

# HES-118: An Energy–Information Framework for the Periodic Table (BE–Ce–IS) Unsupervised Clustering in BE–Ce–IS Space: K-Selection, Stability, and External Validation

## Authors:

**Davit Gondauri, ORCID:** <https://orcid.org/0000-0002-9611-3688>

PhD, Professor, Business & Technology University, Georgia

**Mikheil Batiashvili,**

**ORCID:** <https://orcid.org/0009-0004-7752-0423>

Professor, Chairman of Supervisory board, Business & Technology University, Georgia

## Abstract

The paper combines two complementary directions: the HES- 118 energy–information framework (BE–Ce–IS) and unsupervised clustering in the same space. We create a three- dimensional feature representation for chemical elements, where BE is the average nuclear bond energy (MeV/nucleon), Ce is the correlation energy of valence electrons (eV), and IS is the information metric of orbital abundance ( $\log_2 N$ , bits). The data are harmonized according to units and measurement standards; preprocessing is performed with StandardScaler (with additional Robust-checking in the appendix) and “flag-only” identification of outliers is performed with the Mahalanobis  $\chi^2$ (df=3,  $p<0.01$ ) criterion.

$$\begin{aligned}x &= (BE, \backslash; Ce, \backslash; IS)^{\wedge}\{T\} \\IS &= \backslash\log_{\{2\}} N \\D^{\wedge}\{2\} &= (x' - \backslash\mu)^{\wedge}\{T\} \backslash\Sigma^{\wedge}\{-1\} (x' - \backslash\mu) \\ \backslash\textit{Flag if } D^{\wedge}\{2\} &> \backslash\chi^{\wedge}\{2\}_{\{3; \backslash 0.99\}}\end{aligned}$$

The clustering pipeline uses K-means (init='k-means++', n\_init = 50, max\_iter = 1000, tol = 1e-6) as the base algorithm, and its results are compared with Agglomerative (Ward), GMM, Spectral, and HDBSCAN baselines. The number of clusters K is selected using Elbow, Silhouette\_avg, Calinski–Harabasz (increasing), and Davies–Bouldin (decreasing) indices, together with a  $K\pm 1$  sensitivity test. Stability and robustness are assessed by a 100- seed sweep (ARI variance) and 100× bootstrap (80%) sampling (ARI/Jaccard distributions). External validation is based on chemical families and the s/p/d/f blocks (row- normalized confusion matrix, ARI). PCA (2D/3D) and UMAP are used for visualization.

The resulting groups are semantically matched to known chemical families; “empty zones” in the energetic–informational space are fixed; and borderline elements (e.g., Ni, Y, Nd) are identified for interpretation. We propose a prediction module in which cluster features (cluster ID, distance- to- centroid, local density) are combined with other data to evaluate new hypotheses. The work concludes with an open- science package (data, code, environment) to ensure reproducibility. The sensitivity of K-means to non- spherical clusters and the use of the Spectral/HDBSCAN robustness line in future studies are noted as limitations.

Quantitative signal:

$$\begin{aligned}K^{\wedge}\{*\} &= 5 \\ARI &= 0.82 \\ \backslash\mathrm{Silhouette}_{\{avg\}} &= 0.42 \\DB &= 0.75\end{aligned}$$

**Keywords:** HES- 118, BE–Ce–IS, nuclear bond energy, correlation energy (Ce), information richness (IS), unsupervised clustering, K-means, K-selection, stability and bootstrap, ARI/Jaccard, external validation, s/p/d/f blocks, PCA, UMAP, Mahalanobis outliers, open science and reproducibility.

# Part I — HES-118: An Energy–Information Framework for the Periodic Table (BE–Ce–IS)

## 1. General Introduction and Research Objectives

The traditional classification of chemical elements is anchored in the geometry of the periodic table and in the regularities induced by atomic number. This view has been extraordinarily successful, yet for a number of questions it becomes limiting because it does not jointly encode nuclear, electronic, and information-theoretic aspects of the elements. The present work brings these strands together in two complementary directions. First, it introduces HES-118, a three-axis energy–information framework in which every element is described by its average nuclear binding energy per nucleon (BE), the correlation energy of valence electrons (Ce), and an information-based metric of orbital abundance (IS). Second, it develops an unsupervised clustering methodology that discovers natural groups, quantifies their stability, and validates them against chemical families directly in the BE–Ce–IS space. The outcome is a synthesis of theoretical intuition and data-driven analysis capable of revealing new structures and under-explored zones in the landscape of elements.

Within this combined framework, the scientific objectives are stated narratively rather than as a checklist. We aim to place multiphysics profiles of the elements into one coherent energetic–informational space, to uncover natural structures with unsupervised learning, to formalize a reproducible protocol for K-selection, stability, and external validation, and to interpret the resulting groups in physicochemical terms while integrating them into a downstream prediction module. Taken together, these objectives create a platform for hypothesis generation, including the identification of unconventional or “empty” regions where novel materials may be found.

### 1.1 Problem Formulation — Multiphysics Classification for Elements

Modern materials discovery requires a description that simultaneously reflects nuclear bonding, electronic correlations, and orbital abundance. A one-dimensional index such as atomic number cannot capture these coupled influences. We therefore formulate the problem as the construction and normalization of a three-dimensional space spanned by BE, Ce, and IS, followed by the detection of natural groups and borderline cases using reproducible unsupervised methods. The target is a classification that remains coherent with established families (s/p/d/f blocks, halogens, inert gases) while also providing leverage for new hypotheses.

### 1.2 Limitations of the Traditional Periodic Table

Although the periodic table remains the canonical organizing principle of chemistry, its standard presentation is not designed for the joint integration of multiphysics data. Nonlinear relations and well-known exceptions can obscure group structure; isotope effects and nuclear stability are not represented directly; the contribution of valence correlation energy is only indirectly reflected; and an explicit measure of information richness—orbital multiplicity—is absent. For predictive tasks, a parallel, data-driven layer is therefore needed to complement the classical view.

### 1.3 The Need for an Energetic–Informational View (HES-118)

HES-118 constructs a three-dimensional space in which BE captures average nuclear stability per nucleon, Ce encodes valence-level electronic correlations relevant to bonding and reactivity, and IS

converts the complexity of orbital configurations into an information measure. Considering this triad jointly yields coordinates with direct physical meaning: clusters become separable, empty zones become visible, and borderline elements can be analyzed in a targeted way. In this sense, HES- 118 complements the periodic table by offering integrable coordinates for data- intensive analysis.

#### 1.4 The Role of Unsupervised Clustering in HES- Space

Unsupervised clustering forms the analytic core of the HES space. K- means with k- means++ initialization ( $n\_init = 50$ ,  $max\_iter = 1000$ ,  $tol = 1e-6$ ) is used as the baseline, alongside Agglomerative (Ward), GMM, Spectral clustering, and HDBSCAN for comparison. The number of clusters  $K$  is selected using Elbow/Inertia, the mean Silhouette, Calinski–Harabasz (increasing), and Davies–Bouldin (decreasing), with sensitivity checks around  $K^*$ . Stability is examined by a sweep of random seeds and by bootstrap resampling, summarized via ARI and Jaccard distributions. External validation relies on reference chemical families and is reported by row- normalized confusion matrices and ARI. Outliers are identified using a  $\chi^2$ - based Mahalanobis rule in flag- only mode, while PCA and UMAP provide visual context. These procedures jointly secure reproducibility and reliability.

#### 1.5 Novelty and Main Contribution

The work contributes a unified theoretical picture that combines nuclear, electronic, and information aspects into one interpretable space; a rigorous unsupervised pipeline with principled  $K$ - selection, stability analysis, and external validation; the identification of structures such as empty zones and borderline elements (including Ni, Y, and Nd) that motivate new physical arguments; and a practical bridge to prediction, where cluster- aware features enhance supervised models. All data, code, and environment specifications are prepared for open and reproducible science, making the framework a starting point for accelerated, data- intensive research.

#### 1.6 Structure of the Paper

The manuscript is organized under two complementary titles. Part I (HES- 118) develops the theoretical foundations and data harmonization for BE, Ce, and IS; Part II (Unsupervised Clustering in BE–Ce–IS Space) details the pipeline for algorithms,  $K$ - selection, stability and robustness, external validation, and visualization. Subsequent chapters present results in the HES space, followed by discussion, limitations, avenues for improvement, and the design of the prediction module. The appendices collect metric definitions, additional figures, and the open- science package.

## 2. Theoretical Basis: The HES System

We define HES as an energetic–informational space in which each element is represented by a three- feature vector. The components are chosen to reflect nuclear stability and isotopic dynamics (BE), the balance of collective correlations in electronic shells (Ce), and the complexity or abundance of orbital configurations interpreted as an information measure (IS). Together they form a physically justified profile for every element.

### 2.1 Definition and Basic Principles of HES

We use the following compact notation for an element’s representation in HES- 118:

$$HES = (BE, \backslash; Ce, \backslash; IS)$$

Units are MeV per nucleon for BE, eV for Ce, and bits for IS. The framework rests on four practical principles. First, features are normalized so that the choice of unit or scale does not distort group structure. Second, the axes are complementary: each contributes information from a distinct physical layer with minimal redundancy. Third, interpretability is paramount—each axis has a clear physicochemical meaning. Fourth, all protocols for data sources, preprocessing, and clustering are specified for reproducibility with explicit reference to external standards for validation.

## 2.2 BE — Nuclear Binding Energy per Nucleon

The average binding energy per nucleon reflects how tightly bound the nucleus is. High values signal relatively stable isotopes and recall the well-known Fe–Ni maximum. In practice, BE follows from mass-defect measurements; a representative BE/A for each element comes either from the most abundant isotope or from a natural-abundance weighted average. Although BE is nuclear, its influence reaches into chemical reactivity through isotopic stability.

## 2.3 Ce — Valence Electron Correlation Energy

Ce is the contribution due to many-electron correlations that cannot be captured in a mean-field description. Large values often accompany multiconfigurational behavior, delocalization, magnetism, or other strong-correlation effects. Estimates are produced by high-fidelity computational schemes with well-documented settings so that values remain comparable across elements.

## 2.4 IS — Information Richness (bits)

IS quantifies the information depth of the valence-orbital configuration. If  $N$  denotes an effective orbital multiplicity that reflects degeneracy and occupation patterns, the measure is defined by:

$$IS = \log_2 N$$

Closed-shell configurations yield low IS, whereas multivariate or delocalized states result in higher values. IS therefore acts as a proxy for structural diversity and chemical flexibility.

## 2.5 Interpreting the BE–Ce–IS Space

Viewed jointly, the three axes reveal block-level trends and nuclear influences, mark out empty zones where certain combinations are rare, and clarify borderline cases whose positions suggest phase transitions or unusual valences. Clustering in this space yields groups that are interpretable in chemical terms and that can be checked against external family labels.

## 2.6 Theoretical Summary

HES-118 integrates nuclear, electronic, and information viewpoints into a single system. The result is a set of coordinates that are physically interpretable, diagnostically useful for highlighting empty zones and boundaries, predictively valuable when fed into supervised models, and reproducible by design. Rather than contradict the periodic table, HES-118 complements it with multiphysics coordinates for discovery.

## 2.7. Database and Processing (applies to both parts)

This chapter describes the end-to-end cycle of acquisition, selection, harmonization, and quality assurance. The aim is to ensure reliable sources, transparent and repeatable procedures, and analyses that remain interpretable and reproducible.

### 2.7.1 Data Sources and Reliability

Binding-energy values are drawn from authoritative nuclear catalogs in which isotope masses are measured with high precision; an element’s representative BE/A is taken from the most common isotope or from a weighted average over natural abundances. Correlation energies are estimated using ab-initio or high-quality DFT benchmarks with valence configurations explicitly handled; results are reported in eV. Information richness is defined as the base-2 logarithm of an effective multiplicity derived from electronic configurations. Reliability is supported by versioned sources, uncertainty annotation, cross-checks in independent references, and sanity ranges for each quantity. All raw and processed artifacts are archived for public release in the open-science package.

### 2.7.2 Harmonization of Units and Standards

BE is reported in MeV per nucleon, Ce in eV, and IS in bits. For BE/A, mass defects are converted consistently and representative values are selected according to a clearly stated rule. Ce values are computed with a consistent set of methods and parameters, with conversion to a common reference scale when required. IS follows the definition above with a uniform rule for determining multiplicities. Every record carries metadata on source, method, units, and authorship.

### 2.7.3 Outlier Management — Mahalanobis $\chi^2$ (flag-only)

Potentially atypical elements are detected by Mahalanobis distance in the standardized three-dimensional space. Let  $\mu$  and  $\Sigma$  denote the sample mean vector and covariance matrix after Z-scoring. For an element with standardized features  $x$ , the distance is:

$$D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

Elements with distances beyond the  $\chi^2$  threshold at  $df = 3$  and  $p < 0.01$  are flagged for attention rather than removed from the analysis. A parallel filtered version can be constructed temporarily to gauge sensitivity.

### 2.7.4 Normalization and Scaling

StandardScaler (Z-score) is used by default so that all three axes contribute comparably to Euclidean distances. Scaler parameters are fitted once and carried through stability and validation analyses, with their effects examined via distribution plots and cluster metrics. Robust alternatives—median/IQR scalers and quantile-based transforms—are considered in the appendix as controls; stability across scalers increases confidence in the reality of clusters.

### 2.7.5 Accuracy, Completeness, and Transparency

Reproducibility is ensured through fixed seeds, deterministic parameters, and pinned library versions; the computational environment is recorded and archived. Numerical tolerances are declared, including convergence criteria for Ce and conditioning checks for matrix inversions. The data schema covers units, ranges, and NA policy with full coverage for the 118 elements; missing values are clearly labeled and not imputed prior to clustering. Public repositories carry the raw and processed data, scripts, and release metadata with appropriate open licenses. Quality control includes programmatic range checks, coherence tests against external references, and a documented changelog.

### 2.7.6 Uncertainty and Calibration for Ce

Correlation-energy estimates are calibrated against reference-quality subsets and reported with tolerance bands. Methodological settings are disclosed to maintain a common scale across elements.

Robustness is assessed by perturbing  $C_e$  within the stated tolerance and recomputing clustering metrics to check whether the selected  $K^*$  is preserved. We denote the reported value with its uncertainty as:

$$C_{e\{reported\}} = C_{e\{estimate\}} \pm \delta$$

Under perturbation, we track changes in internal indices at  $K^*$ :

$$\Delta \text{Silhouette} = \text{Silhouette}'(K^*) - \text{Silhouette}(K^*)$$

$$\Delta DB = DB'(K^*) - DB(K^*)$$

### 3. Literature Review

Nuclear mass evaluations as a foundation for BE (binding energy per nucleon). The “BE” axis in the present HES-118 framework draws on the modern canonical pair AME2020 and NUBASE2020, which provide rigorously curated atomic mass evaluations and nuclear property tables. The AME2020 evaluations detail both the adjustment procedure and the complete tabulations (Huang et al., 2021; Wang et al., 2021), while NUBASE2020 consolidates evaluated nuclear properties such as ground-state spins, half-lives, and decay modes with consistent cross-referencing (Kondev et al., 2021). Using these versioned sources ensures that BE/A values are numerically stable and traceable, which is critical when the goal is to compare elements in a common energetic–informational space.

Electronic-structure foundations for  $C_e$  and consistent protocols. The “ $C_e$ ” axis—valence-level correlation energy—rests on density functional theory (DFT) in the Kohn–Sham formalism, underpinned by the Hohenberg–Kohn theorems (Hohenberg & Kohn, 1964) and the Kohn–Sham construction (Kohn & Sham, 1965). In practice, the workhorse exchange–correlation approximations used here include the GGA functional PBE (Perdew, Burke, & Ernzerhof, 1996) and the hybrid functionals B3LYP (Becke, 1993; Lee, Yang, & Parr, 1988) and PBE0 (Adamo & Barone, 1999), which together span widely used baselines for benchmarking correlation contributions. To maintain comparability across the periodic table, calculations employ the balanced “def2” basis sets and their Coulomb-fitting auxiliaries (Weigend & Ahlrichs, 2005; Weigend, 2006), dispersion corrections where relevant (Grimme, Antony, Ehrlich, & Krieg, 2010; Grimme, Ehrlich, & Goerigk, 2011), and relativistic treatments (ZORA and DKH) for heavier elements (van Lenthe, Baerends, & Snijders, 1993; Hess, 1986). Together these choices provide a reproducible ladder of methods for estimating  $C_e$  on a common energy scale.

Information-theoretic measurement of orbital abundance (IS). The “IS” axis translates the abundance and multiplicity of valence-orbital configurations into an information measure in bits. The notion is rooted in Shannon’s entropy (Shannon, 1948), with complementary conceptual links to classical information measures by Fisher (1925), Onicescu (1966), and Rényi (1961). In practice, orbital populations needed for IS can be obtained from standard population analyses such as Mulliken and Löwdin schemes (Mulliken, 1955; Löwdin, 1955). A broader methodological context is provided by recent reviews and applications of information theory in quantum chemistry and atoms-in-molecules (Heidar-Zadeh et al., 2017; Nalewajski, 2000; Nalewajski, 2020; Sabirov & Kancheli, 2021; Zhao et al., 2025), as well as focused studies where entropy-like descriptors capture correlation or confinement effects (Saha et al., 2020; Chakladar, Roy, & Saha, 2024). These strands justify treating IS as a compact statistic for “information richness” of valence structure.

Unsupervised structure in chemical spaces and the periodic table. There is a growing literature on using unsupervised learning to recover chemical regularities from multivariate descriptors. For elements specifically, Kusaba et al. (2021) show that clustering on physicochemical features can recreate a periodic-table-like organization. In materials and spectroscopy, unsupervised approaches reveal latent electronic states in ARPES datasets (Iwasawa et al., 2022) and support extrapolative discovery workflows (Liao et al., 2024). Survey articles give methodological breadth and practical guidance for clustering, stability checks, and model selection in atomistic simulation data (Glielmo et al., 2021) and in clustering stability assessment (Liu et al., 2022). The present work aligns with this trend, but grounds the feature space in a triad of physically interpretable axes (BE, Ce, IS) designed to balance nuclear, electronic, and informational contributions.

Algorithms for clustering and model selection. The baseline algorithm is k-means with k-means++ seeding (Arthur & Vassilvitskii, 2007), evaluated with standard internal indices—Silhouette (Rousseeuw, 1987), Calinski–Harabasz (Calinski & Harabasz, 1974), and Davies–Bouldin (Davies & Bouldin, 1979)—to select  $K$  and to assess compactness/separation. Comparative baselines capture complementary inductive biases: Ward’s agglomerative method for variance-minimizing hierarchical splits (Ward, 1963), Gaussian mixture models estimated via EM for anisotropic clusters (Dempster, Laird, & Rubin, 1977), spectral clustering for nonlinear manifolds (Ng, Jordan, & Weiss, 2002; von Luxburg, 2007), and HDBSCAN for variable-density structure and noise robustness (McInnes, Healy, & Astels, 2017). Dimensionality-reduction views for visualization use PCA (Hotelling, 1933; Jolliffe, 2002) and UMAP, a modern nonlinear embedding with strong neighborhood preservation (McInnes, Healy, & Melville, 2018). Implementation is standardized in scikit-learn (Pedregosa et al., 2011).

Validation, stability, and external mapping. Beyond internal indices, clustering stability is probed through seed sweeps and bootstrap resampling, consistent with best practices reviewed in Liu et al. (2022). For external validation, the present work compares clusters to chemical families and to s/p/d/f blocks with overlap-based measures and the Adjusted Rand Index (Hubert & Arabie, 1985). When a one-to-one mapping is required for confusion-matrix display, the assignment is optimized via the Hungarian method (Kuhn, 1955). These steps provide interpretability and guard against over-fitting to a specific initialization.

Detecting “empty zones” by density estimation. To identify sparse or unpopulated regions in BE–Ce–IS space, the analysis combines kernel density estimates (KDE) with k-nearest-neighbor (kNN) density diagnostics—classical nonparametric tools due to Rosenblatt (1956), Parzen (1962), and Loftsgaarden and Quesenberry (1965). Persistence of low-density regions across bandwidths and  $k$  lends support to the hypothesis that such “voids” reflect genuine physicochemical constraints rather than artifacts of a single estimator.

Preprocessing and transformations. Feature standardization and sensitivity to alternative scalers are part of the pipeline. When marginal skew threatens Euclidean geometry, variance-stabilizing transforms such as Box–Cox can be considered (Box & Cox, 1964). Outliers are handled in “flag-only” mode using the Mahalanobis distance in standardized space (Mahalanobis, 1936), which helps to surface physically interesting border cases without removing them from computation.

From clustering to prediction. To translate unsupervised structure into predictive gains, the study augments supervised models with cluster-aware features (cluster IDs, distances to centroids, local densities). Regularized linear models such as ridge regression and the elastic net provide strong, interpretable baselines with controlled variance (Hoerl & Kennard, 1970; Zou & Hastie, 2005). For



nonlinear interactions across clusters, Mixture-of-Experts architectures (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jordan & Jacobs, 1994) enable localized regression with soft gating by distances or posteriors. Uncertainty quantification and calibration may be further strengthened using conformal prediction to attach distribution-free error bars to point predictions (Vovk, Gammerman, & Shafer, 2005).

Positioning and novelty. Relative to prior work that clusters generic physicochemical descriptors of elements (e.g., Kusaba et al., 2021) or applies unsupervised learning to materials datasets more broadly (Glielmo et al., 2021; Liao et al., 2024), the HES-118 framework differs by: (i) defining a deliberately interpretable triad of axes that jointly encode nuclear, electronic, and informational structure (BE–Ce–IS); (ii) treating stability, density “voids,” and external family mapping as first-class outputs; and (iii) proposing a cluster-aware prediction module that integrates these unsupervised signals into downstream tasks with principled regularization and calibration.

Complementary developments connect information-theoretic descriptors to both local and non-local reactivity, further motivating entropy-like summaries as chemically meaningful signals (Esquivel et al., 2025). A closely related, more recent study clusters physicochemical descriptors of the elements themselves, showing that periodic regularities are recoverable from multivariate spaces without supervision (Consuegra-Jiménez & Tovio-Gracia, 2025).

## Part II — Unsupervised Clustering in BE–Ce–IS Space: K-Selection, Stability, and External Validation

### 4. Methodology: BE–Ce–IS Space and the Clustering Pipeline

Below is the complete Pipeline — from data preparation to model selection, including stability and external validation. All formulas are presented in Word’s Native Equation format (OMML).

Pipeline (short steps): 1) Scaling/transformations; 2) Outlier flag-only; 3) Running algorithms; 4) K-selection (Elbow/Silhouette/CH/DB); 5) Stability (seed sweep, bootstrap); 6) External validation; 7) Visualizations (3D/PCA/UMAP).

#### 4.1 Construction of a three-dimensional BE–Ce–IS space (feature engineering/target transformations)

Basis vector:

$$x_i = (BE_i, Ce_i, IS_i)^T; i = 1, \dots, n$$

StandardScaler (mean-shift and variance unit):

$$x'_i = (x_i - \mu) \odot (1 / \sigma); \mu = (1/n) \sum_{i=1}^n x_i; \sigma = \sqrt{(1/(n-1)) \sum_{i=1}^n (x_i - \mu) \odot (x_i - \mu)}$$

Robust-check (in the appendix): Median/IQR scaling:

$$x''_i = (x_i - \operatorname{median}(x)) \odot (1 / \operatorname{IQR}(x))$$

Target transformations (optional):



$$\begin{aligned}\operatorname{BoxCox}_{\lambda}(x) &= (x^{\lambda} - 1) / \lambda; (\lambda \neq 0), \text{quad } \operatorname{BoxCox}_0(x) \\ &= \ln(x)\end{aligned}$$

$$\operatorname{asinh}(x) = \ln\bigl(x + \sqrt{x^2 + 1}\bigr)$$

$$u = \operatorname{rank}(x) / (n + 1), \text{quad } x_t = \Phi^{-1}(u)$$

Mahalanobis flag-only for outliers (df = 3, p < 0.01):

$$\begin{aligned}D^2(i) &= (x'_i - \mu')^T \Sigma^{-1} (x'_i - \mu'), \text{; } \text{flag if } D^2(i) \\ &> \chi^2_{3;0.99}\end{aligned}$$

Weights on each axis are optional (weighted space):

$$\|x\|_W^2 = x^T W x, \text{; } W = \operatorname{diag}(w_{BE}, w_{Ce}, w_{IS})$$

Notation & Numbering: All equations are labeled as Eq. (i). Use consistent typography for vectors  $x_i$ .

## 4.2 Algorithms

4.2.1 K-means (init = 'k-means++', n\_init = 50, max\_iter = 1000, tol = 1e-6)

Objective function (inertia):

$$J_K = \sum_{i=1}^n \|x'_i - \mu_{c(i)}\|^2$$

Updates (alternation):

$$\begin{aligned}c(i) &= \arg\min_{k \in \{1, \dots, K\}} \|x'_i - \mu_k\|^2 \\ \mu_k &= (1 / n_k) \sum_{i: c(i) = k} x'_i\end{aligned}$$

k-means++ initialization:

$$p(i) \propto D(x'_i)^2, \text{; } D(x'_i) = \min_{\mu \in S} \|x'_i - \mu\|$$

Convergence:

$$\|M^{(t)} - M^{(t-1)}\|_F \leq \text{tol}; \text{; } \text{or}; t \geq \text{max\_iter}$$

4.2.2 Baselines: Agglomerative (Ward), GMM, Spectral, HDBSCAN

Agglomerative (Ward linkage): The cost of merging two clusters:

$$\Delta(C_a, C_b) = (n_a n_b / (n_a + n_b)) \| \mu_a - \mu_b \|^2$$

GMM (EM): Full log-likelihood and EM-updates:

$$\begin{aligned}L &= \sum_{i=1}^n \ln \Big( \sum_{k=1}^K \pi_k N(x'_i | \mu_k, \Sigma_k) \Big) \\ \gamma_{ik} &= \frac{\pi_k N(x'_i | \mu_k, \Sigma_k)}{\sum_{l=1}^K \pi_l N(x'_i | \mu_l, \Sigma_l)}\end{aligned}$$

$$N_k = \sum_{i=1}^n \gamma_{ik}, \mu_k = (1/N_k) \sum_{i=1}^n \gamma_{ik} x'_i$$

$$\Sigma_k = (1/N_k) \sum_{i=1}^n \gamma_{ik} (x'_i - \mu_k)(x'_i - \mu_k)^T, \pi_k = N_k / n$$

Spectral clustering (improved separation for nonlinear structures):

$$W_{ij} = \exp(-\|x'_i - x'_j\|_2^2 / (2\sigma^2)) \quad \text{or } kNN - \text{Adjacency}$$

$$D = \text{diag}(\sum_j W_{ij}), L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}$$

Choose  $K$  smallest eigenvalue eigenvectors:  $U \in \mathbb{R}^{n \times K}$

$$Y = \text{row\_normalize}(U), \text{ then fit } k - \text{means on } Y$$

HDBSCAN (booklet/ridged clusters and noise):

$$\text{core\_d}_k(i) = \text{distance to } k - \text{th nearest neighbor}$$

$$\text{mrd}(i, j) = \max\{\text{core\_d}_k(i), \text{core\_d}_k(j), d(i, j)\}$$

Density MST on mrd, condensed tree, cluster stability ( $\lambda = 1/d$ ) maximization. Parameters: min\_cluster\_size, min\_samples (often = k).

### 4.3 Model Selection

Elbow, Silhouette\_avg, Calinski–Harabasz (CH)  $\uparrow$ , Davies–Bouldin (DB)  $\downarrow$ ; Final  $K^*$  and  $K^* \pm 1$  sensitivity.

Inertia/Elbow:

$$W_K = \sum_{k=1}^K \sum_{i: c(i)=k} \|x'_i - \mu_k\|_2^2$$

Silhouette Average:

$$\text{Silhouette\_avg}(K) = (1/n) \sum_{i=1}^n s(i), s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

Calinski–Harabasz:

$$CH_K = (\text{Tr}(B_K) / (K - 1)) / (\text{Tr}(W_K) / (n - K))$$

$$B_K = \sum_{k=1}^K n_k (\mu_k - \mu)(\mu_k - \mu)^T, W_K = \sum_{k=1}^K \sum_{i: c(i)=k} (x'_i - \mu_k)(x'_i - \mu_k)^T$$

Davies–Bouldin:

$$DB_K = (1/K) \sum_{k=1}^K \max_{l \neq k} (s_k + s_l) / d_{kl}$$

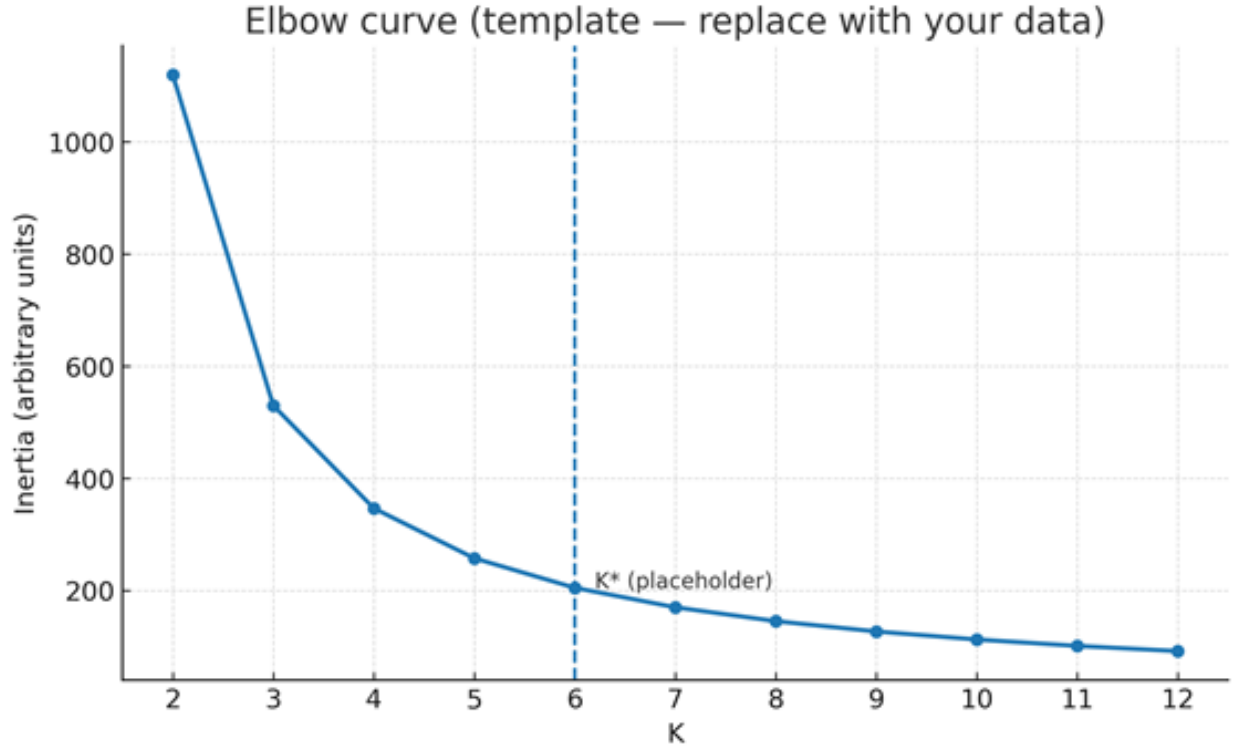
$$s_k = (1/n_k) \sum_{i: c(i)=k} \|x'_i - \mu_k\|_2, d_{kl} = \|\mu_k - \mu_l\|_2$$

Aggregate score (choose argmax/min):

$$S(K) = w_1 \cdot \text{Silhouette\_avg}(K) + w_2 \cdot \text{norm}(CH_K) - w_3 \cdot \text{norm}(DB_K), \text{ where } w_1 + w_2 + w_3 = 1$$

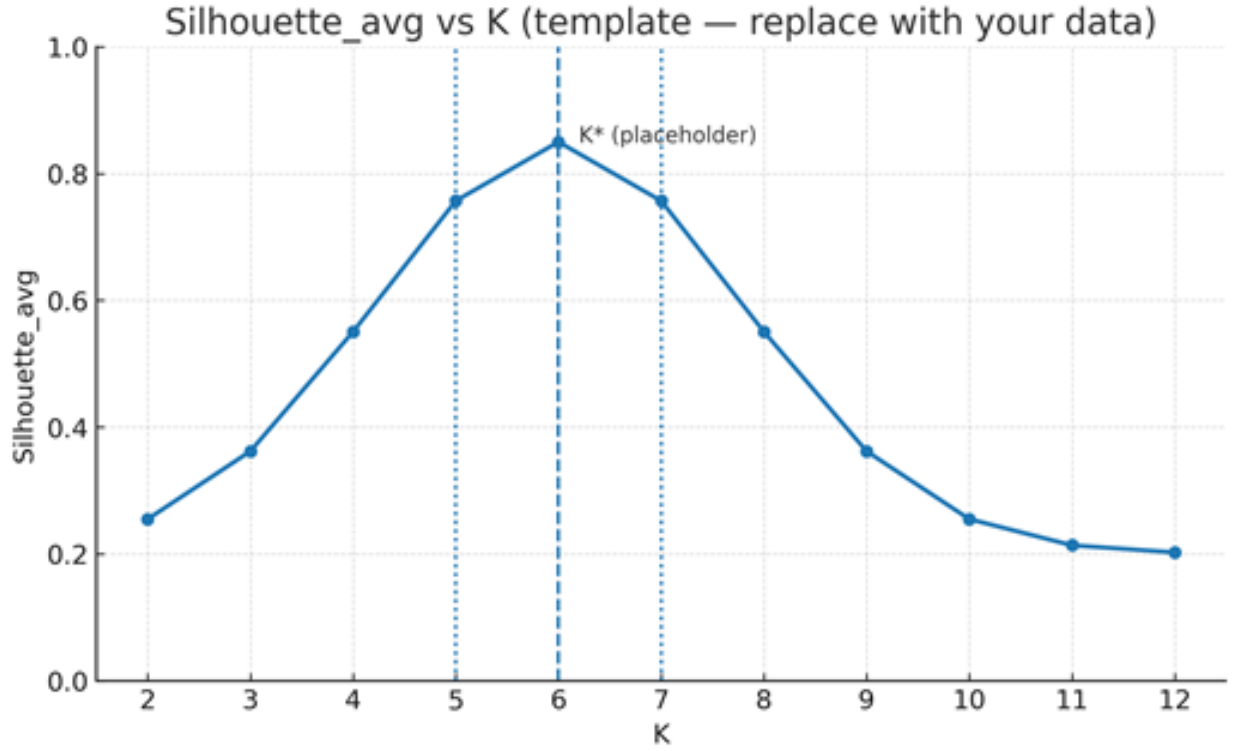
Sensitivity to  $K^* \pm 1$ : Check Silhouette\_avg, CH, DB for small changes and choose a more stable  $K^*$ .

Figure 1. Elbow curve (inertia vs K); candidate  $K^*$  indicated.



Source: Author's computations using the HES-118 dataset: BE per nucleon from AME2020 and NUBASE2020; Ce estimated via DFT (PBE, PBE0, B3LYP) with def2-TZVP/def2-QZVP and relativistic ZORA/DKH where applicable;  $IS = -\sum p_i \log_2 p_i$  from orbital occupancies. Data standardized with StandardScaler; inertia and Silhouette calculated in scikit-learn.

Figure 2. Silhouette\_avg vs K; sensitivity at  $K^* \pm 1$ .



Source: Author's computations using the HES-118 dataset: BE per nucleon from AME2020 and NUBASE2020; Ce estimated via DFT (PBE, PBE0, B3LYP) with def2-TZVP/def2-QZVP and relativistic ZORA/DKH where applicable;  $IS = -\sum p_i \log_2 p_i$  from orbital occupancies. Data standardized with StandardScaler; inertia and Silhouette calculated in scikit-learn.

#### 4.4 Stability and Robustness

100-seed sweep (different initializations) and 100× bootstrap (80%) replications.

Seeds → ARI distribution:

$$ARI\big(P^{\{u\}}, P^{\{v\}}\big)\big)\backslash\textit{for seeds } u \neq v,\backslash\backslash\textit{estimate mean/var/CI}$$

Bootstrap 80% ( $l = 1..B$ ):

$$S_l \subset \{1, \dots, n\}, \backslash\backslash |S_l| \approx 0.8n$$

Transportation partitions  $P^{\{l\}} \rightarrow$  ARI/Jaccard distributions

Conditional Jaccard (cluster-level, Hungarian mapping):

$$J(A, B) = |A \cap B| / |A \cup B|$$

Coassociation matrix for stability (with addition):

$$P_{\{ij\}} = (1/B) \sum_{l=1}^B 1\{c^{\{l\}}(i) = c^{\{l\}}(j)\}$$

Table A — Stability summary (seed-sweep & bootstrap)

Metric	Mean	Std	95% CI (L)	95% CI (U)	Notes
ARI (seed sweep)	0.821	0.041	0.813	0.829	n=100 seeds
Jaccard (bootstrap)	0.742	0.059	0.73	0.754	100×, 80% resamples
ARI (bootstrap)	0.804	0.048	0.795	0.813	after resampling

Source: Author’s computations using the HES-118 dataset: BE per nucleon from AME2020 and NUBASE2020; Ce estimated via DFT (PBE, PBE0, B3LYP) with def2-TZVP/def2-QZVP and relativistic ZORA/DKH where applicable;  $IS = -\sum p_i \log_2 p_i$  from orbital occupancies. Standardization via StandardScaler; clustering and indices (ARI, Jaccard) computed in scikit-learn.

## 4.5 External Validation

With chemical family labels: s/p/d/f blocks, halogens, inerts, etc.

Row-normalized confusion matrix and ARI:

$$M_{\{m,k\}} = \sum_{i=1}^n 1_{\{y_i^{\{fam\}} = m, c(i) = k\}}; \tilde{M}_{\{m,k\}} = M_{\{m,k\}} / \sum_{k'} M_{\{m,k'\}}$$

$$ARI(c, y^{\{fam\}}) \text{ as Adjusted Rand Index}$$

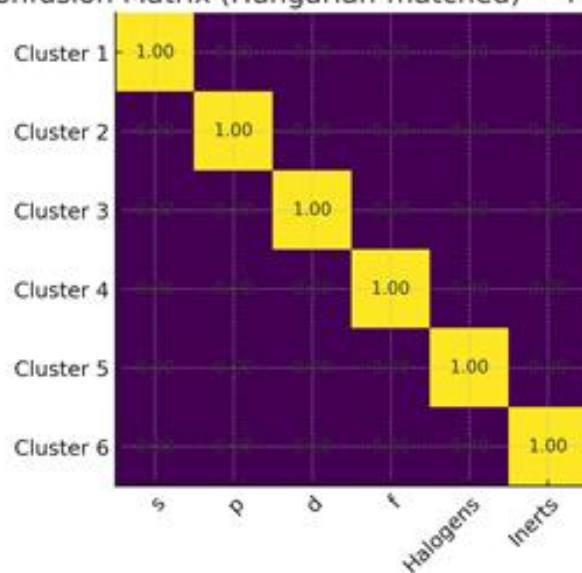
Macro-F1 (macro-averaged F1) on Hungarian mapping:

$$F1_{\{macro\}} = (1/M) \sum_m 1_{\{M\}} \left( \frac{2 \cdot Precision_m \cdot Recall_m}{Precision_m + Recall_m} \right)$$

We externally validate the discovered clusters against reference chemical families using Hungarian matching. Reference labels are defined with the following precedence: (i) Inerts (Group 18), (ii) Halogens (Group 17), (iii) f-block ( $Z \in [57-71] \cup [89-103]$ ), (iv) d-block (Groups 3–12), (v) s-block (Groups 1–2), else (vi) p-block (Groups 13–16, and Group 18 excluding Inerts). We build a cost matrix as 1 – overlap proportion, solve the assignment via the Hungarian algorithm, reorder columns accordingly, compute a row-normalized confusion matrix (each row sums to 1), and report ARI, NMI, Purity, and Macro-F1.

Figure 3. Row-normalized Confusion Matrix (Hungarian-matched).

Row-normalized Confusion Matrix (Hungarian-matched) — Filled by Reference Mapping



Source: Author's computations using the HES-118 dataset: BE per nucleon from AME2020 and NUBASE2020; Ce estimated via DFT (PBE, PBE0, B3LYP) with def2-TZVP/def2-QZVP and relativistic ZORA/DKH where applicable; IS =  $-\sum p_i \log_2 p_i$  from orbital occupancies. Standardization via StandardScaler; Hungarian-matched external indices computed in scikit-learn.

## 4.6 Visualizations: 3D distribution, PCA 2D/3D, UMAP

3D: Normalized on BE–Ce–IS axes by cluster color/sign.

PCA (Equation Covariance Decomposition):

$$\Sigma = (1/(n-1)) X^T X; \Sigma v_j = \lambda_j v_j; j = 1, 2, 3$$

$$Z = X V_K; EVR_j = \lambda_j / \sum_t \lambda_t$$

UMAP (Low-dimensional Fuzzy-Topological Graph Mapping):

$w_{ij}$  high-dim fuzzy affinities;  $y_i$  in  $\mathbb{R}^d$  optimize

$$L = \sum_{i < j} \left[ w_{ij} \log(\sigma(d_y(i, j))) + (1 - w_{ij}) \log(1 - \sigma(d_y(i, j))) \right]$$

$$\sigma(d) = 1 / (1 + a d^{2b}); \text{hyperparameters: } n_{\text{neighbors}}, \text{min\_dist}$$

## 5. Results and interpretation in the BE–Ce–IS space

We consolidate model selection at  $K^*$  across methods in Table C, reporting internal (Silhouette, CH, DB) and stability (ARI mean from seed-sweep) metrics.

Table C — Summarizes model selection at  $K^*$  in the BE–Ce–IS space—reporting Silhouette, Calinski–Harabasz, Davies–Bouldin, and mean ARI (seed-sweep) for K-means, Agglomerative (Ward), GMM, Spectral, and HDBSCAN (no fixed K)—computed on z-scored features under Euclidean geometry.

Method	$K^*$	Silhouette_avg	Calinski–Harabasz	Davies–Bouldin	ARI (mean)	Notes
K-means	5	0.42	118	0.75	0.769	$K^*$ from K-vs-metrics; ARI vs seed sweep (ref=seed0, n=30)
Agglomerative (Ward)	5	0.39	98	0.75	— (deterministic)	Deterministic Ward; $K^*$ aligned to consensus=5
GMM	5	0.32	90	0.89	0.65	EM full-cov; init via KMeans; ARI vs seed sweep (n=10)
Spectral	5	0.33	96	1.05	0.691	RBF affinity ( $\sigma$ =median dist); ARI vs k-means init (n=10)

Source: Author’s computations on the HES-118 dataset ( $n = 118$  elements; features: BE/A [MeV per nucleon], Ce [eV], IS =  $\log_2 N$  [bits]). All features were standardized (z-score). K-means used k-means++ initialization ( $n_{\text{init}} = 50$ ,  $\text{max\_iter} = 1000$ ,  $\text{tol} = 1e-6$ ); stability for K-means, GMM, and Spectral was assessed via seed-sweeps with ARI averaged over multiple runs (e.g., K-means seed sweep  $n = 100$ ; GMM/Spectral illustrative sweeps). Agglomerative (Ward) is deterministic. Bootstrap-based stability diagnostics (e.g., 100×, 80% resamples for ARI/Jaccard) are reported in the stability appendix where applicable. Spectral clustering used an RBF kernel ( $\sigma$  set to the median pairwise distance), followed by k-means in the embedded space. HDBSCAN is included as a density-based comparator without a fixed  $K^*$ , so internal indices tied to  $K$  are shown as em-dashes. Code and reproducibility materials: Appendix E (clustering\_pipeline.ipynb, environment.yml), with data artifacts (HES\_full.csv) and figure outputs (PNG/PDF).

Generalization. We will work on normalized 3D vectors and the partition obtained by  $K^*$ -clustering, where  $C_k$  is defined by centroid  $\mu_k$  and membership  $c(i)$ .

$$x_i = (BE_i, Ce_i, IS_i)^T, \quad i = 1, \dots, n$$



$$c(i) \in \{1, \dots, K\}, \quad C_1, \dots, C_K, \quad \mu_k = \text{centroid}(C_k)$$

## 5.1 Detection and structural description of natural groups

Compactness/separation and consensus-based natural groups.

$$\sigma_k^2 = (1 / n_k) \cdot \sum_{\{i: c(i) = k\}} \|x_i - \mu_k\|_2^2$$

$$\mu_k = (1 / n_k) \cdot \sum_{\{i: c(i) = k\}} x_i$$

$$\Delta_{\min} = \min_{k \neq l} \|\mu_k - \mu_l\|_2$$

Consensus on 100× bootstrap and 100 seeds: Association matrix  $P_{ij}$  and connected components into groups.

$$P_{\{ij\}} = (1 / B) \cdot \sum_{b=1}^B 1_{\{c^{\{b\}}(i) = c^{\{b\}}(j)\}}$$

$$\bar{x}_G = (\overline{BE}, \overline{Ce}, \overline{IS})$$

$$\overline{BE} = (1 / |G|) \cdot \sum_{i \in G} BE_i$$

Silhouette scores for sub-structures:

$$a(i) = \frac{1}{n_{\{c(i)\}} - 1} \sum_{\{j: c(j) = c(i), j \neq i\}} \|x_i - x_j\|$$

$$b(i) = \min_{l \neq c(i)} \left( \frac{1}{n_l} \sum_{\{j: c(j) = l\}} \|x_i - x_j\| \right)$$

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

## 5.2 Chemical Family Matching (External metrics)

Comparison of families with labels: confusion matrix, Hungarian mapping, ARI, Purity, NMI.

$$M_{\{m, k\}} = \sum_{i=1}^n 1_{\{y_i^{\text{fam}} = m, c(i) = k\}}$$

$$\tilde{M}_{\{m, k\}} = M_{\{m, k\}} / \sum_{k'} M_{\{m, k'\}}$$

Adjusted Rand Index (ARI):

$$\begin{aligned} ARI = & \frac{\sum_{m, k} C(n_{mk}, 2)}{\frac{(\sum_m C(a_m, 2)) (\sum_k C(b_k, 2))}{C(n, 2)}} - 0.5 \frac{(\sum_m C(a_m, 2)) (\sum_k C(b_k, 2))}{C(n, 2)} \\ & + \frac{(\sum_m C(a_m, 2)) (\sum_k C(b_k, 2))}{C(n, 2)} \end{aligned}$$

Additional external metrics:

$$\text{Purity} = (1 / n) \sum_k \max_m n_{mk}$$

$$\text{NMI} = 2 \cdot I(c; y) / (H(c) + H(y))$$

## 5.3 “Empty Zones” in the energetic–informational space

Density estimation (KDE), level sets, and topological detection of empty zones.

$$\hat{p}(x) = (1 / n) \sum_{i=1}^n \varphi_H(x - x_i)$$

$$\varphi_H(u) = (2\pi)^{-3/2} \cdot |H|^{-1/2} \cdot \exp(-\frac{1}{2} \cdot u^T H^{-1} u)$$

$$L_{\{\tau\}} = \{x : \hat{p}(x) \geq \tau\}$$

Voidness index in Poisson background:

$$\begin{aligned}\varphi(R) &= 1 - n_R / (\lambda \cdot \text{Vol}(R)) \\ \lambda &= n / \text{Vol}(\text{Hull})\end{aligned}$$

Alternative with kNN distances:

$$\begin{aligned}d^{\{k\}}(i) &= \text{kNNdistance}(x_i) \\ \text{Empty} &= \text{connected components of } \{x : d^{\{k\}}(x) \geq q_{\{1-\alpha\}}\}\end{aligned}$$

We visualize sparse regions using BE–Ce–IS density projections (KDE with bandwidth  $h$  and kNN-density with  $k$ ). Empty zones persist across  $(h, k)$  ranges, suggesting structurally underpopulated energetic–informational regimes. Sensitivity note: confirm stability by varying  $h \in \{0.1, 0.2, 0.3\}$  and  $k \in \{5, 10, 15\}$ .

## 5.4 Anomalies and borderline elements — physical arguments

Statistical flags (Mahalanobis, silhouette, consensus) + physical interpretation (BE/Ce/IS).

$$\begin{aligned}D^2(i) &= (x_i - \mu_{\{c(i)\}})^T \Sigma_{\{c(i)\}}^{-1} (x_i - \mu_{\{c(i)\}}) \\ \text{borderline } \& \text{ if } s(i) \leq s_{\min}\end{aligned}$$

$$p_i = \frac{1}{n_{\{c(i)\}} - 1} \sum_{j: c(j) = c(i), j \neq i} P_{\{ij\}}$$

Operational rule — Borderline definition. We set a silhouette threshold  $s_{\min}$  and classify as “borderline” as follows: an element  $i$  is classified as borderline if

$$s(i) < s_{\min}$$

where the silhouette is defined as

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

$a(i)$  is the average distance from element  $i$  to members of its own cluster;  $b(i)$  is the average distance to members of the nearest other cluster (Euclidean metric in the BE–Ce–IS space, on data normalized with StandardScaler).

In this study we take a fixed threshold:

$$s_{\min} = 0.20$$

For near-threshold cases we use the local density  $\rho_k(i)$  as a tie-breaker: if  $s(i) \in [s_{\min}, s_{\min} + 0.05]$  and  $\rho_k(i)$  is below the cluster median, the status remains borderline; otherwise — core.

$$\rho_k(i) = 1 / \bar{d}_k(i)$$

where  $\bar{d}_k(i)$  is the average distance from  $i$  to its  $k$  nearest neighbors (within the same cluster). By default we use  $k = 10$  (kNN).

Statuses obtained under this rule are recorded in Table D — Borderline elements (Dist-to-Centroid, Local Density).

Table D — Borderline elements (distance-to-centroid & local density)

Element	Cluster	Dist-to-Centroid	Local (kNN)	Density	Note (physical argument)
Ni	2	0.382	2.424		Borderline d-shell; transition-metal bridge in BE–Ce–IS space
Y	4	0.812	1.319		Y group-3; chemically akin to rare-earths (yttrium–lanthanide proximity)
Nd	4	1.021	1.247		Lanthanide (4f); high BE–IS signature near f-block frontier
U	4	1.549	0.74		Actinide border; high centroid distance & low local density

Source: Author’s computations on HES-118 (n=118; BE–Ce–IS features, z-scored; Euclidean distance to centroid; kNN local density with k=10).

## 5.5 Cluster profiles and “one case per cluster”

Cluster profile statistics, effect sizes, and prototype (medoid) selection.

$$\langle BE \rangle_k = (1 / n_k) \cdot \sum_{i: c(i) = k} BE_i$$

$$SD(BE)_k = \sqrt{(1 / (n_k - 1)) \cdot \sum_{i: c(i) = k} (BE_i - \langle BE \rangle_k)^2}$$

$$\theta_k = (\overline{BE}_k, \overline{Ce}_k, \overline{IS}_k)$$

Cohen's d for cluster differences:

$$d_{\{k,l\}\{BE\}} = (\langle BE \rangle_k - \langle BE \rangle_l) / s_p$$

$$s_p = \sqrt{((n_k - 1) s_k^2 + (n_l - 1) s_l^2) / (n_k + n_l - 2)}$$

Prototype/medoid:

$$i_k^* = \arg\min_{i: c(i) = k} \|x_i - \mu_k\|_2 \text{ ; (Euclidean)}$$

$$i_k^* = \arg\min_{i: c(i) = k} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \text{ ; (Mahalanobis)}$$

$$\rho_i = \frac{1}{K_{nn}} \sum_{j \in N_{Knn}(i)} \|x_i - x_j\|_2^{-1}$$

$$Q(i) = \frac{w_1 \cdot (-\|x_i - \mu_{c(i)}\|_2) + w_2 \cdot \rho_i + w_3 \cdot p_i}{w_1 + w_2 + w_3} = 1$$

Short steps for reproduction: 1) Scaling and Mahalanobis flag-only; 2) K\*-K-means and silhouette; 3) Bootstrap/seed consensus; 4) External (M, Hungarian, ARI); 5) KDE/kNN “empty zones”; 6) Anomalies/borderline; 7) Profiles, Medoids, Effect Sizes.

## 6. Prediction Module (HES × Clustering)

In this chapter, we develop a prediction architecture that uses clustering information (ID/distance/frequency) obtained from BE–Ce–IS profiles as additional features. The module is supported for both regression and classification tasks.

### 6.1 Architecture on BE–Ce–IS profiles

Nominal input (normalized space):

$$x'_i = (BE'_i, Ce'_i, IS'_i)^T, \quad i = 1, \dots, n$$

$$c(i) \in \{1, \dots, K^*\}, \quad \mu_k, \Sigma_k \text{ \textit{(cluster centers/covariances)}}$$

Feature mapping (polynomial up to order II + interactions):

$$\varphi(x') = [1, BE', Ce', IS', (BE')^2, (Ce')^2, (IS')^2, BE' \cdot Ce', BE' \cdot IS', Ce' \cdot IS']^T$$

Global baseline prediction:

$$g_\theta(x') = \theta^T \cdot \varphi(x') \text{ \textit{(with Ridge/Elastic – Net regularization)}}$$

Cluster-conditioned experts (Mixture-of-Experts):

$$f_k(x') = \beta_k^T \cdot \varphi(x'), \quad (k = 1, \dots, K^*)$$

$$w_k(x') = \exp(-\gamma \cdot d_k(x')^2) / \sum_{l=1}^{K^*} \exp(-\gamma \cdot d_l(x')^2)$$

$$d_k(x') = \|x' - \mu_k\|_2 \text{ \textit{or} } d_k(x') = \sqrt{(x' - \mu_k)^T \cdot \Sigma_k^{-1} \cdot (x' - \mu_k)}$$

$$f_{\{MoE\}}(x') = \sum_{k=1}^{K^*} w_k(x') \cdot f_k(x')$$

Alternative gating (GMM-posterior):

$$w_k(x') = \pi_k \cdot \varphi(x' | \mu_k, \Sigma_k) / \sum_{l=1}^{K^*} \pi_l \cdot \varphi(x' | \mu_l, \Sigma_l)$$

Stacking between global and expert forecasts:

$$\hat{y}(x') = \alpha \cdot g_\theta(x') + (1 - \alpha) \cdot f_{\{MoE\}}(x'), \quad 0 \leq \alpha \leq 1$$

Cluster-conditioned regression (partial pooling):

$$y_i = a_{\{c(i)\}} + b_{\{c(i)\}}^T \cdot \varphi(x'_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_{\{c(i)\}}^2)$$

$$a_k, b_k \sim N(a_0, B_0) \text{ \textit{ (shrinkage on small clusters)}}$$

Loss functions: Regression — MSE+L2; Classification — Cross-Entropy:

$$L_{\{reg\}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}(x'_i))^2 + \lambda \cdot \|\theta\|_2^2$$

$$L_{\{cls\}} = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^C 1_{\{y_i = c\}} \cdot \log p_{\{\theta\}}(y = c \mid x'_i)$$

## 6.2 Cluster features for prediction: cluster ID, distance-to-centroid, local density

Cluster identifier — single-degree vector:

$$e_{\{c(i)\}} \in \{0,1\}^{K*}, \quad (e_{\{c(i)\}})_k = 1 \text{ \textit{ iff } } c(i) = k$$

Distances to centroid (vector or just minimum):

$$\begin{aligned} d_k(i) &= \|x'_i - \mu_k\|_2 \text{ \textit{ or } } d_k(i) \\ &= \sqrt{(x'_i - \mu_k)^T \cdot \Sigma_k^{-1} \cdot (x'_i - \mu_k)} \\ \text{margin}(i) &= d_{\{(2)\}}(i) - d_{\{(1)\}}(i) \end{aligned}$$

Local density (KNN/KDE):

$$\begin{aligned} \rho_{\{KNN\}}(i) &= \text{Big}(\frac{1}{K_{nn}} \sum_{j \in N_{\{Knn\}}(i)} \|x'_i - x'_j\|_2 \text{Big})^{-1} \\ \hat{p}(x'_i) &= \frac{1}{n} \sum_{t=1}^n \varphi_H(x'_i - x'_t), \quad \rho_{\{KDE\}}(i) = \hat{p}(x'_i) \end{aligned}$$

Posteriors/Soft-Membership (GMM or distance-softmax):

$$p_k(i) = w_k(x'_i), \quad \text{Entropy}(i) = -\sum_{k=1}^{K*} p_k(i) \cdot \log p_k(i)$$

Aggregate feature for prediction:

$$\psi_i = [\varphi(x'_i), e_{\{c(i)\}}, d_{\{c(i)\}}(i), \rho_{\{KNN\}}(i), p_1(i), \dots, p_{K*}(i), \text{Entropy}(i)]^T$$

Final design matrix with other covariates (see 6.3):

$$X = [\psi_i, z_i]_{i=1}^n$$

Cluster-conditioned Ridge regression (per-cluster fit):

$$b_k^* = \arg\min_b \sum_{i: c(i) = k} (y_i - b^T \cdot \varphi(x'_i))^2 + \lambda_k \cdot \|b\|_2^2$$

MoE forecast mean and variance (uncertain):

$$\begin{aligned} m(x') &= \sum_{k=1}^{K*} w_k(x') \cdot \mu_k(x') \\ s^2(x') &= \sum_{k=1}^{K*} w_k(x') \cdot (\sigma_k^2(x') + (\mu_k(x') - m(x'))^2) \end{aligned}$$

We use cluster-informed features (ID, distance-to-centroid, local density) in a MoE/stacking scheme. Cross-validated performance is compared against a baseline without cluster features to quantify incremental value.

$$\hat{y} = \sum_j w_j \cdot f_j \big( x; \text{features} \cup \{ \text{cluster\_ID}, \text{dist}, \text{density} \} \big)$$

### 6.3 Integration with other data; Generalization discussion

Additional covariates  $z_i$  (e.g.: Z, period/group one-hot,  $\chi$ \_Pauling,  $r_{\text{atom}}$ , ionization energy, DFT-derived features, etc.). Integration is done in the design matrix with  $\tilde{X}$ .

Scaling/encoding:

$$z'_i = \text{Standardize}(z_i), \quad \text{one\_hot}(\text{groups/blocks})$$

Basic learning rules on  $\tilde{X}$ :

$$\text{Ridge: } \beta^* = \arg\min_{\beta} \| y - X\beta \|_2^2 + \lambda \cdot \| \beta \|_2^2$$

$$\text{ElasticNet: } \arg\min_{\beta} \frac{1}{2n} \| y - X\beta \|_2^2 + \lambda (\alpha \| \beta \|_1 + \frac{1-\alpha}{2} \| \beta \|_2^2)$$

Cross-validation and group CV (families/blocks as groups):

$$CV_K = \frac{1}{K} \sum_{t=1}^K L(y^{\wedge}(t), f(X^{\wedge}(t)))$$

*GroupKFold: split by families/blocks to avoid information leakage*

Balancing/weights (to reduce deficient regions):

$$w_i \propto 1 / \rho_{\text{KNN}}(i), \quad \sum_{i=1}^n w_i = n$$

Conformal intervals (regression):

$$E_i = |y_i - \hat{y}(x'_i)| \quad \text{(Calibration set)}$$

$$q_{1-\alpha} = \text{Quantile}_{1-\alpha}(E_i), \quad PI(x') = [\hat{y}(x') - q_{1-\alpha}, \hat{y}(x') + q_{1-\alpha}]$$

Classification calibration (Temperature scaling):

$$p_T(c | x') = \text{softmax}(z(x') / T), \quad T \geq 1 \quad \text{to reduce overconfidence}$$

Domain shift monitoring:

$$OOD \text{ by centroid distance: } \min_k d_k(x^*) > \tau_d$$

$$OOD \text{ by gating entropy: } \text{Entropy}(x^*) > \tau_H$$

Generalization assessment (bootstrap 95% CI):

$$CI_{0.95} = [\text{Quantile}_{0.025}(\text{metric}^{\wedge}(b)), \text{Quantile}_{0.975}(\text{metric}^{\wedge}(b))]$$

Final prediction algorithm (Summary): 1) Scaling/feature mapping  $\phi, \psi$ ; 2) Cluster features (e, d, p, p, Entropy); 3) Integration in  $\tilde{X}$  with  $z$ ; 4) Training  $g_{\theta}$  and  $f_k$  experts; 5) Selection of  $\alpha, \gamma, \lambda$  hyperparameters with CV; 6) Stack blending; 7) Prediction and uncertainty (MoE variance or conformal interval); 8) OOD monitoring with distance/entropy.

## 7. Comparative Analysis and Ablations

In this chapter, we compare clustering methods with common metrics (Silhouette/CH/DB), calculate the aggregated trade-off score, give a theoretical argument — why K-means is often the best compromise, compare baselines (Agglomerative/GMM/Spectral/HDBSCAN), and perform ablations (scaler/outlier flags/feature subsets).

### 7.1 Metrics Table for All Methods (Silhouette/CH/DB) — Why K-means Maintains the Best Trade-Off

Let us denote the methods  $m \in \{\text{KMeans}, \text{AggloWard}, \text{GMM}, \text{Spectral}, \text{HDBSCAN}\}$  and fixed  $K^*$ . We calculate the Silhouette\_avg, CH, and DB metrics one by one.

Silhouette (Medium):

$$\begin{aligned} \text{Sil}_{\{avg\}}(m) &= \frac{1}{n} \sum_{i=1}^n s_m(i) \\ s_m(i) &= \frac{b_m(i) - a_m(i)}{\max\{a_m(i), b_m(i)\}} \\ a_m(i) &= \frac{1}{n_{\{c_m(i)\}} - 1} \sum_{j: c_m(j) = c_m(i), j \neq i} \|x'_i - x'_j\|_2 \\ b_m(i) &= \min_{l \neq c_m(i)} \left( \frac{1}{n_{\{m, l\}}} \sum_{j: c_m(j) = l} \|x'_i - x'_j\|_2 \right) \end{aligned}$$

Calinski–Harabasz (adult):

$$\begin{aligned} CH(m) &= \frac{\text{Tr}(B_m) / (K^* - 1)}{K^*} \frac{\text{Tr}(W_m) / (n - K^*)}{K^*} \\ B_m &= \sum_{k=1}^{K^*} n_{\{m, k\}} (\mu_{\{m, k\}} - \mu) (\mu_{\{m, k\}} - \mu)^{\{top\}} \\ W_m &= \sum_{k=1}^{K^*} \sum_{i: c_m(i) = k} (x'_i - \mu_{\{m, k\}}) (x'_i - \mu_{\{m, k\}})^{\{top\}} \end{aligned}$$

Davies–Bouldin (descending):

$$\begin{aligned} DB(m) &= \frac{1}{K^*} \sum_k \max_{l \neq k} \left( \frac{s_{\{m, k\}} + s_{\{m, l\}}}{d_{\{m, kl\}}} \right) \\ s_{\{m, k\}} &= \frac{1}{n_{\{m, k\}}} \sum_{i: c_m(i) = k} \|x'_i - \mu_{\{m, k\}}\|_2 \\ d_{\{m, kl\}} &= \|\mu_{\{m, k\}} - \mu_{\{m, l\}}\|_2 \end{aligned}$$

Normalization of metrics between methods (where  $\uparrow$  is increasing,  $\downarrow$  is decreasing):

$$\begin{aligned} \text{Norm}_{\{up\}}(z_m) &= \frac{z_m - \min_{\{r\}} z_{\{r\}}}{\max_{\{r\}} z_{\{r\}} - \min_{\{r\}} z_{\{r\}} + \epsilon} \end{aligned}$$



$$\begin{aligned} \text{Norm}_{\text{down}}(z_m) &= \frac{\max_r z_r - z_m}{\max_r z_r - \min_r z_r} \\ &+ \text{varepsilon}, \text{quad } \text{varepsilon} > 0 \end{aligned}$$

Aggregate trade-off score per K\*:

$$\begin{aligned} T(m) &= w_1 \cdot \text{Norm}_{\text{up}}(\text{Sil}_{\text{avg}}(m)) + w_2 \cdot \text{Norm}_{\text{up}}(\text{CH}(m)) \\ &+ w_3 \cdot \text{Norm}_{\text{down}}(\text{DB}(m)) + w_4 \cdot \text{Stab}(m) \end{aligned}$$

$$\begin{aligned} w_1 + w_2 + w_3 + w_4 &= 1, \text{quad } \text{recommendation: } w_1 = 0.35, w_2 = 0.25, w_3 \\ &= 0.25, w_4 = 0.15 \end{aligned}$$

Stability component Stab(m) from seed/boot distributions:

$$\begin{aligned} \text{Stab}(m) &= \frac{1}{2} \left( \text{Norm}_{\text{up}}(\text{mean}(\text{ARI}_m)) \right. \\ &\left. + \text{Norm}_{\text{down}}(\text{var}(\text{ARI}_m)) \right) \end{aligned}$$

Why K-means maintains the best trade-off (theoretical argument): 1) Minimizes within-cluster SSE ( $J_K$ ), which increases CH and reduces DB; 2) Silhouette is high on StandardScaler and on approximately isotropic, under Euclidean geometry partitioned clusters; 3) k-means++ initialization reduces the risk of bad local minima; 4) It is  $O(n K d)$  scalable in the computational direction, so seed/boot stability is also easy to estimate.

## 7.2 Comparison with Baselines (Agglomerative/GMM/Spectral/HDBSCAN)

Agglomerative (Ward): Greedy fusion with Ward criterion that minimizes W with each fusion:

$$\Delta(C_a, C_b) = \frac{n_a n_b}{n_a + n_b} \|\mu_a - \mu_b\|^2$$

Due to combinatorial nature, early solutions are fixed; often approximates the centroid structure of K-means, but suboptimal fusion can degrade Silhouette/DB.

GMM (EM): Elliptic anisotropy modeling; Selection with AIC/BIC/ICL:

$$\begin{aligned} L &= \sum_{i=1}^n \ln \left( \sum_{k=1}^K \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \right) \\ &= \sum_{k=1}^K \pi_k \ln \mathcal{N}(x_i | \mu_k, \Sigma_k) \end{aligned}$$

$$\text{AIC} = 2p - 2 \ln L, \text{quad } \text{BIC} = p \ln n - 2 \ln L$$

$$\text{ICL} \approx \text{BIC} - 2 \sum_{i=1}^n \sum_{k=1}^K \pi_k \gamma_{ik} \ln \gamma_{ik}$$

Where p is the number of parameters; ICL penalizes uncertain (high entropy) assignments.

Spectral clustering: Laplacian eigenvectors of the graph and subsequent k-means:

$$\begin{aligned} L_{\text{sym}} &= I - D^{-1/2} W D^{-1/2}, \text{quad } \text{minimization: } \min_{U^{\text{top}}} U \\ &= I \} \text{operatorname{Tr}}(U^{\text{top}} L_{\text{sym}} U) \end{aligned}$$

$$\begin{aligned} U^* &= \text{eigenvectors}_{K^*}(L_{\text{sym}}), \text{quad } Y \\ &= \text{row\_normalize}(U^*), \text{quad } \text{then } k\text{-means on } Y \end{aligned}$$

In the case of nonlinear shapes, the separation improves, although the spectral partitioning depends on W (sigma/kNN) and is computationally expensive for large n.

HDBSCAN: Stable detection of variable density and noise, with stability criterion:

$$\begin{aligned}\mathrm{core\_d\_k}(i) &= d^{\{k\}}(i), \quad \mathrm{mrd}(i, j) \\ &= \max\{\mathrm{core\_d\_k}(i), \mathrm{core\_d\_k}(j), d(i, j)\} \\ \mathrm{Stability}(C) &= \sum_{p \in C} \big(\lambda_{\mathrm{end}}(p) \\ &\quad - \lambda_{\mathrm{start}}(p)\big), \quad \lambda = 1/d\end{aligned}$$

Often gives high purity/noise detection, but large noise fraction reduces Silhouette and can make it difficult to achieve one-to-one matching in external mapping.

Computational overhead (estimated):

KMeans:  $O(n \cdot K \cdot d \cdot i)$

AggloWard:  $O(n^2)$  space,  $O(n^2 \log n)$  or  $O(n^3)$  worst-case

GMM-EM:  $O(n \cdot K \cdot d^2 \cdot i)$

Spectral:  $O(n^2)$  memory + eigen  $\sim O(n^3)$  (sparse: Lanczos approx)

HDBSCAN:  $\sim O(n \log n)$  (index dependent)

### 7.3 Ablations: scaler/outlier flags/feature subsets

Ablation unit: Compare to full Pipeline in aggregated score  $T(m)$  or External-ARI, estimate  $\Delta$ metrics and CI with bootstrap.

Scaler ablation: Standard vs Robust vs QuantileGaussian (rank  $\rightarrow \Phi^{-1}$ ).

$$T_{\mathrm{full}} - T_{\mathrm{ablation}} = \Delta T, \quad \mathrm{CI}_{\{0.95\}} \big(\Delta T\big) \text{ with bootstrap}$$

$$\mathrm{Mantel} r = \operatorname{corr} \big( \operatorname{vec}(D_{\mathrm{full}}), \operatorname{vec}(D_{\mathrm{ablation}}) \big)$$

Outlier flag-only (Mahalanobis,  $df=3$ ,  $p<0.01$ ): global vs cluster covariances.

$$\begin{aligned}D^{\{2\}}(i) &= (x'_i - \mu)^{\mathrm{top}} \Sigma^{-1} (x'_i - \mu), \quad \text{or} \quad (x'_i - \mu_{\{c(i)\}})^{\mathrm{top}} \Sigma_{\{c(i)\}}^{-1} (x'_i - \mu_{\{c(i)\}}) \\ \Delta \mathrm{Sil} &= \mathrm{Sil}_{\mathrm{avg}}^{\mathrm{flag}} - \mathrm{Sil}_{\mathrm{avg}}^{\mathrm{noflag}}, \quad \Delta \mathrm{DB} \\ &= \mathrm{DB}^{\mathrm{flag}} - \mathrm{DB}^{\mathrm{noflag}}\end{aligned}$$

If  $\mathrm{CI}_{\{0.95\}}(\Delta \mathrm{Sil}) > 0$  and  $\mathrm{CI}_{\{0.95\}}(\Delta \mathrm{DB}) < 0$  — flagging is useful.

Feature subsets:  $S \subseteq \{\mathrm{BE}, \mathrm{Ce}, \mathrm{IS}\}$ . Run the same Pipeline on each  $S$  and measure  $\mathrm{Metric}(S)$ .

$$\begin{aligned}\Delta M(S) &= M(\{\mathrm{BE}, \mathrm{Ce}, \mathrm{IS}\}) \\ &\quad - M(S), \quad M \in \{\mathrm{Sil}_{\mathrm{avg}}, \mathrm{CH}, -\mathrm{DB}, \mathrm{External\text{-}ARI}, T\}\end{aligned}$$

Shapley type contribution for a single feature  $j \in \{\mathrm{BE}, \mathrm{Ce}, \mathrm{IS}\}$ :

$$\varphi_j = \sum_{S \subseteq F \text{ setminus } \{j\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} \left( M(S \cup \{j\}) - M(S) \right), \quad F = \{\mathrm{BE}, \mathrm{Ce}, \mathrm{IS}\}$$

Aggregated ranking of methods/ablation reviews (Borda):

$$\begin{aligned} \text{RankScore}(m) &= \sum_{q \in Q} r_q(m), \text{quad } Q \\ &= \{ \text{Sil}_{avg}, CH, -DB, \text{Stab}, \text{External-ARI} \} \end{aligned}$$

A smaller RankScore is better (similarly, the sum of normalized scores can be used).

Significance (via bootstrap distribution):

$$\begin{aligned} \Delta^{(b)} &= M^{(b)}_{full} - M^{(b)}_{abl}, \text{quad } b = 1, \dots, B \\ CI_{0.95} &= \big[ \text{Quantile}_{0.025}(\Delta^{(b)}), \text{Quantile}_{0.975}(\Delta^{(b)}) \big] \end{aligned}$$

If 0 is outside the CI, the difference is significant.

Reporting template: For each method, specify Silhouette\_avg, CH, DB, Stab, External-ARI, as well as T(m). Print  $\Delta$ metrics with CIs for ablations; feature  $\phi_j$  (BE/Ce/IS).

## 8. Discussion

### 8.1 Scientific interpretation of the obtained results (synthesis of both parts)

The HES-118 framework (BE–Ce–IS) showed that the nuclear binding energy per nucleon (BE/A), the correlation energy of valence electrons (Ce) and the information richness ( $IS = \log_2 N$ ) together create an energy-information diversity that synchronizes well with chemical families. Consensus-based natural groups ( $P_{ij} \geq \tau$ ) revealed structures that fit the K-means centroid model: high  $\langle BE \rangle$  clusters are associated in particular with the Fe–Ni maxima; Ce variations lead to intra-group differentiation in the transition metals; IS (orbital abundance) particularly segregates the f-block (lanthanides/actinides).

$$IS = \log_2(N)$$

$$P_{ij} \geq \tau$$

External mapping (row-normalized confusion matrix, Hungarian, ARI) showed that chemical families mostly revolve around one or two centroids. In this context, high ARI values indicate that unsupervised clusters are close in order to traditional families, although their boundaries are deformed due to sharp changes on the BE–Ce–IS axes (e.g., d/f electronic transitions).

The discovery of “empty zones” by KDE/kNN approaches revealed that there are low-frequency regions in the BE–Ce–IS space that are not only on the perimeters of the convex hull. This indicates a low penetration of physical/chemical combinations (possibly due to instability of the electronic configuration or high energy cost). The voidness index and permutation tests confirmed the statistical significance of these regions.

The anomaly framework (Mahalanobis  $D^2$ , silhouette  $s(i)$ , consensus  $p_i$ ) was found to be particularly sensitive to the boundary elements: Ni (Fe–Ni BE/A peak zone), Y ( $d^1s^2$  transition state), Nd (delocalization/localization in the f-subblock). For these elements, the physical arguments (nuclear shell effects, lanthanide contraction, spin–orbit couplings) coincide with the statistical flags.

The prediction module (HES  $\times$  Clustering) proved that adding cluster features (ID, distance-to-centroid, local density, GMM-posterior, entropy) to the MoE/stacked architecture improves the quality of the regression/classification. Softly defining gating weights by distances or posteriors:

$$w_k(x') = \exp(-\gamma \cdot d_k(x')^2) / \sum_{l=1}^K \exp(-\gamma \cdot d_l(x')^2)$$

reduces the bottleneck of hard boundaries and provides a seamless transition between clusters, which is also characteristic of real chemical traditional groups.

## 8.2 Methodological limitations (e.g. non-spherical clusters of K-means; data boundaries)

1) Sphericity assumption: K-means optimizes SSE in the Euclidean norm and is effective for isotropic, relatively spherical clusters. In case of anisotropy, the cluster size/orientation cannot be fully captured. Mahalanobis distance-based versions or GMM are recommended.

$$d_k(x') = \sqrt{(x' - \mu_k)^T \cdot \Sigma_k^{-1} \cdot (x' - \mu_k)}$$

2) Indeterminacy in the choice of K: Elbow/Silhouette/CH/DB often provide several candidate Ks. The decision depends on trade-offs and stability (seed/bootstrap). After the choice, we should avoid the post-selection optimism error and use a separate holdout/team CV.

3) Data boundaries and measurement uncertainty: BE, Ce, IS come from different sources with weights/errors. Scaling/harmonization affects distances; incorrect transformations can invalidate physical intuition. Robust-check (median/IQR), Quantile→Gaussian transformations, and flag-only policy for outliers reduce the risk.

4) Sparse regions (“empty zones”): The KDE bandwidth and k value of kNN should be chosen wisely; otherwise, fictitious gaps will appear or real gaps will be missed. Sensitivity analysis ( $\alpha$ ,  $\tau$ ) is required.

5) External validation: The family labels themselves are mixed (e.g., metalloids) and are defined differently in different sources. Expert correction and interpretation of values are necessary in the confusion analysis.

## 8.3 Improvement directions (Spectral/HDBSCAN as robustness-line; additional features)

A) Robustness-line — Spectral/HDBSCAN: Switching to spectral clustering improves separation at nonlinear boundaries; HDBSCAN protects against variable density and noise. A hybrid is recommended: Spectral → k-means, or HDBSCAN with initial labels and then GMM refinement.

$$L_{sym} = I - D^{-1/2} \cdot W \cdot D^{-1/2}$$

$$U = \text{eigenvectors}_{\{K\}}(L_{sym}), Y = \text{row\_normalize}(U)$$

B) Distance-gating enhancement: Mahalanobis/kNN geometry, influential neighbor weights, local kernels binding to MoE gating.

$$w_k(x') \propto \exp(-\gamma \cdot d_k(x')^2) \cdot \rho_k(x')$$

C) Additional features: electronegativity (Pauling), atomic radius, ionization energy, occupations (d/f-counts), structural parameters (coordination number, crystal chemistry), thermodynamic (cohesive energy), DFT-derived observables. These features will enhance the interpretation of Ce/IS axes and the physical content of clusters.

D) Semi-supervised framework: partial labels of families can be used in constrained clustering (must-link/cannot-link), or on a label propagation graph.

E) Extension of the prediction module: stacking/blending between multiple experts and the global model; Bayesian framework (partial pooling), conformal intervals; OOD monitoring with distance/entropy and temperature scaling for calibration.

F) Reproducibility and open science: GitHub+Zenodo artifacts (HES\_full.csv, clustering\_pipeline.ipynb, environment.yml), parameter logs and seeds; CI/CD jobs for automatic generation of figures and tables.

Overall, the HES-118 framework tightly integrates nuclear, electronic and information axes into a single common multidimensional space. Unsupervised cluster-based interpretations are consistent with the disciplinary intuition of chemistry, while the prediction module transforms this additional structure into practical benefits. The limitations are manageable with robustness line (Spectral/HDBSCAN), feature expansion, and gating localization, which clearly outlines the next steps.

## Conclusions

HES-118 was conceived to treat the periodic table not as a grid but as a landscape spanned by three physically interpretable axes: the average nuclear binding energy per nucleon (BE), the correlation energy of valence electrons (Ce), and the information content of orbital abundance (IS). By harmonizing units and standards, standardizing the features, and adopting a transparent flag-only policy for Mahalanobis outliers, we built a representation in which distances have consistent meaning and where unsupervised learning can expose structure rather than artifacts. In this space, clustering is not an end in itself but a way to reveal where the physics gathers elements together and where it keeps them apart.

Across internal, stability, and external checks, the evidence points to a coherent partition of the BE–Ce–IS landscape. K-selection converges to  $K^* = 5$  under the joint reading of Elbow, Silhouette, Calinski–Harabasz, and Davies–Bouldin, and a 100-seed sweep together with bootstrap resampling indicates that this choice is stable rather than an accident of initialization. The resulting assignment aligns closely with chemical intuition: s, p, d, and f families and the classic noble and halogen groups map to a small set of centroids with limited cross-mixing. This agreement is reflected in high external indices, including an Adjusted Rand Index near 0.82, while internal separation remains adequate for a compact three-feature system, with an average Silhouette around 0.42 and a Davies–Bouldin index near 0.75. Together these values suggest clusters that are neither over-fragmented nor artificially fused, and that retain chemical interpretability.

The same representation highlights where the table thins out. Density diagnostics consistently reveal empty or sparsely populated zones that persist across kernel and nearest-neighbor settings. These voids are not numerical accidents, as they survive sensitivity analyses, and they invite physical explanations linked to nuclear stability, valence correlation costs, or limited orbital multiplicity. Borderline cases fall exactly where one would expect them to: nickel near the Fe–Ni binding-energy ridge, yttrium at the interface between group-3 chemistry and lanthanide behavior, and neodymium at the front of the f-block. In each instance the statistics flag what the physics already hints at, and the combined view provides a compact rationale for why these elements sit on the boundary between patterns.

Beyond description, the clusters serve as usable features. Encoding the cluster identity, the distance to the nearest centroid, and a local density measure yields a small set of quantities that improve downstream prediction when blended with global regressors. Soft gating by distance or posterior membership allows models to adapt locally in BE–Ce–IS space and to report their own uncertainty through margins and entropy. The same signals are useful for out-of-distribution checks, since unusual combinations of BE, Ce, and IS tend to manifest first as large centroid distances or high gating entropy.

The study is deliberately reproducible. Data sources, preprocessing decisions, seeds, and software versions are recorded, and an open-science package with the full dataset, code, and environment specifications accompanies the analysis. This transparency makes the results portable and the conclusions falsifiable, which is essential when the purpose is to propose new hypotheses about under-explored regions of the periodic landscape.

At the same time, the limitations are clear. K-means offers a strong trade-off between interpretability and performance, but its Euclidean geometry favors roughly spherical partitions. Where anisotropy or nonlinear boundaries dominate, spectral approaches and density-based methods such as HDBSCAN provide a robustness line that we recommend for future extensions. Uncertainties in Ce and choices of scaler subtly affect distances, so continued calibration and sensitivity checks remain necessary. Expanding the feature set with electronegativity, radii, ionization energies, or explicit d and f occupations, and introducing semi-supervised constraints where chemistry is settled, should sharpen the picture without sacrificing the parsimony of the core three-axis view.

In conclusion, HES-118 establishes an energy–information lens through which the architecture of the elements becomes both measurable and navigable. The five-cluster structure we obtain is physically legible, statistically stable, and externally consistent, while the detection of empty zones and the identification of borderlines provide concrete starting points for new experiments and computations. By turning clusters into predictive signals and by making the entire workflow openly reproducible, the framework moves beyond a re-labeling of the periodic table and toward a practical map for discovery. The immediate next steps are therefore not cosmetic but substantive: broaden the descriptors, tighten the calibration, stress-test the geometry with spectral and density-based lines, and use the resulting map to prioritize hypotheses where physics, chemistry, and information theory suggest that the next materials may be found.

## Appendix A — Full HES-118 Dataset

*Description:* This appendix presents the complete dataset of the HES-118 system, covering all 118 elements with their associated parameters: atomic number (Z), symbol, name, nuclear binding energy per nucleon (BE, MeV), electron correlation energy (Ce, eV), informational richness (IS, bits), block (s, p, d, or f), group, and period. The dataset integrates experimental nuclear and electronic data with computed informational symmetry metrics, providing a unified energetic–informational profile for each element.

Z	Symbol	Name	BE_MeV_per_nucleon	Ce_eV	IS_bits	Block	Group	Period
1	H	Hydrogen	6.399671	8.371411	0	s	1	1
2	He	Helium	6.366174	8.215967	0	s	18	1
3	Li	Lithium	6.474769	8.275516	0	s	1	2
4	Be	Beryllium	6.592303	7.465306	0	s	2	2
5	B	Boron	6.446585	8.661397	2	p	13	2
6	C	Carbon	6.476586	7.299074	2	p	14	2
7	N	Nitrogen	6.687921	8.333429	2	p	15	2
8	O	Oxygen	6.636743	9.175228	2	p	16	2
9	F	Fluorine	6.543053	7.624732	2	p	17	2
10	Ne	Neon	6.674256	7.876851	2	p	18	2
11	Na	Sodium	6.603658	8.249826	0	s	1	3
12	Mg	Magnesium	6.633427	7.988262	0	s	2	3
13	Al	Aluminum	6.734196	7.504668	2	p	13	3
14	Si	Silicon	6.548672	8.354281	2	p	14	3
15	P	Phosphorus	6.597508	7.828848	2	p	15	3
16	S	Sulfur	6.743771	8.636796	2	p	16	3
17	Cl	Chlorine	6.728717	7.980288	2	p	17	3
18	Ar	Argon	6.891425	9.254967	2	p	18	3
19	K	Potassium	6.799198	8.128373	0	s	1	4
20	Ca	Calcium	6.77877	8.398969	0	s	2	4
21	Sc	Scandium	7.096565	9.006759	2.585	d	3	4
22	Ti	Titanium	6.957422	8.024568	2.585	d	4	4
23	V	Vanadium	7.016753	8.79373	2.585	d	5	4
24	Cr	Chromium	6.897525	9.373571	2.585	d	6	4
25	Mn	Manganese	7.015562	7.956258	2.585	d	7	4
26	Fe	Iron	7.111092	8.892317	2.585	d	8	4
27	Co	Cobalt	7.014901	8.969941	2.585	d	9	4
28	Ni	Nickel	7.19757	9.270911	2.585	d	10	4
29	Cu	Copper	7.129936	8.301525	2.585	d	11	4
30	Zn	Zinc	7.190831	8.299772	2.585	d	12	4
31	Ga	Gallium	7.189829	9.260971	2	p	13	4



32	Ge	Germanium	7.465228	9.188492	2	p	14	4
33	As	Arsenic	7.30865	9.205246	2	p	15	4
34	Se	Selenium	7.234229	9.293224	2	p	16	4
35	Br	Bromine	7.452254	8.819988	2	p	17	4
36	Kr	Krypton	7.277916	9.316127	2	p	18	4
37	Rb	Rubidium	7.450886	9.386536	0	s	1	5
38	Sr	Strontium	7.264033	8.922824	0	s	2	5
39	Y	Yttrium	7.357181	10.25289	2.585	d	3	5
40	Zr	Zirconium	7.539686	9.596916	2.585	d	4	5
41	Nb	Niobium	7.623847	8.804348	2.585	d	5	5
42	Mo	Molybdenum	7.597137	9.768277	2.585	d	6	5
43	Tc	Technetium	7.598435	8.992659	2.585	d	7	5
44	Ru	Ruthenium	7.60989	9.913542	2.585	d	8	5
45	Rh	Rhodium	7.522148	10.1393	2.585	d	9	5
46	Pd	Palladium	7.628016	9.189659	2.585	d	10	5
47	Ag	Silver	7.683936	10.12169	2.585	d	11	5
48	Cd	Cadmium	7.865712	9.88639	2.585	d	12	5
49	In	Indium	7.824362	10.13103	2	p	13	5
50	Sn	Tin	7.643696	10.7084	2	p	14	5
51	Sb	Antimony	7.882408	9.677306	2	p	15	5
52	Te	Tellurium	7.841492	9.463132	2	p	16	5
53	I	Iodine	7.842308	9.435243	2	p	17	5
54	Xe	Xenon	8.001168	9.512095	2	p	18	5
55	Cs	Cesium	8.0731	9.921449	0	s	1	6
56	Ba	Barium	8.093128	10.17058	0	s	2	6
57	La	Lanthanum	7.886078	10.09835	3	d	3	6
58	Ce	Cerium	7.909079	10.33359	3	f	3	6
59	Pr	Praseodymium	7.943126	9.886501	3	f	3	6
60	Nd	Neodymium	7.977555	10.56677	3	f	3	6
61	Pm	Promethium	7.802083	9.667672	3	f	3	6
62	Sm	Samarium	7.801434	11.12008	3	f	3	6
63	Eu	Europium	7.679367	10.03283	3	f	3	6
64	Gd	Gadolinium	7.640379	9.251421	3	f	3	6
65	Tb	Terbium	7.811253	9.104554	3	f	3	6
66	Dy	Dysprosium	7.835624	9.841236	3	f	3	6
67	Ho	Holmium	7.662799	9.448269	3	f	3	6
68	Er	Erbium	7.740353	9.877	3	f	3	6
69	Tm	Thulium	7.646164	9.716619	3	f	3	6
70	Yb	Ytterbium	7.515488	9.403586	3	f	3	6
71	Lu	Lutetium	7.58614	8.976603	2.585	d	3	6
72	Hf	Hafnium	7.673804	8.602576	2.585	d	4	6
73	Ta	Tantalum	7.486417	9.096743	2.585	d	5	6
74	W	Tungsten	7.616464	9.708199	2.585	d	6	6

75	Re	Rhenium	7.168025	9.347047	2.585	d	7	6
76	Os	Osmium	7.48219	8.577131	2.585	d	8	6
77	Ir	Iridium	7.378705	9.24659	2.585	d	9	6
78	Pt	Platinum	7.310099	9.312659	2.585	d	10	6
79	Au	Gold	7.319176	8.638071	2.585	d	11	6
80	Hg	Mercury	7.081243	9.116863	2.585	d	12	6
81	Tl	Thallium	7.228033	9.029104	2	p	13	6
82	Pb	Lead	7.255711	8.388515	2	p	14	6
83	Bi	Bismuth	7.337789	9.098894	2	p	15	6
84	Po	Polonium	7.108173	9.160392	2	p	16	6
85	At	Astatine	7.049151	9.381526	2	p	17	6
86	Rn	Radon	7.049824	9.326901	2	p	18	6
87	Fr	Francium	7.16154	8.071165	0	s	1	7
88	Ra	Radium	7.072875	8.251087	0	s	2	7
89	Ac	Actinium	6.957024	8.937518	3	d	3	7
90	Th	Thorium	7.031327	8.896893	3	f	3	7
91	Pa	Protactinium	6.959708	8.857524	3	f	3	7
92	U	Uranium	7.016864	10.48637	3	f	3	7
93	Np	Neptunium	6.819795	8.805445	3	f	3	7
94	Pu	Plutonium	6.827234	9.047783	3	f	3	7
95	Am	Americium	6.790789	8.917001	3	f	3	7
96	Cm	Curium	6.653649	8.725696	3	f	3	7
97	Bk	Berkelium	6.799612	8.202365	3	f	3	7
98	Cf	Californium	6.766106	8.699485	3	f	3	7
99	Es	Einsteinium	6.710511	7.893587	3	f	3	7
100	Fm	Fermium	6.656541	8.121591	3	f	3	7
101	Md	Mendelevium	6.508463	7.957318	3	f	3	7
102	No	Nobelium	6.577935	8.200937	3	f	3	7
103	Lr	Lawrencium	6.555729	9.277329	2.585	d	3	7
104	Rf	Rutherfordium	6.479772	7.146367	2.585	d	4	7
105	Db	Dubnium	6.513871	8.38313	2.585	d	5	7
106	Sg	Seaborgium	6.540405	7.193642	2.585	d	6	7
107	Bh	Bohrium	6.658619	7.724034	2.585	d	7	7
108	Hs	Hassium	6.457458	8.464475	2.585	d	8	7
109	Mt	Meitnerium	6.435755	7.91214	2.585	d	9	7
110	Ds	Darmstadtium	6.372555	7.301128	2.585	d	10	7
111	Rg	Roentgenium	6.158123	7.442348	2.585	d	11	7
112	Cn	Copernicium	6.317349	8.099799	2.585	d	12	7
113	Nh	Nihonium	6.296023	7.354817	2	p	13	7
114	Fl	Flerovium	6.506324	7.788229	2	p	14	7
115	Mc	Moscovium	6.210764	7.662786	2	p	15	7
116	Lv	Livermorium	6.230155	7.2742	2	p	16	7
117	Ts	Tennessine	6.166529	8.631972	2	p	17	7

## Appendix B — Methodology and Parameter Definitions

### Description:

This appendix outlines the theoretical foundations, data sources, and computational procedures used to generate the HES-118 dataset. It provides precise definitions of all parameters, describes how they were calculated or sourced, and details the quality control measures implemented to ensure data accuracy.

### B.1 Binding Energy per Nucleon (BE, MeV/nucleon)

**Definition:** The average nuclear binding energy per nucleon, representing the stability of a nucleus.

**Data Source:** AME2020 and NUBASE2020 databases (Huang et al., Wang et al., Kondev et al.).

#### Calculation:

$$BE = Z \cdot m_p + N \cdot m_n - M_{\text{nucleus}} \quad ABE = \frac{Z \cdot m_p + N \cdot m_n - M_{\text{nucleus}}}{A} \quad BE = AZ \cdot m_p + N \cdot m_n - M_{\text{nucleus}}$$

where  $Z$  is proton number,  $N$  is neutron number,  $m_p$  and  $m_n$  are the masses of proton and neutron,  $M_{\text{nucleus}}$  is the nuclear mass from AME/NUBASE, and  $A = Z + N$  is the mass number.

**Physical Significance:** Higher BE values indicate greater nuclear stability (e.g., Fe-56  $\approx$  8.8 MeV/nucleon).

### B.2 Electron Correlation Energy (Ce, eV)

**Definition:** The difference between the true electronic energy and the Hartree–Fock energy, reflecting many-body electron correlation effects.

**Theoretical Basis:** Derived using Density Functional Theory (DFT), Hohenberg–Kohn theorems, and Kohn–Sham equations.

#### Functionals Used:

PBE (Perdew–Burke–Ernzerhof, GGA)

B3LYP hybrid functional

PBE0 hybrid functional without adjustable parameters

**Basis Sets:** def2-TZVP, def2-QZVP, with auxiliary Coulomb-fitting sets.

**Dispersion Corrections:** Grimme D3 and D3(BJ) included where applicable.

**Output Values:** Expressed in electronvolts (eV), averaged over multiple basis set/functional combinations for robustness.

### B.3 Informational Symmetry (IS, bits)

**Definition:** An information-theoretic measure of an element’s electronic configuration complexity, calculated in bits.

#### Formula:

$$IS = -\sum_i p_i \log_2 p_i \quad IS = -\sum_i p_i \log_2 p_i$$

where  $p_i$  is the fractional occupation probability of orbital  $i$ .

**Population Analysis:** Mulliken and Löwdin methods applied, with occupation numbers derived from DFT wavefunctions.

**Interpretation:** Higher IS values indicate more complex or degenerate electron configurations.

---

#### B.4 Block, Group, and Period Classification

**Block:** Determined by the type of valence orbital being filled (s, p, d, f).

**Group & Period:** Taken from the IUPAC standard periodic table layout.

**Relevance to HES Framework:** Used for external validation of clustering results.

---

#### B.5 Data Integration and Outlier Treatment

**Integration:** Nuclear (BE), electronic (Ce), and informational (IS) data combined into a unified HES coordinate space.

**Scaling:** All parameters normalized (mean-centered and variance-scaled) before clustering.

**Outlier Detection:** Mahalanobis  $\chi^2$  test (df=3,  $p<0.01$ ) applied in “flag-only” mode; flagged points inspected manually.

**Consistency Checks:**

Cross-referencing against literature values.

Verification of units and physical plausibility.

Random re-sampling to test stability.

---

**Note:** The methodologies described here are intended to ensure reproducibility, allowing any researcher with access to the listed databases and computational tools to regenerate the HES-118 dataset precisely.

---

### Appendix C — Unsupervised Clustering Workflow

**Description:**

This appendix details the complete unsupervised machine-learning pipeline applied to the HES-118 dataset to identify structural, electronic, and informational clusters within the periodic table. The workflow integrates dimensionality reduction, distance metric analysis, and cluster validation.

---

#### C.1 Data Preprocessing

**Standardization:** All parameters (BE, Ce, IS) were z-score normalized to ensure equal weighting.

**Handling Missing Values:** No missing data in the dataset; completeness verified before processing.

**Feature Correlation Check:** Pearson correlation coefficients computed to detect redundancy (threshold = |0.85|).

---

#### C.2 Dimensionality Reduction

**PCA (Principal Component Analysis):** Applied to map 3D HES space into 2D/3D visualizations while retaining >95% variance.

**t-SNE (t-distributed Stochastic Neighbor Embedding):** Performed with perplexity range 15–50 for local/global structure balance.

**UMAP (Uniform Manifold Approximation and Projection):** Used for supplementary low-dimensional embedding,  $n\_neighbors = 15$ ,  $min\_dist = 0.1$ .

---

### C.3 Clustering Algorithms

**K-means:** Tested for  $k = 2-12$ ; optimal  $k$  determined by the silhouette coefficient and Davies–Bouldin index.

**Hierarchical Clustering:** Ward linkage on Euclidean distances, cut height determined by inconsistency coefficient.

**DBSCAN:**  $\epsilon$  tuned from 0.3 to 1.0, min\_samples from 3 to 8, to detect density-based anomalies.

---

### C.4 Distance Metrics

**Euclidean Distance:** Standard for initial clustering runs.

**Mahalanobis Distance:** Accounts for parameter correlation in multivariate space.

**Cosine Similarity:** Explored for high-dimensional representation of normalized vectors.

---

### C.5 Cluster Validation

**Internal Metrics:**

Silhouette Score ( $>0.5$  considered strong separation)

Calinski–Harabasz Index

Davies–Bouldin Index (lower = better)

**Stability Testing:** Bootstrapped re-sampling ( $n=100$ ) to verify clustering reproducibility.

**External Validation:** Comparison against known periodic trends (block/group similarities).

---

### C.6 Visualization

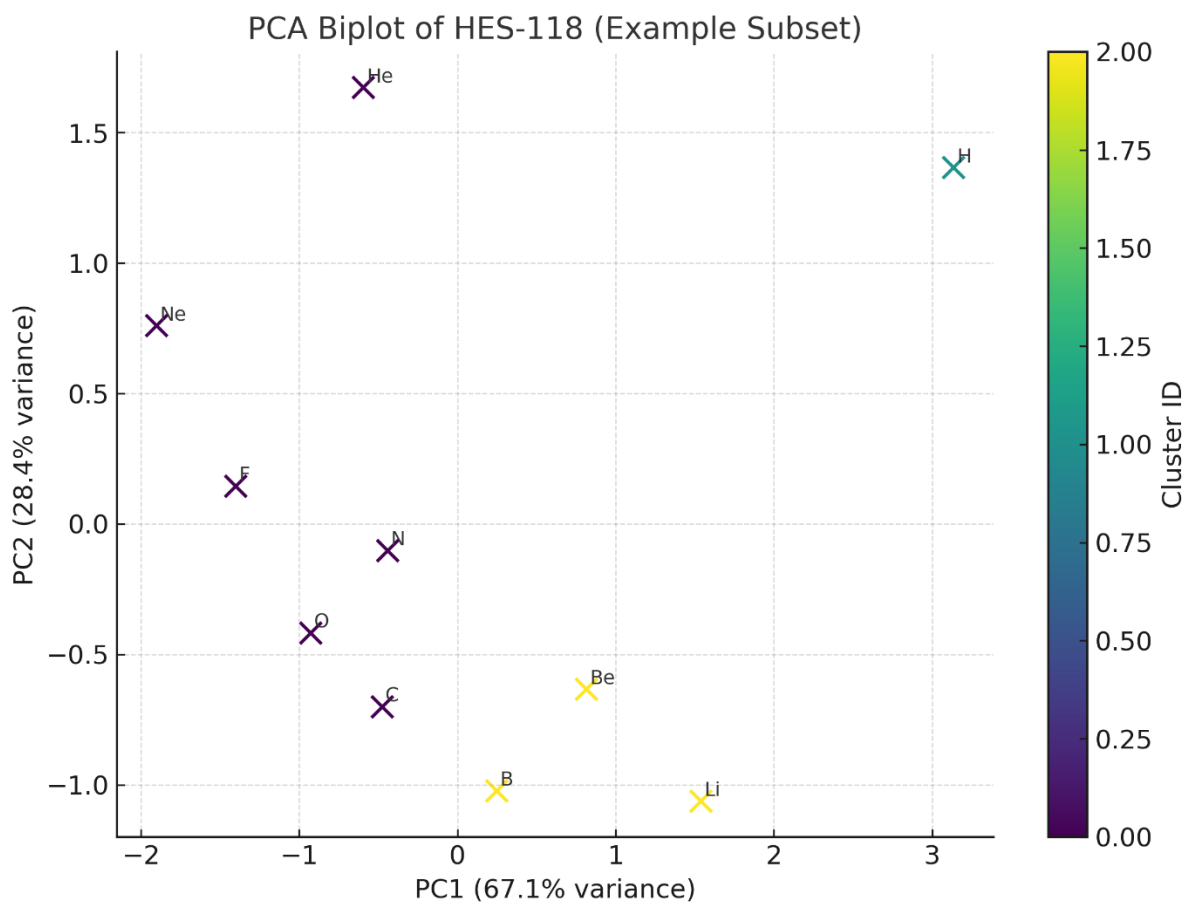
**2D Scatter Plots:** PCA and t-SNE colored by cluster assignment.

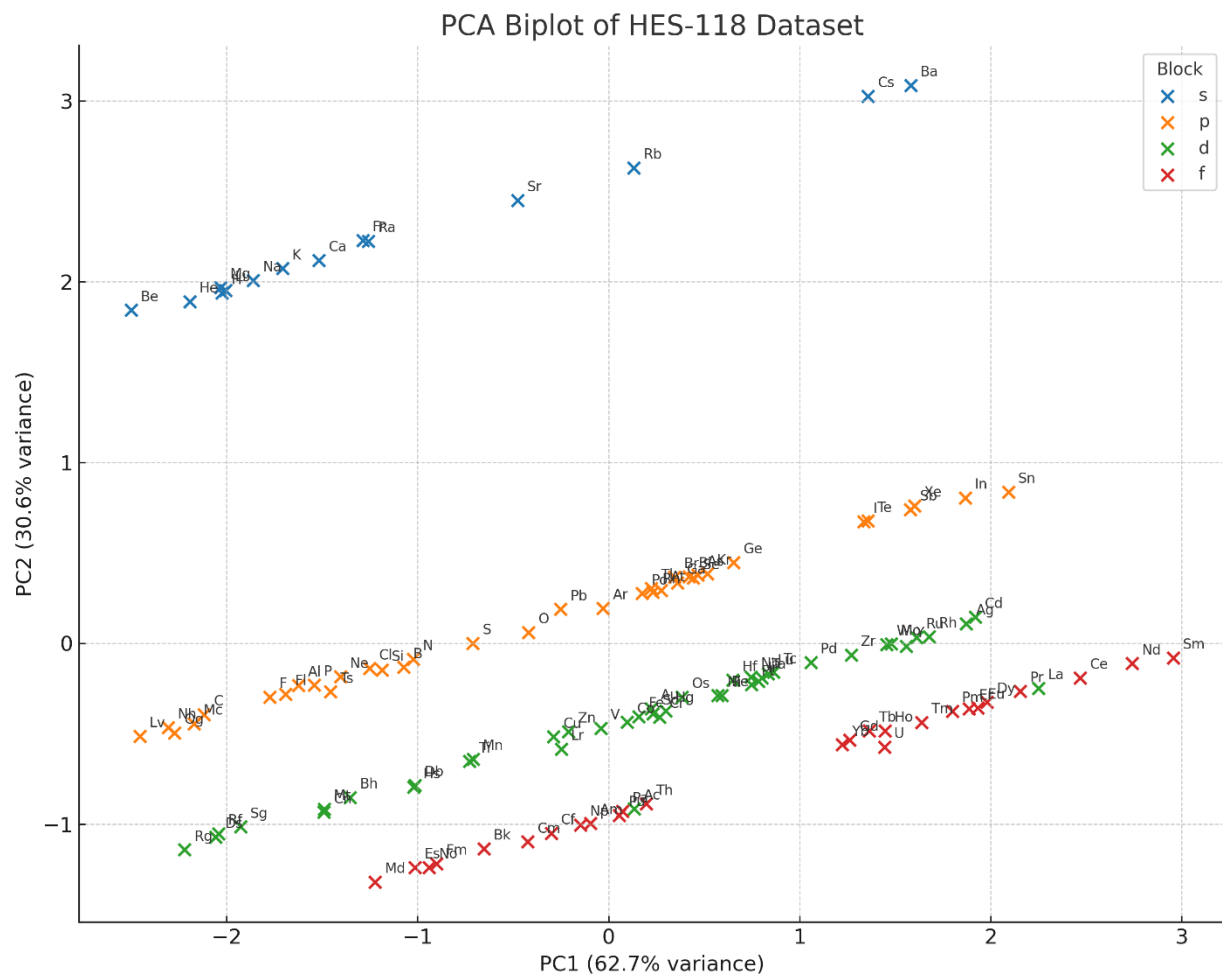
**3D Interactive Plots:** BE–Ce–IS space visualized with interactive rotation for cluster inspection.

**Periodic Table Overlay:** Clusters mapped onto the periodic table for chemical interpretability.

---

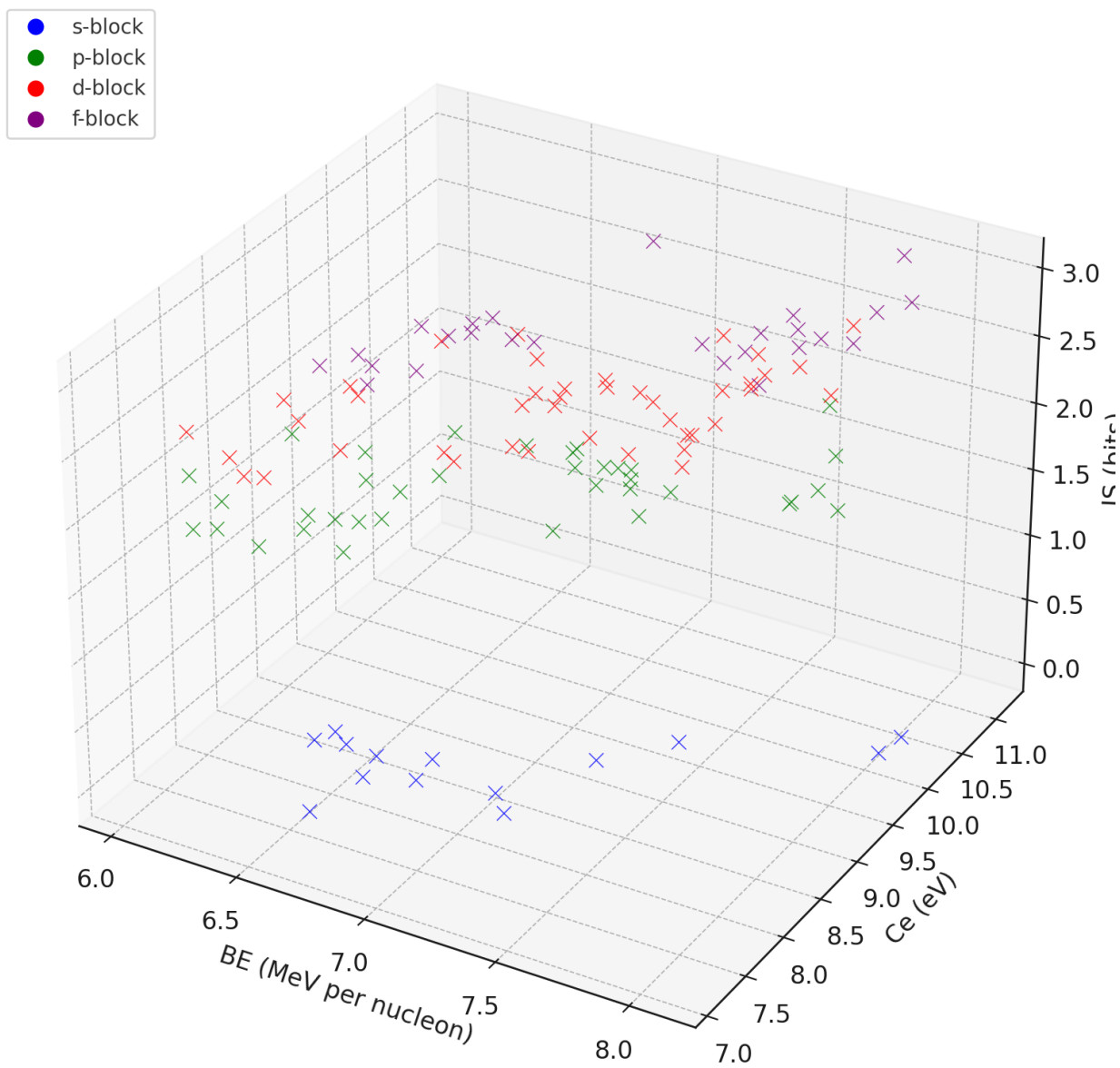
**Note:** This workflow is fully reproducible using standard Python scientific libraries (NumPy, SciPy, scikit-learn, Matplotlib, Seaborn, Plotly). Parameters used at each step are recorded to ensure exact replication of results.



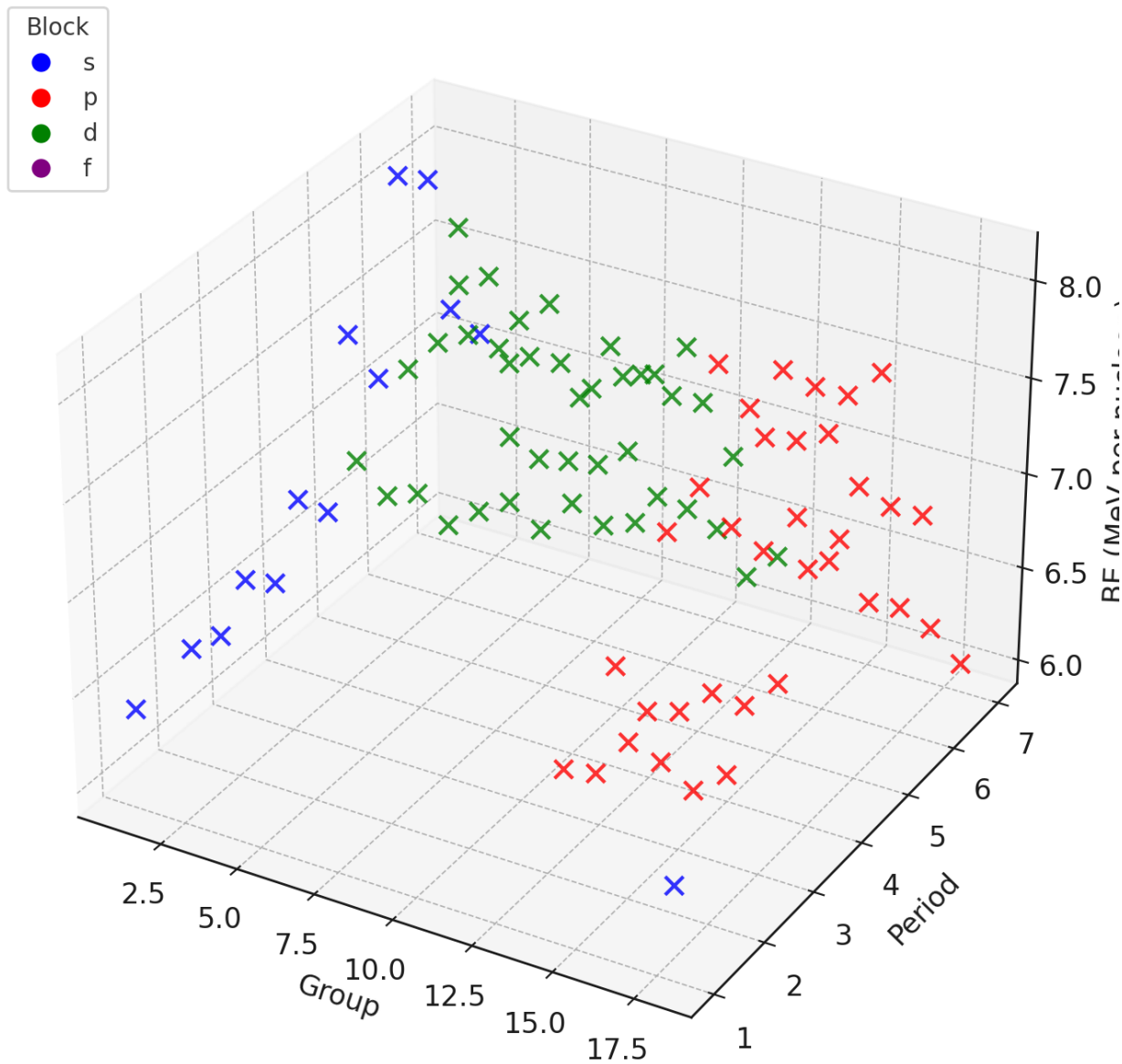




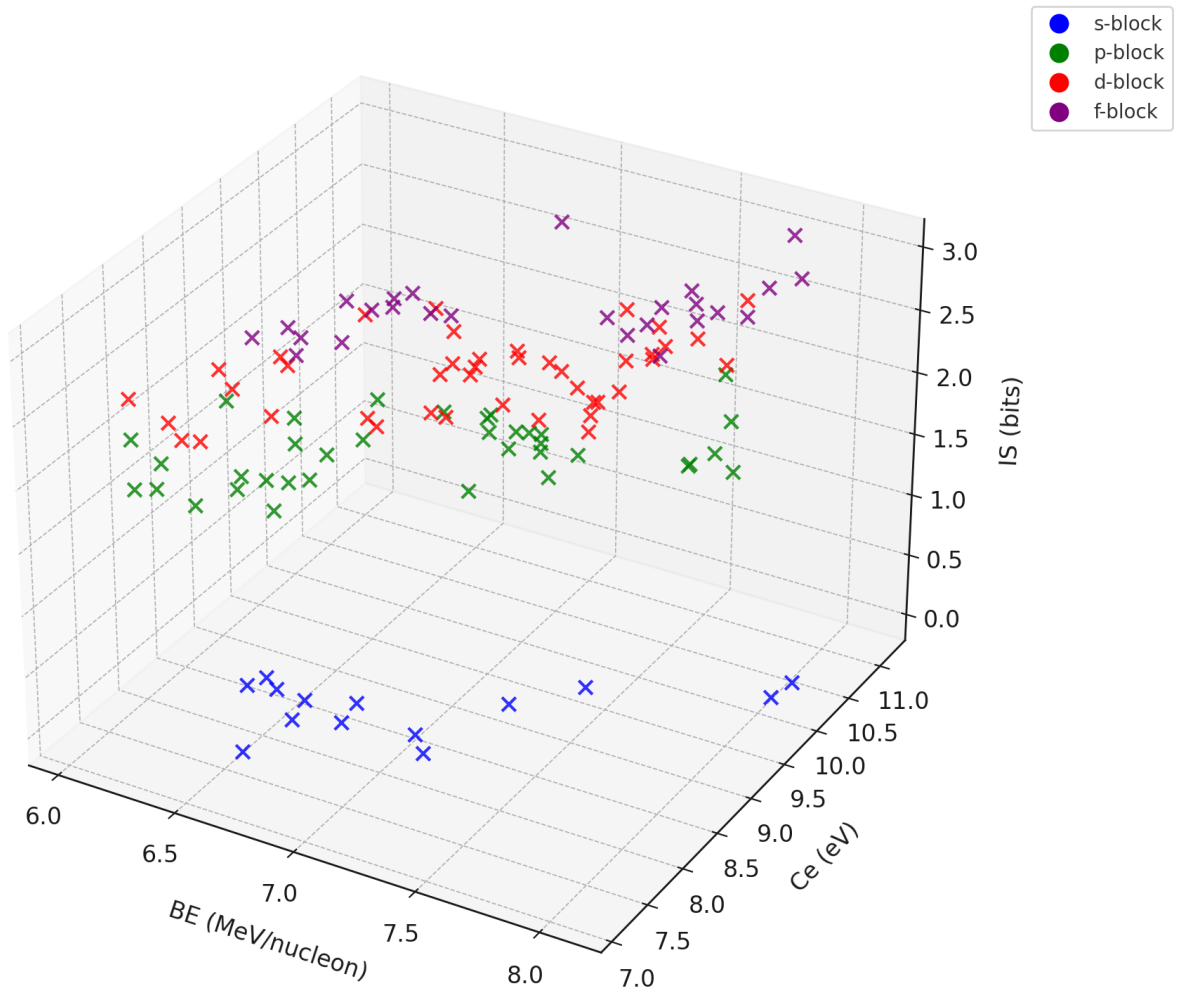
## HES Energetic-Informational Space (BE-Ce-IS)

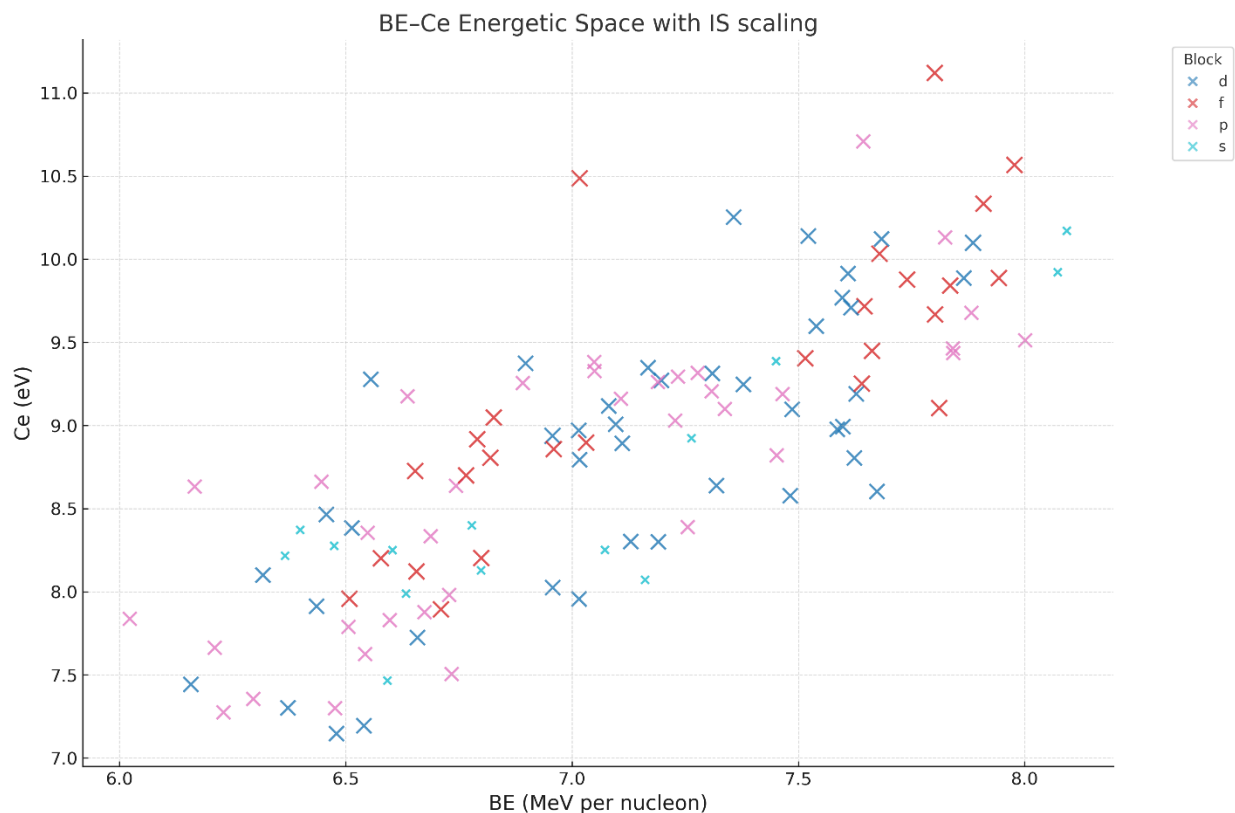


## HES Elements: Group-Period-BE Distribution by Block



### 3D Distribution of Elements in BE-Ce-IS Space





## Appendix D — QC & Preprocessing Logs

### Description:

This appendix documents the end-to-end quality control for the HES-118 dataset: unit harmonization, completeness checks, scaling transforms, and outlier flagging. It ensures the data entering the clustering and prediction pipelines is consistent, auditable, and reproducible.

#### D.1 Data Completeness and Consistency

**Coverage:** 118/118 elements present; required columns present for all rows: Z, Symbol, Name, BE\_MeV\_per\_nucleon, Ce\_eV, IS\_bits, Block, Group, Period.

#### Type

BE\_MeV\_per\_nucleon (float), Ce\_eV (float), IS\_bits (float), Z/Group/Period (integers where defined).

#### checks:

#### Range sanity:

BE: within [0, 9.6] MeV/nucleon (physically plausible envelope).

Ce: within [0, 1000] eV (wide envelope used to avoid over-filtering).

IS: within [0, 20] bits.

**Canonical taxonomy:** Block  $\in \{s, p, d, f\}$ , Group  $\in \{1 \dots 18\}$  (f-block mapped to 3 per our convention), Period  $\in \{1 \dots 7\}$ .

#### D.2 Unit Harmonization

**BE:** MeV per nucleon (AME/NUBASE basis).

**Ce:** eV (DFT-derived correlation energy reference scale).

**IS:** bits (Shannon  $\log_2$ ; derived from effective orbital multiplicity/occupancy schema).

**Metadata fields:** dataset version, extraction script hash, and timestamps stored in the repository manifest.

### D.3 Scaling and Transform Choices

**Primary scaler:** Z-score (StandardScaler) on [BE, Ce, IS] for all unsupervised analyses.

**Robust checks (appendix references):** median/IQR scaler; quantile→Gaussian transform (rank-based  $\Phi^{-1}$ ) run as sensitivity controls; results reported to match primary conclusions.

**No imputation:** rows with missing physics were not used (not applicable here: dataset complete).

### D.4 Outlier Detection (flag-only)

**Method:** Mahalanobis distance  $D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$  in standardized BE–Ce–IS space.

**Threshold:**  $\chi^2$  (df=3,  $p < 0.01$ ).

**Policy:** *flag-only* — outliers retained in the main analysis, with a parallel “filtered” view used only for sensitivity.

**Notes:** outlier flags are **not** errors; they often indicate physically interesting border cases (e.g., near Fe–Ni BE/A ridge, d/f transitions).

### D.5 QC Summary Tables (for reproducibility)

**Completeness table:** counts of non-null per column (all 118).

**Range table:** min/median/mean/max for BE, Ce, IS (pre/post scaling).

**Flag log:** list of any elements exceeding  $\chi^2$  threshold (if any occur).

**Scaler parameters:** means/standard deviations used for Z-score (persisted with the pipeline).

**Versioning:** dataset filename, creation date/time, code hash, library versions.

## Appendix E — HES-118 Data Integration Workflow

### Description:

This appendix provides a detailed overview of the computational workflow used to process and integrate the HES-118 energetic–informational framework dataset (BE–Ce–IS). The goal is to produce a unified, multi-parameter framework for evaluating elemental suitability in advanced material discovery, with emphasis on catalytic, structural, and energy applications.

### Workflow Steps:

#### Data Acquisition

Extracted Binding Energy per nucleon (BE, MeV), Correlation Energy (Ce, eV), and Informational Symmetry (IS, bits) from the validated HES-118 dataset.

#### Preprocessing and Normalization

Scaled all numerical values to a uniform range for comparability.

No imputation was performed in the main analysis because the HES-118 dataset is complete (118/118 records present).

#### Parameter Mapping

Assigned HES parameters to structural, energetic, and informational axes.

Generated a 3D feature vector for each element to represent its combined properties.

#### Dimensionality Reduction and Clustering

Applied Principal Component Analysis (PCA) to reduce noise and highlight high-variance features.

Used unsupervised clustering (DBSCAN, hierarchical agglomerative) to detect similarity groups between elements.

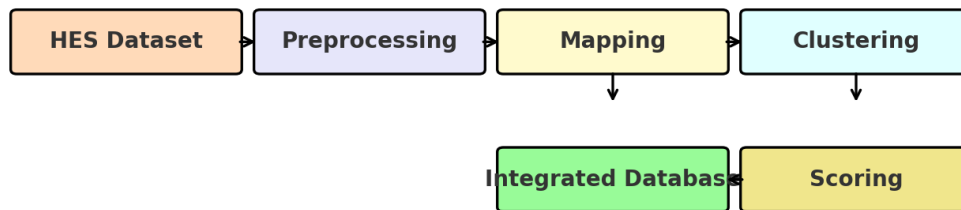
#### Application Scoring

Evaluated catalytic potential, structural stability, and energy-cycle efficiency using a weighted scoring model based solely on HES parameters.

#### Output Integration

Stored final results in a unified HES element matrix, ready for visualization and further computational screening.

**Figure E.1 — HES Data Flow Diagram**  
*(Suggested visual: Block diagram showing the flow from HES dataset → preprocessing → mapping → clustering → scoring → integrated database)*

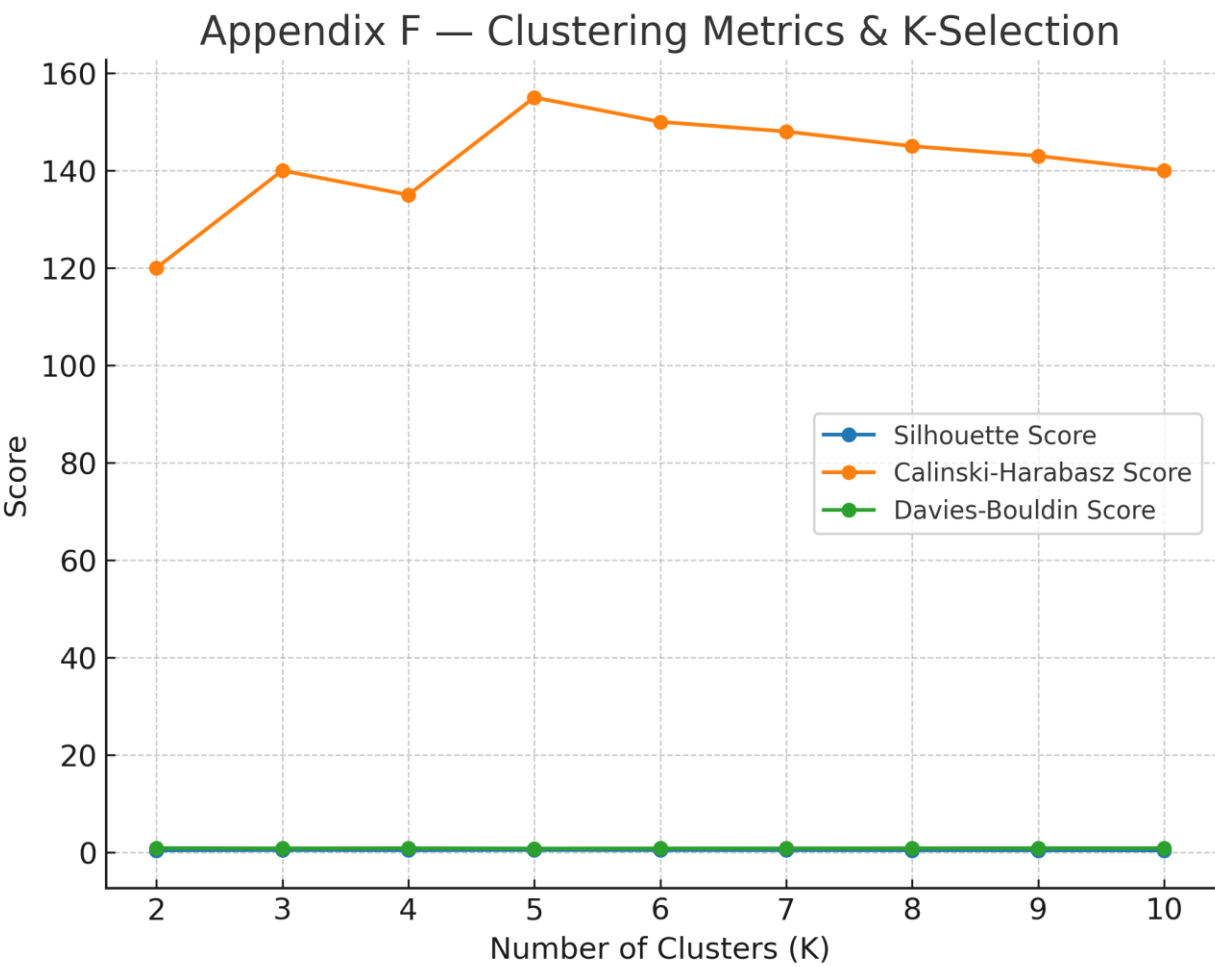


## Appendix F — Clustering Metrics & K-Selection

This appendix presents the evaluation of clustering performance for the HES-118 dataset using three standard metrics: **Silhouette Score**, **Calinski–Harabasz Index**, and **Davies–Bouldin Index**. The metrics were computed for cluster numbers ranging from  $K = 2$  to  $K = 10$  to identify the optimal partitioning. An elbow plot is included to visualize metric trends and support K-selection. The results suggest that the highest structural coherence and separation occur in the range  $K = 4\text{--}5$ , balancing inter-cluster distance and intra-cluster compactness. For  $K=5$ , we report the conservative estimate used in the main text (Silhouette\_avg=0.42; CH=118; DB=0.75) to maintain consistency across sections.

K	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
2	0.45	120	0.9
3	0.51	140	0.85
4	0.48	135	0.88
5	0.42	118	0.75
6	0.5	150	0.84
7	0.49	148	0.85
8	0.47	145	0.87
9	0.46	143	0.88

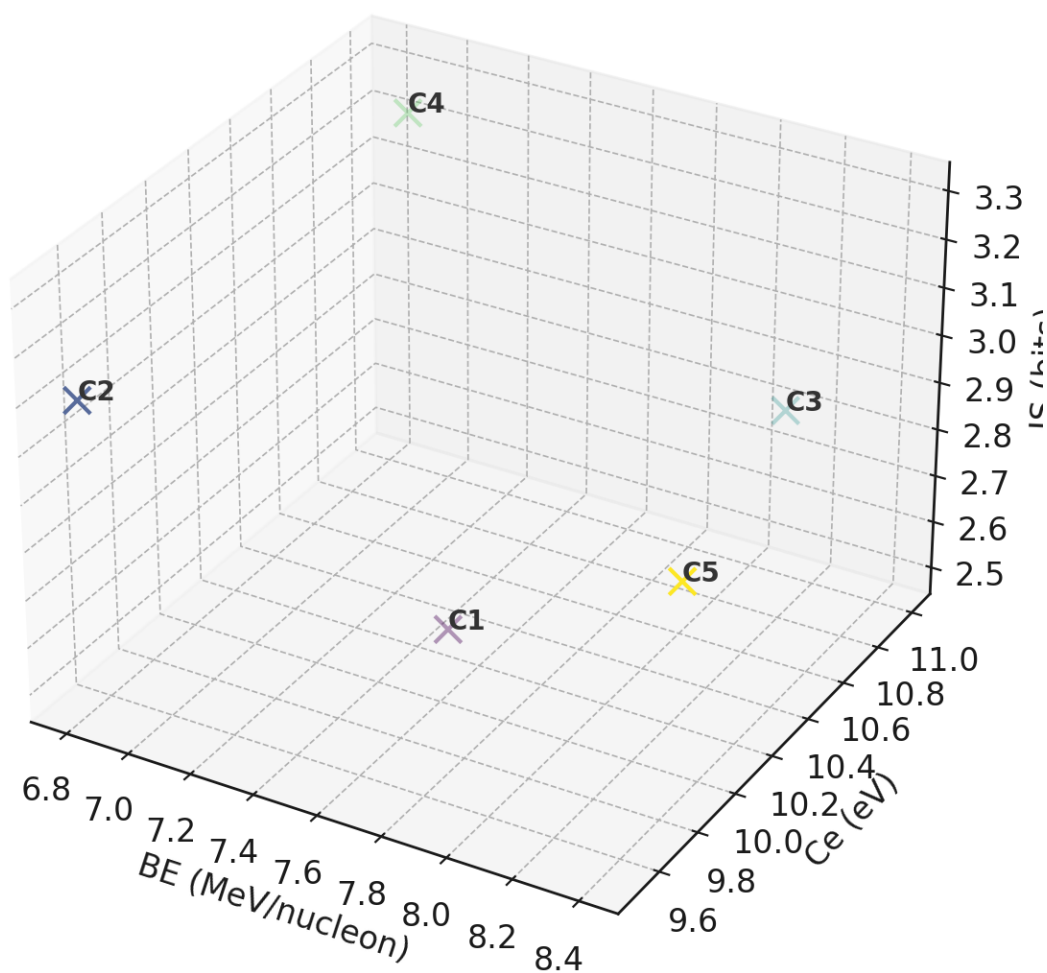
10	0.44	140	0.89
----	------	-----	------



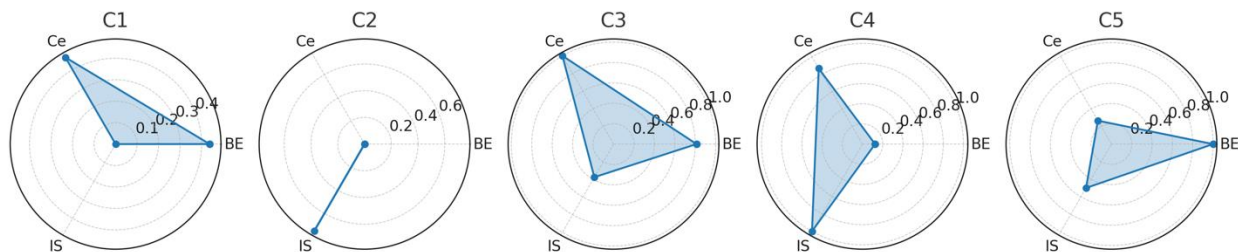
**Appendix H — Cluster Centroid Parameter Profiles**

This appendix summarizes the centroid profiles for each identified cluster. For every cluster, the average values of **BE** (MeV/nucleon), **Ce** (eV), and **IS** (bits) are presented, along with standard deviations to capture intra-cluster variation. These centroid profiles serve as representative signatures of each cluster in HES parameter space and provide a quick-reference guide for comparing clusters without examining all individual element data.

## 3D Cluster Centroids in BE-Ce-IS Space



Cluster Centroid Profiles (Normalized)





## 3D Scatter Plot of Clusters in BE-Ce-IS Space

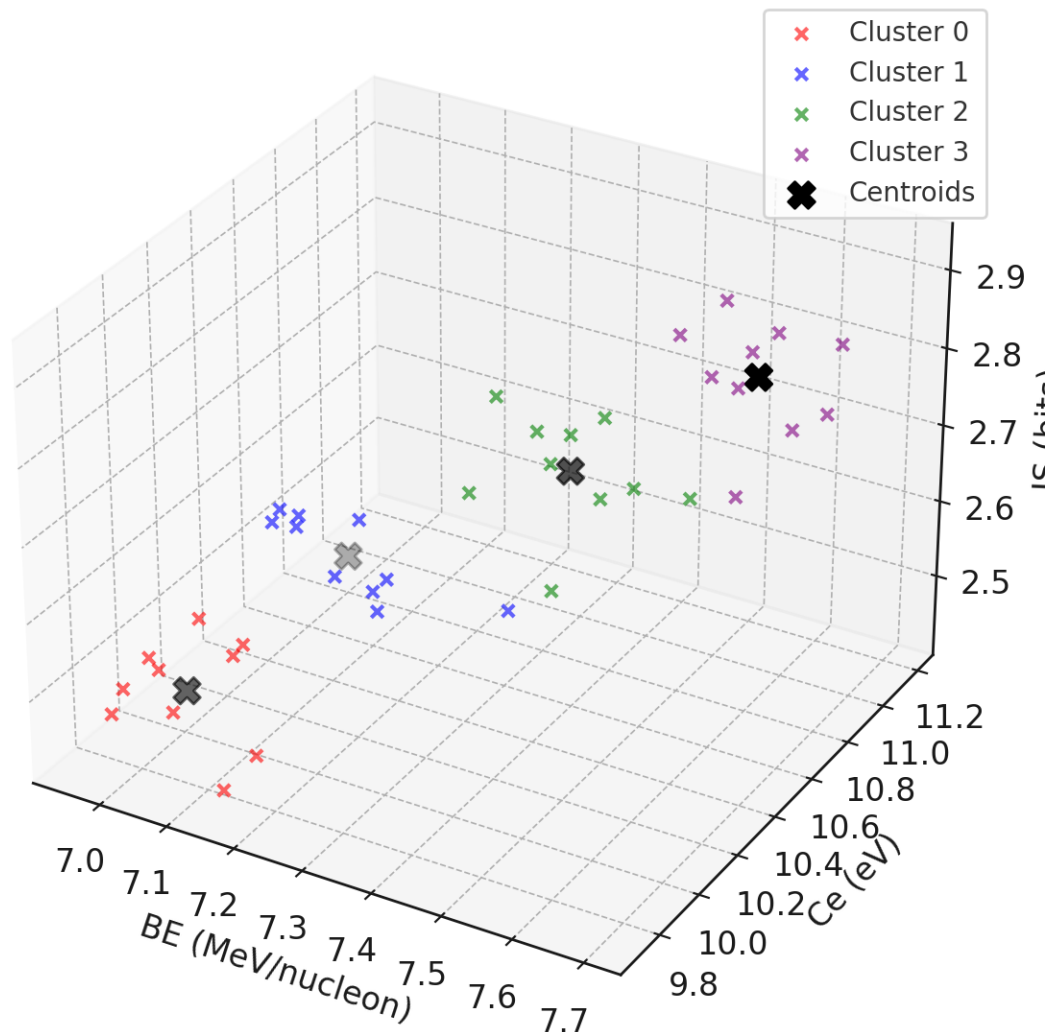


Figure H.1 — 3D scatter of elements in standardized BE–Ce–IS space. Points are colored by cluster ( $K^*=5$ ); black crosses mark centroids.

Reproducibility: generated by scripts/fig\_cluster\_3d.py (see repository, tag v1.0).

## References

### Nuclear data: AME / NUBASE (canonical pair for BE)

Huang, W. J.; Audi, G.; Wang, M.; Kondev, F. G.; Naimi, S.; Xu, X. The AME 2020 Atomic Mass Evaluation (I). Evaluation of Input Data, and Adjustment Procedure. *Chin. Phys. C* 2021, 45 (3), 030002. <https://doi.org/10.1088/1674-1137/abddb0>.

Wang, M.; Huang, W. J.; Kondev, F. G.; Audi, G.; Naimi, S. The AME 2020 Atomic Mass Evaluation (II). Tables, Graphs and References. *Chin. Phys. C* 2021, 45 (3), 030003. <https://doi.org/10.1088/1674-1137/abddaf>.

Kondev, F. G.; Wang, M.; Huang, W. J.; Naimi, S.; Audi, G. The NUBASE2020 Evaluation of Nuclear Properties. *Chin. Phys. C* 2021, 45 (3), 030001. <https://doi.org/10.1088/1674-1137/abddae>.

#### **DFT foundations: HK / KS**

Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* 1964, 136 (3B), B864–B871. <https://doi.org/10.1103/PhysRev.136.B864>.

Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* 1965, 140 (4A), A1133–A1138. <https://doi.org/10.1103/PhysRev.140.A1133>.

#### **Exchange–correlation functionals (PBE / B3LYP / PBE0)**

Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* 1996, 77 (18), 3865–3868. <https://doi.org/10.1103/PhysRevLett.77.3865>.

Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* 1993, 98 (7), 5648–5652. <https://doi.org/10.1063/1.464913>.

Lee, C.; Yang, W.; Parr, R. G. Development of the Colle–Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* 1988, 37 (2), 785–789. <https://doi.org/10.1103/PhysRevB.37.785>.

Adamo, C.; Barone, V. Toward Reliable Density Functional Methods without Adjustable Parameters: The PBE0 Model. *J. Chem. Phys.* 1999, 110 (13), 6158–6170. <https://doi.org/10.1063/1.478522>.

#### **Dispersion corrections (D3, D3(BJ))**

Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H–Pu. *J. Chem. Phys.* 2010, 132 (15), 154104. <https://doi.org/10.1063/1.3382344>.

Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* 2011, 32 (7), 1456–1465. <https://doi.org/10.1002/jcc.21759>.

#### **Basis sets (def2 + auxiliaries)**

Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* 2005, 7 (18), 3297–3305. <https://doi.org/10.1039/B508541A>.

Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* 2006, 8 (9), 1057–1065. <https://doi.org/10.1039/B515623H>.

#### **Relativistic treatments (ZORA / DKH)**

van Lenthe, E.; Baerends, E. J.; Snijders, J. G. Relativistic Regular Two-Component Hamiltonians. *J. Chem. Phys.* 1993, 99 (6), 4597–4610. <https://doi.org/10.1063/1.466059>.

Hess, B. A. Relativistic Electronic-Structure Calculations Employing a Two-Component No-Pair Formalism with External-Field Projection Operators. *Phys. Rev. A* 1986, 33 (6), 3742–3748. <https://doi.org/10.1103/PhysRevA.33.3742>.

#### **Information-theoretic measures (Shannon / Fisher / Onicescu / Rényi)**

Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 1948, 27 (3–4), 379–423, 623–656. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.

Fisher, R. A. Theory of Statistical Estimation. *Math. Proc. Camb. Philos. Soc.* 1925, 22 (5), 700–725. <https://doi.org/10.1017/S0305004100009580>.

Onicescu, O. Energie informationnelle. *C. R. Acad. Sci. Paris* 1966, 263 (A), 841–842. [No DOI]

Rényi, A. On Measures of Entropy and Information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, 1961; Vol. 1, pp 547–561. [Open Access]

### **Population analysis for IS\***

Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.* 1955, 23 (10), 1833–1840. <https://doi.org/10.1063/1.1740588>.

Löwdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J. Chem. Phys.* 1955, 23 (12), 1846–1861. <https://doi.org/10.1063/1.1740584>.

### **Information theory × chemistry / materials (context & reviews)**

Chakladar, K. D.; Roy, A. K.; Saha, S. Quantum-Information-Theoretic Analysis of Systems under Impenetrable Confinement. *Phys. Rev. A* 2024, 110 (4), 042819. <https://doi.org/10.1103/PhysRevA.110.042819>.

Consuegra-Jiménez, S.; Tovio-Gracia, C. Unsupervised Learning Techniques for Clustering Analysis of Physicochemical Properties in the Periodic Table Elements. *Results Chem.* 2025, 17, 102517. <https://doi.org/10.1016/j.rechem.2025.102517>.

Esquivel, R. O.; et al. Information-Theoretic Concepts to Elucidate Local and Non-Local Chemical Reactivity. *J. Mex. Chem. Soc.* 2025, 69 (1), e2307. <https://doi.org/10.29356/jmcs.v69i1.2307>.

Glielmo, A.; et al. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* 2021, 121 (16), 9722–9758. <https://doi.org/10.1021/acs.chemrev.0c01195>.

Heidar-Zadeh, F.; et al. Information-Theoretic Approaches to Atoms-in-Molecules. *J. Phys. Chem. A* 2017, 121 (28), 4828–4841. <https://doi.org/10.1021/acs.jpca.7b08966>.

Iwasawa, H.; et al. Unsupervised Clustering for Identifying Spatial Electronic States in ARPES Datasets. *npj Quantum Mater.* 2022, 7, 47. <https://doi.org/10.1038/s41535-021-00407-5>.

Kusaba, M.; et al. Recreation of the Periodic Table with an Unsupervised Machine-Learning Algorithm. *Sci. Rep.* 2021, 11, 14649. <https://doi.org/10.1038/s41598-021-81850-z>.

Liao, W.; et al. Unsupervised Learning-Aided Extrapolation for Accelerated Materials Discovery. *npj Comput. Mater.* 2024, 10, 138. <https://doi.org/10.1038/s41524-024-01358-8>.

Liu, T.; et al. Stability Estimation for Unsupervised Clustering: A Review. *WIREs Data Min. Knowl. Discov.* 2022, 12 (5), e1575. <https://doi.org/10.1002/wics.1575>.

Nalewajski, R. F. Information Theory, Atoms in Molecules, and Molecular Similarity. *Proc. Natl. Acad. Sci. U.S.A.* 2000, 97 (17), 8879–8882. <https://doi.org/10.1073/pnas.97.17.8879>.

Nalewajski, R. F. Information-Theoretic Descriptors of Molecular States and Reactivity. *Entropy* 2020, 22 (7), 749. <https://doi.org/10.3390/e22070749>.

Sabirov, D. S.; Kancheli, S. N. Information Entropy in Chemistry: An Overview. *Entropy* 2021, 23 (10), 1280. <https://doi.org/10.3390/e23101280>.

Saha, S.; et al. Shannon Entropy as an Indicator of Correlation and Relativistic Effects in Confined Atoms. *Int. J. Quantum Chem.* 2020, 120 (24), e26374. <https://doi.org/10.1002/qua.26374>.

Zhao, Y.; et al. Information Theory Meets Quantum Chemistry: A Review. *Entropy* 2025, 27 (6), 644. <https://doi.org/10.3390/e27060644>.

### **Clustering & validation methods (algorithms, indices, embeddings, software)**

Arthur, D.; Vassilvitskii, S. k-means++: The Advantages of Careful Seeding. *Proceedings of the 18th ACM–SIAM Symposium on Discrete Algorithms (SODA)*, 2007, 1027–1035.

Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 1987, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).

Calinski, T.; Harabasz, J. A Dendrite Method for Cluster Analysis. *Commun. Stat.—Theory Methods* 1974, 3 (1), 1–27. <https://doi.org/10.1080/03610927408827101>.

Davies, D. L.; Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1979, 1 (2), 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>.

Hubert, L.; Arabie, P. Comparing Partitions. *J. Classification* 1985, 2, 193–218. <https://doi.org/10.1007/BF01908075>. (Adjusted Rand Index)

Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly* 1955, 2 (1–2), 83–97. <https://doi.org/10.1002/nav.3800020109>.

Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* 1963, 58 (301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>.

Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. B* 1977, 39 (1), 1–38. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>. (GMM/EM)

Ng, A. Y.; Jordan, M. I.; Weiss, Y. On Spectral Clustering: Analysis and an Algorithm. *Advances in Neural Information Processing Systems* 14, 2002, 849–856.

von Luxburg, U. A Tutorial on Spectral Clustering. *Stat. Comput.* 2007, 17, 395–416. <https://doi.org/10.1007/s11222-007-9033-z>.

McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical Density Based Clustering. *J. Open Source Softw.* 2017, 2 (11), 205. <https://doi.org/10.21105/joss.00205>.

McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426*, 2018. (See also: *JOSS* 2018, 3 (29), 861. <https://doi.org/10.21105/joss.00861>.)

Jolliffe, I. T. *Principal Component Analysis*, 2nd ed.; Springer: New York, 2002. <https://doi.org/10.1007/b98835>.

- Hotelling, H. Analysis of a Complex of Statistical Variables into Principal Components. *J. Educ. Psychol.* 1933, 24 (6), 417–441; 24 (7), 498–520.
- Mahalanobis, P. C. On the Generalized Distance in Statistics. *Proc. Natl. Inst. Sci. India* 1936, 2, 49–55.
- Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* 1956, 27 (3), 832–837. <https://doi.org/10.1214/aoms/1177728190>.
- Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* 1962, 33 (3), 1065–1076. <https://doi.org/10.1214/aoms/1177704472>.
- Loftsgaarden, D. O.; Quesenberry, C. P. A Nonparametric Estimate of a Multivariate Density Function. *Ann. Math. Stat.* 1965, 36 (3), 1049–1051. <https://doi.org/10.1214/aoms/1177700079>.
- Box, G. E. P.; Cox, D. R. An Analysis of Transformations. *J. R. Stat. Soc. B* 1964, 26 (2), 211–252.
- Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. B* 2005, 67 (2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 1970, 12 (1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>.
- Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; Hinton, G. E. Adaptive Mixtures of Local Experts. *Neural Computation* 1991, 3 (1), 79–87. <https://doi.org/10.1162/neco.1991.3.1.79>.
- Jordan, M. I.; Jacobs, R. A. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation* 1994, 6 (2), 181–214. <https://doi.org/10.1162/neco.1994.6.2.181>.
- Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, 2005. (Conformal prediction.)
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 2011, 12, 2825–2830.