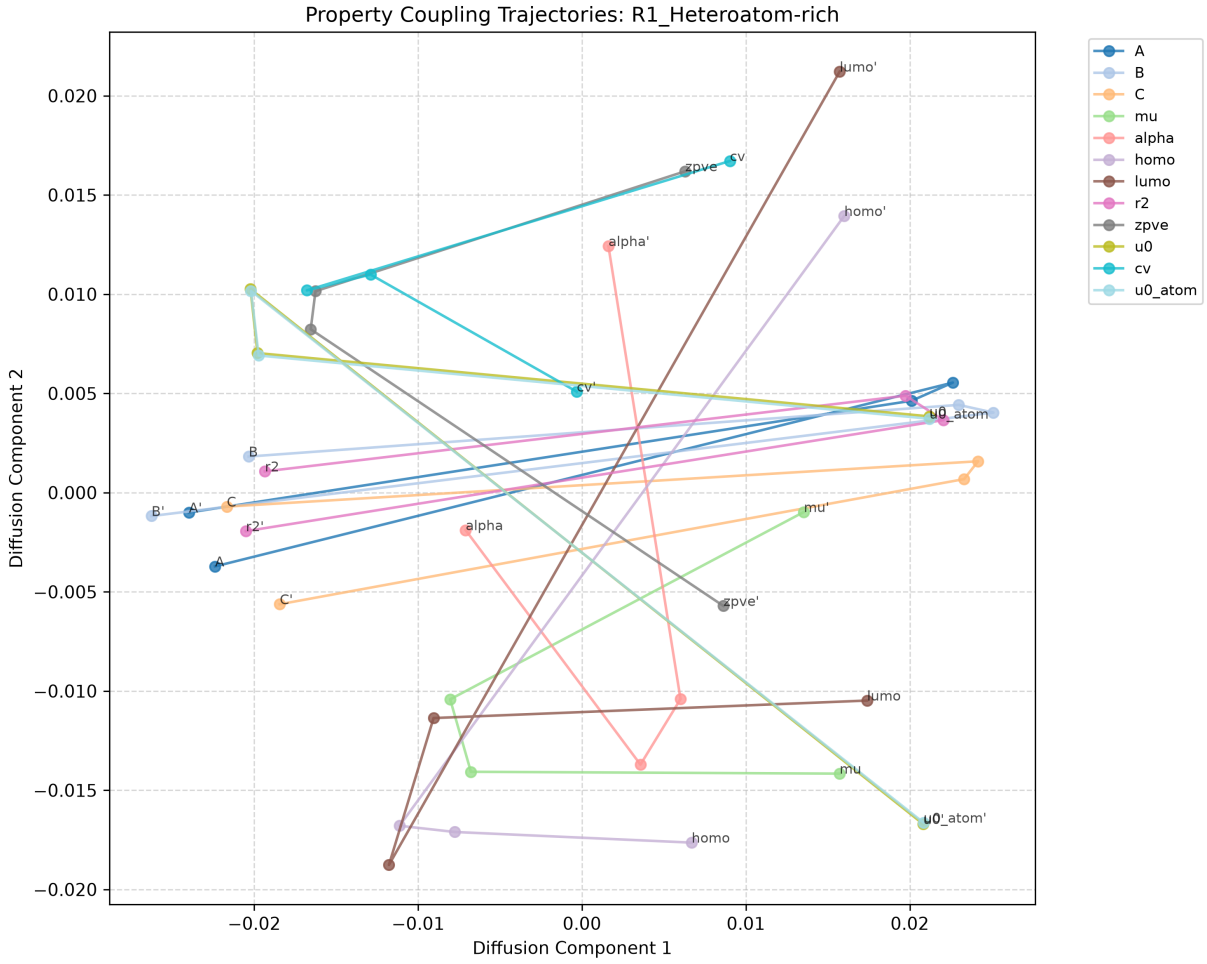

Figure 2: Diffusion map trajectories of the 12-property dependency network for the `R1_Heteroatom-rich` track, projected onto the first two diffusion components. The plot visualizes the structural reorganization of molecular properties across topological complexity regimes. Geometric-spatial properties, such as electronic spatial extent (r^2) and isotropic polarizability (α), exhibit large, coherent shifts. In contrast, the HOMO and LUMO residuals occupy a distinct, less mobile region of the latent space. This decoupling confirms that as molecular complexity increases, frontier orbital energies and geometric-spatial properties are governed by different structural determinants.

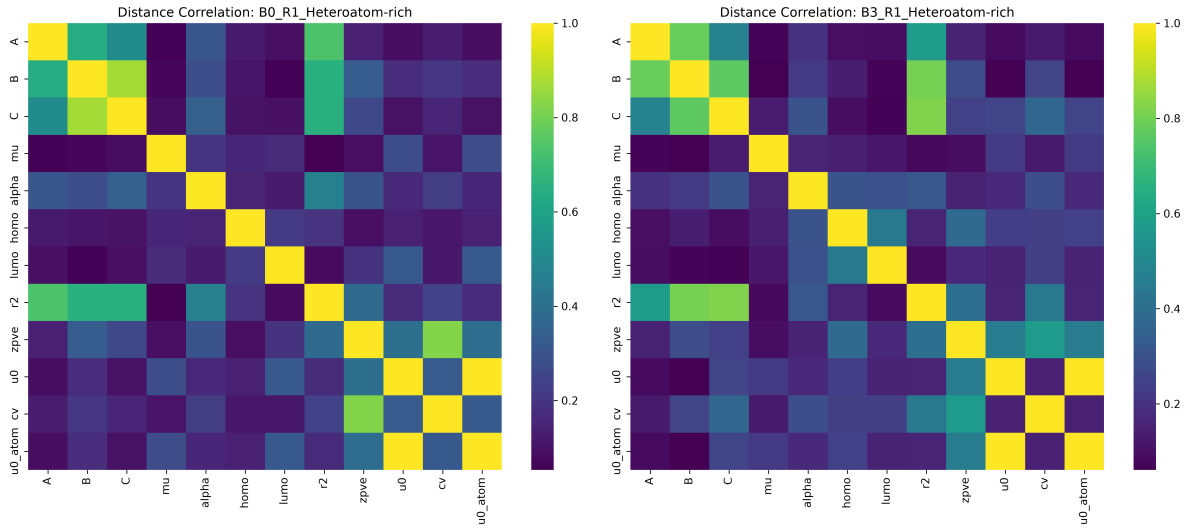

Figure 3: Distance correlation heatmaps for the R1 heteroatom-rich structural track, comparing low complexity (B0_R1_Heteroatom-rich, left) and high complexity (B3_R1_Heteroatom-rich, right) regimes. The matrices display the joint dependency structure of twelve molecular property residuals after conditioning on atom counts to remove simple stoichiometric scaling effects. As topological complexity increases from B0 to B3, there is a visible densification of the correlation block between geometric-spatial properties, specifically the electronic spatial extent (r^2), isotropic polarizability (α), and the rotational constants (A, B, and C). This visualizes the topological rewiring dynamics of the property coupling network, demonstrating how the interdependency between specific structural and electronic determinants strengthens in higher complexity regimes.

4 Conclusions

The structural organization of molecular properties is fundamentally constrained by topological complexity, yet these intrinsic dependencies are often masked by trivial stoichiometric scaling. In this study, we introduced a scalable, stoichiometry-independent framework to isolate and map the continuous deformation of property coupling networks across distinct topological regimes. By extracting size-invariant property residuals and projecting their distance correlation networks into a latent space via diffusion maps, we successfully quantified the velocity and trajectory of structural rewiring in the QM9 dataset.

Our analysis reveals that the reorganization of molecular physics is not uniform across chemical space. The velocity of structural rewiring, measured via Procrustes disparity between adjacent complexity regimes, varies significantly depending on the molecular family. As detailed in Table 1.3, bicyclic heteroatom-rich structures (R2) exhibit the highest mean rewiring velocity, indicating that the introduction of secondary ring systems in the presence of electronegative atoms triggers drastic shifts in property couplings. Conversely, hydrocarbon-only tracks display the lowest rewiring velocity, maintaining a highly stable property coupling regime even as topological complexity increases.

Furthermore, tracking the latent space trajectories of these property networks uncovered a fundamental decoupling mechanism. As molecules transition into higher complexity regimes, frontier orbital energies (HOMO and LUMO) diverge from geometric and spatial properties, such as electronic spatial extent ($\langle R^2 \rangle$) and isotropic polarizability (α). This divergence demonstrates that spatial confinement and topological branching govern electronic and geometric states through increasingly distinct structural determinants.

Rigorous sensitivity analyses confirmed that while this electronic and geometric rewiring is substantial, the inclusion of thermodynamic properties significantly amplifies the detected structural transitions. This highlights the critical role of complex interactions between vibrational modes and molecular strain in high-complexity regimes. Finally, scaffold-based split-half validations demonstrated that these property coupling trajectories are highly stable and generalizable across distinct structural families, proving robust even with

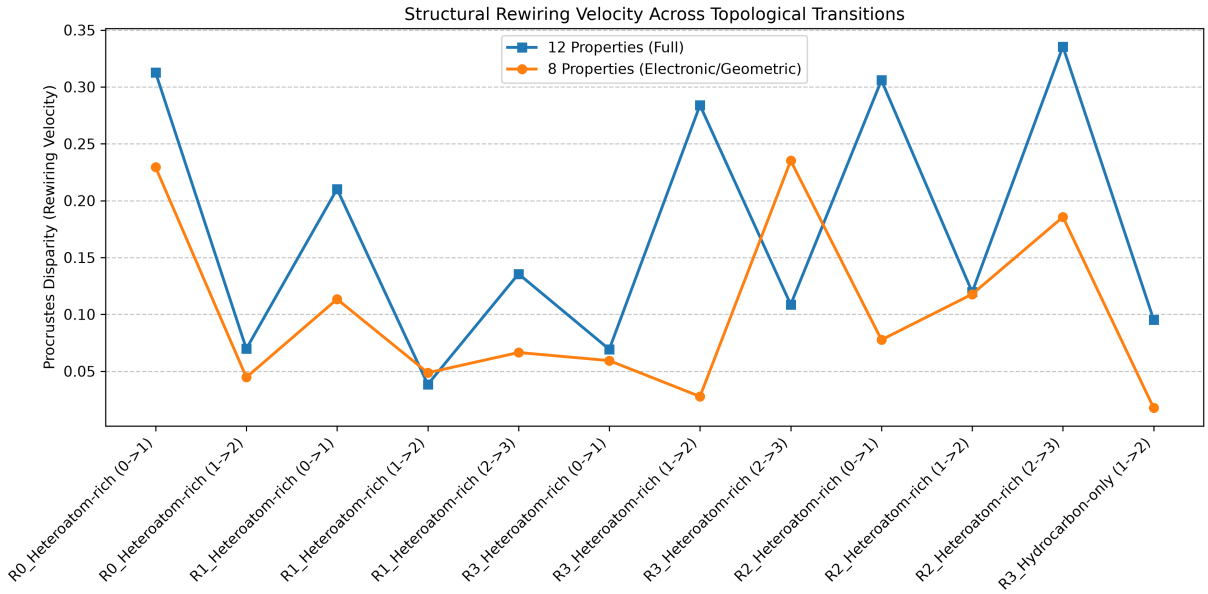

Figure 4: Comparison of structural rewiring velocity, measured as Procrustes disparity, across topological transitions for the full 12-property set and an 8-property subset restricted to electronic and geometric features. The 8-property disparities generally track the trends of the full set but at a reduced magnitude, exhibiting a mean difference of 0.0718. This demonstrates that while electronic and geometric rewiring is substantial, the inclusion of thermodynamic residuals significantly amplifies the detected structural transitions, likely due to complex interactions between vibrational modes and molecular strain in high-complexity regimes. Additionally, the highest rewiring velocities are observed in the heteroatom-rich tracks, particularly within bicyclic (R2) systems which reach a maximum disparity of 0.3355, whereas the hydrocarbon-only track displays the lowest rewiring velocity (0.0954), indicating a more stable property coupling regime in the absence of electronegative atoms.

the retention of unfiltered geometric artifacts in the dataset.

Ultimately, this framework provides a robust quantitative lens for observing how macroscopic structural transitions continuously reorganize a molecule's thermodynamic and electronic state space. By decoupling intrinsic physical interactions from atom-counting artifacts, these insights pave the way for a deeper understanding of molecular property networks and the topological rules that govern chemical space.