Gromov-Wasserstein Alignment of Multivariate Property Networks Across Heteroatom Boundaries

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The joint organization of molecular properties is often assumed to be a static consequence of chemical structure, yet it remains unclear how the introduction of specific heteroatoms fundamentally rewires this physical state space independently of molecular size and stoichiometry. To rigorously quantify this structural rewiring, we analyze the QM9 dataset by extracting non-linear, atom-count-conditioned residuals using gradient-boosted regressors for eight diverse physical properties. We then construct distance correlation matrices for distinct compositional families—such as pure hydrocarbons, nitrogen-only, and oxygen-only populations—and deploy Gromov-Wasserstein optimal transport to compute the exact manifold alignment costs between these boundaries. Our results demonstrate that heteroatoms induce significant deformations in the property dependency network compared to a pure hydrocarbon baseline. The most severe disruptions occur in the geometric-spatial coupling between the rotational constant and electronic spatial extent, as well as in frontier orbital relationships. Additionally, the magnitude of this network rewiring correlates strongly with heteroatom-induced symmetry breaking and increased dipole moments. Finally, participation ratio analyses reveal that while the highest occupied molecular orbital energy remains statistically isolated from the global property network in pure hydrocarbons, the introduction of nitrogen partially reintegrates it. These findings provide a rigorous mathematical mapping of how chemical identity reorganizes global property couplings beyond simple stoichiometric scaling.

1 Introduction

The prediction and understanding of molecular properties lie at the heart of computational chemistry. Traditionally, physical properties—ranging from thermodynamic variables to quantum mechanical observables—are analyzed either in isolation or as direct functions of molecular size and stoichiometry. However, a molecule is not merely a collection of independent features; its properties form a highly coupled, multivariate physical state space. While it is well understood that the introduction of heteroatoms alters individual properties, it remains unclear how these compositional shifts fundamentally rewire the joint dependency network of the molecule, independent of the trivial scaling effects associated with increased mass or atom counts.

Isolating intrinsic chemical physics from stoichiometric determinism presents a significant mathematical challenge. In large chemical spaces, such as the QM9 dataset, the dominant variance in many properties is driven by the sheer number of heavy atoms. Consequently, the underlying structural anomalies and non-linear couplings—such as the interplay between geometric rigidity, electronic spatial extent, and frontier orbital energies—are often masked by size-induced correlations. To rigorously quantify how chemical identity reorganizes global property couplings, one must first strip away the deterministic baseline of atom counts and then deploy a metric capable of comparing the topology of disparate relational manifolds.

In this work, we introduce a framework based on Gromov-Wasserstein optimal transport to measure the exact manifold alignment costs between distinct compositional boundaries. We partition the QM9 dataset into mutually exclusive compositional families, including pure hydrocarbons, nitrogen-only, and oxygen-only populations. To rigorously isolate structural signals, we extract non-linear, atom-count-conditioned residuals for eight physically diverse properties—such as the rotational constant (B), dipole moment (μ), spatial extent (r^2), and frontier orbital energies—using gradient-boosted regressors. By computing distance correlation matrices on these residuals, we capture the intrinsic, non-linear joint structure of the physical properties for each family. The Gromov-Wasserstein optimal transport formulation is then deployed directly

on these matrices, providing a mathematically rigorous distance that quantifies the deformation of the property dependency network across heteroatom boundaries.

Our analysis reveals that the joint organization of molecular properties is not a static consequence of chemical structure but a dynamic manifold that deforms significantly upon the introduction of heteroatoms. We demonstrate that heteroatoms induce severe disruptions in specific property axes, most notably the geometric-spatial coupling between the rotational constant and electronic spatial extent, as well as frontier orbital relationships. Furthermore, we show that the magnitude of this network rewiring correlates strongly with heteroatom-induced symmetry breaking and increased dipole moments. Finally, through participation ratio analyses, we uncover the statistical isolation of the highest occupied molecular orbital (HOMO) energy in pure hydrocarbons and map its partial reintegration into the global property network driven by nitrogen. These findings provide a quantitative mapping of how chemical identity reorganizes global property couplings beyond simple stoichiometric scaling, establishing a new lens through which to view the physical state space of small molecules.

2 Methods

2.1 Data curation and property selection

To establish a rigorous baseline for property network analysis, we utilized the QM9 dataset. To prevent data leakage from stereoisomers that collapse to identical molecular graphs, the dataset was deduplicated based on canonical SMILES strings generated via RDKit. To eliminate high-dimensional noise caused by thermodynamic redundancy and to avoid over-representing the vibrational-energetic axis, we restricted our analysis to a minimalist, physically diverse set of eight properties: the rotational constant (B), dipole moment (μ), isotropic polarizability (α), highest occupied molecular orbital energy (HOMO), lowest unoccupied molecular orbital energy (LUMO), electronic spatial extent (r^2), zero-point vibrational energy ($zpve$), and heat capacity at 298.15 K (c_v). The LUMO energy was selected in place of the highly correlated internal energy at 0 K (u_0) to maintain feature diversity. A $\log(B+1)$ transformation was applied to the rotational constant to tame extreme variance while retaining unflagged geometric artifacts.

The deduplicated dataset was partitioned into mutually exclusive compositional families based on exact atom counts: Pure Hydrocarbons (HC), Nitrogen-only (containing C, H, and N), Oxygen-only (containing C, H, and O), Fluorine-only, and mixed heteroatom cohorts. Only compositional families containing at least 1,000 molecules were retained for downstream statistical robustness.

2.2 Geometric strain and structural descriptors

To capture molecular rigidity and symmetry, 2D structural descriptors including the fraction of sp^3 hybridized carbons, ring count, and topological symmetry classes were computed for each molecule. As a direct geometric proxy for molecular shape symmetry, we calculated the Inertial Tensor Ratio derived from the rotational constants, with specific handling for linear molecules where the rotational constant approaches zero or extreme values.

To address missing geometric context and identify potential artifacts, a 3D conformer was generated for each SMILES string using distance geometry embedding, followed by optimization with the MMFF94 force field. The final strain energy was recorded. Molecules that either failed to embed or fell within the top 5% of MMFF94 strain energy within their respective families were flagged. This flag served as a proxy to verify whether observed network rewirings were driven by genuine chemical physics or underlying geometric inconsistencies.

2.3 Stoichiometric residual extraction

As outlined in the introduction, isolating intrinsic chemical physics from stoichiometric determinism is critical. To achieve this, we trained GPU-accelerated LightGBM regressors using 5-fold cross-validation to predict the eight selected properties. The feature space was strictly limited to the exact counts of C, N, O, F, and H atoms.

Stoichiometric residuals were calculated as the difference between the actual property value and the out-of-fold prediction. This out-of-fold approach rigorously prevents data leakage. The resulting residuals were then standardized (z-scored) within each compositional family. By operating exclusively on these standardized residuals, subsequent analyses measure the intrinsic, non-linear structural variations independent of molecular size and heavy-atom count.

2.4 Network construction and distance correlation

For each compositional family, we constructed an 8×8 pairwise distance correlation ($dCor$) matrix on the standardized residuals. Distance correlation was chosen over linear metrics to capture non-linear and non-monotonic dependencies between physical properties.

To rigorously distinguish between a mere loss of correlation and a true loss of statistical dependency, we validated the network structures by computing the normalized Hilbert-Schmidt Independence Criterion (HSIC) matrix for the same property pairs. The HSIC was computed using a robust Gaussian Radial Basis Function (RBF) kernel, employing a median heuristic to determine the optimal bandwidth. These memory-intensive matrix computations were parallelized across families.

2.5 Gromov-Wasserstein manifold alignment

To quantify the deformation of the property dependency network across heteroatom boundaries, we deployed Gromov-Wasserstein (GW) optimal transport to align the 8×8 $dCor$ matrices of each heteroatom family to the Pure Hydrocarbon (HC) baseline. The cost matrix for the transport problem was defined as the absolute difference in correlation strength between the source and target matrices, intuitively highlighting where relationships strengthened or weakened.

To ensure that the observed GW distances represented true chemical rewiring rather than statistical artifacts of sample size or atom-count distributions, we established a rigorous null baseline. We performed stratified subsampling on the HC cohort to exactly match both the size and the heavy-atom count distribution of each target heteroatom family. The GW distance was then computed between these disjoint HC subsets. Additionally, a strain sensitivity check was performed by re-computing all GW alignments after excluding the molecules flagged for high MMFF94 strain or failed embeddings, ensuring the robustness of the measured topological shifts.

2.6 Transport cost decomposition and symmetry analysis

The global GW transport cost was decomposed to identify the specific property pairs driving the network rewiring. This was achieved by computing the element-wise product of the optimal coupling matrix and the cost matrix. These local contributions were normalized and summed to rank the most severely disrupted property relationships.

To map these mathematical shifts back to physical molecular features, we tested the hypothesis of heteroatom-induced symmetry breaking. We computed the correlation between the magnitude of the GW distance for each family and its average change in molecular symmetry relative to the HC baseline, using the Inertial Tensor Ratio and dipole moment (μ) as primary proxies.

2.7 Participation ratio and permutation testing

To investigate the topological role of specific properties within the network, we computed the Participation Ratio (PR), defined as the sum of squared connections for a given property's row (or column) in the $dCor$ matrix.

To definitively test whether the observed statistical isolation of the HOMO energy was a genuine chemical phenomenon or merely a statistical artifact of its variance in heteroatom spaces, we performed a rigorous sensitivity check. We computed the PR for a known stable property, *zpve*, within the same families to serve as a control. Finally, we validated the HOMO decoupling via a permutation test: the HOMO residual vector was shuffled 1,000 times strictly within its specific heteroatom family to preserve the family-specific distribution. The PR was recomputed for each permutation to build a null distribution, from which the Z-score and corresponding p-value were derived.

3 Results

3.1 Quantitative mapping of network rewiring

The global property dependency network—comprising rotational constants (B), electronic response (α, r^2, μ), frontier orbitals (HOMO, LUMO), and vibrational-thermodynamic variables ($zpve$, c_v)—exhibits significant deformation upon the introduction of heteroatoms. Using the pure hydrocarbon (HC) manifold as a baseline, we quantified the mathematical distance between these networks using the Gromov-Wasserstein distance (GW_{dist}) applied to the distance correlation matrices of the atom-count-conditioned residuals.

The introduction of nitrogen (the N-only family) induces the largest network shift ($GW_{dist} = 0.0331$), followed closely by the oxygen-only and mixed heteroatom populations (both exhibiting $GW_{dist} = 0.0316$). To ensure these results were not statistical artifacts of sample size or heavy-atom count distributions, we established a rigorous null baseline by comparing disjoint subsets of the HC family that were stratified to match the exact heavy-atom distributions of the target heteroatom families. These null distances were consistently orders of magnitude lower (e.g., $GW_{dist} = 0.000182$ for the N-only matched baseline). This confirms that the observed shifts represent genuine "excess rewiring" driven by chemical identity rather than stoichiometric scaling.

3.2 Decomposition of disrupted property pairs

Decomposing the GW transport cost allows us to isolate the specific property relationships that are most severely distorted during the transition from the HC manifold to heteroatom spaces. The network deformation is primarily driven by three critical axes, as illustrated in Figure 1.

First, the geometric-spatial coupling between the rotational constant (B) and the electronic spatial extent (r^2) exhibits the highest aggregated local cost (0.00383). In pure hydrocarbons, molecular volume and the moment of inertia scale uniformly through the carbon skeleton. However, the insertion of nitrogen or oxygen introduces local mass and charge concentrations that disrupt this uniform scaling. Second, the frontier orbital relationship between HOMO and LUMO energies accounts for the next highest cost (0.00259). While the HOMO-LUMO gap is a fundamental electronic identity, the individual orbital residuals become significantly decoupled and less predictable relative to one another in heteroatom spaces. Finally, the response-vibrational divergence between r^2 and zero-point vibrational energy ($zpve$) emerges as the third most disrupted axis (local cost 0.00124), indicating a fundamental shift in how electronic volume translates into vibrational modes.

3.3 Symmetry breaking and dipole induction

The magnitude of the observed network rewiring correlates strongly with heteroatom-induced symmetry breaking. We found a strong positive correlation between the family-specific GW_{dist} and both the average dipole moment (μ, Pearson $r = 0.9432$, $p = 0.0568$) and the relative symmetry index (Pearson $r = 0.9531$, $p = 0.0469$).

The HC baseline is characterized by low dipole moments (mean $\mu = 0.355$ D) and lower relative symmetry indices (0.858), reflecting a population dominated by non-polar, flexible acyclic or simple cyclic structures. In contrast, the introduction of heteroatoms significantly increases both the mean dipole moment (N-only: 3.059 D; O-only: 2.156 D) and the relative symmetry index (N-only: 0.945; O-only: 0.938). This correlation suggests that the rewiring measured by GW transport is the mathematical signature of a physical state space transitioning from a dispersed, size-dominated regime to one structured by local electrostatic dipoles and heteroatom-induced electronic constraints.

3.4 Statistical isolation and reintegration of the HOMO

A recurring anomaly in the property dependency network is the behavior of the HOMO energy. While most properties participate in tightly bound clusters—such as the $zpve$–c_v vibrational block or the B–r^2–α response block—the HOMO residuals remain largely isolated. Participation Ratio (PR) analysis highlights this contrast: network hubs like isotropic polarizability (α) and heat capacity (c_v) exhibit high PR values in

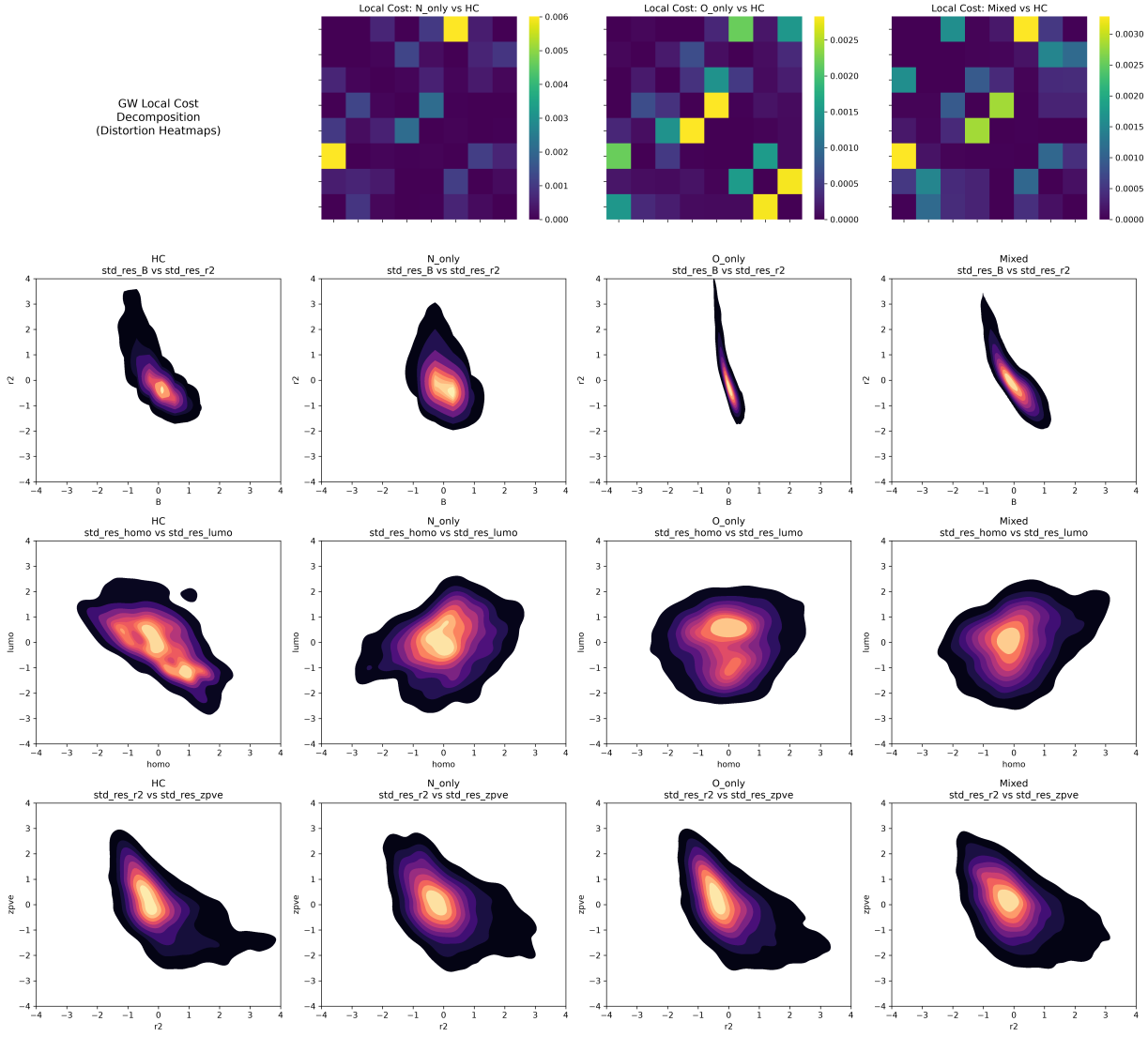

Figure 1: Gromov-Wasserstein (GW) local cost decomposition and joint density distributions of the most severely disrupted property pairs across pure hydrocarbon (HC) and heteroatom-containing (N-only, O-only, and Mixed) manifolds. The top row displays distortion heatmaps quantifying the local cost of transporting the property network from the HC baseline to heteroatom spaces. The subsequent rows illustrate the standardized residuals of the top three rewired axes identified by the GW cost decomposition: the geometric-spatial decoupling of the rotational constant (B) and electronic spatial extent (r^2), the frontier orbital reorganization between HOMO and LUMO, and the response-vibrational divergence between r^2 and zero-point vibrational energy ($zpve$). As evidenced by the shifting density contours, the introduction of nitrogen and oxygen disrupts the uniform scaling of the carbon skeleton, leading to a fundamental rewiring of physical properties driven by local electrostatic dipoles and symmetry breaking rather than stoichiometric scaling.

hydrocarbons (6.97 and 6.57, respectively), whereas the HOMO residual has a PR of only 4.57, the lowest among non-dipole properties.

To determine whether this isolation is a genuine physical phenomenon or a statistical artifact, we performed permutation testing. Shuffling the HOMO residual vector within the HC family yielded a null mean PR of 1.376. The observed HC HOMO PR of 4.408 is significantly higher than this null baseline ($Z = 65.45$), confirming a baseline level of structural coupling. However, upon the introduction of nitrogen, the HOMO PR increases substantially to 5.787 ($Z = 110.30$), indicating a fundamental physical transition.

In pure hydrocarbons, the HOMO energy is largely decoupled from global spatial and thermodynamic properties, likely governed by local C-C bonding environments that do not scale with molecular size. The arrival of nitrogen rewires the HOMO back into the global network, whereas oxygen maintains its relative isolation (PR 4.23). This demonstrates that nitrogen-containing functional groups (e.g., amines, nitriles, heterocycles) impose a more coherent joint structure on frontier orbitals than either oxygen or pure carbon frameworks.

4 Conclusions

The joint organization of molecular properties is frequently assumed to be a static consequence of chemical structure, heavily dominated by stoichiometric scaling. In this study, we challenged this assumption by deploying Gromov-Wasserstein optimal transport to quantify how the introduction of specific heteroatoms fundamentally rewires the multivariate physical state space of small molecules. By analyzing non-linear, atom-count-conditioned residuals from the QM9 dataset, we successfully isolated intrinsic chemical physics from the deterministic effects of molecular size and heavy-atom counts.

Our results demonstrate that the property dependency network is a dynamic manifold that undergoes significant topological deformation across compositional boundaries. The introduction of nitrogen and oxygen induces substantial network shifts compared to a pure hydrocarbon baseline. Through transport cost decomposition, we identified that this rewiring is primarily driven by the disruption of the geometric-spatial coupling between the rotational constant (B) and electronic spatial extent (r^2). Furthermore, heteroatoms cause severe reorganizations in frontier orbital relationships (HOMO and LUMO) and response-vibrational couplings (r^2 and $zpve$), breaking the uniform scaling rules that govern pure carbon skeletons.

We established that these mathematical shifts are deeply rooted in physical mechanisms. The magnitude of the network deformation correlates strongly with heteroatom-induced symmetry breaking and increased dipole moments, marking a transition from a dispersed, size-dominated regime to one structured by local electrostatic constraints. Additionally, participation ratio analyses uncovered a distinct topological role for the highest occupied molecular orbital. While the HOMO energy remains statistically isolated from the global property network in pure hydrocarbons, the arrival of nitrogen partially reintegrates it, imposing a more coherent joint structure on the frontier orbitals than is observed in oxygen-containing or pure hydrocarbon frameworks.

Ultimately, this work provides a rigorous mathematical mapping of how chemical identity reorganizes global property couplings. By demonstrating that heteroatoms stretch and tear the relational manifold of molecular physics independently of stoichiometry, we offer a new quantitative framework for understanding the fundamental rules that govern the joint physical state space of organic molecules.