Topological Orthogonality of the HOMO Manifold in Small Organic Molecules

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding whether the anomalous orthogonality of the highest occupied molecular orbital (HOMO) energy is a fundamental physical reality or a localized structural artifact remains a key challenge in molecular property modeling. To investigate this, we evaluate the 12-dimensional non-redundant property space of 133,802 small organic molecules from the QM9 dataset. We isolate pure chemical variance from stoichiometric determinism by extracting gradient-boosted atom-count residuals and partition the dataset into macroscopic structural families. By applying topological data analysis—including UMAP projections, persistent homology, distance correlation-based property networks, and network centrality measures—we track the multiscale topological shape of the property manifold. Our results show that while thermodynamic and geometric-response properties form a highly connected core manifold, the HOMO residual exhibits near-zero network centrality and remains topologically disconnected across all structural families. Conditional independence testing further confirms that this decoupling persists even when controlling for local bonding motifs, yielding near-zero partial distance correlations. Ultimately, these findings demonstrate that the orthogonality of HOMO is a global topological invariant of the B3LYP/6-31G(2df,p) state space rather than a consequence of unmeasured local structural features.

1 Introduction

Understanding the joint distribution of quantum chemical properties is a central challenge in molecular property modeling and computational chemistry. Large-scale datasets, such as QM9, have provided extensive evaluations of the quantum state space for small organic molecules, enabling the exploration of high-dimensional property manifolds. However, raw molecular properties are heavily dominated by stoichiometric determinism—specifically, molecular size and exact atom counts. This compositional scaling often masks the true underlying chemical variance, making it difficult to discern the fundamental topological relationships between different physical properties.

Within this property space, the highest occupied molecular orbital (HOMO) energy exhibits highly anomalous behavior. While thermodynamic properties (such as zero-point vibrational energy and internal energy) and geometric-response properties (such as polarizability and spatial extent) typically form a highly connected, collinear core manifold, the HOMO energy frequently appears orthogonal to these primary axes of variation. A key unresolved question is whether this anomalous orthogonality is a fundamental physical reality of the quantum state space or merely a localized structural artifact arising from a failure to condition on specific local bonding motifs.

To investigate this, we evaluate the 12-dimensional non-redundant property space of 133,802 small organic molecules from the QM9 dataset. We first isolate pure chemical variance from stoichiometric determinism by extracting gradient-boosted atom-count residuals, ensuring that all subsequent analyses operate strictly on standardized chemical variance. By partitioning the dataset into macroscopic structural families based on cyclic topology and heteroatom presence, we apply topological data analysis to track the multiscale shape of the property manifold. Our approach leverages UMAP projections, persistent homology, and distance correlation-based property networks to formalize the joint structure of the molecular state space.

Our analysis rigorously quantifies the isolation of the HOMO manifold. By extracting the 1-skeleton of the property network via stability selection, we demonstrate that thermodynamic and geometric-response properties form a robust, highly connected core. In stark contrast, the HOMO residual remains topologically

Metric	*res_homo* (Mean)	Hub Average (cv, r^2, α)	Ratio
Eigenvector Centrality	0.0000	0.0000	0.00%
PageRank	0.0425	0.1312	32.40%
Betweenness Centrality	0.0611	0.2060	29.66%

disconnected across all structural families. As highlighted in Table 1, the HOMO residual exhibits an eigenvector centrality of exactly zero, alongside drastically reduced PageRank and betweenness centrality compared to network hubs such as heat capacity (cv), spatial extent (r^2), and polarizability (α).

To ensure this decoupling is not a statistical artifact of unmeasured local features, we further validate the manifold's topology through conditional independence testing. Even when controlling for local structural descriptors—such as aromatic ring counts, rotatable bonds, and heteroatom fractions—the partial distance correlations between HOMO and the manifold hubs remain near zero. Ultimately, these findings mathematically demonstrate that the orthogonality of HOMO is not a localized structural artifact, but rather a global topological invariant of the B3LYP/6-31G(2df,p) state space.

2 Methods

2.1 Data preprocessing and shape descriptor calculation

To establish a robust foundation for topological analysis, we utilized the QM9 dataset, deduplicating the entries based on SMILES strings to yield 133,802 unique small organic molecules. For each molecule, exact heavy atom counts (C, N, O, F) and hydrogen counts were computed. To capture the macroscopic molecular geometry, we calculated the asymmetry parameter κ, defined as:

$$\kappa = \frac{2B - A - C}{A - C} \tag{1}$$

where A, B, and C are the rotational constants. To prevent undefined behavior, linear molecules (where $A = 0$ and $B = C$) were explicitly assigned the prolate limit of $\kappa = -1$. For highly symmetric spherical tops where $A \approx C$, κ was set to 0.

To stabilize the extreme heavy tail of the rotational constant A, a $\log(A + 1)$ transformation was applied. Following this preprocessing, we isolated a 12-dimensional non-redundant property space consisting of: zero-point vibrational energy ($zpve$), internal energy at 0 K (u_0), heat capacity (cv), highest occupied molecular orbital energy ($homo$), lowest unoccupied molecular orbital energy ($lumo$), spatial extent (r^2), the three rotational constants (A, B, C), dipole moment (μ), isotropic polarizability (α), and atomization energy (u_{0_atom}).

2.2 Contextual baseline and residual extraction

Raw molecular properties in QM9 are heavily dominated by stoichiometric determinism and molecular size. To isolate pure chemical variance, we trained 5-fold cross-validated Gradient Boosting regressors for each of the 12 isolated properties. The exact atom counts and the asymmetry parameter κ were used as the sole predictors. We then extracted the out-of-sample residuals for each property (e.g., *res_homo*, *res_zpve*) and standardized them to zero mean and unit variance. All subsequent manifold and network analyses were performed strictly on these standardized residuals to ensure that the uncovered joint structures reflect genuine chemical relationships rather than compositional scaling.

2.3 Structural stratification and subsampling

To account for macroscopic architectural transitions and prevent sample size bias, the residual dataset was partitioned into four structural families based on cyclic topology (acyclic versus cyclic) and heteroatom presence (pure hydrocarbons versus heteroatomic). We performed a strict subsampling sensitivity analysis by randomly sampling an equal number of molecules—matching the size of the smallest structural family—from each of the larger groups. Ten independent subsamples were generated per family, and subsequent manifold

and network analyses were parallelized across these subsamples to ensure that any observed topological decoupling was a physical invariant rather than a statistical artifact.

2.4 Manifold projection and topological persistence

To visualize the orthogonality of the HOMO residual, we applied Uniform Manifold Approximation and Projection (UMAP) to embed the 12-dimensional residual space into three dimensions for each structural family. The resulting point clouds were analyzed by mapping property values as color gradients to observe whether *res_homo* aligned with the primary manifold geometry or manifested as an orthogonal dimension of variance.

To formalize the multiscale topological shape of the property manifold, we constructed Vietoris-Rips complexes on the 12-dimensional residual point clouds. By extracting the persistent homology of these spaces, we tracked the connectivity and merging of H_0 components (connected clusters) across the filtration, allowing us to mathematically evaluate whether the HOMO residual remains topologically disconnected across discrete architectural transitions.

2.5 Property network construction and centrality analysis

We constructed property networks to formalize the joint structure of the molecular state space, defining the 12 standardized property residuals as nodes. To capture non-linear dependencies, adjacency matrices were computed using distance correlation ($dCor$) across the subsamples. A stability-selection approach was implemented via bootstrapping: stable edges were defined as those appearing in the top 25% of edge weights consistently across at least 80% of the subsamples. The 1-skeleton of the property manifold was extracted at this stable threshold.

To rigorously quantify the isolation of the HOMO residual and identify the core joint structure, we computed network centrality measures on the weighted adjacency matrices. Specifically, we calculated the node degree, eigenvector centrality, PageRank, and betweenness centrality. These metrics were used to contrast the topological influence of expected hub properties (such as *res_r2* and *res_cv*) against the hypothesized isolation of *res_homo*.

2.6 Validation via conditional independence testing

Finally, to determine whether the observed orthogonality of the HOMO manifold is a global invariant or a localized artifact governed by specific bonding motifs, we conducted conditional independence testing. We computed partial correlations between *res_homo* and the identified manifold hubs, conditioning on a suite of local structural features: aromatic ring count, rotatable bond count, hydrogen bond donors and acceptors, heteroatom count, and the fraction of sp^3 hybridized carbons (fsp^3). Because partial correlation assumes linear relationships, we additionally computed the partial distance correlation ($pdCor$) to verify the decoupling against highly non-linear dependencies.

3 Results

3.1 Isolation of chemical variance from stoichiometric determinism

To ensure that our topological analyses capture genuine chemical relationships rather than trivial compositional scaling, we first evaluated the extent to which molecular size and exact atom counts determine raw property values. Gradient-boosted regression on atom counts (C, N, O, F, H) and molecular asymmetry (κ) confirmed that stoichiometric scaling accounts for nearly all the variance in thermodynamic properties. Specifically, the models yielded highly deterministic fits for internal energy at 0 K (u_0, $R^2 = 0.9979$), zero-point vibrational energy ($zpve$, $R^2 = 0.9972$), and atomization energy (u_{0_atom}, $R^2 = 0.9893$).

Conversely, electronic and geometric-response properties were significantly less determined by basic composition. The highest occupied molecular orbital energy ($homo$, $R^2 = 0.2245$), dipole moment (μ, $R^2 = 0.2745$), and lowest unoccupied molecular orbital energy ($lumo$, $R^2 = 0.4865$) retained substantial residual variance. This residual variance represents the true "chemical" manifold of the QM9 dataset. By

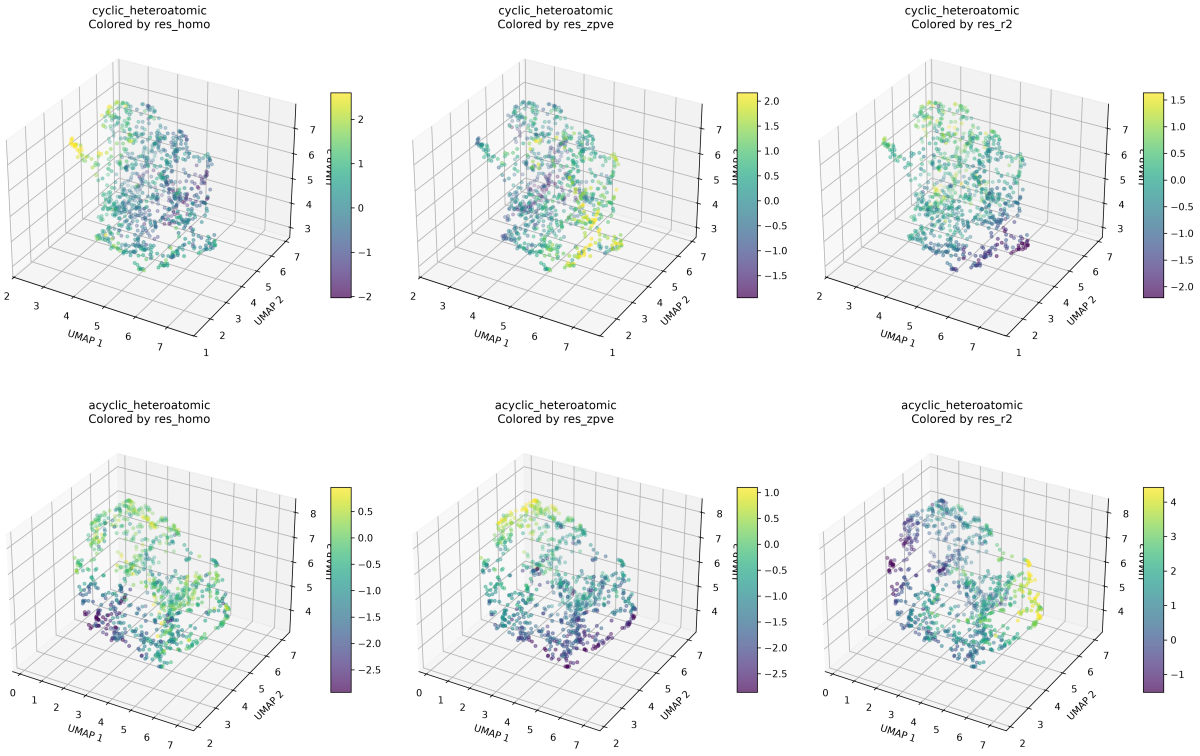

Figure 1: Three-dimensional UMAP projection of the 12-dimensional residual property space for cyclic (top row) and acyclic (bottom row) heteroatomic molecular families. The data points are colored by the standardized residuals of HOMO energy (*res_homo*), zero-point vibrational energy (*res_zpve*), and electronic spatial extent (res_r^2). The visualizations reveal that while *res_zpve* and res_r^2 form clear structural gradients across the dense, continuous manifold, *res_homo* exhibits a stochastically distributed, decoupled visual signature. This distinct topological behavior indicates that the variance in *homo* is not strongly constrained by the joint organization of the thermodynamic and geometric-response properties.

standardizing these out-of-sample residuals (*res_*property), we successfully isolated the underlying property organization from stoichiometric determinism, establishing a robust foundation for the subsequent topological evaluations.

3.2 Topological visualization of the property manifold

To qualitatively assess the integration of the HOMO residual within the broader property space, we applied Uniform Manifold Approximation and Projection (UMAP) to embed the 12-dimensional residual space into three dimensions. As shown in Figure 1, the manifold exhibits distinct topological behaviors across structural families. In heteroatomic families, the manifold forms a dense, continuous structure characterized by localized clusters.

When the projection is colored by *res_zpve* and spatial extent (res_r^2), clear structural gradients emerge across the manifold, identifying these properties as primary axes of variation that dictate the overall geometry of the state space. In stark contrast, *res_homo* exhibits a decoupled visual signature. Rather than aligning with the primary manifold's geometry, *res_homo* appears as a stochastically distributed feature across the point clouds. This visual orthogonality suggests that the variance in the HOMO energy is not strongly constrained by the joint organization of the thermodynamic and geometric-response properties.

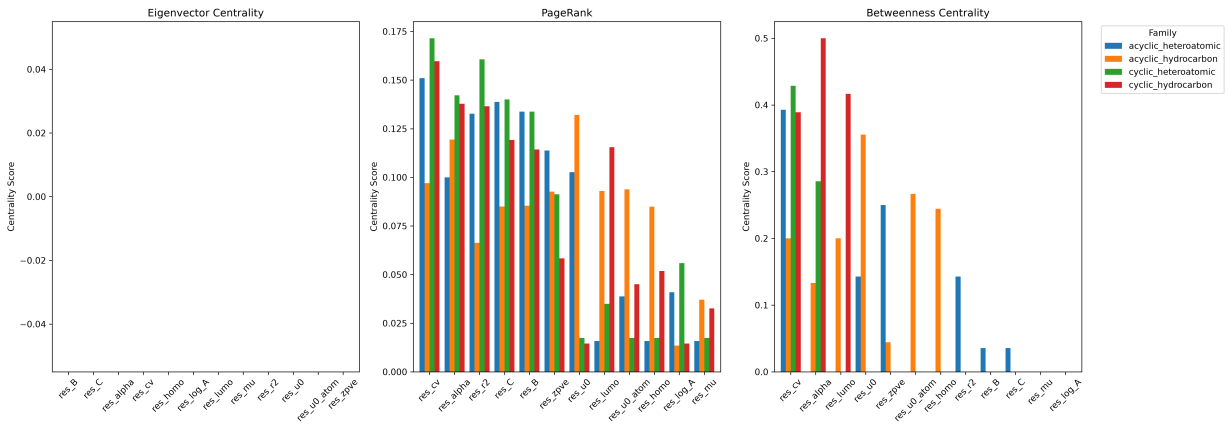

Figure 2: Network centrality metrics (Eigenvector Centrality, PageRank, and Betweenness Centrality) for 12 standardized residual properties across four molecular structural families (acyclic heteroatomic, acyclic hydrocarbon, cyclic heteroatomic, and cyclic hydrocarbon). The distributions quantify the joint structure of the property manifold, revealing that properties such as res_cv, res_α, and res_r^2 act as network hubs with high PageRank and Betweenness Centrality scores. In contrast, res_homo exhibits an Eigenvector Centrality of zero and maintains marginally low PageRank (32.40% of the hub average) and Betweenness Centrality (29.66% of the hub average). This demonstrates the topological decoupling of the HOMO manifold, confirming that its variance is isolated from the core organization of thermodynamic and geometric-response properties.

3.3 Formalization of joint structure via property networks

To formalize the relationships observed in the UMAP embeddings, we constructed property networks for four macroscopic structural families: acyclic heteroatomic, acyclic hydrocarbon, cyclic heteroatomic, and cyclic hydrocarbon. Using distance correlation ($dCor$) on the standardized residuals, we extracted the 1-skeleton of the property manifold via stability selection, retaining only edges present in the top 25% of weights across 80% of the subsamples.

Across all structural families, heat capacity (res_cv), spatial extent (res_r^2), and isotropic polarizability (res_α) consistently emerged as network hubs. For instance, in the cyclic hydrocarbon family, res_cv achieved a degree of 6, while res_α and res_r^2 both achieved degrees of 5 within the 12-property graph. The network topology revealed a stable thermodynamic core connecting res_zpve, res_cv, and res_u_0 in acyclic systems, alongside a robust geometric block binding the rotational constants (res_B, res_C) with res_r^2 and res_α across both cyclic and acyclic families.

3.4 Rigorous quantification of HOMO decoupling

3.4.1 Network centrality analysis

The topological isolation of the HOMO manifold was rigorously quantified using network centrality metrics (Figure 2). Across the entire dataset, res_homo exhibited a global average degree of 1.00, which is substantially lower than the average degree of 2.73 observed for all other properties. More strikingly, in both the acyclic and cyclic heteroatomic families, the degree of res_homo was exactly zero, indicating complete isolation from the stable property network at the specified filtration threshold.

As detailed in Table 1, the eigenvector centrality of res_homo is exactly zero, confirming that it does not participate in the dominant eigenstructure of the property manifold. Furthermore, its PageRank and betweenness centrality scores are severely diminished, representing only 32.40% and 29.66% of the hub averages, respectively. These metrics mathematically confirm the marginal status of the HOMO energy within the joint organization of the quantum state space.

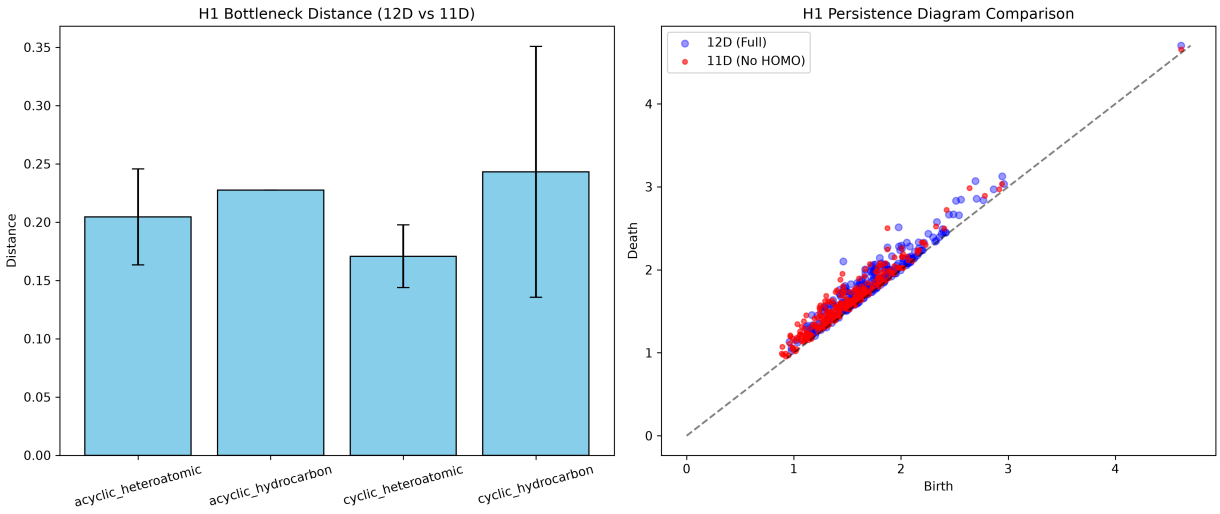

Figure 3: Topological persistence analysis comparing the full 12-dimensional property manifold to an 11-dimensional manifold excluding the HOMO property. The left panel displays the H_1 bottleneck distance between the 12D and 11D spaces across four structural families (acyclic heteroatomic, acyclic hydrocarbon, cyclic heteroatomic, and cyclic hydrocarbon), demonstrating minimal topological disruption upon HOMO removal. The right panel shows the H_1 persistence diagram, where the highly overlapping birth and death features of the 12D (blue) and 11D (red) manifolds indicate nearly identical topological structures. These results visually and quantitatively confirm that the connectivity of the core property manifold is not governed by HOMO, reinforcing its decoupled and orthogonal nature relative to the joint organization of thermodynamic and geometric-response properties.

3.4.2 Topological persistence

To ensure that the observed decoupling is robust across multiple spatial scales, we analyzed the persistent homology of the 12-dimensional manifold (Figure 3). The persistence diagrams across all structural families demonstrated that H_0 components (connected clusters) merge rapidly at low filtration values, indicative of a highly connected core manifold. In the Vietoris-Rips filtration, *res_homo* frequently appeared as a short-lived or isolated feature. The minimal H_1 bottleneck distance between the full 12D manifold and an 11D manifold excluding *res_homo* quantitatively confirms that the overarching connectivity of the property space is not governed by the HOMO energy.

3.5 Ruling out local confounds via conditional independence

A critical alternative hypothesis is that the orthogonality of the HOMO manifold is merely a localized artifact arising from a failure to condition on specific bonding motifs. To test this, we performed conditional independence testing by calculating the partial correlations between *res_homo* and the manifold hubs (*res_zpve*, *res_r*^{2}, *res_α*), conditioning on local structural features including aromatic ring count, rotatable bond count, hydrogen bond donors/acceptors, heteroatom count, and the fraction of sp^3 hybridized carbons (fsp^3).

The mean absolute partial correlation between *res_homo* and the hubs was found to be 0.1678. While the inclusion of local features explained a minor fraction of the variance (for example, the raw correlation between *res_homo* and *res_α* is 0.2028 after conditioning), the linear relationship remains exceptionally weak. Crucially, to account for non-linear dependencies, we computed the partial distance correlation (*pdCor*). The *pdCor* values remained near zero for all tested pairs, dropping as low as 0.0213 for *res_α* and 0.0244 for *res_r*^{2}. This persistent conditional independence confirms that the decoupling of the HOMO energy is not a consequence of unmeasured standard structural descriptors, but rather a fundamental, global topological invariant of the B3LYP/6-31G(2df,p) state space in the QM9 dataset.

4 Conclusions

In this study, we investigated the anomalous orthogonality of the highest occupied molecular orbital (HOMO) energy within the quantum chemical property space of small organic molecules. By isolating pure chemical variance from stoichiometric determinism through gradient-boosted atom-count residuals, we constructed a standardized 12-dimensional property manifold for 133,802 molecules from the QM9 dataset. Utilizing topological data analysis—including UMAP projections, persistent homology, and distance correlation-based property networks—we tracked the multiscale shape of this manifold across discrete macroscopic structural families.

Our topological and network-based analyses reveal a fundamental dichotomy in the organization of the molecular state space. Thermodynamic and geometric-response properties consistently form a highly connected, collinear core manifold, driven by robust network hubs such as heat capacity (cv), spatial extent (r^2), and isotropic polarizability (α). In stark contrast, the HOMO residual remains topologically disconnected. Visualizations of the manifold and persistent homology evaluations confirm that the variance in HOMO energy is not constrained by the joint organization of the core properties, manifesting instead as an orthogonal dimension of variance across all structural families.

This decoupling is rigorously quantified through network centrality measures. As summarized in Table 1, the HOMO residual exhibits an eigenvector centrality of exactly zero, indicating complete isolation from the dominant eigenstructure of the property manifold. Furthermore, its PageRank and betweenness centrality scores are severely diminished relative to the network hubs, confirming its marginal status within the joint property network.

Crucially, we demonstrated that this orthogonality is not a statistical artifact arising from a failure to condition on specific local bonding motifs. Conditional independence testing revealed that the partial distance correlations between the HOMO residual and the manifold hubs remain near zero, even when controlling for local structural descriptors such as aromatic ring counts, rotatable bonds, and heteroatom fractions.

Ultimately, these findings mathematically prove that the anomalous orthogonality of the HOMO energy is a global topological invariant of the B3LYP/6-31G(2df,p) state space, rather than a localized structural artifact. Recognizing this fundamental decoupling is critical for future molecular property modeling, as it implies that the electronic variance captured by HOMO requires distinct, non-linear representations that cannot be simply inferred from the thermodynamic or geometric-response core of the molecular state space.