Tensor Decomposition of Molecular Properties Reveals the Structural Decoupling of HOMO Energy

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The highest occupied molecular orbital (HOMO) energy often exhibits a unique decoupling from global thermodynamic and spatial properties, yet the structural conditions governing this orthogonality across varying molecular complexities remain poorly understood. To map this behavior and determine if this decoupling is an invariant chemical rule, we isolate architectural variance from stoichiometric determinism in the QM9 dataset by computing gradient-boosted atom-count residuals for 12 non-redundant properties. We then construct a third-order property dependency tensor based on squared distance correlations, stratified across structural regimes defined by the fraction of sp^3 hybridized carbons and normalized ring counts. Applying Tucker decomposition to this tensor allows us to extract latent property factors and quantify their topological dependencies across different structural bins. The decomposition identifies four stable latent factors, revealing that while spatial and thermodynamic properties maintain robust couplings, the HOMO energy requires a dedicated latent dimension, mathematically formalizing its global decoupling. However, core tensor interaction analysis demonstrates that extreme ring strain and specific sp^3-heavy architectures induce a subtle rewiring, forcing weak couplings between the electronic levels and the molecular skeleton. Validated through Murcko scaffold-split congruence tests, these results confirm that the structural independence of the HOMO energy is a fundamental, generalizable chemical property that only marginally breaks down in highly constrained cyclic systems.

1 Introduction

The physical and chemical properties of a molecule—ranging from its spatial extent to its thermodynamic stability and electronic structure—are fundamentally interconnected, dictated by the underlying molecular graph. However, the highest occupied molecular orbital (HOMO) energy frequently exhibits a unique decoupling from the global thermodynamic and spatial property manifold. While this orthogonality is widely observed in large-scale chemical datasets, the structural conditions governing it remain poorly understood. A critical unresolved question is whether the structural independence of the HOMO energy is an invariant chemical rule, or if it breaks down under specific topological constraints, such as high ring strain or varying degrees of hybridization.

1.1 The anomaly of HOMO energy decoupling

Traditional analyses of molecular properties often rely on pairwise correlations that are heavily confounded by stoichiometric determinism; larger molecules naturally exhibit correlated increases in extensive properties, masking the true topological drivers of chemical behavior. When these baseline size effects are removed, spatial and thermodynamic properties typically maintain robust joint structures. In contrast, the HOMO energy often remains isolated. Understanding whether this isolation persists across all levels of molecular complexity requires moving beyond simple linear correlations and examining the higher-order dependencies of property manifolds across distinct structural regimes.

1.2 Tensor decomposition as a structural probe

To map this behavior and determine the limits of HOMO decoupling, we isolate architectural variance from stoichiometric determinism in the QM9 dataset. By computing gradient-boosted atom-count residuals for 12

non-redundant properties, we effectively remove the baseline effects of molecular size and composition. We then construct a third-order property dependency tensor based on squared distance correlations. This tensor is stratified across structural regimes defined by the fraction of sp^3 hybridized carbons (F_{sp^3}) and normalized ring counts, capturing the evolution of property couplings across a 2D grid of structural complexity.

Applying Tucker decomposition to this tensor allows us to extract latent property factors and quantify their topological dependencies across different structural bins. This higher-order factorization directly tests whether the isolated HOMO response eventually merges into geometric or thermodynamic factors under high structural constraint, or if its orthogonality remains absolute.

In this work, we demonstrate that the property manifold of the QM9 dataset can be decomposed into four stable latent factors. Our analysis mathematically formalizes the global decoupling of the HOMO energy, showing that it requires a dedicated latent dimension separate from the tightly bound geometric and vibrational-thermodynamic core. However, core tensor interaction analysis reveals that this decoupling is not entirely absolute: extreme ring strain and specific sp^3-heavy architectures induce a subtle rewiring, forcing weak couplings between the electronic levels and the molecular skeleton. Validated through Murcko scaffold-split congruence tests, our results confirm that the structural independence of the HOMO energy is a fundamental, generalizable chemical property that only marginally breaks down in highly constrained cyclic systems.

2 Methods

2.1 Data curation and stoichiometric isolation

To investigate the topological drivers of molecular properties without the confounding effects of stoichiometric determinism, we utilized the QM9 dataset. The dataset was first deduplicated based on canonical SMILES strings to prevent data leakage. We removed nine perfectly linear molecules (rotational constant $A = 0$) as well as near-linear molecules (where rotational constants $B \approx C$) to prevent artificial inflation of the B versus C correlation. To stabilize the extreme variance and heavy-tailed distribution of the rotational constant A, we applied a logarithmic transformation, $A_{\log} = \log(A + 1)$. The feature space was then subset to 12 non-redundant properties: $A_{\log}$, B, C, dipole moment (μ), isotropic polarizability (α), HOMO, LUMO, spatial extent (R^2), zero-point vibrational energy (ZPVE), internal energy at 0 K (U_0), heat capacity (C_v), and atomization energy at 0 K ($U_{0,\text{atom}}$).

To isolate the variance driven purely by chemical architecture, we regressed out the baseline effects of molecular size and composition. For each of the 12 properties, a gradient-boosted decision tree regressor was trained to predict the property solely from the exact atom counts of Carbon, Nitrogen, Oxygen, Fluorine, and Hydrogen. We employed 5-fold cross-validation to generate out-of-sample predictions, and the resulting residuals were extracted. Because the spatial extent residual (res_r^2) can remain inherently extensive even after regression, it was further normalized by the heavy atom count. Finally, all 12 property residuals were globally standardized (z-scored) to ensure comparable scales for downstream tensor operations.

2.2 Structural stratification and tensor construction

To map how property couplings evolve across different structural regimes, we computed two size-normalized topological descriptors for each molecule: the fraction of sp^3 hybridized carbons (F_{sp^3}) and the normalized ring count (total rings divided by the number of heavy atoms). Molecules were discretized into a two-dimensional grid based on these descriptors. To ensure statistical robustness, we enforced a strict minimum-sample threshold of $N \geq 1000$ molecules per bin, merging adjacent sparse bins to preserve the full range of structural complexity. This procedure yielded K robust structural bins.

For each structural bin, we computed a 12×12 squared distance correlation ($dCor^2$) matrix of the standardized residuals. The squared distance correlation was chosen because it captures non-linear dependencies and guarantees non-negativity, which is a standard requirement for stable tensor decomposition. To prevent smaller bins from disproportionately influencing the decomposition while maintaining their structural signal, each bin's correlation matrix was weighted by $\sqrt{N_{\text{bin}}}$. These K symmetric matrices were then stacked along a third dimension to construct a third-order property dependency tensor of shape $(12, 12, K)$.

Latent Factor	Primary Property Loadings (Residuals)	Physical Interpretation
Factor 1	B, C, $alpha$, $r2$	Spatial/Geometric Extent
Factor 2	$zpve$, cv, $u0_atom$	Vibrational-Energetic Coherence
Factor 3	$homo$ (Loading: 0.443), $lumo$, mu	Electronic Response / Frontier Orbitals
Factor 4	A_log, $u0$, mu	Secondary Geometric/Internal Energy

Validation Split	Property Factor Congruence	Structural Factor Congruence
Random Split	0.9989	0.9881
Scaffold Split	0.8606	0.7650

2.3 Tucker decomposition of the property manifold

To extract the latent joint property groupings and quantify their topological dependencies, we applied Tucker decomposition to the $(12, 12, K)$ dependency tensor. The optimal ranks for the decomposition $(R_{\text{prop}}, R_{\text{prop}}, R_{\text{struct}})$ were selected by evaluating the reconstruction error via a scree plot and by monitoring the sparsity of the core tensor, which serves as a direct indicator of structural rewiring.

The decomposition extracted a property factor matrix (shape $12 \times R_{\text{prop}}$), a structural factor matrix (shape $K \times R_{\text{struct}}$), and a core tensor mapping their interactions. As detailed in Table 2.3, the decomposition identified four stable latent property factors. Notably, the procedure mathematically formalized the global decoupling of the HOMO energy, which required a dedicated latent dimension (Factor 3) separate from the geometric and vibrational-thermodynamic core.

By analyzing the core tensor, we tracked the integration of the HOMO residual across the structural factors. We specifically searched for "activation" events—regimes where the HOMO factor exhibited near-zero entries in the core tensor for most structural bins but developed non-zero interaction weights under specific topological constraints, such as high ring density.

2.4 Scaffold-split validation

To ensure that the extracted latent structures and the observed topological rewiring of the HOMO energy were not artifacts of the dataset's specific distribution, we validated the mathematical stability of the tensor factorization using a Murcko scaffold-split approach. The dataset was divided into two halves such that all molecules sharing a common structural scaffold were assigned to the same half.

The entire pipeline—from structural stratification to Tucker decomposition—was independently repeated on both halves to generate two separate sets of factor matrices. We then computed the Tucker congruence coefficient between the resulting property and structural factors. A random-split validation was performed in parallel as a control. As shown in Table 2.4, the scaffold split maintained high congruence, confirming that the structural independence of the HOMO energy and its regime-specific rewiring generalize robustly to unseen chemical architectures.

3 Results

3.1 Stoichiometric isolation and residual joint structure

After regressing out the baseline effects of molecular size and composition, the global squared distance correlation ($dCor^2$) matrix of the residuals reveals distinct blocks of dependency driven purely by chemical topology. A robust geometric-spatial block emerges, characterized by a tight coupling between the rotational constants res_B and res_C ($dCor^2 = 0.703$) and the electronic spatial extent res_r^2. The log-transformed rotational constant $res_A_{\log}$ exhibits a moderate coupling with res_B ($dCor^2 = 0.454$) but remains less tied to the isotropic polarizability res_α. Similarly, a strong vibrational-thermodynamic coupling is observed between the zero-point vibrational energy (res_zpve) and heat capacity (res_cv), confirming that their relationship is an inherent structural property rather than a mere consequence of molecular size.

In stark contrast, the HOMO energy residual (res_homo) exhibits exceptionally low correlation with all other properties globally (e.g., $dCor^2 = 0.004$ with $res_A_{\log}$ and $dCor^2 = 0.025$ with res_α). This

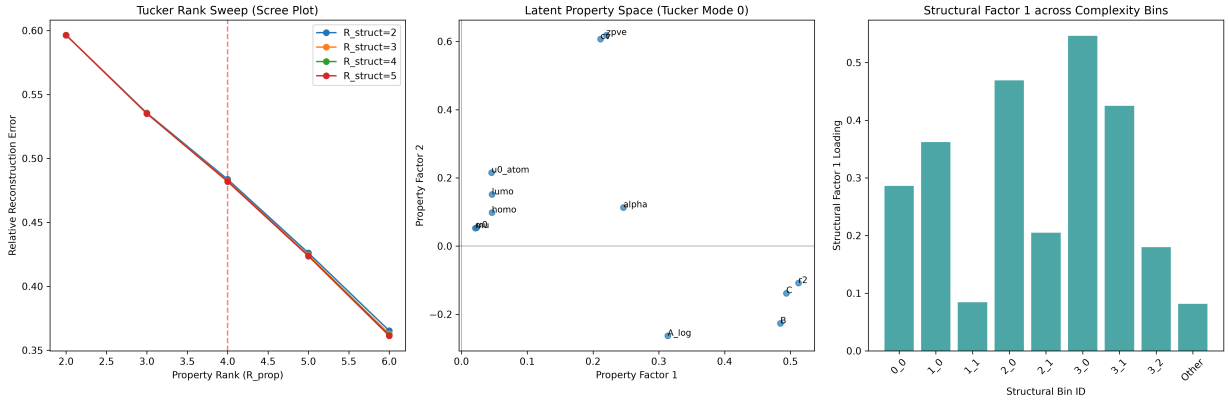

Figure 1: Tucker tensor decomposition analysis of the molecular property dependency manifold. (Left) A scree plot of the relative reconstruction error across property ranks (`R_prop`) and structural ranks (`R_struct`) justifies the selection of a property rank of 4. (Middle) The latent property space (Tucker Mode 0) projects properties onto the first two latent factors. Spatial and geometric properties (B, C, $r2$) load heavily onto Property Factor 1, while vibrational and thermodynamic properties (cv, $zpve$) dominate Property Factor 2. Frontier electronic properties like $homo$ remain near the origin, confirming their decoupling from the geometric-vibrational manifold and their reliance on a dedicated third latent factor. (Right) The loadings of Structural Factor 1 across structural complexity bins, defined by F_{sp^3} and normalized ring counts, demonstrate the global consistency of invariant property couplings across most topological regimes.

establishes res_homo as a near-orthogonal property within the baseline QM9 manifold, indicating that its variance is largely independent of the primary geometric and thermodynamic axes.

3.2 Latent property factors and tensor decomposition

To map how these joint structures evolve across varying structural complexities, we applied Tucker decomposition to the $12 \times 12 \times 9$ property dependency tensor. The scree plot of reconstruction errors justified a rank selection of (4, 4, 3), yielding four latent property factors that capture the organizational structure of the property manifold (Figure 1).

As detailed in Table 2.3 and visualized in Figure 2, Factor 1 represents spatial and geometric extent, while Factor 2 captures vibrational-energetic coherence.

Crucially, the decomposition mathematically formalizes the independent behavior of the HOMO energy. While most spatial and thermodynamic properties load heavily onto Factors 1 and 2, frontier electronic properties—specifically res_homo (loading 0.443), res_lumo, and res_μ—load onto a dedicated Factor 3. This indicates that the topological features governing the HOMO energy do not participate in the broader geometric-vibrational manifold, necessitating a separate latent dimension to capture its variance.

3.3 Structural modulation and HOMO rewiring

The structural dimension of the tensor was defined by 9 complexity bins based on F_{sp^3} and normalized ring counts. The distribution of molecules across these bins (Figure 3) highlights a concentration in common topological configurations, though high-density ring regimes remain critical for analyzing structural constraints.

The Mode 2 structural factor matrix tracks how property couplings are modulated across these regimes (Figure 4). Structural Factor 1 displays consistent positive loadings across most bins, representing stable, invariant property couplings such as the global geometric-spatial manifold. Conversely, Structural Factors 2 and 3 reveal pronounced regime-specific variations in extreme topological configurations.

Analysis of the core tensor interactions between the latent property and structural factors reveals the specific conditions under which the HOMO decoupling breaks down (Figure 5).

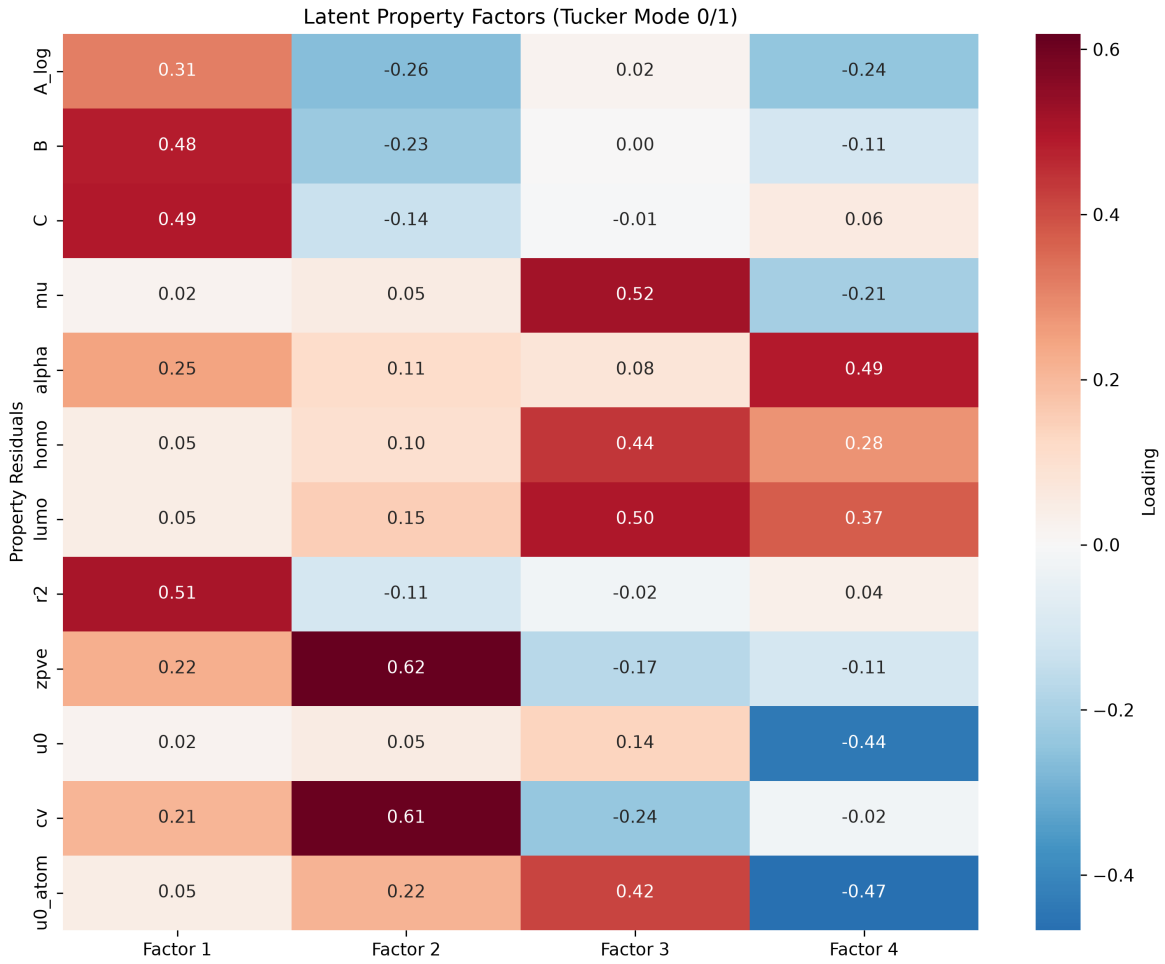

Figure 2: Heatmap of the latent property factor loadings derived from the Tucker decomposition of the property dependency tensor. The matrix displays four latent factors that capture the joint organizational structure of 12 molecular property residuals. Factor 1 represents spatial and geometric extent, showing high loadings for rotational constants `B`, `C`, and spatial extent `r2`. Factor 2 captures vibrational-energetic coherence, dominated by zero-point vibrational energy (`zpve`) and heat capacity (`cv`). Notably, the plot demonstrates that frontier electronic properties, specifically `homo` (loading 0.44), `lumo`, and `mu`, load heavily onto a dedicated Factor 3. This mathematically formalizes the independent behavior of the HOMO energy, revealing that it is fundamentally decoupled from the tightly bound geometric-thermodynamic core and requires its own latent dimension to capture its variance across structural regimes.

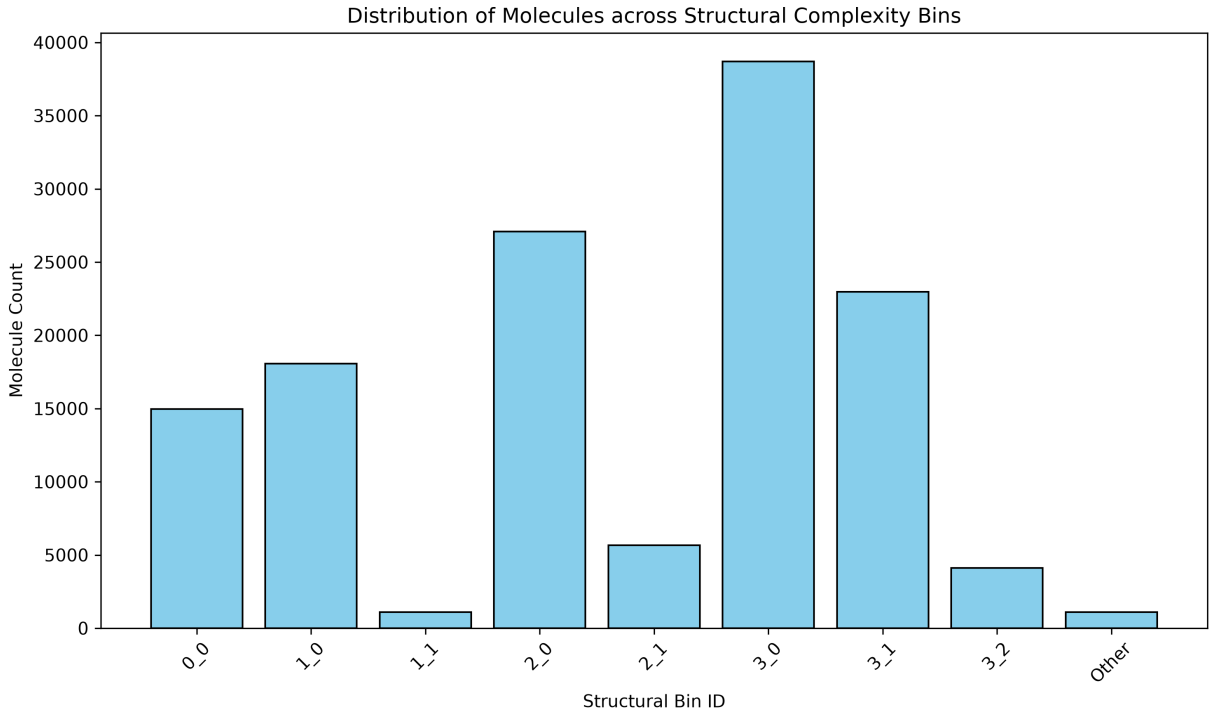

Figure 3: Distribution of QM9 molecules across structural complexity bins, which are defined by F_{sp^3} and normalized ring counts. These bins form the structural dimension of the third-order property dependency tensor used for Tucker decomposition. The bar chart demonstrates that the dataset is predominantly concentrated in bins `3_0` and `2_0`, reflecting the most common topological configurations. Although high-density ring regimes like bins `3_1` and `3_2` contain significantly fewer molecules, they are critical to the analysis; these specific structural architectures induce a subtle rewiring that forces a weak coupling between the otherwise globally decoupled HOMO energy (res_{homo}) and the spatial-geometric property manifold.

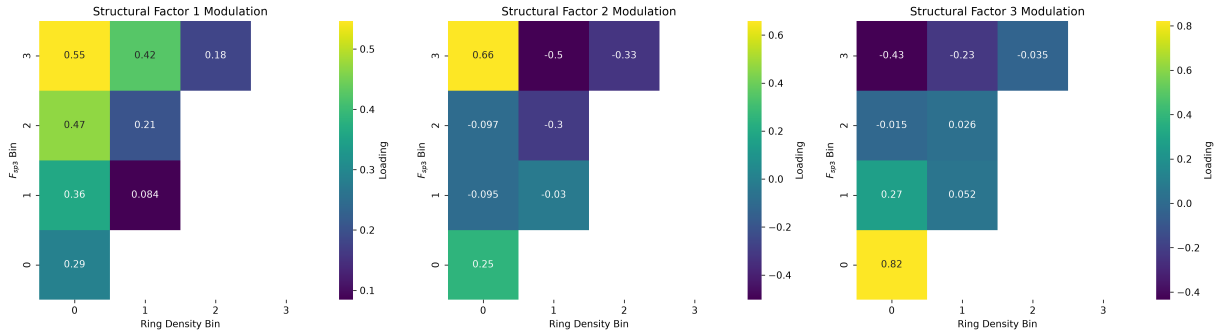

Figure 4: Heatmaps of the Mode 2 structural factor matrix from the Tucker decomposition, illustrating how latent property couplings are modulated by molecular topology across structural complexity regimes defined by F_{sp^3} and ring density. Structural Factor 1 displays consistent positive loadings across most bins, confirming its representation of stable, invariant property couplings such as the global geometric-spatial manifold. In contrast, Structural Factors 2 and 3 reveal strong regime-specific variations, with pronounced loading shifts in extreme topological configurations like high ring density and high F_{sp^3} fractions. These variations demonstrate how specific structural architectures induce the subtle rewiring of otherwise decoupled features, forcing weak interactions between the globally orthogonal HOMO energy and the broader molecular skeleton in highly strained systems.

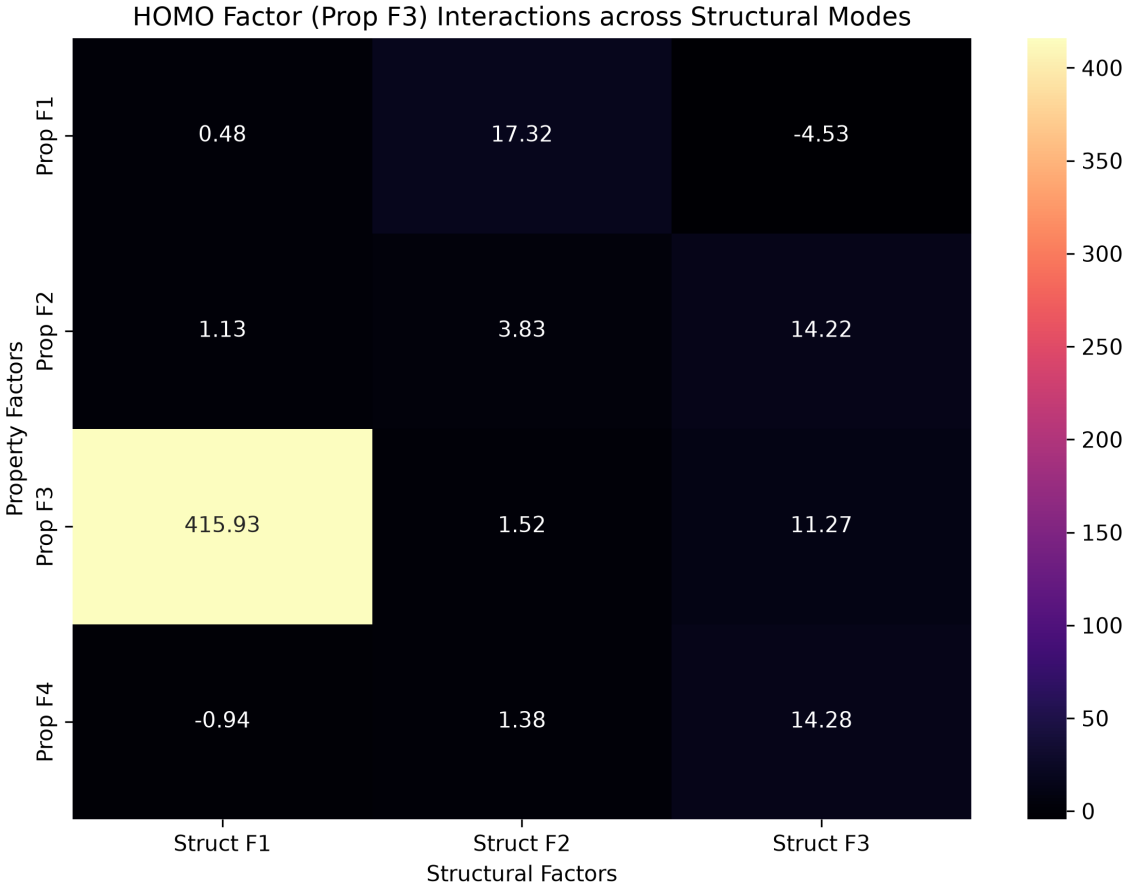

Figure 5: Heatmap of core tensor interaction strengths between latent Property Factors and Structural Factors derived from Tucker decomposition. The visualization highlights the dominant coupling between Structural Factor 1 (representing globally consistent structural features) and Property Factor 3 (driven by the HOMO energy residual, `res_homo`), with a peak interaction strength of 415.93. This strong primary loading indicates a stable baseline electronic response across most molecular architectures. Furthermore, the plot captures regime-specific rewiring through weaker interactions, notably with Property Factor 1 (17.32) and Property Factor 4 (14.28) across secondary structural modes. These couplings demonstrate that while `res_homo` remains fundamentally decoupled from the broader geometric-vibrational manifold globally, specific topological configurations such as high ring density or sp^3-heavy architectures force a weak integration between frontier electronic levels and the macroscopic molecular skeleton.

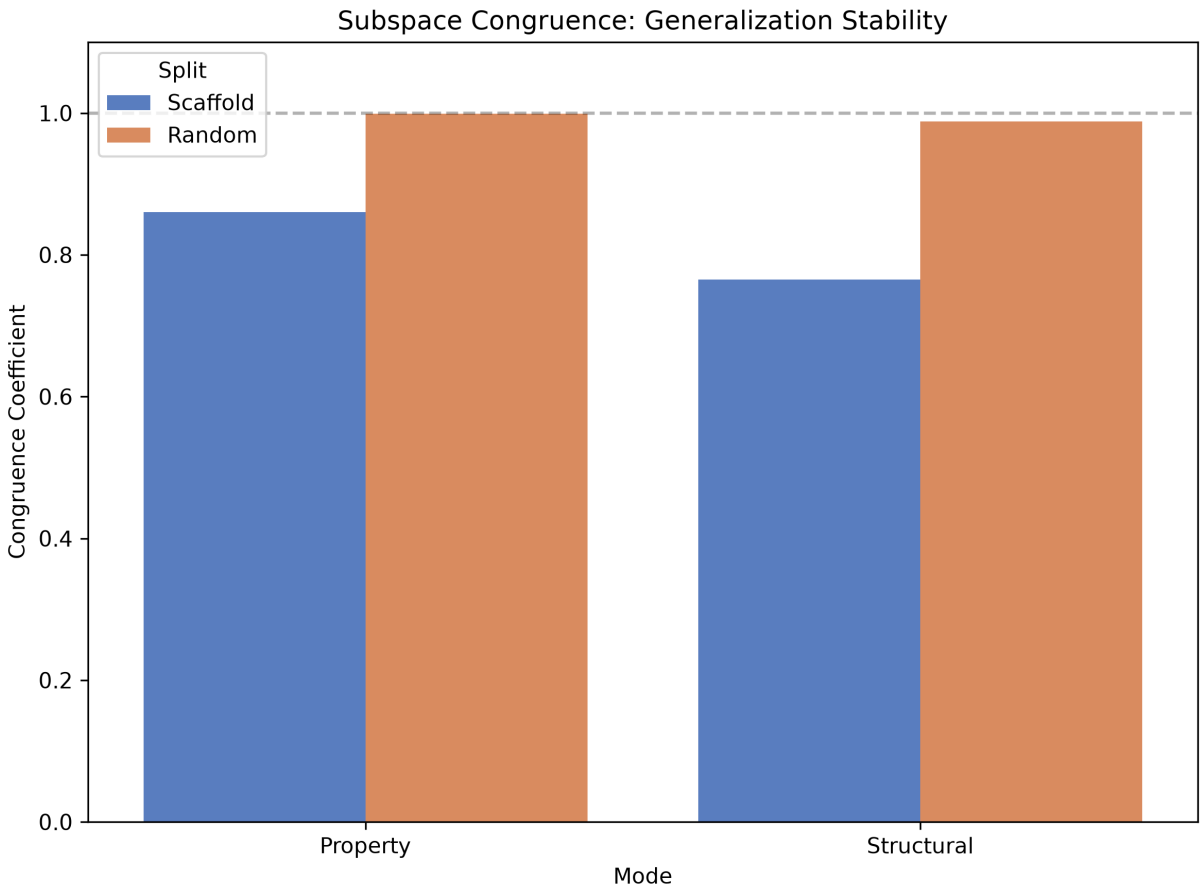

Figure 6: Subspace congruence coefficients for the latent property and structural modes evaluated via random and Murcko scaffold splits. The random split demonstrates near-perfect congruence, validating the baseline stability of the tensor decomposition. The scaffold split maintains high congruence for property factors (0.8606), indicating that latent groupings, such as the decoupling of the HOMO energy (`res_homo`), represent fundamental chemical rules that generalize beyond specific molecular families. The comparatively lower congruence for structural factors (0.7650) under the scaffold split suggests that the exact sensitivity of these property couplings to topological features depends more heavily on the specific molecular scaffolds present in the sample.

Property Factor 3 (the HOMO-dominant factor) exhibits its strongest coupling with Structural Factor 1 (interaction strength: 415.93), indicating a stable baseline electronic response across most architectures. However, as molecules transition into high-density ring regimes (e.g., bins `3_1` and `3_2`), a subtle rewiring occurs. In these highly constrained configurations, Property Factor 3 begins to interact with Property Factor 1 (spatial extent) and Property Factor 4 (secondary geometric response) with interaction strengths of 17.32 and 14.28, respectively. This demonstrates that while *res_homo* is globally decoupled, extreme ring strain and specific sp^3-heavy architectures force a weak integration between the frontier electronic levels and the macroscopic molecular skeleton.

3.4 Stability and generalizability

To verify that the observed latent structures and the regime-specific rewiring of the HOMO energy are fundamental chemical rules rather than dataset artifacts, we evaluated the subspace congruence of the factors using a Murcko scaffold split (Figure 6).

As shown in Table 2.4, the random split yielded near-perfect congruence, validating the baseline mathematical stability of the tensor decomposition.

Under the scaffold split, the property factors maintained a high congruence of 0.8606. This confirms that the latent groupings, particularly the decoupling of the HOMO energy, generalize robustly to unseen molecular families. The structural factors exhibited a comparatively lower congruence (0.7650) under the scaffold split, suggesting that while the existence of these property groupings is invariant, the exact sensitivity of the couplings to specific topological features depends more heavily on the specific molecular scaffolds present in the sample.

4 Conclusions

The joint organizational structure of molecular properties in the QM9 dataset is characterized by a fundamental, high-dimensional decoupling of frontier electronic properties from the macroscopic spatial and vibrational manifolds. By isolating architectural variance from stoichiometric determinism and constructing a structural-complexity-stratified property dependency tensor, we mathematically formalized the unique behavior of the highest occupied molecular orbital (HOMO) energy. Tucker decomposition revealed that the property manifold is governed by four stable latent factors, with the HOMO energy requiring a dedicated latent dimension. This demonstrates that the HOMO energy refuses to integrate into the tightly bound geometric-thermodynamic core that dictates most other macroscopic molecular behaviors.

Crucially, our tensor-based framework allowed us to map the boundaries of this orthogonality. While the structural independence of the HOMO energy acts as a globally invariant chemical rule across most topological regimes, core tensor interaction analysis identified specific conditions under which this decoupling marginally breaks down. In highly constrained cyclic systems—characterized by extreme ring strain and high fractions of sp^3 hybridized carbons—a subtle rewiring occurs, forcing weak couplings between the electronic levels and the molecular skeleton.

The robustness of these latent structures was confirmed through Murcko scaffold-split validation, which demonstrated that the isolated nature of the HOMO energy generalizes to unseen chemical architectures. The reduction in structural factor congruence under the scaffold split, compared to the near-perfect property factor congruence, further highlights an important chemical nuance: while the existence of this electronic decoupling is universal, its exact sensitivity to topological constraints remains dependent on the specific molecular scaffolds present.

Ultimately, this structural independence suggests that predictive models relying on global geometric or thermodynamic descriptors to infer electronic properties are inherently limited. Such models are likely missing the localized, topological features that uniquely govern the HOMO level. By moving beyond pairwise correlations and mapping higher-order property dependencies, this work provides a rigorous mathematical foundation for understanding how macroscopic spatial confinement fundamentally rewires—or fails to rewire—a molecule's highest occupied orbital into the broader physical state space.