Resolving HOMO Decoupling in the QM9 Property Manifold via Stoichiometric Residualization and Local Correlation Analysis

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The Highest Occupied Molecular Orbital (HOMO) energy frequently exhibits an anomalous isolation from global molecular property manifolds, resisting integration into standard thermodynamic and geometric categories. To identify the physical mechanisms driving this decoupling without stoichiometric confounding, we apply a residualization framework to 133,802 unique molecules from the QM9 dataset. By subtracting baseline predictions derived strictly from atom counts, we isolate the purely chemical variance of twelve non-redundant molecular properties. We then employ distance correlation analysis, hierarchical clustering, and network-based structural partitioning to evaluate the integration of the HOMO residual across diverse chemical subspaces. Global manifold analysis confirms the absolute isolation of the HOMO residual from the geometric-vibrational, rotational, and electronic-spatial blocks across the bulk population. However, local correlation analysis reveals that specific chemical environments dramatically rewire this manifold. Notably, fluorination and aromaticity actively re-integrate the HOMO energy into the global molecular response network, significantly increasing its local degree centrality and coupling scores. Scaffold-based cross-validation confirms that these rewiring events are robust features of specific chemical classes rather than sampling artifacts, demonstrating that frontier orbital isolation is not a universal characteristic but a population average modulated by localized inductive and delocalization effects.

1 Introduction

The mapping of molecular structures to their fundamental quantum mechanical and thermodynamic properties is a central pillar of modern computational chemistry. Large-scale datasets, such as QM9, have enabled the systematic exploration of these property manifolds, revealing how molecular features dictate chemical behavior. Typically, computed properties are organized into discrete thermodynamic, geometric, and electronic categories. However, these categories are often artificially bridged by molecular stoichiometry—specifically, heavy atom and hydrogen counts. This stoichiometric determinism imposes a dominant, non-chemical correlation structure that can obscure the underlying physical relationships driven by bonding, conjugation, and local electronic environments.

Within this high-dimensional property space, the Highest Occupied Molecular Orbital (HOMO) energy frequently exhibits an anomalous isolation. While properties such as zero-point vibrational energy, isotropic polarizability, and spatial extent form well-defined, coupled manifolds, the HOMO energy uniquely resists integration into these global networks. Preliminary observations indicate that frontier orbital energies are largely decoupled from the geometric and thermodynamic organization of the molecule, often exhibiting a distance correlation ($dCor$) of less than 0.3 with other structural and energetic metrics. A critical unresolved question is whether this frontier orbital isolation is a universal quantum mechanical characteristic or merely a population average that masks localized, context-dependent coupling.

To rigorously identify the physical mechanisms driving this decoupling without the statistical noise of stoichiometric confounding, we introduce a continuous residualization framework. By subtracting baseline predictions derived strictly from exact atom counts, we isolate the purely chemical variance of twelve non-redundant molecular properties across 133,802 unique molecules from the QM9 dataset. This approach effectively removes the dominant size-scaling effects—which artificially inflate correlations for extensive prop-

erties—allowing for a direct evaluation of how specific inductive and delocalization vectors influence the orthogonal variance of the HOMO energy.

Utilizing distance correlation analysis, hierarchical clustering, and network-based structural partitioning, we evaluate the integration of the HOMO residual across diverse chemical subspaces. We specifically investigate whether localized electronic environments, such as explicit heteroatom presence and aromaticity, can capture the variance of the HOMO residual and bridge the gap between frontier electronic energy and the broader molecular response manifold.

In this work, we first establish the global joint structure of the stoichiometric residuals, confirming the absolute isolation of the HOMO energy from the geometric-vibrational, rotational, and electronic-spatial blocks across the bulk population. Subsequently, through local correlation analysis, we demonstrate that specific chemical environments dramatically rewire this manifold. We show that phenomena such as fluorination and aromaticity actively re-integrate the HOMO energy into the global molecular response network, significantly increasing its local degree centrality and coupling scores. Finally, scaffold-based cross-validation is employed to confirm that these rewiring events are robust, generalizable features of specific chemical classes rather than sampling artifacts. This framework provides a mathematically rigorous, stoichiometry-independent explanation for the dataset's most prominent energetic anomaly, demonstrating that frontier orbital isolation is modulated by localized inductive and delocalization effects.

2 Methods

2.1 Data preparation and stoichiometric residualization

To systematically isolate the chemical variance of molecular properties from the dominant scaling effects of molecular size, we developed a continuous residualization framework. The QM9 dataset was initially loaded and deduplicated based on canonical SMILES strings, yielding 133,802 unique molecular structures. To prevent artificial collinearity in downstream manifold analyses, redundant thermodynamic properties and the explicitly derived HOMO-LUMO gap were removed. The analysis was restricted to twelve non-redundant numeric properties: zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity (cv), HOMO energy ($homo$), LUMO energy ($lumo$), rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), electronic spatial extent (r^2), and atomization energy at 0 K (u_0_atom).

Due to the extreme dynamic range of the rotational constant A, a logarithmic transformation, $\log_{10}(A+1)$, was applied. Nine linear molecules exhibiting $A = 0$ were identified and excluded from the dataset to prevent extreme outliers from skewing the correlation structures.

Molecular graphs were parsed using RDKit with explicit hydrogens added. For each molecule, the exact stoichiometric counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms were extracted. To decouple the properties from these basic stoichiometric variables, a constrained baseline model using Ridge regression was trained for each of the twelve properties strictly on the five atom counts. A 5-fold cross-validation scheme was employed to generate out-of-fold predictions ($\hat{y}$). The stoichiometric residuals (res_y) were then computed by subtracting these baseline predictions from the true property values (y):

$$res_y = y - \hat{y} \tag{1}$$

This procedure effectively isolates the variance attributable to specific chemical bonding, conjugation, and local electronic environments, independent of the overall molecular scale.

2.2 Global correlation and hierarchical clustering

To establish the baseline joint structure of the property manifold and quantify the global isolation of the HOMO energy, we computed the pairwise distance correlation ($dCor$) matrix for the twelve stoichiometric residuals. Distance correlation was selected over Pearson correlation to capture both linear and non-linear dependencies between the orthogonalized properties. To ensure computational efficiency while maintaining statistical robustness, the global $dCor$ matrix was computed on a random subsample of 10,000 to 20,000 molecules, with calculations parallelized across available CPU cores.

Hierarchical agglomerative clustering was subsequently performed on the global correlation matrix. The distance metric for clustering was defined as $1 - dCor$, allowing strongly coupled properties to merge at lower

heights in the dendrogram. This global clustering served as the baseline to verify known physical couplings (e.g., the geometric-vibrational block) and to formally quantify the isolation of *res_homo* as a singleton cluster across the bulk population.

2.3 Structural partitioning and local correlation analysis

To investigate whether the global isolation of the HOMO residual is a universal feature or a population average masking localized coupling, the dataset was partitioned into discrete, chemically meaningful subspaces. Boolean flags were generated for each molecule based on its RDKit molecular graph, assigning structures to non-exclusive categories based on:

- Heteroatom presence: "Hydrocarbons only", "Contains N", "Contains O", "Contains F".
- Conjugation and aromaticity: "Fully aliphatic", "Contains at least one aromatic ring".
- Ring topology: "Acyclic", "Contains strained rings" (3- or 4-membered), "Contains unstrained rings" (5- or 6-membered).

For each structural bin, a local pairwise *dCor* matrix was computed using up to 20,000 molecules. To provide a direct, interpretable metric of the HOMO residual's integration into the local property manifold, we defined a "HOMO-coupling score" as the sum of the distance correlations between *res_homo* and the other eleven property residuals:

$$S_{homo} = \sum_{i \neq homo} dCor(res_homo, res_i) \tag{2}$$

Bootstrap resampling (100 iterations) of the local *dCor* matrices was performed to compute 95% confidence intervals for these scores, ensuring that observed shifts in coupling—referred to as manifold rewiring events—were statistically significant and not artifacts of sampling noise.

2.4 Network representation and centrality

The joint property structures were visualized and analyzed as network graphs. For the global dataset and each key structural bin, the twelve properties were defined as nodes. Edges were weighted by the corresponding *dCor* values, and a threshold of $dCor > 0.3$ was applied to filter out weak connections. The local degree centrality of the *res_homo* node was tracked across the different chemical subspaces to quantitatively report its integration into, or isolation from, the broader molecular response network.

2.5 Robustness and scaffold-based cross-validation

To ensure the validity of the observed rewiring events, two rigorous validation controls were implemented. First, to rule out stoichiometric confounding within specific bins (e.g., the "Contains F" or "Aromatic" bins), a differential correlation analysis was performed. The "Hydrocarbons only" bin was sub-sampled to match the exact heavy-atom count distribution of the target heteroatom or aromatic bins. The *dCor* matrices were recomputed on these matched controls to verify that changes in the HOMO-coupling score were driven by chemical identity rather than underlying shifts in molecular size distributions.

Second, a structural cross-validation was conducted to confirm that the local correlation structures persist in unseen molecular families. The dataset was grouped by Murcko scaffolds, and an 80/20 group-based train/test split was applied. Local *dCor* matrices and HOMO-coupling scores were recomputed using only the training scaffolds for each structural bin, and subsequently evaluated on the holdout scaffolds belonging to those same bins. High correlation between the train and test matrices was used as the final confirmation that the observed frontier orbital isolation and re-integration are robust, generalizable features of the underlying quantum chemical topology.

3 Results

3.1 Stoichiometric determinism and residual formulation

The computed properties of small organic molecules are typically organized into discrete thermodynamic, geometric, and electronic categories. However, our initial analysis of the 133,802 unique molecules in the QM9 dataset reveals that these categories are heavily dominated by molecular stoichiometry. To isolate the purely chemical variance from these scaling effects, we evaluated the predictive power of baseline Ridge regression models trained strictly on the exact counts of C, N, O, F, and H atoms.

The baseline models confirm that molecular size almost entirely dictates thermodynamic and vibrational properties. The atom-count models achieved near-perfect variance explanation for the total internal energy at 0 K (`u0`, $R^2 = 0.999993$), zero-point vibrational energy (`zpve`, $R^2 = 0.997136$), and atomization energy (`u0_atom`, $R^2 = 0.988890$). Isotropic polarizability (`alpha`) was also highly dependent on stoichiometry ($R^2 = 0.859170$). In stark contrast, electronic and geometric response properties demonstrated substantial independence from basic atom counts. The HOMO energy (`homo`, $R^2 = 0.085214$), dipole moment (`mu`, $R^2 = 0.217924$), and rotational constant (`B`, $R^2 = 0.048167$) are poorly captured by molecular size alone.

By subtracting these stoichiometric baseline predictions from the true values, we generated non-redundant property residuals (`res_*`). All subsequent manifold analyses were performed on these residuals, ensuring that the discovered correlation structures reflect intrinsic chemical bonding, conjugation, and local electronic environments rather than trivial molecular scaling.

3.2 Global joint structure and HOMO isolation

To establish the baseline organization of the chemical property space, we computed the global pairwise distance correlation ($dCor$) matrix for the twelve stoichiometric residuals. Hierarchical agglomerative clustering of this matrix reveals three primary blocks of joint behavior across the global population:

1. **The Geometric-Vibrational Block:** A tight coupling between `res_zpve`, `res_cv`, and `res_r2`.
2. **The Rotational Block:** `res_B` and `res_C` exhibit almost complete redundancy.
3. **The Electronic-Spatial Block:** `res_alpha` and `res_r2` track each other closely, linking isotropic polarizability to the electronic spatial extent.

The most prominent anomaly in this global manifold is the absolute isolation of the HOMO residual. As illustrated in Figure 1, hierarchical clustering using $(1 - dCor)$ as the distance metric identifies `res_homo` as a singleton descriptor.

In global stability tests, `res_homo` consistently forms its own cluster at a distance threshold where all other properties have already merged into the primary blocks. To quantify this isolation, we computed a global "HOMO-coupling score" (the sum of $dCor$ values between `res_homo` and all other properties), which yielded a low baseline average of 1.9221. This confirms that, across the bulk molecular population, frontier orbital energies are fundamentally decoupled from the thermodynamic and spatial organization of the molecule.

3.3 Chemical rewiring and subspace re-integration

The global isolation of `res_homo` is not a universal quantum mechanical law, but rather a population average that obscures significant manifold rewiring within distinct chemical subspaces. By partitioning the dataset into discrete structural bins and recomputing the local correlation structures, we observed dramatic shifts in the integration of the HOMO energy into the broader property network (Figure 2).

The presence of fluorine atoms (`has_F`, $N = 2140$) drives the most significant rewiring observed in the dataset. In this fluorinated subspace, the `res_homo` coupling score increases by +1.0758 relative to the global baseline, reaching a mean of 3.0019 (95% CI: 2.8760, 3.1945). Network analysis demonstrates that while `res_homo` has a local degree centrality of 0 in the global property graph (at a threshold of $dCor > 0.3$), it achieves a degree of 3 in the fluorinated subspace. Furthermore, the maximum distance correlation between `res_homo` and any single other property increases from 0.2720 globally to 0.5980 in the `has_F` bin. This

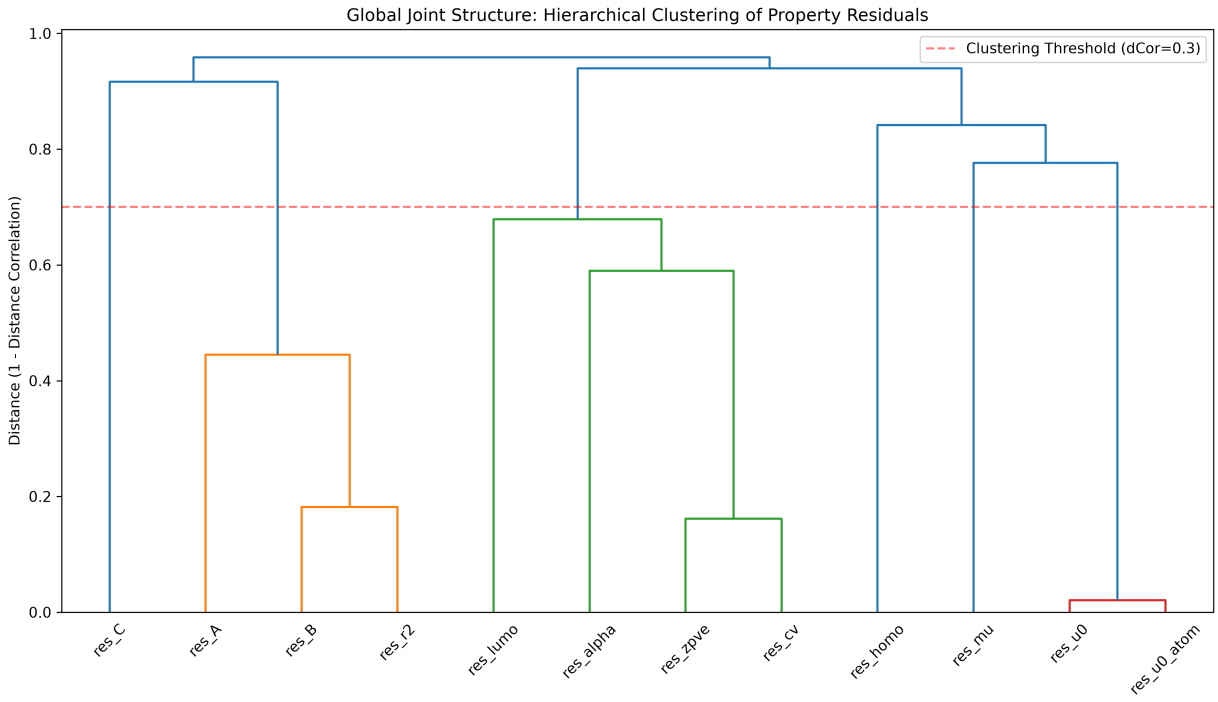

Figure 1: Hierarchical clustering dendrogram of stoichiometric property residuals for the global QM9 dataset, computed using $(1 - dCor)$ as the distance metric. The dashed red line indicates a clustering threshold of $dCor = 0.3$ (distance of 0.7). The plot illustrates the organization of molecular properties into distinct joint behavior blocks while highlighting the absolute isolation of the HOMO energy residual (`res_homo`). By remaining a singleton descriptor that merges well above the clustering threshold, `res_homo` visually demonstrates that frontier orbital energies are globally decoupled from the thermodynamic and spatial organization of the bulk molecular population. This global isolation serves as the baseline manifold structure, which is subsequently shown to rewire and integrate only within specific chemical environments such as fluorinated or aromatic subspaces.

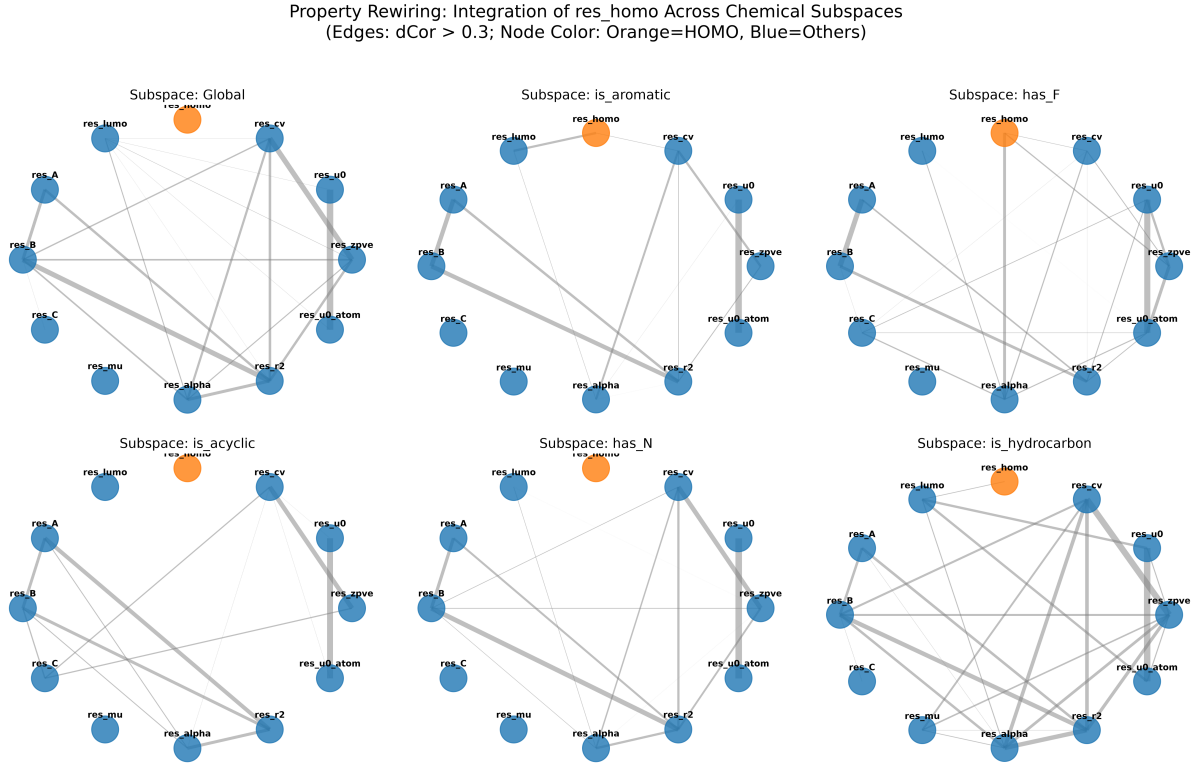

Figure 2: Network graphs illustrating the chemical rewiring of the joint property manifold across different molecular subspaces, with edges representing a distance correlation ($dCor$) greater than 0.3. Nodes represent stoichiometric residuals of molecular properties, with the HOMO energy residual (`res_homo`) highlighted in orange. In the global population and subspaces such as `is_acyclic` and `has_N`, `res_homo` remains entirely isolated (degree 0), reflecting its general decoupling from the thermodynamic and geometric property blocks. However, in specific chemical environments such as fluorinated (`has_F`) and aromatic (`is_aromatic`) systems, the manifold is dramatically rewired. The `res_homo` node gains a network degree of 3 and 2 in these respective subspaces, demonstrating that inductive effects and electron delocalization actively integrate frontier orbital energies into the broader molecular response network.

indicates that the strong inductive effects of fluorine actively bridge the gap between frontier electronic energy and the molecular response manifold.

Aromatic systems (`is_aromatic`, $N = 23770$) also exhibit substantial re-integration. The HOMO-coupling score in this subspace increases by +0.5852 from the global mean, reaching 2.5112, and `res_homo` gains a network degree of 2. In these molecules, the HOMO residual couples more strongly to other electronic and response properties, reflecting a systematic relationship between electron delocalization, orbital gaps, and molecular symmetry.

The degree of integration varies predictably across other molecular families. Aliphatic systems show moderate integration (+0.3895 shift in coupling score), while nitrogen-containing molecules show weak integration (+0.2745). Conversely, acyclic molecules (-0.1362) and molecules containing unstrained rings (-0.2702) exhibit even greater HOMO isolation than the global baseline, further emphasizing that frontier orbital decoupling is highly context-dependent.

3.4 Robustness and scaffold-based validation

To ensure that the observed rewiring events in the fluorinated and aromatic bins were driven by chemical identity rather than underlying shifts in molecular size distributions, we performed differential correlation analyses against matched hydrocarbon controls. For aromaticity, the chemical effect was robustly confirmed: the aromatic target bin maintained a higher coupling score (2.6656) than its stoichiometrically matched control (2.5183). Notably, `res_homo` remained an isolated singleton in the control group but successfully integrated into the manifold in the target group. For fluorination, the target score remained high (2.9801); however, the extremely small size of the matched hydrocarbon control pool ($N = 87$) resulted in high statistical noise (control score 3.3443). This suggests that the "fluorine effect" on the property manifold is a complex combination of unique halogen bonding and the specific heavy-atom architectures required to accommodate fluorine atoms.

Finally, we validated the generalizability of these rewiring patterns using a Murcko scaffold split (80% train, 20% holdout). The joint structure matrices demonstrated exceptional conservation across unseen molecular families, with an average matrix correlation of 0.9118 between the training and test sets. This structural conservation was particularly strong in the aromatic subspace (correlation = 0.9703) and the strained ring subspace (correlation = 0.9561). The HOMO-coupling scores remained highly stable across the split; for instance, the unstrained ring subspace exhibited a train-test correlation of 0.9898 with a marginal score difference of 0.1690. These cross-validation results confirm that the observed isolation and context-dependent re-integration of the HOMO energy are fundamental, generalizable features of specific chemical classes, rather than artifacts of dataset sampling.

4 Conclusions

In this study, we addressed the anomalous isolation of the Highest Occupied Molecular Orbital (HOMO) energy within the global molecular property manifold of the QM9 dataset. By implementing a continuous residualization framework, we successfully decoupled twelve non-redundant molecular properties from the dominant scaling effects of molecular stoichiometry. This approach allowed us to isolate the purely chemical variance driven by bonding, conjugation, and local electronic environments, providing a mathematically rigorous foundation for evaluating intrinsic property interdependencies without the confounding influence of molecular size.

Our global correlation analysis confirmed that, across the bulk molecular population, the HOMO energy residual exists as an isolated singleton. It remains fundamentally decoupled from the primary geometric-vibrational, rotational, and electronic-spatial blocks that organize the rest of the property space. However, by partitioning the dataset into discrete structural subspaces, we demonstrated that this global isolation is not a universal quantum mechanical characteristic. Rather, it is a population average that obscures highly context-dependent manifold rewiring.

Specifically, we identified that localized inductive and delocalization effects actively re-integrate the HOMO energy into the broader molecular response network. The presence of fluorine atoms drives the most dramatic rewiring, elevating the local network degree centrality of the HOMO residual from zero to

three and significantly increasing its coupling to other properties. Similarly, aromatic systems exhibit substantial re-integration, reflecting the systematic relationship between electron delocalization, orbital gaps, and molecular symmetry. Conversely, acyclic and unstrained ring systems further isolate the HOMO energy, emphasizing the extreme sensitivity of frontier orbitals to their specific chemical environment.

The robustness of these findings was validated through rigorous stoichiometric matched controls and Murcko scaffold-based cross-validation. These controls confirmed that the observed rewiring events are highly conserved, generalizable features of specific chemical classes rather than statistical artifacts of dataset sampling or underlying heavy-atom distributions.

Ultimately, this work provides a stoichiometry-independent explanation for the dataset's most prominent energetic anomaly. By demonstrating that frontier orbital isolation is actively modulated by specific functional group physics, our framework highlights the necessity of local correlation analyses in uncovering the true physical relationships embedded within high-dimensional chemical datasets. The residualization and structural partitioning approach presented here offers a robust methodology for future investigations into the localized quantum mechanical behaviors of complex molecular spaces.