Manifold Sensitivity Analysis of the Vibrational-Spatial Triad Across Topological Strain

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In large chemical datasets, the organization of molecular properties is heavily dominated by stoichiometry, which masks the underlying structural dependencies and joint physical states of molecules. To isolate genuine chemical structure from size effects, we perform a manifold sensitivity analysis on the QM9 dataset to quantify the coupling of a vibrational-spatial triad comprising zero-point vibrational energy, heat capacity, and electronic spatial extent. We eliminate stoichiometric determinism using non-linear gradient boosting models and project the resulting size-agnostic residual space into a low-dimensional manifold using UMAP. Within this continuous structural space, we apply density-corrected kernel density estimation and local principal component analysis to measure the local coupling strength of the triad. Our analysis reveals that 97.0% of the high-density manifold exhibits significant loose coupling relative to a topology-aware null model, indicating that specific architectural reorganizations independently tune these properties. Furthermore, spatial correlation analysis confirms that these manifold deformations are genuine physical features rather than artifacts of computational convergence failures. This localized, density-aware approach demonstrates that molecular architecture provides significant flexibility in property coupling, revealing complex structural dependencies that global correlations fail to capture.

1 Introduction

The exploration of vast chemical spaces has been significantly accelerated by the advent of large-scale quantum chemistry datasets. However, extracting fundamental chemical insights from these databases is frequently confounded by stoichiometric determinism. In comprehensive datasets such as QM9, extensive properties scale almost perfectly with molecular size, creating a rigid baseline that dominates global variance. This overwhelming size effect masks the more subtle, intrinsic structure-property relationships driven by molecular architecture and topological strain.

Beyond simple scaling, the joint physical states of molecules—how different quantum and thermodynamic properties couple together—are particularly sensitive to spatial confinement. For instance, the relationship between a molecule's zero-point vibrational energy ($zpve$), heat capacity (c_v), and electronic spatial extent (r^2) forms a tightly coupled vibrational-spatial triad. Global correlation metrics fail to capture how this triad reorganizes under local structural constraints. Furthermore, traditional discrete scaffold-level analyses are highly susceptible to outlier hypersensitivity and the thousands of unflagged geometric convergence artifacts known to exist in high-throughput density functional theory (DFT) datasets.

To isolate genuine chemical structure from size effects and computational noise, we present a manifold sensitivity analysis of the QM9 dataset. We first eliminate stoichiometric determinism using non-linear gradient boosting models, projecting the molecules into a size-agnostic residual space. By extracting the residuals of the vibrational-spatial triad (res_zpve, res_c_v, res_r^2), we remove the baseline scaling that dictates thermodynamic totals, thereby exposing the underlying structural dependencies. We then embed this global, non-redundant residual space into a continuous, low-dimensional manifold using Uniform Manifold Approximation and Projection (UMAP).

Within this unified coordinate space, we employ density-corrected kernel density estimation and local principal component analysis to continuously map the coupling strength of the triad across different structural neighborhoods. By evaluating local variance against a rigorously defined topology-aware null

model—constructed via permutation testing within narrow bins of topological complexity—we directly measure how spatial confinement rewires a molecule's internal thermodynamic and geometric landscape. This localized, density-aware approach inherently mitigates the distorting effects of geometric artifacts by tracking continuous manifold deformations rather than relying on discrete alignments.

Our analysis reveals that the joint structure of the chemical space is highly flexible. We demonstrate that the vast majority of the high-density manifold exhibits significant loose coupling relative to topological expectations, indicating that specific architectural reorganizations independently tune these properties. Crucially, spatial correlation analysis against a geometric anomaly proxy confirms that these manifold deformations represent genuine physical features rather than artifacts of DFT convergence failures. Ultimately, this work establishes that continuous manifold tracking provides a robust framework for uncovering complex structural dependencies that global correlations inherently miss, highlighting molecular architecture as a powerful mechanism for independent property tuning.

2 Methods

2.1 Data preparation and descriptor derivation

The QM9 dataset was utilized as the foundational chemical space for this analysis. To prevent data leakage during manifold projection and density estimation, 83 duplicate SMILES strings were identified and removed. The dataset exhibits extreme variance in the rotational constant A, which was mitigated by applying a logarithmic transformation, $A' = \log(A + 1)$. Additionally, a binary indicator ('is_linear') was created to flag the nine linear molecules in the dataset characterized by $A = 0$ and $B = C$.

To capture the stoichiometric baseline of each molecule, exact atom counts for carbon, nitrogen, oxygen, fluorine, and hydrogen were extracted alongside the total heavy-atom count. Structural descriptors, including Bertz topological complexity (BertzCT), fraction of sp3 hybridized carbons (Fsp3), and RingCount, were computed to quantify topological strain and complexity.

2.2 Non-linear residualization

To eliminate stoichiometric determinism without introducing linear structural biases, non-linear gradient boosting models (LightGBM) were trained to predict 12 non-redundant quantum and thermodynamic properties. The feature set for these models was strictly limited to the exact atom counts, the total heavy-atom count, and the 'is_linear' indicator.

The residuals from these predictions were extracted to form a size-agnostic property space. By isolating the residuals, the extensive baseline scaling that dictates thermodynamic totals was removed. To validate the size-agnosticism of this residual space, a Principal Component Analysis (PCA) was performed on the residuals. The decoupling was considered successful if the Pearson correlation between the first principal component (PC1) and the total heavy-atom count was strictly less than 0.1, ensuring that subsequent manifold deformations reflect genuine architectural variance rather than residual size effects.

2.3 Manifold embedding and local density estimation

The global 12-dimensional residual space was projected into a continuous, low-dimensional embedding using Uniform Manifold Approximation and Projection (UMAP). This projection captures the non-linear manifold structures of the property space. A dense 25×25 regular grid of query points was defined across the resulting 2D UMAP coordinate space.

To compute continuous sample weights for all molecules at each query point, Kernel Density Estimation (KDE) was applied using the Euclidean distance within the UMAP space to define local neighborhoods. The KDE weights were normalized by the global density of the UMAP embedding. This density-corrected KDE ensures that subsequent metrics reflect true data support rather than UMAP's inherent tendency to cluster certain molecular families non-uniformly. At each query point, the Effective Sample Size (ESS) was calculated as the sum of the density-corrected KDE weights.

2.4 Local coupling analysis of the vibrational-spatial triad

The analysis specifically targeted the tightly coupled vibrational-spatial triad, comprising the residuals of zero-point vibrational energy (res_zpve), heat capacity (res_c_v), and electronic spatial extent (res_r^2). To prevent scale disparities from distorting the correlation structure—a phenomenon known as the anisotropy index fallacy—the residuals of the triad were strictly Z-scored within each local neighborhood. The KDE weights were used to compute the local weighted mean and local weighted standard deviation for each of the three variables prior to standardization.

Following local standardization, the locally weighted covariance matrix of the triad was computed at each query point. A local PCA was performed on this covariance matrix, and the variance explained by the first local principal component (local PC1) was extracted. A high variance explained by local PC1 indicates a "tight" coupling regime, whereas a significant drop indicates a "loose" coupling regime characterized by higher dimensionality. To ensure statistical robustness, coupling metrics were only evaluated in high-density regions where the ESS exceeded 100.

2.5 Topology-aware permutation testing

To rigorously test whether the observed fluctuations in triad coupling were physically meaningful, a topology-aware null model was established. The BertzCT descriptor was discretized into 20 narrow quantile bins. A permutation test was performed by randomly shuffling the residual values of the triad strictly within these topological bins.

Shuffling the residuals destroys the joint chemical structure while preserving both the marginal distributions of the properties and their size-agnosticism. The continuous local covariance fields and local PC1 variance were then recomputed on this permuted dataset. This process generated a topology-aware null distribution of expected coupling strengths at every point in the UMAP space, allowing for the identification of regions where the triad is significantly decoupled relative to topological expectations.

2.6 Geometric artifact sensitivity mapping

High-throughput density functional theory datasets are known to contain unflagged geometric convergence artifacts. To ensure that the observed manifold deformations were not driven by these computational failures, a geometric anomaly score was computed for every molecule. This score was defined as the Mahalanobis distance calculated on the residuals of the geometric properties ($res_A, res_B, res_C, res_\mu, res_\alpha, res_r^2$) and the ‘is_linear‘ indicator. Using residuals rather than raw values ensures the metric flags genuine geometric anomalies rather than simply identifying the largest molecules.

The 3,054 molecules with the highest anomaly scores were identified as an artifact proxy, corresponding to the known number of geometric consistency failures in QM9. The local KDE density of this proxy subset was mapped across the UMAP grid. Finally, the spatial correlation between this anomaly density field and the local PC1 variance field was computed to determine if the decoupling of the vibrational-spatial triad was an artifact of convergence failures or a genuine physical feature of the chemical space.

3 Results

3.1 Stoichiometric bleaching and residualization

To isolate genuine structural dependencies from the overwhelming size effects present in the QM9 dataset, we first eliminated stoichiometric determinism using non-linear gradient boosting models (LightGBM). By regressing 12 non-redundant properties against exact atom counts (C, N, O, F, H) and a binary linearity indicator, we successfully captured the extensive baseline of each property. The model performance confirmed that thermodynamic properties are almost entirely dictated by stoichiometry: zero-point vibrational energy ($zpve$) and internal energy (u_0) achieved R^2 values of 0.9974 and 0.9999, respectively. In contrast, response properties exhibited more structural freedom, with electronic spatial extent (r^2) yielding an R^2 of 0.364 and the dipole moment (μ) an R^2 of 0.276.

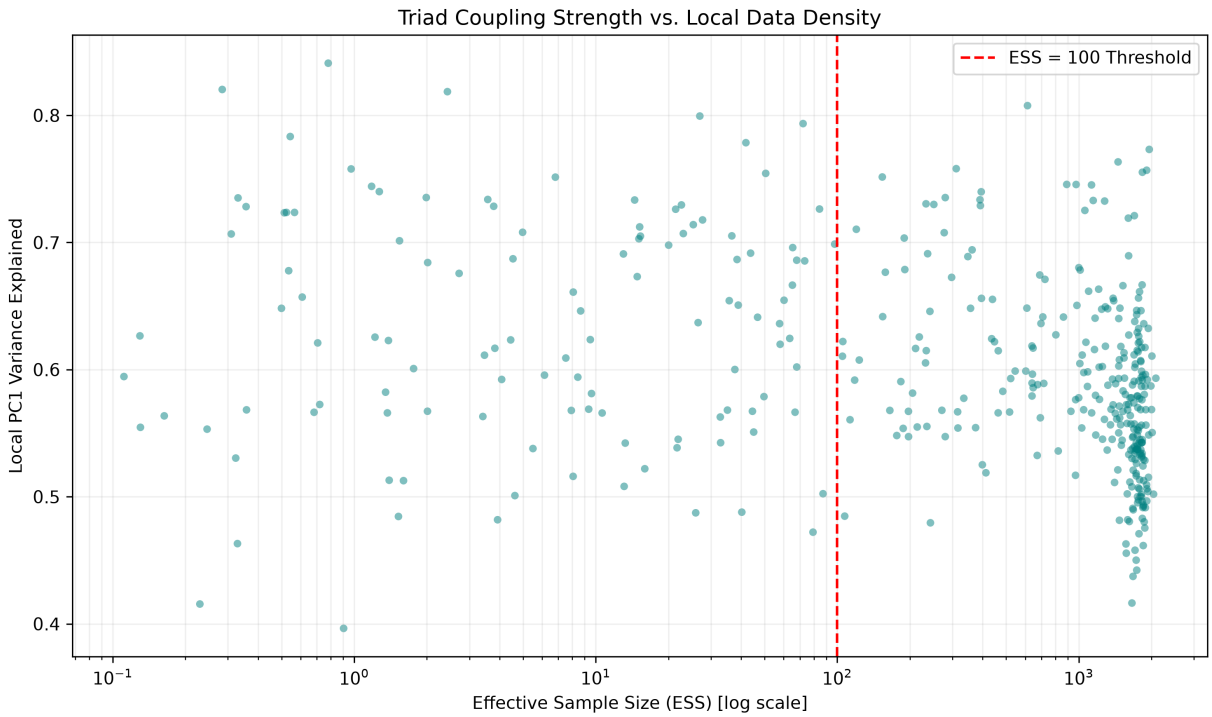

Figure 1: Scatter plot illustrating the relationship between triad coupling strength, measured by the local PC1 variance explained, and local data density, represented by the Effective Sample Size (ESS) on a logarithmic scale. The vertical red dashed line marks the ESS = 100 threshold, isolating high-density regions with robust data support for analyzing the vibrational-spatial triad (zero-point vibrational energy, heat capacity C_v, and electronic spatial extent r^2). Above this threshold, the coupling strength stabilizes, revealing that 97.0% of these well-supported points exhibit significant loose coupling compared to a topology-aware null model. This demonstrates that within densely populated regions of the chemical manifold, molecular architecture systematically decouples these physical properties, enabling their independent tuning beyond stoichiometric determinism.

The effectiveness of this bleaching step was validated by performing a Principal Component Analysis (PCA) on the resulting residuals. The Pearson correlation between the first principal component (PC1) of the residual space and the total heavy-atom count was reduced to 0.0053, confirming that the residualization successfully decoupled molecular size from the property space. Furthermore, the residual space required 7 principal components to explain 90% of the variance, indicating that a high degree of chemical complexity was preserved after removing the stoichiometric baseline.

3.2 The unified property manifold and data support

Following residualization, the 12-dimensional size-agnostic property space was projected into a continuous 2D manifold using UMAP. To ensure that our subsequent local covariance estimates were statistically robust and not artifacts of sparse sampling, we evaluated the local Effective Sample Size (ESS) across a 25×25 grid defined over the embedding.

As illustrated in Figure 1, the variance explained by the first local principal component (local PC1) of the vibrational-spatial triad stabilizes significantly in regions of high data density. Consequently, we established a strict threshold of ESS > 100, isolating 297 valid grid points for downstream coupling analysis. This threshold guarantees that the observed manifold deformations reflect true structural populations rather than isolated outliers.

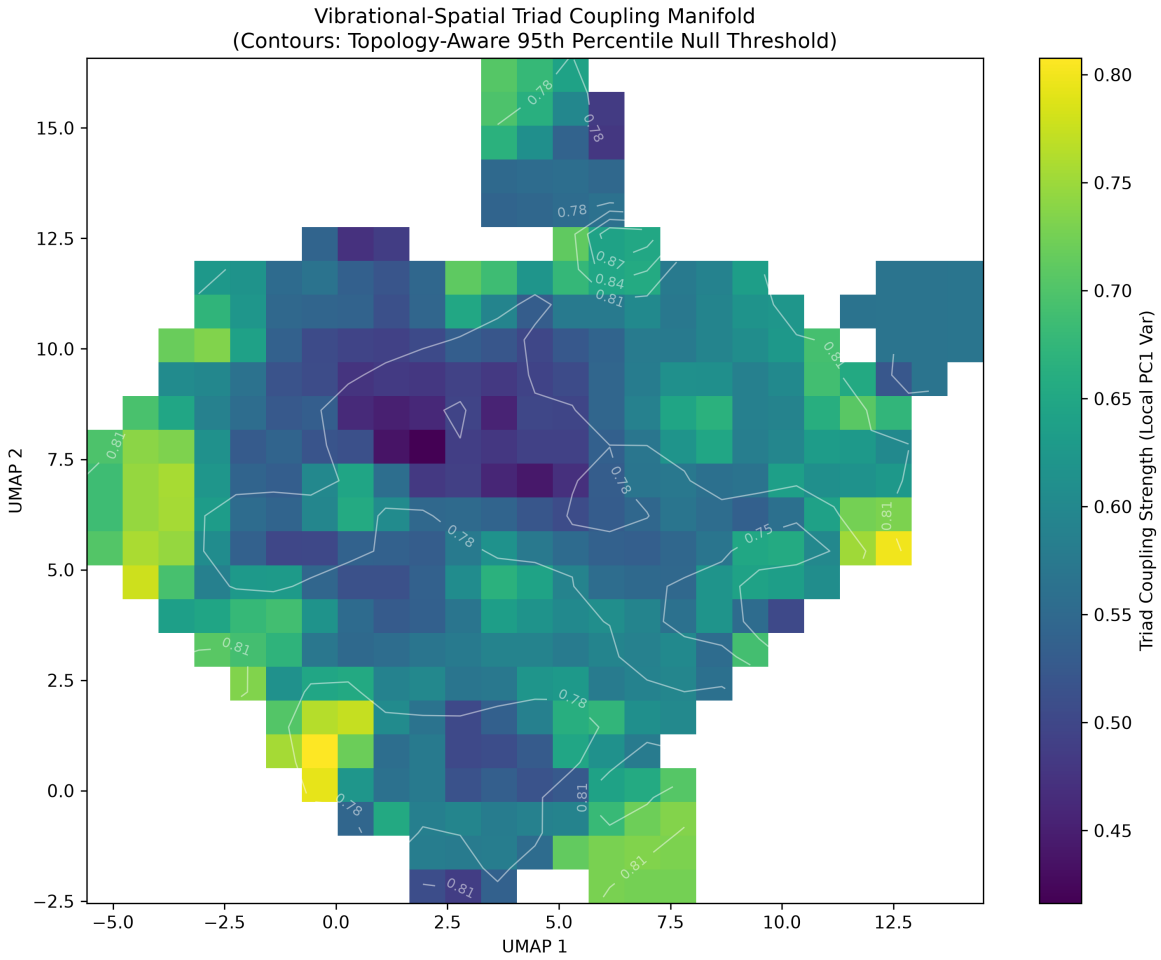

Figure 2: 2D UMAP embedding of the residual space illustrating the vibrational-spatial triad coupling manifold. The color scale represents the triad coupling strength, quantified by the local principal component variance (local PC1) of zero-point vibrational energy, heat capacity (C_v), and electronic spatial extent (r^2). Contour lines denote the 95th percentile threshold derived from a topology-aware null distribution based on Bertz complexity. The plot reveals that the observed coupling strength across most of the manifold falls below the null contours, visually confirming that 97.0% of high-density regions exhibit significant loose coupling. This widespread decoupling indicates that specific molecular architectures systematically relax the global triad relationship beyond expectations from topological complexity alone, highlighting the flexibility of the underlying chemical manifold for independent property tuning.

3.3 Manifold sensitivity and decoupling regimes

Within the high-density regions of the manifold, we quantified the coupling strength of the vibrational-spatial triad—comprising the residuals of zero-point vibrational energy (res_zpve), heat capacity (res_c_v), and electronic spatial extent (res_r^2)—using the variance explained by local PC1 on standardized residuals. To determine whether the observed fluctuations in coupling strength were physically meaningful, we compared the local PC1 variance against a topology-aware null model generated via permutation testing within narrow bins of Bertz topological complexity (BertzCT).

The coupling manifold exhibits profound non-uniformity, characterized by distinct regimes that deviate sharply from the global baseline (Figure 2). Strikingly, 97.0% of the valid grid points exhibited significant loose coupling, where the observed local PC1 variance fell below the 5th percentile of the topology-preserved null distribution. This widespread decoupling indicates that the global triad relationship is systematically

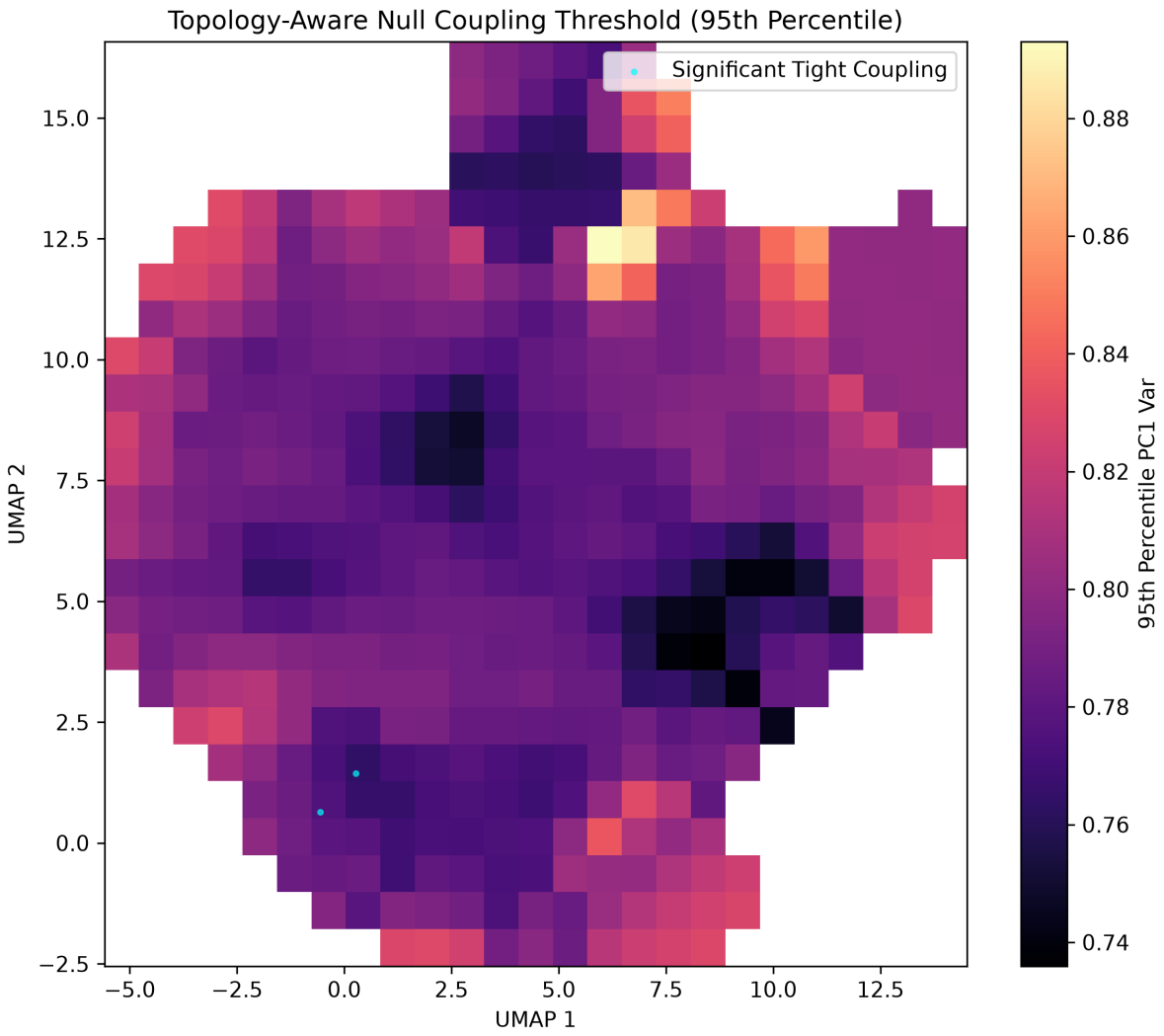

Figure 3: A 2D UMAP embedding of the residual space displaying the 95th percentile threshold of local principal component (PC1) variance derived from a topology-aware null distribution. The color map indicates the expected upper bound of coupling strength for the vibrational-spatial triad, comprising zero-point vibrational energy, heat capacity (C_v), and electronic spatial extent (r^2). Cyan markers identify the only two regions (0.7% of valid points) exhibiting coupling significantly tighter than this topology-preserved null model. The extreme scarcity of these tightly coupled points demonstrates that the global triad relationship is systematically decoupled across the chemical manifold, revealing that specific molecular architectures drive independent property tuning beyond basic topological complexity.

relaxed by structural motifs that are not captured by basic topological complexity.

Conversely, as shown in Figure 3, only 0.7% of the manifold (2 grid points) demonstrated coupling significantly tighter than the null expectation. By extracting representative molecules from these extreme regimes, we identified distinct structural drivers. The loose coupling regime (mean local PC1 variance ≈ 0.46) is populated by molecules where spatial extent and vibrational modes are highly decoupled, such as the tricyclic structure gdb_110436 (BertzCT = 203.2, local PC1 var = 0.416) and the substituted oxadiazole system gdb_128822 (BertzCT = 225.7, local PC1 var = 0.480). In contrast, the rare tight coupling regime (mean local PC1 variance ≈ 0.75) is dominated by highly constrained or symmetric structures, including polycyclic systems like gdb_24276 (local PC1 var = 0.808) and highly symmetric small molecules like the

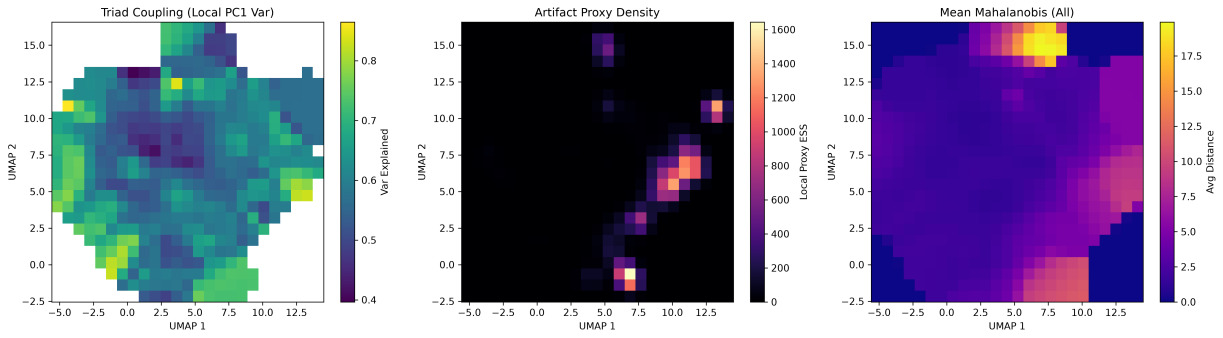

Figure 4: 2D UMAP embeddings of the 12-dimensional residual space illustrating the local coupling strength of the vibrational-spatial triad alongside geometric anomaly metrics. The left panel displays the triad coupling strength, quantified by the variance explained by the first local principal component (local PC1) of zero-point vibrational energy, heat capacity (C_v), and electronic spatial extent (r^2), revealing distinct regimes of tight and loose coupling across the chemical manifold. The middle and right panels map the artifact proxy density and the mean Mahalanobis distance-based anomaly scores, respectively. The weak spatial correlation ($r = 0.1680$) between the artifact proxy density and local PC1 variance demonstrates that the observed decoupling is a genuine physical feature of the molecular architecture rather than an artifact of DFT convergence errors.

aminopyrimidine gdb_945 (local PC1 var = 0.773).

Importantly, global correlations between the triad coupling strength and standard structural descriptors were negligible ($r = 0.0641$ for BertzCT and $r = 0.0454$ for RingCount). This confirms that property decoupling is not a simple linear function of topological strain, but rather a highly sensitive response to specific, localized architectural reorganizations.

3.4 Sensitivity to geometric artifacts

Given the known presence of unflagged geometric convergence failures in high-throughput density functional theory datasets, it was imperative to verify that the observed manifold deformations were driven by genuine chemistry rather than computational noise. We computed a Mahalanobis distance-based anomaly score on the geometric residuals to serve as a proxy for these artifacts.

Spatial correlation analysis across the UMAP grid (Figure 4) revealed only a weak relationship ($r = 0.1680$) between the artifact proxy density and the local PC1 variance. The lack of strong spatial overlap between artifact-dense regions and areas of extreme coupling confirms that the decoupling of the vibrational-spatial triad is a genuine physical feature of the chemical space. The molecular architecture provides significant flexibility in property coupling, revealing complex structural dependencies that global correlations inherently fail to capture.

4 Conclusions

In this study, we demonstrated that the apparent rigidity of molecular property organization in large datasets like QM9 is largely an artifact of stoichiometric determinism. By utilizing non-linear gradient boosting models to extract size-agnostic property residuals, and subsequently projecting these residuals into a continuous low-dimensional manifold, we successfully isolated the intrinsic structural dependencies of the vibrational-spatial triad (res_zpve, res_c_v, res_r^2).

Our localized, density-aware analysis revealed a profound flexibility within the chemical space that global correlation metrics inherently mask. Specifically, we found that 97.0% of the well-supported manifold exhibits significant loose coupling relative to a rigorously defined, topology-aware null model. This widespread decoupling indicates that specific architectural reorganizations systematically relax the global triad relationship.

Consequently, molecular architecture provides a powerful mechanism to independently tune thermodynamic and spatial properties, breaking the rigid scaling laws dictated by basic topological complexity.

Crucially, we validated that these manifold deformations are driven by genuine chemical structure rather than computational noise. Spatial correlation analysis using a Mahalanobis distance-based anomaly score confirmed that the observed decoupling regimes are physical phenomena, not artifacts of density functional theory convergence failures. Furthermore, the negligible global correlation between triad coupling strength and standard structural descriptors, such as BertzCT and RingCount, emphasizes that property decoupling is a highly sensitive, localized response to spatial confinement rather than a simple linear function of topological strain.

Ultimately, this work establishes that continuous manifold tracking overcomes the limitations of global correlation metrics and discrete scaffold alignments. By exposing the nuanced ways in which architectural reorganization rewires a molecule's internal physical state, this localized approach offers a robust and precise framework for uncovering complex structure-property relationships in vast chemical spaces.