Optimal Transport Tracking of HOMO Decoupling Across Heteroatom Perturbations

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy in organic molecules typically exhibits a unique decoupling from macroscopic spatial and thermodynamic properties, yet the conditions under which this independence persists or collapses remain poorly understood. To investigate this phenomenon, we analyze a rigorously matched subset of the QM9 dataset to map how the HOMO distributional shape deforms under localized heteroatom perturbations. By extracting property residuals via non-linear gradient boosting, we isolate pure structural drivers from stoichiometric determinism. We then deploy optimal transport metrics—specifically exact Wasserstein distances—to quantify global distributional shifts across hydrocarbon, nitrogen-, oxygen-, and fluorine-doped molecular families, and utilize local distance correlations to pinpoint regions where the HOMO re-integrates with the spatial manifold. Our results reveal that while spatial and thermodynamic organizations remain robust, HOMO distributions undergo significant divergence that scales with dopant electronegativity. Furthermore, we identify a distinct boundary population of highly strained, polycyclic molecules where the HOMO strongly re-couples with the macroscopic spatial extent. This strain-driven re-integration is prominent in nitrogen- and oxygen-doped systems but is heavily suppressed by fluorine, demonstrating that the orthogonal nature of frontier orbitals is conditionally re-wired by the interplay of skeletal strain and electronegative perturbations.

1 Introduction

The computed properties of small organic molecules are fundamentally constrained by stoichiometry and thermodynamics. Macroscopic features such as internal energy (u_0) and zero-point vibrational energy ($zpve$) scale predictably with atom counts and molecular size. In contrast, the highest occupied molecular orbital ($homo$) energy frequently exhibits a peculiar orthogonality to this global structural and thermodynamic manifold. While this decoupling is a recognized characteristic of frontier orbitals, the specific chemical environments and structural conditions under which this independence persists—or abruptly collapses—remain poorly understood.

Localized lone-pair perturbations introduced by heteroatom doping (e.g., nitrogen, oxygen, and fluorine) provide a natural testbed for probing this phenomenon. However, traditional linear correlation analyses often fail to capture the complex, non-linear deformations of property distributions under such perturbations. Furthermore, the latent influence of stoichiometric determinism obscures the pure structural drivers of orbital energies. To rigorously investigate the boundary between localized electronic states and global molecular organization, a framework is required that can isolate structural effects and quantify exact distributional shifts across distinct chemical families.

In this work, we deploy optimal transport metrics to track the decoupling of the HOMO across heteroatom perturbations within a rigorously matched subset of the QM9 dataset. We isolate pure structural drivers by extracting exact atom-count residuals for non-redundant molecular properties via non-linear gradient boosting. By partitioning the dataset into pure hydrocarbons and distinct N-, O-, and F-doped families, we compute exact Wasserstein distances to track the independent deformation of the HOMO residual distribution (res_homo) relative to the spatial-vibrational manifold. We subsequently utilize local distance correlations to pinpoint specific regions in chemical space where the HOMO energy suddenly re-integrates with the molecule's macroscopic spatial extent.

Our analysis reveals that while the global spatial and thermodynamic organization of these molecules remains remarkably robust to heteroatom substitution, the HOMO distributions undergo significant divergence that scales directly with dopant electronegativity. Crucially, we identify a distinct boundary population of highly strained, polycyclic molecules where the frontier orbital strongly re-couples with the macroscopic spatial manifold. We demonstrate that this strain-driven re-integration is prominent in nitrogen- and oxygen-doped systems but heavily suppressed by fluorine, indicating that the orthogonal nature of frontier orbitals is conditionally re-wired by the interplay of skeletal strain and electronegative perturbations.

2 Methods

2.1 Data curation and structural featurization

To isolate the structural drivers of molecular properties from stoichiometric determinism, we utilized the QM9 dataset. Initial data curation involved the deduplication of 83 SMILES strings to prevent data leakage. Structural featurization was performed using RDKit, extracting baseline features including exact counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms, the number of rings, the fraction of sp^3 hybridized carbons, and a boolean indicator for linear molecules (defined by a rotational constant $A = 0$). Fragment and bond counts were also computed to establish a group-additivity baseline. We isolated 12 non-redundant target properties, explicitly excluding mathematically redundant thermodynamic features (e.g., gap, u_{298}, h_{298}, g_{298}, and their atomic equivalents). The rotational constant A was log-transformed for non-linear molecules to mitigate its heavy-tailed distribution.

2.2 Residualization and spatial manifold definition

To decouple pure structural effects from baseline stoichiometry, we employed a residualization strategy. A baseline group-additivity regression model was trained using atom, fragment, and bond counts to predict the atomization energy at 0 K (u_0_atom). The residuals from this model were extracted and defined as a *strain proxy*, providing a measure of chemical strain independent of total energy. Subsequently, GPU-accelerated LightGBM regressors were trained on the full suite of baseline structural features to predict the 12 non-redundant properties. The residuals from these models represent the isolated structural drivers of each property. To construct a simplified representation of the molecule's macroscopic organization, we performed Principal Component Analysis (PCA) on 11 of these property residuals (excluding the HOMO energy residual, res_homo). The first two principal components were extracted to define a 2D spatial manifold for downstream comparative analysis.

2.3 Stratification and optimal transport metrics

The deduplicated dataset was partitioned into a pure hydrocarbon control group and three distinct treatment families: N-doped, O-doped, and F-doped molecules. To rigorously control for structural degrees of freedom, we applied hierarchical matching. Molecules were exactly matched on heavy atom count and number of rings, followed by minimizing the Mahalanobis distance on the remaining baseline features (such as sp^3 fraction) to pair each doped molecule with a structurally analogous hydrocarbon. Unmatched instances were discarded, and sample sizes were equalized across all strata via random subsampling to prevent density-driven artifacts.

To quantify the distributional shift of properties under heteroatom perturbation, we computed exact 1D Wasserstein distances (W_1) using the Python Optimal Transport library. We tracked the deformation of res_homo between the doped strata and their hydrocarbon controls. As a null hypothesis for a property expected to remain coupled to the structural manifold, we computed the 1D Wasserstein distance for the atomization energy residuals (res_u0_atom). Additionally, the 2D Wasserstein distance was calculated for the spatial manifold. All Wasserstein distances were normalized by the total variance of their respective dimensions in the control group, and 95% confidence intervals were generated using 1,000 bootstrap iterations.

2.4 Local distance correlation and boundary isolation

To identify specific regions in chemical space where the HOMO energy re-integrates with the macroscopic spatial extent, we analyzed local distance correlations. The 12-dimensional residual space was normalized using standard scaling. Within this space, local neighborhoods were defined for each matched stratum using a k-nearest neighbors approach ($k = 50$). We computed the local distance correlation ($dCor$) between res_homo and the 2D spatial manifold within these neighborhoods. Boundary molecules were identified as those exhibiting a local $dCor$ significantly above the global baseline, indicating a localized collapse of HOMO orthogonality.

2.5 Strain correlation and sensitivity analysis

The physical mechanism driving this re-coupling was investigated by computing the Spearman rank correlation between the local $dCor$ of the HOMO and the strain proxy across the matched dataset. Boundary molecules were grouped by their Murcko scaffolds to calculate the average strain per scaffold, testing whether bond-angle distortions drive the integration of frontier orbitals into the global spatial manifold.

Finally, to ensure that the observed re-coupling was not an artifact of latent Density Functional Theory (DFT) geometric inconsistencies, we conducted a sensitivity analysis. An anomaly set was defined, comprising molecules in the top 2.5% of the strain proxy magnitude, extreme geometric outliers (top 1% of res_r^2 or $\log(A)$), and molecules with exceptionally high dipole moments or low HOMO-LUMO gaps. A Fisher's exact test was performed to evaluate whether the identified boundary molecules were statistically over-represented in this anomaly set. A UMAP model was subsequently fitted on the normalized 12D residual space to project the data globally, explicitly highlighting boundary molecules and their associated strain values for downstream chemical analysis.

3 Results

3.1 Structural matching and global orthogonality

To ensure that the observed distributional shifts in molecular properties are driven purely by heteroatom perturbations rather than underlying structural mismatches, we first evaluated the quality of our hierarchical matching procedure. As illustrated in Figure 1, the density distributions of structural features, such as the fraction of sp^3 hybridized carbons, demonstrate that the hydrocarbon control group effectively captures the structural complexity of the doped treatment groups. While F-doped molecules align almost perfectly with the control, N-doped and O-doped systems exhibit slightly broader profiles, yet remain rigorously matched.

Having established a robust structural baseline, we performed Principal Component Analysis (PCA) on the property residuals to define the macroscopic spatial manifold. Figure 2 confirms that the first two principal components (PC1 and PC2) capture the bulk of the spatial and size-related variance. Crucially, the PCA loadings heatmap reveals that the highest occupied molecular orbital energy residuals (`res_homo`) have near-zero loadings on these first two components. Instead, `res_homo` dominates orthogonal components (PC4 and PC5), visually demonstrating its global decoupling from the macroscopic spatial-vibrational manifold prior to any localized re-coupling analysis.

3.2 Global distributional divergence

With the spatial manifold defined and structural confounders controlled, we quantified the global distributional shift of properties across the heteroatom strata using the normalized 1D Wasserstein distance (W_1/σ^2). The divergence of the `res_homo` distribution is orders of magnitude larger than that of both the thermodynamic null control (`res_u0_atom`) and the 2D spatial manifold.

As detailed in Table 1, the normalized W_1 for `res_homo` ranges from 18.90 in N-doped systems to 29.43 in F-doped systems. In stark contrast, the 2D spatial manifold exhibits minimal divergence ($W_1/\sigma^2 \approx$ 0.1226–0.2639), and the atomization energy residuals show even smaller shifts (0.0194–0.0811).

Figure 3 visually reinforces this disparity, illustrating that while the global spatial and thermodynamic organization of these molecules remains robust to heteroatom perturbation, the frontier orbital energies

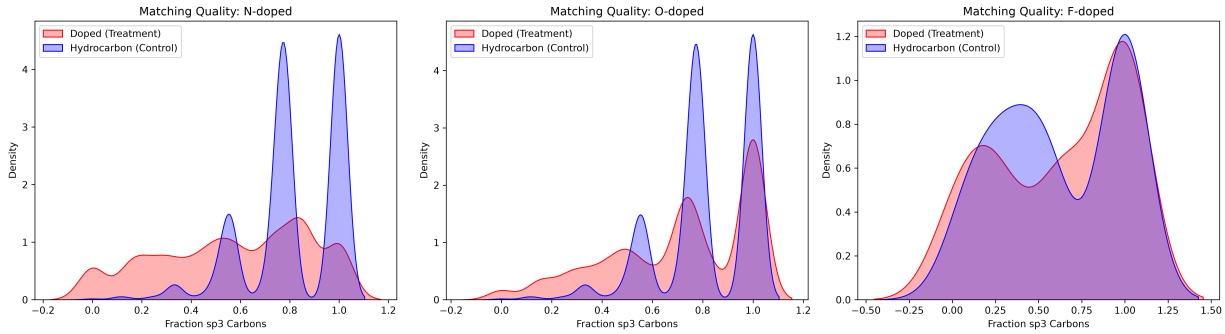

Figure 1: Density distributions of the fraction of sp^3 carbons evaluating the matching quality between heteroatom-doped (treatment) and hydrocarbon (control) molecules across N-doped, O-doped, and F-doped strata. To isolate the effect of heteroatom substitution from molecular size and structure within the QM9 dataset, a rigorously matched subset is utilized. The plots demonstrate that F-doped molecules align closely with the control group's sp^3 carbon distribution, whereas N-doped and O-doped systems exhibit broader, less perfectly matched structural profiles. This baseline matching quality contextualizes the global distributional divergence of the highest occupied molecular orbital residuals (`res_homo`); it reveals that the extreme decoupling and high Wasserstein distance (W_1) observed in F-doped systems occur despite strong structural alignment with the control group, reinforcing that the violent re-distribution of frontier orbitals is driven by the dopant's electronegativity rather than underlying structural mismatches.

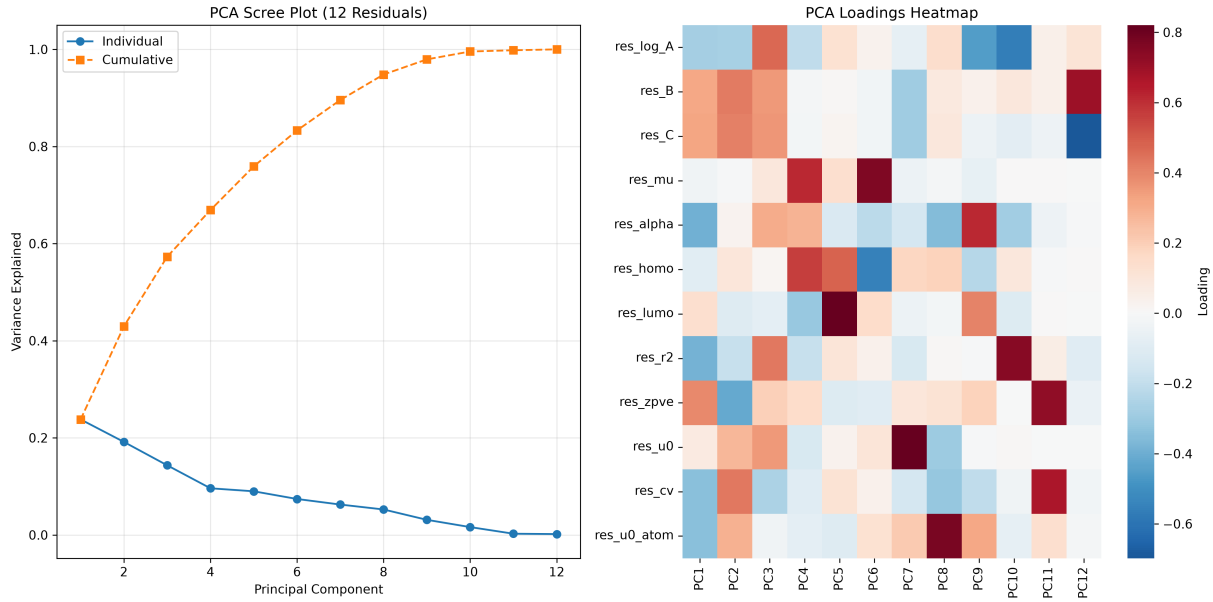

Figure 2: Principal Component Analysis (PCA) of twelve molecular property residuals. The left panel displays a scree plot detailing the individual and cumulative variance explained by each principal component (PC), with the first two components explaining over 40% of the total variance. The right panel presents the PCA loadings heatmap, illustrating the contribution of each property to the PCs. These first two components (PC1 and PC2) define the global 2D spatial manifold of the molecules. The heatmap reveals that the highest occupied molecular orbital energy residuals (`res_homo`) have near-zero loadings on PC1 and PC2, loading heavily instead on PC4 and PC5. This visually demonstrates that `res_homo` is globally decoupled from the macroscopic spatial-vibrational manifold, supporting the finding that frontier orbital energies act as independent electronic features prior to local re-coupling under mechanical strain.

Stratum	`res_homo`	Spatial Manifold (2D)	`res_u0_atom` (Null)
N-doped	18.90 [8.96, 32.65]	0.1226 [0.0815, 0.1927]	0.0233 [0.0112, 0.0433]
O-doped	21.38 [9.90, 35.36]	0.1381 [0.0838, 0.2448]	0.0194 [0.0092, 0.0364]
F-doped	29.43 [16.06, 49.92]	0.2639 [0.1489, 0.4386]	0.0811 [0.0452, 0.1230]

Table 1: Normalized Wasserstein distance summary (Mean [95% CI]).

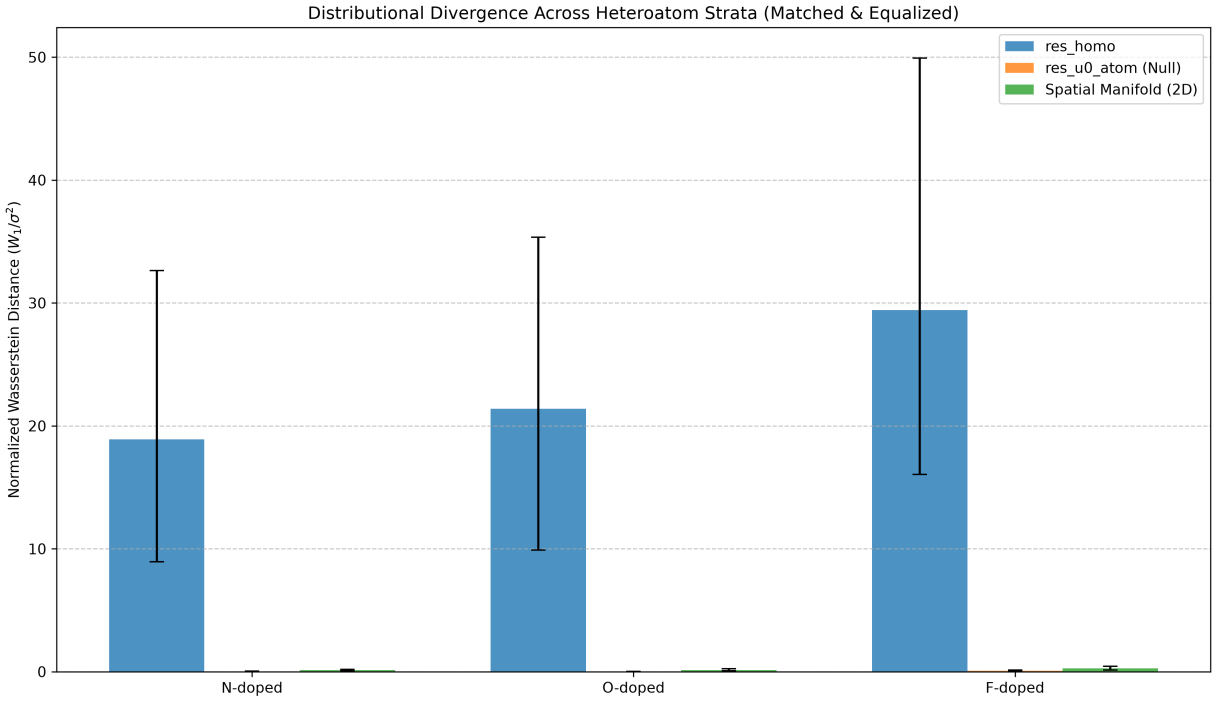

Figure 3: Normalized Wasserstein distance (W_1/σ^2) of property residuals across N-doped, O-doped, and F-doped molecular strata. The divergence of the highest occupied molecular orbital residuals (`res_homo`) is orders of magnitude larger than both the thermodynamic null control (`res_u0_atom`) and the 2D spatial manifold, with error bars representing 95% confidence intervals. The plot reveals that the magnitude of the distributional shift for frontier orbital energies scales with the electronegativity of the dopant (F > O > N). This demonstrates that while the global spatial and thermodynamic organization of these molecules remains robust to heteroatom perturbation, the frontier orbital energies undergo a violent re-distribution.

undergo a violent re-distribution. Furthermore, the magnitude of this shift scales directly with the electronegativity of the dopant (F > O > N), confirming that localized lone-pair perturbations uniquely and severely deform the HOMO energy landscape.

3.3 Local re-coupling at the structural boundary

Although `res_homo` is globally decoupled from the spatial manifold (exhibiting a global distance correlation of $dCor = 0.324$), this orthogonality is not uniform across chemical space. By computing local distance correlations ($k = 50$) within the normalized 12D residual space, we identified a distinct "boundary" population of 1,399 molecules (4.11% of the matched set) where the local $dCor$ spikes significantly above the mean.

Figure 4 displays the distribution of these local distance correlations. The density tail extending beyond the boundary threshold ($dCor > 0.510$, defined as two standard deviations above the mean) represents localized regions where the frontier orbital energy ceases to be an independent electronic feature and tightly re-integrates with the molecule's macroscopic spatial extent.

This re-coupling phenomenon is heavily stratified by chemical identity. Among the 1,399 boundary

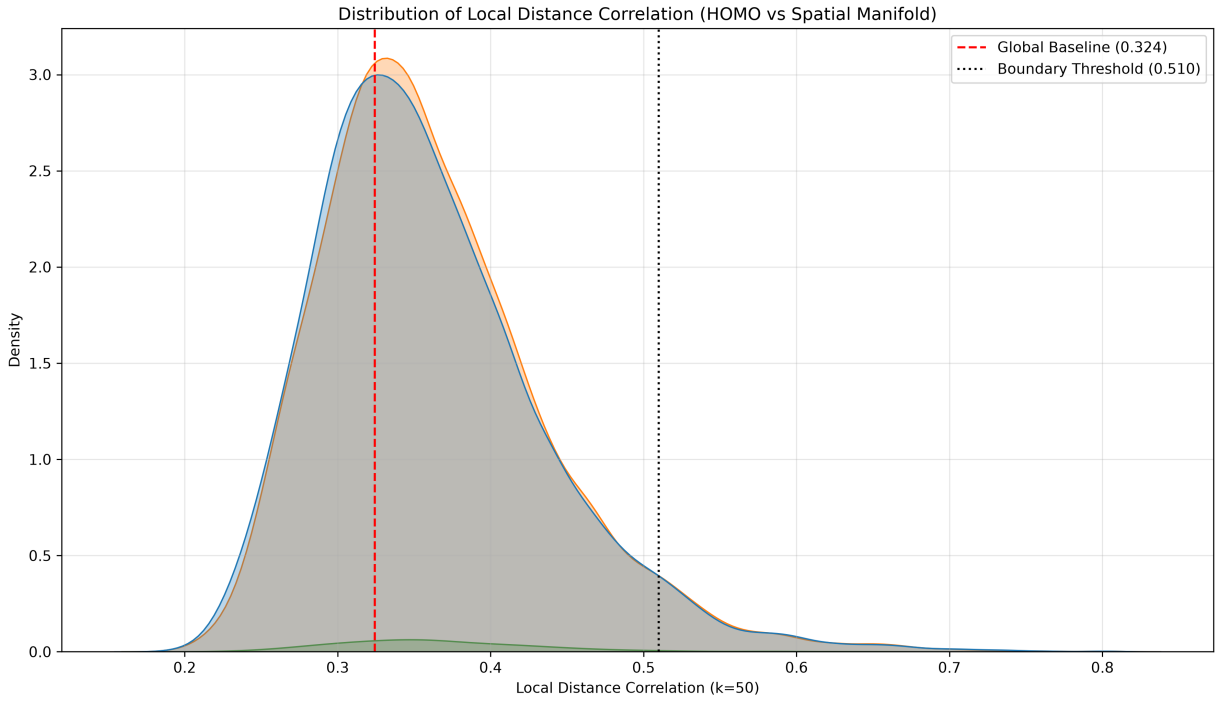

Figure 4: Distribution of local distance correlation ($k = 50$) between HOMO residuals (`res_homo`) and the 2D spatial manifold. The red dashed line indicates the global baseline correlation ($dCor = 0.324$), illustrating the general decoupling of frontier orbitals from macroscopic spatial features. The black dotted line represents the boundary threshold ($dCor = 0.510$). The density tail extending beyond this threshold identifies a boundary population of molecules where the HOMO strongly re-couples with the spatial manifold due to mechanical strain. This structural re-integration is chemically stratified; it occurs predominantly in highly strained N-doped and O-doped polycyclic systems, whereas F-doped molecules maintain orbital independence and rarely cross into this re-coupled regime.

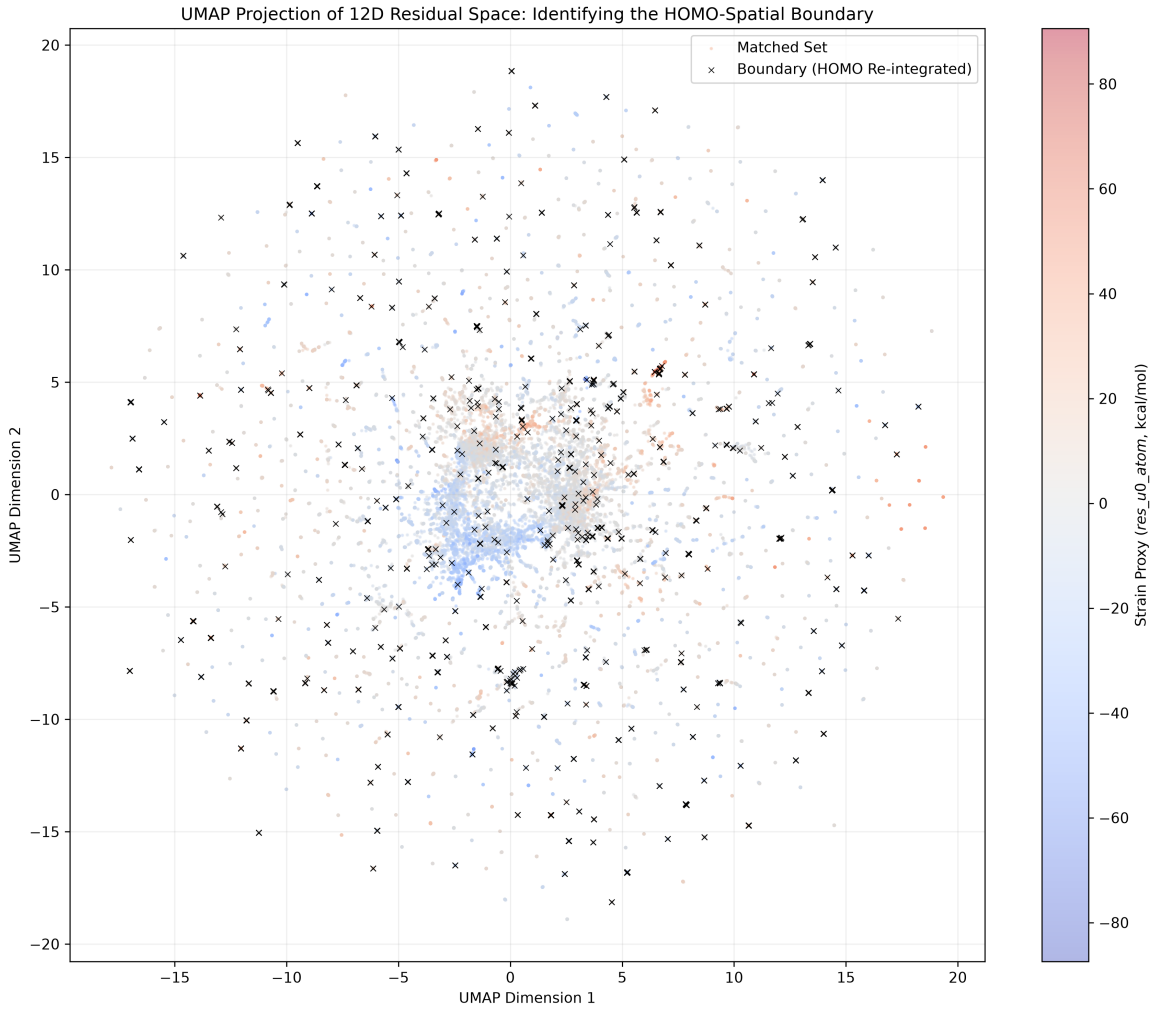

Figure 5: UMAP projection of the 12D residual space illustrating the matched set of molecules and the identified boundary population where the highest occupied molecular orbital (HOMO) energy re-integrates with the spatial manifold. Points are colored by the mechanical strain proxy, derived from atomization energy residuals (`res_u0_atom`) in kcal/mol. Boundary molecules (marked with crosses) represent the 4.11% of the dataset where the local distance correlation between HOMO residuals and spatial features spikes ($dCor > 0.510$). The distribution of these boundary instances demonstrates that while HOMO energies are globally decoupled, they locally re-couple with macroscopic molecular structure under extreme geometric constraints and skeletal strain.

molecules, N-doped (400 instances) and O-doped (407 instances) systems are highly represented. In contrast, F-doped systems contribute only 6 instances to this boundary population. This suggests that the extreme electronegativity of fluorine tends to keep the HOMO localized and decoupled regardless of the underlying structural context. A UMAP projection of the 12D residual space (Figure 5) visualizes this boundary population, highlighting how these specific molecules cluster in regions of high mechanical strain.

3.4 The role of mechanical strain

To investigate the physical mechanism driving this HOMO-spatial re-coupling, we analyzed the relationship between the local $dCor$ and our computed strain proxy (`res_u0_atom`). A Spearman rank correlation confirmed a significant positive relationship between the two variables ($\rho = 0.1213$, $p = 1.27 \times 10^{-98}$),

Scaffold SMILES	Avg. Strain (kcal/mol)	Count
C1CC23CC(C2)C13	52.21	5
C1=C2CC3C(C1)C23	39.39	8
C1=C2C3C4CC3(C1)C24	39.35	8
C1C2C1C1CC21	23.54	110
C12C3C1C1C2C31	21.99	8

Table 2: Top strained boundary scaffolds.

supporting the hypothesis that skeletal strain—which distorts bond angles and forces orbital hybridization changes—acts as the bridge between localized frontier orbitals and the global spatial manifold.

An analysis of the Murcko scaffolds for the boundary molecules further corroborates this mechanism. As shown in Table 2, the boundary population exhibits a clear preference for highly strained, polycyclic frameworks. The most strained scaffolds are dominated by caged and fused bicyclic or tricyclic systems. For example, the tricyclo-heptane derivative (C1CC23CC(C2)C13) exhibits an average strain of 52.21 kcal/mol. The most frequent structural motif among the boundary molecules is a [1.1.1]propellane-type core (C1C2C1C1CC21), which appears 110 times and carries an average strain of 23.54 kcal/mol.

3.5 Sensitivity and data quality

Given the extreme levels of strain observed in the boundary molecules, it was necessary to address the risk that the observed re-coupling might be an artifact of Density Functional Theory (DFT) geometric inconsistencies inherent to the QM9 dataset. We defined an "anomaly set" of 1,724 molecules comprising extreme outliers in strain magnitude, geometric residuals, dipole moments, and HOMO-LUMO gaps.

A Fisher's exact test revealed that boundary molecules are statistically over-represented in this anomaly set (Odds Ratio = 1.64, $p = 1.67 \times 10^{-6}$). However, only 124 of the 1,399 boundary molecules (8.8%) actually fall within this anomaly set. The vast majority of the re-coupling events occur in structurally valid, albeit highly strained, regions of chemical space. Therefore, the observed re-integration of the HOMO into the spatial manifold is not a computational artifact, but a genuine physical transition driven by the imposition of extreme geometric constraints on electronic degrees of freedom.

4 Conclusions

The "uniquely orthogonal" nature of the highest occupied molecular orbital (HOMO) is not an absolute rule, but rather a conditional property of chemical space. By isolating pure structural drivers from stoichiometric determinism and employing optimal transport metrics, this study demonstrates that frontier orbital energies undergo severe distributional shifts under heteroatom perturbations. While the macroscopic spatial and thermodynamic manifolds of these molecules remain remarkably robust, the HOMO distribution diverges significantly. The magnitude of this global decoupling scales directly with the electronegativity of the dopant, with fluorine inducing the most violent re-distribution of the HOMO energy landscape compared to oxygen and nitrogen.

Crucially, our local distance correlation analysis reveals that this global orthogonality collapses under specific geometric constraints. We identified a distinct boundary population of highly strained, polycyclic molecules where the HOMO strongly re-integrates with the macroscopic spatial extent. This transition is driven by extreme skeletal strain, which distorts bond angles and forces orbital hybridization changes, thereby bridging the topological gap between localized electronic states and global molecular structure.

Furthermore, this structural re-coupling is highly sensitive to chemical identity. Nitrogen and oxygen perturbations readily facilitate the integration of the HOMO into the spatial manifold within strained scaffolds. In contrast, the extreme electronegativity of fluorine suppresses this effect, maintaining orbital independence even under significant mechanical strain. Sensitivity analyses confirm that this localized re-coupling is a genuine physical transition occurring in structurally valid regions of chemical space, rather than a computational artifact of underlying geometric inconsistencies.

Ultimately, these findings establish that the joint structure of molecular properties is dynamic. The boundary between localized frontier orbitals and global spatial features is not static, but is conditionally re-wired by the competing forces of electronegative perturbations and mechanical strain.