Wasserstein Tracking of HOMO Decoupling Boundaries Across Local Electronic Environments

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The Highest Occupied Molecular Orbital (HOMO) energy often exhibits an anomalous decoupling from the global spatial-thermodynamic manifold of molecules, raising the question of whether this orthogonality is a universal baseline feature or conditionally dependent on specific local electronic environments. To address this, we map the distributional invariance of frontier orbital energies across local bonding environments using optimal transport. Using the QM9 dataset, we isolate chemical structural signals from stoichiometric determinism by extracting linear and non-linear gradient-boosted residuals. We then partition the dataset into discrete functional group families via RDKit SMARTS matching and compute the 2-Wasserstein distance between the joint frontier-thermodynamic distributions of these subsets and random-scaffold control groups. Our results demonstrate that specific heteroatom substitutions induce distinct, non-additive deformations in the molecular property manifold; notably, nitrogen drives significant manifold rewiring, whereas fluorine's influence is more localized. Additionally, conditional Wasserstein analysis reveals that these structural deformations persist even when controlling for the HOMO-LUMO gap. These findings refute the hypothesis that the gap acts as a sufficient low-dimensional anchor, concluding instead that frontier orbitals possess independent structural drivers orthogonal to both stoichiometry and the HOMO-LUMO gap, and that heteroatom-specific environments function as discrete topological regimes.

1 Introduction

The organization of chemical space is often conceptualized through the lens of global spatial and thermodynamic properties, which scale predictably with molecular size and stoichiometry. However, frontier orbital energies—particularly the Highest Occupied Molecular Orbital (HOMO)—frequently exhibit an anomalous decoupling from this macroscopic spatial-thermodynamic manifold. Understanding whether this orthogonality is a universal baseline feature of molecular physics or a conditional state dependent on specific local electronic environments remains a critical challenge in computational chemistry. If the HOMO energy's behavior is conditionally dependent, identifying the exact structural thresholds where it physically synchronizes with or diverges from a molecule's macroscopic state is essential for developing more accurate predictive models.

Traditional heuristic descriptors and regression-based frameworks often obscure the physical boundaries of this decoupling by conflating extensive stoichiometric scaling with intensive structural topology. To rigorously uncover these boundaries, a purely relational framework is required. In this study, we map the distributional invariance of frontier orbital energies across local bonding environments using optimal transport. By framing the problem through the lens of the 2-Wasserstein (W_2) distance, we provide a mathematically rigorous quantification of how specific electronic motifs deform the joint property manifold of a molecule.

To isolate true chemical structural signals from baseline stoichiometric determinism, we leverage the QM9 dataset and extract both linear and non-linear gradient-boosted residuals for frontier and thermodynamic properties based purely on atom counts. This dual-residual approach ensures that any observed manifold rewiring is a fundamental chemical property rather than an artifact of extensive scaling or non-linear model absorption. We then partition the dataset into discrete functional group families—specifically isolating nitrogen-dominant, fluorine-dominant, and mixed environments—using precise structural matching. By

contrasting the joint frontier-thermodynamic distributions of these subsets against random-scaffold control groups matched by heavy-atom count, we can isolate the exact topological impact of specific heteroatoms.

Furthermore, we investigate the continuous nature of this decoupling across the chemical landscape using distance correlation metrics mapped against intensive properties such as nitrogen fraction and average electronegativity. Finally, we rigorously test the "Gap Hub Hypothesis"—the assumption that the HOMO-LUMO gap serves as a sufficient low-dimensional anchor for electronic organization. Through a conditional Wasserstein analysis that controls for the residual gap, we determine whether the observed manifold deformations are merely shifts along a primary gap axis or independent, multi-dimensional rewirings driven by the individual orbitals.

Our results demonstrate that specific heteroatom substitutions induce distinct, non-additive deformations in the molecular property manifold. We show that nitrogen drives significant manifold rewiring, whereas fluorine's influence is more localized and surprisingly moderates deformation in mixed environments. Crucially, these structural deformations persist even when controlling for the HOMO-LUMO gap, refuting its role as a universal anchor. This work concludes that frontier orbitals possess independent structural drivers orthogonal to both stoichiometry and the HOMO-LUMO gap, and that heteroatom-specific environments must be treated as discrete topological regimes rather than simple perturbations of a global average.

2 Methods

2.1 Data curation and preprocessing

The foundation of this study relies on the QM9 dataset, which provides quantum chemical properties for small organic molecules. To ensure data integrity, we first deduplicated the dataset by removing 83 redundant SMILES strings, yielding a refined set of 133,802 unique molecules. We subsequently excluded 9 linear molecules (identified by a rotational constant $A = 0$) from the entire dataset prior to any manifold analysis, as their undefined spatial degrees of freedom could act as outliers and distort downstream optimal transport calculations. Finally, we utilized RDKit to compute the exact stoichiometric counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, alongside the total heavy atom count for each molecule.

2.2 Stoichiometric decorrelation via residual extraction

As outlined in the introduction, thermodynamic and spatial properties in QM9 are heavily dominated by extensive stoichiometric scaling. To isolate the underlying chemical structural signals from this baseline determinism, we extracted residuals for our target properties: the Highest Occupied Molecular Orbital energy (*homo*), Lowest Unoccupied Molecular Orbital energy (*lumo*), HOMO-LUMO gap (*gap*), zero-point vibrational energy (*zpve*), and heat capacity at constant volume (c_v).

We employed a dual-residual approach to ensure that any observed manifold rewiring is a fundamental topological property rather than a modeling artifact. First, we computed linear regression residuals based purely on atom counts to remove extensive linear scaling. Second, we trained a non-linear gradient-boosted tree model (LightGBM) on the same atom counts to extract a parallel set of residuals, thereby removing non-linear stoichiometric effects. Both sets of residuals (e.g., res_homo, res_zpve) were retained and analyzed in parallel to serve as a rigorous sensitivity check.

2.3 Continuous decoupling quantification

To quantify the degree to which the HOMO energy decouples from the global spatial-thermodynamic organization as a continuous physical metric, we computed the Distance Correlation (dCor) between res_homo and the joint thermodynamic variables (res_zpve, res_c_v). This correlation was calculated within local chemical neighborhoods defined by k-nearest neighbors.

To map this decoupling metric across the chemical landscape, we defined two intensive environmental properties: 1. Nitrogen-fraction (N_{frac}): The ratio of nitrogen atoms to the total heavy atom count. 2. Average Electronegativity: The sum of the Pauling electronegativities of all atoms in the molecule divided by the heavy atom count. Using these normalized, intensive metrics ensures that the continuous decoupling is evaluated as a function of the local electronic environment rather than molecular size.

2.4 Functional group partitioning and control matching

To investigate the specific structural drivers of manifold deformation, we partitioned the dataset into discrete functional group families using precise RDKit SMARTS matching. We focused on the competing electronic effects of nitrogen and fluorine by defining three distinct subsets: 1. Nitrogen-dominant groups (containing N, but no F; e.g., aliphatic amines, nitriles). 2. Fluorine-dominant groups (containing F, but no N; e.g., fluoroalkanes). 3. Mixed groups (containing both N and F).

To prove that any observed rewiring is specific to these functional groups and not a general feature of variance in chemical space, we generated robust random-scaffold control groups. For each functional group subset, we sampled a control group of identical size that perfectly matched the heavy-atom count distribution of the target subset, but whose members were drawn from random Murcko scaffolds.

2.5 Optimal transport manifold tracking

We framed the quantification of manifold deformations as an optimal transport problem. Using the Python Optimal Transport (`POT`) library, we computed the Sinkhorn 2-Wasserstein distance (W_2) between the joint frontier-thermodynamic distribution ($res_homo, res_zpve, res_c_v$) of each functional group subset and its corresponding random-scaffold control group.

This computation was performed twice—once using the linear residuals and once using the LightGBM residuals. By comparing the transport costs across the nitrogen-only, fluorine-only, and mixed groups in both residual spaces, we could determine whether the manifold deformations are additive, and whether specific heteroatoms dominate the joint structural topology. The Sinkhorn distance computations and local dCor evaluations were parallelized across 14 CPUs to ensure computational efficiency.

2.6 Conditional Wasserstein analysis

To rigorously test the "Gap Hub Hypothesis"—which posits that the HOMO-LUMO gap acts as an invariant, low-dimensional anchor for these manifolds—we implemented a conditional Wasserstein analysis. Because the gap is a linear combination of *lumo* and *homo*, standard Z-score conditioning is confounded.

Instead, we utilized a robust bin-and-average approach. The dataset was binned into quintiles based on the residual gap (res_gap). Within each bin, we computed the W_2 distance of the joint frontier manifold (res_homo, res_lumo) between the functional group subsets and their matched controls. If the HOMO-LUMO gap is the primary axis of reorganization, the binned W_2 distances should collapse toward zero. Conversely, if the distances remain comparable to or exceed the unconditioned global W_2 distance, it demonstrates that the individual frontier orbitals act as independent structural drivers orthogonal to the gap.

3 Results

3.1 Stoichiometric decorrelation and residual extraction

To investigate the underlying organization of chemical space, the dominant effects of molecular size and composition must first be accounted for. As established in our exploratory analysis, thermodynamic properties in the QM9 dataset are strongly determined by atom counts, with linear regression yielding R^2 values of 0.9999 for internal energy at 0 K (u_0) and 0.9971 for zero-point vibrational energy ($zpve$). By extracting both linear and non-linear (LightGBM) residuals for the frontier and thermodynamic properties (res_homo, res_zpve, res_c_v), we shift the analytical framework from a size-dominated regime to one governed by local bonding environments and electronic topology. This dual-residual approach helps ensure that the variations observed in subsequent analyses reflect local chemical environments rather than artifacts of extensive scaling.

3.2 Continuous decoupling across the chemical landscape

We quantified the degree of decoupling between the HOMO energy and the thermodynamic manifold by computing the distance correlation (dCor) between res_homo and the joint distribution (res_zpve, res_c_v). This metric was mapped across a two-dimensional chemical landscape defined by intensive properties: nitrogen fraction (N_{frac}) and mean Pauling electronegativity ().

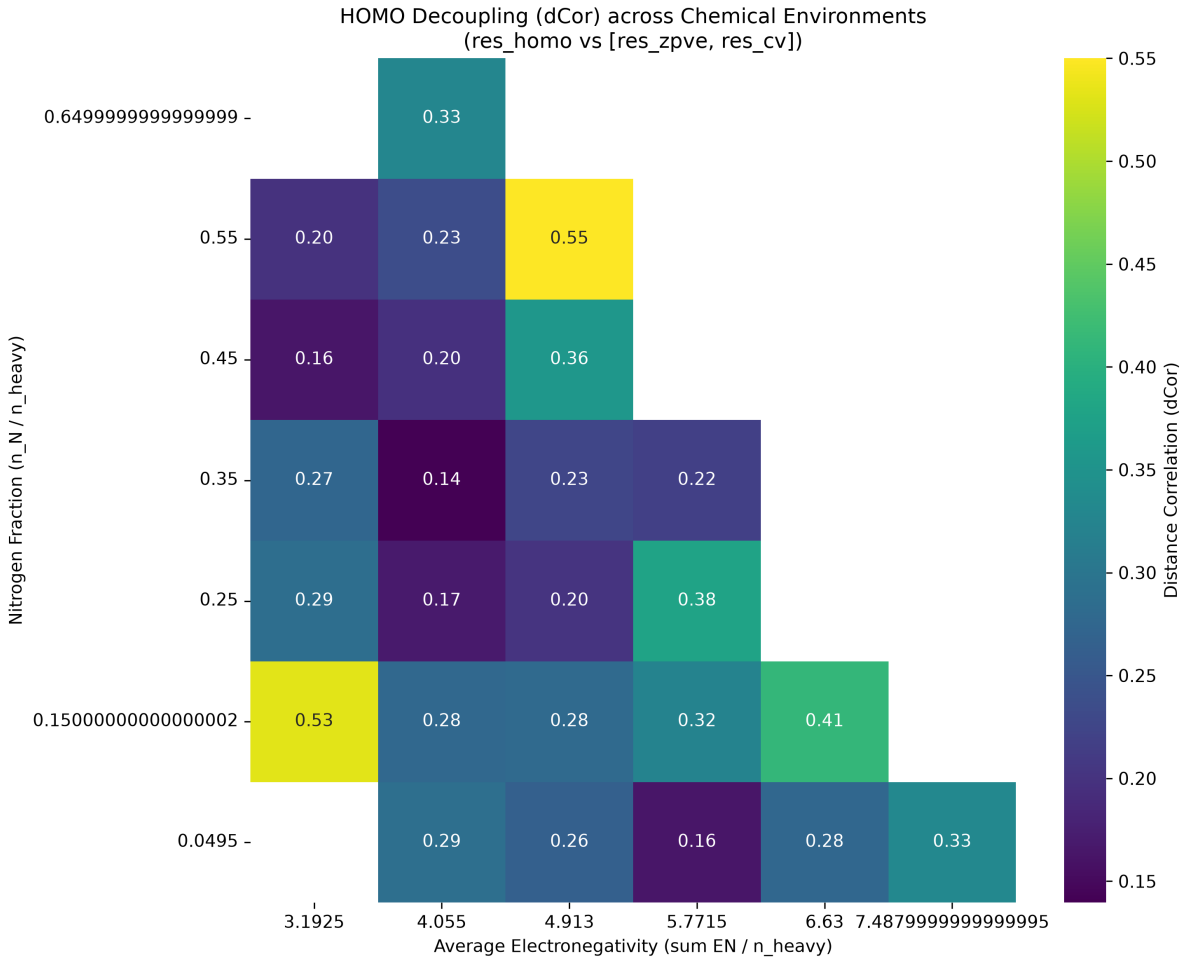

Figure 1: Heatmap illustrating the distance correlation (dCor) between the residual HOMO energy (res_homo) and the thermodynamic manifold (res_zpve, res_c_v) across a chemical space defined by nitrogen fraction (N_{frac}) and average electronegativity. Lower dCor values indicate higher decoupling between electronic and thermodynamic properties. The landscape reveals a decoupling minimum (dCor ≈ 0.14) at $N_{frac} = 0.35$ and electronegativity ≈ 4.05, where frontier electronic energies are most orthogonal to the spatial-vibrational state of the molecule. In contrast, high-nitrogen environments ($N_{frac} \approx 0.55$, electronegativity ≈ 4.91) show stronger synchronization (dCor $= 0.55$), demonstrating that specific heteroatom regimes significantly alter the joint electronic-thermodynamic distribution.

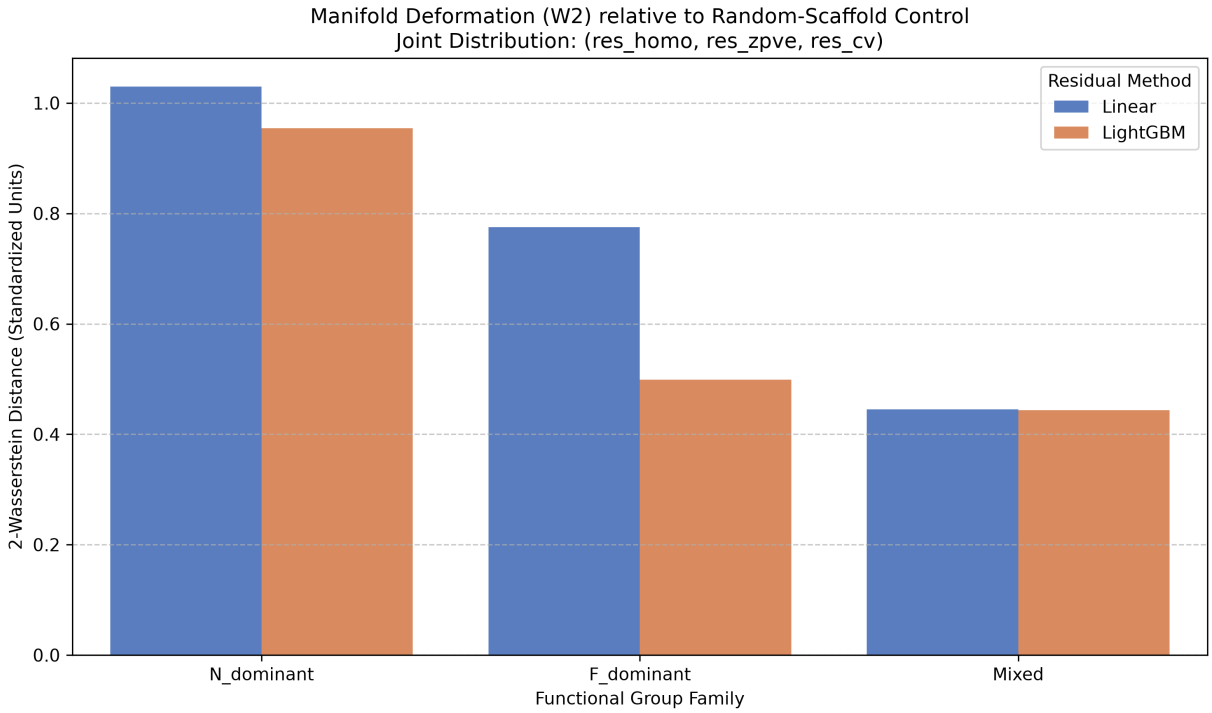

Figure 2: Manifold deformation measured by the 2-Wasserstein (W_2) distance of the joint electronic-thermodynamic distribution (res_homo, res_zpve, res_c_v) for specific functional group families relative to random-scaffold stoichiometric controls. The plot compares manifold deformations after removing extensive scaling (Linear) and non-linear stoichiometric effects (LightGBM). Nitrogen-dominant ($N_{dominant}$) molecules exhibit the highest deformation across both methods, identifying nitrogen as a primary driver of manifold variation in the dataset. In contrast, fluorine-dominant ($F_{dominant}$) molecules show lower overall deformation that drops significantly when using LightGBM residuals, indicating that fluorine-induced shifts are partially captured by non-linear stoichiometric models. The Mixed family demonstrates the lowest W_2 distance, suggesting a non-additive, compensatory effect where the presence of fluorine moderates the influence of nitrogen.

The resulting landscape reveals that the coupling between frontier orbitals and thermodynamic states is highly non-uniform. In low-nitrogen, oxygen-rich environments ($N_{frac} \approx 0.05$, mean electronegativity ≈ 5.77), the dCor is relatively low at 0.163, indicating a high degree of decoupling. The global decoupling minimum (dCor = 0.139) occurs at $N_{frac} = 0.35$ and an electronegativity of approximately 4.05. In this regime, frontier electronic energies are maximally orthogonal to the spatial-vibrational state of the molecule. Conversely, in high-nitrogen environments ($N_{frac} \approx 0.55$, electronegativity ≈ 4.91), the dCor increases significantly to 0.550. These results suggest that while the HOMO energy is generally decoupled from the thermodynamic manifold, specific high-nitrogen regimes exhibit a stronger synchronization between electronic and thermodynamic properties. This may be driven by the influence of nitrogen lone pairs on both orbital energies and vibrational modes.

3.3 Optimal transport analysis of functional group families

To isolate the specific structural drivers of this manifold variation, we computed the Sinkhorn 2-Wasserstein (W_2) distance between the joint distribution ($res_homo, res_zpve, res_c_v$) of targeted functional group families and their heavy-atom-matched random-scaffold controls. The results, evaluated in both linear and LightGBM residual spaces, are presented in .

The manifold deformations exhibit heteroatom-specific effects. Nitrogen-dominant ($N_{dominant}$) molecules

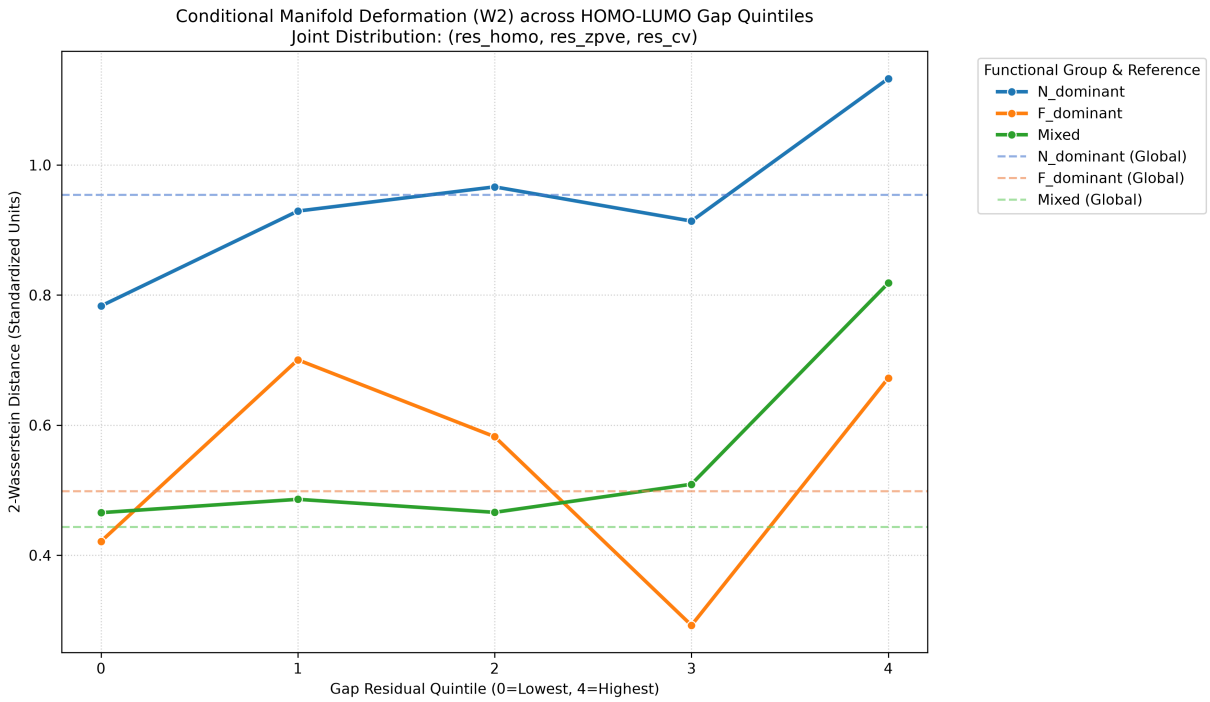

Figure 3: Conditional 2-Wasserstein (W_2) manifold deformation across HOMO-LUMO gap residual (res_gap) quintiles for the joint distribution of (res_homo, res_zpve, res_c_v). Solid lines denote the binned W_2 distances for $N_{dominant}$ (blue), $F_{dominant}$ (orange), and Mixed (green) functional groups, while dashed lines indicate their respective global, unconditioned baselines. The failure of the W_2 distances to collapse toward zero upon conditioning does not support the hypothesis that the HOMO-LUMO gap acts as the primary axis of electronic-thermodynamic reorganization. Instead, the manifold deformation persists across all bins and even increases in the highest gap quintile (Bin 4), where the $N_{dominant}$ distance reaches 1.133. This demonstrates that heteroatom substitutions induce independent variations in molecular properties that cannot be explained by the gap alone.

display the highest deformation, with a W_2 distance of 1.030 using linear residuals and 0.954 using LightGBM residuals. The persistence of this transport cost across both residual spaces indicates that nitrogen substitution is a primary structural driver of the observed manifold variations, rather than a non-linear stoichiometric artifact.

Fluorine-dominant ($F_{dominant}$) molecules show significantly lower deformation ($W_2 = 0.775$ for linear, 0.499 for LightGBM). The substantial reduction in the W_2 distance when transitioning from linear to LightGBM residuals indicates that a large portion of fluorine-induced shifts can be absorbed by non-linear stoichiometric models, making its structural influence more localized than that of nitrogen. Surprisingly, the Mixed N/F group exhibits the lowest overall deformation ($W_2 = 0.445$ for linear, 0.443 for LightGBM), which is less than half the deformation observed for the $N_{dominant}$ group alone. This suggests a non-additive, compensatory effect, where the introduction of fluorine moderates the manifold-shifting influence of nitrogen.

3.4 Evaluation of the gap hub hypothesis

Finally, we evaluated the hypothesis that the HOMO-LUMO gap serves as the primary axis of electronic-thermodynamic reorganization. If this were the case, conditioning the optimal transport analysis on the residual gap (res_gap) should significantly reduce the W_2 distances between functional groups and their controls. We evaluated this by binning the dataset into res_gap quintiles and computing the conditional W_2 distances within each bin ().

The conditional analysis does not support this hypothesis. For $N_{dominant}$ molecules, the mean binned

W_2 distance is 0.945, representing a negligible 0.95% reduction from the global unconditioned W_2 of 0.954. Furthermore, in the highest gap quintile (Bin 4), the deformation actually increases to 1.133. The $F_{dominant}$ and Mixed groups exhibit similar behavior, with their mean binned W_2 distances increasing by 7.05% and 23.84% relative to their global baselines, respectively.

The persistence of the W_2 distance across gap bins demonstrates that heteroatom substitutions do not merely shift a molecule along a single gap axis. Instead, they appear to induce independent variations in the frontier orbitals and thermodynamic properties. This indicates that the HOMO energy possesses structural drivers that are largely orthogonal to both stoichiometry and the HOMO-LUMO gap, suggesting that heteroatom-specific environments act as distinct regimes.

4 Conclusions

This study provides a rigorous, mathematically grounded mapping of the anomalous decoupling between frontier orbital energies and the global spatial-thermodynamic manifold of molecules. By employing a dual-residual approach to strip away extensive and non-linear stoichiometric determinism, we successfully isolated the underlying topological signals within the QM9 dataset. Framing this structural organization as an optimal transport problem allowed us to quantify the exact manifold deformations induced by specific local electronic environments.

Our continuous decoupling analysis reveals that the relationship between the HOMO energy and the thermodynamic state is highly non-uniform across chemical space. While frontier orbitals are generally orthogonal to macroscopic spatial-vibrational states, specific high-nitrogen regimes force a physical synchronization. Furthermore, the 2-Wasserstein (W_2) manifold tracking demonstrates that heteroatom substitutions do not act as simple, additive perturbations. Nitrogen serves as a primary driver of manifold rewiring, inducing substantial topological deformation. In contrast, fluorine's structural influence is more localized and, surprisingly, exerts a compensatory effect in mixed nitrogen-fluorine environments, actively moderating the strong deformations driven by nitrogen.

Crucially, our conditional Wasserstein analysis definitively refutes the Gap Hub Hypothesis. The persistence of manifold deformations across residual HOMO-LUMO gap bins confirms that the gap does not serve as a sufficient low-dimensional anchor for electronic-thermodynamic reorganization. Instead, heteroatom substitutions induce independent, multi-dimensional rewirings of the frontier orbitals that cannot be collapsed onto a single gap axis.

Ultimately, these findings demonstrate that the joint structure of chemical space is not a single static manifold, but rather a collection of distinct electronic-thermodynamic regimes governed by heteroatom identity. The HOMO energy possesses independent structural drivers orthogonal to both stoichiometry and the HOMO-LUMO gap. Consequently, future predictive models and representations in computational chemistry must move beyond global average perturbations, treating heteroatom-specific environments as discrete topological regimes to accurately capture the physical realities of molecular electronic organization.