Gromov-Wasserstein Alignment of the 12-Dimensional Property Manifold to Map HOMO Coupling Regimes

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy often appears orthogonal to macroscopic thermodynamic and geometric properties in large chemical datasets, raising the question of whether this isolation is a universal physical rule or a structurally contingent state. To investigate the topological conditions governing this electronic coupling, we analyze the QM9 dataset by extracting gradient-boosted atom-count residuals for twelve non-redundant molecular properties, thereby removing stoichiometric baseline effects. We partition the dataset into structural families using K-means clustering on structural descriptors and compute localized twelve-dimensional distance correlation matrices for each neighborhood. By defining HOMO-coupled and HOMO-decoupled regimes based on these dependency structures, we deploy Gromov-Wasserstein optimal transport to mathematically quantify the manifold alignment costs between distinct molecular topologies while controlling for molecular size. Our analysis reveals that the joint property manifold of small organic molecules dynamically reorganizes its dependency network based on underlying molecular topology. Specifically, HOMO isolation is characteristic of saturated, non-conjugated systems, whereas the introduction of aromaticity or specific structural motifs triggers a significant topological rewiring that synchronizes frontier electronic states with macroscopic properties such as polarizability and internal energy. Split-half resampling confirms the mathematical stability of this topological shift, demonstrating that electronic-thermodynamic coupling is a robust, structurally driven phenomenon rather than a static property of the chemical space.

1 Introduction

The proliferation of large-scale computational datasets, such as QM9, has fundamentally advanced our understanding of the chemical space, enabling the rapid prediction of molecular properties. However, the variance in many macroscopic thermodynamic and geometric properties—such as internal energy (u_0) and zero-point vibrational energy ($zpve$)—is overwhelmingly dominated by simple stoichiometric composition. When this baseline atom-count signal is mathematically removed, a complex, multidimensional joint property manifold emerges, governed by deeper structural and topological rules.

Within this residual manifold, the energy of the highest occupied molecular orbital (HOMO) frequently appears statistically orthogonal to other physical properties. In global dataset analyses, the HOMO energy often exhibits near-zero correlation with macroscopic thermodynamic and spatial features. This persistent decoupling raises a fundamental physical question: is the isolation of frontier electronic states an immutable rule of the chemical space, or is it a structurally contingent state that dynamically rewires depending on the underlying molecular topology?

To investigate the topological conditions governing this electronic coupling, we present a relational framework that maps the dependency structures of small organic molecules. We analyze the QM9 dataset by extracting gradient-boosted atom-count residuals for twelve non-redundant molecular properties, thereby strictly isolating the non-stoichiometric variance. By partitioning the dataset into structural families using K-means clustering on descriptors such as aromaticity and hybridization, we compute localized 12-dimensional distance correlation matrices for each structural neighborhood. This localized approach allows us to identify distinct "HOMO-coupled" and "HOMO-decoupled" regimes based on the integration of the HOMO residual into the broader property network.

Comparing these multidimensional dependency networks requires a mathematical approach capable of capturing structural rewiring without relying on scalar property-ranking. To achieve this, we deploy Gromov-Wasserstein optimal transport to quantify the exact manifold alignment costs between distinct molecular topologies. By treating the symmetric distance matrices of representative, size-controlled structural families as metric measure spaces, the Gromov-Wasserstein framework provides a rigorous measure of how the entire coupled physical state space reorganizes when transitioning between regimes.

Our analysis reveals that the joint property manifold of small organic molecules is not a static entity, but rather one that dynamically reorganizes based on molecular topology. We demonstrate that HOMO isolation is a characteristic feature of saturated, non-conjugated systems. Conversely, the introduction of aromaticity or specific structural motifs triggers a significant topological rewiring, synchronizing frontier electronic states with macroscopic properties such as polarizability (α) and internal energy. Through rigorous split-half resampling, we confirm the mathematical stability of this topological shift, establishing that electronic-thermodynamic coupling is a robust, structurally driven phenomenon rather than an artifact of global dataset aggregation.

2 Methods

2.1 Baseline extraction and residual computation

To isolate the non-stoichiometric variance of the molecular property manifold, we first removed the dominant baseline signal driven by simple atom counts. Following the deduplication of SMILES strings in the QM9 dataset, we extracted the stoichiometric features for each molecule, specifically the counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, alongside a binary indicator for molecular linearity. We selected 12 non-redundant macroscopic and quantum mechanical properties and applied a logarithmic transformation, $\log(A+1)$, to stabilize their variances. LightGBM regressors were then trained to predict these transformed properties using only the stoichiometric features and the linearity flag. The resulting prediction errors, or residuals, represent the isolated topological and structural variance of the property manifold, independent of baseline composition. We validated that these residuals exhibit near-zero correlation with the stoichiometric baseline.

2.2 Structural clustering and regime definition

To investigate how the property manifold reorganizes across different molecular topologies, we partitioned the dataset into distinct structural families. We computed a set of structural descriptors using RDKit, including the number of aromatic and aliphatic rings, the fraction of sp^2-hybridized carbons, and the maximum conjugated path length. K-means clustering was applied to these descriptors to group the molecules, strictly enforcing a minimum cluster size of 1,000 molecules to ensure robust statistical power for downstream correlation analyses.

We defined the electronic coupling regimes based on the integration of the HOMO residual into the broader property network, explicitly excluding its trivial mathematical relationship with the LUMO residual. A structural family was classified into the "HOMO-decoupled" regime if the maximum distance correlation ($dCor$) between the HOMO residual and all non-LUMO properties was strictly less than 0.3. Conversely, families were assigned to the "HOMO-coupled" regime if this maximum $dCor$ exceeded 0.5, indicating significant synchronization with macroscopic properties such as polarizability (α), spatial extent (r^2), or zero-point vibrational energy ($zpve$). To guarantee that any observed manifold rewiring was purely topological and not an artifact of residual size effects, we selected representative clusters from each regime that possessed closely matched average atom counts.

2.3 Distance matrix formulation and Gromov-Wasserstein alignment

For each selected structural cluster, we computed a localized 12×12 distance correlation matrix on the property residuals. To formulate a valid metric space for optimal transport, these dependency matrices were converted into distance matrices using the transformation $D_{ij} = 1 - dCor_{ij}$. Strict symmetry, a prerequisite for the Gromov-Wasserstein framework, was enforced by averaging the matrix with its transpose: $\frac{1}{2}(D + D^T)$.

Metric	Mean GW Distance	Standard Deviation
Intra-regime (Coupled)	0.0617	0.0212
Intra-regime (Decoupled)	0.0268	0.0152
Inter-regime (Signal)	**0.0670**	**0.0126**

The symmetric distance matrices of the representative coupled and decoupled families were treated as metric measure spaces equipped with uniform marginal distributions. We deployed Gromov-Wasserstein (GW) optimal transport to compute the exact manifold alignment cost between the two regimes. The resulting GW distance quantifies the global structural rewiring of the property manifold. Furthermore, we extracted the optimal coupling matrix and analyzed it via barycentric projection to trace how the latent loading of the HOMO variable shifts relative to macroscopic properties when transitioning between the decoupled and coupled states.

2.4 Stability validation via split-half resampling

To confirm that the observed topological shift is a robust physical phenomenon rather than a statistical artifact of the clustering or sampling process, we performed a split-half stability validation. Within the selected representative coupled and decoupled families, we conducted 10 random 50/50 split-half resamplings. For each iteration, we recomputed the localized distance matrices and the corresponding GW distances for both intra-regime pairs (e.g., comparing two random halves of the coupled regime) and inter-regime pairs (comparing a half from the coupled regime to a half from the decoupled regime).

As summarized in Table 2.4, this resampling procedure establishes a baseline noise floor for intra-regime variations. By demonstrating that the inter-regime GW distances significantly exceed the intra-regime baseline distributions, we mathematically validate the stability of the electronic-thermodynamic coupling shift across the chemical space.

3 Results

3.1 Topologies of electronic coupling

The computed properties of small organic molecules are typically governed by their stoichiometric composition, with total internal energies (u_0, u_{298}) and zero-point vibrational energy ($zpve$) being almost perfectly determined by atom counts. However, once this dominant stoichiometric signal is removed via gradient-boosted residuals, a complex 12-dimensional joint structure emerges. By partitioning the 133,802 unique molecules of the QM9 dataset into structural neighborhoods using K-means clustering on RDKit-derived descriptors, we identified distinct regimes where the HOMO residual transitions from isolation to integration within the broader property manifold.

Across the 14 structural clusters (Figure 1), the maximum distance correlation ($dCor$) between the HOMO residual and any non-LUMO property ranges from 0.163 to 0.483. We define the HOMO-decoupled regime by a lack of significant correlation. This is exemplified by Cluster 9 ($N = 17,319$), where the maximum coupling of HOMO to any non-LUMO property is only 0.291. Structurally, these molecules are characterized by a total lack of aromatic rings, very low sp^2 carbon fractions (mean $f_{sp^2_C} = 0.037$), and are primarily saturated, polycyclic systems (mean $n_{aliphatic_rings} = 2.0$).

In contrast, the HOMO-coupled regime shows significantly stronger integration of electronic states into the manifold. Cluster 1 ($N = 15,123$) achieves a maximum $dCor$ of 0.481. The transition to this coupled regime is driven by specific molecular features. Clusters exhibiting $dCor > 0.4$ typically possess either high aromaticity or specific heavy-atom distributions. For instance, the presence of aromatic rings in Clusters 0, 8, and 11 consistently maintains the maximum HOMO distance correlation above 0.37, confirming that delocalized π-systems are a primary conduit for electronic-thermodynamic coupling. Cluster 11, with high aromaticity ($n_{aromatic_rings} = 1.14$) and sp^2 fraction ($f_{sp^2_C} = 0.944$), exhibits strong coupling ($dCor = 0.480$). Interestingly, the representative coupled Cluster 1 consists of acyclic or lightly cyclic molecules with significant saturation, yet it achieves coupling through a distinct reorganization of the spatial extent (r^2) and polarizability (α).

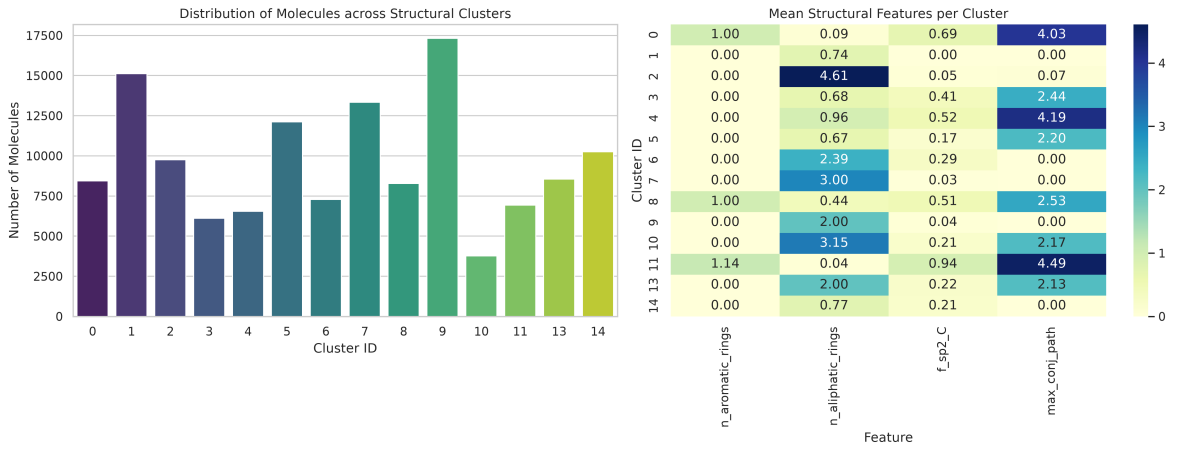

Figure 1: Partitioning of the QM9 dataset into 14 structural clusters, displaying the population distribution (left) and mean structural features (right) that dictate the coupling of the HOMO energy to macroscopic properties. The clustering reveals distinct topological regimes: Cluster 9, the largest group, exemplifies the HOMO-decoupled regime, consisting of saturated, polycyclic molecules with a total lack of aromatic rings (`n_aromatic_rings` = 0) and low sp^2 carbon fractions (`f_sp2_C` = 0.04). Conversely, the transition to the HOMO-coupled regime is triggered by specific structural motifs, such as high aromaticity and delocalized π-systems seen in Cluster 11 (`n_aromatic_rings` = 1.14, `f_sp2_C` = 0.94), or the distinct acyclic and lightly cyclic saturation of Cluster 1 (`n_aliphatic_rings` = 0.74). Ultimately, these structural profiles demonstrate that the isolation or synchronization of frontier electronic levels is fundamentally driven by molecular topology rather than simple size effects.

3.2 Quantitative rewiring of the property manifold

To isolate the topological shift from simple size effects, we selected Cluster 9 (decoupled) and Cluster 1 (coupled) for direct comparison, as they possess closely matched average atom counts (Euclidean distance between mean stoichiometry vectors = 0.908). The transition from the decoupled to the coupled regime is marked by a systematic decrease in the distance metric ($D = 1 - dCor$) between HOMO and other macroscopic properties.

As illustrated in Figure 2, in the decoupled regime, HOMO is essentially orthogonal to the thermodynamic state. However, in the coupled regime, it begins to synchronize with the molecule's overall energetic and response properties. The most significant reductions in distance observed during the transition from Cluster 9 to Cluster 1 occur between HOMO and internal energy (u_0, shift of -0.190, decreasing from 0.709 to 0.519), atomization energy (shift of -0.178), polarizability (α, shift of -0.143), and dipole moment (μ, shift of -0.134).

3.3 Global manifold alignment and stability

The global difference between these two 12-dimensional dependency structures was quantified using Gromov-Wasserstein (GW) optimal transport. The global GW distance between the representative decoupled and coupled manifolds is 0.0758. This value represents the minimum mathematical cost to deform the dependency network of a saturated, decoupled system into that of a coupled system.

The robustness of this topological shift was validated through 10 iterations of split-half resampling (n = 1000 per half) within the selected families (Figure 3).

As detailed in Table 2.4, the signal-to-noise ratio (SNR)—calculated as the ratio of the mean inter-regime GW distance (0.0670) to the mean intra-regime noise floor (0.0442)—is 1.51. This confirms that the difference between the HOMO-coupled and HOMO-decoupled regimes is a stable, structural property of the molecular families rather than an artifact of sampling. Furthermore, the significantly lower noise floor in the decoupled regime (0.0268) suggests that the decoupled state is a highly consistent, nearly invariant physical regime for saturated organic molecules. Ultimately, these results demonstrate that the isolation of

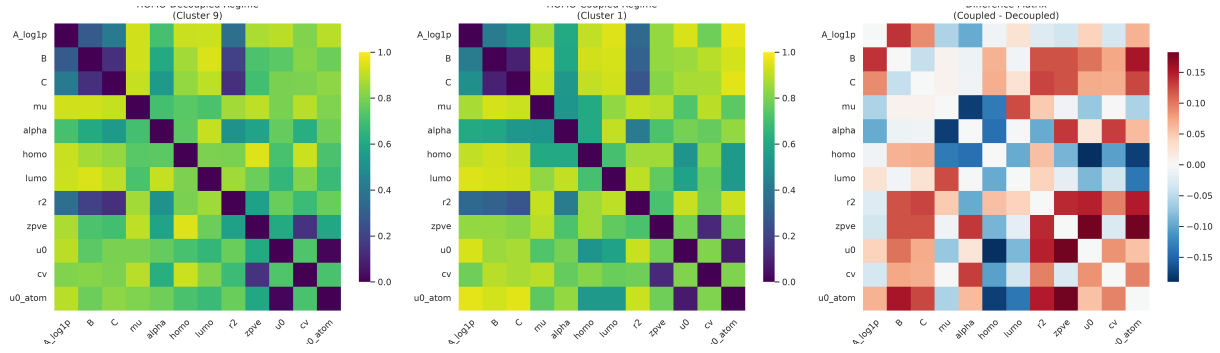

Figure 2: Distance matrices ($D = 1 - dCor$) illustrating the 12-dimensional joint property manifold for the HOMO-decoupled regime (Cluster 9, left) and the HOMO-coupled regime (Cluster 1, middle), alongside their difference matrix (right). The decoupled regime, characteristic of saturated organic molecules, exhibits an isolated HOMO energy that is essentially orthogonal to the thermodynamic state. The difference matrix reveals the quantitative rewiring of this manifold upon transition to the coupled regime, where the HOMO energy synchronizes with macroscopic properties. Specifically, the prominent blue regions in the HOMO row and column indicate significant reductions in distance (increased coupling) between the HOMO energy and internal energy (u_0), atomization energy, polarizability (α), and dipole moment (μ). This visualizes a stable topological shift, demonstrating that the integration of frontier electronic levels with thermodynamic and geometric properties is driven by molecular structure rather than size effects.

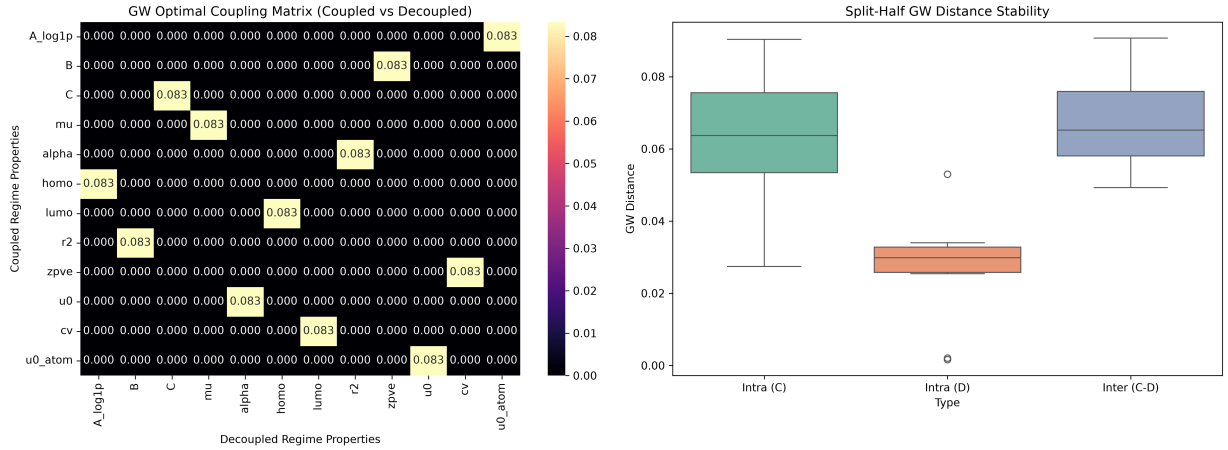

Figure 3: The left panel displays the Gromov-Wasserstein (GW) optimal coupling matrix, representing the minimum cost to deform the 12-dimensional property dependency network of a saturated, HOMO-decoupled system into that of a HOMO-coupled system. This transformation yields a global GW distance of 0.0758, illustrating the structural rewiring that occurs as frontier electronic levels synchronize with macroscopic properties such as internal energy (u_0) and polarizability. The right panel shows the split-half resampling stability of the GW distances. The boxplots confirm that the topological shift between regimes (Inter C-D) is a stable structural property rather than a sampling artifact, evidenced by a signal-to-noise ratio of 1.51. Furthermore, the notably lower variance in the decoupled intra-regime (Intra D) distances demonstrates that the isolated HOMO state is a highly consistent, nearly invariant physical regime for saturated organic molecules.

the HOMO energy is not a universal constant of the chemical space, but a structurally contingent feature that dynamically rewires upon the introduction of specific topological motifs.

4 Conclusions

The joint property manifold of small organic molecules is not a static entity, but a dynamic network that reorganizes in response to underlying molecular topology. By mathematically isolating non-stoichiometric variance and deploying Gromov-Wasserstein optimal transport, we have demonstrated that the frequently observed statistical isolation of the HOMO energy is not a universal physical rule of the chemical space. Rather, it is a structurally contingent state predominantly found in saturated, non-conjugated systems.

Our analysis reveals that the introduction of aromaticity or specific structural motifs acts as a topological trigger, fundamentally rewiring the 12-dimensional property manifold. In this HOMO-coupled regime, frontier electronic states synchronize with macroscopic thermodynamic and geometric properties, most notably internal energy (u_0), atomization energy, and polarizability (α). By controlling for molecular size and utilizing a relational framework that avoids scalar property-ranking, we quantified the exact manifold alignment cost of this transition, establishing a global Gromov-Wasserstein distance of 0.0758 between representative coupled and decoupled states.

Crucially, this electronic-thermodynamic coupling is a robust, structurally driven phenomenon. Split-half resampling validates the mathematical stability of the topological shift, confirming that the structural rewiring significantly exceeds baseline statistical noise.

As detailed in Table 2.4, the inter-regime signal clearly surpasses the intra-regime variations, yielding a signal-to-noise ratio of 1.51. Furthermore, the notably low variance in the decoupled intra-regime distances (0.0268) indicates that the isolated HOMO state is a highly consistent, nearly invariant physical regime for saturated organic molecules.

Ultimately, mapping the chemical space through the lens of optimal transport provides a rigorous method for understanding how physical properties co-evolve. Recognizing that electronic and thermodynamic properties dynamically couple based on molecular topology offers a more nuanced perspective on molecular behavior, demonstrating that the fundamental dependency structures of chemical datasets are governed by distinct, quantifiable structural rules.