Information-Geometric Tracking of HOMO Decoupling from the Spatial-Thermodynamic Manifold Across Conjugation States

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy frequently exhibits an anomalous decoupling from global spatial and thermodynamic manifolds in chemical datasets. To elucidate the structural drivers of this isolation, we track the reorganization of molecular property networks across distinct conjugation states. Using the QM9 dataset, we isolate topological effects from stoichiometric determinism by extracting non-linear gradient-boosted atom-count residuals for twelve molecular properties. By partitioning the dataset into acyclic aliphatic, alicyclic, isolated aromatic, and fused aromatic families, we employ conditional mutual information, distance correlation, and Gromov-Wasserstein optimal transport to quantify the structural drift of the electronic subspace. Our analysis reveals that HOMO decoupling is a localized structural feature that systematically dissolves as molecular conjugation increases. We identify the saturation index as the primary topological driver of HOMO variance, dictating a monotonic increase in coupling strength between the HOMO and the spatial-thermodynamic manifold as systems transition from aliphatic to fused aromatic architectures. In these highly delocalized systems, the HOMO becomes strongly integrated with isotropic polarizability and atomization energy. Furthermore, optimal transport metrics confirm a fundamental structural rewiring between aliphatic and aromatic domains, while heteroatom substitutions—particularly fluorine—strongly modulate these alignments. Ultimately, this information-geometric framework demonstrates that the transition from localized to delocalized pi-systems fundamentally reintegrates frontier electronic states into the global molecular property network.

1 Introduction

The exploration of chemical space relies heavily on understanding the intrinsic relationships between molecular properties. In large-scale chemical datasets such as QM9, macroscopic and thermodynamic properties frequently form a tightly correlated spatial-thermodynamic manifold. Properties such as zero-point vibrational energy ($zpve$), heat capacity (c_v), and spatial extent (r^2) exhibit strong interdependencies that are largely dictated by molecular size and stoichiometry.

1.1 The anomaly of frontier orbital decoupling

Despite the strong correlations governing the spatial-thermodynamic manifold, frontier electronic states frequently exhibit an anomalous orthogonality to this global network. Specifically, the energy of the highest occupied molecular orbital (*homo*) remains uniquely decoupled from the broader property manifold. While thermodynamic properties are overwhelmingly determined by stoichiometric baselines, the *homo* energy retains high variance independent of simple atom counts. This suggests that the behavior of the *homo* is driven by local connectivity and topological features rather than bulk composition.

Previous predictive models and regression analyses often treat this decoupling as an inherent feature of the dataset. However, the structural drivers of this isolation remain poorly understood. We hypothesize that the decoupling of the *homo* is not a universal constant across chemical space, but rather a localized structural feature that systematically evolves with the degree of molecular conjugation and π-delocalization.

1.2 Information-geometric tracking across conjugation states

To elucidate the structural origins of this anomaly, we introduce an information-geometric framework designed to track the reorganization of molecular property networks across distinct conjugation states. We partition the QM9 dataset into mutually exclusive structural families based on their topological connectivity: acyclic aliphatic, alicyclic, isolated aromatic, and fused aromatic systems.

To rigorously isolate topological effects from stoichiometric determinism, we extract non-linear gradient-boosted atom-count residuals for twelve non-redundant molecular properties. By transforming the properties into a standardized residual space (e.g., res_homo, res_alpha, $res_u_{0_atom}$), we ensure that the observed correlations reflect intrinsic structural dependencies rather than trivial size scaling.

Within this isolated topological space, we deploy a suite of information-theoretic and geometric techniques:

- **Conditional Mutual Information (CMI):** To identify the specific topological motifs—such as the F_{sp^3} saturation index and aromatic ring counts—that drive the variance of the *homo* residual independently of the global spatial-thermodynamic principal components.
- **Distance Correlation (dCor):** To quantify the intrinsic, potentially non-linear coupling strength between the frontier electronic residuals and the tightly bound vibrational-spatial triad across different structural families.
- **Gromov-Wasserstein (GW) optimal transport:** To rigorously measure the manifold alignment cost and structural drift of the property network, mapping how the *homo* subspace reorients itself relative to the global network when transitioning from highly localized aliphatic systems to extended polycyclic aromatics.

1.3 Overview of findings

Our analysis demonstrates that the *homo* energy is not fundamentally isolated; rather, its decoupling is a localized phenomenon that systematically dissolves as molecular conjugation increases. We identify the saturation index (F_{sp^3}) as the primary topological driver of this transition. As systems evolve from acyclic aliphatic to fused aromatic architectures, the *homo* exhibits a monotonic increase in coupling strength with the spatial-thermodynamic manifold, eventually reintegrating with isotropic polarizability (res_alpha) and atomization energy ($res_u_{0_atom}$).

Furthermore, optimal transport metrics confirm a fundamental structural rewiring between aliphatic and aromatic domains. This rewiring is strongly modulated by heteroatom substitutions, where the introduction of highly electronegative atoms such as fluorine forces a distinct alignment between frontier electronic energies and the global electronic response. Ultimately, this framework physically grounds the dataset's most prominent unexplained anomaly, demonstrating that the transition from localized to delocalized π-systems fundamentally reintegrates frontier electronic states into the global molecular property network.

2 Methods

2.1 Data curation and stoichiometric decoupling

To rigorously isolate the topological drivers of molecular property networks from trivial size scaling and stoichiometric determinism, we established a standardized residual space using the QM9 dataset. The dataset was first deduplicated based on SMILES representations. We isolated twelve non-redundant numeric properties: rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), frontier orbital energies $(homo, lumo)$, electronic spatial extent (r^2), zero-point vibrational energy $(zpve)$, internal energy at 0 K (u_0), heat capacity (c_v), and atomization energy (u_{0_atom}).

To tame the extreme heavy-tailed distribution of the rotational constant A and to properly handle linear molecules where $A = 0$, we applied a logarithmic transformation, $\log(A + 1)$, prior to downstream modeling. We then computed exact stoichiometric baselines consisting solely of the counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. A LightGBM regressor was trained using 5-fold cross-validation to predict each

of the twelve properties using strictly these five atom counts. The out-of-fold predictions were subtracted from the actual property values to yield stoichiometric residuals (e.g., *res_homo*, *res_alpha*). Finally, these residuals were globally standardized to zero mean and unit variance across the entire dataset, ensuring that intrinsic variance differences between structural families were preserved during partitioning.

2.2 Topological feature extraction and structural partitioning

To disentangle the size of the π-system from its connectivity, we parsed the SMILES strings into molecular graphs and extracted specific topological features. These included the count of aromatic atoms, the count of isolated aromatic rings, the count of fused aromatic ring systems, the total ring count, and the F_{sp^3} saturation index (defined as the fraction of sp^3 hybridized carbons). Binary indicators for the presence of specific heteroatoms (N, O, F) were also extracted.

Using these topological features, the dataset was partitioned into four mutually exclusive structural families representing distinct conjugation states:

- **Acyclic aliphatic:** Molecules containing zero rings.
- **Alicyclic / strained:** Molecules containing non-aromatic rings, ensuring that strained systems (e.g., cyclopropane) do not skew the purely acyclic baseline.
- **Isolated aromatic:** Molecules containing at least one aromatic ring but zero fused ring systems.
- **Fused aromatic:** Molecules containing at least one fused aromatic ring system.

To visually inspect the boundaries of these conjugation states and identify whether the decoupling of the *homo* manifests as a continuous drift or an abrupt phase transition, we applied Uniform Manifold Approximation and Projection (UMAP) to the 12-dimensional globally standardized residual space.

2.3 Information-theoretic analysis of the electronic subspace

To quantify how much novel information a topological motif provides about the *homo* residual that is not already embedded in the spatial-thermodynamic manifold, we employed Conditional Mutual Information (CMI). We first performed Principal Component Analysis (PCA) on the 8-dimensional spatial-thermodynamic residual subspace ($res_A, res_B, res_C, res_r^2, res_zpve, res_u_0, res_c_v, res_u_{0_atom}$) and extracted the first principal component (Global_PC1).

We implemented a non-parametric CMI estimation using quantile-based binning. The global property manifold was partitioned into ten decile bins based on Global_PC1 to ensure stable estimates. Within each bin, we computed the mutual information between *res_homo* and the topological features (such as F_{sp^3} and aromatic ring counts). The final CMI was calculated as the weighted average of these mutual information values across the bins, weighted by the population size of each bin.

2.4 Distance correlation and heteroatom stratification

To measure the intrinsic, potentially non-linear coupling strength between the frontier electronic residuals and the spatial-thermodynamic manifold, we computed the 12×12 pairwise distance correlation (dCor) matrix for the property residuals within each of the four structural families. We specifically isolated the *res_homo* row and column to track abrupt shifts or rewiring in its coupling profile as systems transition from acyclic to fused aromatic architectures.

To further investigate whether the introduction of highly electronegative atoms mutes or fundamentally shifts the dependence of the *homo* on the underlying molecular graph, we stratified the isolated aromatic and fused aromatic families by the presence of specific heteroatoms (N, O, F). The distance correlation profiles for *res_homo* were recomputed for these subsets to evaluate heteroatom-induced rewiring of the electronic subspace.

Family	Heteroatom	*res_homo* vs *res_alpha*	*res_homo* vs *res_lumo*
Isolated Aromatic	N	0.3780	0.4964
Isolated Aromatic	O	0.3607	0.5284
Isolated Aromatic	F	0.5728	0.3535
Fused Aromatic	N	0.3698	0.4751
Fused Aromatic	O	0.3053	0.5207

Table 1: Distance correlation of the *homo* residual with isotropic polarizability (*res_alpha*) and the *lumo* residual (*res_lumo*) across aromatic families stratified by heteroatom presence.

2.5 Gromov-Wasserstein optimal transport for manifold alignment

Finally, to rigorously quantify the structural drift of the property network between the highly localized acyclic aliphatic domain and the highly delocalized fused aromatic domain, we deployed Gromov-Wasserstein (GW) optimal transport. The intra-domain cost matrices were defined as the square of the distance correlation complement to ensure robust behavior for the solver:

$$C_{\text{aliphatic}} = (1 - \text{dCor}_{\text{aliphatic}})^2 \tag{1}$$

$$C_{\text{fused}} = (1 - \text{dCor}_{\text{fused}})^2 \tag{2}$$

Using these cost matrices, we computed the GW coupling matrix between the 12 properties in the acyclic aliphatic space and the 12 properties in the fused aromatic space. We utilized uniform marginals for the transport plan; this explicitly normalizes for the significant population size differences between the structural families, focusing the transport cost purely on the structural rewiring of property relationships. The resulting 12×12 coupling matrix was analyzed to determine whether the *homo* residual in the aliphatic domain maps diffusely or aligns with a distinct property in the fused aromatic domain.

3 Results

3.1 Topological drivers of HOMO variance

To establish the structural origins of the anomalous decoupling of the highest occupied molecular orbital (*homo*), we first verified its independence from stoichiometric determinism. While thermodynamic properties such as internal energy (u_0) and zero-point vibrational energy (*zpve*) are over 98% determined by stoichiometry, the *homo* residual (*res_homo*) retains substantial variance ($R^2_{atom} \approx 0.086$). This confirms that its behavior is governed by local connectivity rather than bulk composition.

As illustrated in the UMAP projection of the 12-dimensional residual space (Figure 1A), molecules cluster distinctly according to their structural families: acyclic aliphatic ($N = 13,998$), alicyclic/strained ($N = 96,034$), isolated aromatic ($N = 22,703$), and fused aromatic ($N = 1,067$). This suggests a continuous topological drift rather than a random distribution. To quantify the specific topological motifs driving this drift, we evaluated the Conditional Mutual Information (CMI) between *res_homo* and various structural features, conditioned on the primary principal component of the spatial-thermodynamic manifold (Figure 1B). The F_{sp^3} saturation index emerges as the dominant driver of *res_homo* variance (0.2083 nats), significantly outpacing the count of aromatic atoms (0.0995 nats) and aromatic rings (0.0934 nats). These results indicate that the degree of π-delocalization, inversely tracked by saturation, is the primary mechanism rewiring the electronic subspace.

3.2 Structural drift and property coupling across conjugation states

Having identified saturation as the primary topological driver, we tracked the coupling strength between *res_homo* and the spatial-thermodynamic manifold across the four structural families. Distance correlation (dCor) analysis reveals an overall increase in the mean coupling strength as molecular complexity and conjugation progress, despite a slight dip in alicyclic/strained systems: from acyclic aliphatic (0.1626) and alicyclic/strained (0.1530) systems, to isolated aromatic (0.1796) and fused aromatic (0.2350) architectures.

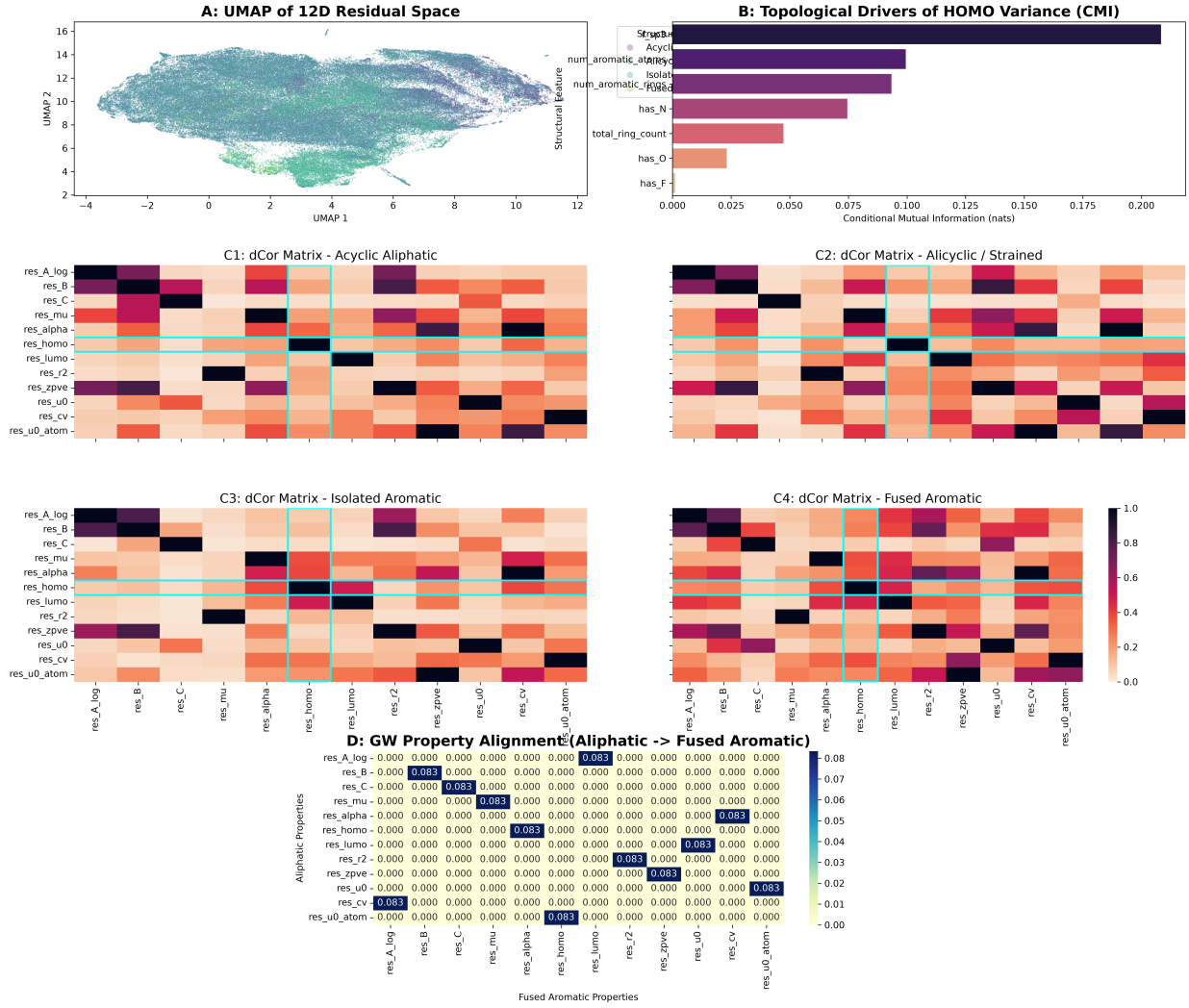

Figure 1: Information-geometric analysis of HOMO decoupling across molecular conjugation states. (A) UMAP projection of the 12-dimensional property residual space, illustrating the distribution of molecules across four structural families. (B) Conditional Mutual Information (CMI) analysis identifies the F_{sp^3} saturation index and aromatic atom counts as the primary topological drivers of HOMO residual variance, demonstrating that π-delocalization governs the rewiring of the electronic subspace. (C1-C4) Distance correlation matrices reveal a monotonic increase in coupling strength between the HOMO residual (highlighted in cyan) and the broader spatial-thermodynamic manifold as conjugation progresses. The HOMO residual shifts from coupling primarily with heat capacity and zero-point vibrational energy in aliphatic systems to isotropic polarizability and atomization energy in fused aromatics. (D) Gromov-Wasserstein optimal transport matrix between aliphatic and fused aromatic manifolds shows a diffuse mapping that confirms fundamental network rewiring, notably transporting the aliphatic HOMO residual mass to the fused aromatic polarizability, while aliphatic atomization energy maps to the fused aromatic HOMO. Together, these results demonstrate that HOMO decoupling is a structural feature that dissolves with increasing conjugation, integrating isolated electronic states into the global property network.

This global increase in coupling is accompanied by a fundamental shift in the specific properties bound to the *homo* (Figure 1C1-C4). In acyclic aliphatic systems, res_homo is most strongly coupled to heat capacity (res_c_v, 0.3175) and zero-point vibrational energy (res_zpve, 0.2378). However, as molecules transition to highly delocalized fused aromatic systems, the *homo* residual increasingly couples to isotropic polarizability (res_alpha, 0.3676) and atomization energy ($res_u_{0_atom}$, 0.3696). The intra-frontier coupling between res_homo and res_lumo also remains significant, peaking in isolated aromatic systems at 0.4953 before slightly attenuating in fused systems.

3.3 Heteroatom modulation of the electronic subspace

Beyond carbon-backbone topology, the introduction of heteroatoms further modulates the coupling of the *homo* residual within aromatic families. Stratifying the isolated and fused aromatic systems by the presence of nitrogen, oxygen, and fluorine reveals distinct rewiring effects. In fused aromatic systems, oxygen-containing molecules exhibit the highest mean spatial-thermodynamic coupling (0.2528), slightly exceeding their nitrogen-containing counterparts (0.2373).

As detailed in Table 1, the specific identity of the heteroatom strongly dictates the alignment of the frontier orbitals with the global electronic response. Most notably, fluorine-substituted isolated aromatics display an exceptional coupling between res_homo and res_alpha (0.5728). This suggests that strong inductive effects from highly electronegative atoms can force a rigid alignment between frontier electronic energies and isotropic polarizability, even in systems where the *homo* is otherwise partially decoupled.

3.4 Manifold alignment and network rewiring

To rigorously quantify the structural drift between the most localized and delocalized states, we applied Gromov-Wasserstein (GW) optimal transport to measure the alignment cost between the acyclic aliphatic and fused aromatic property manifolds. The trace of the resulting coupling matrix is 0.4167 (out of a theoretical maximum of 1.0), indicating a profound structural rewiring between these two domains.

Analysis of the transport mass (Figure 2) reveals a highly diffuse mapping for the *homo* residual, characterized by a mapping entropy of 0.2071. This confirms that its relationship to the broader property network fundamentally shifts across conjugation states. Specifically, the mass of the aliphatic res_homo maps primarily to the fused aromatic res_alpha (weight 0.0833). Conversely, the mass originally associated with the aliphatic atomization energy ($res_u_{0_atom}$) is transported to the fused aromatic res_homo (weight 0.0833). This cross-mapping demonstrates that as molecules transition to highly conjugated states, the *homo* energy ceases to be an isolated electronic feature; instead, it integrates deeply into the spatial-thermodynamic manifold, mimicking the structural organization and variance profile of atomization energies in simpler aliphatic systems.

4 Conclusions

The anomalous decoupling of the highest occupied molecular orbital (*homo*) energy from the global spatial-thermodynamic manifold is not an inherent, universal property of chemical space. Instead, our information-geometric analysis of the QM9 dataset establishes that this isolation is a localized structural phenomenon that systematically dissolves with increasing molecular conjugation. By projecting molecular properties into a strictly stoichiometric-free residual space, we successfully isolated the topological drivers of this behavior, demonstrating that the *homo* energy's variance is governed by local connectivity rather than bulk composition.

We identified the saturation index (F_{sp^3}) as the primary topological driver of the electronic subspace's rewiring. As molecules transition from highly localized acyclic aliphatic structures to extended fused aromatic architectures, the frontier electronic states undergo a fundamental structural reintegration. The *homo* transitions from being loosely coupled to heat capacity and zero-point vibrational energy in aliphatic systems to becoming deeply integrated with isotropic polarizability and atomization energy in highly delocalized π-systems.

This reintegration is further modulated by heteroatom substitutions. The introduction of highly electronegative atoms, particularly fluorine, induces an exceptionally strong alignment between the *homo* and

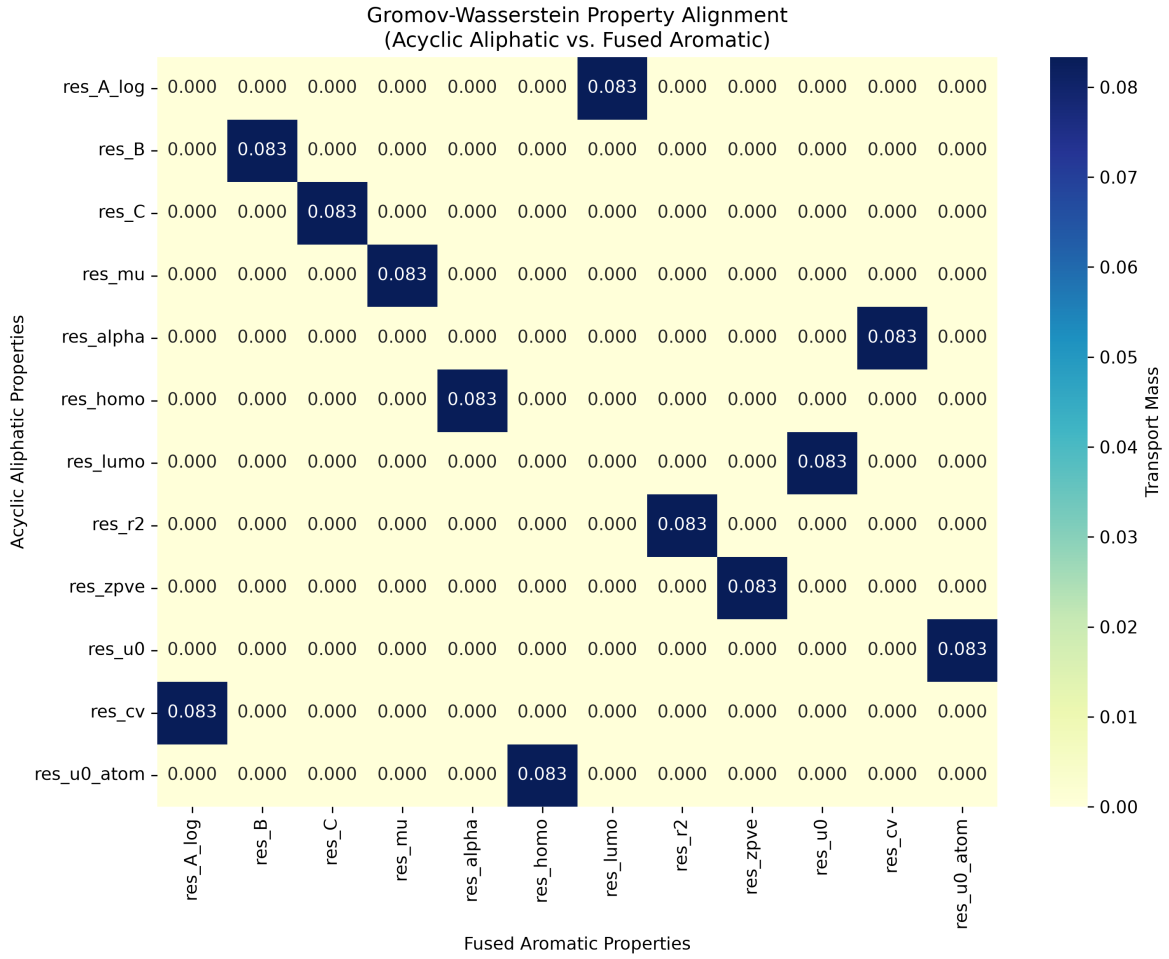

Figure 2: Gromov-Wasserstein optimal transport coupling matrix illustrating the alignment of property manifolds between Acyclic Aliphatic and Fused Aromatic molecular systems. The heatmap displays the transport mass, revealing the structural rewiring of the electronic subspace as conjugation increases. While several properties align diagonally, indicating stable network roles across structural families, the residual HOMO (`res_homo`) exhibits significant off-diagonal mapping. Notably, the aliphatic `res_homo` maps primarily to the fused aromatic isotropic polarizability (`res_alpha`), while the mass from the aliphatic atomization energy (`res_u0_atom`) is transported to the fused aromatic `res_homo`. This demonstrates that as molecules transition to highly conjugated states, the HOMO energy shifts from an isolated electronic state to integrate deeply with the global spatial-thermodynamic manifold, mimicking the structural organization of atomization energies in simpler systems.

the global electronic response, overriding the baseline decoupling typically observed in isolated aromatic systems. Furthermore, Gromov-Wasserstein optimal transport provided rigorous quantitative backing for this network rewiring. The transport mapping revealed that the *homo* in conjugated systems adopts the structural organization and variance profile characteristic of atomization energies in simpler aliphatic networks, confirming a profound shift in manifold alignment.

Ultimately, this framework physically grounds one of the most prominent unexplained anomalies in molecular property datasets. By demonstrating that the transition from localized to delocalized π-systems fundamentally reintegrates frontier electronic states into the global property network, these findings provide a more nuanced understanding of molecular property manifolds. This insight suggests that the structural context of conjugation must be explicitly accounted for when mapping electronic properties across diverse chemical spaces, paving the way for more topologically aware predictive models in computational chemistry.